Classificação da Disponibilidade de Vagas de Estacionamento Usando Aprendizagem Profunda

Calebe Pereira Lemos

SERVIÇO DE PÓS-GRADUAÇÃO DA FACOM-UFMS

Data de Depósito:

Assinatura:_____

Classificação da Disponibilidade de Vagas de Estacionamento Usando Aprendizagem Profunda

Calebe Pereira Lemos

Orientador: *Prof. Dr. Wesley Nunes Gonçalves*

Dissertação apresentada ao Programa de Pós-Graduação em Computação da Faculdade de Ciência da Computação - UFMS como parte dos requisitos necessários à obtenção para do título de Mestre em Ciência da Computação.

UFMS - Campo Grande
Março/2024

Agradecimentos

Gostaria de agradecer primeiramente a Deus por Sua constante presença ao meu lado, por me conceder a oportunidade de vivenciar cada momento desta jornada. Agradeço à minha família e amigos por todo apoio, vocês foram fundamentais em todo o processo.

Um agradecimento especial ao meu orientador, Professor Dr. Wesley Gonçalves, que foi responsável por me proporcionar muita troca de conhecimento e inspirações nessa caminhada. Agradeço ao Professor Dr. José Marcato Junior, por todo apoio e oportunidades proporcionadas no laboratório de geomática. Aos meus colegas do laboratório, muito obrigado por sua colaboração e companheirismo ao longo dessa trajetória.

Durante este período, tiveram várias pessoas que foram essenciais para a conclusão deste trabalho. Agradeço ao Mário de Araújo Carvalho, Ronaldo Silva e Maurício de Souza pela parceria durante as disciplinas. Também sou grato ao Ederson Costa pela ajuda com o drone e na captura das imagens. Suas contribuições foram fundamentais nessa caminhada.

Gostaria de expressar minha profunda gratidão ao Professor Dr. Amaury Antônio de Castro Junior, cujo exemplo como profissional e como ser humano é verdadeiramente inspirador. Seu constante incentivo e apoio durante a graduação e o mestrado foram inestimáveis, e cada conversa que tivemos foi fundamental para o meu progresso até o momento atual. Agradeço também por sua generosidade ao disponibilizar seu tempo e equipamento para concluir a captura de imagens na UFMS. Sua contribuição foi crucial para o sucesso deste projeto.

E um agradecimento mais que especial às pessoas que constituem minha base, pois foram elas que me proporcionaram todas as oportunidades para me dedicar aos estudos, desde a infância até o momento presente. Cada detalhe e incentivo que recebi ao longo desse caminho fez-me acreditar que posso chegar cada vez mais longe. Aos meus pais, Eder Dias Lemos e Karine Pereira Cavalcante Lemos, dedico meu mais profundo agradecimento. Vocês são meu

maior exemplo de determinação e amor incondicional. Obrigado por cada ensinamento, por toda dedicação e por estarem sempre ao meu lado. Amo vocês!

Finalmente gostaria de agradecer à CAPES pela minha bolsa de mestrado e a Faculdade de Computação - UFMS, pelo suporte e estrutura disponibilizados para o desenvolvimento de minha formação.

Abstract

The emission from motor vehicles is one of the most significant for atmospheric pollution. In this context, knowing the availability of parking spaces plays an important role in reducing air pollution, as the search time is shorter. In addition, these systems can contribute to improving traffic efficiency, as they prevent drivers from circulating unnecessarily in search of a parking space. However, automating this task presents challenges, mainly related to image capture with different lighting, weather seasons and obstructed view. This work aims to evaluate recent deep learning methods for classifying available parking spaces from images. The results highlighted Res2Net, with accuracy greater than 99% in experiments with the public dataset (CNR-Park+EXT) and 100% for the constructed dataset (UFMS-Park).

Resumo

A emissão proveniente de veículos automotores é uma das mais consideráveis para poluição atmosférica. Neste contexto, conhecer a disponibilidade de vagas de estacionamento desempenha um papel importante para redução da poluição do ar, pois o tempo de busca é menor. Além disso, esses sistemas podem contribuir para a melhoria da eficiência do tráfego, pois evitam que os motoristas circulem sem necessidade em busca de uma vaga. Entretanto, a automatização dessa tarefa apresenta desafios, principalmente relacionados com a captura da imagem com diferentes iluminações, estações climáticas e visão obstruída. Este trabalho tem como objetivo avaliar métodos recentes de aprendizagem profunda para classificação de vagas de estacionamento disponíveis a partir de imagens. Os resultados mostraram destaque para o Res2Net, com acurácia superior a 99% nos experimentos com o dataset público (CNR-Park+EXT) e 100% para o dataset construído (UFMS-Park).

