Abordagens Multimodais com Fusão de Dados em Aprendizado Profundo

Lucas de Souza Rodrigues

Abordagens Multimodais com Fusão de Dados em Aprendizado Profundo

Lucas de Souza Rodrigues

Orientador: *Prof. Dr. Edson Takashi Matsubara*

Tese apresentada à Faculdade de Computação - FACOM-UFMS como parte dos requisitos necessários à obtenção do título de Doutor em Ciência da Computação.

UFMS - Campo Grande
junho/2023

Dedicatória

Ao meu filho,
Heitor Passone Rodrigues,

À minha família.

Agradecimentos

Cada realização é resultado de um grande empenho e comprometimento! Ao longo desse período, tive a valiosa oportunidade de aprofundar meus conhecimentos e ampliar meus horizontes no campo da computação. Cada momento foi uma experiência enriquecedora, onde conheci professores comprometidos e dedicados em transmitir seu saber com entusiasmo.

Desejo expressar minha mais profunda gratidão pela concretização deste sonho, e, acima de tudo, agradecer à minha companheira zelosa. Ao longo dessa trajetória acadêmica, compartilhamos momentos significativos, enfrentamos desafios e crescemos juntos, tornando sua presença fundamental para o alcance desse sonho. Ao meu querido Heitor, meu amado filho que veio ao mundo para encher nossas vidas de alegria e renovar nossas forças constantemente na possibilidade de acreditar em um futuro melhor. Meu pai, Luis Carlos ou Giovani, conhecido por sua incansável dedicação em todas as suas empreitadas, é um exemplo de paternidade, sempre abdicando de tudo em prol do bem-estar e sucesso de seus filhos. Minha mãe, Laiz, uma mãe dedicada, uma inspiração de mulher, que sempre foi o alicerce da nossa família, trazendo amor e harmonia para todos nós. Às minhas queridas irmãs, Priscila e Fernanda, expresso minha gratidão pelo apoio incondicional ao longo deste período. Vocês são pessoas especiais e, sem dúvida alguma, as melhores irmãs que alguém poderia ter. Aos meus sogros, pelo suporte e não medirem esforços durante essa fase de estudos.

Os colegas do Laboratório de Inteligência Artificial (LIA), Pedro, Lucas, André, Edilene, Eliton, Edmar e Mauro, obrigado pelo apoio nos estudos, em especial, ao Kenzo que sempre me auxiliou nos experimentos e resultados, foram incansáveis noites de escrita de artigos e pesquisas. Gostaria de expressar meu profundo agradecimento à Faculdade de Computação (FACOM) por disponibilizar toda sua infraestrutura e recursos para o desenvolvimento deste trabalho. Também sou grato ao Instituto Federal de Mato Grosso do Sul (IFMS) pela concessão do afastamento necessário para viabilizar a realização

deste sonho.

Agradeço ao meu orientador professor Edson Takashi, pela recepção calorosa durante todo este período. Sua orientação, as reuniões e o conhecimento compartilhado foram inestimáveis, proporcionando-me uma visão impressionante sobre os diversos aspectos que a ciência pode oferecer. Sou imensamente grato por ter tido a oportunidade de trabalhar sob sua orientação e pela contribuição valiosa que você trouxe para o meu crescimento acadêmico. Ao professor Eraldo pelos preciosos momentos que tivemos juntos e por ter disponibilizado suas GPUS para os experimentos realizados neste trabalho.

Minha gratidão a Deus que tem sido meu amparo ao longo desta jornada, pois: "*O conselho da sabedoria é: Procure obter sabedoria, use tudo o que você possui para adquirir entendimento.*" (Provérbios 4:7). Por fim, meus sinceros agradecimentos a todas as pessoas que, de maneira direta ou indireta, contribuíram para o desenvolvimento deste trabalho.

Abstract

Deep neural networks, especially language and vision models, have been widely used in real problems in recent years. Usually models apply the use of only one type of data/information (text, image, video, audio) in learning problems, also called unimodal models. However, given the growing amount of unstructured information and the variety of existing data formats, new approaches have been developed with the aim of establishing strategies that enable the use of multiple data in the same learning model. This work explores data fusion in Multimodal Machine Learning (ML) models. The proposal of this thesis explores a simple strategy that uses mathematical operations to merge the different types of data between the layers of the multimodal architecture, mechanisms of attention and residual connections. Another proposal explores the use of multimodal knowledge distillation to optimize the performance of deep learning models, transferring knowledge between modalities of the same domain. The main advance of this work was to use arithmetic operations, attention mechanisms and residual connections in multimodal approaches with data fusion. This allowed obtaining complementary representations about the modalities, which led to a better convergence without significant difference with the state-of-the-art.

Keywords: data fusion, multimodal model, deep neural networks, language model, vision model, attention mechanisms.

Resumo

As redes neurais profundas, especialmente os modelos de linguagem e visão, têm sido amplamente utilizados em problemas reais nos últimos anos. Geralmente modelos aplicam o uso de apenas um tipo de dado/informação (texto, imagem, vídeo, áudio) em problemas de aprendizado, também chamados de modelos unimodais. No entanto, dada a quantidade crescente de informações não estruturadas e a variedade de formatos de dados existentes, novas abordagens têm sido desenvolvidas com o objetivo de estabelecer estratégias que viabilizem a utilização de múltiplos dados em um mesmo modelo de aprendizado. Este trabalho explora a fusão de dados em modelos de Aprendizado de Máquina Multimodal (AM). A proposta desta tese explora uma estratégia simples que utiliza operações matemáticas para fundir os diversos tipos de dados entre as camadas da arquitetura multimodal, mecanismos de atenção e conexões residuais. Uma outra proposta explora o uso da destilação de conhecimento multimodal para otimizar o desempenho de modelos de aprendizado profundo, transferindo conhecimento entre modalidades de um mesmo domínio. O principal avanço deste trabalho foi usar as operações aritméticas, mecanismos de atenção e conexões residuais em abordagens multimodais com a fusão de dados. Isso permitiu obter representações complementares sobre as modalidades, o que levou a uma melhor convergência sem diferença significativa com o estado-da-arte.

Palavras-chave: fusão de dados, modelo multimodal, redes neurais profundas, modelo de linguagem, modelo de visão, mecanismos de atenção.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO

1

Introdução

Este capítulo apresenta uma visão geral desta tese de doutorado juntamente com os objetivos de pesquisa produzido neste estudo.

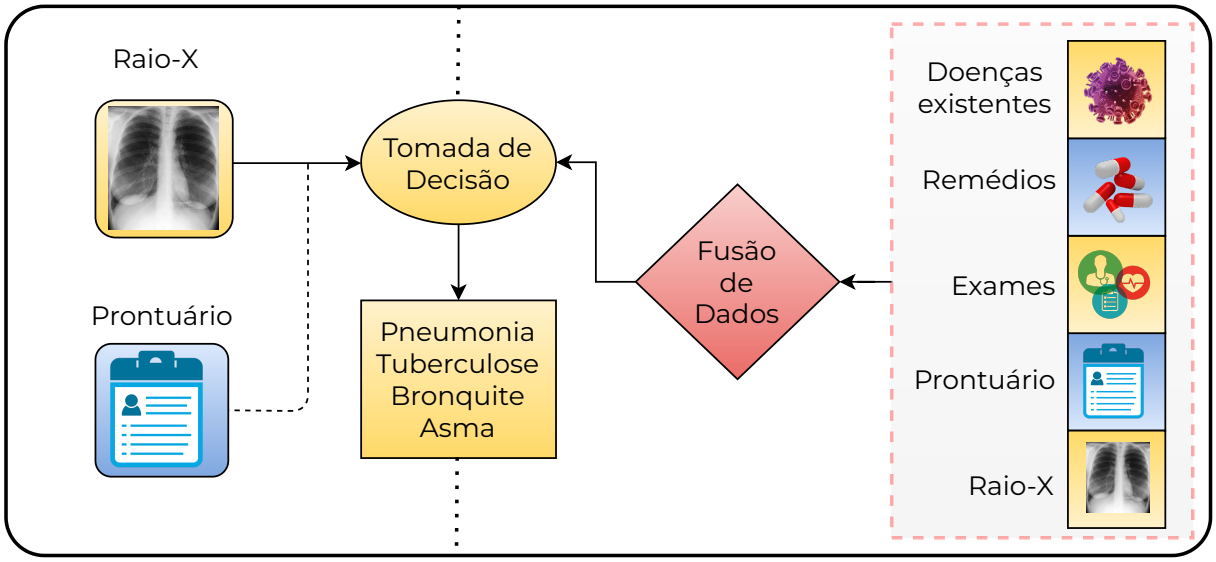
1.1 Contextualização

O Aprendizado de Máquina (AM) tradicional utiliza uma única modalidade para a indução de um modelo. No entanto, pesquisadores têm explorado novas abordagens no campo do (AM) para lidar com a presença de informações não estruturadas e frequentemente combinadas com diversas fontes de informação, conhecidas como modalidades, tais como texto, imagem, áudio e vídeo (Zhang et al., 2018a). O Aprendizado Multimodal, por exemplo, possibilita a utilização dessas modalidades em um único modelo de aprendizado (Gao et al., 2020). Segundo Ramachandram and Taylor (2017) as diferentes modalidades geram uma representação compartilhada que podem proporcionar ganhos de desempenho em comparação a uma única modalidade.

Neste trabalho o Aprendizado de Máquina Multimodal com foco em Redes Neurais Profundas (*Deep Learning*) (Wang et al., 2020d), explora o Processamento de Linguagem Natural (PLN) (Chowdhary and Chowdhary, 2020) e Visão Computacional (VC) (Sebe et al., 2005) por meio da fusão de dados. Essas áreas podem ser promissoras na resolução de vários problemas por meio de abordagens multimodais, muitas vezes superiores em relação as abordagens unimodais. Contudo, grande parte dos modelos de aprendizado fazem uso de uma única modalidade, subtilizando o potencial da multimodalidade dos dados disponíveis. A multimodalidade pode ser comparada a diferentes áreas do conhecimento que se complementam para uma compreensão mais

ampla. Assim como a combinação dos sentidos humanos (visão, olfato, paladar, audição, tato) é essencial para uma percepção completa do ambiente, a fusão de dados multimodais permite obter uma representação mais completa e abrangente dos dados (Krogh, 2008; Baltrušaitis et al., 2018).

Segundo Gao et al. (2020) a fusão de dados visa integrar os dados de diferentes modalidades de um mesmo domínio com o objetivo de compartilhar informações. Para ilustrar o funcionamento da fusão de dados descrita na Figura 1.1, suponha o seguinte problema: considere um conjunto de possíveis doenças respiratórias $D \in \{\text{Pneumonia, Tuberculose, Bronquite, Asma}\}$, na qual o objetivo é classificar uma das possíveis doenças por meio de uma análise clínica. Uma série de dados podem ser considerados nesta avaliação, porém normalmente o diagnóstico de um profissional ligado a área médica analisa apenas imagens de ressonância magnética e/ou um prontuário com informações do paciente que podem conter (nome, idade, altura, peso, endereço e telefone). Para resolver problemas cotidianos como este a fusão de dados por meio de uma abordagem multimodal pode ser eficaz quando há disponibilidade de várias fontes de dados em um mesmo domínio.



a) Diagnóstico sem Fusão de Dados

b) Diagnóstico com Fusão de Dados

Figura 1.1: Fusão de Dados.

Suponha agora que um profissional da área médica consiga relacionar as imagens de ressonância magnética juntamente com dados importantes obtidos previamente, como doenças hereditárias provenientes do paciente, uso de medicações, sintomas recentes e exames de anos anteriores. Logo, o profissional da área médica pode relacionar todas essas informações para a tomada de decisão.

Atualmente problemas relacionados a PLN e VC, utilizam arquiteturas de múltiplas camadas em redes neurais com a possibilidade de fundir informa-

ções sobre o mesmo contexto, mas com entradas de dados diferentes. Por exemplo, a combinação de texto e imagem pode fornecer uma compreensão mais completa de uma cena visual, incorporando tanto informações visuais quanto descritivas. Isso é especialmente relevante pois a contextualização dos dados é essencial para uma percepção mais precisa e abrangente do problema.

O presente trabalho introduz novas abordagens sobre o contexto do aprendizado multimodal, sobretudo o conceito de fusão de dados e suas técnicas no campo da IA.

1.2 Problema

A ideia central do aprendizado multimodal consiste na combinação de diferentes modalidades, uma forma já bem consolidada em aprendizado multimodal é a concatenação (Guo et al., 2019; Gao et al., 2020; Kiela et al., 2020; Wang et al., 2020c). Entretanto, até o presente momento, há poucos estudos na literatura além da concatenação para o aprendizado multimodal. Embora existam diversas abordagens que exploram a fusão de dados em diferentes estágios e pontos de uma rede neural (Tian et al., 2019; Lobantsev et al., 2020; Bakkali et al., 2020; Wang et al., 2021b), ainda há espaço para examinar o estudo de técnicas de fusão mais avançadas. O uso de operações aritméticas para combinar as modalidades ou mecanismos de atenção para capturar as relações entre as modalidades são exemplos promissores a serem investigados nesse contexto. Na busca pelo avanço neste tema central em aprendizado multimodal, este trabalho irá explorar as seguintes ideias:

1. A concatenação tem sido amplamente explorada como o método predominante na fusão de modalidades no contexto de redes neurais profundas. No entanto, outras operações têm sido encontradas em arquiteturas de redes neurais, como a soma (*Skip Connection, Positional Encoding*) e a multiplicação (Mecanismo de Atenção). Diante desse cenário, quais os benefícios as operações aritméticas podem proporcionar na fusão entre as modalidades? Seria possível obter ganhos ao empregar operadores aritméticos na fusão de dados multimodais?

2. Os mecanismos de atenção realizam a multiplicação das entradas "consulta" (*query*) e "chave" (*key*) para obter ativações que buscam identificar combinações significativas de palavras. Nesse contexto, surge a questão sobre a aplicabilidade dos mecanismos de atenção em combinações entre as modalidades. Similarmente, é possível explorar o potencial dos mecanismos de atenção para identificar as combinações entre as modalidades?

3. As Conexões Residuais (*Skip Connections*) são descritas na literatura por sua capacidade de melhorar a convergência de modelos em aprendizado profundo. No entanto, esse benefício se estende também às redes multimodais? Será que a aplicação de *Skip Connections* em redes multimodais pode proporcionar melhorias semelhantes, facilitando a convergência do modelo?

4. Na destilação de conhecimento, um modelo maior é utilizado para auxiliar no treinamento de um modelo menor. Essa abordagem tem se mostrado eficaz em transferir o conhecimento adquirido pelo modelo maior para o modelo menor. No contexto multimodal é possível aplicar esse mecanismo, onde um modelo de uma modalidade auxilia de forma complementar o treinamento de outra modalidade?

1.3 Objetivo

O objetivo desta tese de doutorado é aprimorar o estudo dos operadores de fusão multimodal. Portanto, propomos as seguintes hipóteses a serem investigadas:

Hipótese 1 - As operações aritméticas em abordagens multimodais são promissoras para a fusão dos dados em relação a outras técnicas tradicionais já difundidas na literatura.

Hipótese 2 - Em um cenário com diversos operadores de fusão de dados sem um resultado claramente superior, a utilização dos mecanismos de atenção pode contribuir para a combinação desses operadores e possibilitar a obtenção de um melhor desempenho.

Hipótese 3 - A utilização de *Skip Connections* em modelos multimodais resulta em uma melhor convergência dos modelos, semelhante ao que é observado em modelos unimodais na literatura.

Hipótese 4 - Modelos de Aprendizado baseados em Destilação Multimodal resultam em uma transferência de conhecimento entre modalidades, com potencial de melhorar a capacidade de generalização entre as diversas modalidades.

1.4 Contribuições e Organização do Trabalho

A organização do trabalho e as contribuições são descritas a seguir:

1. O Capítulo 2 aborda noções fundamentais sobre os modelos de aprendizado e explora os principais conceitos adotados em trabalhos relacionados a redes multimodais, bem como os diferentes tipos de fusão de dados. Há também uma descrição das técnicas e métodos empregados em conjunto as redes multimodais, como os mecanismos de atenção, conexões residuais, destilação de conhecimento multimodal e algumas aplicações multimodais presentes na literatura.

2. No Capítulo 3 foi desenvolvido dois experimentos para mesclar dados textuais e visuais, possibilitando a construção de uma representação conjunta entre camadas de entrada recebidas de modelos unimodais. Também foi adotado um conjunto de operações aritméticas para avaliar não só o desempenho da fusão de dados, mas também a relação existente entre a junção de múltiplas fontes. Além disso, dez mecanismos de atenção foram empiricamente avaliados em vários conjuntos de dados multimodais. Outras técnicas como a normalização de dados em algumas camadas da rede e o uso das redes de compreensão e excitação incorporados a modelos de última geração são exemplos de métodos que foram usados para aperfeiçoar a fusão de dados.

3. No Capítulo 4 este estudo se dedicou a investigar novas direções na área da Destilação de Conhecimento Multimodal, uma técnica que permite a fusão de informações de distintas modalidades. Ao utilizar uma estratégia simples de transferência de conhecimento obtido por meio de modelos pré-treinados para modelos mais compactos e eficientes, a destilação multimodal se apresenta como uma opção promissora em domínios com ampla variedade de dados.

4. O Capítulo 5 descreve um caso de estudo, na qual as redes multimodais demonstram a capacidade de potencializar os resultados de predição e adquirir uma representação conjunta ao integrar mais de um modelo de aprendizado em uma rede multimodal. Os experimentos conduzidos nesta pesquisa evidenciam que a utilização dos modelos multimodais e a fusão de dados em determinados domínios apresentam resultados promissores.

5. A conclusão deste trabalho é abordada no Capítulo 6, destacando algumas das contribuições desta pesquisa e a exploração dos objetivos mencionados na Seção 1.3.

CAPÍTULO

2

Procedimentos e Conceitos Fundamentais

Neste capítulo são apresentados os procedimentos e conceitos fundamentais relacionados ao Aprendizado Multimodal adotados neste trabalho. São discutidos os Modelos de Aprendizado, incluindo as áreas de linguagem e visão, bem como algumas métricas de avaliação. Além disso, são abordados os conceitos fundamentais utilizados para a implementação das Redes Multimodais. Na Figura 2.1 é apresentada a linha de estudo envolvendo o uso de redes neurais multimodais, técnicas e aplicações para a resolução de tarefas em alguns domínios, descritas no Capítulo 3. Por fim, uma organização da pesquisa é elaborada para nortear os objetivos deste trabalho na Seção 2.8.

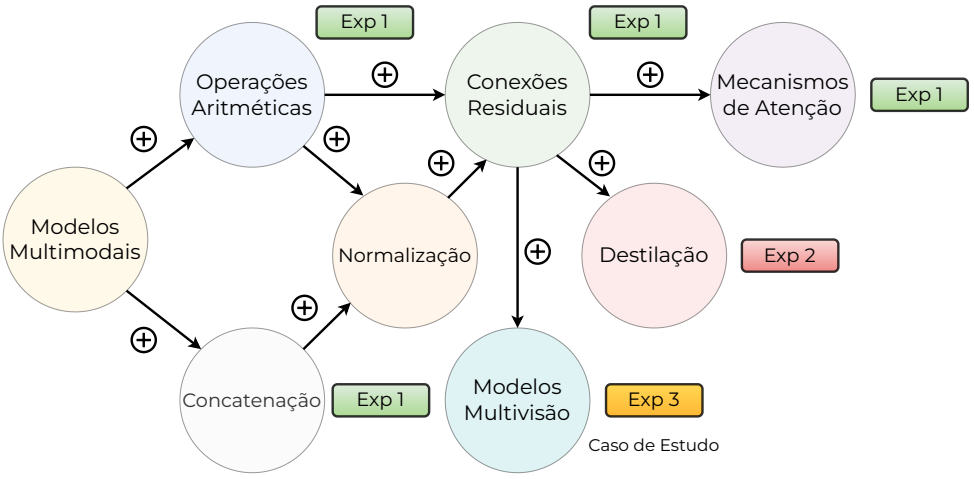


Figura 2.1: Visão geral da linha de estudo, onde "Exp" indica os experimentos realizados neste trabalho.

2.1 Redes Multimodais

Conjuntos de dados multimodais são coleções de dados que contêm informações de diferentes modalidades ou fontes, como planilhas, imagens, textos, áudios e vídeos. Esses conjuntos são utilizados em abordagens multimodais, onde técnicas de aprendizado são aplicadas para explorar a interação e a complementaridade entre as diferentes modalidades. Um das primeiras técnicas relacionadas a abordagens multimodais foi o uso de métodos *ensembles* (Dietterich, 2000), em que consistia em combinar as previsões de vários classificadores individuais, chamados de membros do *ensemble*, para chegar a uma previsão final mais confiável e precisa. Neste contexto, o aprendizado profundo multimodal, consiste em uma série de técnicas que buscam unificar várias modalidades de dados, descrita também como fusão de dados ou multimodal (Gao et al., 2020).

O uso da fusão de dados cobre diversos domínios e permite a combinação de várias fontes de dados, conforme ilustrado pela Figura 2.2. Uma das vantagens apresentadas no trabalho de Ramachandram and Taylor (2017) é que o uso de várias fontes de dados em redes neurais profundas possibilita aprender hierarquicamente representações com um controle refinado sobre os dados. Uma prática no aprendizado profundo multimodal é construir uma camada de representação mesclando as várias fontes de dados, de modo que a rede compreenda uma representação conjunta de suas entradas. Uma série de trabalhos foram propostos na literatura para lidar com as redes multimodais (Atrey et al., 2010; Khaleghi et al., 2013; Yang et al., 2017; Williams et al., 2018; Yu et al., 2020), os quais descrevem o processo da fusão multimodal, o uso de correlação e independência, nível de confiança, informações contextuais, sincronização entre diferentes modalidades e seleção das fontes de dados.

A literatura destaca uma série de combinações que podem ser representadas na fusão multimodal, a Tabela 2.1 descreve de forma simplificada os aplicativos e fontes correlacionadas para o uso do aprendizado profundo multimodal extraído do trabalho (Guo et al., 2019).

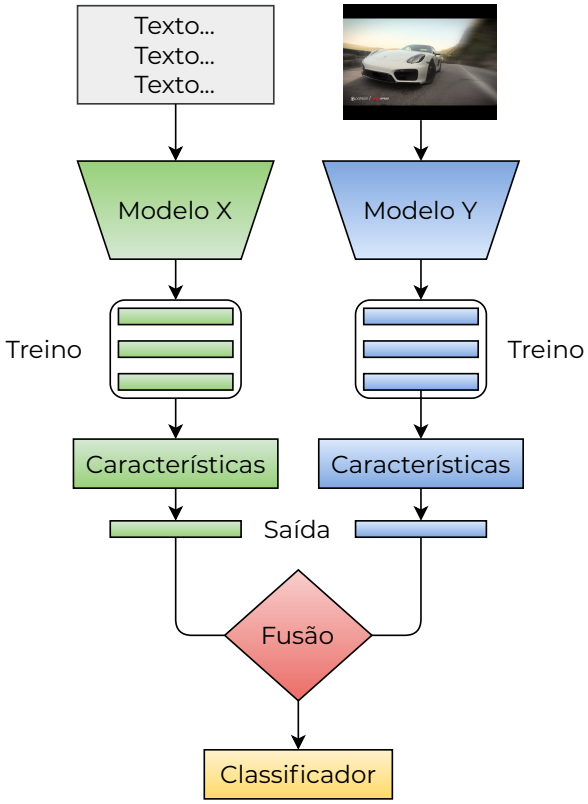
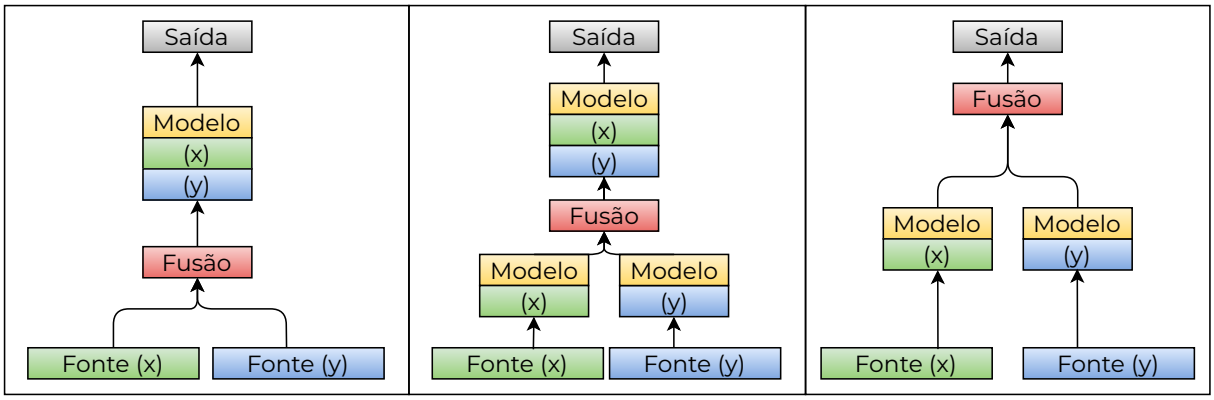


Figura 2.2: Visão geral do funcionamento de uma Rede Multimodal com Fusão Tardia.

Tabela 2.1: Resumo das aplicações e a relação com as fontes de texto, imagem, áudio e vídeo.

Aplicações	**Fontes**
Legenda da Imagem	Imagem, Vídeo
Classificação de Vídeo	Áudio, Vídeo, Texto
Descrição de Vídeo	Vídeo, Texto
Texto para Síntese de Imagem	Texto, Imagem
Detecção de eventos	Áudio, Vídeo, Texto
Análise de sentimento	Áudio, Vídeo, Texto
Resposta Visual a Perguntas	Imagem, Texto
Reconhecimento de Emoções	Áudio, Vídeo, Texto
Reconhecimento de Fala	Áudio, Vídeo
Transferência de Aprendizado	Imagem, Texto

A Figura 2.3 ilustra os três tipos de fusão multimodal populares descrito em diversos trabalhos (Fusão Precoce, Intermediária ou Tardia), os quais serão objeto de discussão na próxima Seção 2.2.



(a) Fusão Precoce (b) Fusão Intermediária (c) Fusão Tardia

Figura 2.3: Tipos de Fusão Multimodal.

2.2 Tipos de Fusão

2.2.1 Fusão Precoce

A fusão precoce é uma estratégia usada para fundir as diversas fontes a nível de dados na entrada de uma rede neural, seguida pela aplicação de um único modelo de aprendizado. É um processo que envolve a integração de várias fontes, diferentes em tamanho e escala em um único vetor de recursos. Logo, caso os dados brutos não sejam pré-processados e enviados diretamente a rede neural, torna-se uma tarefa desafiadora extrair as características em um modelo multimodal. O pré-processamento é fundamental na fusão precoce, pois algumas fontes de dados podem não estar disponíveis em alguns domínios ou a taxa de amostragem entre as fontes pode variar. Outro aspecto envolvido na extração dos dados sem o pré-processamento é que algumas informações inseridas diretamente na camada de entrada de um modelo multimodal pode produzir dados discretos, enquanto outras dados contínuos. Logo, pode haver um problema de discrepância na representação e na manipulação dos dados.

Para resolver questões relacionadas a entrada de dados representações de nível superior podem ser extraídas na primeira camada em modelo multimodal com fusão precoce, também chamadas de "*handcrafted features*" ou recursos manuais (cor, textura, grafia, representação textual, segmentação de áudio e vídeo) (Nanni et al., 2017). Normalmente é um conjunto de características recomendado por um especialista, que posteriormente são fundidas para camada subsequente.

A concatenação é a forma mais comum em fundir os dados em uma rede neural profunda resultante de várias fontes de dados. Na fusão precoce essa operação pode criar vetores de entrada esparsos com diversas redundâncias (Ramachandram and Taylor, 2017). A redução de dimensionalidade é uma das

estratégias descritas na literatura para resolver redundâncias em dados brutos, os estudos de Wang et al. (2015a) e Masci et al. (2013) estendem esse conceito para criar um espaço incorporado com representações de dados multimodais.

A maior parte dos modelos de fusão precoce parte da premissa de que existe independência condicional entre as fontes de dados (Ramachandram and Taylor, 2017). Contudo, a maioria dos domínios disponíveis para abordagens com fusão, possuem dados altamente correlacionados. De acordo com Sebe et al. (2005) a correlação possibilita a convergência de informações provenientes de diversas fontes para transmitir uma representação de alto nível. Outro problema enfrentado pela fusão precoce é lidar com o sincronismo dos dados ao longo do tempo entre as diferentes fontes. A literatura prevê um série de mecanismos que podem contornar este problema por meio de re-amostragem dos dados para integrar séries de eventos discretos com sinais contínuos (Martínez and Yannakakis, 2014).

Um exemplo clássico de fusão precoce é encontrado no estudo de Poria et al. (2015) com uso de vetores de recursos combinados com as modalidades textuais, visuais e de áudio para treinar um classificador baseado em aprendizado de vários núcleos/*kernels* para análise de sentimento. Outro estudo publicado por Williams et al. (2018) também demonstra o uso da fusão precoce em dois contextos, o primeiro sem uso de redução de dimensionalidade e o segundo com redução por meio da Análise de Componentes Principais ou *Principal Component Analysis* (PCA) (Wold et al., 1987), um procedimento matemático que propõe converter um conjunto de observações possivelmente correlacionadas em um conjunto com variáveis linearmente não correlacionadas, chamadas de componentes principais, conforme ilustra a Figura 2.4.

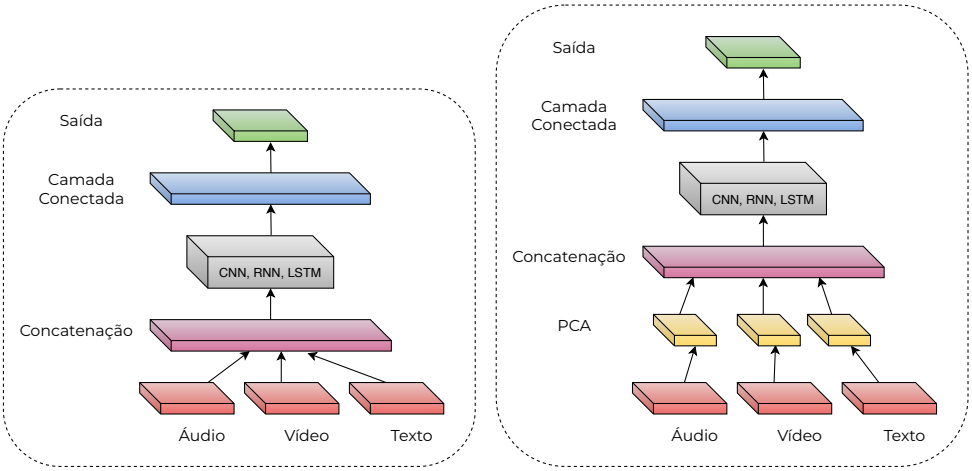


Figura 2.4: Fusão Precoce com e sem redução de dimensionalidade. Adaptado de Williams et al. (2018).

2.2.2 Fusão Intermediária

A fusão intermediária está presente na maioria dos trabalhos que utilizam redes neurais profundas multimodais, pois compreende o uso de modelos unimodais que possuem o melhor desempenho para cada entrada de dados específica (texto, áudio, imagem, vídeo) em suas camadas iniciais. Em seguida, as camadas subsequentes utilizam as representações intermediárias compartilhada pelas camadas iniciais, a fim de concatenar tais representações na etapa de treinamento nas camadas seguintes por meio de um modelo multimodal.

Uma proposta para fusão intermediária é demostrada pelo trabalho de Williams et al. (2018), descrito na Seção 2.2.1. Sua abordagem é utilizar os pesos intermediários a partir dos modelos unimodais inseridos nas camadas iniciais e então fundi-los em camadas totalmente conectadas, conforme ilustrado na Figura 2.5.

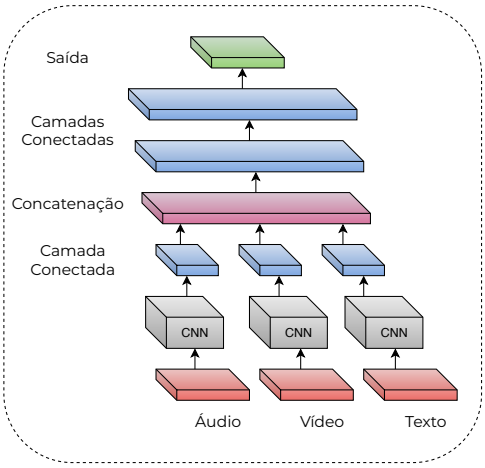


Figura 2.5: Fusão Intermediária. Adaptado de Williams et al. (2018).

A camada de representação compartilhada pode ser compreendida por uma única camada ou uma série de camadas gradualmente fundidas ao longo da rede neural. No entanto, uma simples concatenação de recursos ou pesos pela camada de representação compartilhada sem um critério moderador, pode ocasionar erro na interpretação da rede ao aprender as relações entre os modelos unimodais ou até *overfitting* no conjunto de treinamento (Ramachandram and Taylor, 2017). Uma das soluções utilizadas para resolver problemas de aprendizado em redes multimodais é o uso de camadas adicionais *autoencoders*, uma técnica que permite treinar a rede para reduzir a dimensionalidade e ignorar o ruído do sinal entre as camadas. Outra alternativa é o uso do PCA referenciado na fusão precoce, Seção 2.2.1.

Contudo, uma das vantagens no uso da fusão intermediária é a flexibilidade em fundir várias fontes de dados em camadas compartilhas ao longo da rede, também chamada de fusão lenta, conforme ilustrado pelo Figura 2.6.

Essa flexibilidade permite obter resultados significativos para certas tarefas, na qual há um processo temporal entre os dados, por exemplo no trabalho de Karpathy et al. (2014), que utiliza abordagens empíricas para classificação de vídeo em grande escala, utilizando um conjunto de dados com um milhão de vídeos da rede social *YouTube* (Burgess, 2011), relacionadas a 487 classes. Sua abordagem foi estender a conectividade de uma rede CNN no domínio do tempo para obter vantagens em relação ao espaço-temporal das informações.

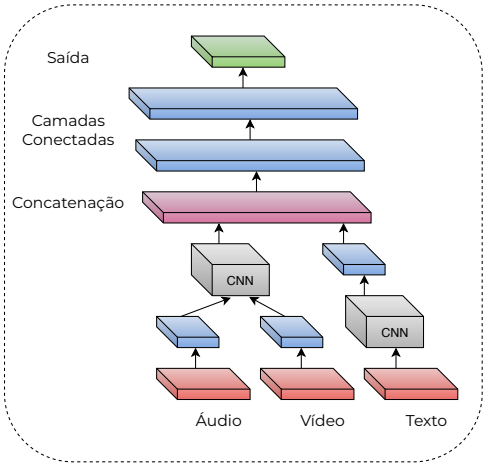


Figura 2.6: Fusão Lenta. Adaptado de Williams et al. (2018).

Um modelo de última geração que utiliza a fusão intermediária é o *ShuffleNet* (Zhang et al., 2018b), projetado especialmente para dispositivos móveis com capacidade limitada em recursos computacionais. Esta arquitetura usa convoluções em grupo e troca de canais para reduzir o custo computacional, conforme ilustrado na Figura 2.7.

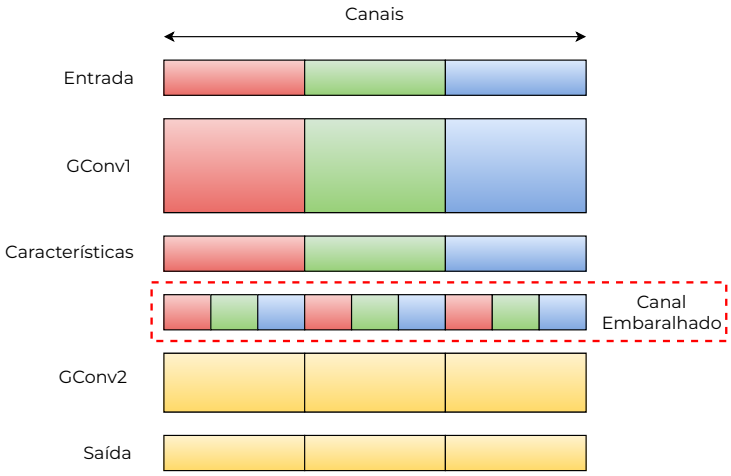


Figura 2.7: Arquitetura *ShuffleNet*. Adaptado de Zhang et al. (2018b).

Quando há convolução em grupo a informação é bloqueada porque as saídas de um determinado grupo se relacionam apenas entre si. Para resolver esse problema é adicionado uma operação de *shuffling* de canais, assim as

informações são repassadas a diferentes grupos na camada de convolução. Logo, o conceito principal da arquitetura *ShuffleNet* é aplicar a convolução em grupo juntamente com operações de *shuffling* de canal, pois há melhora de desempenho quando o modelo compõe uma rede menor com mapas de recursos mais finos.

Para o uso da fusão intermediária é necessário analisar quais as fontes de dados podem ser concatenadas neste processo, pois requer uma estrutura representativa dos dados, com abordagens específicas para cada domínio.

2.2.3 Fusão Tardia

A fusão tardia, também chamada nível de decisão, compreende o uso de treinamento em redes unimodais e uma fusão multimodal nas últimas camadas do modelo. Logo, requer múltiplos estágios de treinamento sem interações de baixo nível entre as fontes de dados com compartilhamento das representações unimodais. Conforme estudo de Ramachandram and Taylor (2017) esta arquitetura possui uma vantagem em relação a outros mecanismos, porque erros de múltiplas redes unimodais representada por seus classificadores tendem a não estar correlacionados, e o método é independente de características.

A fusão tardia possui várias estratégias para que decisões de diferentes classificadores sejam combinadas, conforme descrito na Seção 2.2.1, a concatenação é a forma mais comum em relacionar duas redes unimodais, porém há inúmeros estudos com abordagens empíricas para realizar esta tarefa, desde uma combinação matemática (máximo, mínimo, média) até métodos como a regra de *Bayes* ou votação majoritária, na qual um classificador analisa as saídas das redes unimodais e define as relações predominantes na maioria das redes (Atrey et al., 2010). Esse processo ocorre por que diferentemente da fusão precoce, uma arquitetura de decisão final já possui informações de erro e classificação de ambas redes unimodais.

De acordo com Dashtipour et al. (2021) a vantagem em uma arquitetura com fusão tardia é obter predições pelos modelos unimodais com mesma taxa de amostragem ou dimensões equivalentes, logo as predições extraídas podem ser facilmente concatenadas sem aumento de amostragem ou redução de dimensionalidade. Conforme ilustra a Figura 2.8 uma possível abordagem para fusão tardia pode ser representada por três entradas individuais (áudio, vídeo e texto) concatenadas no nível de decisão e uma entrada direta com dados tabulados por exemplo variáveis extraídas de uma planilha.

Há uma série de trabalhos de última geração que utilizam fusão tardia no aprendizado de redes multimodais, pois quando as fontes de dados não são correlacionadas, com dimensionalidade e taxas de amostragem diferentes, o

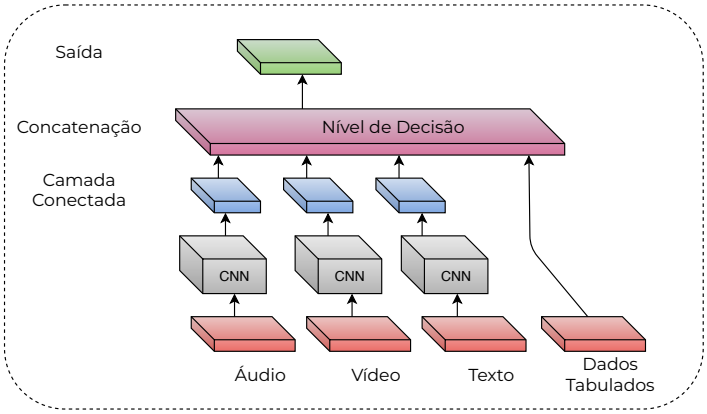


Figura 2.8: Fusão Tardia. Adaptado de Williams et al. (2018).

uso da fusão tardia simplifica o processo e a implementação da rede neural. No entanto, em seu estudo Ramachandram and Taylor (2017), afirma não ver evidências conclusivas de que a fusão tardia seja melhor que a fusão precoce, pois o desempenho das redes multimodais dependem do problema e domínio da aplicação.

2.3 *Skip Connection*

Skip Connection, também chamada de conexões residuais, é uma técnica de conexão direta entre duas camadas de uma rede neural e permite que as informações ignorem uma ou mais camadas (He et al., 2016; Adaloglou, 2020). Portanto, não é necessário percorrer todas as camadas intermediárias, isso só é possível com a inclusão de uma conexão entre a entrada de uma camada e a saída de uma camada posterior, que pode ser alocada entre as várias camadas à frente na rede neural. A Figura 2.9 ilustra este conceito em um modelo codificador-decodificador para segmentação de imagens, também chamado de modelo *U-Nets* (Ronneberger et al., 2015).

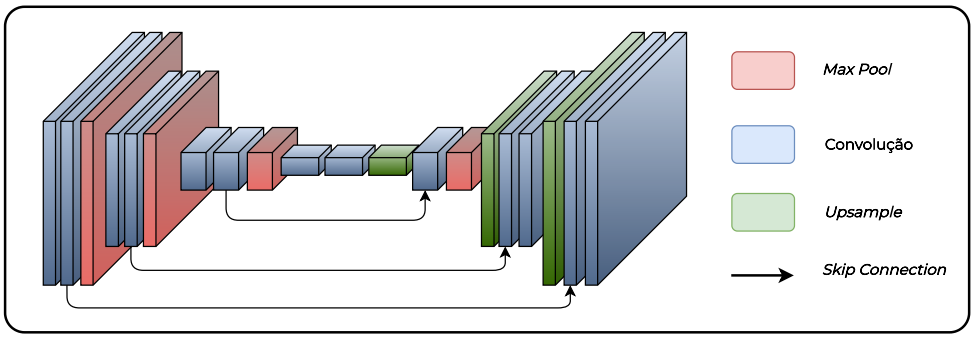


Figura 2.9: Arquitetura *U-Net* com uso de *Skip Connection*.

Sua importância em uma rede neural é que à medida que mais camadas são adicionadas em um modelo, as informações tendem a se perder entre as camadas. Esse problema afeta o gradiente usado para ajustar os pe-

sos da rede, pois seu valor diminui conforme a informação é propagada pela rede, uma disfunção chamada degradação do gradiente (Grosse, 2017; Hanin, 2018). A consequência para o ajuste do gradiente é que as camadas iniciais na etapa de retropropagação podem não ser atualizadas com informações relevantes para fazer os ajustes significativos nos pesos, gerando uma perda de desempenho na rede neural.

Outro fator relevante é que a preservação da informação é um aspecto crítico que justifica o uso do *Skip Connection*, pois possibilita que a informação seja enviada diretamente para as camadas intermediárias ou finais, permitindo ajustes significativos na atualização dos pesos e na rede neural como um todo (Liu et al., 2020). Seu uso é comum em arquiteturas de redes neurais profundas, como as utilizadas em tarefas de processamento de imagem, texto e áudio.

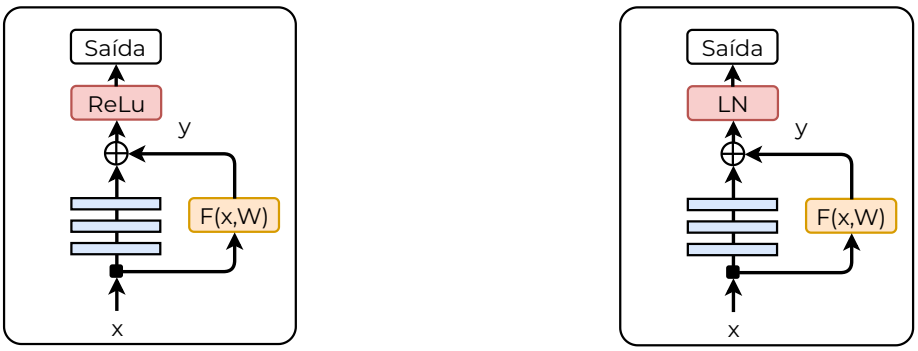
As conexões residuais foram inicialmente empregadas em arquiteturas de redes neurais profundas, como a *ResNet* (He et al., 2016), permitindo que as camadas posteriores aprendessem como transformar a entrada em um espaço de recursos mais apropriado, em vez de tentar reconstruir a entrada original. No trabalho de He et al. (2016) as conexões residuais são descritas em blocos, também chamados de blocos básicos de construção de redes. Assim, um bloco residual pode conter duas ou mais camadas convolucionais, seguidas por uma conexão residual. Esse conceito permite que as informações avancem diretamente pelo bloco, evitando a degradação do gradiente.

A contribuição do *Skip Connection* também pode auxiliar redes neurais com conexões de atenção, já utilizadas na arquitetura *Transformer*
(Vaswani et al., 2017). Pois permitem que a rede concentre sua atenção em diferentes partes da entrada durante a fase de processamento, portanto, usar *Skip Connection* garante que informações importantes sejam retidas e possam ser usadas por camadas posteriores da rede (Schneider and Vlachos, 2023). A Figura 2.10 ilustra duas alternativas para lidar com o problema de otimização nos modelos *ResNet* e *Transformer*, explorando *Skip Connection* juntamente com outros métodos de normalização. Liu et al. (2020) em seu trabalho descreve a formulação ilustrada por meio da equação Equação 2.1:

$$y = G(\lambda x + F(x, W)), \tag{2.1}$$

onde x representa a entrada do bloco junto com *Skip Connection*, F realiza a transformação não linear induzida pela rede neural parametrizada por W, G representa a função de normalização, y denota a saída do bloco residual e λ denota o fator de modulação que controla a importância relativa do *Skip Connection*.

Essa abordagem tem sido amplamente utilizada em tarefas de reconheci-



(a) ResNet. (b) Transformer.

Figura 2.10: Representação do *Skip Connection*: (a) Unidade Residual Convencional (He et al., 2016), (b) Unidade Residual em *Transformers* (LN = Normalização de Camadas) (Vaswani et al., 2017). Adaptado de Liu et al. (2020) .

mento de imagem e processamento de linguagem natural. Liu et al. (2020) propõe em seu trabalho uma forma adaptativa de ajustar a escala de entrada aplicando recursivamente o *Skip Connection* com normalização de camadas, seus resultados indicam um aumento substancial de desempenho e eficácia na generalização de diversas tarefas. Xu et al. (2021) menciona em seus estudos com *Graph Neural Networks* (GNNs), que o treinamento de GNNs é implicitamente acelerado por *Skip Connection* com resultados promissores em termos de otimização.

2.4 Mecanismos de Atenção

Aplicar mecanismos de atenção em redes neurais para resolução de tarefas é uma técnica usada atualmente nos modelos de aprendizado. A atenção permite encontrar correlações entre diferentes entradas de um problema, possui a capacidade de concentrar-se em informações relevantes e processa-las seletivamente dentro da rede neural (Vaswani et al., 2017; Niu et al., 2021). Sua aplicação prática pode ser encontrada nas Redes Neurais Recorrentes (RNN) (Wang and Tax, 2016) e Redes Neurais Convolucionais (CNN) (Tian et al., 2020; Zhu et al., 2021), pois melhora significativamente a eficiência e a precisão (*Accuracy*) no processamento de informações perceptivas (Xu et al., 2015). Os mecanismos de atenção mais comuns foram implementados para resolução de problemas em PLN (Hu, 2019) como tradução de texto, análise de sentimento, questionários com perguntas e respostas e classificação de texto. Porém novos estudos propagaram seu conceito para outras áreas da ciência, com diversas propostas para área da visão computacional (Sun et al., 2020). Segundo estudo de Guo et al. (2022) as tarefas que se destacam neste segmento são classificação de imagens, detecção de objetos, segmentação semântica, reconhecimento facial, processamento de imagem médica e tarefa

multimodal.

Na seção a seguir é apresentada uma breve descrição de dez mecanismos de atenção, utilizados em pesquisas avançadas, com uma descrição resumida dos tipos de atenção empregados, estrutura e operações envolvidas em cada mecanismo. O termo "bloco" foi adotado para descrever o componente/módulo de cada mecanismo de atenção, há também abreviações e termos específicos da linguagem Python (Lutz, 2013) e biblioteca Pytorch (Paszke et al., 2019b) para as operações de agrupamento, média, normalização e funções de ativação.

Módulos de Atenção

1 - SE-Net - *Squeeze-and-Excitation Networks* (Hu et al., 2018) inclui um bloco adicional na construção de CNNs tradicionais, que propõe melhorar as interdependências dos canais, conforme ilustra a Figura 2.11(1). Sua proposta é que para qualquer camada de uma rede neural convolucional é possível construir um bloco SE correspondente que recalibra os mapas de recursos. Essa abordagem foi utilizada no desafio de classificação de imagens (ILSVRC 2017 Russakovsky et al. (2015)) e obteve um ganho de 25% em relação ao modelo campeão em 2016. Dentro das operações de seu bloco há uma dimensão espacial que é aplicada por uma operação *AdaptiveAvgPool*, em seguida a atenção de canal é aprendida por meio de duas camadas conectadas. Para normalizar os dados no canal é aplicado uma função sigmoide e finalmente o mapa de atenção do canal é combinado com o entrada original para obter os recursos ponderados.

2 - CBAM - *Convolutional Block Attention Module* é um mecanismo que usa atenção de canal e espacial ao mesmo tempo e correlaciona as duas atenções em série (Woo et al., 2018). Em termos estruturais é semelhante ao mecanismo *Squeeze-and-Excitation* (SE) *Attention*, seu diferencial está no uso das operações de agrupamento *AvgPool* e *MaxPool* com diferentes efeitos de representação. Os autores propõe executar *AvgPool* e *MaxPool* nos recursos originais da dimensão espacial, em seguida usar a representação do mecanismo SE para extrair a atenção de canal. A Equação 2.2 descreve resumidamente as operações usadas nas duas atenções:

$$\begin{aligned} FeatureMap\,(F) &= F \in \mathbb{R}^{HxWxC} \\ ChannelAttention\,(CA) &= M_{ca}(F) \in \mathbb{R}^{1x1xC} \\ SpacialAttention\,(SA) &= M_{sa}(F) \in \mathbb{R}^{HxWx1} \end{aligned} \tag{2.2}$$

onde H, W, C representam largura, altura e quantidade de canais respectivamente, deste modo CBAM infere sequencialmente um mapa de atenção de canal 1D e um mapa de atenção espacial 2D, durante a etapa de multipli-

cação os valores de atenção do canal são transmitidos ao longo da dimensão espacial. Assim a saída obtida pelo mapa de recursos após as operações são $F' = M_{ca}(F) * F$ para o mapa de recursos refinado do canal de atenção e $F'' = M_{sa}(F') * F'$ para atenção espacial. A Figura 2.11(2) representa as duas atenções utilizadas no CBAM, os parâmetros são compartilhados entre as duas operações *AvgPool* e *MaxPool*, em seguida uma convolução é usada para extrair a atenção espacial e finalmente uma normalização é efetuada para obter a matriz de atenção.

3 - CooAtt - *Coordinate Attention* incorpora informações posicionais no canal de atenção, descrita pelos pesquisadores de "Atenção Coordenada" (Hou et al., 2021). Sua estrutura fatora a atenção de canal em dois processos de codificação de recurso 1D que agregam informações ao longo das duas direções espaciais. Diferente de outros mecanismos de atenção o CooAtt não usa apenas um vetor de recurso por meio de agrupamento global 2D. Seu mecanismo por meio da fatoração paralela no canal de atenção permite integrar efetivamente informações de coordenadas espaciais nos mapas de atenção gerados. Inicialmente esse mecanismo foi proposto para modelos clássicos usados em redes móveis *MobileNetV2* (Sandler et al., 2018), *MobileNeXt* (Daquan et al., 2007) e *EfficientNet* (Koonce, 2021a) com foco em localizar e reconhecer com mais precisão os objetos de interesse. A Figura 2.11(3) ilustra um bloco CooAtt com um mapa de características de entrada, em seguida uma fatoração para duas coordenadas (Horizontal = $X\ GAP$ e Vertical = $Y\ GAP$) são realizadas ao longo do canal, há uma codificação separada para cada par de mapas de atenção sensíveis à direção e posição que podem ser aplicados ao mapa de recursos de entrada para melhorar as representações dos objetos. A atenção coordenada é produzida pelas operações de concatenação e uma convolução 2D respectivamente. Finalmente depois da normalização e função de ativação, ela é segmentada em dois mapas de características com recursos de direção espacial, seguida por uma operação de convolução 2D e uma etapa residual é adicionada na saída da rede neural.

4 - Eca-Net - *Efficient Channel Attention* (Wang et al., 2020a) foi implementado para otimizar a atenção entre os canais de entrada com redução no número de parâmetros e cálculos, sua proposta é baseada no *Squeeze-and-Excitation* (SE) *Attention* porém ao invés de usar duas camadas totalmente conectadas para obter a atenção de canal, sua estratégia requer apenas uma convolução unidimensional para interação local, chamado de *crosschannel*, sem redução de dimensionalidade conforme ilustra a Figura 2.11(4). Uma operação de média global é executada para extrair as características dos recursos originais da dimensão de entrada, em seguida há uma ponderação de pesos por meio de uma convolução 1D de tamanho k (*kernel*). Os auto-

res afirmam que mecanismos mais sofisticados aumentam a complexidade do modelo, assim sua proposta implementa um método para selecionar o tamanho do *kernel* (k) da convolução 1D de forma adaptativa para determinar a cobertura da interação local entre os canais.

5 - PNA - *Parallel Network Attention* (Goyal et al., 2021) é um mecanismo que propõe usar sub-redes paralelas em vez de empilhar uma camada após a outra. Os autores afirmam que o uso em larga escala da profundidade em redes neurais acarreta em mais computação sequencial e maior latência, seus estudos evidenciam que usar uma rede profunda com apenas 12 camadas é capaz de atingir uma precisão superior a 80% no conjunto de dados *ImageNet* Deng et al. (2009), 96% no CIFAR10 e 81% no CIFAR100 (Krizhevsky et al., 2009). Um bloco PNA ilustrado na Figura 2.11(5) consiste em três ramificações paralelas: convolução 1×1, convolução 3×3 e uma camada chamada *Skip-Squeeze-and-Excitation* (SSE). Os autores propõe usar blocos parecidos com o modelo *VGG* (Simonyan and Zisserman, 2014), pois afirmam reduzir a latência durante a inferência dos dados. O uso de várias ramificações é feito por blocos de convolução 3×3 durante o treinamento, em seguida são fundidos em uma nova convolução 3×3, consistindo de apenas um bloco 3×3 e não linear. O SSE é aplicado ao lado de uma conexão de salto e usa uma única camada totalmente conectada, sua proposta é aumentar o campo receptivo uma vez que o uso de uma rede profunda com apenas 3×3 convoluções é bastante limitado.

6 - PSA - *Polarized Self-Attention* (Liu et al., 2021a) propõe resolver problemas quando há *pixel-wise regression*, ou seja, uma abordagem para prever o valor de cada pixel de uma imagem de acordo com as informações presentes nos *pixels* vizinhos. Uma bloco PSA pode ser dividido em duas etapas representado pela Figura 2.11(6), a primeira consiste em fazer uma filtragem polarizada com alta resolução interna no cálculo da atenção de canal e espacial, assim reduz significativamente o tamanho da entrada ao longo de suas dimensões. A segunda é uma composição de não linearidade para ajustar à distribuição de saída do canal para delinear pontos-chaves em mapas de calor e máscaras de segmentação. A polarização ocorre pela fusão da composição entre as operações *softmax* e sigmoide na atenção de canal e atenção espacial.

7 - RAN - *Residual Attention Network* (Wang et al., 2017) é um bloco que possui módulos de atenção e pode ser incorporado em diversas redes com arquitetura *feed forward* de última geração. A estratégia em usar RAN é empilhar os módulos de atenção que geram recursos de reconhecimento por meio do aprendizado residual. Um módulo de atenção é dividido em duas ramificações chamadas de ("*mask branch*") e ("*trunk branch*"). Conforme a Figura 2.11(7) três hiperparâmetros são usados para a configuração de um bloco

RAN, primeiro p é o número de unidades residuais de pré-processamento antes da divisão em ramificação (*mask* e *trunk*), t descreve o número de unidades residuais na ramificação do "*trunk*" ou tronco, e r o número de unidades residuais entre a camada de agrupamento adjacente na ramificação da máscara. Um bloco RAN padrão possui três estágios de repetição, sua natureza incremental em rede empilhada permite gradualmente refinar o mecanismo de atenção.

8 - S2Att - *Spatial-Shift Attention* (Yu et al., 2021) também denominado *S2-MLP*^{2} é um mecanismo de atenção que propõe melhorar o *backbone* de visão de seu antecessor *S2-MLP* (Yu et al., 2022), sua ideia central é expandir o mapa de recursos ao longo da dimensão do canal, em seguida dividir o mapa de recursos expandido em várias partes ilustrado na Figura 2.11(8). A construção do seu bloco possui diferentes operações de deslocamento espacial em partes divididas, que são fundidas novamente por operações de atenção. Diferente do *backbone S2-MLP* o *S2Att* adota *patches* de menor escala e uma estrutura piramidal para aumentar a precisão do reconhecimento de imagem.

9 - SHA - *Shuffle Attention* (Zhang and Yang, 2021) é um mecanismo que usa estratégias de embaralhamento para combinar módulos de atenção, conforme ilustra a Figura 2.11(9). A estratégia adotada é divida em duas etapas, a primeira canaliza dimensões em vários sub-recursos antes de processá-los em paralelo. Em seguida cada sub-recurso utiliza um módulo de atenção por meio da unidade de embaralhamento para extrair dependências de recursos em dimensões espaciais e de canal. Depois destas duas etapas todos os sub-recursos são fundidos e um operador, chamado "*channel shuffle*", que é selecionado para que a comunicação de informações entre os diferentes sub-recursos se relacionem.

10 - TripleAtt - *Triplet Attention* (Misra et al., 2021) é uma estratégia que usa métodos para calcular pesos de atenção capturando a interação entre as dimensões dos canais por meio de uma estrutura de três ramificações, descrito na Figura 2.11(10). Os autores afirmam ser um mecanismo mais leve e eficaz que outros módulos de atenção. Sua ideia principal é aplicar uma atenção tripla nos canais de entrada, esse tipo de atenção extrai dependências interdimensionais por meio de operações de rotação seguida por transformações residuais, por fim há uma codificação das informações entre intercanais e espaciais.

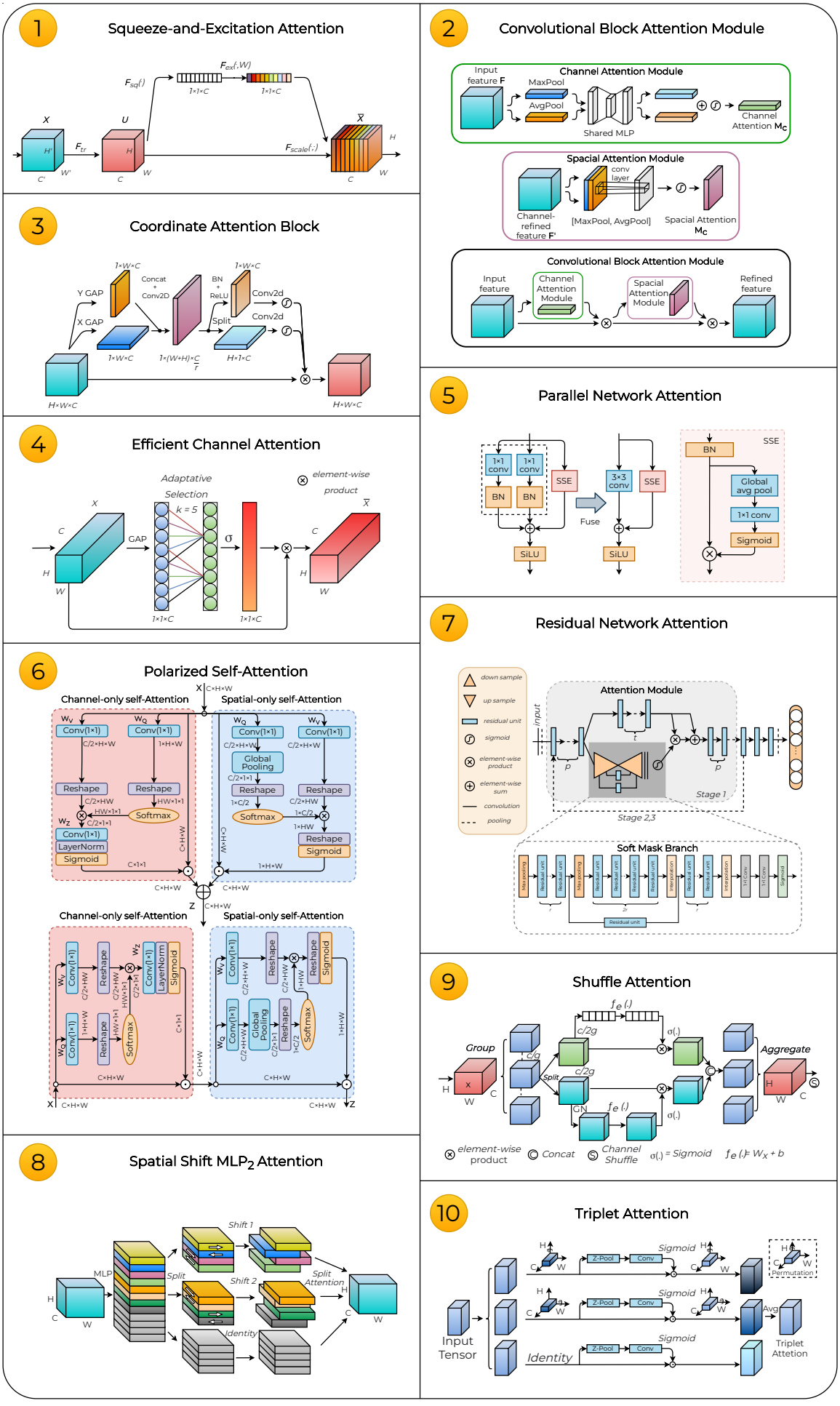


Figura 2.11: Estrutura dos Mecanismos de Atenção.

2.5 Destilação do Conhecimento Multimodal

Destilação de conhecimento multimodal, também chamado *Multimodal Knowledge Distillation* (MKD), é uma técnica para otimizar o desempenho de modelos de aprendizado profundo, transferindo conhecimento de várias modalidades ou fontes (Wang et al., 2020b). Nesse método o conhecimento de várias fontes, como imagens, áudio, texto e vídeo, é destilado em um único modelo, permitindo fazer previsões com base em informações de várias modalidades. Essa abordagem é baseada na destilação do conhecimento tradicional (KD), em que há um processo para compactar o conhecimento de um modelo complexo (grande) em um modelo mais simples (pequeno) (Xue et al., 2021). No MKD, esse processo é aplicado a várias modalidades, permitindo que o modelo aprenda com várias fontes de informação, conforme ilustrado em Figura 2.12.

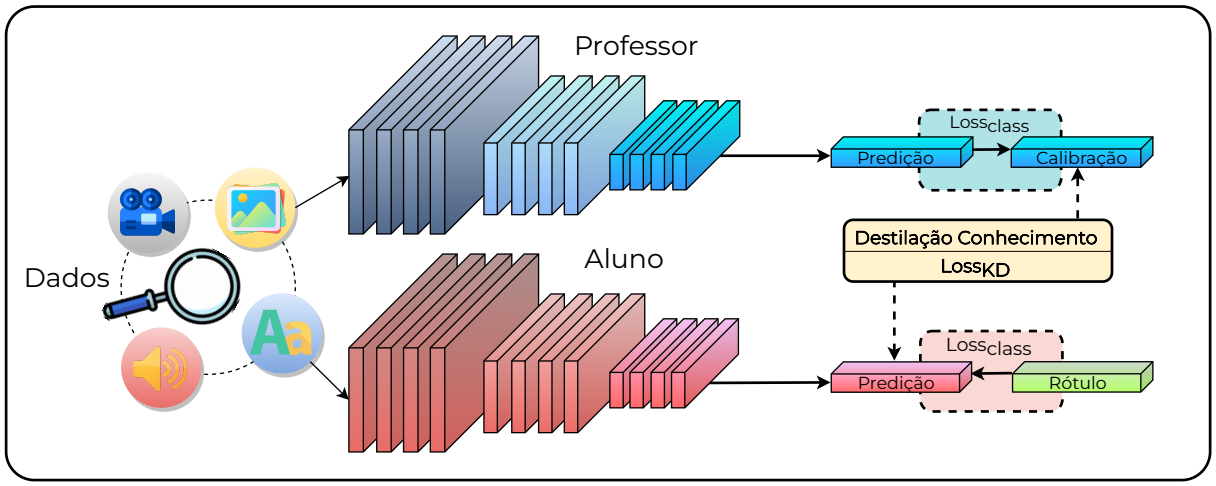


Figura 2.12: Arquitetura *Multimodal Knowledge Distillation* (MKD).

A principal vantagem do MKD é ajustar o treinamento do modelo aluno com outro tipo de informação presente no mesmo domínio de aplicação por diferentes categorias e técnicas de destilação (Wang et al., 2020b). Ao combinar o conhecimento de várias modalidades o modelo pode fazer previsões mais refinadas, pois tem acesso a uma gama mais ampla de informações para ajustar seu modelo (Dou et al., 2020). Isso pode levar a melhores resultados em uma variedade de aplicações, algumas das quais já estão presentes na destilação de conhecimento tradicional, incluindo classificação de imagem (Tung and Mori, 2019; Park et al., 2019; Peng et al., 2019a; Zhang et al., 2022), segmentação de imagem (Mullapudi et al., 2019), classificação de vídeo (Pan et al., 2020; Hu et al., 2020; Liyuan et al., 2021), reconhecimento de fala (You et al., 2021) e processamento de linguagem natural (Sun et al., 2019).

O processo da técnica MKD pode ser dividido em várias etapas: (i) primeiramente, os dados de cada modalidade são processados e transformados em uma representação adequada para o modelo; (ii) em seguida, as representa-

ções de cada modalidade são treinadas individualmente em modelos unimodais. Existe também a possibilidade de obter modelos pré-treinados para cada modalidade e realizar um ajuste fino para que seja possível realizar a etapa de destilação nas últimas camadas (Gou et al., 2021). Ao selecionar e treinar os modelos unimodais, a destilação do conhecimento pode ser combinada entre o modelo professor e o modelo aluno, transferindo o conhecimento adquirido pelo modelo professor para o modelo aluno por meio de diferentes modalidades (Wang et al., 2020b; Xue et al., 2021, 2022); (iii) Finalmente, por meio desse processo de otimização, uma representação combinada é inserida no modelo de predição, também chamado de modelo destilado.

Tabela 2.2: Visão geral de trabalhos publicados para KD e MKD.

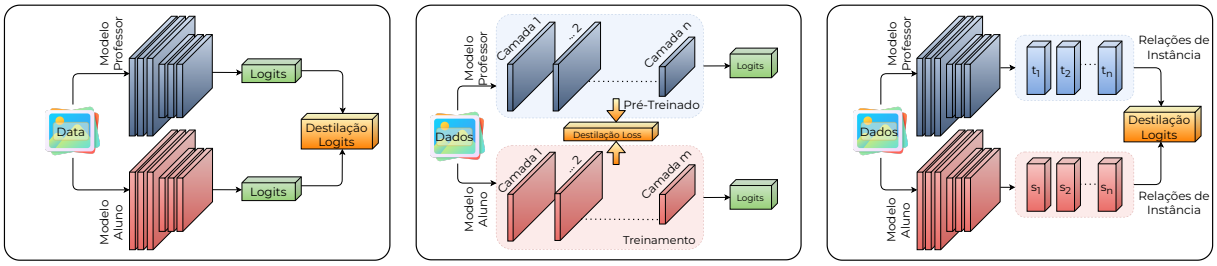
Artigo	**Problema**	**Modalidade**				**Destilação**	
		Texto	**Imagem**	**Áudio**	**Vídeo**	**Unimodal**	**Multimodal**
(Garcia et al., 2019)	Reconhecimento de Ação		✓			✓	
(Hu et al., 2020)	Segmentação de Imagem		✓			✓	
(Dai et al., 2022)	Visão-Linguagem (VLP)	✓	✓				✓
(You et al., 2021)	Resposta a Perguntas Faladas	✓		✓			✓
(Li et al., 2018a)	Compreensão Auditiva	✓		✓			✓
(Rao et al., 2021)	Visão-Linguagem (VLP)	✓	✓				✓
(Zhang et al., 2022)	Sumarização Multimodal	✓	✓				✓
(Liao et al., 2018)	Sistemas de Diálogo	✓	✓				✓
(Wang et al., 2020b)	Aprendizado Multimodal	✓	✓				✓
(Liyuan et al., 2021)	Reconhecimento de *Streamer*		✓		✓	✓	

A Tabela 2.2 resume trabalhos semelhantes que usam MKD ou referências ao KD tradicional como linha de base (*baseline*). Vários estudos apontam a usabilidade do MKD como uma técnica para otimizar modelos de aprendizado em redes neurais. Liu et al. (2021b) propõe uma nova estrutura chamada KD *Vision-and-language pretraining* (KD-VLP), que melhora a eficiência da visão e pré-treinamento de linguagem incorporando detecção de objeto e transferência de conhecimento de objeto em modelos VLP. You et al. (2021) descreve um novo método chamado *Multi-modal Residual Knowledge Distillation* (MRD) para treinar modelos de resposta a perguntas faladas usando destilação de conhecimento multimodal e aprendizado residual. Os autores demonstram por meio de experimentos no conjunto de dados *Spoken SQuAD* (Li et al., 2018a) que seu modelo MRD-Net supera as técnicas de última geração por uma margem significativa. Rao et al. (2021) desenvolveu um procedimento de treinamento em dois estágios, onde um grande modelo de visão e linguagem (professor) é primeiro pré-treinado em um conjunto de dados de grande escala e, em seguida, é usado para destilar o conhecimento para um modelo menor (aluno) para recuperação *cross-modal*. O modelo do aluno é treinado para imitar a saída do modelo do professor em um conjunto de dados menor. Os autores mostram que o modelo aluno destilado pode alcançar desempenho competitivo em vários *benchmarks* de recuperação *cross-modal*, sendo muito menor e

mais rápido que o modelo professor original.

Métodos de Destilação de Conhecimento

Três categorias de conhecimento foram estabelecidas no trabalho de Gou et al. (2021): Conhecimento Baseado em Resposta ("*Response-Based Knowledge*"), Conhecimento Baseado em Recurso "*Resource-Based Knowledge*"), e Conhecimento Baseado em Relacionamento ("*Relationship-Based Knowledge*").



(a) *Response-Based.* (b) *Feature-Based.* (c) *Relation-Based.*

Figura 2.13: Métodos de Destilação de Conhecimento. Adaptado de Gou et al. (2021).

Response-Based Knowledge - A camada de saída final do modelo do professor é o foco do conhecimento baseado em resposta, com o objetivo de fazer com que o modelo do aluno imite suas previsões. Isso é obtido por meio de uma função de perda de destilação, que calcula a diferença entre os *logits* dos modelos do aluno e professor (Ba and Caruana, 2014; Hinton et al., 2015), conforme ilustrado por Figura 2.13a. O modelo do aluno melhora suas habilidades de previsão, pois a perda de destilação é minimizada durante o treinamento.

Para problemas de visão computacional, como classificação de imagens, alvos flexíveis incorporam conhecimento baseado em resposta (Thoker and Gall, 2019; Hu et al., 2020; Li et al., 2020). Esses alvos são distribuições de probabilidade sobre as classes de saída, geradas usando a função *softmax*, descrita na Equação 2.3, Equação 2.4 (Gou et al., 2021):

$$p(z_i, T) = \frac{exp(z_i/T)}{\Sigma_j exp(z_j/T)} \tag{2.3}$$

$$KD_{Response}(p(z_t, T), p(z_s, T)) = \mathscr{L}_R(p(z_t, T), p(z_s, T)), \tag{2.4}$$

onde z_i é o *logit* para a i-ésima classe, e um fator de temperatura T é introduzido para controlar a importância de cada alvo suave. O parâmetro de temperatura modula a contribuição de cada "classe" para o conhecimento. Assim, a destilação de conhecimento baseada em resposta é frequentemente empregada no aprendizado supervisionado. Reescrevendo a perda de destilação em

termos do *logits* suaves, $(p(z_t,T)$ é *logit* para o modelo professor e $(p(z_s,T)$ é *logit* para o modelo aluno, onde a função de perda mais comum $\mathscr{L}_R(.)$ usado neste tipo de situação é a divergência *Kullback Leibler* (Joyce, 2011).

Avanços recentes no conhecimento baseado em resposta permitiram que os modelos de aprendizado profundo utilizassem com mais eficiência as informações ricas contidas nos rótulos fundamentais com alvos condicionais (Meng et al., 2019). Hinton et al. (2015) propõe destilar o "conhecimento obscuro" aprendido pelo modelo aluno em relação ao modelo professor, usando uma abordagem baseada em resposta para avaliar a qualidade das previsões do modelo do aluno. Wang et al. (2021a) também apresenta um novo método de destilação baseado em resposta para tradução automática neural, no qual o modelo aluno é treinado para produzir a mesma saída que o modelo professor para uma determinada entrada.

Feature-Based Knowledge - Em redes neurais profundas um modelo professor também possui conhecimento dos dados em suas camadas intermediárias. Essas camadas são treinadas para diferenciar recursos específicos e esse conhecimento pode ser empregado para treinar um modelo aluno (Yim et al., 2017; Heo et al., 2019). O objetivo, conforme demonstrado na Figura 2.13b, é ensinar o modelo aluno a capturar as mesmas ativações de recursos que o modelo professor. A função de perda de destilação minimiza a diferença entre as ativações de recursos dos modelos professor e aluno para adquirir conhecimento das camadas intermediárias. Normalmente, a perda de destilação para transferência de conhecimento baseada em recursos é representada pela Equação 2.5 (Gou et al., 2021):

$$KD_{Feature}(f_t(x),f_s(x)) = \mathscr{L}_F(\Phi_t(f_t(x)),\Phi_s(f_s(x))), \tag{2.5}$$

onde os mapas de recursos das camadas intermediárias dos modelos professor e aluno são denotados como $f_t(x)$ e $f_s(x)$, respectivamente. As funções de transformação $\Phi_t(f_t(x))$ e $\Phi_s(f_s(x))$ são empregadas quando os mapas de recursos dos modelos professor e aluno têm formas diferentes. Para corresponder a esses mapas de recursos, uma função de similaridade $\mathscr{L}_F(.)$ é usada.

Desde então, pesquisadores introduziram uma série de estudos para combinar as camadas intermediárias usando o conhecimento baseado em recursos. Heo et al. (2019) propõem uma métrica chamada "distância do limite de ativação" para medir a semelhança entre os limites de ativação do professor e aluno. Eles também introduzem uma nova técnica para gerar os limites de ativação, chamadas de redes neurais de busca de limite, que são treinadas para maximizar a distância entre os limites de ativação de diferentes classes. O método proposto por Passban et al. (2021), introduz uma projeção de camada baseada em atenção que projeta o conhecimento aprendido por um modelo

professor em um modelo aluno. Essa projeção é obtida calculando mapas de atenção entre os mapas de recursos dos modelos professor e aluno, que são usados para orientar o processo de destilação. Chen et al. (2021a) também propôs um método chamado de destilação de camada cruzada com calibração semântica, que envolve a destilação de conhecimento em diferentes camadas de um modelo professor para um modelo aluno. Além disso, o método incorpora um mecanismo de calibração semântica, que visa reduzir a lacuna semântica entre os modelos de professor e aluno.

Relation-Based Knowledge - Além do conhecimento armazenado nas camadas de saída e camadas intermediárias de uma rede neural, o conhecimento por associação entre mapas de recursos também pode ser aplicado para treinar um modelo aluno, conforme ilustrado na Figura 2.13c. É uma relação que pode ser modelada através de correlações entre mapas de recursos, grafos, matrizes de similaridade, *embeddings* ou distribuições probabilísticas baseadas em representações de recursos (Gou et al., 2021).

Uma maneira comum de formular a perda de destilação de conhecimento baseado em relações, baseado entre os mapas de recursos, é representada pela Equação 2.6 (Gou et al., 2021):

$$KD_{Relation}(f_t, f_s) = \mathscr{L}_R{}^1(\Psi_t(\hat{f}_t, \check{f}_t), \Psi_s(\hat{f}_s, \check{f}_s)), \tag{2.6}$$

onde os mapas de recursos dos modelos professor e aluno são referidos como f_t e f_s, respectivamente. Os pares de mapa de recursos são selecionados a partir do modelo professor $\hat{f}_t$ e $\check{f}_t$ e do modelo aluno $\hat{f}_s$ e $\check{f}_s$). As funções de similaridade $\Psi_t(.)$ e $\Psi_s(.)$ são aplicadas a pares de mapas de características de ambos os modelos, e $\mathscr{L}_R{}^1(.)$ é a função de correlação para os mapas de recursos do professor e do aluno.

Recentemente houve vários avanços na destilação de conhecimento baseada em relacionamento que melhoraram sua eficácia. Park et al. (2019) propõe perdas de destilação de distância e ângulo que penalizam diferenças estruturais nas relações. O método é chamado "*Relation Distillation*" e é baseado na ideia de que os mapas de recursos do modelo do professor possuem uma estrutura relacional que captura informações importantes sobre os dados de entrada. Outro avanço na destilação do conhecimento relacional é proposto por Yang et al. (2022), com o conceito *Cross-Image Relational* KD (CIRKD), que se concentra na transferência de relacionamentos estruturados *pixel* por *pixel* e *pixel* por região entre imagens inteiras. Mecanismos de atenção (Guo et al., 2022) também são propostos na literatura para melhor capturar as relações entre os recursos aprendidos pelo modelo professor. Os mecanismos de atenção permitem que o modelo aluno atenda seletivamente às partes mais informativas dos mapas de recursos do modelo professor, o que tem se mostrado

eficaz na melhoria da transferência de conhecimento (Zhou et al., 2020).

2.6 Aplicações Multimodais

Uma série de estudos empíricos emergiu com aplicações multimodais, pois é um campo multidisciplinar de grande importância e com crescente potencial (Baltrušaitis et al., 2018). Para o propósito desta revisão uma visão geral dos aplicativos que utilizam o aprendizado multimodal serão abordados.

Atualmente existem vários domínios que envolvem diferentes modalidades de dados, os quais são descritos abaixo:

- **Multimídia** - Aplicativos incluindo classificação de imagens, agrupamento de imagens, reconhecimento de voz e processamento de linguagem natural. O aprendizado multimodal em multimídia elenca as áreas descritas nos tópicos subsequentes, como também tarefas que incluem detecção de conceito semântico, audiovisual, detecção de fala, rastreamento humano e detecção de eventos. O estudo de Tian et al. (2019) apresenta uma nova estrutura de aprendizado profundo multimodal para detecção de eventos a partir de vídeos que integram diferentes representações de dados em dois níveis, ou seja, nível de quadro e nível de vídeo. Sua proposta prevê o uso do modelo *InceptionV3* (Szegedy et al., 2016) para dados visuais, modelo *AENet* (Takahashi et al., 2017) para extração de áudio dos vídeos e *GloVe* (Pennington et al., 2014) para incorporação de palavras extraídas das descrições dos vídeos com base na fatoração de uma matriz de palavras estatísticas de co-ocorrência. A taxa de acerto (acurácia) em relação aos modelos unimodais e multimodais de última geração é melhorada em mais de 16% e 7% respectivamente.

- **Reconhecimento da Atividade Humana** - Busca identificar e analisar automaticamente as atividades humanas usando informações adquiridas de vários tipos de dados e incorpora aplicativos de amplo alcance, como vigilância, robótica e monitoramento de saúde pessoal. O trabalho de Ehatisham-Ul-Haq et al. (2019) utiliza uma abordagem multimodal para reconhecimento de ações humanas, que extraí dados de vários sensores, incluindo câmera RGB, sensor de profundidade e sensores inerciais vestíveis. A representação dos vários sensores é então fundida e os classificadores *KNN* e *SVM* são usados para treinar e testar o modelo de fusão proposto, conforme ilustrado na Figura 2.14. O trabalho apresenta resultados promissores e obteve uma taxa de acerto (acurácia) de 97,6% nas 27 ações humanas diferentes.

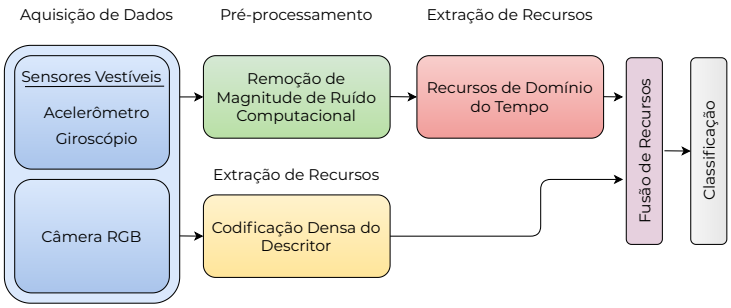


Figura 2.14: Diagrama de blocos para reconhecimento humano. Adaptado de Ehatisham-Ul-Haq et al. (2019).

- **Sistemas Autônomos** - São aplicações integradas em diversos setores da engenharia: automotivo, aéreo, navegação e internet das coisas. Recentemente, algumas transformações na indústria automobilística tem impactado a forma como os seres humanos estão interagindo com seus veículos. O estudo de Caesar et al. (2020) descreve o uso de dados multimodais na direção autônoma de veículos, com possibilidade de avaliar tarefas, como detecção de objetos, rastreamento e comportamento do veículo em uma variedade de condições. Sua principal contribuição é a criação de um conjunto de dados (*nuScenes*) com imagens em 360 graus e uso de sensores (radar, lidar e gps) instalados nos veículos de última geração, conforme ilustrado na Figura 2.15.

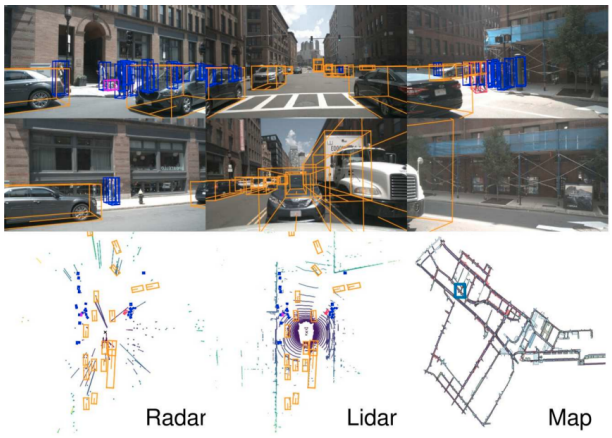


Figura 2.15: Uso de sensores e imagens em sistemas de direção autônoma. Fonte: (Caesar et al., 2020).

- **Aplicações Médicas** - Aplicações médicas multimodais consistem em várias varreduras do mesmo objeto/imagem usando múltiplos métodos de extração, as quais podem ser usadas no diagnóstico, planejamento de tratamento, comunicação médico-paciente, bem como orientação interdisciplinar. Há diferentes modalidades de imagens médicas, como ressonância magnética, tomografia computadorizada, tomografia por emissão

de pósitrons (PET), ressonância magnética funcional (*fMRI*), raio-X e ultrassom.

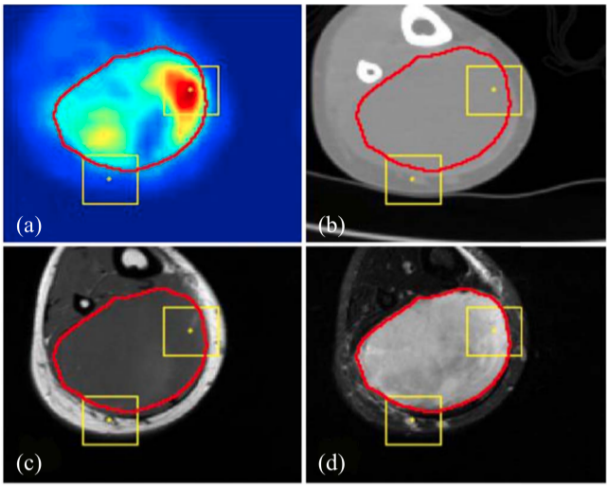


Figura 2.16: Imagens médicas sobre diferentes perspectivas. Fonte: (Guo et al., 2019).

A pesquisa de Guo et al. (2019) relata o uso de CNNs para contornar as lesões de sarcomas com tecidos moles, usando diversos tipos de imagens. A Figura 2.16, ilustra uma abordagem multimodal com imagens extraídas na mesma posição de um objeto, sobre diferentes perspectivas visuais, os pontos em amarelo representam tumores identificados. O trabalho descreve que para a tarefa de segmentação de tumor, realizar a fusão de imagens entre as camadas de uma rede neural, geralmente é melhor do que fundir imagens na saída da rede.

Limitações em Aplicações Multimodais

Alguns estudos permitem relacionar desafios encontrados nas aplicações multimodais, pois como ainda há modelos em estágio preliminar, há uma série de itens que precisam ser solucionados e que dão suporte para a investigação deste trabalho.

- **Arquitetura** - O estudo de Gao et al. (2020) descreve em suas conclusões que há um grande número de pesos livres nos modelos de aprendizado profundo multimodal, especialmente, parâmetros redundantes que têm pouco efeito na tarefa de interesse. Treinar dados multimodais em dispositivos de computação de alto desempenho com arquitetura computacional atual, pode não aprender estruturas de recursos dos dados multimodais devido ao grande volume de dados.
- **Relação Semântica** - Outro aspecto envolvido neste estudo é que para aprender informações de intermodalidade a maioria dos modelos de aprendizado profundo existentes para fusão de dados multimodais, primeiro

usa um modelo profundo unimodal para obter uma representação específica de cada entrada de dado. Porém, ao usar este método, os modelos de aprendizado profundo multimodal podem não capturar o conhecimento totalmente semântico dos dados multimodais. Associado a este fato, as representações de intermodalidade são concatenadas de forma linear, de modo que podem não caber nos relacionamentos complexos em modalidades múltiplas. Logo, uma pesquisa que investigue modelos de aprendizado profundo para dados multimodais e que levem em consideração as relações semânticas são objeto de estudo.

- **Pesos** - Segundo Atrey et al. (2010) o problema da atribuição de peso ideal às diferentes modalidades é um problema em aberto. Pois normalmente há diferentes níveis de confiança nas diferentes modalidades para realizar as várias tarefas de aprendizado.

- **Conjunto de Dados** - Alguns trabalhos também notificam a baixa qualidade dos dados multimodais, muitas vezes com presença de ruído, dados incompletos e *outliers* (Ramachandram and Taylor, 2017). Um conjunto de dados com informações faltantes, pode influenciar na correlação entre as diversas fontes de uma domínio, impactando fortemente no desempenho do aprendizado multimodal. Estratégias para mitigar problemas como este em aprendizado multimodal são factíveis de estudo, isso por que um dos problemas relacionados a baixa qualidade dos dados é que dados multimodais são limitados devido ao alto custo da rotulagem manual. A aquisição de conjuntos de dados rotulados de alta qualidade consome muito trabalho.

- **Modelos Pré-Treinados** - Em relação aos tipos de fontes de dados o estudo de Guo et al. (2019) afirma que há modelos de última geração para modalidades ligadas a fontes textuais e de imagem, porém carece de modelos pré-treinados disponíveis para as modalidades de áudio ou vídeo. Logo, as redes profundas usadas para extrair recursos de áudio ou vídeo sofrem com *overfitting* devido às instâncias de treinamento limitadas.

- **Transferência Intermodal** - Ainda segundo Guo et al. (2019) há uma demanda por estudos que caracterizem a fusão de dados no uso da transferência intermodal, que visa transferir conhecimento de uma modalidade para outra. Outro ponto um pouco mais desafiador é o aprendizado de transferência entre conjuntos de dados multimodais, que pode ser referência para estudos posteriores.

2.7 Medidas de Avaliação dos Modelos

Nesta seção são abordadas as medidas de avaliação de desempenho para mensurar a qualidade do treinamento dos modelos. A literatura aborda uma série de funções matemáticas para avaliar a capacidade de erro e acerto em modelos de aprendizado (Burkov, 2019). Cada métrica possui uma finalidade diferente, logo é importante levar em consideração a aplicação prática do modelo (classificação, probabilidade, ranking). Note que alguns termos, métricas e nomenclaturas citadas ao longo deste trabalho, serão mantidas no idioma inglês para uma melhor compreensão do leitor.

2.7.1 Métricas de Classificação

As métricas de avaliação para um modelo de classificação relacionam a comparação entre as classes preditas pelo modelo e as classes verdadeiras de cada exemplo. Portanto, as métricas de classificação têm como objetivo mensurar quão distante o modelo está de uma classificação perfeita. Essas métricas estão associadas a problemas que expressam valores discretos para os atributos relacionados a classe.

Matriz de Confusão

Para avaliação dos modelos e aplicação das métricas é necessário descrever a Matriz de Confusão, uma mecanismo utilizado para realizar o cálculo obtido por cada avaliação. A Matriz de Confusão é caracterizada por uma tabela que mostra as frequências de classificação para cada classe do modelo. Sua representação é definida por uma matriz com linhas e colunas. Cada coluna desta matriz representa as instâncias previstas em uma classe do conjunto de dados, enquanto que as linhas descrevem as instâncias reais de cada classe do conjunto.

Sua representação é válida para conjuntos que possuem apenas duas classes, porém para problemas multi-classe é possível a generalização do conjunto (ex: um contra todos) para duas classes. A Matriz de Confusão é definida por duas linhas e duas colunas com as seguintes situações:

(VP)	verdadeiro positivo	: exemplo positivo	predito como positivo
(FP)	falso positivo	: exemplo negativo	predito como positivo
(VN)	verdadeiro negativo	: exemplo negativo	predito como negativo
(FN)	falso negativo	: exemplo positivo	predito como negativo

Sua definição pode ser também compreendida pela Tabela 2.3, na qual os dados do conjunto são dispostos entre exemplos positivos e negativos. Onde,

VP, FP, VN e **FN** representam para o modelo de classificação, respectivamente, o número de exemplos verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos; **Pos** o total de exemplos positivos; **Neg** o total de exemplos negativos; **PPos** o total de exemplos preditos como positivos; **PNeg** o total de exemplos como negativos; **Total** o número total de exemplos no conjunto de exemplos.

Tabela 2.3: Matriz de Confusão.

	Preditos Positivos	Preditos Negativo	
Exemplos Positivos	VP (Correto)	FN (Incorreto)	**Pos**
Exemplos Negativos	FP (Incorreto)	VN (Correto)	**Neg**
	PPos	**PNeg**	**Total**

PRECISION

A medida de avaliação *Precision*, Equação 2.7, descreve a taxa com que todos os exemplos avaliados como positivos são realmente positivos.

$$Precision = \frac{VP}{FP+VP} \tag{2.7}$$

RECALL ou TVP

Enquanto *Precision* avalia os exemplos preditos, *Recall* avalia entre os exemplos positivos quantos realmente foram preditos como positivos. Logo, expressa a qualidade de classificação da classe positiva.

$$Recall = \frac{VP}{FN+VP} \tag{2.8}$$

TFP

Expressa a qualidade de classificação da classe negativa, no entanto avalia a classe com uma taxa de erro. Assim quanto menor o valor, melhor será a classificação da classe negativa.

$$TFP = \frac{FP}{FP+VN} \tag{2.9}$$

F1

A medida de avaliação *F1* possui como estimativa a média harmônica obtida entre os valores *Recall* e *Precision* e sua fórmula pode ser definida na Equação 2.10.

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{2.10}$$

ACCURACY

A *Accuracy* é baseada na proporção das instâncias classificadas corretamente pelo volume total de instâncias do conjunto. Sua formulação é definida pela Equação 2.11.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.11}$$

Análise ROC

O gráfico ROC (*Receiver Operating Characteristic*) (Fan et al., 2006), Figura 2.17, é um espaço bidimensional formado pelos eixos X e Y representados respectivamente por TFP (Taxa de Falso Positivos) e TVP (Taxa de Verdadeiro Positivos). Uma linha diagonal aos eixos representa um classificador/modelo aleatório. Os pontos inseridos no gráfico por meio da Matriz de Confusão formam a curva ROC de um classificador/modelo. Esta é uma técnica prática para a visualização de classificadores baseado no seu desempenho, um instrumento que permite estudar a variação da sensibilidade e especificidade de um classificador/modelo. O gráfico ROC é descrito no espaço bidimensional por duas zonas importantes: Céu ROC, representado pelos pontos próximos a (0,1), caracteriza uma classificação próxima da ideal na qual todos exemplos são preditos corretamente, e Inferno ROC representado pelos pontos próximos a (1,0), em que estabelece uma condição oposta ao Céu ROC com registro dos piores resultados para um classificador.

A variação na escala AUC abrange os limites (lim = $0 \leftrightarrow 1$). A Figura 2.18, exibe a construção da curva ROC para o dataset diabetes (Asuncion and Newman, 2007) e ilustra um classificador aleatório com uso do método de validação cruzada (Browne, 2000) para n-folds, em que consiste dividir o conjunto total em n subconjuntos de mesmo tamanho, na qual um subconjunto é usado para testes e os $n-1$ subconjuntos usados na etapa de treino.

LOG LOSS

Perda Logarítmica ou LOG LOSS (Perda de Entropia Cruzada) é uma métrica usada para as estimativas de probabilidade. É uma métrica que corrige previsões incorretas muito confiantes, logo é considerada uma boa métrica para comparar modelos em problemas binários e multiclasse. Para qualquer

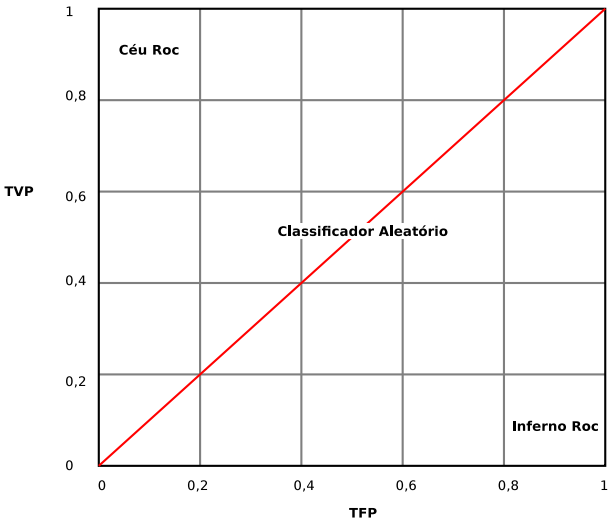


Figura 2.17: Gráfico ROC. Adaptado de Flach (2003).

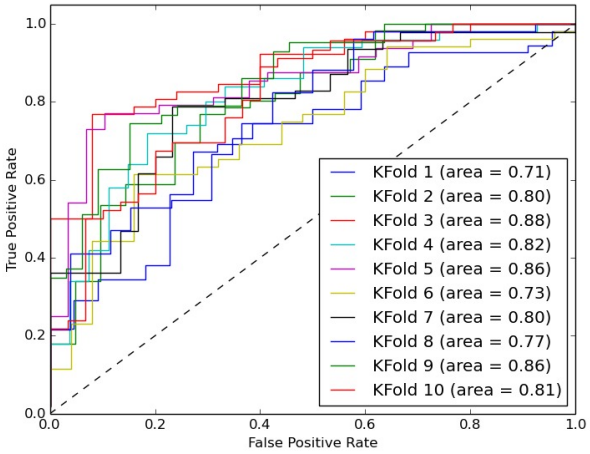


Figura 2.18: Gráfico ROC *dataset diabete* (Asuncion and Newman, 2007).

tipo de problema, um menor valor para perda logarítmica significa obter melhores previsões. Para um conjunto com rótulo verdadeiro $y \in 0,1$ e uma estimativa de probabilidade $p = \Pr(y = 1)$, a perda de log é:

$$\text{LOG LOSS} = L_{\log}(y,p) = -(y\log(p) + (1-y)\log(1-p)) \tag{2.12}$$

2.7.2 Métricas de Regressão

Métricas de Regressão lidam com valores contínuos, aplicados normalmente aos números reais. Sua finalidade é estimar a relação entre as variáveis do problema, por meio de uma equação matemática que descreve o comportamento do modelo. Essas métricas permitem tanto prever a classe do dado de entrada, como também prever o seu valor e são importantes para medir o erro de modelos de previsão em valores numéricos.

Mean Squared Error

Mean Squared Error ou Erro Quadrático Médio (MSE) é uma métrica muito utilizada em problemas de regressão linear, quando há uma linearidade entre os dados de entrada e os dados a serem preditos. Sua abordagem, calcula a diferença entre os resultados obtidos e o resultado real, eleva cada diferença ao quadrado, e depois calcula a média.

$$\text{MSE} = \frac{1}{n}\sum(y_{observado} - y_{predito})^2 \tag{2.13}$$

Há uma derivação da métrica MSE chamada *Root Mean Squared Error* (RMSE), que simplifica e facilita a interpretação. Sua fórmula adiciona a raiz quadrada ao MSE, logo o erro reproduz a mesma unidade de medida da variável dependente.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_{observado} - y_{predito})^2} \tag{2.14}$$

Mean Absolute Error

Mean Absolute Error ou Erro Médio Absoluto (MAE) é uma medida de erros entre observações pareadas que expressam o mesmo fenômeno. Logo, é a diferença absoluta média entre $y_{observado}$ e $y_{predito}$, descrito pela fórmula 2.15:

$$\text{MAE} = \frac{1}{n}\sum|y_{observado} - y_{predito}| \tag{2.15}$$

2.7.3 Métricas de Recuperação de Informação

As métricas usadas em abordagens com fusão de dados são equivalentes as métricas de classificação e regressão. Porém para cada experimento é necessário avaliar sua aplicação prática. Neste sentido é possível relacionar algumas métricas de recuperação de informação.

A *Accuracy* é a métrica de avaliação mais utilizada em trabalhos recentes com fusão de dados, isso por que é muito empregada em problemas de classificação (Yang et al., 2017; Yu et al., 2020). Pois, reflete o número de exemplos corretos sobre todos exemplos de um conjunto de dados. Portanto, ao fundir dados pesquisadores buscam obter um maior número de previsões de acerto em modelos multimodais quando comparado a modelos unimodais em um mesmo conjunto de dados.

Outra derivação muito usada para comparação entre modelos unimodais e multimodais é o uso do *Average Precision* (AP), que tem o objetivo de encontrar a área acima sob a curva da *Precision-Recall* (Yohanandan, 2020). Em alguns contextos, o AP é calculado para cada classe e a média é calculada para ob-

ter a *mean average precision* (mAP). A pontuação mAP é calculada tomando o AP médio em todas as classes, conforme ilustrado na Figura 2.19. Os trabalhos de Hu et al. (2019) e Pang et al. (2015), utilizam mAP como métrica de avaliação para validação de seus experimentos.

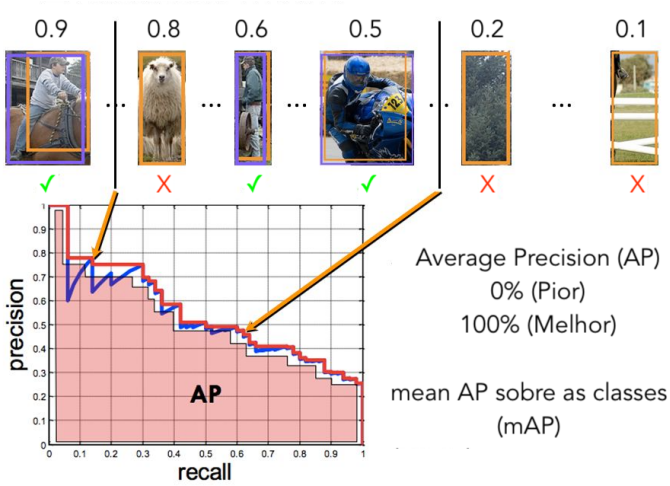


Figura 2.19: Métrica mAP. Adaptado de Lanqing (2019).

Há também para problemas multimodais o uso da *loss* para detecção de objetos ou *object detection*, que consiste em duas partes (Jiang and Learned-Miller, 2017). Primeiro a perda de localização para previsão de deslocamento de uma caixa delimitadora, em seguida a perda de classificação para probabilidade das classes. Ambas as partes são calculadas como a soma dos erros quadrados. Outras variações desta métrica podem ser encontradas na literatura com o nome Intersecção sobre União ou *Intersection over Union* (IoU), uma métrica de avaliação usada para medir a precisão de um detector de objeto em um conjunto de dados específico (Jiang et al., 2020). O IoU calcula a interseção sobre a área de sobreposição entre as caixas delimitadoras pela união da área total entre as caixas, conforme ilustrado na Figura 2.20.

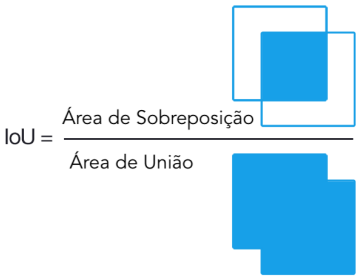


Figura 2.20: Métrica IoU. Adaptado de Rosebrock (2016).

A métrica IoU também é utilizada em Redes Neurais Convolucionais Baseadas em Região (R-CNN) com abordagens multimodais. O objetivo da R-CNN é obter uma imagem de entrada e produzir um conjunto de caixas delimitadoras como saída, onde cada caixa delimitadora contém um objeto. Em

seu estudo Ren et al. (2015), desenvolveu uma arquitetura similar a *Fast* R-CNN (Girshick, 2015), chamada de *Faster* R-CNN para detecção de objetos composto por dois módulos. O primeiro módulo é uma rede profunda totalmente convolucional para mapear regiões em uma imagem, e o segundo módulo é o detector *Fast* R-CNN incorporado em sua arquitetura. Para problemas como este, uma forma de avaliar sua taxa de acerto é mesclar métricas de regressão e classificação, sendo a regressão a probabilidade de uma determinada região na imagem ser um objeto e a classificação predizer se o objeto delimitado por uma caixa ser de fato um objeto, conforme descrito na Fórmula 2.16.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \tag{2.16}$$

Neste contexto, i representa o índice dentro de um lote de entrada no modelo e p_i a probabilidade desse elemento ser um objeto. Logo, a primeira parte da equação classifica os elementos em objetos de forma binária (0,1) pelo termo p_i^*, sendo 0 = (não objeto) e 1 = (objeto). A segunda parte da equação descreve o termo t_i, como uma representação vetorial das 4 coordenadas parametrizadas da caixa delimitadora ser prevista e t_i^* a previsão correta das caixas delimitadores, associada a um elemento verdadeiro pelo termo p_i^*. Para cálculo das métricas, a LOG LOSS é usada na classificação para prever as duas classes (0,1) e para regressão a função de perda robusta (*smooth* L_1) definida no trabalho de Girshick (2015) é empregada pelo termo R, logo tem-se $L_{reg}(t_i,t_i^*) = R(t_i - t_i^*)$ e significa que a função de perda para regressão é ativada apenas para elementos positivos, ou seja, classificados como objetos. Por fim, os dois termos são normalizados por N_{cls} e N_{reg} e ponderado por um parâmetro de equilíbrio λ.

No entanto, quando as pesquisas buscam elencar o efeito multimodal em modelos descritos na literatura, abordagens baseadas no erro de predição são amplamente usadas. Pois, métricas que envolvem o erro de classificação ou regressão, quantificam sua precisão penalizando predições falsas, por exemplo o uso da métrica LOG LOSS.

2.8 Organização da Pesquisa

A Figura 2.21 representa a estrutura deste estudo mostrando as atividades realizadas em cada etapa, algumas das quais foram concluídas e outras serão objeto de estudos futuros.

No Capítulo 3 são discutidos alguns experimentos com aprendizado multimodal. Esta abordagem torna-se interessante devido ao uso das redes pré-treinadas já descritas neste capítulo, como ponto de partida para o aprendi-

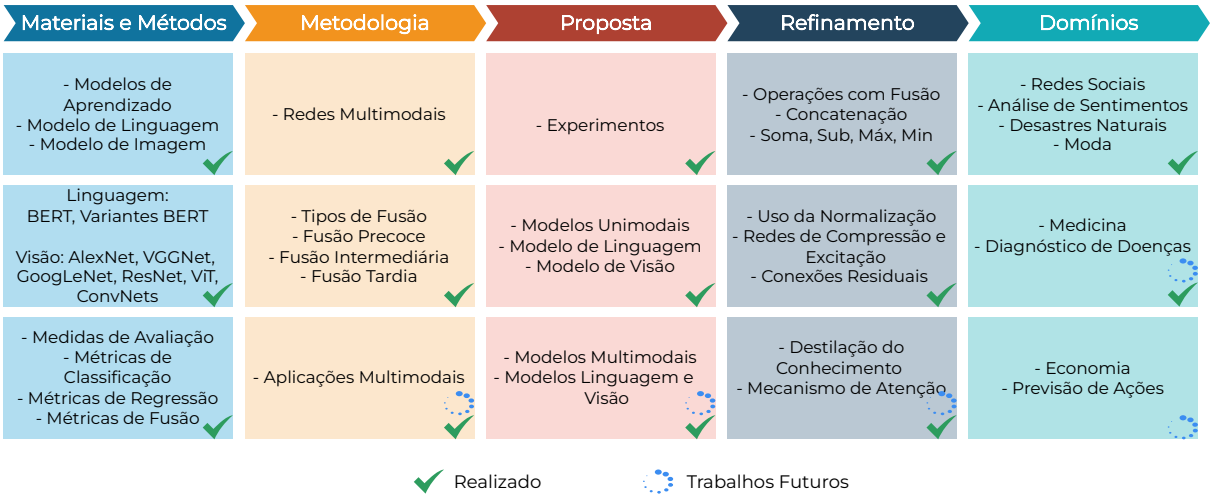


Figura 2.21: Estrutura e Planejamento.

zado de uma nova tarefa. Utilizar uma rede pré-treinada com transferência de aprendizado é normalmente muito mais rápido e fácil do que treinar uma rede do zero (Tan et al., 2018). As abordagens exploradas neste trabalho são ilustradas na Figura 2.22 e serão discutidos nos experimentos dos Capítulos 3,4 e 5. O mapa do estudo apresenta os tópicos explorados na pesquisa, bem como as interconexões entre eles. O nós centrais representam os conceitos teóricos estudados, enquanto as ramificações indicam as relações entre eles. As cores indicam a natureza dos tópicos. Os itens em vermelho compõe temas relacionados ao uso do Aprendizado Multimodal.

Para uma melhor contextualização da pesquisa as próximas seções e capítulos são divididos em duas etapas: (i) para extrair características relevantes de cada modalidade, o Apêndice A consiste em uma série de trabalhos publicados por Modelos Unimodais. Durante esta etapa pesquisas exploratórias foram conduzidas para entender o comportamento de cada modalidade, bem como estratégias, otimizações e modelos estado da arte para a resolução de problemas atuais. Com o objetivo de otimizar o espaço neste trabalho, os estudos unimodais foram incluídos como fonte de referência na seção de Apêndice; (ii) os experimentos descritos nos Capítulos 3,4 e 5 propõe aprender uma representação conjunta usando as modalidades de imagem, texto para problemas de classificação e regressão, as características de cada modalidade são mescladas utilizando uma série métodos propostos neste capítulo, neste sentido o uso de operações aritméticas, normalizações, mecanismos de atenção, destilação de conhecimento, aprendizado multivisão, entre outras abordagens foram empregadas para descrição dos Modelos Multimodais.

Figura 2.22: Mapa do Estudo.

CAPÍTULO

3

Estudos com Operações Aritméticas, Mecanismos de Atenção e Conexões Residuais

Contemplado os estudos individualizados sobre os modelos de linguagem e visão descritos no Apêndice A, abordagens com modelos multimodais são descritas neste capítulo.

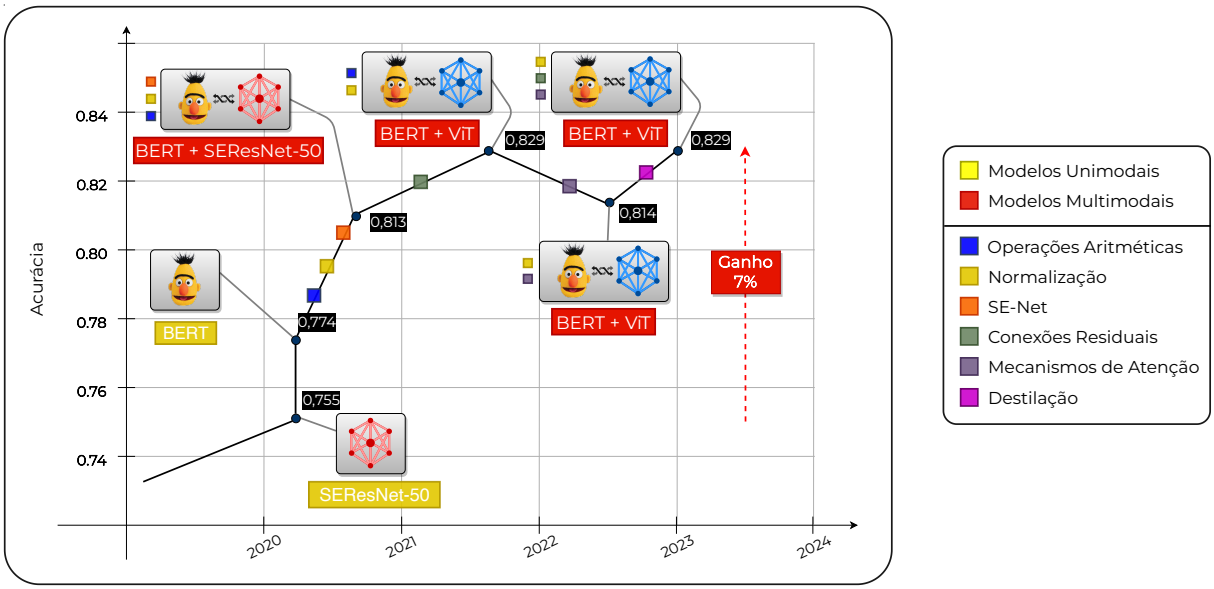


Figura 3.1: Evolução das técnicas do Aprendizado Multimodal realizados por este trabalho utilizando o conjunto de dados multimodal (*Top Speed*).

A Figura 3.1 ilustra um gráfico temporal com a evolução da pesquisa em termos de *Accuracy* ao longo do tempo, no período de 2019 a 2023. Durante

esse período foram abordados diversos assuntos relacionados ao Aprendizado Multimodal, tais como operações aritméticas, normalizações, conexões residuais, mecanismos de atenção e destilação do conhecimento multimodal. O conjunto de dados utilizado nesta ilustração (*Top Speed*) foi o primeiro conjunto multimodal (texto + imagem) a ser empregado nos experimentos com modelos multimodais. A incorporação de novas técnicas ao longo deste período, permitiu uma evolução gradativa de desempenho. Os resultados obtidos foram relevantes em diversos domínios, no entanto apenas o conjunto de dados (*Top Speed*) foi usado intensivamente nos estudos preliminares e permite traçar uma perspectiva evolutiva deste trabalho.

Neste capítulo serão conduzidos dois experimentos para investigar diferentes abordagens multimodais. A Seção 3.1 descreve o Experimento (1.0) com objetivo de analisar o uso das operações aritméticas em relação a concatenação para a fusão de dados multimodais. Algumas combinações com as operações matemáticas são propostas para avaliar a eficácia na obtenção de representações complementares entre as modalidades. Já o Experimento (1.1) descrito na Seção 3.2 visa realizar uma análise mais abrangente, envolvendo cinco conjuntos de dados distintos. Além das operações aritméticas, os mecanismos de atenção e conexões residuais serão explorados na fusão multimodal. Essa abordagem mais ampla permitirá uma melhor compreensão das potencialidades e limitações dessas técnicas, bem como seu desempenho em relação aos modelos de aprendizado multimodal.

3.1 Experimento 1.0 - Fusão de Dados por meio de Operações Aritméticas, Normalização e Redes de Compressão e Excitação

Esta seção aborda o tema da fusão de dados por meio de operações aritméticas, normalização e redes de compressão e excitação. Essas técnicas permitem combinar características complementares das modalidades e capturar informações mais abrangentes sobre as abordagens já explanadas neste trabalho.

Identificação bem-sucedida de vídeos do Youtube usando aprendizado profundo multimodal

Um único vídeo de uma plataforma de compartilhamento de conteúdo pode fornecer muito mais dados do que o próprio vídeo. O estudo proposto usa dados obtidos de um canal popular de automobilismo do *YouTube* (Top Speed) [1],

[1] https://www.youtube.com/user/CanalTopSpeed/

para estimar as visualizações de vídeo usando classes binárias. Para este propósito, recursos de imagem e texto são extraídos dos dados do *Youtube* (Burgess, 2011) para estimar as visualizações usando uma abordagem de aprendizado profundo multimodal. A estrutura multimodal integra todos os recursos disponíveis em um vídeo (por exemplo, *frames*, miniaturas, legendas, descrição, título e áudio) e extrai informações de diferentes fontes para prever e avaliar seu conteúdo. Texto de títulos e transcrições de áudio, miniaturas de imagens, número de curtidas, *likes* e visualizações são exemplos de dados disponíveis em um vídeo do *YouTube*. Apesar da variabilidade a maioria dos modelos de aprendizado utiliza apenas um tipo de dado (texto, imagem, áudio ou vídeo). Além disso o uso simultâneo de múltiplas fontes de dados para tais problemas ainda é raro.

Ao considerar a literatura multimodal usando plataformas de multimídia, o estudo de Yu and Shi (2020) concatena recursos incorporados de quadros, títulos, descrições e áudios para selecionar imagens visualmente atraentes. O recurso concatenado é processado por *gating* de contexto semelhante aos módulos de auto-atenção do *Transformer* (Vaswani et al., 2017) e submetido a camadas totalmente conectadas. Além da concatenação, este estudo propõe o uso de oito operações aritméticas para mesclar dados textuais e visuais, pois o uso de operações simples pode trazer resultados significativos em relação a estratégia de concatenação.

Estudos recentes são o ponto de partida para relacionar operações de fusão em redes multimodais. O trabalho de Chen et al. (2020) usa métodos de aprendizado profundo para extrair recursos de *frames* de vídeo, informações de movimento e recursos de sequência de vídeo. A análise de sentimento também foi estudada em abordagens de vídeo, nos trabalhos de Pandeya and Lee (2020) e Jin et al. (2020), o aprendizado profundo multimodal baseado em fusão é utilizado para classificar o sentimento em vídeos.

Pré-processamento e Configuração Experimental

Para este experimento o título do vídeo foi usado para extrair recursos textuais. Algumas variações com descrições e comentários foram analisadas nos testes iniciais. No entanto, manter apenas o título do vídeo como entrada textual, foi a solução mais simples encontrada. Como entrada de dados visuais as imagens estáticas de vídeos do *Youtube*, chamadas *thumbnails*, cujo termo é utilizado por designers gráficos foi utilizada como representação visual. Depois de uma etapa de pré-processamento ilustrada pela Figura 3.2 passando por etapas de redimensionamento, normalização de *pixels* e conversão de escala, 462 exemplos foram usados para os experimentos, separados por limite de visualizações de vídeo para dividir as classes. Para medir o sucesso do vídeo

foi estabelecido o limite em 100 mil visualizações (12,5% dos inscritos). Após essa marca, segundo o desenvolvedor de conteúdo, o vídeo tem mais chances de monetizar e fazer sucesso para a audiência do canal. Este limite também está associado ao saldo e número de exemplos disponíveis no conjunto de dados, totalizando 462 vídeos com aproximadamente 50% destes acima de 100 mil visualizações. Outro fator que impede de tratar este trabalho como um problema de regressão é que os vídeos possuem temporalidade e não são sazonais. Inicialmente esse problema foi modelado usando a temporalidade das visualizações em um período de 30 dias, mas a variação nos resultados não foi significativa para os experimentos. Assim, constitui-se um problema de classificação binária ($>$ 100k visualizações positivas e $<$ 100k visualizações negativas), conforme ilustrado na Tabela 3.1.

(a) Imagem Original (b) Imagem Processada

Figura 3.2: Imagem pré-processada para uso em rede neural para o conjunto de dados *Top Speed*.

Tabela 3.1: Número de exemplos por classe.

Classes	Vídeos	Visualizações
0	220	$<$100k
1	242	$>$100k

Ao todo oito métodos de combinação diferentes foram testados para avaliar o desempenho do aprendizado profundo multimodal. Nesse cenário a fusão dos modelos possui dois tipos básicos de integração, concatenação ou combinação de dados por operações aritméticas. A arquitetura inicial para os dois modelos de dados (imagem e texto), foi respectivamente *Bertimbau-base* como extrator de recursos de texto (Souza et al., 2020) pré-treinado em um grande corpus português-brasileiro. E um modelo *SE-Net* baseado em *ResNet* (He et al., 2016) (Hu et al., 2018) para extrair recursos de imagem. A *SE-Net* recalibra de forma adaptativa as respostas dos recursos do canal, modelando explicitamente as interdependências entre os canais. Em seguida, uma camada para fusão é aplicada nas camadas de saída de ambos os modelos, pois

geram vetores de recursos unidimensionais (*embeddings*) o que torna o processo de fusão tardia útil entre dois modelos unimodais.

Operações aritméticas

Para fins de notação as seguintes equações são descritas no formato abaixo:

Considere $\mathbf{z}^{img} = f(\mathbf{x}^{img})$ um mapeamento de uma imagem $\mathbf{x}^{img}$ para um espaço de incorporação $\mathbf{z}^{img} \in \mathrm{R}^d$ onde d denota a dimensão deste mapeamento de imagem. Tomando o texto como entrada, seja $\mathbf{z}^{texto} = g(\mathbf{x}^{texto})$ o mapeamento de um texto $\mathbf{x}^{texto}$ para uma espaço de incorporação $\mathbf{z}^{texto} \in \mathrm{R}^n$ onde n denota a dimensão deste mapeamento de texto.

Concatenação - Concatena a sequência de *embeddings* fornecida na dimensão fornecida. Todos os *embeddings* devem ter a mesma forma (exceto na dimensão de concatenação) ou estar vazios. A dimensão resultante é a soma das dimensões de entrada.

$$\text{Concatenação}(\mathbf{z}^{texto}, \mathbf{z}^{img}) = \mathbf{z}^{texto} \frown \mathbf{z}^{img} \tag{3.1}$$

Observe que a concatenação permite a operação de dois *embeddings* com dimensões diferentes. Quando uma operação requer o mesmo número de dimensões uma camada totalmente conectada $fc : \mathrm{R}^d \to \mathrm{R}^n$ é definida para fusão. Para o propósito deste estudo dimensões com mesmo tamanho são descritas como: $\mathbf{rz}^{img} = fc(\mathbf{z}^{img})$, onde as *embeddings* de imagem são redimensionadas.

As outras operações de fusão exploradas neste estudo são definidas nas seguintes equações:

Divisão - Divide os elementos da primeira entrada pela segunda entrada.

$$\text{Divisão}(\mathbf{z}^{texto}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} \frac{z_i^{texto}}{rz_i^{img}} \tag{3.2}$$

Máximo - Compara todos os elementos nas *embeddings* de entrada e retorna os valores máximos para cada posição .

$$\text{Máximo}(\mathbf{z}^{texto}, \mathbf{rz}^{img}) = \max_{1 \le i \le n} \begin{cases} z_i^{texto}, & \text{se } z_i^{texto} > rz_i^{img}. \\ rz_i^{img}, & \text{caso contrário.} \end{cases} \tag{3.3}$$

Mínimo - Compara todos os elementos nas *embeddings* de entrada e retorna os valores mínimos para cada posição.

$$\text{Mínimo}(\mathbf{z}^{texto}, \mathbf{rz}^{img}) = \min_{1 \le i \le n} \begin{cases} z_i^{texto}, & \text{se } z_i^{texto} < rz_i^{img}. \\ rz_i^{img}, & \text{caso contrário.} \end{cases} \tag{3.4}$$

Multiplicação - Multiplica os elementos da primeira entrada pela segunda

entrada.

$$\text{Multiplicação}(\mathbf{z}^{texto}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} z_i^{texto} * rz_i^{img} \tag{3.5}$$

Soma - Retorna a soma dos elementos de todas as entradas.

$$\text{Soma}(\mathbf{z}^{texto}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} (z_i^{texto} + rz_i^{img}) \tag{3.6}$$

Subtração - Retorna a subtração dos elementos de todas as entradas.

$$\text{Subtração}(\mathbf{z}^{texto}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} (z_i^{texto} - rz_i^{img}) \tag{3.7}$$

Potenciação - Concatena a entrada e obtém a potência de cada elemento na entrada resultante com o expoente (2) e retorna um novo vetor de incorporação com o resultado.

$$\text{Potenciação}(\mathbf{z}^{texto}, \mathbf{z}^{img}) = (\mathbf{z}^{texto} \frown \mathbf{z}^{img})^2 \tag{3.8}$$

A classificação é realizada na etapa final do processo da fusão de dados. A Figura 3.3 ilustra o modelo do aprendizado multimodal, com a aplicação da técnica de fusão tardia em todas as combinações geradas. Inicialmente os modelos unimodais são executados em paralelo e nas camadas subsequentes ocorre a fusão por uma das oito operações.

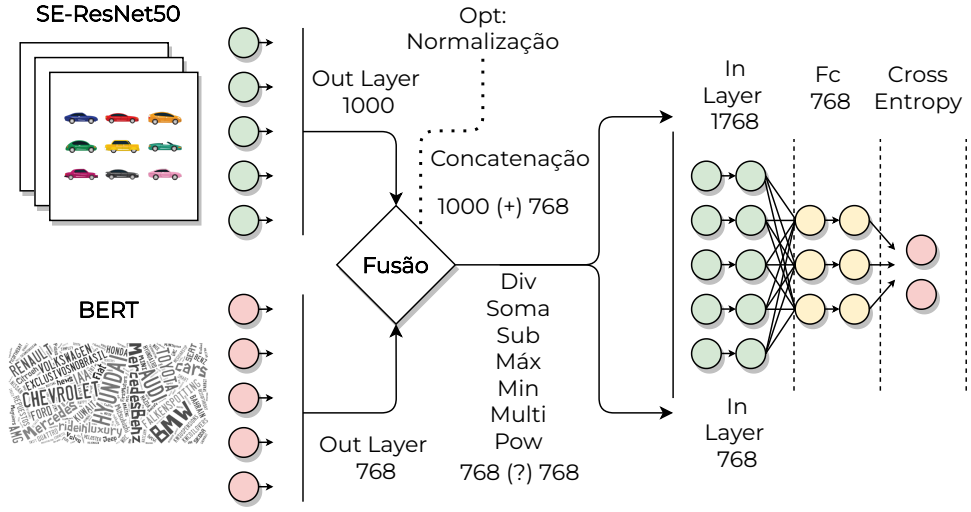


Figura 3.3: Modelo de Fusão - Representação gráfica do modelo multimodal com fusão tardia sobre dois modelos unimodais (*BERT* e *SE-ResNet50*). Na etapa de fusão uma das oito operações aritméticas é selecionada e em seguida uma camada totalmente conectada é aplicada na saída do modelo.

Por se tratar de um conjunto de dados com classe binária foi usado a entropia cruzada (De Boer et al., 2005), como critério de minimização. A Fórmula 3.9 é definida como uma função de perda de entropia cruzada binária

que calcula a perda de um exemplo calculando a média dos erros. Onde $\hat{y}_i$ é o valor esperado e y_i é o valor observado. A entropia cruzada binária mede o quão longe do valor real (0 ou 1) a previsão está para cada uma das classes e, em seguida, calcula a média dos erros de classe para obter a perda final.

$$\underset{\textit{Cruzada Binária}}{\overset{\textit{Entropia}}{}} = -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \tag{3.9}$$

Pré-processamento e Configuração Experimental

A Tabela 3.2 descreve os parâmetros usados no experimento seguindo o protocolo de treinamento de Ramachandram and Taylor (2017), que utiliza fusão tardia com a agregação de decisões de vários classificadores, cada um treinado em modalidades separadas. Os modelos não tiveram suas camadas congeladas durante o treinamento e suas dimensões de saída foram alteradas conforme descrito na Seção 3.1 (operações aritméticas). Um teste de intervalo de taxa de aprendizado foi realizado para encontrar a taxa de aprendizado mais alta que minimiza a perda e não faz com que ela exploda (Smith, 2017). A taxa de aprendizado encontrada foi usada como o limite superior da política de um ciclo usada para treinamento (Smith and Topin, 2019). A política de taxa de aprendizado de um ciclo auxilia na alteração da taxa de aprendizado após cada lote. A validação cruzada de 10 partições foi usada para avaliar os modelos estudados, e cada experimento usou 100 épocas por vez. As métricas foram avaliadas ao final das 100 épocas. Finalmente, cada resultado é a média dos 10 *folds*. O otimizador *AdamW* (Loshchilov and Hutter, 2017) foi selecionado para os experimentos a fim de definir a taxa de aprendizado.

Tabela 3.2: Conjunto de Parâmetros.

Parâmetros	
Épocas	10
Validação Cruzada	10 *Folds*
Otimizador	*AdamW*
Taxa de Aprendizado	3e-5 <n <3e-2
Weight Decay	1e-4
Critério	*CrossEntropy*

Resultados Experimentais

A Tabela 3.3 mostra os resultados em termos das métricas escolhidas (*ROC-AUC, Accuracy, Recall, Precision, F1 e ROC-AUC*). Primeiro os resultados entre texto (*BERTimbau*) e imagem (*SE-ResNet50*) indicam que os títulos são mais discriminativos do que as miniaturas em nossa configuração. Em segundo lugar nossa análise revela que a fusão tardia usando subtração, soma, concatenação, máximo e mínimo tem desempenho superior ao dos modelos

unimodais. No entanto, algumas operações como potenciação, multiplicação e divisão, não são tão eficazes ao vincular incorporações de várias modalidades.

Tabela 3.3: Métricas dos modelos analisados, usando validação cruzada de 10 partições.

Modelos	***Loss***	***Accuracy***	***Precision***	***Recall***	***F1***	***ROC-AUC***
BERTimbau	0.572 ± 0.052	0.774 ± 0.040	0.779 ± 0.039	0.774 ± 0.040	0.772 ± 0.041	0.774 ± 0.040
SE-ResNet50	0.588 ± 0.046	0.755 ± 0.067	0.758 ± 0.067	0.755 ± 0.066	0.754 ± 0.066	0.755 ± 0.066
Fusão	***Loss***	***Accuracy***	***Precision***	***Recall***	***F1***	***ROC-AUC***
Subtração	0.536 ± 0.068	0.813 ± 0.067	0.819 ± 0.068	0.813 ± 0.067	0.812 ± 0.067	0.813 ± 0.067
Soma	0.529 ± 0.068	0.809 ± 0.067	0.815 ± 0.068	0.809 ± 0.067	0.807 ± 0.067	0.809 ± 0.067
Concatenação	0.537 ± 0.065	0.794 ± 0.078	0.799 ± 0.080	0.795 ± 0.078	0.793 ± 0.078	0.795 ± 0.078
Máximo	0.540 ± 0.063	0.793 ± 0.041	0.798 ± 0.044	0.793 ± 0.040	0.793 ± 0.041	0.793 ± 0.040
Mínimo	0.530 ± 0.067	0.791 ± 0.070	0.802 ± 0.076	0.790 ± 0.071	0.789 ± 0.071	0.790 ± 0.071
Potenciação	0.629 ± 0.079	0.702 ± 0.095	0.709 ± 0.102	0.699 ± 0.097	0.696 ± 0.098	0.699 ± 0.097
Multiplicação	0.690 ± 0.016	0.531 ± 0.124	0.532 ± 0.149	0.526 ± 0.120	0.508 ± 0.130	0.526 ± 0.120
Divisão	0.697 ± 0.006	0.493 ± 0.050	0.478 ± 0.141	0.492 ± 0.040	0.402 ± 0.058	0.492 ± 0.040

A proposta atingiu 81,3% de taxa de acerto usando a operação de "*Subtração*" com um ganho de 3,9% sobre o modelo textual e 5,8% para o modelo visual. Logo, parece viável usar o aprendizado profundo multimodal para estimar as visualizações de vídeos do *Youtube*. A Figura 3.4 mostra as métricas adotadas entre os 10 *folds* utilizados no experimento. É possível observar que a fusão por *Subtração* mostra valores mais altos em *Accuracy, Precision, Recall, F1 e ROC-AUC* do que *BERTimbau* e *SE-ResNet50*.

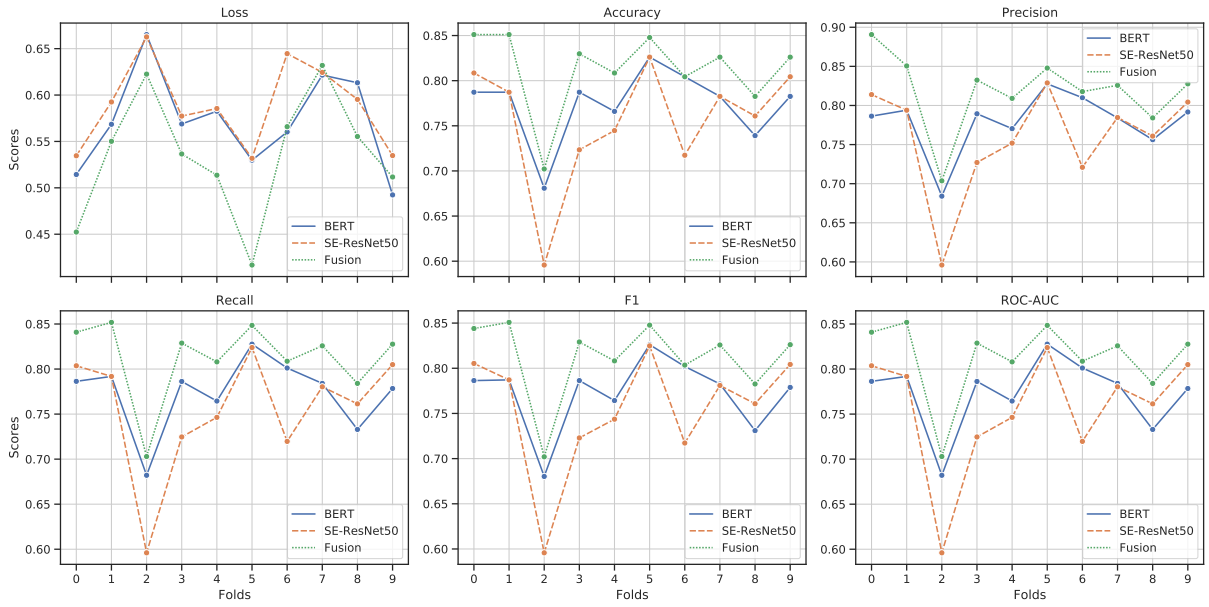


Figura 3.4: *Loss*, *Accuracy*, *Precision*, *Recall*, *F1* e *ROC-AUC* para etapa de validação, com o número de 100 épocas para os modelos *BERTimbau*, *SE-ResNet50* and Fusão = *Subtração*.

Ganho de Desempenho

Considerando os resultados obtidos neste trabalho uma comparação de desempenho é descrita na Tabela 3.4 a qual obteve uma melhoria significativa em relação aos modelos unimodais.

Tabela 3.4: Comparação de Desempenho - Modelos Unimodais versus Multimodais.

Modelos	***Accuracy* (%)**	**STD**	**BERTimbau**	**SE-ResNet50**
BERTimbau Base	**77.43**	0.040	-	-
SE-ResNet50	**75.51**	0.067	-	-
Subtração	**81.30**	**0.067**	**3.86**	**5.79**
Soma	**80.86**	**0.067**	**3.42**	**5.35**
Concatenação	**79.37**	**0.078**	**1.94**	**3.86**
Máximo	**79.35**	**0.041**	**1.91**	**3.84**
Mínimo	**79.14**	**0.070**	**1.70**	**3.63**
Potenciação	70.16	0.095	-7.27	-5.35
Multiplicação	53.14	0.124	-24.29	-22.37
Divisão	49.29	0.050	-28.15	-26.22

A hipótese para este resultado é que o uso de fusão tardia por meio de operações aritméticas é frequentemente favorecido porque os erros de vários classificadores tendem a não ser correlacionados, e o método é independente (Ramachandram and Taylor, 2017). Outro resultado interessante é que a operação "*Máximo*" se assemelha à operação de *pooling* máximo (*MaxPooling*), e a operação de "*Soma*" ocorre em muitas arquiteturas diferentes de aprendizado profundo (conexões residuais e codificação posicional do BERT). Os resultados indicam que a soma dos *embeddings* pode melhorar os resultados quando comparados com a operação "*Máximo*". Um teste de significância foi realizado usando *T-test* com um valor de p = 0.05 comparando o "*Subtração*" com abordagens unimodais (*BERTimbau* e *SE-ResNet50*). *Subtração* é significativamente diferente com *SE-ResNet50*, mas não com *BERTimbau*. Portanto, a operação de "*Subtração*" melhorou significativamente a modalidade visual.

Para complementar essa abordagem foram adicionadas duas estratégias neste trabalho: (i) normalização em algumas etapas no processo de fusão; (ii) e inserção de Redes de Compressão e Excitação, que serão descritas a seguir.

Fusão com Normalização

Conforme descrito no Capítulo 2 o aprendizado multimodal pode fornecer taxas de acerto superiores, quando comparadas a redes unimodais em determinados casos. Porém, outro fator de interesse observado no experimento descrito na Tabela 3.3, foi a adição da normalização em algumas etapas da fusão de dados, com resultados mais expressivos em relação as outras abordagens.

Computacionalmente a normalização consiste em regularizar os vetores de ativação das camadas ocultas. Como os computadores não possuem memória infinita para carregar todo conjunto de dados, os dados são carregados em lote, logo a normalização não é feita em todo o conjunto e sim lote por lote (Ioffe and Szegedy, 2015). Algumas bibliotecas já possuem nativamente funções e procedimentos para normalizar camadas em redes neurais sem a necessidade de implementação, como é o caso do *Pytorch* (Paszke et al., 2019a).

A normalização em lote, também chamada de *BatchNorm* (BN), utiliza o primeiro e segundo momento estatístico de uma rede neural para regularizar as camadas de interesse. O primeiro momento é chamado de média, e o segundo de variância. Por fim, há neste processo o cálculo do desvio padrão. A normalização é iniciada antes ou depois de uma função linear, conforme ilustra a Figura 3.5.

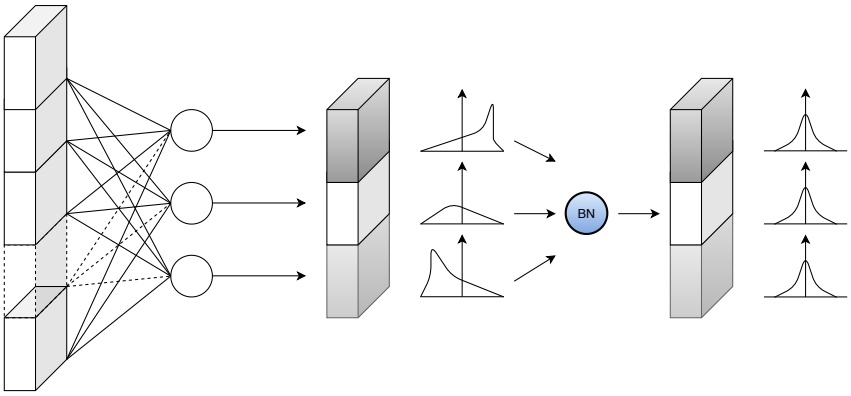


Figura 3.5: Sequência de passos para normalização das camadas em um rede neural.

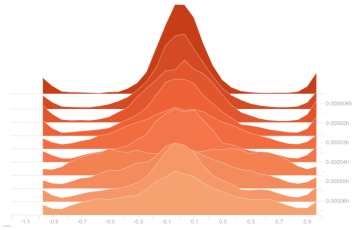
A Fórmula 3.10 descreve a normalização padrão em um conjunto de dados, onde E é igual a média dos elementos do conjunto (x), Var a variância dos elementos do conjunto (x), γ permite ajustar o desvio padrão e β ajustar o bias, deslocando a curva à direita ou à esquerda (Ioffe and Szegedy, 2015).

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \varepsilon}} * \gamma + \beta \tag{3.10}$$

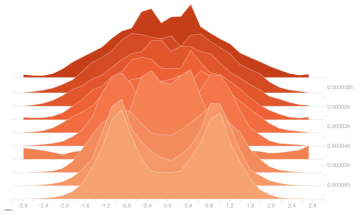
Em geral a normalização suaviza os valores em lotes permitindo que o modelo use taxas de aprendizado mais altas sem comprometer a convergência do treinamento, atua também como regulador em alguns casos eliminando a necessidade de *dropout*. Para analisar graficamente os resultados obtidos na Tabela 3.3 foi extraída representações de um experimento completo com 10 execuções em um *fold*.

Para este experimento a operação de "*Concatenação*" foi escolhida, pois representa a operação mais utilizada na literatura para fusão de dados. As representações são histogramas das saídas da última camada dos modelos de

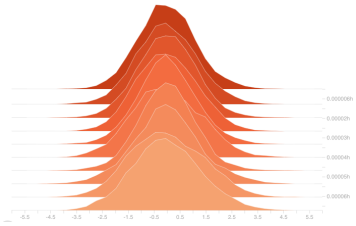
texto e visão, em seguida a aplicação da normalização nestas saídas e por fim as etapas de fusão acrescidas da normalização, representadas pela Figura 3.6. Há também uma representação em relação as predições de cada execução, ou seja, o comportamento dos dados a cada iteração na saída do modelo. Os eixos (x, y) representam respectivamente os *bins* e intervalo de tempo de cada execução, para esta tarefa foram usadas as bibliotecas *TensorBoard* (TensorFlow, 2020) e *Pytorch* (Paszke et al., 2019a).



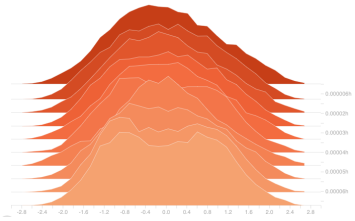
(a) Saída Texto



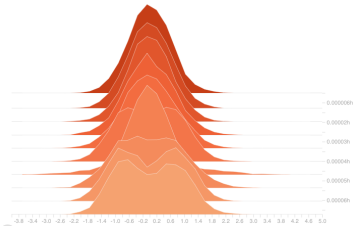
(b) Saída Texto com Normalização



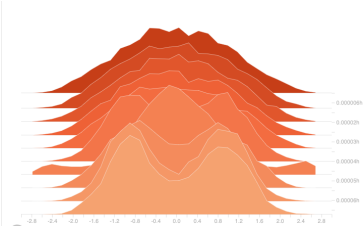
(c) Saída Imagem



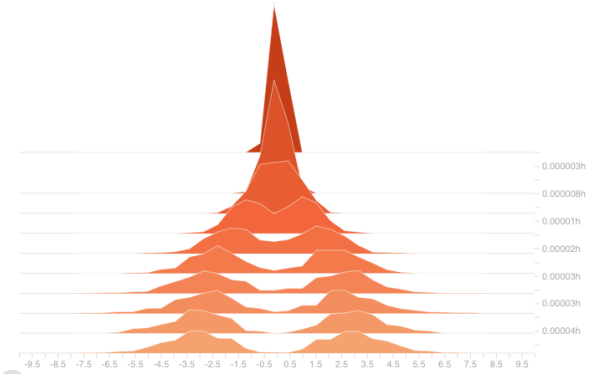
(d) Saída Imagem com Normalização



(e) Fusão



(f) Fusão com Normalização



(g) Predição

Figura 3.6: Representação da Normalização com operação de Concatenação.

Analisando as representações empiricamente, as informações de saída para o modelo de texto possuem uma distribuição simétrica com forma unimodal,

porém suas bordas possuem um pico, algo que parece ser característico do modelo. Ao normalizar seus dados o modelo sofre um achatamento com uma simetria mais suave. Já o modelo de imagem sofre menos impacto com a normalização, porém ao analisar sua curva sobre o eixo (x) antes da normalização seu desenho por vezes parece ser assimétrico. A normalização ajusta essa saída com uma distribuição mais simétrica, porém com uma variação menos achatada. A fusão dos dados segue a mesma linha do modelo de imagem, com uma suavização da curva e pouco impactado sobre a variação da distribuição. Por fim, a saída do modelo para predição apresenta uma distribuição com dois picos, ou seja, bimodal. Na primeira época esses dois picos tendem a estar interligados e na décima época estão mais esparsos um do outro devido ao treinamento do modelo durante as execuções.

Em geral o uso da normalização para este experimento melhorou o desempenho na construção da rede neural multimodal. Assim como em outros estudos da área (Ioffe and Szegedy, 2015; Ba et al., 2016; Ulyanov et al., 2016; Salimans and Kingma, 2016), a aplicação da normalização no contexto multimodal pode ser realizada nas saídas das camadas compartilhadas, com o objetivo de normalizar as ativações antes de serem passadas para as camadas subsequentes. Essa abordagem tem como propósito manter a distribuição das ativações estável durante o treinamento, reduzindo a covariância entre as modalidades e contribuindo para a redução do tempo de treinamento do modelo.

Redes de Compressão e Excitação

Uma outra abordagem relacionada nos experimentos deste trabalho é o uso de redes de compreensão e excitação, conhecidas como *Squeeze-and-Excitation Networks* (SE-Net), já citada na Seção 2.4. Elas incluem um bloco adicional na construção de CNNs tradicionais que propõe melhorar as interdependências dos canais, conforme ilustra a Figura 3.7. Logo, para as camadas de uma rede neural convolucional é possível construir um bloco *SE-Net* correspondente que recalibra os mapas de recursos. Essa abordagem foi utilizada no desafio de classificação de imagens (ILSVRC 2017) e obteve um ganho de 25% em relação ao modelo campeão em 2016.

Os aspectos envolvidos nas redes de compreensão e excitação estão relacionados da seguinte forma:

- **Entrada dos Dados** - Considere uma camada (X) da rede neural convolucional representada por uma altura H', largura W' e os canais C'. Sob o aspecto computacional, pode ser visto como uma matriz tridimensional em forma de tensores. Os canais nesta etapa podem ser considerados a profundidade de uma imagem, por exemplo fontes visuais em RGB.

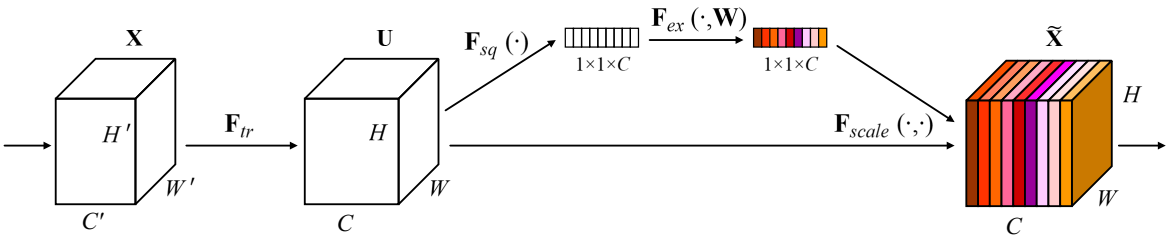


Figura 3.7: Bloco *SE-Net*. Fonte: (Hu et al., 2018)

- **Camada Convolucional** - Na próxima etapa ocorre uma transformação F_{tr} dos tensores de entrada e aplicação de alguns filtros, também compreendida como uma camada convolucional.

- **Extração de Características** - A camada (U) é resultante da transformação F_{tr} e pode ter tamanhos equivalentes ou modificados em relação a camada de entrada, logo é representado agora por altura H, largura W e os canais C. Os canais nesta fase podem ser compreendidos pelo mapa de atributos extraídos pelos filtros de convolução da etapa anterior.

- **Componente *Squeeze*** - Esta etapa compreende modificar o relacionamento das camadas subsequentes em relação aos filtros de convolução gerados pela transformação F_{tr}. O componente *squeeze* recalibra a rede e analisa as entradas com pesos e probabilidades de forma individual. Neste sentido, ao invés de aplicar pesos de forma genérica, esta técnica analisa a interdependência entre as camadas. Logo, a camada (U) passa por um processo de redução de dimensionalidade pela função $F_{sq}(.)$, na qual cada canal da camada (U) corresponde a um único valor. A função $F_{sq}(.)$ pode ser representada por uma medida estatística (média, máximo, mínimo) com saída unidimensional $[1 \text{ x } 1 \text{ x } C]$.

- **Componente *Excitation*** - Este componente faz com que a relação de ordem entre as camadas seja mais representativa, para isso a saída unidimensional gerada pelo componente *squeeze* é agora processado por um módulo $F_{ex}(.,W)$. Esse módulo é ilustrado pela Figura 3.8, e descreve a sequência de passos do componente *excitation*, que pode ser incorporada em qualquer rede convolucional profunda. A ilustração abaixo mostra o uso de um bloco SE-Net no modelo *Inception* e *ResNet* como exemplo. Para fazer uso das informações agregadas na operação *squeeze*, a função deve ser capaz de relacionar uma interação não linear entre canais e aprender uma relação não mutuamente exclusiva.

  O módulo da operação $F_{ex}(.,W)$ é representado por uma operação de *pooling* e ligado a uma camada totalmente conectada, seguida por uma

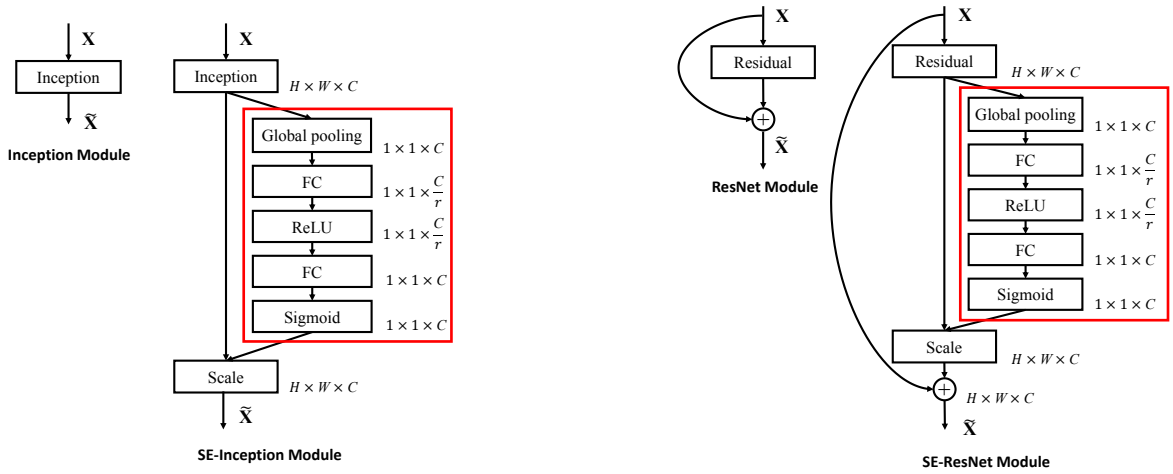


Figura 3.8: Operações em um bloco *SE-Net*. Fonte: (Hu et al., 2018).

função de ativação ReLU com o critério de anular/zerar todos os valores negativos da saída da camada anterior. O próximo passo depois da função de ativação ReLU é conectar os neurônios/tensores novamente a uma camada totalmente conectada com tamanho de canal igual a camada de entrada (X) com bloco de dimensão $[1 \text{ x } 1 \text{ x } C]$. Veja que a primeira camada conectada e ReLU podem ter dimensões diferentes da camada de entrada (X), pois compreendem um bloco $[1 \text{ x } 1 \text{ x } \frac{C}{r}]$. O último passo é aplicar novamente uma função de ativação sigmoide, pois diferente da ativação ReLU ela normaliza os valores de saída para escala $\{0 <= x <= 1\}$.

- **Camada de Saída** - Realizadas as operações pelo bloco *SE-Net* é gerada uma camada de saída representada pela multiplicação do valor resultante do bloco *SE-Net* pela camada de entrada (X). Essa etapa é descrita com uma função F_{scale} que retorna uma camada de saída $(\tilde{X})$.

Essa técnica pode ser empregada no uso de fusão de dados, pois esses blocos podem ser empilhados nas redes unimodais ou simplesmente aplicados em uma fusão tardia, pois se generalizam de forma extremamente eficaz em diferentes conjuntos de dados (Hu et al., 2018).

Um experimento inicial foi desenvolvido para verificar a eficácia dos blocos *SE-Net* em modelos unimodais. A primeira etapa foi analisar a taxa de ganho para o modelo *ResNet50*, aqui chamado de *SE-ResNet50*. A Tabela 3.5 mostra os resultados em termos das métricas escolhidas (*ROC-AUC, Accuracy, F1, Recall, Precision e ROC-AUC*). Os parâmetros utilizados são os mesmos abordados na Tabela 3.2 e a implementação usada foi desenvolvida por Hatay (2019).

Tabela 3.5: Operações com blocos *SE-Net* usando validação cruzada de 10 partições para o conjunto de dados *Top Speed*.

Modelos	***Loss***	***Accuracy***	***Precision***	***Recall***	***F1***	***ROC-AUC***
ResNet50	0.645 ± 0.037	0.684 ± 0.055	0.708 ± 0.061	0.682 ± 0.057	0.672 ± 0.065	0.682 ± 0.057
SE-ResNet50	0.588 ± 0.046	0.755 ± 0.067	0.758 ± 0.067	0.755 ± 0.066	0.754 ± 0.066	0.755 ± 0.066

Os resultados indicam que a inserção do bloco *SE-Net* no modelo *ResNet50* foi eficaz em todas as métricas avaliadas, pois há uma melhora significativa em relação ao modelo padrão *ResNet50*. Os próximos experimentos deste estudo visam inserir mecanismos de atenção nas operações de fusão, pois a expectativa é que o conjunto de todas essas técnicas aliadas as redes multimodais possam responder as perguntas mensuradas na Seção 1.3.

Este estudo foi publicado no "*Symposium on Knowledge Discovery, Mining and Learning* (KDMiLe 2022)", sob o título "*Successful Youtube video identification using multimodal deep learning*", a conferência ainda não publicou sua versão digital para o público.

3.2 Experimento 1.1 - Fusão de Dados por meio de Operações Aritméticas, Conexões Residuais e Mecanismos de Atenção

Esta seção apresentará um experimento detalhado sobre as técnicas de fusão de dados por meio das operações aritméticas, conexões residuais e mecanismos de atenção. Este experimento explora dez mecanismos de atenção encontrados na literatura e estratégias empregadas nessas abordagens. Além disso, este estudo utiliza cinco conjuntos de dados de domínios distintos para avaliar três abordagens multimodais empregadas neste experimento, visando a avaliação dessas técnicas.

Experimento 2 - MASK: Uma abordagem de convergência rápida usando conexões de Atenção Multimodal + SKip

Ao lidar com o aprendizado multimodal surge um conjunto de perguntas: Qual é a melhor estratégia para fundir os dados multimodais? Quanta melhoria é possível alcançar com operações aritméticas simples e mecanismos de atenção? Para responder a essas perguntas, uma avaliação extensiva foi realizada com abordagens de fusão multimodal familiares e inexploradas para mesclar informações textuais e visuais em tarefas de classificação. Em particular, um estudo exploratório com uma rede de fusão multimodal baseada em operações aritméticas e mecanismos de atenção foi utilizado para avaliar modelos multimodais e unimodais em cinco conjuntos de dados multimodais

de diferentes domínios. Este resultado experimental revela que é possível alcançar uma convergência mais rápida ao combinar o Aprendizado de Atenção Multimodal com conexões *Skip Connection*, chamada neste trabalho de MASK.

A literatura destaca diversas combinações que podem ser representadas no aprendizado multimodal. A Figura 3.9 descreve de forma simplificada os aplicativos e fontes correlacionadas para o uso de aprendizado profundo multimodal extraído do trabalho de Guo et al. (2019). Existe uma categorização para os métodos de representação multimodal profunda: (i) representação conjunta, que mantém as representações unimodais com a possibilidade de mesclar as características multimodais; (ii) representação coordenada, que visa aprender representações separadas para cada modalidade em um subespaço coordenado; (iii) modelos de codificador-decodificador, usados para gerar uma representação intermediária e mapear uma modalidade em relação a outra; (iv) representação multimodal usando operações aritméticas e mecanismos de atenção para unificar modalidades de dados e gerar uma representação compartilhada.

	Aplicações	Texto	Imagem	Vídeo	Áudio
Representação Conjunta	Classificação de Vídeo			✓	✓
	Detecção de Evento	✓		✓	✓
	Análise de Sentimento	✓		✓	✓
	Resposta a perguntas visuais	✓	✓		
	Reconhecimento de emoções	✓		✓	✓
	Reconhecimento de fala			✓	✓
Representação Coordenada	Recuperação Cross-modal	✓	✓		
	Legenda da imagem	✓	✓		
	Incorporação Cross-modal	✓	✓	✓	
	Transferência de aprendizagem	✓	✓		
Codificador-Decodificador	Legenda da imagem	✓	✓		
	Descrição de Vídeo	✓		✓	
	Síntese de texto para imagem	✓	✓		
Operações Aritméticas com Atenções Proposta	Classificação de Texto	✓	✓		
	Classificação de Imagem	✓	✓		
	Recuperação Cross-modal*	✓	✓		
	Incorporação Cross-modal*	✓	✓		
	Transferência de aprendizagem	✓	✓		

*Domínios com diversidade de dados.

Figura 3.9: Em cada aplicação, a categorização de métodos de representação multimodal profunda pode incluir algumas das modalidades como: áudio, vídeo, imagem e texto. Adaptado de Guo et al. (2019).

Esses métodos de representação multimodal profunda podem ser correlacionados aos dados usando três tipos básicos de fusão. De acordo com Ramachandram and Taylor (2017), arquiteturas profundas se conectam à fusão multimodal por fusão precoce, intermediária ou tardia. A fusão precoce é

uma estratégia usada para mesclar as várias fontes de dados na entrada de uma rede neural, seguida da aplicação de um único modelo de aprendizado. A fusão intermediária é adotada na maioria dos trabalhos que utilizam redes neurais profundas multimodais, pois compreende o uso de modelos unimodais que possuem o melhor desempenho para cada entrada de dados específica (texto, áudio, imagem, vídeo) em suas camadas iniciais. Por fim, a fusão tardia compreende o uso de treinamento em redes unimodais e uma fusão multimodal nas últimas camadas do modelo. O uso da fusão de dados abrange vários domínios de aplicação e permite a combinação de várias fontes de dados, conforme ilustrado na Figura 3.10.

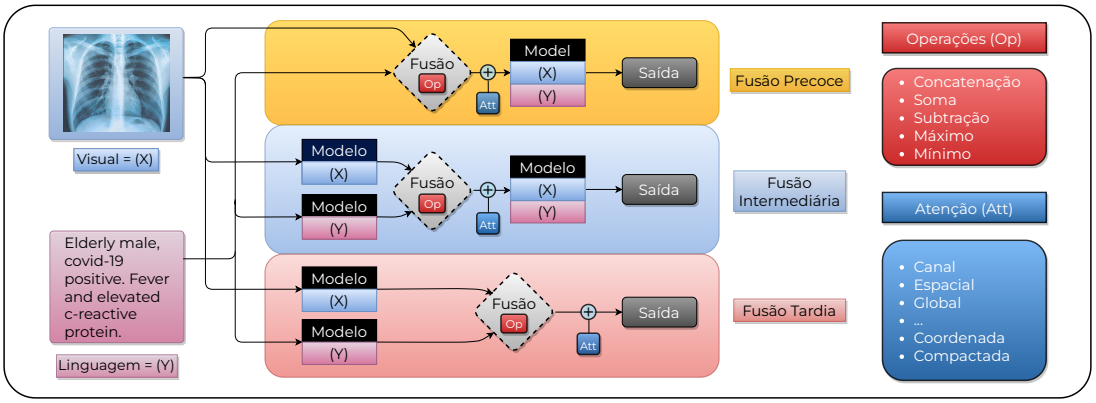


Figura 3.10: Modelos de Fusão para o Aprendizado Multimodal com Operações Aritméticas e Mecanismos de Atenção.

Outra abordagem eficaz na área de Aprendizado Profundo que tem ganhado cada vez mais espaço nas pesquisas de ponta é a utilização dos Modelos de Atenção que utilizam características obtidas de diferentes partes de uma rede para resolver um problema sob diversas perspectivas (de Santana Correia and Colombini, 2022). Basicamente a atenção pode ser classificada em dois estágios (Chen et al., 2021b): (i) atenção global, na qual o modelo foca todas as entradas consideradas igualmente importantes para a tomada de decisão em uma rede neural; (ii) Atenção local, em que a rede dá mais atenção a certas partes da entrada, priorizando informações importantes e ignorando informações irrelevantes.

Aliado a esse contexto o uso de conexões residuais, também chamado do *Skip Connection*, pode auxiliar nos modelos de atenção para que as informações obtidas nas primeiras camadas não sejam perdidas durante o processamento da rede, garantindo que a rede não perca informações e não sofra degradação do gradiente (He et al., 2016). Nesse contexto, explorar as capacidades da rede multimodal com modelos de atenção pode agregar informações interessantes e aumentar a taxa de desempenho (*Accuracy*) para problemas tratados com aprendizado unimodal (Niu et al., 2021). A primeira implementação bem-sucedida do *Skip Connection* foi proposta por He et al. (2016), na

qual implementaram o conceito chamado *Residual Network* (*ResNet*) com blocos residuais para adição de atalhos. Essa estratégia adiciona uma transformação de identidade à entrada, permitindo que informações relevantes não sejam perdidas durante o processo de convolução em redes recorrentes. Outro uso eficaz é capturar informações de várias camadas em diferentes níveis de granularidade. Da *ResNet*, surgiram várias propostas que exploram o uso do *Skip Connection* em diferentes tipos de redes neurais *DenseNet* (Huang et al., 2017), *Transformers* (Vaswani et al., 2017), *BERT* (Devlin et al., 2018), *ViT* (Dosovitskiy et al., 2020), com resultados promissores.

Meel and Vishwakarma (2021) definiram tarefas de classificação usando a fusão multimodal de ramificações textuais e visuais. Os autores fornecem uma análise detalhada sobre a concatenação de vetores de características e a média ponderada de probabilidades para resolver seu problema, mas não consideram outras operações aritméticas na composição de suas pesquisas que possam contribuir para o aprendizado multimodal. Baltrušaitis et al. (2018) também investigam os vários desafios enfrentados pelo aprendizado multimodal com um ponto de partida na representação do modelo, sua estrutura, fusão de dados e co-aprendizagem. Guo et al. (2019) traz algumas reflexões sobre como utilizar modelos codificador-decodificador, redes geradoras adversárias e mecanismos de atenção no aprendizado da representação multimodal. Portanto, em vez de focar na estrutura básica, aprendizado e nas cenas de aplicação, este estudo possui enfoque em tarefas usando operações aritméticas e mecanismos de atenção. Três abordagens diferentes foram usadas para identificar as melhores combinações entre operações aritméticas, modelos multimodais e mecanismos de atenção.

Proposta

Basicamente é possível dividir o método proposto em três etapas. A primeira etapa consiste em usar modelos unimodais para extrair características relevantes de cada modalidade. Na segunda etapa, as características de cada modalidade são mescladas por meio das operações aritméticas (*Soma, Subtração, Máximo e Mínimo*), além da concatenação que é predominantemente utilizada na literatura. Dadas as fusões decorrentes das operações aritméticas, os mecanismos de atenção são aplicados na terceira etapa para correlacionar as funcionalidades em um único modelo capaz de identificar quais informações são relevantes selecionando entre os canais de cada operação. Os modelos utilizados neste trabalho são amplamente utilizados em diversas áreas de pesquisa de ponta, por isso foram categorizadas como extratores de características (*embeddings*) de última geração para cada modalidade. Para entrada textual, o modelo *BERTimbau* com 12 camadas de codificador em-

pilhadas umas sobre as outras foi utilizado e, para entrada visual, o modelo *ViT vit_base_patch16_224* pré-treinado em *ImageNet-21k* (Ridnik et al., 2021) (14 milhões de imagens, 21.843 classes) em resolução 224x224 e ajustado no conjunto de dados de imagens *ImageNet* 2012 (Krizhevsky et al., 2012) (1 milhão de imagens, 1.000 classes) em resolução 224x224. A Tabela 3.6 resume as informações e parâmetros estabelecidos para a pesquisa realizada neste artigo. As três etapas também são detalhadas nos tópicos abaixo.

Tabela 3.6: Visão geral dos modelos selecionados para linguagem e visão pré-treinados.

Modelos	**Parâmetros**	**Camadas**	**Tamanho Camadas Ocultas**
BERTimbau base	110M	12	768
ViT base	86.7M	12	768

O primeiro modelo compreende a execução da fusão intermediária através de operações aritméticas, neste cenário é avaliado a eficácia de cada operação individual nos modelos unimodais. No segundo experimento são adicionados alguns mecanismos de atenção nas operações aritméticas já descritos na Seção 2.4, esta etapa realiza a fusão intermediária para cada operação e armazena sua saída em um canal, então um mecanismo de atenção é utilizado para manipular as saídas obtidas pelas operações. Por fim, o último experimento também faz uso dos mecanismos de atenção, mas utiliza a técnica *Skip Connection* (He et al., 2016; Targ et al., 2016) para ignorar algumas conexões dentro do modelo e medir sua eficácia em relação aos outros experimentos realizados, denominado Atenção Multimodal com *Skip Connection* (MASK). Este estudo prevê que o último modelo é altamente promissor quando comparado a modelos que não utilizam *Skip Connection* ou apenas são fundidos por operações aritméticas.

Para fundir os dois modelos em suas camadas finais uma configuração simples descrita na Figura 3.11 foi aplicada. Especificamente, optou por ajustar as últimas camadas de cada modelo unimodal removendo o agrupamento final (*AvgPooling*), as normalizações e as camadas totalmente conectadas. Essa abordagem permitiu maior flexibilidade no compartilhamento de informações entre os modelos, visto que ambos são baseados em *Transformers*: (i) para o modelo *BERT*, foi utilizado um método de *pooling* (*AvgPooling*) que combinou suas 12 camadas de atenção; (ii) da mesma forma, reduziu-se a dimensionalidade das *embeddings* de *token/patch* no modelo *ViT* para corresponder ao tamanho de saída do modelo *BERT*; (iii) Em última análise ambos os modelos foram configurados para produzir tensores do mesmo tamanho, [Tamanho do lote, 64,768], com exceção da dimensão do tamanho do lote, que variava dependendo das fontes de dados empregadas para cada modelo.

Operações Aritméticas - (Op) - Figura 3.12a ilustra o modelo do aprendi-

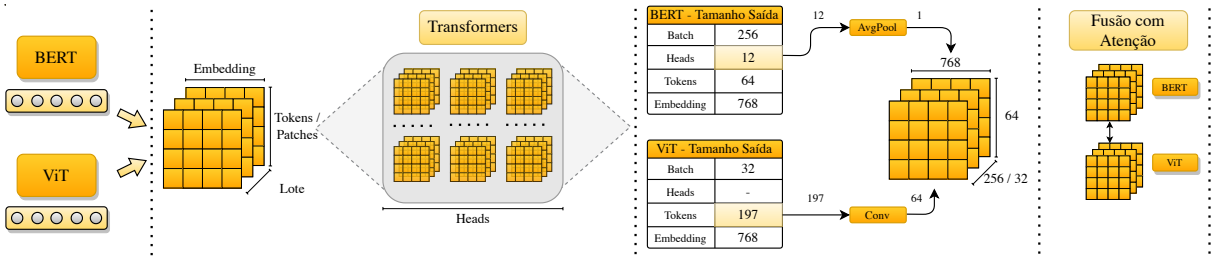


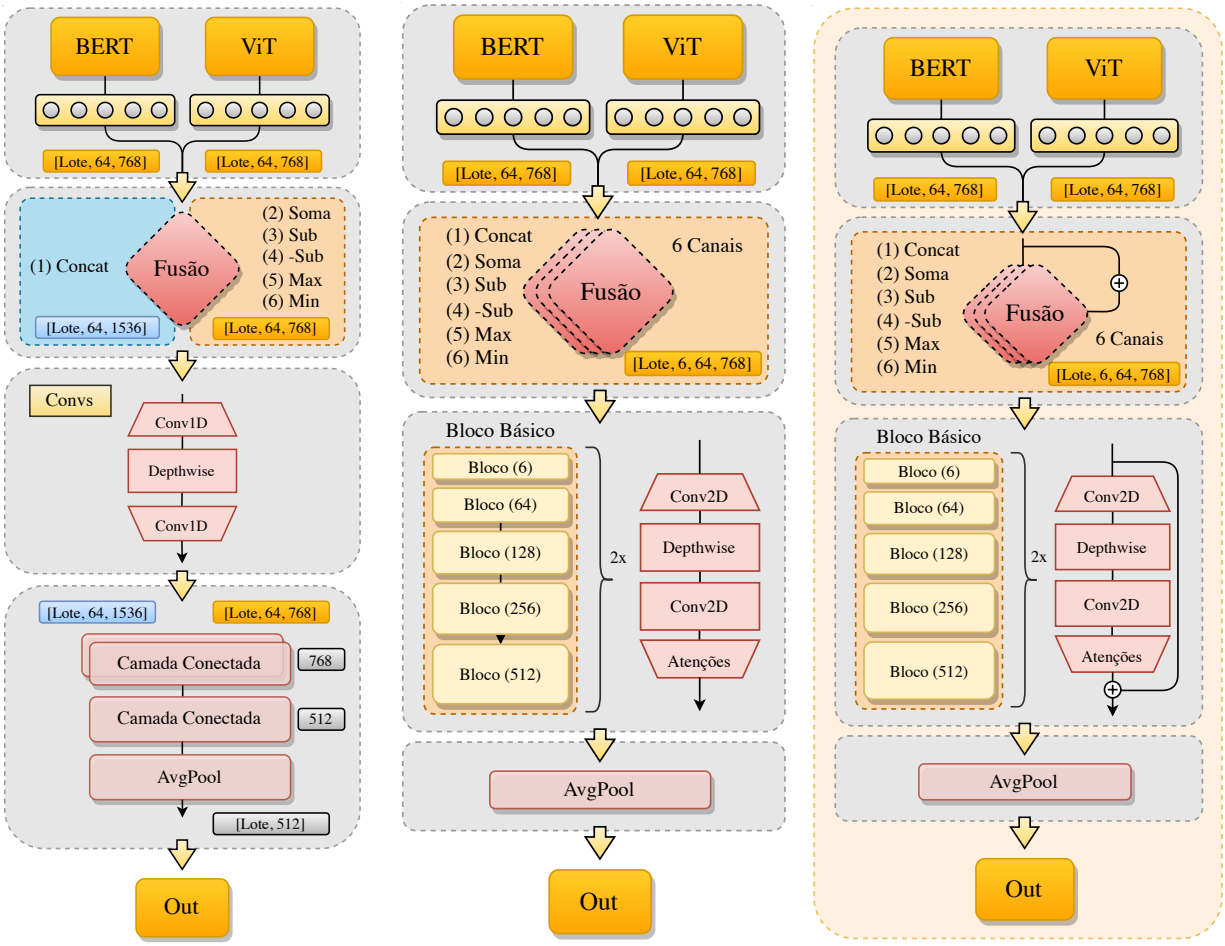
Figura 3.11: Configuração dos modelos *BERT* e *ViT*

zado multimodal com a aplicação da técnica de fusão intermediária em todas as combinações geradas. Inicialmente os modelos unimodais são executados em paralelo e nas camadas subsequentes, a fusão ocorre por uma das seis operações. As saídas dos modelos unimodais são representadas respectivamente por: (i) 12 camadas de atenção do modelo *BERT* com dimensão $[64 \times 768]$, (ii) 197 canais do modelo *ViT* com dimensão $[768]$. Para permitir que os modelos tenham a mesma dimensão alguns passos antes da fusão foram predefinidos, há a remoção de uma dimensão do *BERT* (camadas de atenção) através da operação *AvgPool* e redução dos canais de $[197]$ para $[64]$ no modelo *ViT*, então os dois modelos têm a mesma dimensão $[Batch, 64, 768]$. Em seguida, um bloco de convolução com a adição de uma convolução de profundidade é adicionado, uma técnica recente utilizada em diversos trabalhos (Woo et al., 2018). Por fim, camadas totalmente conectadas são adicionadas seguidas de um agrupamento médio na saída da rede neural.

Mecanismos de Atenção sem Skip Connection (Att) - O segundo modelo lista todas as operações aritméticas em canais individuais, então o modelo adiciona um dos dez tipos de mecanismo de atenção já mencionados na Seção 2.4. Conforme ilustrado na Figura 3.12b, neste modelo existe uma estrutura básica de blocos que utiliza a mesma ideia implementada no trabalho de Tan and Le (2021). Em cada bloco básico uma sequência de convoluções mais um mecanismo de atenção em sua saída é introduzido, cada bloco é executado duas vezes e, finalmente, um agrupamento médio *AvgPool* é adicionado à saída da rede neural.

Mecanismos de Atenção com Skip Connection (MASK) - O terceiro modelo tem as mesmas características do segundo modelo, sua diferença está na utilização do blocos residuais utilizando a técnica *Skip Connection* (He et al., 2016; Targ et al., 2016), ignorando algumas conexões da rede durante o processo de treinamento do modelo. Na Figura 3.12c, esta técnica é empregada com aplicação da fusão intermediária, após cada operação é adicionado uma conexão residual. Para criar a matriz identidade a estratégia utilizada foi somar a saída dos dois modelos unimodais. Há fortes indícios que este terceiro modelo pode obter vantagens sobre o primeiro e o segundo modelo, pois sua característica é generalizar melhor a rede durante as etapas de fusão com os

mecanismos de atenção, eliminando algumas conexões.



(a) operações aritméticas (b) Atenções (c) MASK

Figura 3.12: Modelos Multimodais: (a) Operações Aritméticas, (b) Mecanismos de Atenção, (c) Mecanismos de Atenção com *Skip Connection*.

As operações aritméticas utilizadas neste experimento são as mesmas já descritas no Experimento (1) 3.1, porém operações que não tiveram um desempenho significativo foram suprimidas neste trabalho (Divisão, Multiplicação e Potenciação). Houve também a duplicação da operação de *Subtração* $(-sub)$ invertendo a ordem das modalidades neste trabalho, pois a ordem afeta o resultado do cálculo e sua representação ao mesclá-las na rede neural. Observe que para as operações *Máximo, Mínimo, Subtração, -Subtração e Soma*, as dimensões das *embeddings* de entrada devem ser as mesmas e neste trabalho são $n = 768$ e $d = 768$.

Conjunto de Dados

Cinco conjuntos de dados compostos por textos e imagens de diferentes domínios foram avaliados, são conjuntos coletados de pesquisas recentes que possuem dados multimodais em larga escala. A seguir, há um breve resumo de cada conjunto de dados e suas aplicações:

Top Speed - É um conjunto de dados obtido de uma popular plataforma de vídeos na internet (*Youtube*) e possui uma série de dados rotulados (miniaturas, título do vídeo, descrição, tempo de publicação, visualizações, CTR, exibição tempo, assinantes, curtidas, desgostos e comentários). Seu conteúdo aborda questões do automobilismo e as novas tendências do mercado. Tivemos autorização do especialista *Youtuber* para obter os dados para prever através de *thumbnails* e descrição do vídeo o impacto no interesse das pessoas. A proposta apresentada utiliza o número de visualizações como atributo de classe, vídeos com mais de 100 mil visualizações são rotulados como "sucesso" e os menores com "sem sucesso". Esse limite foi definido pelo especialista, pois vídeos com mais de 100 mil visualizações têm maior monetização exclusivamente para esse canal.

ChestXRay - Um conjunto de dados público aberto de radiografias de tórax e imagens de tomografia computadorizada de pacientes positivos ou suspeitos de COVID-19, também existem outras doenças, como pneumonias virais e bacterianas (MERS, SARS e ARDS) (Cohen et al., 2020; Maguolo and Nanni, 2021; Tartaglione et al., 2020). São imagens coletadas de fontes públicas por hospitais e médicos, e juntamente com as imagens, estão disponíveis uma série de dados sobre os pacientes (sexo, idade, temperatura, PCR, saturação, laudo médico, entre outras informações). Existem também várias segmentações pulmonares, escores de gravidade de pneumonia, escores de *brixia*, caixas de limites pulmonares e segmentação de raios X de tórax. O objetivo inicial desse conjunto de dados é melhorar as previsões de doenças pulmonares no estágio de triagem e gerenciar o atendimento ao paciente. Este conjunto de dados originalmente tem cinco classes (Vírus, Bactéria, Fungo, Aspiração Lipoide, Desconhecido), mas para fins de avaliação, o conjunto de dados foi redefinido para duas classes (positivo, negativo) para pacientes diagnosticados com COVID-19.

Disaster - Este conjunto de dados reúne uma coleção de imagens e postagens relacionadas a desastres naturais, danos à infraestrutura, risco ambiental e identificação de vítimas humanas (Mouzannar et al., 2018). São dados obtidos diretamente das redes sociais capazes de transmitir informações e localizar sobreviventes em situações de desastres. A ideia central desse conjunto de dados é automatizar a extração de informações por meio de postagens em redes sociais para direcionar recursos de socorro de forma eficiente. O conjunto de dados possui seis classes para detectar (infraestrutura danificada, natureza danificada, incêndios, inundações, danos humanos e sem danos).

Fashion-iq - De acordo com os criadores deste conjunto de dados, *Fashion-iq* é um conjunto de dados de moda com legendas geradas por humanos para diferenciar pares semelhantes de imagens de roupas junto com informações

secundárias na palavra de descrições de produtos e rótulos de atributos visuais (Wu et al., 2021). O principal objetivo deste conjunto de dados é apoiar a pesquisa sobre recuperação interativa de imagens de moda. As imagens são representadas por três categorias de roupas (vestido, camisa, top).

ROCO - *Radiology Objects in Context* (ROCO) é um conjunto de dados de imagens publicamente disponíveis do *PubMed Central Open Access* (Walport and Kiley, 2006), que foram automaticamente analisadas e definidas como radiológicas ou não radiológicas (Pelka et al., 2018). A abordagem deste conjunto de dados é construir modelos generativos para legendas de imagens, modelos de classificação para categorização, marcação de imagens ou sistemas de recuperação de imagens baseados em conteúdo.

A Tabela 3.7 apresenta uma visão geral dos conjuntos de dados usados nesse experimento. Há uma descrição sobre a quantidade de exemplos disponíveis para cada partição (treinamento, validação e teste) usada nos experimentos de estudo, menção de quais conjuntos de dados foram estratificados. Depois do pré-processamento dos dados há a remoção de exemplos com dados ausentes. Alguns ajustes foram necessários para a construção dos conjuntos de dados: (i) a descrição dos vídeos foi usada como entrada textual para o conjunto de dados *Top Speed*, (ii) os relatórios médicos e notas clinicas técnicas foram unificadas para compor a entrada textual do conjunto de dados *ChestXRay*, (iii) para o conjunto de dados *Fashion-iq*, foi usada as imagens categorizadas como (candidata) e unificação da entrada textual com os dados (*caption0* e *caption1*), (iv) Para o conjunto de dados de *Disaster*, as postagens publicadas na rede social *Twitter* sobre desastres foram coletadas e segmentadas em imagem (entrada visual) e *tweets* (entrada textual), (v) finalmente, o conjunto de dados ROCO teve as variáveis (legendas, palavras-chave e *semtypes*) como entrada textual.

Tabela 3.7: Estatísticas dos conjuntos de dados de treinamento, validação e teste.

Conjuto de Dados	Treino	Validação	Teste	Total	Classes	Estratificado	Dado Texto	Dado Visual
Top Speed[1]	325	70	70	465	2	Sim	Título e descrição do vídeo	*Thumbnails*
ChestXRay (Tartaglione et al., 2020)	522	111	112	745	2	Sim	Laudos médicos e clínicos	*X-ray* and CT
Disaster (Mouzannar et al., 2018)	4081	875	875	5831	6	Sim	Postagens publicadas	Imagens publicadas
Fashion-iq (Wu et al., 2021)	18000	6118	6016	30134	3	Sim	Descrição do bate-papo	Imagem do bate-papo
ROCO (Pelka et al., 2018)	68738	8572	8560	85871	2	Não	Legendas, palavras-chave e *semtypes*	Imagem radiológica

[1] Nota: *Top Speed* é um conjunto de dados privado para o público em geral.

Pré-processamento e Configuração Experimental

Tabela 3.8 descreve os parâmetros usados para os conjuntos de dados de treinamento. Um teste de intervalo de taxas de aprendizado (Smith, 2017) foi realizado para encontrar a taxa de aprendizado mais alta que minimiza a perda e não faz com que o gradiente exploda. Essa taxa foi incorporada à política de

um ciclo (Smith and Topin, 2019) durante o treinamento. Os dados ausentes foram removidos dos conjuntos de dados e divido respectivamente [70%, 15%, 15%] para partições de treinamento, validação e teste, respectivamente. Um implementação empírica para política de parada (Prechelt, 1998) foi elaborada durante a etapa de treinamento, garantindo que o modelo execute um número suficiente de épocas de treinamento.

Tabela 3.8: Parâmetros.

Parâmetros	**BERT**	**ViT**	**Multimodal**
Tamanho do Lote	256	64	16
Épocas	100	100	100
Early Stop	14	14	14
Otimizador	AdamW	AdamW	AdamW
Agendador	One Cycle LR	One Cycle LR	One Cycle LR
Taxa de Aprendizado	3e-3 <n <3e-2	3e-4 <n <3e-3	3e-4 <n <3e-3
Weight Decay	1e-02	1e-02	1e-02
Criterion	CrossEntropy	CrossEntropy	CrossEntropy
Tokenizer	WordPiece	-	WordPiece
Sequência do *Token*	64	-	64
Resolução Imagem	-	224x224	224x224

Os modelos foram implementados em *PyTorch* (Paszke et al., 2019b) e os experimentos foram realizados em quatro GPUs GeForce GTX 1080 Ti individuais com uso de memória limitado a 64GB. Os modelos unimodais pré-treinados estão listados na biblioteca de código aberto *HuggingFace* (Wolf et al., 2019) e pelo pesquisador Ross Wightman (Wightman, 2019). mecanismos de atenção foram extraídos do repositório (Xiaoma, 2022) com uma série de publicações envolvendo modelos de atenção (Ma et al., 2022a; Ji et al., 2022; Ma et al., 2022b).

Resultados Experimentais

A Tabela 3.9a e Tabela 3.9b mostram os resultados em função das métricas escolhidas (*Accuracy* e *ROC-AUC*). Ao usar apenas estas métricas, o estudo propõe enfatizar o desempenho dos modelos em relação aos conjuntos de dados avaliados e a capacidade de um modelo em distinguir as classes de um problema. Os resultados obtidos em alguns cenários mostraram-se descalibrados, o que levou a algumas interpretações erradas. Portanto, houve a necessidade em manter a *Accuracy* como uma métrica mais conhecida, mas as análises foram realizada utilizando a métrica *ROC-AUC*. Primeiro, os resultados entre os modelos de texto (*BERTimbau*) e imagem (*ViT*) indicam que as imagens são mais discriminatórias do que o conteúdo textual nesta configuração para a maioria dos conjuntos de dados. Em segundo lugar, a análise revela que os modelos multimodais com fusão intermediária superam os modelos unimodais em todas as operações aritméticas.

Ao inserir os mecanismos de atenção é possível obter resultados semelhantes aos das operações aritméticas, porém sem a necessidade de avaliar individualmente cada operação, neste cenário a própria rede é responsável por gerar uma representação conjunta entre os canais de fusão através dos mecanismos de atenção. Segundo o estudo de Niu et al. (2021) os resultados obtidos nesta avaliação podem ser denotados por duas combinações: (i) o mecanismo de atenção pode resolver o problema de sobrecarga de informação através de um esquema de alocação de recursos; (ii) e processar informações mais relevantes por meio de canais de atenção.

Ao analisar os mecanismos de atenção os resultados indicam que o uso do *Skip Connection* auxilia na convergência do modelo, pois informações captadas nas camadas iniciais podem auxiliar camadas posteriores no aprendizado. Esta proposta pode ser vista nos trabalhos de He et al. (2016) e Huang et al. (2017) com os modelos *ResNet* e *DenseNet*. Nesse sentido, o uso do *Skip Connection* pode fornecer um caminho alternativo para o gradiente com retropropagação, pois através da função identidade usando apenas uma adição de vetores é possível preservar o gradiente. No entanto, o conjunto de dados *ChestXRay* não obteve resultados satisfatórios usando *Skip Connection*, uma suposição para esta análise é que o domínio sem uso das segmentações e pontuações disponíveis para imagens médicas em sua documentação oficial torna o conjunto de dados pobre em informações relevantes. A configuração para este experimento analisou apenas as imagens brutas sem adicionar informações secundárias, mesmo assim, o desempenho dos mecanismos de atenção foi bem-sucedido sobre os modelos unimodais na maioria dos conjuntos de dados.

Além disso, como parte desta avaliação, há uma comparação significativa para os ganhos de classificação, Tabela 3.9a e Tabela 3.9b mostram a diferença entre resultados unimodais versus multimodais. Esta análise compara o melhor modelo multimodal com modelos unimodais (texto e imagem) para cada conjunto de dados individualmente. A proposta alcançou uma variação de 0,70% a 22,10% de ganho em *Accuracy* (ambas as modalidades) e 0,50% a 12,80% para *ROC-AUC*. Neste sentido, o uso da fusão intermediária por meio de operações aritméticas e mecanismos de atenção pode gerar uma representação mais rica ao modelo, pois os erros de vários classificadores tendem a não ser correlacionados e o método é independente (Ramachandram and Taylor, 2017). Usando *ROC-AUC* as melhorias médias sobre texto e visão são 6,35 e 4,2, respectivamente. No entanto, a escolha errada da operação de fusão pode tornar os resultados piores do que os modelos unimodais.

Tabela 3.9: Métricas (*Accuracy* e *ROC-AUC*) entre Modelos Unimodais, Operações Aritméticas e Mecanismos de Atenção.

Accuracy					
Modelos	**Top Speed**	**ChestXRay**	**Disaster**	**Fashion-iq**	**ROCO**
BERT	0.757	0.848	0.851	0.708	0.947
ViT	0.686	0.866	0.903	0.910	0.966
Operações aritméticas					
CONCAT	0.786	0.813	0.887	0.921	**0.973**
MAX	0.771	0.839	0.899	0.914	0.972
MIN	0.771	0.821	**0.913**	0.918	0.949
-SUB	**0.829**	0.866	0.902	0.921	0.972
SUB	0.786	**0.875**	0.905	0.922	0.969
SOMA	0.786	0.804	0.895	**0.923**	0.972
Mecanismos de Atenção sem *Skip Connection*					
CBAM	0.714	0.866	0.880	0.900	0.953
COOATT	0.800	0.920	**0.915**	0.895	0.965
ECA	0.771	0.938	0.854	0.913	0.966
PNA	0.714	0.875	0.817	0.884	0.964
PSA	0.814	0.929	0.874	**0.915**	0.939
RAN	0.700	0.634	0.831	0.877	0.954
S2ATT	0.729	0.625	0.525	0.437	0.939
SENET	**0.814**	**0.973**	0.769	0.914	0.956
SHA	0.714	0.964	0.886	0.868	**0.969**
TRIPLEATT	0.743	0.955	0.893	0.903	0.960
Mecanismos de Atenção com *Skip Connection*					
CBAM	0.829	0.875	0.913	0.928	**0.973**
COOATT	**0.829**	0.866	0.910	0.922	0.972
ECA	0.814	0.848	0.912	0.927	0.971
PNA	0.814	0.884	0.907	0.925	0.972
PSA	0.814	0.866	0.910	**0.929**	0.972
RAN	0.814	0.857	0.915	0.929	0.971
S2ATT	0.757	0.848	**0.925**	0.926	0.971
SENET	0.829	0.875	0.911	0.926	0.972
SHA	0.814	**0.884**	0.913	0.927	0.972
TRIPLEATT	0.814	0.857	0.910	0.927	0.971
Ganho (Unimodal vs Multimodal)					
Texto	7.20%	12.50%	7.40%	22.10%	2.60%
Visão	14.30%	10.70%	2.20%	1.90%	0.70%

(a) *Accuracy*.

ROC-AUC					
Modelos	**Top Speed**	**ChestXRay**	**Disaster**	**Fashion-iq**	**ROCO**
BERT	0.836	0.924	0.958	0.879	0.929
ViT	0.794	0.902	0.985	0.980	0.970
Operações aritméticas					
CONCAT	0.863	0.884	0.974	0.983	**0.981**
MAX	0.861	0.909	**0.984**	0.982	0.979
MIN	0.858	0.909	0.981	0.982	0.980
-SUB	**0.879**	0.909	0.983	0.983	**0.981**
SUB	0.861	**0.922**	0.977	**0.984**	0.978
SOMA	0.848	0.920	0.970	0.983	0.979
Mecanismos de Atenção sem *Skip Connection*					
CBAM	0.828	0.929	0.982	0.974	0.948
COOATT	0.862	0.968	**0.990**	0.976	0.965
ECA	0.876	0.970	0.976	**0.980**	0.970
PNA	0.841	0.945	0.974	0.945	0.968
PSA	0.862	0.956	0.979	0.977	0.945
RAN	0.831	0.882	0.971	0.975	0.958
S2ATT	0.794	0.674	0.592	0.578	0.864
SENET	**0.922**	0.965	0.984	**0.980**	0.967
SHA	0.867	0.961	0.980	0.962	**0.978**
TRIPLEATT	0.880	**0.977**	0.981	0.976	0.949
Mecanismos de Atenção com *Skip Connection*					
CBAM	0.862	0.912	0.985	**0.986**	**0.981**
COOATT	**0.864**	0.899	**0.986**	0.985	0.980
ECA	0.861	0.893	0.980	**0.986**	0.980
PNA	0.862	0.911	0.985	0.944	**0.981**
PSA	0.863	0.902	**0.986**	**0.986**	0.980
RAN	0.862	**0.933**	0.985	**0.986**	0.979
S2ATT	0.849	0.894	**0.984**	0.985	0.979
SENET	0.863	0.918	0.984	**0.986**	0.980
SHA	**0.864**	**0.919**	0.984	**0.986**	0.980
TRIPLEATT	0.862	0.905	0.985	**0.986**	0.980
Ganho (Unimodal vs Multimodal)					
Texto	8.60%	4.10%	3.20%	10.70%	5.20%
Visão	12.80%	6.30%	0.50%	0.60%	1.10%

(b) *ROC-AUC*.

Épocas

Os resultados experimentais ilustrados na Tabela 3.10 indicam as variações usadas em cada experimento e o número de épocas realizadas. Para evitar o *overfitting* da rede neural a técnica de *Early Stopping* (Prechelt, 1998) foi usada para definir o número máximo de períodos de treinamento. Os resultados indicam que o uso de diferentes modalidades impacta no número de épocas executadas por cada modelo, os modelos (Op) e (Att) possuem um valor relativamente maior que o modelo (MASK). Intuitivamente, o uso do *Skip Connection* pode ajudar a regularizar o modelo com um número significativamente menor de épocas, porém para conjuntos de dados maiores (ex: *Fashion-iq* e *ROCO*) esse efeito torna-se nulo devido à quantidade expressiva de exemplos usados no treinamento.

Representação Visual e Desempenho

A Figura 3.13 mostra os experimentos utilizados através da representação visual, esta análise permite mapear os modelos em relação ao desempenho de sua *Accuracy, Precision, Recall, F1 e ROC-AUC*. Para experimentos com

Tabela 3.10: Efeitos de cada componente em relação às épocas de cada conjunto de dados.

	Diferente Variações				
Texto	✓		✓	✓	✓
Imagem		✓	✓	✓	✓
Operações Aritméticas			✓	✓	✓
Sem *Skip Connection*				✓	
Com *Skip Connection*					✓
Épocas			**I***	**II***	**III***
Top Speed	13	6	22	38	8
ChestXRay	11	7	15	85	5
Disaster	6	7	36	29	6
Fashion-iq	3	2	3	42	1
ROCO	2	1	1	1	1

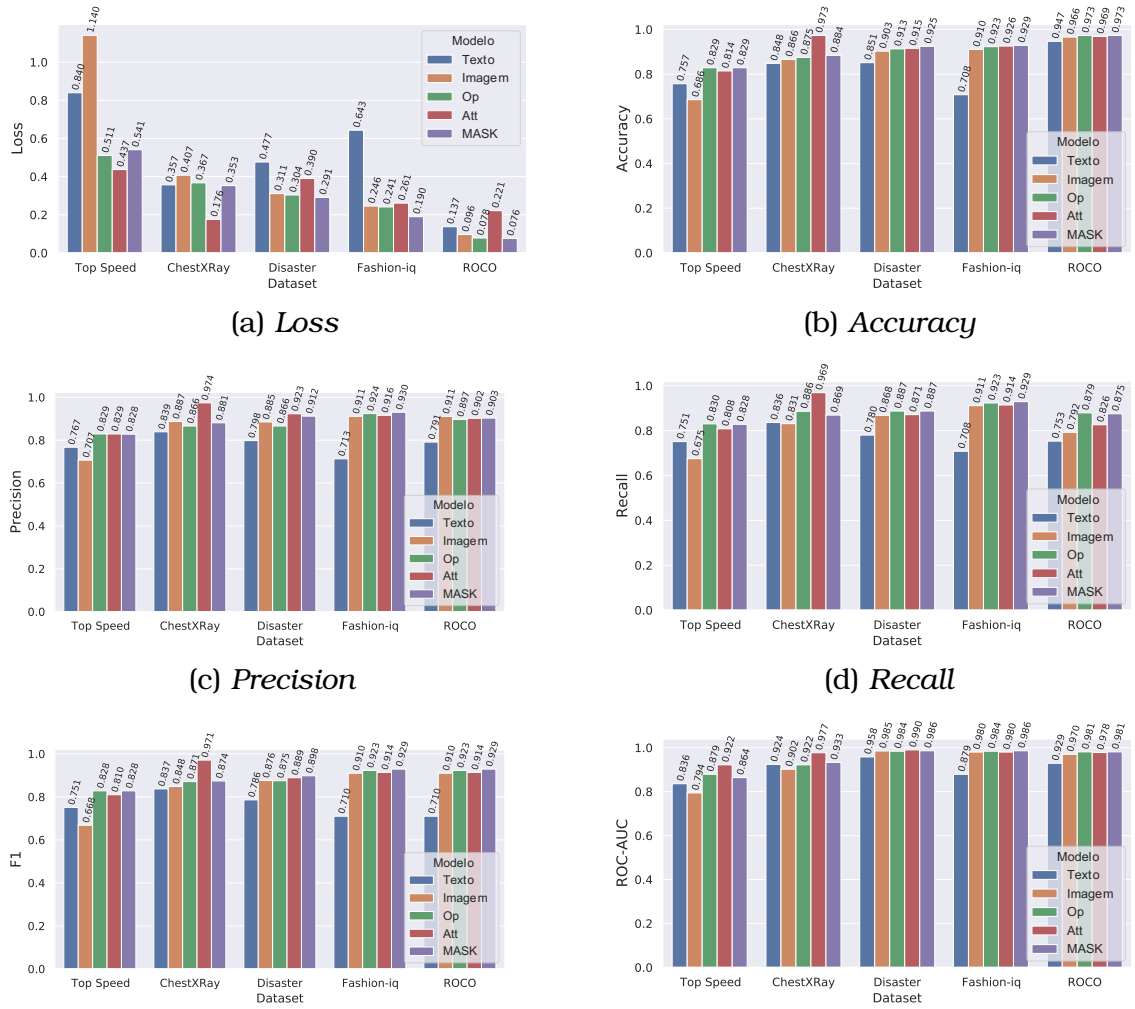
* Modelos: [I - Op] [II - Att] [III - MASK]

operações aritméticas e mecanismos de atenção os gráficos representam em cada conjunto de dados o melhor resultado obtido por cada modelo (Op, Att e MASK) em relação à métrica de *Accuracy*. É possível observar que os três modelos propostos neste trabalho possuem os maiores valores para todas as métricas apresentadas. A perda também é menor em comparação com modelos unimodais para a maioria dos conjuntos de dados.

A exploração das operações aritméticas indica um aumento significativo nas métricas dos classificadores aprendidos por fusão. Ao contrário de outros trabalhos (Meel and Vishwakarma, 2021; Kiela et al., 2020) propor outras operações além da concatenação para mesclar dados pode ter resultados satisfatórios em diversos domínios. A operação de "*Subtração*" também foi eficaz em comparação a operação de "*Concatenação*". Ao subtrair os resultados de diferentes modelos é possível identificar padrões comuns e aumentar a confiabilidade das previsões. Os resultados indicam que a subtração das *embeddings* pode melhorar os resultados quando comparado com a concatenação. Por outro lado é perceptível que modelos multimodais dependem diretamente do poder computacional para processar múltiplas modalidades em tempo de execução, experimentos mostram que o gasto computacional para executar lotes multimodais é 16x maior para entradas textuais e 4x maior para entradas visuais.

Convergência ROC-AUC

Para avaliar a evolução da função de perda (*ROC-AUC*) ao longo do tempo durante o treinamento e validação dos experimentos, um gráfico de perda foi usado como representação visual para medir o quão bem o modelo está aprendendo a tarefa específica para o qual foi treinado. A Figura 3.14 mostra a curva de perda do conjunto de treinamento e validação. A configuração experimental utilizou parada antecipada com *Early Stop* de 14 iterações. A abordagem MASK (em vermelho) converge claramente mais rápido do que (Op) e (Att). O cálculo da média do número de iterações para ambos modelos Op,



(a) *Loss* (b) *Accuracy*

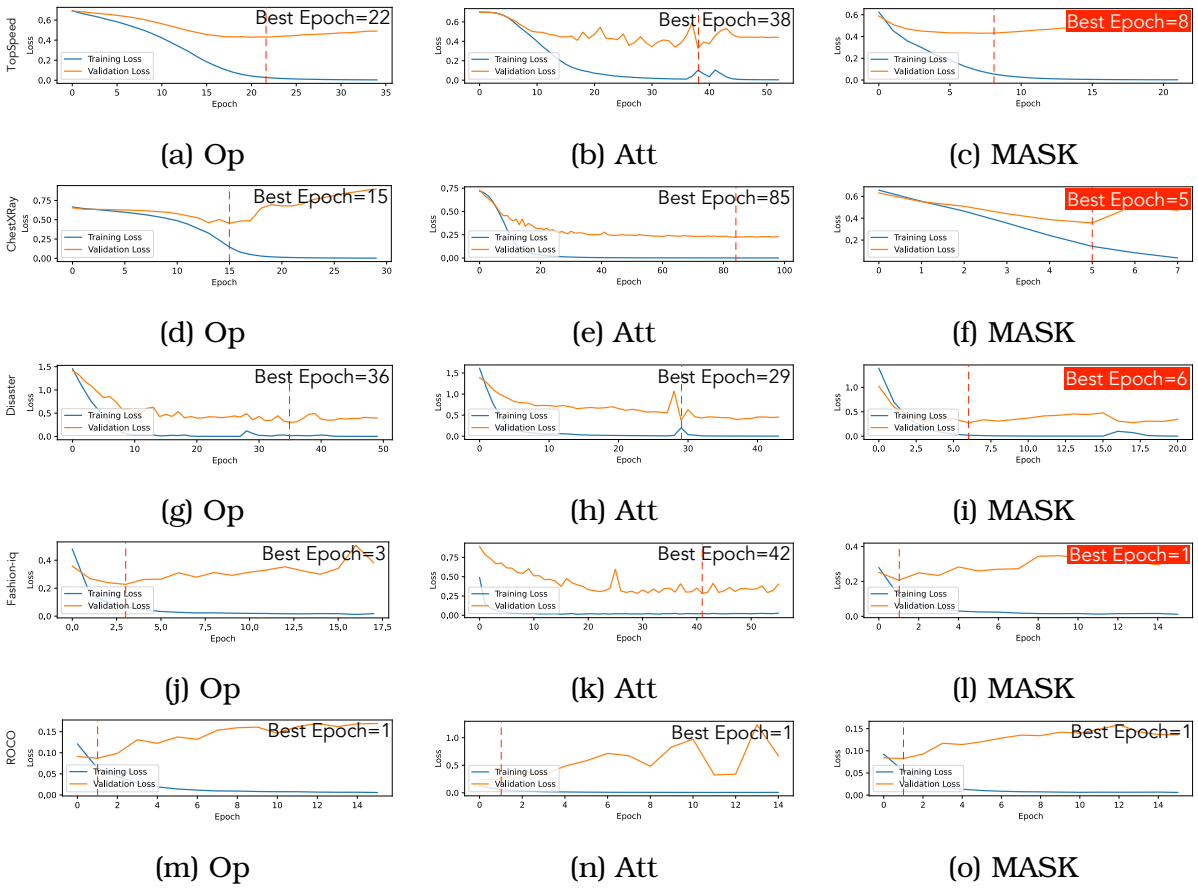
(c) *Precision* (d) *Recall*

(e) *F1* (f) *ROC-AUC*

Figura 3.13: Scores: *Loss, Accuracy, Precision, Recall, F1, ROC-AUC*.

Att e MASK é respectivamente, 15.4, 39.0 e 4.2. Assim MASK é 3,6 vezes mais rápido que (Op) e 9,2 vezes mais rápido que (Att). Para conjuntos de dados maiores (ex: *Fashion-iq* e *ROCO*), a convergência é mais suave e com pouca margem de diferença em comparação com outros modelos.

Este estudo também ilustra um gráfico violino apresentado na Figura 3.15 para visualizar a distribuição dos dados numéricos. Esta representação resume os dados em uma função de densidade, caracterizando melhor os modelos em relação a sua distribuição (Hintze and Nelson, 1998). Para este experimento a métrica *ROC-AUC* descrita na Tabela 3.9b é usada para mapear os resultados obtidos pelos modelos unimodais e pelos três modelos apresentados neste trabalho. De forma geral a distribuição dos picos são semelhantes para o modelo (Att) com quartis muito próximos e valores que ocorrem com mais frequência. Para os modelos textual, visual, (Op) e (MASK), há maior discrepância entre seus quartis, pois regiões mais estreitas do gráfico de densidade indicam valores que ocorrem com menor frequência. Essa representação permite dizer que modelos com discrepâncias maiores podem ter dificuldade em



(a) Op (b) Att (c) MASK

(d) Op (e) Att (f) MASK

(g) Op (h) Att (i) MASK

(j) Op (k) Att (l) MASK

(m) Op (n) Att (o) MASK

Figura 3.14: *ROC-AUC* de treinamento versus *ROC-AUC* de validação dos três modelos implementados.

encontrar um padrão nos dados, esse fato pode estar relacionado ao contexto do problema ou implementação do modelo.

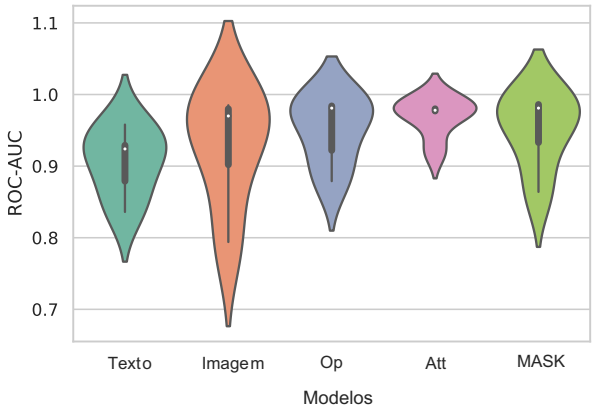


Figura 3.15: Visualização do Gráfico Violino para métrica *ROC-AUC*.

Análise de Variância (ANOVA)

Para comparar médias simultaneamente em várias populações há também uma descrição de análise de variância como ponto de referência (St et al., 1989). O teste visa comparar duas estimativas independentes de σ^2 de sub-

populações k para avaliar diferenças entre subpopulações. Esta abordagem ocorre através dos desvios entre as estimativas médias das subpopulações ($\mu_1, \mu_2, ..., \mu_k$) e as estimativas da média populacional (μ) (Soma dos Quadrados Entre Grupos). A análise dos desvios entre as observações e a média amostral de seu respectivo grupo (*Group Sum of Squares*) é realizada pela segunda estimativa, que visa avaliar a diferença dentro dos grupos. Para avaliar as diferenças significativas entre os modelos, os resultados obtidos foram submetidos a um teste ANOVA *one-way* neste experimento (Tabela 3.11). O $valor-p$ não é inferior a 0,05, portanto não é possível rejeitar a hipótese nula. Este resultado já era esperado considerando a avaliação de apenas cinco amostras, pois é difícil obter significância nas diferenças com esse tamanho de amostra. Assim as classificações médias correspondentes em cada modelo não são suficientes para tirar qualquer conclusão sobre os modelos, porém é possível afirmar que os modelos (Att e MASK) para *Accuracy* e *ROC-AUC*, apenas superaram marginalmente os modelos unimodais uma vez que consistentemente alcançaram classificações mais altas.

Tabela 3.11: Teste ANOVA.

ANOVA			
Métrica	**f-estatística**	**p-valor**	**Resultado**
Accuracy	1.3080960035861906	0.30086763493441776	Incapaz de rejeitar a hipótese nula
ROC-AUC	1.0489391969615502	0.4071180125119972	Incapaz de rejeitar a hipótese nula

Ranking de Performance

Os modelos foram classificados separadamente pela média ponderada sobre todos os conjuntos de dados. A Tabela 3.12 mostra a classificação média de *Accuracy* e *ROC-AUC*. Para as operações aritméticas (Op), a "Subtração Inversa": $\sum_{i=1}^{n}(rz_i^{img} - z_i^{texto})$ foi a que obteve os resultados mais expressivos para quase todos os conjuntos de dados para as duas métricas avaliadas. Uma hipótese para este resultado é que as imagens são mais discriminativas do que o conteúdo textual para configuração deste estudo, pois as representações compartilhadas possuem informações altamente correlacionadas, assim a operação de subtração se assemelha a uma operação de agrupamento (*MaxPool*, *AvgPool*), permitindo que sejam feitas suposições sobre os recursos contidos nas sub-regiões agrupadas.

Os modelos (Att) e (MASK) apresentam resultados diferentes sendo variada a classificação dos mecanismos de atenção em relação às métricas avaliadas. No modelo (Att) o mecanismo que obteve os melhores resultados foi *CooAtt* e *SE-Net*, há indícios que esse mecanismos mesmo sem o uso do *Skip Connection* são eficazes na convergência do modelo. Conforme mencionado em Seção 2 o módulo *CooAtt* com seu processo de fatoração e atenção de canal paralelo permite integrar efetivamente informações de coordenadas espaciais em ma-

Tabela 3.12: Ranking dos Modelos (Op), (Att) e (MASK).

	Accuracy		***ROC-AUC***	
Rank	**Operações aritméticas**			
1	**-SUB**	0.898	**-SUB**	0.947
2	**SUB**	0.891	**SUB**	0.944
3	**MAX**	0.879	**MAX**	0.943
4	**SOMA**	0.876	**MIN**	0.942
5	**CONCAT**	0.876	**SUM**	0.940
6	**MIN**	0.874	**CONCAT**	0.937
	Mecanismos de Atenção sem *Skip Connection*			
1	**COOATT**	0.899	**SENET**	0.964
2	**PSA**	0.894	**ECA**	0.954
3	**TRIPLEATT**	0.891	**TRIPLEATT**	0.953
4	**ECA**	0.888	**COOATT**	0.952
5	**SENET**	0.885	**SHA**	0.950
6	**CBAM**	0.863	**PSA**	0.944
7	**PNA**	0.859	**PNA**	0.941
8	**SHA**	0.774	**CBAM**	0.932
9	**RAN**	0.691	**RAN**	0.923
10	**S2ATT**	0.651	**S2ATT**	0.700
	Mecanismos de Atenção com *Skip Connection*			
1	**CBAM**	0.904	**RAN**	0.949
2	**SENET**	0.903	**SHA**	0.947
3	**SHA**	0.902	**SENET**	0.946
4	**PNA**	0.900	**CBAM**	0.945
5	**COOATT**	0.900	**TRIPLEATT**	0.944
6	**PSA**	0.898	**PSA**	0.943
7	**RAN**	0.897	**COOATT**	0.943
8	**TRIPLEATT**	0.896	**ECA**	0.940
9	**ECA**	0.894	**S2ATT**	0.938
10	**S2ATT**	0.885	**PNA**	0.937

pas de atenção, ao contrário de outros mecanismos que usam apenas um vetor de recurso por meio de agrupamento global 2D para geração de mapas de atenção. *CBAM* e *RAN* foram os mecanismos que obtiveram os melhores resultados no modelo (MASK). O *CBAM* é considerado uma evolução do mecanismo de atenção *Squeeze-and-Excitation* (SE), e o uso da atenção espacial e do canal é significativamente superior ao uso da atenção canalizada isoladamente. O módulo *RAN* tem a flexibilidade de ser incorporado em várias redes com uma arquitetura avançada de alimentação, com a capacidade de empilhar módulos de atenção que geram recursos de reconhecimento por meio de aprendizado residual. Assim resultados obtidos por modelo (MASK) favorecem a utilização do *Skip Connection* como moderador entre as camadas da rede e os mecanismos de atenção propostos neste trabalho.

Comparação SOTA

Para medir o potencial dos três modelos apresentados em nosso estudo, uma comparação foi realizada com quatro modelos SOTA presentes na literatura. O desempenho foi analisado sobre as duas métricas adotadas neste trabalho (*Accuracy* e *ROC-AUC*) com a seleção das melhores combinações deste experimento para cada um dos três modelos. Os modelos selecionados para

comparação são os seguintes:

Kiela et al. (2020) - Descreve um método de fusão simples no qual a média das pontuações de saída unimodal de *ResNet152* (He et al., 2016) e *BERT* (fusão tardia) são concatenadas. Para uma comparação justa, neste experimento foi usado *SE-ResNet152* (Hu et al., 2018) como modelo de visão, pois é uma melhoria do modelo e obtém melhores pontuações do que o *ResNet-152* padrão. As saídas do modelo foram redimensionadas para: *BERT* ($768 \rightarrow 300$) e *SE-ResNet152* ($1000 \rightarrow 300$).

Meel and Vishwakarma (2021) - É um método que compreende a utilização de treinamento em redes unimodais e fusão tardia nas últimas camadas do modelo. Os autores propõem o uso de pesos para cada entrada unimodal como forma de ajustar a fusão e suas probabilidades. Para este trabalho, foi adotados os seguintes pesos $W_1 = 0,6$ e $W_2 = 0,4$ para saída de texto e visão, respectivamente. Os autores usaram o modelo *BERT* e *ALBERT* (Lan et al., 2019) para entrada textual e *Inception-ResNetV2* (Szegedy et al., 2017) para visão, como atualização neste experimentou optou-se por usar *InceptionV4* (Szegedy et al., 2017) como modelo para entrada visual.

Li et al. (2019) - *VisualBERT* consiste em uma pilha de camadas *Transformer* que alinham implicitamente elementos de um texto de entrada e regiões em uma imagem de entrada associada com mecanismo de auto-atenção. Este modelo foi pré-treinado no conjunto de dados COCO (Lin et al., 2014), contendo pares de imagens e suas legendas.

Tabela 3.13: Comparação com Modelos Estado-da-arte.

Accuracy					
Modelos	**Top Speed**	**ChestXRay**	**Disaster**	**Fashion**	**ROCO**
Operações Aritméticas (Op)	**0.829**	0.875	0.913	0.923	**0.973**
Mecanismos de Atenção (Att)	0.814	**0.973**	0.915	0.926	0.969
MASK	**0.829**	0.884	0.925	**0.929**	**0.973**
BERT + SE_ResNet152 (Kiela et al., 2020)	0.729	0.866	0.907	0.912	0.964
BERT + InceptionV4 (Meel and Vishwakarma, 2021)	0.786	0.902	0.744	0.754	0.934
ALBERT + InceptionV4 (Meel and Vishwakarma, 2021)	0.471	0.375	0.761	0.851	0.968
VisualBERT (Kiela et al., 2020)	0.771	0.795	**0.927**	0.881	0.955
AUC-ROC					
Modelos	**Top Speed**	**ChestXRay**	**Disaster**	**Fashion**	**ROCO**
Operações Aritméticas (Op)	0.879	0.922	0.984	0.984	**0.981**
Mecanismos de Atenção (Att)	**0.922**	**0.977**	**0.990**	0.980	0.978
MASK	0.864	0.933	0.986	**0.986**	**0.981**
BERT + SE_ResNet152 (Kiela et al., 2020)	0.793	0.952	0.982	0.981	0.966
BERT + InceptionV4 (Meel and Vishwakarma, 2021)	0.878	0.975	0.929	0.920	0.891
ALBERT + InceptionV4 (Meel and Vishwakarma, 2021)	0.855	0.973	0.952	0.952	0.913
VisualBERT (Kiela et al., 2020)	0.857	0.843	**0.990**	0.970	0.934

Com exceção do modelo *VisualBERT* a abordagem deste estudo foi mais eficiente para quase todos os conjuntos avaliados. Alguns conjuntos de dados não apresentaram resultados satisfatórios para o modelo multimodal (*ALBERT + InceptionV4*), o que pode estar relacionado ao modelo pré-treinado e ajus-

tes de parâmetros. A pesquisa seguiu fielmente os parâmetros e estratégias adotados por cada estudo citado na Tabela 3.13.

Predições

Para fins de análise comparativa as predições dos melhores modelos para métrica *Accuracy* relatados na Tabela 3.12 foram coletadas para cada conjunto de dados avaliado. Neste experimento a matriz de confusão foi plotada pelo mapa de calor disponível na biblioteca *Seaborn* (Bisong, 2019). A matriz de confusão é considerado um indicador apropriado para avaliar o desempenho e a efetividade de um classificador categórico ou binário. A Figura 3.16 ilustra os dois modelos unimodais (*BERT* e *ViT*) e os modelos multimodais: **Op** = $-Sub$, **Att** = $CooAtt$, **MASK** = $CBAM$. Nesta análise o modelo unimodal *ViT* obteve mais acertos que o *BERT*, o que indica que as imagens podem ser mais representativas. Dentre os três modelos multimodais desenvolvidos neste trabalho, o modelo (MASK) com o mecanismo de atenção CBAM foi capaz de prever melhor os exemplos em quase todos os conjuntos de dados, exceto para o conjunto de dados *ChestXRay*. É possível observar que para conjuntos de dados binários, todos os modelos são mais bem sucedidos em prever a classe de prioridade em relação à classe adversária. Esse resultado pode ser justificado pelo fato de que algumas classes possuem mais representações durante a fase de treinamento, o que ajuda a enriquecer o modelo para a etapa de predição.

Discussão

O aprendizado multimodal mostra um viés mais forte do que as abordagens unimodais para problemas com mais de uma modalidade. Usando uma das três abordagens descritas neste estudo, pretendemos demonstrar o potencial de construir arquiteturas multimodais, mantendo o desempenho comparável ao dos métodos de aprendizado multimodais de última geração. Futuramente o estudo será expandido para investigar se os dados apresentam interdependência e analisar as diferenças entre os modelos multimodais. Com o avanço dos modelos unimodais pretende-se também substituir alguns dos modelos utilizados neste trabalho, principalmente o modelo *BERT* por seus modelos derivados como *ALBERT* (Lan et al., 2019), *ROBERTA* (Liu et al., 2019), *XL-Net* (Yang et al., 2019) e *ELECTRA* (Clark et al., 2020), pois podem fornecer resultados melhores do que os obtidos neste experimento. Outro ponto relevante é que a integração das conexões residuais e mecanismos de atenção pode conduzir a resultados comparáveis aos métodos mais avançados conhecidos (SOTA), além de possuir convergência mais rápida. (Liu et al., 2020; Xu et al., 2021).

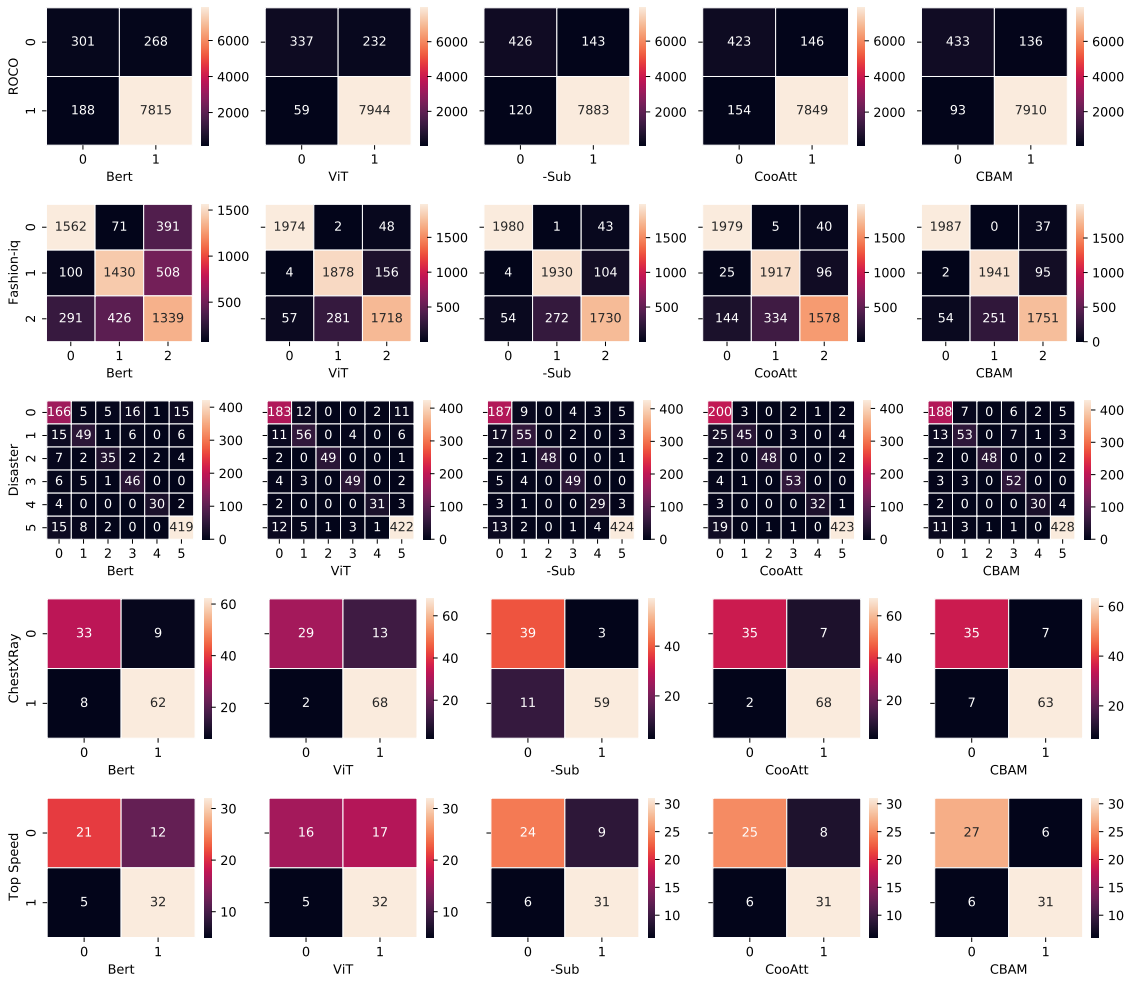


Figura 3.16: Matriz de Confusão - Resultados da classificação com os cinco conjuntos de dados avaliados neste experimento.

Este estudo foi submetido para publicação no periódico "*Knowledge-Based Systems - Elsevier*", sob o título "*MASK: a faster convergence approach using Multimodal Attention + SKip connections*".

CAPÍTULO

4

Destilação de Conhecimento Multimodal

Neste capítulo serão discutidos os métodos de extração e representação de informações multimodais, bem como as estratégias de fusão e transferência de conhecimento entre as modalidades.

4.1 Experimento 2 - Explorando a eficácia da destilação de conhecimento multimodal: descobertas e implicações

A Destilação de Conhecimento, também chamado "*Knowledge Distillation*" (KD), tem proporcionado aos modelos de aprendizado profundo a capacidade de comprimir modelos por meio de técnicas de destilação, tornando-os mais eficientes para implantação em hardware de baixo custo e com recursos limitados. O objetivo deste estudo é avaliar a eficácia da arquitetura KD em combinações de texto e imagem para conjuntos de dados multimodais. Esses experimentos possibilitam a avaliação da destilação de conhecimento multimodal (MKD) por meio de variações entre modelos de aprendizagem (textual e visual) com diferentes tamanhos.

O objetivo do KD é transferir o conhecimento aprendido por um modelo grande para um modelo pequeno, permitindo que esse modelo destilado tenha um desempenho melhor e seja facilmente implementado. Esse processo pode ser compreendido pela metodologia de ensino-aprendizagem, em que o modelo grande tem o papel de (professor) e o modelo menor e mais simples o papel do

(aluno) (Gou et al., 2021). A ideia por trás dessa metodologia é treinar o modelo do aluno para imitar as saídas do modelo do professor, dado um conjunto de entradas. O modelo professor atua como um guia e fornece supervisão ao modelo aluno durante o treinamento (Yim et al., 2017). A principal vantagem do KD é a capacidade de incorporar o conhecimento de um modelo preexistente, que é frequentemente um ponto de partida vantajoso para o treinamento de um novo modelo. Isso pode economizar tempo e recursos em comparação com o treinamento de um modelo do zero.

Este processo é normalmente aplicado a modelos unimodais em domínios que possuem apenas uma fonte de dados para treinamento (Cho and Hariharan, 2019). No entanto, algumas abordagens ganharam destaque com o uso da fusão de dados (Ramachandram and Taylor, 2017; Guo et al., 2019; Gao et al., 2020), que permite a combinação de informações de múltiplas fontes, geralmente de diferentes modalidades, para realizar tarefas como classificação, detecção de objetos, tradução, entre outros. Esta é uma área promissora e novos métodos estão sendo desenvolvidos para melhor combinar informações de diferentes modalidades. No entanto, alguns desafios comuns incluem a falta de dados, a heterogeneidade das fontes de informação e a complexidade no processo de integração de informações de diferentes modalidades (Ramachandram and Taylor, 2017). Novas abordagens também estão sendo investigadas para resolver problemas relacionados à escalabilidade e eficiência computacional no processo de integração de informações multimodais (Eitel et al., 2015; Liu et al., 2016; Williams et al., 2018; Oramas et al., 2018; Radu et al., 2018).

Este trabalho lança luz sobre novas direções no campo da Destilação de Conhecimento Multimodal (MKD) (Wang et al., 2020b; Xue et al., 2021; Dai et al., 2022), uma técnica que permite a integração de informações de diferentes modalidades. Embora outras pesquisas sigam uma direção semelhante (Dou et al., 2020; Garcia et al., 2021; Xue et al., 2022; Zhang et al., 2022) este estudo propõe uma avaliação experimental entre duas variações da arquitetura MKD para medir a eficiência de quatro redes neurais profundas, também chamada de "*Deep Neural Networks*" (DNN), incorporando mais de uma fonte de informações em comparação com o KD tradicional. Com uma estratégia simples de transferir conhecimento aprendido por modelos pré-treinados para modelos menores e mais eficientes, a destilação de conhecimento multimodal é uma alternativa viável para domínios que possuem maior diversidade de dados (Xue et al., 2021). O objetivo do estudo descrito é investigar o uso de MKD em cinco conjuntos de dados multimodais de diferentes domínios, com uma avaliação experimental que permita diagnosticar fragilidades e potencialidades em comparação com outras abordagens de integração de informação e desti-

lação de conhecimento. A ideia central desta pesquisa é explorar variações entre modelos de aprendizagem (textual e visual) com tamanhos diferentes de modelos de aprendizado, em seguida, descrever uma análise comparativa com o KD tradicional.

Proposta

Para extração de características neste experimento, quatro DNNs são usadas consistindo em dois modelos minúsculos (*tiny*) e dois modelos básicos (*base*). Além disso, as métricas de classificação como *Accuracy, Recall, Precision, F1, e ROC-AUC* são utilizadas para prever a classe nos cinco conjuntos de dados usados. O *pipeline* deste estudo é ilustrado na Figura 4.1 com três experimentos diferentes. Ele começa com a extração de dados multimodais que são alimentados na camada de entrada dos modelos. O pipeline então segue as seguintes etapas: (i) extração e seleção de atributos para dados textuais e imagens; (ii) treinamento de cada DNN; (iii) implementação do KD com três experimentos: KD tradicional, MKD com Professor (modelo *base*) e Aluno (modelo *tiny*), e MKD com Professor (modelo *base*) e Aluno (modelo *base*); (iv) avaliação de diferentes modelos e variações de KD e MKD; (v) finalmente é realizada uma avaliação experimental considerando os diferentes aspectos de comparação entre as arquiteturas KD, incluindo as métricas de classificação, convergência e padrões observados neste estudo usando diferentes conjuntos de dados.

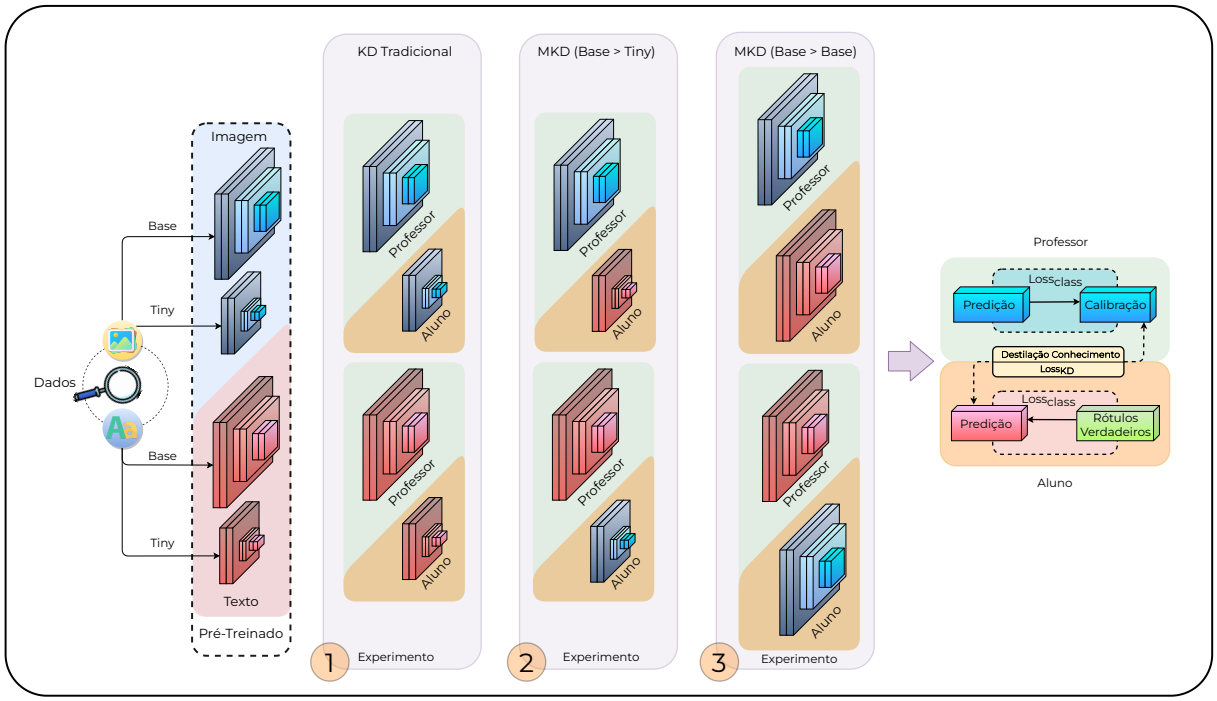


Figura 4.1: Arquitetura MKD - Pipeline.

A proposta da arquitetura MKD permite a avaliação de modelos e dados

multimodais sob as seguintes perspectivas:

Correlação - Destilar o conhecimento de uma modalidade para fornecer informações adicionais a outra modalidade pode conter uma relação intrínseca de dependência entre as duas fontes de dados, muitas vezes mais representativa em uma das modalidades investigadas em um domínio multimodal (Guo et al., 2019). Essa correlação em determinados cenários pode contribuir positivamente para a construção de um modelo baseado em KD, pois busca entender como uma modalidade se comporta em um cenário onde há variação de outras modalidades. A fim de avaliar essa questão, os Experimentos 2 e 3 foram realizados em dobras para simular o impacto na destilação dos modelos de imagem em relação ao texto e vice-versa.

Diversidade de Modelos - Foi adicionada uma abordagem não convencional na construção de KD, na qual grandes modelos foram incorporados na função do aluno, como mostrado na Experimento 3. Esta adição permite explorar se modelos de diferentes modalidades podem obter melhor desempenho tanto em configurações quanto em parâmetros, seja um modelo pequeno ou grande (Hu et al., 2021). É uma avaliação extensa que ajuda a mensurar se é necessário utilizar modelos comparáveis no processo de KD ou se é possível agregar informações ricas em modelos menores por meio da multimodalidade.

Variedade de dados - A abordagem adotada neste estudo considera apenas duas fontes de dados singulares no mesmo domínio (Experimentos 2 e 3), pois acredita-se que é possível cruzar informações relevantes para mapear as classes de interesse em um modelo destilado pela multimodalidade. Pesquisas recentes apontam para esse caminho com resultados satisfatórios (Xue et al., 2022; Zhang et al., 2022; Li et al., 2020), mas há espaço para explorar novas técnicas em KD, que visam criar abordagens com diferentes fontes de dados e diferentes combinações arquitetônicas.

Arquitetura da Destilação de Conhecimento Multimodal

MKD (base $\rightarrow$ tiny) - Foi empregada uma abordagem KD tradicional no Experimento 1 que teve como objetivo destilar conhecimento de modelos (*base*) para modelos (*tiny*). Para isso, foram utilizadas duas combinações multimodais, sendo a primeira para extrair conhecimento da imagem (papel professor) para o texto (papel aluno) e a segunda para destilar o texto (papel professor) para imagem (papel aluno).

MKD (base $\rightarrow$ base) - O objetivo do segundo experimento foi examinar as diferenças entre os modelos e o número de parâmetros utilizados em cada variação. Para isso, há a destilação de conhecimento de modelos (*base*) para modelos (*base*) entre diferentes modalidades, seguindo os mesmos passos da abordagem MKD (*base* $\rightarrow$ *tiny*).

A avaliação experimental utiliza uma arquitetura que se baseia em pesquisas atuais sobre destilação de conhecimento (Romero et al., 2014; Hinton et al., 2015; Li, 2018; Tang and Wang, 2018), usando estruturas e especificações técnicas apropriadas. O conhecimento baseado em resposta ("*Response-Based Knowledge*") descrito no trabalho de Li (2018) foi utilizado para avaliar a precisão dessa abordagem, representado pela Equação 4.1.

$$KD_{loss} = Diff_{loss}(Softmax(\frac{aluno_f}{\mathscr{T}}), (Softmax(\frac{professor_f}{\mathscr{T}}) * (\alpha * \mathscr{T}^2)), \tag{4.1}$$

onde $Diff_{loss}$ calcula a perda devido à divergência *Kullback-Leibler* (Joyce, 2011), $aluno_f$ e $professor_f$ representam as características obtidas nas camadas de saída de cada modelo. Há também a adição de hiperparâmetros, com ($\mathscr{T}$) representando uma temperatura para escalar a incerteza das previsões do professor e (α) para ajustar o peso da perda de destilação do aluno. Um (α) igual a 0 significa que é considerada apenas a perda de destilação e vice-versa. Para temperatura ($\mathscr{T}$) valores maiores tendem a gerar uma distribuição mais suave entre as classes de saída.

Para comparar a diferença e o ajuste de perda entre os modelos professor e aluno a Equação 4.1 é usada neste estudo e depois é realizada uma calibração do modelo aluno usando a Equação 4.2, otimizando a função de perda com base nos dados de treinamento. Essa abordagem provou ser eficaz para melhorar as distribuições de probabilidade nas classes de saída.

$$\text{Calibração} = KD_{loss} + CrossEntropy(aluno_f, classe) * (1 - \alpha), \tag{4.2}$$

Pré-processamento e Configuração Experimental

Neste estudo, foram avaliados cinco conjuntos de dados já citados no Experimento (1.1) da Seção 3.2, contendo informações de texto e imagem de diferentes domínios, conforme ilustrado na Figura 4.2. Esses conjuntos de dados foram coletados recentemente a partir de pesquisas, representando uma amostra de dados multimodais em grande escala.

Os parâmetros adotadas nesta pesquisa, assim como a configuração experimental seguiu o mesmo critérios já citados no Experimento (1.1) da Seção 3.2, porém houve a adição de modelos com tamanhos variados, a Tabela 4.1 resume as informações sobre os modelos estabelecidos nesta pesquisa.

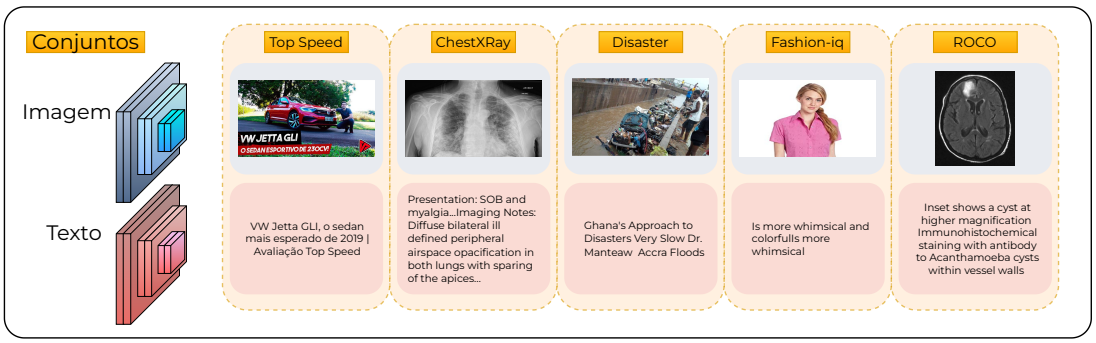


Figura 4.2: Conjuntos de dados Multimodal.

Tabela 4.1: Visão geral dos modelos selecionados para Linguagem e Visão.

Modelos	**Parâmetros**	**Camadas**	**Tamanho Camada Oculta**
BERT *tiny*	4.4M	2	128
BERT *base*	110M	12	768
ViT *tiny*	5.7M	12	192
ViT *base*	86.7M	12	768

Resultados Experimentais

Cada um dos quatro modelos selecionados foi treinado individualmente nos conjuntos de dados especificados, e em seguida, várias combinações foram realizadas durante a fase de destilação de conhecimento. O objetivo foi cobrir todas as combinações possíveis e analisar as dificuldades encontradas na destilação utilizando a técnica "*Response-Based Knowledge*". Para avaliar esta abordagem a pesquisa foi divida em três partes que envolveu: (i) avaliação dos resultados dos quatro modelos individuais e destilação de conhecimento tradicional usando métricas padrão como *Accuracy, Precision, Recall, F1 e ROC-AUC*; (ii) avaliar as combinações multimodais da arquitetura MKD; e (iii) realizar uma análise comparativa entre a destilação de conhecimento tradicional e a arquitetura proposta MKD.

Calibração - A Tabela 4.2 descreve a calibração dos hiperparâmetros usados para cada experimento com destilação de conhecimento. Optou-se por usar a calibração porque ela ajuda a melhorar a confiabilidade das previsões de forma mais justa para cada combinação realizada (Nixon et al., 2019). Usando empiricamente a técnica "*Grid Search*" (Scikit-learn, 2007), testes foram conduzidos com uma variação (0,0 a 0,9) para o hiperparâmetro α = *alpha* e (1,0 a 9,0) para $\mathscr{T}$ = temperatura. Para cada conjunto α e $\mathscr{T}$ houve uma execução sobre o modelo de destilação no conjunto de treinamento correspondente e avaliação no conjunto de validação. Os melhores conjuntos de parâmetros foram obtidos classificando as pontuações de validação para *ROC-AUC* em cada conjunto de dados. Assim é possível garantir que os valores definidos na tabela foram as melhores indicações para cada conjunto de dados.

Métricas KD - As métricas *Accuracy, Precision, Recall, F1 e ROC-AUC* obti-

Tabela 4.2: Calibração dos Modelos KD e MKD.

Hiperparâmetros	**TopSpeed**		**ChestXRay**		**Disaster**		**Fashion-iq**		**ROCO**	
	α	$\mathscr{T}$	α	$\mathscr{T}$	α	$\mathscr{T}$	α	$\mathscr{T}$	α	$\mathscr{T}$
KD - Linguagem	0.0	9	0.3	7	0.5	1	0.5	1	0.1	3
KD - Visão	0.9	5	0.6	3	0.4	3	0.7	3	0.8	1
MKD - Linguagem$_{(tiny)}$	0.0	9	0.2	5	0.8	1	0.0	3	0.3	5
MKD - Visão$_{(tiny)}$	0.9	5	0.5	3	0.2	1	0.0	7	0.1	3
MKD - Linguagem$_{(base)}$	0.6	9	0.6	1	0.2	1	0.5	1	0.7	1
MKD - Visão$_{(base)}$	0.1	7	0.2	5	0.0	5	0.0	7	0.4	1

das para avaliação nos conjuntos de teste são fornecidas na Tabela 4.3. Modelos de texto e imagem foram treinados separadamente, e posteriormente, aplicou-se a técnica KD convencional para aprimorar os modelos menores de cada modalidade. Informações sobre a calibração dos modelos e as configurações de parâmetros podem ser encontradas na Tabela 3.8 e Tabela 4.2. A principal meta é avaliar o desempenho dos modelos nos conjuntos de dados escolhidos e sua habilidade de distinguir entre as diferentes classes do problema. De acordo com os resultados obtidos pelos modelos *BERTimbau* e *ViT* as imagens são mais discriminantes do que o conteúdo textual para a maioria dos conjuntos de dados. Adicionalmente foi observado que a aplicação de KD nos modelos *BERT* e *ViT* é superior aos modelos não destilados em quase todos os conjuntos de dados selecionados, exceto no conjunto de dados *ChestXRay*.

Tabela 4.3: Métricas KD

Modalidade	**Modelo** $_{(teacher \rightarrow student)}$	***Precision***	***Recall***	***F1***	***Accuracy***	***ROC-AUC***
		TopSpeed				
Linguagem	BERT_{tiny}	0.729	0.733	0.732	0.729	0.801
KD - Linguagem	$\text{BERT}_{base} \rightarrow \text{BERT}_{tiny}$	0.730	0.730	0.729	0.729	**0.809**
Visão	ViT_{tiny}	0.743	0.743	0.740	0.741	0.842
KD - Visão	$\text{ViT}_{base} \rightarrow \text{ViT}_{tiny}$	0.749	0.737	0.738	0.743	**0.855**
		ChestXRay				
Linguagem	BERT_{tiny}	0.893	0.886	0.886	0.886	**0.964**
KD - Linguagem	$\text{BERT}_{base} \rightarrow \text{BERT}_{tiny}$	0.889	0.881	0.885	0.893	0.933
Visão	ViT_{tiny}	0.857	0.853	0.838	0.844	**0.890**
KD - Visão	$\text{ViT}_{base} \rightarrow \text{ViT}_{tiny}$	0.852	0.788	0.804	0.830	**0.890**
		Disaster				
Linguagem	BERT_{tiny}	0.825	0.769	0.744	0.753	0.941
KD - Linguagem	$\text{BERT}_{base} \rightarrow \text{BERT}_{tiny}$	0.851	0.842	0.844	0.902	**0.975**
Visão	ViT_{tiny}	0.839	0.808	0.751	0.773	0.954
KD - Visão	$\text{ViT}_{base} \rightarrow \text{ViT}_{tiny}$	0.800	0.808	0.800	0.848	**0.976**
		Fashion-iq				
Linguagem	BERT_{tiny}	0.694	0.693	0.694	0.690	0.859
KD - Linguagem	$\text{BERT}_{base} \rightarrow \text{BERT}_{tiny}$	0.699	0.702	0.699	0.701	**0.873**
Visão	ViT_{tiny}	0.890	0.890	0.890	0.889	0.974
KD - Visão	$\text{ViT}_{base} \rightarrow \text{ViT}_{tiny}$	0.908	0.906	0.905	0.905	**0.979**
		ROCO				
Linguagem	BERT_{tiny}	0.941	0.768	0.703	0.730	0.922
KD - Linguagem	$\text{BERT}_{base} \rightarrow \text{BERT}_{tiny}$	0.803	0.752	0.775	0.949	**0.932**
Visão	ViT_{tiny}	0.966	0.878	0.837	0.856	0.969
KD - Visão	$\text{ViT}_{base} \rightarrow \text{ViT}_{tiny}$	0.904	0.817	0.855	0.968	**0.973**

A abordagem proposta neste estudo levou a uma melhoria de 0,8% para 3,4% no *ROC-AUC*, que é a métrica mais adequada para avaliar o desempenho dos modelos para problemas de classificação. Trabalhos anteriores de Gou et al. (2021) e Tang et al. (2020) também relataram resultados comparáveis,

indicando que a destilação de conhecimento pode melhorar a representação dos modelos destilados ao incorporar a relação professor-aluno. A abordagem envolve a geração de uma representação coletiva do conhecimento por meio da função de perda por destilação, que calcula a diferença entre os *logits* dos modelos professor e aluno. De acordo com Hinton et al. (2015) a avaliação desse método envolve duas combinações: (i) transferência de conhecimento para o modelo destilado treinando-o em um conjunto de transferência e usando uma distribuição de destino flexível para cada instância no conjunto de dados; (ii) quando rótulos precisos são conhecidos para todo ou parte do conjunto de transferência, treinar o modelo destilado para produzir os rótulos corretos pode resultar em melhoria significativa.

Métricas MKD - A Tabela 4.4 mostra as variações implementadas na arquitetura MKD. A abordagem inicial consistiu na destilação convencional de conhecimento entre modelos de professor e aluno em diversas modalidades. Na primeira variação o modelo professor teve mais parâmetros do que o modelo aluno, semelhante ao KD. Na segunda variação a arquitetura MKD foi utilizada em modelos com um número semelhante de parâmetros para avaliar como o modo professor afeta os resultados da destilação em relação ao tamanho do modelo. Esta técnica permitiu analisar os resultados e determinar os efeitos do modelo selecionado nos resultados da destilação.

Tabela 4.4: Métricas MKD

TopSpeed						
Modalidade	**Modelo** $_{(teacher \rightarrow student)}$	***Precision***	***Recall***	***F1***	***Accuracy***	***ROC-AUC***
MKD - Linguagem$_{(tiny)}$	ViT$_{base}$ $\rightarrow$ BERT$_{tiny}$	0.743	0.744	0.743	0.743	0.722
MKD - Linguagem$_{(base)}$	ViT$_{base}$ $\rightarrow$ BERT$_{base}$	0.700	0.697	0.697	0.700	**0.805**
MKD - Visão$_{(tiny)}$	BERT$_{base}$ $\rightarrow$ ViT$_{tiny}$	0.756	0.735	0.735	0.743	0.860
MKD - Visão$_{(base)}$	BERT$_{base}$ $\rightarrow$ ViT$_{base}$	0.829	0.808	0.810	0.814	**0.876**
ChestXRay						
MKD - Linguagem$_{(tiny)}$	ViT$_{base}$ $\rightarrow$ BERT$_{tiny}$	0.886	0.886	0.886	0.893	0.938
MKD - Linguagem$_{(base)}$	ViT$_{base}$ $\rightarrow$ BERT$_{base}$	0.934	0.898	0.911	0.920	**0.953**
MKD - Visão$_{(tiny)}$	BERT$_{base}$ $\rightarrow$ ViT$_{tiny}$	0.727	0.705	0.641	0.643	**0.882**
MKD - Visão$_{(base)}$	BERT$_{base}$ $\rightarrow$ ViT$_{base}$	0.796	0.612	0.592	0.705	0.867
Disaster						
MKD - Linguagem$_{(tiny)}$	ViT$_{base}$ $\rightarrow$ BERT$_{tiny}$	0.853	0.836	0.844	0.898	**0.977**
MKD - Linguagem$_{(base)}$	ViT$_{base}$ $\rightarrow$ BERT$_{base}$	0.753	0.765	0.756	0.819	0.960
MKD - Visão$_{(tiny)}$	BERT$_{base}$ $\rightarrow$ ViT$_{tiny}$	0.767	0.788	0.773	0.834	0.969
MKD - Visão$_{(base)}$	BERT$_{base}$ $\rightarrow$ ViT$_{base}$	0.880	0.865	0.871	0.909	**0.987**
Fashion-iq						
MKD - Linguagem$_{(tiny)}$	ViT$_{base}$ $\rightarrow$ BERT$_{tiny}$	0.693	0.696	0.694	0.695	0.867
MKD - Linguagem$_{(base)}$	ViT$_{base}$ $\rightarrow$ BERT$_{base}$	0.708	0.711	0.700	0.710	**0.877**
MKD - Visão$_{(tiny)}$	BERT$_{base}$ $\rightarrow$ ViT$_{tiny}$	0.882	0.883	0.882	0.882	0.972
MKD - Visão$_{(base)}$	BERT$_{base}$ $\rightarrow$ ViT$_{base}$	0.907	0.907	0.907	0.907	**0.979**
ROCO						
MKD - Linguagem$_{(tiny)}$	ViT$_{base}$ $\rightarrow$ BERT$_{tiny}$	0.816	0.715	0.754	0.949	**0.931**
MKD - Linguagem$_{(base)}$	ViT$_{base}$ $\rightarrow$ BERT$_{tiny}$	0.805	0.701	0.740	0.946	0.928
MKD - Visão$_{(tiny)}$	BERT$_{base}$ $\rightarrow$ ViT$_{tiny}$	0.870	0.826	0.846	0.964	0.960
MKD - Visão$_{(base)}$	BERT$_{base}$ $\rightarrow$ ViT$_{tiny}$	0.902	0.824	0.858	0.968	**0.973**

A hipótese é que a vantagem do KD decorre da combinação dos *embeddings* com a representação compartilhada pelos modelos unimodais (Ramachandram and Taylor, 2017), por meio da destilação de conhecimento baseado

na multimodalidade. A exploração das duas variações presentes na Tabela 4.4 indicam que em alguns casos através da arquitetura MKD é possível melhorar os resultados dos modelos rotulados como aluno, em geral a destilação de modelos com quantidade comparável de parâmetros garante melhores resultados. No entanto, não há garantia de que a arquitetura MKD possa abranger todos os domínios possíveis, pois podem existir modalidades com maior influência em um domínio e dificultar a destilação de modalidades que não sejam tão representativas. Outra questão está relacionada com a eficácia dos modelos utilizados na aprendizagem e que podem impactar significativamente nos resultados (Blasch et al., 2018), uma avaliação experimental antes da criação de uma arquitetura MKD pode ser necessária para avaliar a implicação dos modelos SOTA na destilação de conhecimento.

Comparação de desempenho - Para medir o potencial da arquitetura MKD em relação às duas variações apresentadas em nosso estudo, resultados obtidos neste trabalho foram comparados com a destilação de conhecimento tradicional (KD). A métrica escolhida para medir a taxa de acerto dos exemplos em relação às suas respectivas classes foi calculada para avaliar o desempenho. Como critério foi adotado os parâmetros descritos na Equação 4.1 e Equação 4.2, descritos no trabalho de Li (2018) para destilar os modelos em todas as execuções.

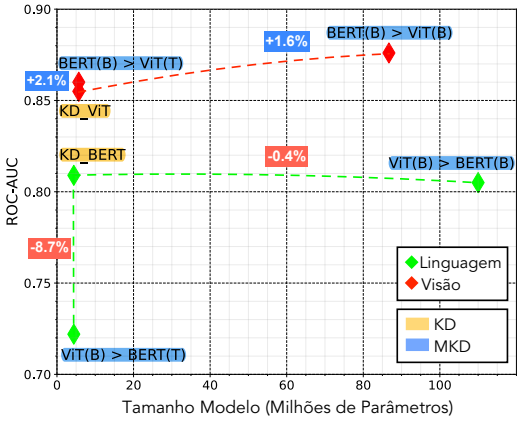
Os resultados do estudo apontam que a abordagem MKD proposta é uma alternativa promissora em relação ao KD tradicional. A métrica *ROC-AUC* foi utilizada para avaliar os modelos e apresentou melhora significativa em alguns conjuntos, superando os resultados obtidos pelo KD tradicional. No entanto, houve exceções em que a abordagem proposta não foi capaz de obter pontuações mais altas, principalmente para conjuntos de dados menores ou com modalidades não representativas.

Tabela 4.5: Comparação de Desempenho - *ROC-AUC*

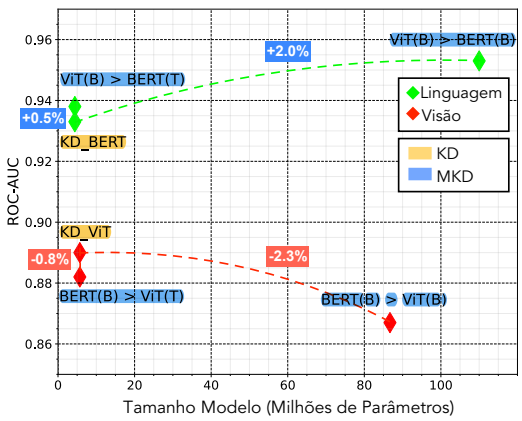
		Linguagem				
Modality	**Modelo** $_{(teacher \rightarrow student)}$	**TopSpeed**	**ChestXRay**	**Disaster**	**Fashion-iq**	**ROCO**
KD - Linguagem	$\text{BERT}_{base} \rightarrow \text{BERT}_{tiny}$	**0.809**	0.933	0.975	0.873	**0.932**
MKD - Linguagem$_{(tiny)}$	$\text{ViT}_{base} \rightarrow \text{BERT}_{tiny}$	0.722	0.938	**0.977**	0.867	0.931
MKD - Linguagem$_{(base)}$	$\text{ViT}_{base} \rightarrow \text{BERT}_{base}$	0.805	**0.953**	0.960	**0.877**	0.928
		Visão				
KD - Visão	$\text{ViT}_{base} \rightarrow \text{ViT}_{tiny}$	0.855	**0.890**	0.976	**0.979**	**0.973**
MKD - Visão$_{(tiny)}$	$\text{BERT}_{base} \rightarrow \text{ViT}_{tiny}$	0.860	0.882	0.969	0.972	0.960
MKD - Visão$_{(base)}$	$\text{BERT}_{base} \rightarrow \text{ViT}_{base}$	**0.876**	0.867	**0.987**	**0.979**	**0.973**

A significância paramétrica requer amostras normalmente distribuídas. Cinco conjuntos de dados, o que significa cinco amostras, são muito pequenos para testar a normalidade das distribuições. No entanto, buscando excelência e avaliação rigorosa, um teste ANOVA (*One-way*) foi realizado neste experimento e obteve uma estatística (f) de 0,431916730 e um ($p-valor$) de 0,821812084. Portanto, não é possível rejeitar a hipótese nula.

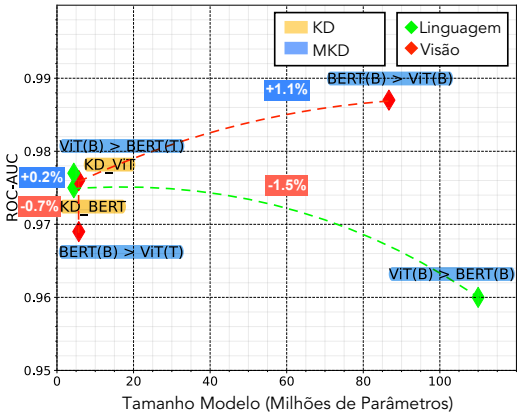
A Figura 4.3 apresenta um gráfico comparando as arquiteturas KD e MKD. Este gráfico permite uma análise dos resultados obtidos em relação ao número de parâmetros utilizados em cada modelo DNN testado neste experimento. Os resultados apontaram que a destilação dos modelos textuais apresentou melhor desempenho em relação aos modelos visuais, provavelmente devido às descrições e legendas que auxiliaram na precisão (*Accuracy*) dos modelos visuais destilados.



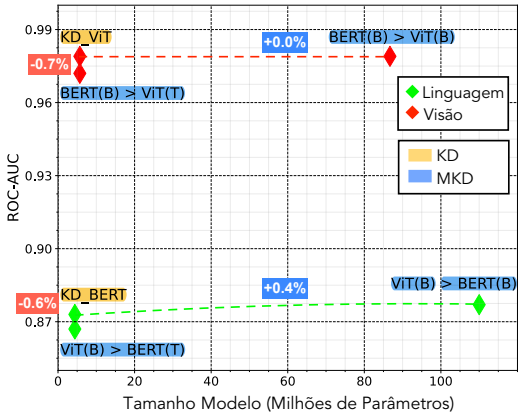
(a) TopSpeed



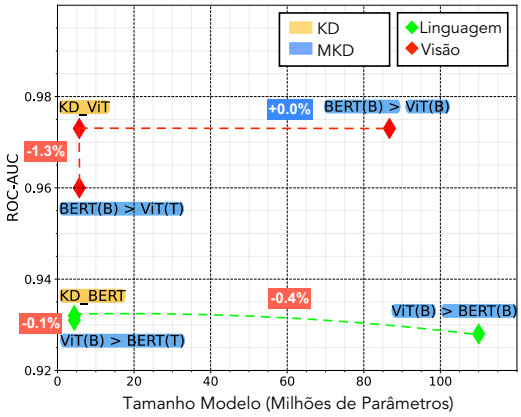
(b) ChestXRay



(c) Disaster



(d) Fashion-iq



(e) ROCO

Figura 4.3: Gráfico - Complexidade dos Modelos versus métrica *ROC-AUC*.

Discussão

A destilação entre modelos com tamanho comparável (*base*) obteve os melhores resultados em nossa pesquisa, no entanto, a destilação de modelos (*tiny*) superou os modelos (*base*) quando destilados na arquitetura KD tradicional e MKD em alguns conjuntos de dados. Gou et al. (2021) argumenta

na formulação de seus experimentos que naturalmente as configurações dos modelos professor e aluno são pré-fixadas com tamanhos e estruturas invariáveis, normalmente com o uso de um modelo maior para o papel do professor e um menor para o papel do aluno (Hinton et al., 2015). De fato, esta configuração suporta os resultados obtidos neste experimento, proporcionando um desempenho significativo aos modelos (*tiny*) no processo de destilação. No entanto, o autor justifica que não existe um padrão conceitual para projetar especificamente arquiteturas de professores e alunos, e muitas dessas configurações estão quase ausentes na literatura.

A investigação de modelos destilados baseados na arquitetura KD multimodal mostrou o potencial para construir novas configurações (Xue et al., 2021; You et al., 2021), mantendo desempenho comparável ao método KD tradicional de última geração. No Experimento (1): os resultados obtidos pela métrica *ROC-AUC* indicam que KD é satisfatório para a maioria dos conjuntos de dados utilizados nesta pesquisa. O conjunto de dados *ChestXRay* (Cohen et al., 2020) provou ser insatisfatório para execução com KD. Para este caso, o conjunto foi limitado apenas às imagens brutas e não incorporou as segmentações e pontuações específicas do domínio disponíveis na documentação oficial do conjunto de dados, considerado ser apropriado para extração de características.

Os Experimentos (2) e (3) permitem analisar a arquitetura MKD sob a ótica dos modelos e da fusão das fontes visuais e textuais presentes nos conjuntos de dados multimodais. A diferença de desempenho foi examinada ao avaliar o papel do aluno com modelos de tamanho, estrutura e número de parâmetros variáveis. Neste estudo optou-se por usar modelos (*base* e *tiny*) e os resultados apontam uma pequena vantagem ao usar modelos (*base*) para papel do aluno na arquitetura KD multimodal, logo a extração de recursos de cada modalidade é variável e independente entre as fontes de dados obtidas nos conjuntos de dados (Wang et al., 2020b). Portanto, usar um modelo maior e com mais parâmetros pode ter um efeito satisfatório quando submetido à destilação, mas essa abordagem não satisfaz os princípios básicos da arquitetura KD que garantem que o modelo destilado seja menor e com o mesmo desempenho em relação aos modelos maiores (Hinton et al., 2015). Para este cenário a justificativa é que esta configuração permite auxiliar a arquitetura KD multimodal no processo de compartilhamento de informações complementares entre as modalidades para que o modelo destilado alcance previsões mais refinadas.

Este estudo foi submetido para publicação no periódico "*Expert Systems with Applications*", sob o título "*Exploring the effectiveness of multimodal knowledge distillation: findings and implications*".

CAPÍTULO

5

Caso de Estudo - Fusão entre Modelos de Aprendizado e dados Unimodais

A fusão entre modelos de aprendizado e dados unimodais tem o potencial de enriquecer a representação dos dados, capturar relações complexas e fornecer abordagens mais robustas e abrangentes sobre o problema em questão. Neste capítulo há um caso de estudo com exploração de técnicas de fusão entre modelos de aprendizado e dados unimodais.

5.1 Experimento 3 - Uma Rede Multivisão para predição de fenotipagem de alto rendimento para matéria verde e seca

A espécie *Panicum maximum Jacq.* (sin. *Megathyrsus maximus*) é uma das forrageiras mais importantes para a produção animal nas regiões tropicais e subtropicais do mundo (Jank et al., 2014). A adoção de cultivares melhoradas dessa espécie foi uma das principais estratégias para aumentar a produção pecuária (Gomide and Gomide, 1999; Jank et al., 2008). Para continuar lançando novos cultivares de forma mais eficiente os programas de melhoramento precisam aplicar estratégias inovadoras para aumentar sua eficiência em tempo, custo e precisão. Com o uso da agricultura de precisão aliada a um conjunto de tecnologias como sensores, sistemas de informação e ges-

tão informatizada para otimizar a produção é possível aumentar a eficiência da fenotipagem nos programas de melhoramento (Gebremedhin et al., 2019; Castro et al., 2020; de Oliveira et al., 2021), que contribuirá para o lançamento de cultivares forrageiras melhoradas que melhor atendam às necessidades da pecuária.

Dada a importância do processo de fenotipagem e a busca por plantas melhoradas, estudos anteriores propuseram maneiras novas e mais eficientes de avaliar características por meio do uso de algoritmos de aprendizado de máquina alimentados por imagens aéreas de *drones* (Singh et al., 2018; Stewart et al., 2019; Ampatzidis and Partel, 2019; Xiong et al., 2021). Nas forragens a visualização do rendimento total de matéria verde ("*Total Green Matter Yield*" - TGMY), matéria seca foliar ("*Leaf Dry Matter Yield*" - LDMY) e matéria seca total ("*Total Dry Matter Yield*" - TDMY) permite prever o nível de produção da cultura por meio de modelos de aprendizado avançados (Castro et al., 2020; de Oliveira et al., 2021). O uso de *drones* é motivado pelo baixo custo operacional, flexibilidade e baixo risco para obtenção de imagens de uma grande área cultivada, fundamental para a fenotipagem de alto rendimento (HTP) (Tsouros et al., 2019). Os quadricópteros podem ser equipados com câmeras convencionais que fornecem as bandas de espectro vermelho, verde e azul (RGB) ou sensores multiespectrais, que geram dados de espectro de luz visível e não visível.

Vários estudos apontam para a usabilidade do "Veículo Aéreo Não Tripulado" (VANTs) acoplados a sensores, como um dispositivo de alto impacto na coleta de informações em relação aos métodos tradicionais de manejo e monitoramento agrícola (Rokhmana, 2015; Maes and Steppe, 2019; Negash et al., 2019). É uma tecnologia de sucesso no setor agrícola pois suas aplicações se intensificaram como importante ferramenta na agricultura de precisão e no reconhecimento da exploração agrícola por meio de imagens aéreas (Mogili and Deepak, 2018; Radoglou-Grammatikis et al., 2020). No entanto, prever padrões e estimar massa de forragem em programas de melhoramento é um processo altamente complexo. Algumas abordagens de aprendizado profundo destacam o uso de modelos pré-treinados para estimar a massa de forragem, geralmente com resultados divergentes entre os modelos de aprendizado profundo (Zhang et al., 2019; Narayanan et al., 2021; Pache et al., 2022). A explicação para a diferença desses resultados está na dificuldade de alguns modelos em encontrar um padrão de reconhecimento, erro na atualização das camadas com precisão insuficiente, implementação do modelo para domínios específicos e em alguns casos a utilização de um conjunto de dados relativamente pequeno para o treinamento.

Para tentar resolver alguns desses problemas este trabalho propõe o uso de aprendizagem de representação multivisão baseada em fusão de características em Redes Neurais Convolucionais (CNNs) (Li et al., 2018b), uma técnica explorada em vários trabalhos (Peng et al., 2019b; Feng et al., 2019; Amin et al., 2020; Li et al., 2021; Ying et al., 2021), que abrange vários domínios de aplicativos e permite a combinação de várias fontes de dados e recursos (Ramachandram and Taylor, 2017). Nessa perspectiva é possível mapear todas essas informações em um sistema de aprendizado, para que a tomada de decisão seja mais assertiva em problemas de classificação e previsão de resultados. O uso adequado dessas técnicas permite estimar a massa de forragem em programas de melhoramento com uma perspectiva de informações complementares de diversos recursos. Esta abordagem visa correlacionar as imagens coletadas pelo UAV-RGB e a massa de forragem medida no campo por meio de modelos de aprendizado em visão computacional (Chen et al., 2021b). Posteriormente um modelo é criado para combinar as saídas produzidas pelos modelos de visão, a fim de aprimorar a interconexão entre as camadas das redes neurais e alcançar representações mais refinadas (Guo et al., 2019). Alguns estudos vão na mesma direção deste trabalho, porém com abordagens multimodais focadas na agricultura de precisão em problemas específicos que não consideram o tema deste trabalho (Ali et al., 2015; Viljanen et al., 2018; Chen et al., 2021c). Esta pesquisa examina seis redes neurais convolucionais de última geração para reconhecimento de padrões e imagens, com foco em sensoriamento remoto e agricultura de precisão. A metodologia proposta pode ser empregada para melhorar a maioria dos modelos de aprendizado em visão computacional.

A prática adotada no aprendizado multimodal e multivisão é construir uma camada de representação mesclando dados de várias fontes ou multivisões da mesma entrada, para que a rede componha uma representação compartilhada desses recursos (Atrey et al., 2010), conforme ilustrado pela Figura 5.1.

Vários trabalhos foram propostos na literatura para lidar com redes multimodais (Eitel et al., 2015; Liu et al., 2016; Radu et al., 2018; Oramas et al., 2018; Guo et al., 2019) e redes multivisão (Wang et al., 2015b; Hao et al., 2017; Zheng et al., 2023) que descrevem questões que influenciam o processo da fusão multimodal, representação multivisão, correlação, independência dos dados, sincronização e seleção entre diferentes modalidades.

A extração de características neste experimento foi realizada por meio da utilização de duas CNNs, uma fusão tardia e métricas de regressão para estimar a massa de forragem em *Panicum maximum* Jacq. A Figura 5.2 exemplifica um modelo multivisão, começando com a extração de imagens usando

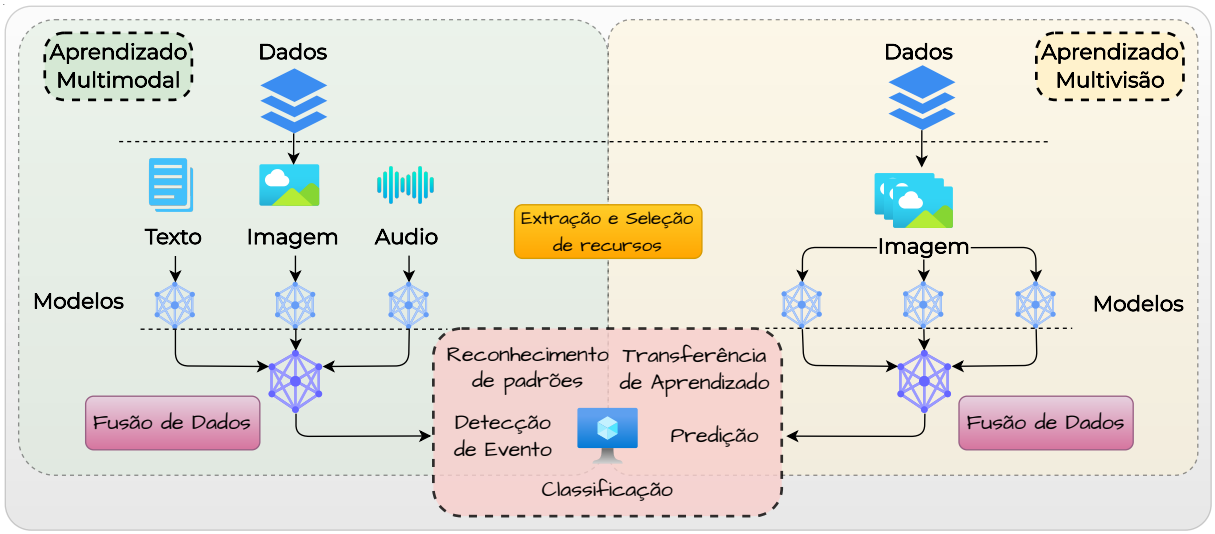


Figura 5.1: Aprendizado Multimodal versus Aprendizado Multivisão.

um VANT, então: (i) há mapeamento e processamento de imagens; (ii) extração e seleção de atributos; (iii) execução das duas redes neurais convolucionais (CNN) que obtiveram o melhor desempenho entre os modelos selecionados neste trabalho; (iv) aplicação de um modelo multivisão para mesclar a saída das redes CNN; (v) finalmente, gerar uma predição compartilhada.

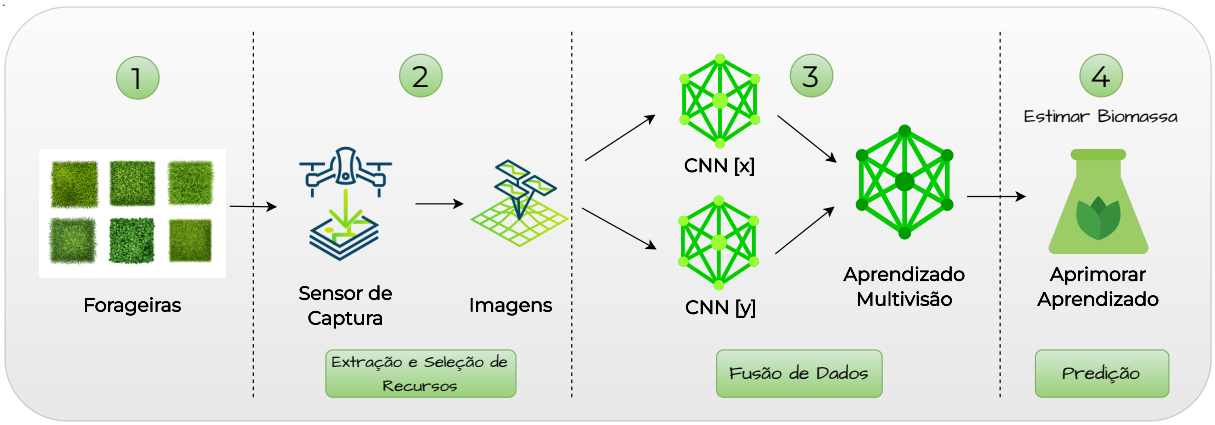


Figura 5.2: Visão geral do Aprendizado Multivisão aplicado à estimativa de massa de forragem.

Pré-processamento e Configuração Experimental

O conjunto de dados foi obtido através do processo de fotogrametria por um VANT a bordo com uma câmera digital RGB com resolução de imagem de 5472x3649 para captura dos genótipos *Panicum maximum*, já citados em experimentos unimodais anteriores (Apêndices A.0.6, A.0.7). A área de estudo está localizada na Embrapa Gado de Corte [1] (Embrapa, 2020), Campo Grande, Mato Grosso do Sul, Brasil – latitude $20°26'46''S$, longitude $54°43'16''W$ e altitude

[1] https://www.embrapa.br/gado-de-corte

535m, presente na Figura 5.3. O voo foi realizado no dia 23 de janeiro de 2019, por volta das 9 horas da manhã, com múltiplas alturas, implicando diferentes distâncias amostrais do solo: 0,5, 1,0, 1,5 cm/pixel.

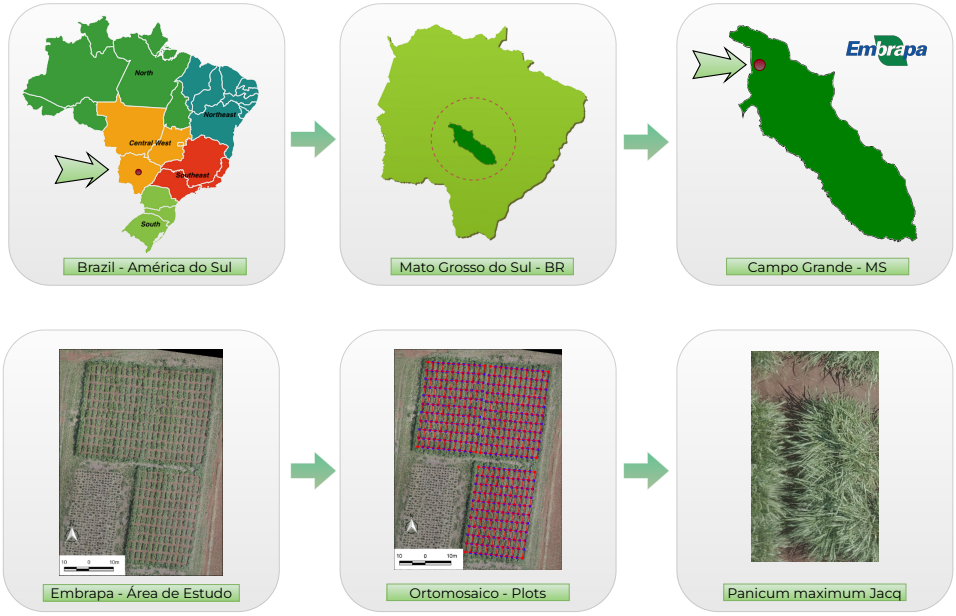


Figura 5.3: Visualização da área de estudo.

O *pipeline* para esta abordagem é mostrado na Figura 5.4 e segue os mesmos parâmetros usados em experimentos unimodais anteriores (Apêndices A.0.6, A.0.7), porém imagens com três GSD diferentes (0.5, 1.0 e 1.5 $cm/pixel$) são usadas neste experimento. Resumidamente, as imagens RGB foram coletadas em um experimento com 330 parcelas com diferentes genótipos da espécie por um UAV. Cada parcela constou de duas linhas de 2,0m de comprimento e 0,5m de distância. Cada linha contém cinco plantas com 0,5m de distância entre as plantas, com dez plantas por parcela. As parcelas distam 1,0m uma da outra, representando uma área de 4,5m^2. Em seguida ortoimagens foram criadas por meio de processamento computacional usando as imagens RGB obtidas com base na técnica de fotogrametria UAV (Tsouros et al., 2019), usando o software Pix4D baseado nas técnicas SfM (Structure-from-Motion) e MVS (Multi-view Stereo). Ao final desse processo, a ortoimagem de cada plotagem é encaminhada como entrada para as arquiteturas baseadas em regressão CNN.

Embora a criação do ortomosaico seja uma etapa essencial no processo utilizado até agora, trabalhar diretamente com as ortoimagens geradas não é fácil devido ao seu grande tamanho (*gigabytes*) e exigiria uma abordagem completamente diferente. Foi utilizado um *script* em Python denominado Field Plot Cropper (fPlotCropper)[2], cujo objetivo é extrair porções do ortomosaico, presentes na Figura 5.3 (Ortomosaico - *Plots*), de forma a resultar em pequenas

[2] https://github.com/wvmcastro/tiffviewer

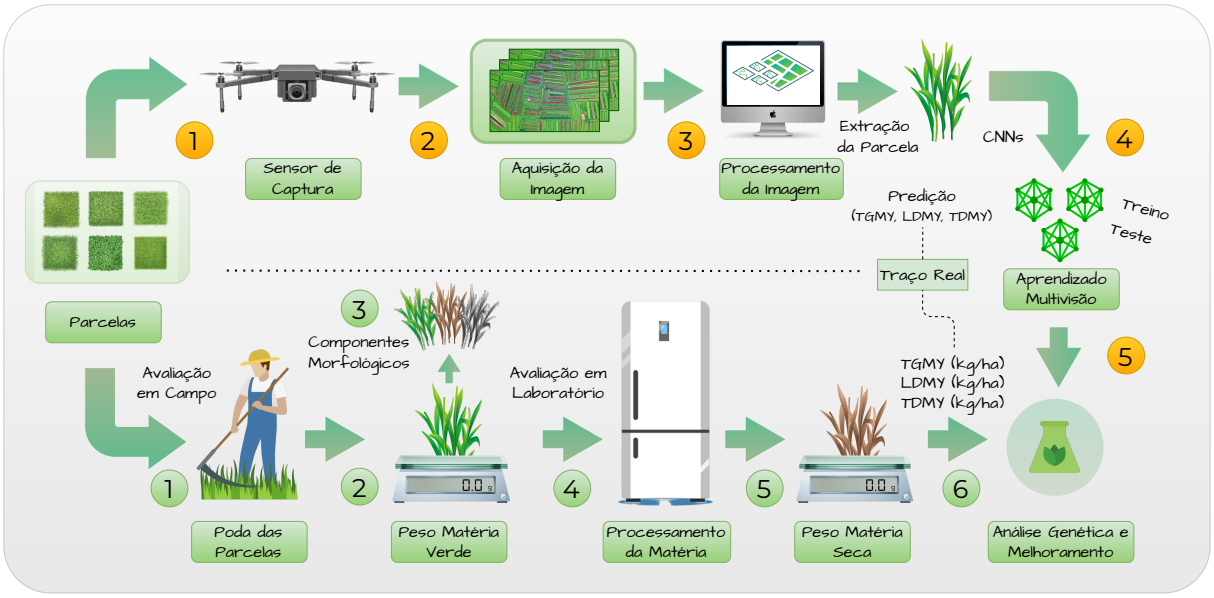


Figura 5.4: Pipeline.

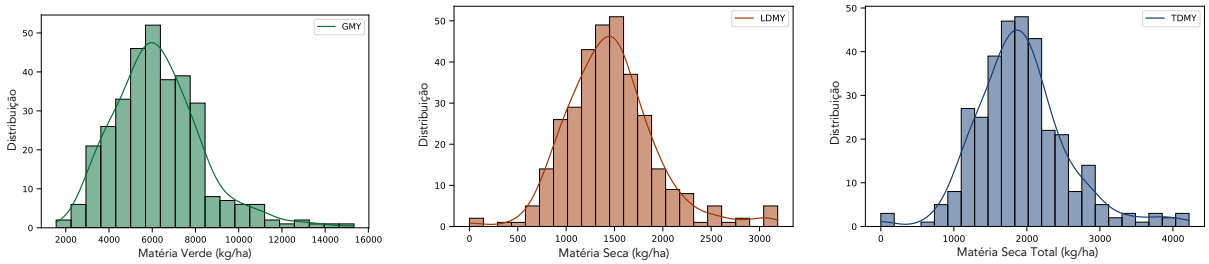
imagens com cada fenótipo isolado e identificado. Para o processamento de dados foi usado uma configuração de hardware tipo *desktop* com uma *CPU AMD Threadripper* 1900X (3,8 GHz), GPU 2x *Nvidia RTX* 3080 Ti e 64 GB de RAM DDR4 (*Quad-Channel*).

Os conjuntos de dados criados neste experimento correspondem a três GSDs diferentes, com imagens capturadas no mesmo local e no mesmo dia pela manhã. A comparação da produção de biomassa verde e matéria seca pelos modelos de aprendizado será realizada considerando as variações dos GSDs. A Tabela 5.1 resume as informações e parâmetros estabelecidos para a pesquisa realizada neste artigo. A técnica de aumento de dados foi utilizada para tornar o banco de dados mais robusto. As imagens foram invertidas da esquerda para a direita *(h incorporado)* e de cima para baixo (horizontalmente/verticalmente) *(hv incorporado)*, ao final do processo o conjunto de dados final obteve 990 imagens.

Tabela 5.1: Modelos e descrições dos conjuntos de dados.

Modelos	**Pré-Treinado**	**Número de Camadas**	**Número de Parâmetros**	**Tamanho do Lote**
AlexNet (Krizhevsky et al., 2017)	sim/não	8	62M	256
DarkNet53 (Redmon, 2016)	sim/não	53	42M	64
ResNext50 (Xie et al., 2017a)	sim/não	50	25M	64
MobileNetV3-Large (Koonce, 2021b)	sim	20	5.4M	256
SE-ResNet152 (Hu et al., 2018)	sim	152	60M	32
ViT-Base (Dosovitskiy et al., 2020)	sim	16	86M	64
GSD	**Resolução**	**Imagens**	**Tipo do Conjunto**	**Data**
0.5	479 x 791	330	incorporado hv	23/01/19
1.0	141 x 237	330	incorporado hv	23/01/19
1.5	220 x 368	330	incorporado hv	23/01/19

Como em trabalhos anteriores (Castro et al., 2020; de Oliveira et al., 2021) que usaram o mesmo conjunto de dados, a faixa de valores varia de 500 a 16000 $kg.ha^{-1}$ para essas características, conforme mostrado na Figura 5.5.



(a) TGMY (b) LDMY (c) TDMY

Figura 5.5: Atributo alvo y para Distribuição de Biomassa em $kg.ha^{-1}$

Resultados

A avaliação é composta por três etapas: (i) avaliação dos resultados usando as métricas padrão MAE, RSME, Correlação de *Pearson*; (ii) uma representação visual dos GSDs para medir sua densidade versus as métricas de avaliação; (iii) e uma análise de variância (ANOVA) (St et al., 1989) como método estatístico para comparar as variâncias entre as medianas de diferentes grupos.

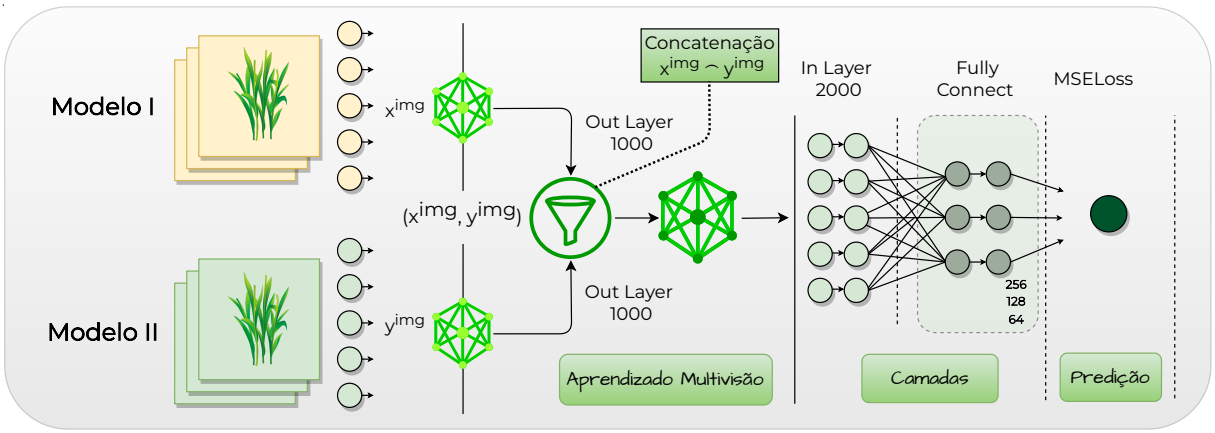


Figura 5.6: Modelo *Deep4Fusion* Multivisão utilizado neste trabalho.

A Figura 5.6 mostra a construção do modelo de aprendizado multivisão chamado neste trabalho de "*Deep4Fusion*", inicialmente cada parcela idêntica é enviada individualmente para os modelos de visão, então os dois modelos com a menor taxa de erro (MAE) são selecionados para formar uma rede multivisão. Em seguida as camadas de saída de cada modelo são concatenadas e passam por uma etapa de normalização (Ioffe and Szegedy, 2015) que permite utilizar maiores taxas de aprendizado, economizando a etapa de inicialização dos pesos. Finalmente camadas totalmente conectadas são aplicadas com uma redução no tamanho da dimensionalidade para calcular a perda. O otimizador selecionado para os experimentos foi *MSELoss* presente na biblioteca *PyTorch* (Paszke et al., 2019b). Por se tratar de um conjunto de dados reduzido foi necessário aplicar 10 vezes a validação cruzada para avaliar os

modelos estudados, e cada experimento utilizou 100 épocas por vez.

A Tabela 5.2, Tabela 5.3 e Tabela 5.4 apresentam o MAE, RMSE e coeficiente de correlação de *Pearson* para cada arquitetura CNN em relação às características da matéria verde e seca. Os modelos com um asterisco foram selecionados para compor a rede multivisão e simbolizam o modelo "*Deep4Fusion*" na tabela. Os resultados apresentados descrevem uma faixa de MAE para todos os GSDs entre 552,675 (*Deep4Fusion*) a 2918,362 (*ViT*) $kg.ha^{-1}$ para o matéria TGMY, 130,755 (*Deep4Fusion*) a 345,764 (*ViT*) $kg.ha^{-1}$ para matéria LDMY e 183.422 (*Deep4Fusion*) para 473.610 (*ViT*) $kg.ha^{-1}$ para o matéria TDMY. Neste cenário os modelos de visão que obtiveram melhor desempenho foram *MobileNetV3* e *SE-ResNet152* para TGMY, LDMY e TDMY, com os menores valores absolutos para MAE e RMSE, apenas para a característica TGMY com GSD 1.5 o modelo *AlexNet* foi superior em relação ao *SE-ResNet152*. Utilizando a estratégia de multivisão, foi possível maximizar ainda mais os resultados obtidos pelos modelos de visão. O ganho percentual obtido em relação aos modelos de visão para todos os GSDs foi de aproximadamente: (i) TGMY - 21%, (ii) LDMY - 37% e (iii) TDMY - 33%.

Tabela 5.2: Resultados - TGMY.

GSD 0.5						
Modelos	**Pré-Treinado**	**Épocas**	**Tamanho do Lote**	**MAE**	**RMSE**	**Pearson**
AlexNet	não	50	256	1054.336 ± 186.630	1398.346 ± 282.596	0.770 ± 0.084
AlexNet	sim	72	256	763.281 ± 147.449	1040.336 ± 191.322	0.878 ± 0.063
Darknet53	não	59	64	1135.241 ± 355.935	1457.416 ± 419.206	0.813 ± 0.058
***MobileNetV3**	**sim**	**52**	**256**	**708.448 ± 136.545**	**921.239 ± 179.276**	**0.896 ± 0.043**
ResNext50	não	54	64	1516.708 ± 832.245	1819.739 ± 919.388	0.758 ± 0.104
ResNext50	sim	71	64	832.739 ± 193.582	1113.225 ± 308.125	0.841 ± 0.087
***SE-ResNet152**	**sim**	**68**	**32**	**735.327 ± 105.151**	**964.135 ± 142.353**	**0.907 ± 0.024**
ViT-Base	sim	100	64	2918.362 ± 322.534	3515.553 ± 402.125	0.070 ± 0.143
Deep4Fusion	**sim**	**44**	**32**	**552.675 ± 97.072**	**703.677 ± 109.210**	**0.947 ± 0.022**
GSD 1.0						
AlexNet	não	75	256	1049.935 ± 243.279	1405.131 ± 392.281	0.758 ± 0.077
AlexNet	sim	67	256	887.662 ± 175.966	1137.448 ± 215.742	0.832 ± 0.065
Darknet53	não	62	64	1106.623 ± 336.317	1440.625 ± 394.416	0.796 ± 0.066
***MobileNetV3**	**sim**	**54**	**256**	**718.710 ± 65.271**	**951.728 ± 140.528**	**0.884 ± 0.041**
ResNext50	não	69	64	1038.428 ± 426.431	1278.732 ± 463.707	0.828 ± 0.085
ResNext50	sim	67	64	973.530 ± 169.273	1205.722 ± 218.489	0.820 ± 0.067
***SE-ResNet152**	**sim**	**70**	**32**	**755.175 ± 123.286**	**991.228 ± 152.007**	**0.900 ± 0.032**
ViT-Base	sim	100	64	2855.683 ± 320.917	3457.162 ± 405.324	0.094 ± 0.242
Deep4Fusion	**sim**	**55**	**32**	**616.125 ± 115.616**	**785.540 ± 146.253**	**0.926 ± 0.028**
GSD 1.5						
AlexNet	não	40	256	1033.770 ± 212.389	1389.809 ± 334.745	0.758 ± 0.070
***AlexNet**	**sim**	**73**	**256**	**835.987 ± 114.278**	**1086.609 ± 165.698**	**0.856 ± 0.034**
Darknet53	não	72	64	1170.312 ± 391.679	1478.530 ± 480.280	0.783 ± 0.099
***MobileNetV3**	**sim**	**49**	**256**	**803.512 ± 164.283**	**1053.758 ± 209.377**	**0.860 ± 0.049**
ResNext50	no	68	64	1058.451 ± 261.849	1352.384 ± 301.377	0.783 ± 0.079
ResNext50	yes	67	64	977.287 ± 203.612	1250.148 ± 270.235	0.818 ± 0.057
SE-ResNet152	yes	72	32	900.929 ± 339.212	1140.949 ± 364.060	0.882 ± 0.043
ViT-Base	sim	100	64	2888.560 ± 323.425	3487.867 ± 404.086	0.067 ± 0.177
Deep4Fusion	**sim**	**43**	**32**	**680.961 ± 133.845**	**866.488 ± 199.693**	**0.909 ± 0.065**

Tabela 5.3: Resultados - LDMY.

Modelos	Pré-Treinado	Épocas	Tamanho do Lote	MAE	RMSE	Pearson
GSD 0.5						
AlexNet	não	41	256	286.920 ± 39.164	393.134 ± 65.594	0.579 ± 0.090
AlexNet	sim	59	256	211.806 ± 35.422	294.004 ± 73.734	0.743 ± 0.154
Darknet53	não	57	64	259.770 ± 46.599	356.696 ± 79.716	0.682 ± 0.124
***MobileNetV3**	**sim**	**48**	**256**	**208.862 ± 26.953**	**290.610 ± 68.900**	**0.790 ± 0.083**
ResNext50	não	56	64	240.877 ± 33.184	326.774 ± 57.657	0.735 ± 0.056
ResNext50	sim	58	64	211.859 ± 28.746	291.820 ± 49.853	0.778 ± 0.091
***SE-ResNet152**	**sim**	**51**	**32**	**193.057 ± 35.033**	**278.092 ± 67.196**	**0.780 ± 0.222**
ViT-Base	sim	74	64	345.764 ± 78.955	461.035 ± 107.455	0.381 ± 0.190
Deep4Fusion	**sim**	**45**	**32**	**130.755 ± 35.844**	**173.086 ± 45.772**	**0.922 ± 0.061**
GSD 1.0						
AlexNet	não	56	256	280.671 ± 53.181	390.692 ± 85.885	0.562 ± 0.194
AlexNet	sim	61	256	230.070 ± 32.262	318.772 ± 57.524	0.735 ± 0.079
Darknet53	não	56	64	245.411 ± 34.406	335.352 ± 70.882	0.727 ± 0.073
***MobileNetV3**	**sim**	**46**	**256**	**215.918 ± 46.543**	**295.462 ± 85.708**	**0.771 ± 0.088**
ResNext50	não	53	64	247.987 ± 55.138	344.503 ± 77.790	0.680 ± 0.176
ResNext50	sim	56	64	233.093 ± 30.173	325.092 ± 68.743	0.707 ± 0.155
***SE-ResNet152**	**sim**	**50**	**32**	**217.267 ± 41.574**	**291.914 ± 78.349**	**0.811 ± 0.107**
ViT-Base	sim	76	64	333.811 ± 81.101	451.074 ± 109.789	0.361 ± 0.200
Deep4Fusion	**sim**	**44**	**32**	**151.392 ± 56.389**	**203.962 ± 84.928**	**0.876 ± 0.116**
GSD 1.5						
AlexNet	não	39	256	275.052 ± 57.531	375.221 ± 76.894	0.642 ± 0.101
AlexNet	sim	59	256	230.460 ± 38.403	321.110 ± 76.890	0.740 ± 0.040
Darknet53	não	54	64	254.345 ± 36.588	344.695 ± 71.500	0.698 ± 0.110
***MobileNetV3**	**sim**	**46**	**256**	**219.756 ± 37.928**	**302.254 ± 58.132**	**0.745 ± 0.143**
ResNext50	não	54	64	238.628 ± 50.378	328.606 ± 75.091	0.694 ± 0.134
ResNext50	sim	55	64	220.436 ± 43.508	295.578 ± 78.724	0.751 ± 0.149
***SE-ResNet152**	**sim**	**51**	**32**	**212.481 ± 28.271**	**299.726 ± 68.126**	**0.777 ± 0.094**
ViT-Base	sim	78	64	344.430 ± 79.643	459.285 ± 108.066	0.376 ± 0.171
Deep4Fusion	**sim**	**36**	**32**	**178.397 ± 29.816**	**243.739 ± 63.905**	**0.853 ± 0.086**

Tabela 5.4: Resultados - TDMY.

Modelos	Pré-Treinado	Épocas	Tamanho do Lote	MAE	RMSE	Pearson
GSD 0.5						
AlexNet	não	41	256	380.910 ± 53.152	537.386 ± 83.091	0.549 ± 0.134
AlexNet	sim	55	256	305.514 ± 90.954	415.959 ± 149.466	0.764 ± 0.177
Darknet53	não	55	64	389.320 ± 82.380	523.379 ± 122.399	0.637 ± 0.183
***MobileNetV3**	**sim**	**44**	**256**	**284.060 ± 53.268**	**396.601 ± 114.472**	**0.807 ± 0.094**
ResNext50	não	59	64	340.677 ± 33.175	465.017 ± 80.249	0.686 ± 0.130
ResNext50	sim	61	64	312.945 ± 69.646	439.338 ± 102.134	0.710 ± 0.161
***SE-ResNet152**	**sim**	**59**	**32**	**275.530 ± 56.250**	**377.998 ± 115.943**	**0.822 ± 0.095**
ViT-Base	sim	96	64	473.610 ± 81.412	643.433 ± 130.297	0.457 ± 0.145
Deep4Fusion	**sim**	**44**	**32**	**214.849 ± 47.056**	**297.982 ± 81.995**	**0.867 ± 0.088**
GSD 1.0						
AlexNet	não	45	256	413.684 ± 74.151	570.535 ± 97.791	0.453 ± 0.105
AlexNet	sim	67	256	294.954 ± 39.250	415.849 ± 95.007	0.736 ± 0.132
Darknet53	não	57	64	347.814 ± 52.272	465.765 ± 78.580	0.702 ± 0.114
***MobileNetV3**	**sim**	**49**	**256**	**294.453 ± 41.027**	**420.559 ± 85.597**	**0.759 ± 0.119**
ResNext50	não	56	64	329.197 ± 38.055	448.948 ± 67.035	0.706 ± 0.111
ResNext50	sim	58	64	317.707 ± 70.428	432.891 ± 122.653	0.722 ± 0.146
***SE-ResNet152**	**sim**	**56**	**32**	**310.351 ± 57.480**	**428.037 ± 100.509**	**0.773 ± 0.140**
ViT-Base	sim	97	64	472.660 ± 81.043	642.448 ± 129.795	0.349 ± 0.098
Deep4Fusion	**sim**	**39**	**32**	**246.067 ± 55.351**	**323.963 ± 93.330**	**0.875 ± 0.058**
GSD 1.5						
AlexNet	não	52	256	364.867 ± 70.952	520.811 ± 92.606	0.599 ± 0.107
AlexNet	sim	59	256	326.049 ± 68.554	454.985 ± 124.139	0.713 ± 0.162
Darknet53	não	64	64	355.348 ± 55.833	477.868 ± 72.688	0.655 ± 0.116
***MobileNetV3**	**sim**	**46**	**256**	**313.079 ± 42.516**	**445.714 ± 82.230**	**0.713 ± 0.156**
ResNext50	não	66	64	344.733 ± 76.600	461.241 ± 133.879	0.678 ± 0.232
ResNext50	sim	58	64	329.299 ± 63.948	454.241 ± 129.616	0.693 ± 0.160
***SE-ResNet152**	**sim**	**54**	**32**	**300.863 ± 57.812**	**425.205 ± 117.391**	**0.770 ± 0.147**
ViT-Base	sim	99	64	471.962 ± 79.764	642.001 ± 128.861	0.398 ± 0.204
Deep4Fusion	**sim**	**42**	**32**	**183.422 ± 71.703**	**229.775 ± 91.391**	**0.926 ± 0.060**

Para métrica de correlação de *Pearson* houve uma variação em todos os GSDs de 0,067 (*ViT*) a 0,947 (*Deep4Fusion*) para TGMY, 0,361 (*ViT*) a 0,922 (*Deep4Fusion*) para LDMY e 0,349 (*ViT*) a 0,926 (*Deep4Fusion*) para TDMY. Essa diferença descreve uma alta correlação significativa entre os dados estimados pelos modelos de visão e multivisão com os dados reais obtidos em campo. Portanto, há evidências que garantem uma forte relação entre as ima-

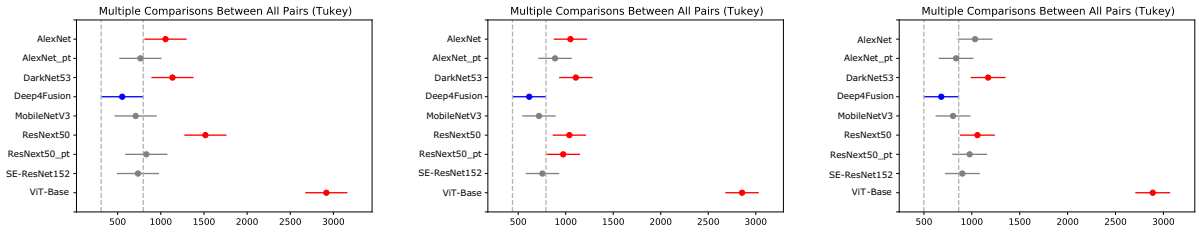
gens coletadas pelo VANT com os dados de matéria verde e seca utilizando as CNNs descritas na literatura, juntamente com os modelos multivisão inseridos neste trabalho. No geral, as três principais redes com os melhores resultados foram *MobileNetV3*, *SE-ResNet152* e *Deep4Fusion*. *ViT* obteve os piores resultados para ambas as características, seguido por *AlexNet* e *DarkNet53*.

Análise de Variância (ANOVA)

Tabela 5.5: ANOVA *one-way*.

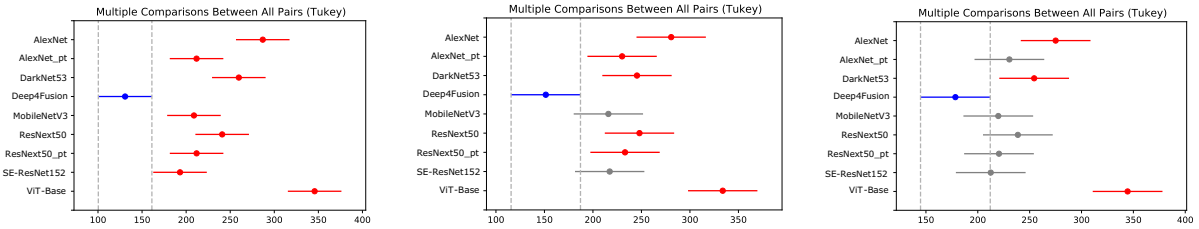
ANOVA				
Matéria	**GSD**	**f-estatística**	**p-valor**	**Resultado**
TGMY	0.5	45,1793	9,7084080492E-27	Rejeitar hipótese nula
	1.0	74,6585	3,6917607500E-34	Rejeitar hipótese nula
	1.5	68,6202	7,1370153025E-33	Rejeitar hipótese nula
LDMY	0.5	20,5771	1,2278149366E-16	Rejeitar hipótese nula
	1.0	9,7528	2,2213792599E-09	Rejeitar hipótese nula
	1.5	9,9757	1,4607996208E-09	Rejeitar hipótese nula
TDMY	0.5	13,3998	3,7078637396E-12	Rejeitar hipótese nula
	1.0	13,6566	2,4448902638E-12	Rejeitar hipótese nula
	1.5	12,7782	1,0333824371E-11	Rejeitar hipótese nula

Assim como no Experimento (1.1) descrito na Seção 3.2, para este estudo uma Análise de Variância (ANOVA) *one-way* foi realizado com base nos dados de cada modelo para mapear as diferenças significativas avaliadas neste trabalho. A literatura descreve este teste como uma extensão do *T-Test Student* não pareado (Kalpić et al., 2011) para mais de dois grupos. A análise permite rejeitar a hipótese nula de que os modelos funcionam igualmente usando o MAE, representado graficamente por Figura 5.7, Figura 5.8 e Figura 5.9 para o Teste de *Tukey*.



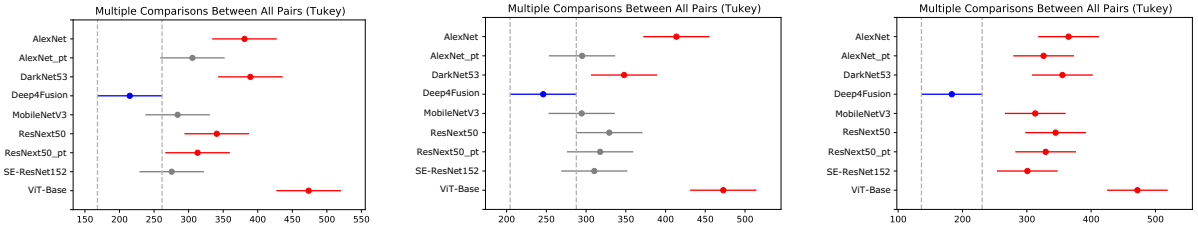
(a) GSD 0.5 (b) GSD 1.0 (c) GSD 1.5

Figura 5.7: Média e intervalo de confiança de 95% para o teste *pos-hoc Tukey* HSD sobre a métrica MAE - TGMY.



(a) GSD 0.5 (b) GSD 1.0 (c) GSD 1.5

Figura 5.8: Média e intervalo de confiança de 95% para o teste *pos-hoc Tukey* HSD sobre a métrica MAE - LDMY.



(a) GSD 0.5 (b) GSD 1.0 (c) GSD 1.5

Figura 5.9: Média e intervalo de confiança de 95% para o teste *pos-hoc Tukey* HSD sobre a métrica MAE - TDMY.

Avaliação Comparativa

Para avaliar o potencial do modelo multivisão sobre os modelos de visão investigados em neste estudo, uma comparação estatística foi realizada por meio da análise de trabalhos anteriores que compõem os mesmos critérios, conjunto de dados e modelos descritos na literatura. As métricas coletadas para medir o desempenho dos modelos foram (MSE, RSME e Correlação de *Pearson*). A Tabela 5.6, Tabela 5.7 e Tabela 5.8 apresentam os melhores resultados obtidos nos trabalhos de Castro et al. (2020) e de Oliveira et al. (2021) para GSD 0.5. Os resultados apresentados fornecem uma boa indicação da eficácia dos modelos de visão múltipla. Não houve comparação estatística com os GSDs 1.0 e 1.5 por não haver dados publicados em trabalhos anteriores, porém empiricamente GSDs com fator acima de 0,5 podem ter resultados significativos quando analisados sob a ótica de modelos multivisão. Os resultados indicam uma diferença mínima entre os GSDs analisados e que podem otimizar o processo de avaliação, visto que GSDs acima de 0.5 consomem menos processamento de dados na construção de ortomosaicos.

Tabela 5.6: TGMY - MAE, RMSE e Correlação de *Pearson* para comparação de desempenho com trabalhos anteriores.

GSD	Modelos	MAE	RMSE	Pearson
0.5	AlexNet (Castro et al., 2020)	730.000 ± 59.000	-	0.880 ± 0.040
	MobileNetV3	708.448 ± 136.545	921.239 ± 179.276	0.896 ± 0.043
	SE-ResNet152	735.327 ± 105.151	964.135 ± 142.353	0.907 ± 0.024
	Deep4Fusion	**552.675 ± 97.072**	**703.677 ± 109.210**	**0.947 ± 0.022**
1.0	MobileNetV3	718.710 ± 65.271	951.728 ± 140.528	0.884 ± 0.041
	SE-ResNet152	755.175 ± 123.286	991.228 ± 152.007	0.900 ± 0.032
	Deep4Fusion	**616.125 ± 115.616**	**785.540 ± 146.253**	**0.926 ± 0.028**
1.5	AlexNet	835.987 ± 114.278	1086.609 ± 165.698	0.856 ± 0.034
	MobileNetV3	803.512 ± 164.283	1053.758 ± 209.377	0.860 ± 0.049
	Deep4Fusion	**680.961 ± 133.845**	**866.488 ± 199.693**	**0.909 ± 0.065**

Tabela 5.7: LDMY - MAE, RMSE e Correlação de *Pearson* para comparação de desempenho com trabalhos anteriores.

GSD	Modelos	MAE	RMSE	Pearson
0.5	AlexNet (de Oliveira et al., 2021)	204.390 ± 56.460	286.240 ± 80.390	0.790 ± 0.120
	MobileNetV3	208.862 ± 26.953	290.610 ± 68.900	0.790 ± 0.083
	SE-ResNet152	193.057 ± 35.033	278.092 ± 67.196	0.780 ± 0.222
	Deep4Fusion	**130.755 ± 35.844**	**173.086 ± 45.772**	**0.922 ± 0.061**
1.0	MobileNetV3	215.918 ± 46.543	295.462 ± 85.708	0.771 ± 0.088
	SE-ResNet152	217.267 ± 41.574	291.914 ± 78.349	0.811 ± 0.107
	Deep4Fusion	**151.392 ± 56.389**	**203.962 ± 84.928**	**0.876 ± 0.116**
1.5	MobileNetV3	219.756 ± 37.928	302.254 ± 58.132	0.745 ± 0.143
	SE-ResNet152	212.481 ± 28.271	299.726 ± 68.126	0.777 ± 0.094
	Deep4Fusion	**178.397 ± 29.816**	**243.739 ± 63.905**	**0.853 ± 0.086**

Tabela 5.8: TDMY - MAE, RMSE e Correlação de *Pearson* para comparação de desempenho com trabalhos anteriores.

GSD	Modelos	MAE	RMSE	Pearson
0.5	AlexNet (de Oliveira et al., 2021)	289.660 ± 96.280	419.950 ± 136.930	0.750 ± 0.200
	MobileNetV3	284.060 ± 53.268	396.601 ± 114.472	0.807 ± 0.094
	SE-ResNet152	275.530 ± 56.250	377.998 ± 115.943	0.822 ± 0.095
	Deep4Fusion	**214.849 ± 47.056**	**297.982 ± 81.995**	**0.867 ± 0.088**
1.0	MobileNetV3	294.453 ± 41.027	420.559 ± 85.597	0.759 ± 0.119
	SE-ResNet152	310.351 ± 57.480	428.037 ± 100.509	0.773 ± 0.140
	Deep4Fusion	**246.067 ± 55.351**	**323.963 ± 93.330**	**0.875 ± 0.058**
1.5	MobileNetV3	313.079 ± 42.516	445.714 ± 82.230	0.713 ± 0.156
	SE-ResNet152	300.863 ± 57.812	425.205 ± 117.391	0.770 ± 0.147
	Deep4Fusion	**183.422 ± 71.703**	**229.775 ± 91.391**	**0.926 ± 0.060**

Discussão

Neste estudo foi implementada uma rede multivisão formada por dois modelos de visão com fusão tardia (Gao et al., 2020). Conforme descrito nas seções anteriores os resultados indicam que o *MobileNetV3* e o *SE-ResNet152* apresentam o melhor desempenho em relação aos demais modelos. A *AlexNet* (pré-treinada) obteve os melhores resultados em trabalhos anteriores foi

superada por modelos que utilizam redes *Squeeze-and-Excitation* (*SE-Net*) em suas camadas (Hu et al., 2018). Uma observação interessante é que modelos pequenos como o *MobileNetV3*, utilizado para rodar modelos em dispositivos móveis (Koonce, 2021b), tiveram resultados satisfatórios em relação a modelos mais robustos, como é o caso do *SE-ResNet152* e do *ViT*. De acordo com a pesquisa os modelos pré-treinados para o domínio de dados estudado apresentam maior eficiência do que aqueles que não passaram por pré-treinamento.

A investigação de modelos de visão múltipla mostrou o potencial para construir novas arquiteturas, mantendo o desempenho comparável aos métodos de aprendizado de visão única de última geração. No entanto, vários estudos devem ser repetidos e ampliados para reafirmar as validações sobre os modelos multivisão. Há fortes evidências de que modelos SOTA baseados em mecanismos de atenção, compressão e excitação, representação visual em escala, entre outros métodos atuais, podem ser mais eficientes do que modelos legados já descritos na literatura moderna (Chen et al., 2021b). Aumentar o conjunto de dados com imagens mais precisas pode ampliar os resultados obtidos, uma vez que dispositivos móveis e de sensoriamento remoto estão mais acessíveis aos pesquisadores da área. No entanto, como mencionado anteriormente, aumentaria proporcionalmente o poder computacional para avaliar experimentos dessa magnitude.

Este estudo foi publicado no periódico "*Computers and Electronics in Agriculture - Elsevier*", sob o título "*Deep4Fusion: a Deep FORage Fusion framework for high-throughput phenotyping for green and dry matter*".

CAPÍTULO

6

Conclusões

Neste capítulo são apresentadas as conclusões deste trabalho. Na 6.1 são apresentadas as principais contribuições desta tese; na 6.2 são discutidas algumas limitações dos métodos propostos neste trabalho para abordagens multimodais; por fim, na 6.3 são apresentadas algumas ideias para trabalhos futuros.

6.1 Principais Contribuições

O trabalho desenvolvido nesta tese explora o uso das redes multimodais, sobretudo destaca-se o uso das operações de fusão em problemas com domínios onde há várias fontes de dados. A Tabela 6.1 apresenta uma lista completa de trabalhos que foram realizados na fase de experimentação e contribuíram para o aperfeiçoamento desta pesquisa. Uma variedade de publicações foi aceita como resultado de estudos individuais, parcerias e coparticipações com os pesquisadores do Laboratório de Inteligência Artificial (UFMS-FACOM) (LIA, 2019) e Bioinspirada (ICMC-USP) (BIOCOM, 2020). Algumas propostas avaliadas ao longo do processo de pesquisa ainda não foram submetidas ou estão em processo de avaliação em periódicos da área. Os estudos e propostas avaliados incluem tanto abordagens já consolidadas na literatura, quanto contribuições inovadoras para a área do estudo em questão.

As hipóteses levantadas na introdução a respeito dos objetivos deste trabalho são apresentados novamente, desta vez com algumas das soluções encontradas.

Tabela 6.1: Trabalhos produzidos dentro do Doutorado.

	Título	Dados			Aprendizado	Área	Ano	Revista/Conferência	Primeiro Autor	Status
		Tabular	Texto	Imagem						
1	BERT for Stock Market Sentiment Analysis		✓		Unimodal	Financeira	2019	ICTAI		Publicado
2	NER com classes desbalanceadas e o impacto na manipulação de pesos		✓		Unimodal	Jurídico	2019	Não Publicado	✓	-
3	Can Twitter data estimate Reality Show outcomes?		✓		Unimodal	Rede Social	2020	BRACIS	✓	Publicado
4	Deep Learning Applied to Phenotyping of Biomass in Forages with UAV-Based RGB Imagery			✓	Unimodal	Sensoriamento	2020	MDPI - *Sensors*		Publicado
5	Deep Learning Multimodal applied to phenotyping of biomass inforages with UAV-based RGB Imagery	✓		✓	Multimodal	Sensoriamento	2020	Não Publicado	✓	-
6	Convolutional Neural Networks to Estimate Dry Matter Yield in a Guineagrass Breeding Program Using UAV Remote Sensing			✓	Unimodal	Sensoriamento	2021	MDPI - *Sensors*		Publicado
7	Entity extraction from Portuguese legal documents using distant supervision		✓		Unimodal	Jurídico	2022	PROPOR		Publicado
8	Identification of Controversial Political Topics using Twitter Data		✓		Unimodal	Rede Social	2022	BRACIS		Publicado
9	Successful Youtube video identification using multimodal deep learning		✓	✓	Multimodal	Rede Social	2022	KDMile	✓	Publicado
10	Deep4Fusion: a Deep FORage Fusion framework for high-throughput phenotyping for green and dry matter			✓	Multivisão	Sensoriamento	2023	Computers and Electronics in Agriculture	✓	Publicado
11	Exploring the effectiveness of multimodal knowledge distillation: findings and implications		✓	✓	Multimodal	Multi-área	2023	Expert Systems with Applications	✓	Submetido
12	MASK: a faster convergence approach using Multimodal Attention + SKip connections		✓	✓	Multimodal	Multi-área	2023	Neuralcomputing	✓	Submetido

Hipótese 1 - As operações aritméticas em abordagens multimodais são promissoras para a fusão dos dados em relação a outras técnicas tradicionais já difundidas na literatura.

Resposta: As operações aritméticas em abordagens multimodais mostraram-se promissoras para a fusão dos dados, superando o uso da concatenação uma técnica amplamente discutida na literatura (Guo et al., 2019; Gao et al., 2020; Kiela et al., 2020; Wang et al., 2020c). Em particular a operação de subtração teve maior destaque em relação as outras combinações de fusão. Enquanto a concatenação simplesmente combina as características das modalidades, a operação de subtração permite destacar as diferenças entre as modalidades. Ao subtrair as características entre as modalidades, a operação enfatiza as discrepâncias e variações presentes nos dados, muitas vezes útil quando as modalidades têm informações complementares. Essa afirmação pode ser evidenciada no Experimento 3.1, pois ao relacionar mais de uma fonte de dados para prever o número de visualizações em um vídeo, houve um ganho de um ganho de 3,9% sobre o modelo textual e 5,8% para o modelo visual com a operação de subtração.

A pesquisa sobre esse tema resultou na publicação do artigo (9) e submissão do artigo (11) descritos na Tabela 6.1.

Hipótese 2 - Em um cenário com diversos operadores de fusão de dados sem um resultado claramente superior, a utilização dos mecanismos de atenção pode contribuir para a combinação desses operadores e possibilitar a obtenção de um melhor desempenho.

Resposta: Os mecanismos de atenção tiveram um papel importante na relação com a fusão de dados. Sua estrutura utiliza características obtidas de diferentes partes de uma rede para resolver um problema sob diversas perspectivas (de Santana Correia and Colombini, 2022). Essa abordagem expandiu este trabalho sob o aspecto estrutural da rede multimodal reafirmando os resultados preliminares obtidos no primeiro experimento deste trabalho com ênfase para a operação aritmética de subtração em relação as outras combinações. Além dos mecanismos de atenção as etapas de normalização foram benéficas para estabilizar o treinamento e ajustar as saídas das camadas compartilhadas entre as modalidades.

A pesquisa sobre esse tema resultou na publicação do artigo (9) e submissão do artigo (11) descritos na Tabela 6.1.

Hipótese 3 - A utilização de conexões residuais (*Skip Connections*) em modelos multimodais resulta em uma melhor convergência dos modelos, semelhante

ao que é observado em modelos unimodais na literatura.

A incorporação de conexões residuais nos modelos multimodais no primeiro experimento proporcionou uma melhora na convergência do modelo, semelhante ao que é observado em modelos unimodais descritos na literatura (He et al., 2016; Adaloglou, 2020). A abordagem MASK foi 3,6 vezes mais rápida que modelo (Op) e 9,2 vezes mais rápida que modelo (Att), ilustradas na Figura 3.14.

A pesquisa sobre esse tema resultou na na publicação do artigo (10) e submissão do artigo (11 e 12) descritos na Tabela 6.1.

Hipótese 4 - Modelos de Aprendizado baseados em Destilação Multimodal resultam em uma transferência de conhecimento entre modalidades, com potencial de melhorar a capacidade de generalização entre as diversas modalidades.

A destilação de conhecimento permitiu treinar modelos menores e mais leves para tarefas específicas usando abordagens multimodais. O Experimento 4.1 aponta que a transferência de conhecimento entre modalidades pode conter uma relação de dependência e complementaridade. Os ganhos com a destilação neste trabalho foram mais eficazes quando modelos de linguagem foram caracterizados como professor sobre os modelos de visão. Logo, há indícios de que certas modalidades são mais importantes em certos domínios e estão diretamente ligadas à forma como as informações são organizadas e estruturadas em diferentes camadas no modelo multimodal.

A pesquisa sobre esse tema resultou na submissão do artigo (11) descrito na Tabela 6.1.

6.2 Limitações

Construir arquiteturas para unificação de várias fontes de dados é uma proposta promissora, sobretudo com o avanço das redes neurais, modelos pré-treinados e técnicas de fusão de dados. Porém, não é garantia cobrir perfeitamente todos os domínios e problemas encontrados no mundo real. Nesta seção são apresentadas algumas das limitações das soluções propostas neste trabalho:

1. Modelos de Aprendizado: A complexidade dos modelos de aprendizado utilizados tornou o processo de treinamento e avaliação mais dispendioso. Isso se deve ao grande número de parâmetros e GPUs necessários para executar versões maiores do *BERT* e *ViT*. Como resultado, o estudo

foi limitado a modelos menores. No entanto, os resultados do estudo foram promissores e fornecem uma base para estudos com modelos mais complexos.

2. Otimização de Hiperparâmetros: Um fator limitante neste trabalho foi a otimização dos hiperparâmetros para os modelos multimodais, ajustar dois modelos unimodais com taxas de aprendizados distintas, tamanhos variáveis (camadas, arquitetura, lote) e fontes de dados variadas para construção de um modelo multimodal, exige um cuidado adicional para que as etapas de fusão deem importância equivalente para ambas as modalidades. Outra limitação dos hiperparâmetros é que eles podem ser sensíveis para cada modalidade. O melhor conjunto de valores para uma modalidade pode não ser eficaz para outras modalidades de um mesmo domínio.

3. Métricas de Avaliação: As métricas usadas neste trabalho limitaram-se ao uso de métricas de classificação e regressão, já contempladas em estudos anteriores. Não foi possível testar algumas métricas relacionadas a recuperação de informação descritas na Seção 2.7.3, pois os cenários e conjuntos listados neste estudo restringiram-se a problemas de predição e análise de ganho de desempenho. Porém, algumas limitações perceptíveis obtidas neste estudo são: (i) dificuldade em comparar diretamente a relevância das informações de cada modalidades; (ii) informações ausentes em uma ou mais modalidades podem afetar a avaliação e análise das métricas; (iii) as informações compartilhadas entre as modalidades podem melhorar a capacidade de generalização ou reduzir a precisão das previsões, dependendo da natureza dos dados e do modelo utilizado.

4. Arquitetura Multimodal: A fusão de dados juntamente com outras técnicas para construção de modelos multimodais proporcionou uma série de experimentos, porém nesta pesquisa não houve uma estrutura específica para qualquer domínio multimodal. Pois as relações entre as modalidades de dados podem ser altamente dependentes do problema em questão, o que dificulta a definição de uma estratégia única para a fusão de dados. Os mecanismos de atenção fornecidos neste estudo também possuem resultados variados devido a complexidade e variabilidade dos dados em cada domínio.

6.3 Trabalhos Futuros

Abaixo são apresentadas algumas sugestões de possíveis refinamentos e extensões dos modelos multimodais e métodos apresentados neste trabalho.

1. Os modelos multimodais apresentados neste trabalho restringiram-se apenas a dados textuais e visuais, porém novas abordagens de aprendizado multimodal podem ser exploradas em tarefas como tradução automática, reconhecimento de fala, detecção de objetos, reconhecimento da atividade humana e sistemas autônomos.

2. Além das métricas já citadas neste estudo é possível relacionar novos indicadores para medir a eficiência dos modelos multimodais em relação aos modelos unimodais já descritos na literatura. Apesar deste estudo considerar várias métricas de avaliação, como ganho de informação, uma exploração mais ampla com outras métricas ainda é necessária.

3. Outra abordagem possível para expansão deste trabalho é investigar o desempenho de outras arquiteturas de redes neurais convolucionais e/ou recorrentes em tarefas multimodais. Explorar o uso de técnicas de aprendizado por reforço para melhorar o desempenho das redes em tarefas multimodais e destilar conhecimento por meio de estruturas multimodais, pode ser o primeiro ponto de partida para estudo futuros.

4. Algumas estratégias como normalização, conexões residuais e destilação de conhecimento foram essenciais para o aprimoramento desta pesquisa. Porém os mecanismos de atenção tiveram o maior destaque neste trabalho, pela sua capacidade de melhorar os modelos no sentido de capturar e integrar informações de diferentes fontes de forma mais eficiente e precisa. Um possibilidade futura é explorar outros mecanismos de atenção de última geração que permitam uma melhorar as relações entre as diferentes modalidades.

APÊNDICE

A

Estudos com Modelos Unimodais

Para um melhor entendimento das redes multimodais alguns estudos individualizados foram realizados juntamente com o grupo de pesquisadores do Laboratório de Inteligência Artificial (LIA-FACOM), Laboratório de Computação Bioinspirada (BIOCOM-ICMC) e Empresa Brasileira de Pesquisa Agropecuária (Embrapa, 2020), alguns dos trabalhos foram publicados em periódicos e conferências da área e outros estão em processo de experimentação e análise. O estudo foi dividido em fonte de dados textuais e visuais, sob a perspectiva das redes neurais e modelos de última geração.

A etapa textual compreende a aplicação de tarefas em PLN para o modelo *BERT*, como forma de mensurar os conjuntos de dados avaliados. Domínios específicos citados na Figura 2.21 foram utilizados na tarefa de análise de sentimentos para prever a reação de um determinado grupo em relação a questões econômicas, políticas e programas de televisão. Algumas abordagens para Reconhecimento de Entidade Nomeada (NER) definida como uma sub-tarefa de extração de informações que busca localizar e classificar entidades, foram realizadas com o intuito de melhorar a representatividade de entidades minoritárias em conjuntos de dados disponíveis na literatura, sobretudo conjuntos com temas jurídicos.

Dados visuais foram submetidos a diversos modelos de visão computacional descritos na literatura e abordados no Capítulo 2. Os experimentos com modelos visuais estão elencados em alguns trabalhos de campo realizados por pesquisadores da Embrapa com apoio de estudantes do

LIA. A pesquisa de campo teve como objetivo predizer a fenotipagem de biomassa, por meio de métodos baseados em aprendizado profundo e imagens RGB com uso de veículo aéreo não tripulado (UAV), os modelos selecionados foram *AlexNet, ResNet, VGGNet, MobileNet e ViT*.

As abordagens e técnicas envolvidas em cada experimento, são descritas nas seções seguintes divididos em sete experimentos individualizados.

A.0.1 Experimento 1 - Análise de Sentimento no Mercado Financeiro

Uma das tarefas relacionadas a PLN é a análise de sentimento, neste contexto há uma série de domínios que podem ser relacionados aos sentidos humanos para predizer uma ação futura. Prever o mercado financeiro é uma tarefa desafiadora, pois é influenciada por um conjunto de fatores externos e internos. Um das estratégias para verificar a volatilidade das ações é extrair informações em tempo real de reportagens postadas em veículos de imprensa e sites especializados no mercado de ações. Para este propósito o modelo *BERT* foi utilizado para realizar a análise de sentimento de reportagens e fornecer informações relevantes para a tomada de decisão, com um conjunto de dados anotados manualmente de artigos de notícias sobre ações consideradas positivas, neutras ou negativas.

Proposta

O Corpus anotado foi retirado dos seguintes sites e veículos de imprensa: CNBC, *Forbes*, *Investopedia*, *New York Times*, *Washington Post*, *Business Insider* e outros sites de notícias. Em seguida, um ajuste fino foi realizado com o modelo pré-treinado *BERT* (Face, 2022) para o domínio do problema e uma avaliação experimental comparativa entre Máquina de Vetores de Suporte (SVM) (Vapnik, 1999), *Naive Bayes* (NB) (McCallum et al., 1998) e *TextCNN* (Rosenthal et al., 2017) foi efetuada. Por fim, uma análise do modelo *BERT* em relação a sua predição com o índice *Dow Jones* (S&P, 2019), um forte indicador financeiro de ações norte-americano que avalia as 30 grandes ações industriais, foi elaborado para verificar a correlação das predições. A Tabela A.1, descreve o percentual de notícias extraídas por sites e veículos de imprensa, totalizando 582 artigos de notícias.

Tabela A.1: Fontes usadas para coletar notícias para construção do Corpus.

Fonte	**Artigos**	**%**
Business Insider	51	8,7
CNBC	77	13.2
Forbes	32	5.5
Investopedia	41	7.0
New York Times	45	7.7
Washington Post	31	5.3
Others	305	52.4
Total	**582**	**100.0**

Resultados Experimentais e Discussão

A análise dos dados avalia o teor das notícias financeiras antes do horário de abertura do mercado de ações. A ideia é reproduzir o cenário em que um agente financeiro se restringe a operar no horário de abertura da bolsa. Os experimentos foram ajustados com modelo *BERT-Base* pré-treinado em um conjunto rotulado. Para fins de avaliação experimental a validação cruzada de 10 partições foi utilizada. A Tabela A.2 descreve a avaliação experimental entre SVM, NB e *TextCNN* com textos convertidos na estrutura *Bag of Words* (BOW) (Zhang et al., 2010) e representação de frequência inversa do documento de frequência (TF-IDF) (Aizawa, 2003). Para *TextCNN*, um vetor médio de *embeddings* de palavras foi obtido pela biblioteca *fastText* (Bojanowski et al., 2017).

Tabela A.2: Resultados experimentais com validação cruzada de 10 partições.

Modelo	**Accuracy**	**Precision**	**Recall**	**F1**
NB bow	0.610 ± 0.060	0.593 ± 0.196	0.557 ± 0.069	0.503 ± 0.103
SVM bow	0.628 ± 0.063	0.627 ± 0.074	0.609 ± 0.066	0.601 ± 0.071
NB tfidf	0.610 ± 0.062	0.607 ± 0.102	0.568 ± 0.065	0.542 ± 0.080
SVM tfidf	0.624 ± 0.076	0.631 ± 0.104	0.595 ± 0.083	0.578 ± 0.099
TextCNN	0.739 ± 0.05	0.703 ± 0.18	0.500 ± 0.14	0.569 ± 0.12
BERT	$\mathbf{0.825 \pm 0.04}$	$\mathbf{0.750 \pm 0.17}$	$\mathbf{0.713 \pm 0.16}$	$\mathbf{0.725 \pm 0.15}$

Outro experimento também foi executado para verificar a relação entre o sentimento do mercado e a série temporal da bolsa de valores usando o *BERT* para classificar as notícias que foram coletadas durante aproximadamente um mês juntamente com o índice *Down Jones*. Uma série temporal com taxa de notícias positivas a cada hora e variação da bolsa de valores foi traçada, sob uma média móvel considerando as 10 horas anteriores, conforme ilustrado pela Figura A.1. No início de cada dia, 5 horas antes da abertura da bolsa foi calculado o sentimento médio das notícias. A premissa para este estudo é que o sentimento médio que precede a abertura do mercado de ações indique mais fortemente o humor do mercado no período em que a bolsa de valores está fechada.

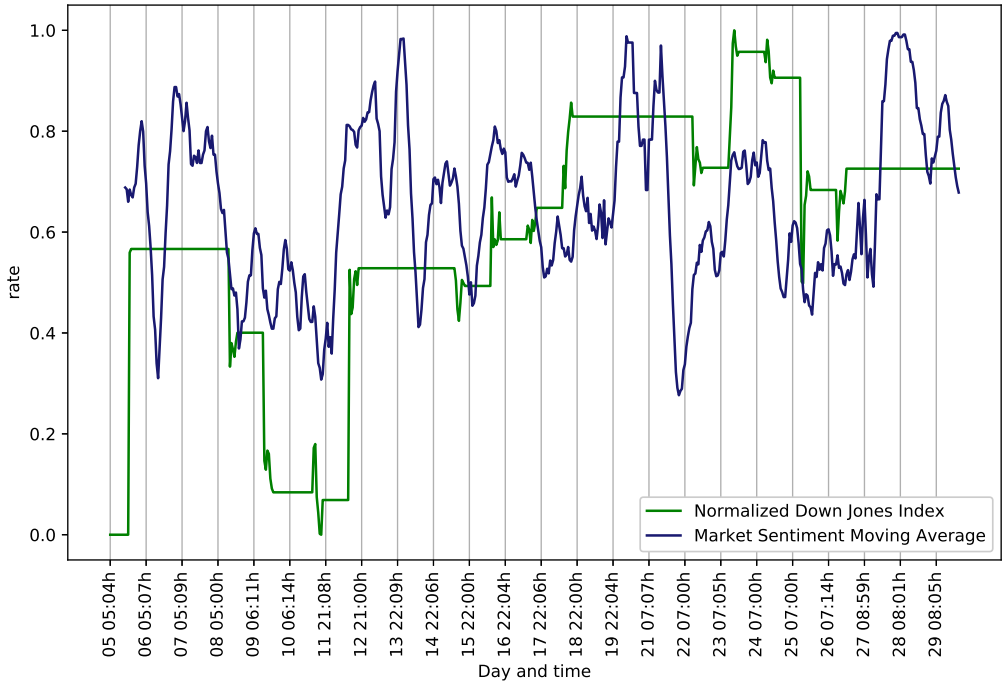


Figura A.1: Variação do Índice Normalizado Down Jones (Verde) e Média Móvel do Sentimento do Mercado (Azul). Fonte: Sousa et al. (2019)

Os resultados indicaram que *BERT* teve um desempenho superior ao das redes neurais convolucionais e abordagem com *embeddings* de palavras na ordem de 8,6% quando comparada à taxa de acerto (acurácia). No entanto, os resultados das séries temporais com análise de sentimento das notícias e o índice *Down Jones* possuem ruídos e são difíceis de analisar. Logo, é perceptível em determinados picos uma correlação entre a análise de sentimentos e o índice *Down Jones*.

Este estudo foi publicado na "*International Conference on Tools with Artificial Intelligence* (ICTAI 2019)", sob o título "*BERT for Stock Market Sentiment Analysis*" (Sousa et al., 2019).

A.0.2 Experimento 2 - NER com classes desbalanceadas e o impacto na manipulação de pesos

Ainda sobre o contexto de PLN, o reconhecimento de entidades nomeadas (NER) (Mohit, 2014) permite extrair informações que buscam localizar e classificar entidades em um texto com categorias predefinidas. Uma coleção de trabalhos (Souza et al., 2019; Hakala and Pyysalo, 2019; Liang et al., 2020) destacam o uso de NER em suas pesquisas aliadas ao modelo *BERT*, contudo há cenários em que a quantidade de termos relacionados a uma entidade nomeada é restrita ou escassa no conjunto de dados.

Técnicas como *oversamplig*, *undersampling* e *SMOTE*, são abordagens eficientes para tratar o conjunto de dados (Shelke et al., 2017) em abordagens NER. Seja para duplicar os dados existentes, sobretudo as entidades com poucos rótulos, reduzir o conjunto, balancear as entidades ou gerar dados sintéticos. Há também técnicas especializadas para manipular as entradas e pesos das entidades nomeadas em modelos de aprendizagem profunda, por exemplo a entropia cruzada ou *cross entropy* (CE) (De Boer et al., 2005), definida como uma medida de diferença entre duas distribuições de probabilidade para uma determinada variável aleatória ou um conjunto de eventos.

Proposta

Um experimento empírico foi elaborado a fim de verificar o impactado na manipulação dos pesos de entrada sobre as entidades para tarefas com NER. Inicialmente a implementação nativa do modelo *BERT* prevê o uso da entropia cruzada para relacionar as entidades nomeadas a um conjunto de dados. Com base nesta premissa, aplicar uma função estatística sobre as entidades nomeadas e adicionar pesos na entrada da entropia cruzada, é uma forma eficiente para determinar entidades de interesse. Logo, penalizar entidades minoritárias com pesos maiores, faz com que o modelo ajuste o desequilíbrio do conjunto.

Pré-processamento e Configuração Experimental

Os conjuntos utilizados neste experimentos são corpus estruturados em um formato NER, conhecido como (IOB2 - *Inside, Outside, Beginning*) (Ramshaw and Marcus, 1999). A Tabela A.3, descreve os quatro corpus separados por entidades e divididos em conjunto de treino, teste e validação. O corpus *LeNER-BR* possui reconhecimento de entidades nomeadas em documentos legais brasileiros (de Araujo et al., 2018), *HAREM* é um evento para avaliação de modelos de NER para o idioma português (Linguateca, 2020), *WikiNER* é um corpus formado por artigos da Wikipédia (Nothman et al., 2012) e Paramopama é um corpus que estende a versão em português do corpus *WikiNER* com revisões de entidades incorretas (Júnior et al., 2015).

Resultados Experimentais e Discussão

A Tabela A.4 descreve alguns métodos utilizados para melhorar a precisão de entidades minoritárias utilizando o modelo *BERT* pré-treinado

Tabela A.3: Entidades dos Corpus *LeNER-BR, HAREM, WikiNER* e Paramopama divididos em conjuntos de treino, teste e validação.

LeNER-BR				
ID	**Entidades**	**Treino**	**Validação**	**Teste**
0	JURISPRUDENCIA	4087	622	660
1	LEGISLACAO	13764	1883	2669
2	LOCAL	1448	211	132
3	ORGANIZACAO	7206	1072	1367
4	PESSOA	4774	730	735
5	TEMPO	2515	369	260
6	O	201899	29853	41807
	TOTAL	**235693**	**34740**	**47630**
HAREM				
ID	**Entidades**	**Treino**	**Validação**	**Teste**
0	ABSTRACCAO	679	178	628
1	ACONTECIMENTO	347	84	235
2	COISA	161	35	254
3	LOCAL	1696	447	1339
4	OBJECTO	1	0	0
5	OBRA	570	149	555
6	ORGANIZACAO	1789	389	1278
7	OUTRO	53	15	31
8	PESSOA	1661	367	1578
9	TEMPO	636	175	609
10	VALOR	807	173	599
11	O	79133	19743	59519
	TOTAL	**87533**	**21755**	**66625**
Wikiner				
ID	**Entidades**	**Treino**	**Validação**	**Teste**
0	LOC	138184	34611	56383
1	MISC	37619	9259	16979
2	ORG	25308	6226	9538
3	PER	63343	15675	27856
4	O	1838389	458283	762024
	TOTAL	**2102843**	**524054**	**872780**
Paramopama				
ID	**Entidades**	**Treino**	**Validação**	**Teste**
0	LOCAL	9805	2558	5097
1	ORGANIZACAO	3867	1005	2282
2	PESSOA	4320	1074	1932
3	TEMPO	6251	1542	3033
4	O	148207	37231	78491
	TOTAL	**172450**	**43410**	**90835**

no idioma português (Souza et al., 2020). Um total de quatro execuções foram utilizadas para avaliação experimental, sendo elas uma execução padrão sem adição de pesos como ponto de partida, uso do método "*Bangla*", descrito no trabalho de (Ashrafi et al., 2020) com penalidade para as entidades minoritárias em relação a entidade majoritária em vez de todo o conjunto de dados, aplicação do método "*Vandit*", definido no estudo de (Cui et al., 2019) com uso de reponderação que usa o número efetivo de exemplos para cada entidade para reequilibrar a perda e por último um método com pesos manuais com escala $\{0 \iff 10\}$. Os valores são representados na tabela pela métrica de taxa de acerto (acurácia) e desvio padrão. A coluna "*Dist Treino*", descreve a distribuição de exemplos por entidade nos experimentos realizados.

Os experimentos demonstram que usar medidas estatísticas para manipular pesos pode ajudar entidades minoritárias a obter uma maior taxa de acerto (acurácia), porém não há evidências conclusivas que modelos

Tabela A.4: Experimento NER com classes desbalanceadas sobre os conjuntos *HAREM, LeNER-BR, Paramopana e WikiNER.*

HAREM					
Entidades	**Dist Treino**	**BaseLine**	**Bangla**	**Vandit**	**Manual**
Outro	1789	0.152 ± 0.046	0.141 ± 0.032	**0.152 ± 0.042**	0.116 ± 0.052
Acontecimento	347	0.476 ± 0.044	0.487 ± 0.040	**0.489 ± 0.040**	0.395 ± 0.057
Coisa	161	0.477 ± 0.033	0.482 ± 0.027	0.493 ± 0.037	**0.534 ± 0.016**
Abstração	679	0.530 ± 0.027	**0.537 ± 0.016**	0.529 ± 0.011	0.514 ± 0.015
Obra	1	**0.572 ± 0.019**	0.570 ± 0.031	0.565 ± 0.036	0.525 ± 0.050
Organização	570	0.777 ± 0.008	**0.782 ± 0.011**	0.779 ± 0.006	0.755 ± 0.019
Valor	636	0.804 ± 0.010	**0.825 ± 0.011**	0.802 ± 0.011	0.787 ± 0.023
Pessoa	53	**0.843 ± 0.005**	0.840 ± 0.006	0.838 ± 0.008	0.835 ± 0.005
Local	1696	0.848 ± 0.008	**0.855 ± 0.007**	0.846 ± 0.003	0.837 ± 0.007
Tempo	1661	0.908 ± 0.005	**0.916 ± 0.008**	0.907 ± 0.005	0.902 ± 0.007
Average micro		0.778 ± 0.006	**0.783 ± 0.006**	0.778 ± 0.005	0.764 ± 0.007
Average macro		0.779 ± 0.005	**0.784 ± 0.006**	0.778 ± 0.005	0.766 ± 0.006
LeNER-BR					
Entidades	**Dist Treino**	**BaseLine**	**Bangla**	**Vandit**	**Manual**
LOCAL	1448	**0.755 ± 0.034**	0.715 ± 0.048	0.724 ± 0.070	0.693 ± 0.055
JURISPRUDENCIA	4087	**0.815 ± 0.015**	0.816 ± 0.015	0.803 ± 0.014	0.674 ± 0.035
ORGANIZACAO	7206	**0.879 ± 0.009**	0.876 ± 0.011	0.871 ± 0.004	0.795 ± 0.018
PESSOA	4774	0.948 ± 0.006	0.950 ± 0.006	**0.950 ± 0.005**	0.929 ± 0.006
TEMPO	2515	0.949 ± 0.009	**0.954 ± 0.009**	0.950 ± 0.014	0.895 ± 0.010
LEGISLACAO	13764	0.957 ± 0.010	**0.957 ± 0.007**	0.953 ± 0.010	0.939 ± 0.015
Average micro		**0.904 ± 0.005**	0.903 ± 0.006	0.899 ± 0.005	0.840 ± 0.010
Average macro		**0.905 ± 0.005**	0.905 ± 0.006	0.900 ± 0.005	0.845 ± 0.010
WikiNER					
Entidades	**Dist Treino**	**BaseLine**	**Bangla**	**Vandit**	**Manual**
MISC	37619	0.850 ± 0.001	**0.853 ± 0.002**	0.849 ± 0.002	0.687 ± 0.008
ORG	25308	**0.882 ± 0.002**	**0.882 ± 0.002**	0.882 ± 0.003	0.798 ± 0.019
LOC	138184	**0.943 ± 0.000**	0.943 ± 0.001	0.941 ± 0.001	0.838 ± 0.005
PER	63343	**0.949 ± 0.001**	**0.949 ± 0.001**	0.948 ± 0.001	0.909 ± 0.003
Average micro		**0.929 ± 0.000**	**0.929 ± 0.000**	0.928 ± 0.001	0.833 ± 0.002
Average macro		**0.929 ± 0.000**	**0.929 ± 0.000**	0.927 ± 0.001	0.834 ± 0.002
Paramopama					
Entidades	**Dist Treino**	**BaseLine**	**Bangla**	**Vandit**	**Manual**
TEMPO	4320	**0.826 ± 0.005**	0.824 ± 0.004	0.822 ± 0.007	0.702 ± 0.009
LOCAL	9805	**0.916 ± 0.001**	0.915 ± 0.002	0.915 ± 0.002	0.861 ± 0.011
PESSOA	3867	0.918 ± 0.003	0.920 ± 0.004	**0.921 ± 0.002**	0.895 ± 0.007
Average micro		**0.899 ± 0.001**	0.898 ± 0.002	0.898 ± 0.002	0.833 ± 0.007
Average macro		**0.899 ± 0.001**	0.898 ± 0.002	0.898 ± 0.002	0.837 ± 0.007

personalizados sejam melhores que *BERT* pré-treinado sem uso de pesos. Como as entidades são correlacionadas, a adição de pesos manipula todas as entidades de forma com que classes majoritárias também possam sofrer ajustes em sua taxas de acerto. Experimentos futuros com outros modelos descritos na literatura, podem descrever se a adição de pesos é eficaz ou o modelo *BERT* pré-treinado consegue generalizar de maneira satisfatória as entidades em um conjunto NER.

Este estudo ainda aguarda publicação em um periódico ou conferência na área de PLN. Outros estudos relacionados a Modelos de Linguagem e NER foram investigados, porém não serão elencados neste trabalho.

A.0.3 Experimento 3 - Extração de entidade de documentos jurídicos portugueses usando supervisão distante

A maioria das abordagens para extração de entidades com função e preenchimento, também chamada de *"Role-filler Entity Extraction"* (REE), depende de grandes corpos de treinamento rotulados nos quais as menções de entidade são anotadas diretamente no documento de entrada. Uma pesquisa exploratória foi realizada em uma base de conhecimento igualmente denominada de *"Knowledge Base"* (KB), sobre entidades nomeadas para realizar REE em nível de documento de petições relacionadas a apreensão de drogas. O estudo propõe um sistema que aprende a extrair entidades de petições para preencher 29 papéis de um evento de apreensão de drogas. A base de conhecimento abrange mais de 170 mil entidades e seis mil petições, de forma que cada entidade no KB esteja vinculada a uma petição específica, as menções a uma entidade no texto de uma petição não são anotadas. Logo, a falta dessas anotações traz desafios relacionados a incompatibilidades entre valores de entidades previstos no KB e menções de entidades nos documentos. Neste contexto, um método de anotação distante é proposto para superar esses desafios e rotular automaticamente os documentos de petição usando o KB disponível. A Figura A.2, descreve uma petição anotada que descreve menções de entidades presentes na pesquisa abordada.

(...) no dia 23 de dezembro de 2019, por volta das 18h37min, na Rua 11, a denunciada ANA DA SILVA foi flagrada portando 01 porção de MACONHA (...)

Data da apreensão | Hora da apreensão | Local | Nome do Suspeito | Tipo de droga

Figura A.2: *Texto*: fragmento de petição de apreensão de drogas (anônima) com menções de entidades destacadas. *Rótulo*: legenda de cores em destaque para rótulos de entidade.

Proposta

O sistema proposto compreende duas fases: treinamento e predição. Na fase de treinamento o sistema aplica um um Método de Anotação Distante ou *"Distant Annotation Method"* (DAM) que utiliza o KB para criar um corpus rotulado e então treina um *BERT Sequence Labeling Model* (BSLM) para rotular um texto de petição com as entidades mencionadas. Na fase de predição dado um texto de petição, o *BSLM* é empregado para rotular o texto com possíveis menções de entidades e então um procedimento de padronização é aplicado para extrair entidades únicas. Em

seguida por meio de abordagens simples de padronização, as menções rotuladas são convertidas em entidades únicas, resultando em um modelo preenchido que pode ser inserido no KB ou comparado com as entidades já existentes para validação. Para resolver este problema, foram adotadas as métricas *Precision*, *Recall* e *F1*.

A primeira etapa do estudo consiste na criação da DAM, logo para cada entidade presente no KB tem-se a geração de um ou mais padrões a serem buscados no texto da petição. Cada padrão representa uma frase possível para a entidade, sendo uma possível menção à entidade. Um algoritmo de correspondência de *string* é então usado para procurar todas as correspondências de cada padrão no texto da petição. Três algoritmos de correspondência de strings são usados neste contexto: *Simple Match* (SM), *Regular Expression* (REGEX) e *Fuzzy Search* (FS). Para execução dos algoritmos uma função de similaridade (S), limites (L) para controle da distância máxima de edição entre o padrão fornecido, uma *substring* do texto e pesos (β) foram definidos para ajustar os métodos propostos.

Resultados Experimentais e Discussão

Para avaliar diferentes componentes do DAM as seguintes variações foram avaliadas: (i) SM: *Simple Match* para todos os papéis (sem heurística); (ii) FS: *Fuzzy Search* para todos os papéis com similaridade fixa (S = 0,75 foi ajustado para maximizar F1 no conjunto validação); (iii) SM+H: uso de Heurísticas para os respectivos papéis e SM para os demais; (iv) FS+H: uso de Heurísticas para os respectivos papéis e FS (com S = 0,75) para os demais; e (v) FSRS+H: construção de sistema completo com Heurísticas e FS com similaridade específica de função abordadas nesta pesquisa, onde (S é definido de acordo com β = 6, onde β é o peso da métrica *Recall*). A Tabela A.5, descreve as métricas de desempenho para os sistemas correspondentes no conjunto de desenvolvimento.

Tabela A.5: Desempenho no conjunto de desenvolvimento de sistemas obtidos por diferentes versões do DAM. Médias entre cinco execuções. Os desvios padrão estão entre parênteses.

Versão DAM	**Precision**	**Recall**	**F_1-Score**
SM	74.78 (1.52)	54.28 (0.45)	62.89 (0.35)
FS	74.10 (1.09)	59.36 (0.49)	65.91 (0.49)
SM+H	82.16 (1.19)	74.33 (0.98)	78.04 (0.35)
FS+H	**82.84 (0.45)**	73.92 (0.53)	78.13 (0.23)
FSRS+H	82.41 (0.43)	**78.59 (0.49)**	**80.45 (0.23)**

A versão DAM proposta nesta pesquisa (FSRS+H) apresentou a maior pontuação F1 por entidade de 80,45 no conjunto de teste com precisão

superior a 82%. Logo, com uma estratégia que controla o equilíbrio entre as métricas *Precision* e *Recall* a fim de otimizar o desempenho, é possível identificar menções de entidades e rotulá-las. A pergunta a ser respondida no futuro é se modelos mais complexos, como os baseados na arquitetura do *Transformer* (Vaswani et al., 2017), podem superar métodos mais simples de rotulagem.

Este estudo foi publicado no "*Processamento Computacional da Língua Portuguesa* (PROPOR 2022)", sob o título "*Entity extraction from Portuguese legal documents using distant supervision*" (Navarezi et al., 2022).

A.0.4 Experimento 4 - Predições a partir de análise de opiniões extraídas de Redes Sociais

O uso das redes sociais para avaliar as opiniões dos usuários em relação a marca de produtos, assuntos políticos, economia, programas de TV e pessoas públicas, é uma das áreas de interesse em aprendizado de máquina, pois obter informações sobre questões de utilidade pública, produtos ou serviços é uma tarefa difícil. Há custos financeiros, ferramentas de software e equipe de campo para realizar estudos detalhados sobre opiniões e sentimento do público em geral.

Um estudo foi elaborado para explorar essa relação entre as opiniões das pessoas e os eventos do mundo real. Foi utilizado a rede social Twitter (Murthy, 2018) para extrair informações de um *reality* show sobre possíveis candidatos eliminados durante sua programação. Para esta abordagem um *reality* show que usa a opinião do público para definir a eliminação de participantes, foi selecionado para avaliação. Pois, há um relação causal entre a opinião do público e eventos do mundo real.

Com o avanço do PLN e das técnicas de aprendizado de máquina o uso da análise de sentimento permite extrair informações de redes sociais, microblogs e sites. Com a rede social Twitter é possível formalizar um conjunto de dados com recursos simples, como a contagem de curtidas, retuítes, seguidores, *hashtags* específicas e contagens de análise de sentimento obtidas pelo modelo *BERT*, conforme ilustrado pela Figura A.3.

Proposta

Para este experimento três conjuntos de experimentos foram utilizados. O primeiro objetivo foi encontrar um modelo de análise de bons sentimentos, depois prever a rejeição do participante como um problema de regressão e predizer o candidato eliminado. Foi utilizado o processador de texto

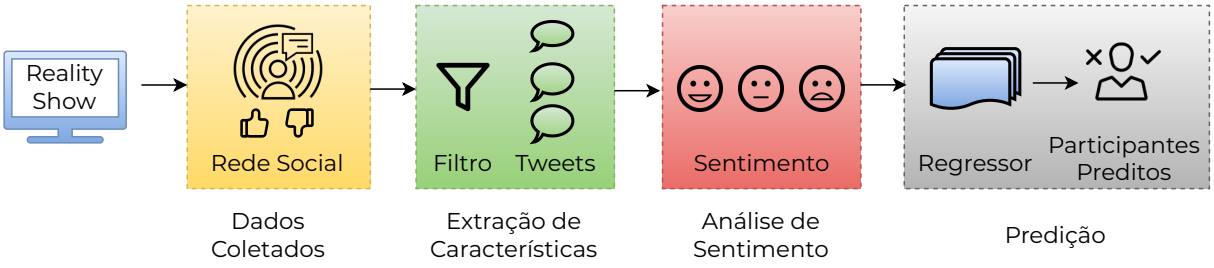


Figura A.3: Sequência de passos para predicação de participantes eliminados.

e o *tokenizer* implementado pelo vencedor do *SemEval-2017 Task 4* "Sentiment Analysis in Twitter" (Rosenthal et al., 2019), para pré-processar os *tweets*. Ao todo foram realizadas 17 coletas correspondendo a todas as 17 eliminações do *reality* show, totalizando aproximadamente 3,5 milhões de *tweets*. O modelo *BERT* pré-treinado foi usado para identificar *tweets* positivos, neutros e negativos dos participantes. Em seguida o número de exemplos com *tweets* positivos, neutros e negativos foi usado para melhorar as previsões de resultados reais. Logo, o candidato com a maior porcentagem de votos é eliminado do *reality* show, desta forma a porcentagem representa o valor alvo e o problema é tratado como um modelo de regressão.

Pré-processamento e Configuração Experimental

BERT pré-treinado no idioma português (Souza et al., 2020) foi comparado em uma avaliação experimental com *BERT* multilinguagem, NB (Gaussiana e Multinomial), *SVM* e *K-Nearest Neighbor* (KNN) (Guo et al., 2003), sobre o conjunto *TweetSentBR* (Brum and das Graças Volpe Nunes, 2018), com o objetivo de avaliar modelos de diferentes complexidades. Esses modelos usam diferentes representações de texto, como *BOW*, *TF-IDF*, *embeddings* em *Word2Vec* e *Fasttext*.

Resultados Experimentais e Discussão

A diferença nas métricas é estaticamente significativa de acordo com a análise de variância e são representadas na Tabela A.6. Todas as métricas foram obtidas usando a validação cruzada de 10 partições.

Para predizer possíveis candidatos os seguintes modelos de regressão foram utilizados: Linear, *Ridge, Lasso, Elastic Net, SVR (Support Vector Regressor), KNN (K-Nearest Neighbours), SGD (Stochastic Gradient Descent) e Random Forest*, descritos na documentação da biblioteca *Sklearn* (Pedregosa et al., 2011), juntamente com três *ensembles* combinados, *Ada*

Tabela A.6: Avaliação experimental sobre o conjunto *TweetSentBR*.

Modelos	***Accuracy*** mean ± std	***Precision*** mean ± std	***Recall*** mean ± std	***F1*** mean ± std
KNN (BOW)	0.574 ± 0.012	0.550 ± 0.016	0.528 ± 0.013	0.529 ± 0.014
MNB (BOW)	0.638 ± 0.013	0.610 ± 0.016	0.601 ± 0.014	0.597 ± 0.014
SVM (BOW)	0.647 ± 0.013	0.621 ± 0.015	0.614 ± 0.015	0.614 ± 0.015
KNN (TF-IDF)	0.611 ± 0.009	0.592 ± 0.012	0.562 ± 0.011	0.565 ± 0.013
MNB (TF-IDF)	0.634 ± 0.006	0.638 ± 0.012	0.571 ± 0.007	0.563 ± 0.009
SVM (TF-IDF)	0.647 ± 0.013	0.620 ± 0.015	0.611 ± 0.014	0.612 ± 0.140
KNN (avg-fasttext)	0.590 ± 0.010	0.571 ± 0.015	0.548 ± 0.012	0.531 ± 0.013
GNB (avg-fasttext)	0.552 ± 0.021	0.548 ± 0.021	0.548 ± 0.021	0.538 ± 0.021
SVM (avg-fasttext)	0.659 ± 0.006	0.635 ± 0.007	0.628 ± 0.006	0.628 ± 0.006
TextCNN (word2vec)	0.660 ± 0.009	0.633 ± 0.012	0.629 ± 0.011	0.630 ± 0.012
TextCNN (fasttext)	0.684 ± 0.009	0.659 ± 0.012	0.656 ± 0.009	0.656 ± 0.010
Bert-Multilingual-Cased	0.659 ± 0.011	0.635 ± 0.013	0.632 ± 0.014	0.632 ± 0.014
BERTimbau	**0.720 ± 0.017**	**0.698 ± 0.017**	**0.696 ± 0.018**	**0.697 ± 0.017**

Boost e Árvores de Decisão, Bagging e SVR e média das previsões de modelos individuais para gerar previsão final (conjunto de votação). Por fim, os hiperparâmetros foram definidos usando a ferramenta *grid-search* disponibilizada pela biblioteca *Sklearn*. Foram realizadas 10 repetições de 17 validações cruzadas e calculada a média dos resultados das métricas para estabilidade numérica, a métrica de avaliação utilizada foi o Erro Quadrático Médio (MSE) e Erro Médio Absoluto (MAE), conforme descrito na Tabela A.7. Os resultados indicam que a diferença nas métricas não é estatisticamente significativa de acordo com a análise de variância, no entanto os menores MAE e MSE foram obtidos usando *Random Forest*.

Tabela A.7: Avaliação experimental para predição de candidatos eliminados.

Modelo	**MSE (média,stddev)**	**MAE (média,stddev)**
Linear_Regression	(0.150 , 0.247)	(0.260 , 0.164)
Lasso	(0.090 , 0.061)	(0.257 , 0.097)
Elastic_Net	(0.086 , 0.051)	(0.252 , 0.086)
KNN	(0.083 , 0.060)	(0.233 , 0.100)
SGD	(0.079 , 0.053)	(0.236 , 0.084)
Ridge	(0.072 , 0.050)	(0.227 , 0.086)
Ensamble2 (Bagging, SVR)	(0.057 , 0.047)	(0.194 , 0.079)
Ensamble3 (Voting of SVR, KNN and Rigde)	(0.056 , 0.040)	(0.193 , 0.076)
SVR	(0.047 , 0.035)	(0.176 , 0.070)
Ensamble1 (Adaboost, DT)	(0.043 , 0.051)	(0.143 , 0.090)
Random_Forest	**(0.032 , 0.034)**	**(0.134 , 0.072)**

Por fim, o último experimento foi usar um procedimento de validação de 17 cruzamentos com uma amostragem de dados mais próxima do problema do mundo real. Para este procedimento cada eliminação foi enumerada de 1 a 17. Cada eliminação no *reality* show tem em média 3 participantes nomeados, logo esses participantes foram usados como exemplo no conjuntos de teste. Para uma eliminação N um treino no modelo de regressão foi realizado paras as outras $N-1$ eliminações. A Figura A.4 mostra quantas vezes o modelo previu uma eliminação N corretamente nas 10 execuções em média.

Os resultados indicaram que o uso da regressão teve um desempenho sa-

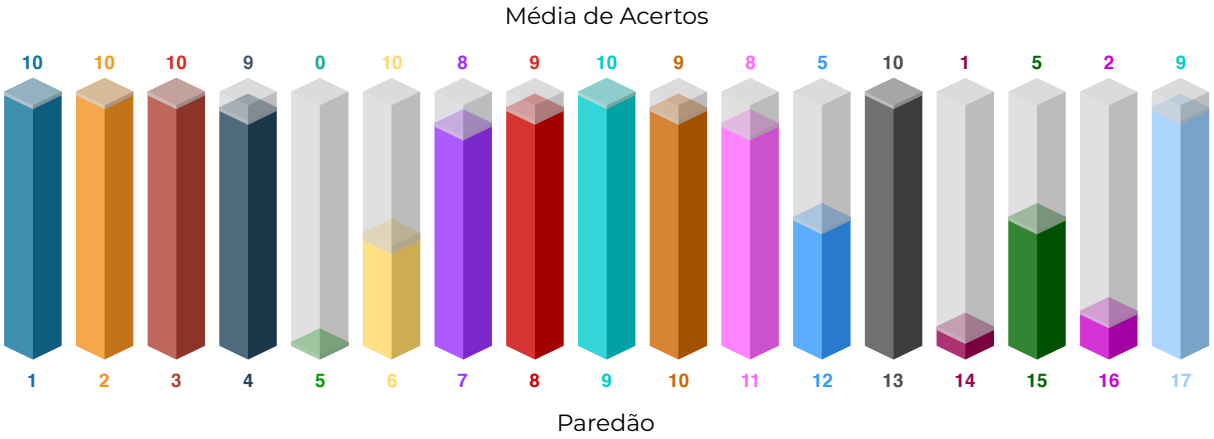


Figura A.4: Média de acertos do modelo por eliminação.

tisfatório na maioria das eliminações, no entanto abordagens utilizando redes sociais representam apenas uma fração do público total de um *reality* show. É possível usar o conjunto de dados produzido neste trabalho para treinar eliminações em versões posteriores do *reality* show ou domínios similares.

Informações adicionais sobre esta pesquisa podem ser encontradas no artigo publicado no "*Brazilian Conference on Intelligent Systems* (BRACIS 2020)", sob o título "*Can Twitter data estimate Reality Showoutcomes?*" (Sakiyama et al., 2020).

A.0.5 Experimento 5 - Identificação de temas políticos controversos utilizando dados de Redes Sociais

As redes sociais tornaram-se o principal palco de discussão sobre temas políticos. Em particular os processos eleitorais costumam trazer opiniões polarizadas sobre temas abordados pelos candidatos. Diante disso, uma melhor compreensão dos temas polêmicos de alto impacto e da opinião trazida por publicações em redes sociais, pode ajudar entender a dinâmica das discussões políticas para que a sociedade compreenda como elas afetam seu dia a dia. Neste sentido, um estudo investigativo foi elaborado para aplicar abordagens com modelagem de tópicos baseada em agrupamento, a fim de produzir informações sobre avaliação pública em temas políticos controversos. O estudo proposto enriquece as representações de texto combinando técnicas não supervisionadas de última geração (*HDBSCAN*) (McInnes et al., 2017) e técnicas supervisionadas por meio de um modelo de linguagem (BERTimbau) (Souza et al., 2020) para identificar temas políticos polêmicos, sobretudo no idioma português brasileiro em publicações de mídia social.

No Brasil uma pesquisa apresentada pelo Instituto Data Senado Senado (10) mostra que os principais canais de comunicação para busca de informações sobre política são: TV (37%), redes sociais (24%) e sites da Internet (23%). Portanto, o atual contexto brasileiro indica ser uma excelente oportunidade para analisar dados de mídias sociais sobre discussões políticas. Como mostrado na Figura A.5, usando as publicações das mídias sociais, seria possível entender melhor a discussão política usando um modelo de tópico para identificar e estudar os diferentes tópicos nos dados. Em seguida, usando a análise de sentimento, é possível estimar a avaliação do usuário sobre diferentes temas.

Proposta

Sobre este contexto, a abordagem da modelagem de tópicos nesta pesquisa, foi empregada usando publicações do Twitter sobre o cenário político no Brasil. Por fim, o objetivo foi identificar automaticamente tópicos potenciais de alto impacto relacionados à política, combinando duas tarefas de PLN: modelagem de tópicos e análise de sentimentos. A proposta é apresentada na Figura A.6.

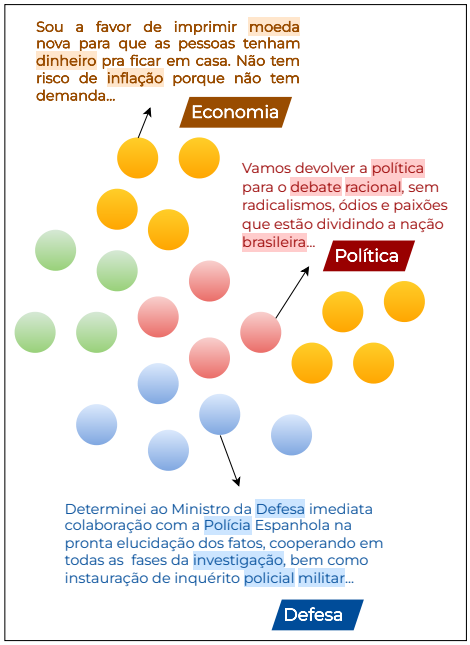


Figura A.5: Técnicas de Modelagem de Tópicos em PLN.

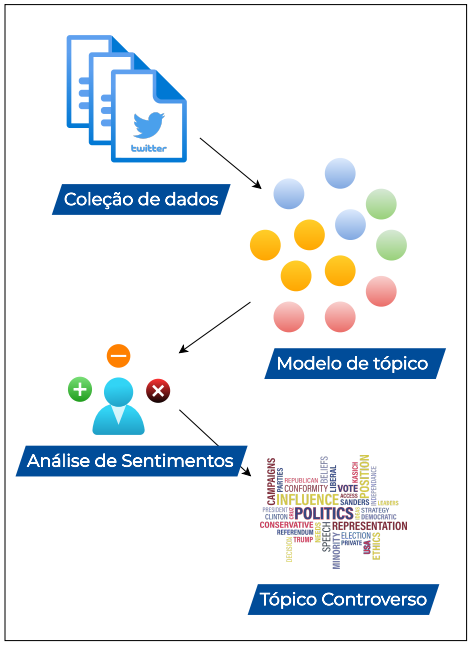


Figura A.6: Identificação do Tópico Controverso.

Pré-processamento e Configuração Experimental

Foram realizadas coletas semanais na rede social Twitter e foi possível identificar temas polêmicos para cada data analisada. Em seguida os

temas polêmicos foram relacionados com notícias do mundo real para validar as descobertas e comparar o método proposto com métodos tradicionais descritos na literatura. As coletas foram realizadas em horários e dias onde políticos discursaram em programas ao vivo ou *podcasts*. Dado o horário de início das transmissões ao vivo, as coletas de postagens do Twitter estiveram relacionadas até 3 horas após o início das transmissões. O estudo focou em transmissões realizadas em maio de 2022 e a coleta foi realizada através do aplicativo de raspagem de mídia social *Snscrape*[1].

Tabela A.8: Exemplos coletados para cada transmissão ao vivo de Maio/2022.

Data	***Tweets* coletados**	**Após o pré-processamento**
05 de Maio	9.243	7.987
12 de Maio	9.257	8.221
19 de Maio	8.827	6.883
27 de Maio	8.041	7.029

A Tabela A.8 descreve o número de *tweets* coletados para cada transmissão ao vivo. No total 30.120 *tweets* únicos foram analisados com uma média em torno de 20 *tokens* (separados por espaço) por *tweet*. A Figura A.7 ilustra uma visão geral da metodologia adotada nesta pesquisa, entre elas os principais componentes são: Coleta de Dados, Pré-processamento do *Tweet*, Calibração e *Clusterização*, e a Análise de *Cluster* realizada.

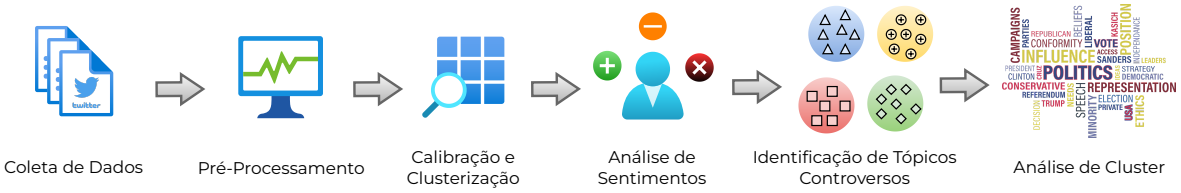


Figura A.7: Visão geral da identificação e análise do Tópico Controverso.

A detecção de tópicos controversos é feita combinando o agrupamento e a análise de sentimento dos *tweets* coletados. Para cada data de análise, *HDBSCAN* é usado para gerar os rótulos de *cluster*. Em seguida, o BERTimbau pré-treinado extrai os sentimentos de cada *tweet* nos *clusters*. Como próximo passo, os *clusters* são classificados pela porcentagem de exemplos positivos, neutros e negativos nos *clusters*. Por fim, cada *cluster* corresponde a um determinado assunto, consequentemente os temas potencialmente polêmicos geralmente estão localizados em *clusters* com grande quantidade de publicações negativas. Portanto, um *cluster* controverso é definido como sendo um *cluster* com uma porcentagem negativa acima de um limite C ($C \in \mathbb{R}$, $0 \leq C \leq 1$), onde C é um parâmetro confi-

[1] https://github.com/JustAnotherArchivist/snscrape

gurável. Observou-se que um limiar $C = 0,7$ foi suficiente para identificar *clusters* controversos nos dados coletados. Seguindo essa metodologia, ao identificar um *cluster* controverso, é possível quantificar o quão ruim o tópico foi recebido pelos usuários do Twitter, analisando o valor de C.

Resultados Experimentais e Discussão

Para avaliar quantitativamente as descobertas um cálculo sobre a medida de coerência C_V (Röder et al., 2015) é relatado para cada tópico controverso identificado. Com base na hipótese distributiva das palavras, a medida de coerência C_V visa quantificar o quanto o tema é destacado pelos documentos analisados. A avaliação do C_V é obtida extraindo do tópico controverso os seus 10 principais *tokens* com os maiores TF-IDFs do *cluster*.

Tabela A.9: *Clusters* controversos detectados e seus respectivos tamanhos (número de exemplos), porcentagem negativa e C_V.

Data	**Tamanho**	**Negativo(%)**	C_V
05 de Maio	588	72.2	0.429
12 de Maio	310	77.1	0.681
19 de Maio (i)	137	83.5	0.402
19 de Maio (ii)	62	80.6	0.666
27 de Maio	96	86.4	0.972

Na Tabela A.9 são apresentadas informações sobre os *clusters* controversos identificados. Há uma grande variação no número de exemplos dos *clusters* controversos e na métrica C_V entre as datas de análise. Além disso, 19 de maio foi a única data com mais de um *cluster* controverso identificado. Para visualizar os tópicos controversos identificados pelo HDBSCAN foi elaborado uma ilustração do tipo nuvens de palavras descrita na Figura A.8.

(a) 05 de Maio.

(b) 12 de Maio.

(c) 19 de Maio(i).

(d) 19 de Maio(ii).

(e) 27 de Maio.

Figura A.8: Visualizações da nuvem de palavras para os 50 *tokens* com o maior *cluster-TF-IDF* no *cluster TF-IDF*. Quanto maior o *token*, maior sua pontuação TF-IDF.

Para avaliar os resultados obtidos nesta pesquisa utilizando a mesma metodologia o algoritmo de agrupamento K-*Means* (Ahmed et al., 2020) e sem redução de dimensão foi usado para construir *clusters* controversos. Por meio de uma avaliação estatística ao comparar as duas abordagens diferentes para identificar *clusters* controversos os resultados foram favoráveis usando o método UMAP e HDBSCAN, disponíveis na Tabela A.10. No entanto por questões de publicação e outras análises ainda não fornecidas, alguns dados desta pesquisa serão suprimidos neste trabalho.

Este estudo foi aceito no "*Brazilian Conference on Intelligent Systems* (BRACIS 2023)", sob o título "*A Framework for Controversial Political Topics Identification using Twitter Data*".

Tabela A.10: Os 10 principais *tokens* e C_V, com base no *cluster* TF-IDF para cada data de coleta de dados. Os principais *tokens* HDBSCAN à esquerda e K-*Means* os principais *tokens* à direita.

	HDBSCAN		K-Means	
Data	**Top-10**	**Cv**	**Top-10**	**Cv**
05 de Maio	votar, eleição, urna, eleitoral, ser, voto, governar, fraudar, tse, eleito	0,429	votar, eleição, ser, eleitoral, urna, governar, país, campanha, ano, auditoria	0,307
12 de Maio	preço, imposto, impor, aumentar, redução, combustível, lucrar, governar, reduzir, mercar	0,681	votar, eleição, ser, governar, ciro, ter, dizer, pt, falar, ano,	0,328
19 de Maio	governar, político, corrupção, poder, presidência, corrupto, ano, ser, público, apoiar	0,402	governar, votar, ser, ter, ano, político, eleição, stf, dinheiro, país	0,327
27 de Maio	educação, cortar, universidade, bilião, orçamentar, governar, 3.2, federar, bloquear, mec	0,972	pesquisar, governar, votar, ser, ver, eleição, ter, perder, ir, ruir	0,271

A.0.6 Experimento 6 - Aprendizado profundo aplicado à fenotipagem de biomassa em forragens com imagens RGB baseadas em UAV

Proposta

O uso da visão computacional tem intensificado diversos estudos na área de monitoramento em várias regiões geográficas, por exemplo o controle da biomassa de forragens, fazendas de gado, focos de incêndios, entre outros. Uma das estratégias no monitoramento pode ser vista no trabalho de Castro et al. (2020), com a finalidade de obter fenotipagem não destrutivas e rápidas para produção de biomassa. Este experimento propôs a avaliação de métodos baseados em aprendizado profundo e imagens RGB baseadas em UAV, para estimar o valor da produção de biomassa por diferentes genótipos da espécie da grama forrageira *Panicum maximum Jacq* (Gomide and Gomide, 2000).

Os experimentos foram conduzidos no cerrado brasileiro com 110 genótipos com três repetições, totalizando 330 parcelas, conforme ilustra a Figura A.9.

Resultados Experimentais

Dois modelos de visão baseados em CNNs denominados *AlexNet* e *ResNet* foram avaliados e comparados ao *VGGNet*, já adotado em trabalhos anteriores com mesmo tema para outras espécies de gramíneas. As previsões retornadas pelos modelos alcançaram uma correlação de 0,88 e um erro absoluto médio de 12,98% usando *AlexNet*, conforme descrito na Tabela A.11.

O experimento indica quatro grupos de resultados. O primeiro grupo com os melhores resultados representa o modelo pré-treinado *AlexNet*, descri-

Figura A.9: Área de pesquisa - Embrapa (MS).

Tabela A.11: Resultados experimentais sobre produção de biomassa utilizando Redes CNNs: *AlexNet, ResNet e VGGNet*.

Experimento	Modelo	MAE	MSE (%)	Correlação (r)
1	AlexNet	837 ± 106	14.58± 2.52	0.84± 0.03
2	AlexNet h	880 ± 202	15.11± 3.24	0.83± 0.06
3	AlexNet hv	924 ± 143	15.48± 2.30	0.82± 0.05
4	ResNet18	1086 ± 219	17.70± 3.41	0.74± 0.06
5	ResNet18 h	1046 ± 107	19.01± 2.77	0.74± 0.06
6	ResNet18 hv	1031 ± 153	18.76± 4.28	0.75± 0.06
7	AlexNet Pré-treinado	759 ± 102	13.23± 2.23	0.87± 0.05
8	AlexNet Pré-treinado h	768 ± 123	13.54± 2.88	0.87± 0.03
9	**AlexNet Pré-treinado** hv	**730 ± 59**	**12.98± 2.18**	**0.88± 0.04**
10	ResNet18 Pré-treinado	1206 ± 233	19.46± 5.15	0.73± 0.04
11	ResNet18 Pré-treinado h	1205 ± 194	23.16± 4.80	0.71± 0.07
12	ResNet18 Pré-treinado hv	1012 ± 128	18.58± 2.34	0.77± 0.05
13	VGGNet11 Pré-treinado	825 ± 152	13.89± 3.09	0.84± 0.04

tos nos experimentos (7,8,9). O segundo grupo é também composto pelo modelo *AlexNet* sem pré-treinamento (1,2,3). O terceiro e o quarto grupos são os resultados obtidos pelo modelo *ResNet18*. O resultados expressos na coluna MAE possuem uma variação de erro em Kilogramas(kg), na faixa de 1.556,00 kg · ha $^{-1}$ a 15.333,00 kg · ha $^{-1}$, portanto MAE de 730 representa um variação de 730 kg · ha $^{-1}$. Os símbolos (h, v e hv) reportados nos resultados da Tabela A.11, indicam que os modelos foram aumentados com imagens viradas da esquerda para a direita ou vice-versa (h, v) e com imagens invertidas de cima para baixo (hv). Em geral, os resultados do *AlexNet* são melhores do que o *ResNet18* e *VGGNet* que foi considerado apenas como linha de partida para o estudo.

Discussão

Os resultados indicaram que o método *AlexNet* teve melhor desempenho. Uma possível explicação para isso é que o método *ResNet18* apesar de ser uma rede mais profunda que a *AlexNet* em sua implementação, não foi

capaz de representar adequadamente o problema com seus filtros convolucionais mesmo com o modelo pré-treinado. Logo, não foi capaz de modificar suas camadas com precisão suficiente. Os resultados sem etapas de pré-treinamento também não foram suficientes. Isso demonstra como a falta de dados para o treinamento impactou o desempenho. No entanto, a avaliação de diferentes etapas de pré-processamento (com e sem aumento de dados e pré-treinamento) resultou em implicações essenciais para a integração de medições agronômicas coletadas no campo com métodos robustos em imagens RGB de sensoriamento remoto.

Este estudo foi publicado na "MDPI-*Sensors*", sob o título "*Deep Learning Applied to Phenotyping of Biomass in Forages with UAV-Based RGB Imagery*" (Castro et al., 2020).

A.0.7 Experimento 7 - Redes neurais convolucionais para estimar o rendimento de matéria seca em um programa de criação de capim-guiné usando sensoriamento remoto UAV

Ainda com foco em sensoriamento remoto um segundo experimento foi realizado em parceria com a Embrapa. O destaque foi intensificar pesquisas relacionadas a matéria seca de forragem (Gomide and Gomide, 2000), pois é a principal fonte de nutrientes na dieta de animais ruminantes. Assim, esta característica é avaliada na maioria dos programas de melhoramento de forrageiras com o objetivo de aumentar a produtividade.

Proposta

O principal objetivo deste estudo foi propor uma abordagem em rede neural convolucional (CNN) usando imagens UAV-RGB para estimar características de produção de matéria seca em um programa de melhoramento de capim-guiné. Para isso foi utilizado um experimento composto por 330 parcelas conduzido na Embrapa Gado de Corte (Embrapa, 2020), com informações da mesma área de pesquisa do Experimento (1). O conjunto de dados de imagens foi composto por imagens obtidas com um sensor RGB embutido em um *drone* do tipo *Phantom* 4 PRO (Peppa et al., 2019). As variáveis de rendimento de matéria seca foliar ("*Leaf Dry Matter Yield*" - LDMY) e rendimento total de matéria seca ("*Total Dry Matter Yield*" - TDMY) foram obtidos por metodologia agronômica convencional e considerados como dados reais para pesquisa.

Pré-processamento e Configuração Experimental

Apesar de pesquisas anteriores já citadas na Seção A.0.6 - Experimento (1) apontarem uma maior eficiência do *AlexNet* em relação a redes neurais profundas, optou-se por avaliar CNNs mais avançadas em relação aos avanços da literatura e quantidade de parâmetros, para extração de características da matéria seca. Foram escolhidas as arquiteturas *ResNeXt50* (Xie et al., 2017b), *DarkNet53* proposto por Redmon and Farhadi (2018), além de outras duas CNNs propostas recentemente para tarefas similares denominadas *MaCNN* (Ma et al., 2019) e *LF-CNN* (Barbosa et al., 2020). Para uma melhor comparação entre as diferentes arquiteturas, a Tabela A.12 apresenta as arquiteturas utilizadas neste estudo com o número de camadas e parâmetros.

Tabela A.12: Comparação entre os diferentes modelos em termos de número de camadas e parâmetros.

Modelos	**Número de Camadas**	**Número de Parâmetros**
AlexNet	8	62 M
AlexNet Pré-treinado	8	62 M
MaCNN	5	1.1 M
LF-CNN	10	3.6 K
ResNeXt50	50	25 M
ResNext50 Pré-treinado	50	25 M
DarkNet53	53	42 M

Resultados Experimentais e Discussão

As Tabelas A.13 e A.14 apresentam o erro absoluto médio (MAE), a raiz quadrada do erro-médio (RMSE) e o coeficiente de correlação de *Pearson r* para cada arquitetura CNN em relação às características de matéria seca. Em relação as informações obtidas pelas tabelas, os valores de MAE variaram entre 204,39 (AlexNet pré-treinado) a 266,77 (LF-CNN) kg.ha^{-1} para a característica LDMY, e para TDMY a variação é 289,66 (*AlexNet* pré-treinado) a 366,93 (*LF-CNN*) kg.ha^{-1}. Assim como no Experimento (1) citado na Seção A.0.6, é importante mencionar que os valores predominantes para esta matéria seca possui uma variação entre 500 a 4000 kg.ha^{-1}. Logo, é possível analisar um melhor desempenho do *AlexNet* pré-treinado para característica LDMY, pois apresentou os menores valores para MAE e RMSE, com erro médio absoluto de 204,39 kg.ha^{-1}. Para característica TDMY, *AlexNet* pré-treinado apresentou o menor MAE em relação aos demais modelos, para as demais métricas, *ResNeXt50* pré-treinado apresentou os melhores resultados, com RMSE de 413,07 kg.ha^{-1}.

Tabela A.13: Resultados para LDMY.

Modelos	**Erro Absoluto Médio**	**Raiz Quadrada do Erro-médio**	**Correlação de *Pearson* (r)**
AlexNet	248.41 ± 47.58	340.70 ± 64.85	0.70 ± 0.09
AlexNet Pré-treinado	**204.39 ± 56.46**	**286.24 ± 80.39**	**0.79 ± 0.12**
MaCNN	240.74 ± 65.09	333.60 ± 86.93	0.71 ± 0.11
LF-CNN	266.57 ± 89.19	366.93 ± 110.33	0.62 ± 0.13
ResNeXt50	221.04 ± 54.44	319.98 ± 98.47	0.72 ± 0.12
ResNext50 Pré-treinado	231.66 ± 63.41	319.58 ± 87.06	0.73 ± 0.10
DarkNet53	217.30 ± 57.09	311.76 ± 76.68	0.76 ± 0.12

Tabela A.14: Resultados para TDMY.

Modelos	**Erro Absoluto Médio**	**Raiz Quadrada do Erro-médio**	**Correlação de *Pearson* (r)**
AlexNet	311.37 ± 88.58	441.31 ± 131.29	0.73 ± 0.17
AlexNet Pré-treinado	**289.66 ± 96.28**	419.95 ± 136.93	0.75 ± 0.20
MaCNN	345.11 ± 97.94	477.12 ± 136.34	0.68 ± 0.20
LF-CNN	364.44 ± 145.38	506.56 ± 176.24	0.60 ± 0.17
ResNeXt50	306.09 ± 137.15	449.07 ± 175.06	0.71 ± 0.17
ResNext50 Pré-treinado	294.73 ± 78.83	**413.07 ± 117.77**	**0.76 ± 0.24**
DarkNet53	291.12 ± 80.26	419.50 ± 131.87	0.75 ± 0.23

É possível concluir que o sensoriamento remoto com veículos aéreos não tripulados de baixo custo embarcados com sensores RGB de alta resolução, juntamente com redes neurais convolucionais, é uma técnica promissora a ser utilizada para estimar a produção de matéria seca no programa de melhoramento de capim-guiné. Além disso, a *ResNeXt50* com pré-treinamento apresenta os melhores resultados, pois é uma rede capaz de estimar com mais precisão os parâmetros genéticos. Em investigações futuras espera-se aumentar o conjunto de dados e sua variabilidade avaliando outros campos experimentais com outras características ambientais. Uma possível implementação com modelos multimodais para estimar a produção de matéria pode ser aplicada como ferramenta para aumentar a eficiência na seleção em programas de melhoramento de forrageiras.

Este estudo foi publicado na "MDPI-*Sensors*", sob o título "*Convolutional Neural Networks to Estimate Dry Matter Yield in a Guineagrass Breeding Program Using UAV Remote Sensing*" (de Oliveira et al., 2021).

Referências Bibliográficas

Adaloglou, N. (2020). Intuitive explanation of skip connections in deep learning. *https://theaisummer.com/*. 15, 106

Ahmed, M., Seraj, R., e Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295. 125

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65. 111

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., e Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, 7(12):16398–16421. 91

Amin, S. U., Muhammad, G., Abdul, W., Bencherif, M., e Alsulaiman, M. (2020). Multi-cnn feature fusion for efficient eeg classification. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, páginas 1–6. IEEE. 91

Ampatzidis, Y. e Partel, V. (2019). Uav-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. *Remote Sensing*, 11(4):410. 90

Ashrafi, I., Mohammad, M., Mauree, A. S., Nijhum, G. M. A., Karim, R., Mohammed, N., e Momen, S. (2020). Banner: a cost-sensitive contextualized model for bangla named entity recognition. *IEEE Access*, 8:58206–58226. 114

Asuncion, A. e Newman, D. (2007). Uci machine learning repository. xvi, 34, 35

Atrey, P. K., Hossain, M. A., El Saddik, A., e Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379. 8, 14, 31, 91

Ba, J. e Caruana, R. (2014). Do deep nets really need to be deep? *Advances in neural information processing systems*, 27. 25

Ba, J. L., Kiros, J. R., e Hinton, G. E. (2016). Layer normalization. 52

Bakkali, S., Ming, Z., Coustaty, M., e Rusiñol, M. (2020). Visual and textual deep feature fusion for document image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, páginas 562–563. 3

Baltrušaitis, T., Ahuja, C., e Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443. 2, 28, 58

Barbosa, A., Trevisan, R., Hovakimyan, N., e Martin, N. F. (2020). Modeling yield response to crop management using convolutional neural networks. *Computers and Electronics in Agriculture*, 170:105197. 129

BIOCOM, I. (2020). Laboratório de computação bioinspirada. http://www.biocom.icmc.usp.br/. (Accessed on 04/06/2020). 103

Bisong, E. (2019). Matplotlib and seaborn. In *Building machine learning and deep learning models on google cloud platform*, páginas 151–165. Springer. 73

Blasch, E., Liu, S., Liu, Z., e Zheng, Y. (2018). Deep learning measures of effectiveness. In *NAECON 2018-IEEE National Aerospace and Electronics Conference*, páginas 254–261. IEEE. 85

Bojanowski, P., Grave, E., Joulin, A., e Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. 111

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132. 34

Brum, H. e das Graças Volpe Nunes, M. (2018). Building a Sentiment Corpus of Tweets in Brazilian Portuguese. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., e Tokunaga, T., editors, *Proceedings of the Eleventh International Conference*

on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA). 119

Burgess, J. (2011). Youtube. In Meyer, L. H., editor, *Oxford Bibliographies Online*, páginas 1–1. Oxford University Press, United Kingdom. 13, 43

Burkov, A. (2019). *The hundred-page machine learning book*, volume 1. Andriy Burkov Canada. 32

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., e Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, páginas 11621–11631. xv, 29

Castro, W., Marcato Junior, J., Polidoro, C., Osco, L. P., Gonçalves, W., Rodrigues, L., Santos, M., Jank, L., Barrios, S., Valle, C., et al. (2020). Deep learning applied to phenotyping of biomass in forages with uav-based rgb imagery. *Sensors*, 20(17):4802. 90, 94, 99, 100, 126, 128

Chen, D., Mei, J.-P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., e Chen, C. (2021a). Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, páginas 7028–7036. 27

Chen, H., Li, H., e Wu, X. (2020). Research on feature extraction and multimodal fusion of video caption based on deep learning. In *Proceedings of the 2020 4th International Conference on Management Engineering, Software Engineering and Service Sciences*, páginas 73–76. 43

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., e Miao, Y. (2021b). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712. 57, 91, 101

Chen, Y., Guerschman, J., Shendryk, Y., Henry, D., e Harrison, M. T. (2021c). Estimating pasture biomass using sentinel-2 imagery and machine learning. *Remote Sensing*, 13(4):603. 91

Cho, J. H. e Hariharan, B. (2019). On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, páginas 4794–4802. 78

Chowdhary, K. e Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, páginas 603–649. 1

Clark, K., Luong, M.-T., Le, Q. V., e Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*. 73

Cohen, J. P., Morrison, P., e Dao, L. (2020). Covid-19 image data collection. *arXiv 2003.11597*. 62, 88

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., e Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, páginas 9268–9277. 114

Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q., e Fung, P. (2022). Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*. 24, 78

Daquan, Z., Hou, Q., Chen, Y., Feng, J., e Yan, S. (2007). Rethinking bottleneck structure for efficient mobile network design. arxiv 2020. *arXiv preprint arXiv:2007.02269*. 19

Dashtipour, K., Gogate, M., Cambria, E., e Hussain, A. (2021). A novel context-aware multimodal framework for persian sentiment analysis. *arXiv preprint arXiv:2103.02636*. 14

de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., e Bermejo, P. (2018). Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, páginas 313–323. Springer. 113

De Boer, P.-T., Kroese, D. P., Mannor, S., e Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67. 46, 113

de Oliveira, G. S., Marcato Junior, J., Polidoro, C., Osco, L. P., Siqueira, H., Rodrigues, L., Jank, L., Barrios, S., Valle, C., Simeão, R., et al. (2021). Convolutional neural networks to estimate dry matter yield in a guineagrass breeding program using uav remote sensing. *Sensors*, 21(12):3971. 90, 94, 99, 100, 130

de Santana Correia, A. e Colombini, E. L. (2022). Attention, please! a survey of neural attention models in deep learning. *Artificial Intelligence Review*, páginas 1–88. 57, 105

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., e Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, páginas 248–255. Ieee. 20

Devlin, J., Chang, M.-W., Lee, K., e Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 58

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, páginas 1–15. Springer. 8

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 58, 94

Dou, Q., Liu, Q., Heng, P. A., e Glocker, B. (2020). Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425. 23, 78

Ehatisham-Ul-Haq, M., Javed, A., Azam, M. A., Malik, H. M., Irtaza, A., Lee, I. H., e Mahmood, M. T. (2019). Robust human activity recognition using multimodal feature-level fusion. *IEEE Access*, 7:60736–60751. xv, 28, 29

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., e Burgard, W. (2015). Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, páginas 681–687. IEEE. 78, 91

Embrapa (2020). A embrapa - portal embrapa. https://www.embrapa.br/. (Acessado em 05/06/2020). 92, 109, 128

Face, H. (2022). bert-base-multilingual-cased hugging face. https://huggingface.co/bert-base-multilingual-cased. (Accessed on 08/31/2022). 110

Fan, J., Upadhye, S., e Worster, A. (2006). Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, 8(1):19–20. 34

Feng, J., Chen, J., Liu, L., Cao, X., Zhang, X., Jiao, L., e Yu, T. (2019). Cnn-based multilayer spatial–spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image

classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1299–1313. 91

Flach, P. A. (2003). The geometry of roc space: understanding machine learning metrics through roc isometrics. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, páginas 194–201. xv, 35

Gao, J., Li, P., Chen, Z., e Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864. 1, 2, 3, 8, 30, 78, 100, 105

Garcia, N. C., Bargal, S. A., Ablavsky, V., Morerio, P., Murino, V., e Sclaroff, S. (2019). Dmcl: Distillation multiple choice learning for multimodal action recognition. *arXiv preprint arXiv:1912.10982*. 24

Garcia, N. C., Bargal, S. A., Ablavsky, V., Morerio, P., Murino, V., e Sclaroff, S. (2021). Distillation multiple choice learning for multimodal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, páginas 2755–2764. 78

Gebremedhin, A., Badenhorst, P. E., Wang, J., Spangenberg, G. C., e Smith, K. F. (2019). Prospects for measurement of dry matter yield in forage breeding programs using sensor technologies. *Agronomy*, 9(2):65. 90

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, páginas 1440–1448. 38

Gomide, C. A. d. M. e Gomide, J. A. (1999). Análise de crescimento de cultivares de panicum maximum jacq. *Revista Brasileira de Zootecnia*, 28:675–680. 89

Gomide, C. A. M. e Gomide, J. A. (2000). Morfogênese de cultivares de panicum maximum jacq. *Revista Brasileira de Zootecnia*, 29(2):341–348. 126, 128

Gou, J., Yu, B., Maybank, S. J., e Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819. xv, 24, 25, 26, 27, 78, 83, 87

Goyal, A., Bochkovskiy, A., Deng, J., e Koltun, V. (2021). Non-deep networks. *arXiv preprint arXiv:2110.07641*. 20

Grosse, R. (2017). Lecture 15: Exploding and vanishing gradients. *University of Toronto Computer Science*. 16

Guo, G., Wang, H., Bell, D., Bi, Y., e Greer, K. (2003). Knn model-based approach in classification. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, páginas 986–996. Springer. 119

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., e Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, páginas 1–38. 17, 27

Guo, W., Wang, J., e Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394. xv, xvi, 3, 8, 30, 31, 56, 58, 78, 80, 91, 105

Hakala, K. e Pyysalo, S. (2019). Biomedical named entity recognition with multilingual bert. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, páginas 56–61. 112

Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? *Advances in neural information processing systems*, 31. 16

Hao, T., Wu, D., Wang, Q., e Sun, J.-S. (2017). Multi-view representation learning for multi-view action recognition. *Journal of Visual Communication and Image Representation*, 48:453–460. 91

Hatay, R. (2019). senet.pytorch/senet at master · moskomule/senet.pytorch · github. https://github.com/moskomule/senet.pytorch/tree/master/senet. (Acessado em 10/03/2021). 54

He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 770–778. xv, 15, 16, 17, 44, 57, 59, 60, 65, 72, 106

Heo, B., Lee, M., Yun, S., e Choi, J. Y. (2019). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, páginas 3779–3787. 26

Hinton, G., Vinyals, O., e Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*. 25, 26, 81, 84, 88

Hintze, J. L. e Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184. 68

Hou, Q., Zhou, D., e Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, páginas 13713–13722. 19

Hu, D. (2019). An introductory survey on attention mechanisms in nlp problems. In *Proceedings of SAI Intelligent Systems Conference*, páginas 432–448. Springer. 17

Hu, J., Shen, L., e Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 7132–7141. xvi, 18, 44, 53, 54, 72, 94, 101

Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., e Gori, P. (2020). Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, páginas 772–781. Springer. 23, 24, 25

Hu, P., Zhen, L., Peng, D., e Liu, P. (2019). Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, páginas 635–644. 37

Hu, X., Chu, L., Pei, J., Liu, W., e Bian, J. (2021). Model complexity of deep learning: A survey. *Knowledge and Information Systems*, 63:2585–2619. 80

Huang, G., Liu, Z., Van Der Maaten, L., e Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 4700–4708. 58, 65

Ioffe, S. e Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*. 50, 52, 95

Jank, L., Barrios, S. C., do Valle, C. B., Simeão, R. M., e Alves, G. F. (2014). The value of improved pastures to brazilian beef production. *Crop and Pasture Science*, 65(11):1132–1137. 89

Jank, L., Resende, R. M. S., Valle, C. d., Resende, M. d., Chiari, L., Cançado, L. J., e Simioni, C. (2008). Melhoramento genético de panicum maximum. *Melhoramento de forrageiras tropicais*, 1:55–87. 89

Ji, J., Ma, Y., Sun, X., Zhou, Y., Wu, Y., e Ji, R. (2022). Knowing what to learn: A metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31:4321–4335. 64

Jiang, H. e Learned-Miller, E. (2017). Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, páginas 650–657. IEEE. 37

Jiang, S., Qin, H., Zhang, B., e Zheng, J. (2020). Optimized loss functions for object detection and application on nighttime vehicle detection. *arXiv preprint arXiv:2011.05523*. 37

Jin, B. T., Abdelrahman, L., Chen, C. K., e Khanzada, A. (2020). Fusical: Multimodal fusion for video sentiment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, páginas 798–806. 43

Joyce, J. M. (2011). Kullback-leibler divergence. In *International encyclopedia of statistical science*, páginas 720–722. Springer. 26, 81

Júnior, C. M., Macedo, H., Bispo, T., Santos, F., Silva, N., e Barbosa, L. (2015). Paramopama: a brazilian-portuguese corpus for named entity recognition. *Encontro Nac. de Int. Artificial e Computacional*. 113

Kalpić, D., Hlupić, N., e Lovrić, M. (2011). *Student's t-Tests*, páginas 1559–1563. Springer Berlin Heidelberg, Berlin, Heidelberg. 98

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., e Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, páginas 1725–1732. 13

Khaleghi, B., Khamis, A., Karray, F. O., e Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44. 8

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., e Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624. 3, 67, 72, 105

Koonce, B. (2021a). Efficientnet. In *Convolutional neural networks with swift for tensorflow*, páginas 109–123. Springer. 19

Koonce, B. (2021b). Mobilenetv3. In *Convolutional Neural Networks with Swift for Tensorflow*, páginas 125–144. Springer. 94, 101

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. 20

Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105. 59

Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90. 94

Krogh, A. (2008). What are artificial neural networks? *Nature biotechnology*, 26(2):195–197. 2

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., e Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*. 72, 73

Lanqing, L. (2019). Implementação de voc map python. https://zhuanlan.zhihu.com/p/68806221. (Acessado em 12/12/2020). xvi, 37

Li, C.-H., Wu, S.-L., Liu, C.-L., e Lee, H.-y. (2018a). Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*. 24

Li, H. (2018). Exploring knowledge distillation of deep neural nets for efficient hardware solutions. *CS230 Report*. 81, 85

Li, K., Yu, L., Wang, S., e Heng, P.-A. (2020). Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, páginas 775–783. 25, 80

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., e Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*. 72

Li, M.-a., Han, J.-f., e Yang, J.-f. (2021). Automatic feature extraction and fusion recognition of motor imagery eeg using multilevel multiscale cnn. *Medical & Biological Engineering & Computing*, 59(10):2037–2050. 91

Li, Y., Yang, M., e Zhang, Z. (2018b). A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883. 91

LIA, U. F. (2019). Laboratório de inteligência artificial. http://lia.facom.ufms.br/. (Acessado em 15/03/2019). 103

Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., e Zhang, C. (2020). Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, páginas 1054–1064. 112

Liao, L., Ma, Y., He, X., Hong, R., e Chua, T.-s. (2018). Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, páginas 801–809. 24

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., e Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, páginas 740–755. Springer. 72

Linguateca (2020). Linguateca - harem. https://www.linguateca.pt/HAREM/. (Acessado em 05/05/2020). 113

Liu, F., Ren, X., Zhang, Z., Sun, X., e Zou, Y. (2020). Rethinking skip connection with layer normalization. In *Proceedings of the 28th international conference on computational linguistics*, páginas 3586–3598. xv, 16, 17, 73

Liu, H., Liu, F., Fan, X., e Huang, D. (2021a). Polarized self-attention: towards high-quality pixel-wise regression. *arXiv preprint arXiv:2107.00782*. 20

Liu, W., Zheng, W.-L., e Lu, B.-L. (2016). Emotion recognition using multimodal deep learning. In *International conference on neural information processing*, páginas 521–529. Springer. 78, 91

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., e Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. 73

Liu, Y., Wu, C., Tseng, S.-y., Lal, V., He, X., e Duan, N. (2021b). Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*. 24

Liyuan, W., Jing, Z., Jiacheng, Y., e Li, Z. (2021). Porn streamer recognition in live video based on multimodal knowledge distillation. *Chinese Journal of Electronics*, 30(6):1096–1102. 23, 24

Lobantsev, A., Gusarova, N., Vatian, A. S., Kapitonov, A. A., e Shalyto, A. A. (2020). Comparative assessment of text-image fusion models for medical diagnostics. - , (5 (108)):70–79. 3

Loshchilov, I. e Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. 47

Lutz, M. (2013). *Learning python: Powerful object-oriented programming*. "O'Reilly Media, Inc.". 18

Ma, J., Li, Y., Chen, Y., Du, K., Zheng, F., Zhang, L., e Sun, Z. (2019). Estimating above ground biomass of winter wheat at early growth stages using digital images and deep convolutional neural network. *European Journal of Agronomy*, 103:117–129. 129

Ma, Y., Ji, J., Sun, X., Zhou, Y., Wu, Y., Huang, F., e Ji, R. (2022a). Knowing what it is: Semantic-enhanced dual attention transformer. *IEEE Transactions on Multimedia*. 64

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., e Ji, R. (2022b). X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*. 64

Maes, W. H. e Steppe, K. (2019). Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. *Trends in plant science*, 24(2):152–164. 90

Maguolo, G. e Nanni, L. (2021). A critic evaluation of methods for covid-19 automatic detection from x-ray images. *Information Fusion*, 76:1–7. 62

Martínez, H. P. e Yannakakis, G. N. (2014). Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International conference on multimodal interaction*, páginas 34–41. 11

Masci, J., Bronstein, M. M., Bronstein, A. M., e Schmidhuber, J. (2013). Multimodal similarity-preserving hashing. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):824–830. 11

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, páginas 41–48. Citeseer. 110

McInnes, L., Healy, J., e Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205. 121

Meel, P. e Vishwakarma, D. K. (2021). Multi-modal fusion using fine-tuned self-attention and transfer learning for veracity analysis of web information. *arXiv preprint arXiv:2109.12547*. 58, 67, 72

Meng, Z., Li, J., Zhao, Y., e Gong, Y. (2019). Conditional teacher-student learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 6445–6449. IEEE. 26

Misra, D., Nalamada, T., Arasanipalai, A. U., e Hou, Q. (2021). Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, páginas 3139–3148. 21

Mogili, U. R. e Deepak, B. (2018). Review on application of drone systems in precision agriculture. *Procedia computer science*, 133:502–509. 90

Mohit, B. (2014). Named entity recognition. In *Natural language processing of semitic languages*, páginas 221–245. Springer. 112

Mouzannar, H., Rizk, Y., e Awad, M. (2018). Damage identification in social media posts using multimodal deep learning. In *ISCRAM*. 62, 63

Mullapudi, R. T., Chen, S., Zhang, K., Ramanan, D., e Fatahalian, K. (2019). Online model distillation for efficient video inference. In *Proceedings of the IEEE/CVF International conference on computer vision*, páginas 3573–3582. 23

Murthy, D. (2018). *Twitter*. Polity Press Cambridge, UK. 118

Nanni, L., Ghidoni, S., e Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71:158–172. 10

Narayanan, B., Saadeldin, M., Albert, P., McGuinness, K., e Mac Namee, B. (2021). Extracting pasture phenotype and biomass percentages using weakly supervised multi-target deep learning on a small dataset. *arXiv preprint arXiv:2101.03198*. 90

Navarezi, L. M., Sakiyama, K., Rodrigues, L. S., Robaldo, C. M., Lobato, G. R., Vilela, P. A., Matsubara, E. T., e Fernandes, E. R. (2022). Entity extraction from portuguese legal documents using distant supervision. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, páginas 166–176. Springer. 118

Negash, L., Kim, H.-Y., e Choi, H.-L. (2019). Emerging uav applications in agriculture. In *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*, páginas 254–257. IEEE. 90

Niu, Z., Zhong, G., e Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62. 17, 57, 65

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., e Tran, D. (2019). Measuring calibration in deep learning. In *CVPR workshops*, volume 2. 82

Nothman, J., Ringland, N., Radford, W., Murphy, T., e Curran, J. R. (2012). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. 113

Oramas, S., Barbieri, F., Nieto Caballero, O., e Serra, X. (2018). Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.* 78, 91

Pache, M. C. B., Sant'Ana, D. A., Rozales, J. V. A., de Moraes Weber, V. A., Junior, A. d. S. O., Garcia, V., Pistori, H., e Naka, M. H. (2022). Prediction of fingerling biomass with deep learning. *Ecological Informatics*, 71:101785. 90

Pan, B., Cai, H., Huang, D.-A., Lee, K.-H., Gaidon, A., Adeli, E., e Niebles, J. C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, páginas 10870–10879. 23

Pandeya, Y. R. e Lee, J. (2020). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, páginas 1–19. 43

Pang, L., Zhu, S., e Ngo, C.-W. (2015). Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia*, 17(11):2008–2020. 37

Park, W., Kim, D., Lu, Y., e Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, páginas 3967–3976. 23, 27

Passban, P., Wu, Y., Rezagholizadeh, M., e Liu, Q. (2021). Alp-kd: Attention-based layer projection for knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 35, páginas 13657–13665. 26

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., e Chintala, S. (2019a). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., e Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, páginas 8024–8035. Curran Associates, Inc. 50, 51

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019b). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32. 18, 64, 95

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., e Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 119

Pelka, O., Koitka, S., Rückert, J., Nensa, F., e Friedrich, C. M. (2018). Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, páginas 180–189. Springer. 63

Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., e Zhang, Z. (2019a). Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, páginas 5007–5016. 23

Peng, Y., Liao, M., Song, Y., Liu, Z., He, H., Deng, H., e Wang, Y. (2019b). Fb-cnn: Feature fusion-based bilinear cnn for classification of fruit fly image. *IEEE Access*, 8:3987–3995. 91

Pennington, J., Socher, R., e Manning, C. D. (2014). Glove: Global vectors for word representation. In *In EMNLP*. 28

Peppa, M., Hall, J., Goodyear, J., e Mills, J. (2019). Photogrammetric assessment and comparison of dji phantom 4 pro and phantom 4 rtk small unmanned aircraft systems. *ISPRS Geospatial Week 2019*. 128

Poria, S., Cambria, E., e Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-

level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, páginas 2539–2544. 11

Prechelt, L. (1998). Early stopping-but when? In *Neural Networks: Tricks of the trade*, páginas 55–69. Springer. 64, 66

Radoglou-Grammatikis, P., Sarigiannidis, P., Lagkas, T., e Moscholios, I. (2020). A compilation of uav applications for precision agriculture. *Computer Networks*, 172:107148. 90

Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., e Kawsar, F. (2018). Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–27. 78, 91

Ramachandram, D. e Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108. 1, 8, 10, 11, 12, 14, 15, 31, 47, 49, 56, 65, 78, 84, 91

Ramshaw, L. A. e Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, páginas 157–176. Springer. 113

Rao, J., Qian, T., Qi, S., Wu, Y., Liao, Q., e Wang, X. (2021). Student can also be a good teacher: Extracting knowledge from vision-and-language model for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, páginas 3383–3387. 24

Redmon, J. (2013–2016). Darknet: Open source neural networks in c. http://pjreddie.com/darknet/. 94

Redmon, J. e Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. 129

Ren, S., He, K., Girshick, R., e Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*. 38

Ridnik, T., Ben-Baruch, E., Noy, A., e Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*. 59

Röder, M., Both, A., e Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, páginas 399–408. 124

Rokhmana, C. A. (2015). The potential of uav-based remote sensing for supporting precision agriculture in indonesia. *Procedia Environmental Sciences*, 24:245–253. 90

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., e Bengio, Y. (2014). Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*. 81

Ronneberger, O., Fischer, P., e Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, páginas 234–241. Springer. 15

Rosebrock, A. (2016). Intersection over union (iou) - pyimagesearch. https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/. (Acessado em 12/12/2020). xvi, 37

Rosenthal, S., Farra, N., e Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, páginas 502–518. 110

Rosenthal, S., Farra, N., e Nakov, P. (2019). Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*. 119

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252. 18

Sakiyama, K., de Souza Rodrigues, L., e Matsubara, E. T. (2020). Can twitter data estimate reality show outcomes? In *Brazilian Conference on Intelligent Systems*, páginas 466–482. Springer. 121

Salimans, T. e Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29. 52

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., e Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 4510–4520. 19

Schneider, J. e Vlachos, M. (2023). A survey of deep learning: From activations to transformers. *arXiv preprint arXiv:2302.00722*. 16

Scikit-learn (2007). sklearn.model_selection.gridsearchcv — scikit-learn 1.2.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. (Accessed on 12/13/2022). 82

Sebe, N., Cohen, I., Garg, A., e Huang, T. S. (2005). *Machine learning in computer vision*, volume 29. Springer Science & Business Media. 1, 11

Senado, I. D. (10). Datasenado — portal institucional do senado federal. https://www12.senado.leg.br/institucional/datasenado/publicacaodatasenado?id=panorama-politico-2022. (Accessed on 06/10/2022). 122

Shelke, M. S., Deshmukh, P. R., e Shandilya, V. K. (2017). A review on imbalanced data handling using undersampling and oversampling technique. *International Journal of Recent Trends in Engineering and Research*, 3(4):444–449. 113

Simonyan, K. e Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 20

Singh, A. K., Ganapathysubramanian, B., Sarkar, S., e Singh, A. (2018). Deep learning for plant stress phenotyping: trends and future perspectives. *Trends in plant science*, 23(10):883–898. 90

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, páginas 464–472. IEEE. 47, 63

Smith, L. N. e Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, páginas 369–386. SPIE. 47, 64

Sousa, M. G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P. H., Fernandes, E. R., e Matsubara, E. T. (2019). Bert for stock market sentiment analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, páginas 1597–1601. IEEE. xvii, 112

Souza, F., Nogueira, R., e Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*. 112

Souza, F., Nogueira, R., e Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, páginas 403–417. Springer. 44, 114, 119, 121

S&P, D. J. I. (2019). S&p dow jones indices. https://www.spglobal.com/spdji/en/. (Acessado em 16/09/2019). 110

St, L., Wold, S., et al. (1989). Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272. 69, 95

Stewart, E. L., Wiesner-Hanks, T., Kaczmar, N., DeChant, C., Wu, H., Lipson, H., Nelson, R. J., e Gore, M. A. (2019). Quantitative phenotyping of northern leaf blight in uav images using deep learning. *Remote Sensing*, 11(19):2209. 90

Sun, J., Jiang, J., e Liu, Y. (2020). An introductory survey on attention mechanisms in computer vision problems. In *2020 6th International Conference on Big Data and Information Analytics (BigDIA)*, páginas 295–300. IEEE. 17

Sun, S., Cheng, Y., Gan, Z., e Liu, J. (2019). Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*. 23

Szegedy, C., Ioffe, S., Vanhoucke, V., e Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31. 72

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., e Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 2818–2826. 28

Takahashi, N., Gygli, M., e Van Gool, L. (2017). Aenet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia*, 20(3):513–524. 28

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., e Liu, C. (2018). A survey on deep transfer learning. In *International conference on artificial neural networks*, páginas 270–279. Springer. 39

Tan, M. e Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, páginas 10096–10106. PMLR. 60

Tang, J., Shivanna, R., Zhao, Z., Lin, D., Singh, A., Chi, E. H., e Jain, S. (2020). Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*. 83

Tang, J. e Wang, K. (2018). Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, páginas 2289–2298. 81

Targ, S., Almeida, D., e Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*. 59, 60

Tartaglione, E., Barbano, C. A., Berzovini, C., Calandri, M., e Grangetto, M. (2020). Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data. *International Journal of Environmental Research and Public Health*, 17(18):6933. 62, 63

TensorFlow (2020). Tensorboard. https://www.tensorflow.org/tensorboard?hl=pt-br. (Acessado em 10/12/2020). 51

Thoker, F. M. e Gall, J. (2019). Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, páginas 6–10. IEEE. 25

Tian, C., Xu, Y., Li, Z., Zuo, W., Fei, L., e Liu, H. (2020). Attention-guided cnn for image denoising. *Neural Networks*, 124:117–129. 17

Tian, H., Tao, Y., Pouyanfar, S., Chen, S.-C., e Shyu, M.-L. (2019). Multimodal deep representation learning for video classification. *World Wide Web*, 22(3):1325–1341. 3, 28

Tsouros, D. C., Bibi, S., e Sarigiannidis, P. G. (2019). A review on uav-based applications for precision agriculture. *Information*, 10(11):349. 90, 93

Tung, F. e Mori, G. (2019). Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, páginas 1365–1374. 23

Ulyanov, D., Vedaldi, A., e Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*. 52

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999. 110

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., e Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, páginas 5998–6008. xv, 16, 17, 43, 58, 118

Viljanen, N., Honkavaara, E., Näsi, R., Hakala, T., Niemeläinen, O., e Kaivosoja, J. (2018). A novel machine learning method for estimating biomass of grass swards using a photogrammetric canopy height model, images and vegetation indices captured by a drone. *Agriculture*, 8(5):70. 91

Walport, M. e Kiley, R. (2006). Open access, uk pubmed central and the wellcome trust. *Journal of the Royal Society of Medicine*, 99(9):438–439. 63

Wang, D., Cui, P., Ou, M., e Zhu, W. (2015a). Learning compact hash codes for multimodal representations using orthogonal deep structure. *IEEE Transactions on Multimedia*, 17(9):1404–1416. 11

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., e Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 3156–3164. 20

Wang, F. e Tax, D. M. (2016). Survey on the attention based rnn model and its applications in computer vision. *arXiv preprint arXiv:1601.06823*. 17

Wang, F., Yan, J., Meng, F., e Zhou, J. (2021a). Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*. 26

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., e Hu, Q. (2020a). Supplementary material for 'eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA*, páginas 13–19. 19

Wang, Q., Zhan, L., Thompson, P., e Zhou, J. (2020b). Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, páginas 1828–1838. 23, 24, 78, 88

Wang, W., Arora, R., Livescu, K., e Bilmes, J. (2015b). On deep multi-view representation learning. In *International conference on machine learning*, páginas 1083–1092. PMLR. 91

Wang, W., Tran, D., e Feiszli, M. (2020c). What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, páginas 12695–12705. 3, 105

Wang, X., Zhao, Y., e Pourpanah, F. (2020d). Recent advances in deep learning. 1

Wang, Y., Xu, X., Yu, W., Xu, R., Cao, Z., e Shen, H. T. (2021b). Combine early and late fusion together: A hybrid fusion framework for image-text matching. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, páginas 1–6. IEEE. 3

Wightman, R. (2019). Pytorch image models. https://github.com/rwightman/pytorch-image-models. 64

Williams, J., Comanescu, R., Radu, O., e Tian, L. (2018). Dnn multimodal fusion techniques for predicting video sentiment. In *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, páginas 64–72. xv, 8, 11, 12, 13, 15, 78

Wold, S., Esbensen, K., e Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52. 11

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*. 64

Woo, S., Park, J., Lee, J.-Y., e Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, páginas 3–19. 18, 60

Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., e Feris, R. (2021). Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, páginas 11307–11317. 63

Xiaoma (2022). External attention pytorch. https://github.com/xmu-xiaoma666/External-Attention-pytorch. (Accessed on 09/01/2022). 64

Xie, S., Girshick, R., Dollár, P., Tu, Z., e He, K. (2017a). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 1492–1500. 94

Xie, S., Girshick, R., Dollár, P., Tu, Z., e He, K. (2017b). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 1492–1500. 129

Xiong, J., Yu, D., Liu, S., Shu, L., Wang, X., e Liu, Z. (2021). A review of plant phenotypic image recognition technology based on deep learning. *Electronics*, 10(1):81. 90

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., e Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, páginas 2048–2057. PMLR. 17

Xu, K., Zhang, M., Jegelka, S., e Kawaguchi, K. (2021). Optimization of graph neural networks: Implicit acceleration by skip connections and more depth. In *International Conference on Machine Learning*, páginas 11592–11602. PMLR. 17, 73

Xue, Z., Gao, Z., Ren, S., e Zhao, H. (2022). The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*. 24, 78, 80

Xue, Z., Ren, S., Gao, Z., e Zhao, H. (2021). Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, páginas 854–863. 23, 24, 78, 88

Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., e Zhang, Q. (2022). Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, páginas 12319–12328. 27

Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., e Luo, J. (2017). Deep multimodal representation learning from temporal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 5447–5455. 8, 36

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., e Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*. 73

Yim, J., Joo, D., Bae, J., e Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 4133–4141. 26, 78

Ying, L., Qian Nan, Z., Fu Ping, W., Tuan Kiang, C., Keng Pang, L., Heng Chang, Z., Lu, C., Jun, L. G., e Nam, L. (2021). Adaptive weights learning in cnn feature fusion for crime scene investigation image classification. *Connection Science*, 33(3):719–734. 91

Yohanandan, S. (2020). map (mean average precision) might confuse you! https://www.xailient.com/post/map-mean-average-precision-might-confuse-you. (Acessado em 05/12/2020). 36

You, C., Chen, N., e Zou, Y. (2021). Mrd-net: Multi-modal residual knowledge distillation for spoken question answering. In *IJCAI*, páginas 3985–3991. 23, 24, 88

Yu, T., Li, X., Cai, Y., Sun, M., e Li, P. (2021). S^2-mlp-mlpv2: Improved spatial-shift mlp architecture for vision. *arXiv preprint arXiv:2108.01072*. 21

Yu, T., Li, X., Cai, Y., Sun, M., e Li, P. (2022). S^2-mlp: Spatial-shift mlp architecture for vision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, páginas 297–306. 21

Yu, Z., Cui, Y., Yu, J., Wang, M., Tao, D., e Tian, Q. (2020). Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, páginas 3743–3752. 8, 36

Yu, Z. e Shi, N. (2020). A multi-modal deep learning model for video thumbnail selection. *arXiv preprint arXiv:2101.00073*. 43

Zhang, Q., Yang, L. T., Chen, Z., e Li, P. (2018a). A survey on deep learning for big data. *Information Fusion*, 42:146–157. 1

Zhang, Q.-L. e Yang, Y.-B. (2021). Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 2235–2239. IEEE. 21

Zhang, S., Yao, L., Sun, A., e Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38. 90

Zhang, X., Zhou, X., Lin, M., e Sun, J. (2018b). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 6848–6856. xv, 13

Zhang, Y., Jin, R., e Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52. 111

Zhang, Z., Meng, X., Wang, Y., Jiang, X., Liu, Q., e Yang, Z. (2022). Unims: A unified framework for multimodal summarization with knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, páginas 11757–11764. 23, 24, 78, 80

Zheng, Q., Zhu, J., Li, Z., Tian, Z., e Li, C. (2023). Comprehensive multiview representation learning. *Information Fusion*, 89:198–209. 91

Zhou, Z., Zhuge, C., Guan, X., e Liu, W. (2020). Channel distillation: Channel-wise attention for knowledge distillation. *arXiv preprint arXiv:2006.01683*. 28

Zhu, H., Xie, C., Fei, Y., e Tao, H. (2021). Attention mechanisms in cnn-based single image super-resolution: A brief review and a new perspective. *Electronics*, 10(10):1187. 17