Sumário

Sumário xi
Lista de Figuras xiii
Lista de Tabelas 1

1 Introdução **3**
1.1 Motivação 3
1.2 Objetivos 4
1.3 Principais Resultados 4
1.4 Estrutura do Trabalho 5

2 Revisão de Literatura **7**

3 Materiais e Métodos **11**
3.1 Conjunto de dados 11
3.2 Métodos de Aprendizado Profundo 14
3.3 Métricas de Avaliação 18

4 Experimentos e Resultados **21**
4.1 Delineamento Experimental 21
4.2 Resultados e Discussão 22
4.2.1 Resultados Quantitativos 22
4.2.2 Resultados Qualitativos 25
4.2.3 Complexidade Computacional 27

5 Considerações Finais **33**
5.1 Resumo dos Objetivos e Principais Resultados 33
5.2 Trabalhos Futuros 33

Referências **37**

Lista de Figuras

3.1 Exemplo de imagens do conjunto de dados e seus respectivos rótulos. 12
3.2 Exemplo de imagens da UFMS. 13
3.3 Exemplo de imagens do dataset da UFMS. 13
3.4 Bloco da ResNet. 14
3.5 Bloco da ResNeXt com cardinalidade 32. 15
3.6 Bloco *Squeeze-and-Excitation* da SE-ResNeXt. 15
3.7 Imagem ilustrativa da Res2Net. 16
3.8 Imagem ilustrativa da ConvNeXt V2. 16
3.9 Imagem ilustrativa do *Vision Transformer*. 17
3.10 Imagem ilustrativa do *HiViT*. 18

4.1 Precisão dos modelos de acordo com a condição climática. 23
4.3 Exemplos de erros dos modelos. 26
4.4 Mapa de ativação das imagens erradas. 26
4.6 Mapa de ativação das imagens erradas pelo HiViT (UFMS-Park). . 27
4.7 FPS x Taxa de Acerto. 29
4.2 Precisão de cada modelo para cada clima. 30
4.5 Imagens erradas do conjunto de dados da UFMS. 31

Lista de Tabelas

2.1 Resultados e características dos trabalhos citados. 8
2.2 Resultados da revisão de (Martynova et al., 2024) em termos de F1-Score. 9

3.1 Difrenças entre CNR-Park e UFMS-Park. 14

4.1 Configurações para treinamento. 22
4.2 Resultados dos modelos para o dataset ALL. 23
4.3 Resultados dos modelos para o dataset UFMS-Park. 24
4.4 Tempo para médio para classificar um conjunto de imagens e FPS. 28
4.5 Tempo para médio para classificar imagens do segundo experimento e FPS. 28

CAPÍTULO

1

Introdução

Neste capítulo é apresentada uma breve introdução sobre o tema, trazendo uma visão geral do problema tratado e dos principais objetivos da pesquisa. O capítulo está organizado da seguinte maneira: na Seção 1.1 é apresentada a motivação sobre o tema de pesquisa; na Seção 1.2 é apresentado o objetivo do trabalho; e na Seção 1.4 é apresentada a estrutura da dissertação.

1.1 Motivação

A poluição atmosférica constitui-se em um dos mais graves problemas associados à qualidade de vida dos habitantes. De acordo com (Derísio, 2017), tal poluição é resultante das interações entre homem e o meio em que vive, pois essas interações produzem boa parte dos resíduos responsáveis pela poluição atmosférica. Segundo (Gomes, 2009), dentre as matérias consideradas poluentes atmosféricos, destacam-se os gases e material particulado proveniente de veículos automotores, indústrias e da incineração de resíduos sólidos. Para (Derísio, 2017), a fonte emissora mais significativa são os automóveis, pois o rápido e contínuo processo de urbanização observado no Brasil tem levado ao aumento da motorização individual, além do planejamento urbano adotado pela maioria das cidades, que incentiva o transporte motorizado individual.

Neste cenário, a preocupação com os efeitos gerados pelas emissões veiculares, torna-se maior. Com isso, sistemas inteligentes para gerenciamento de estacionamentos desempenham um papel importante para melhorar a vida na cidade em termos de redução da poluição do ar, emissão de gases e congestionamento de tráfego (Oliveira et al., 2020). Esse tipo de sistema visa oferecer

informações em tempo real sobre a disponibilidade de vagas com o objetivo de reduzir o tempo de busca de um motorista em grandes centros. Além disso, a sugestão de vagas disponíveis pode reduzir o tráfego nas regiões centrais, pois reduz a circulação de motoristas em busca de uma vaga.

Em particular, métodos de aprendizagem profunda estão sendo aplicados com objetivo de detectar carros em estacionamentos, como o trabalho de (Ding and Yang, 2019) ou classificar uma vaga em ocupada ou desocupada, como a proposta de (Kyu Park and Young Park, 2022), de forma rápida a partir de imagens capturadas por câmeras de segurança. No entanto, essa tarefa ainda é desafiadora. (Mahmud et al., 2020) apresentou trabalhos relacionados a classificação e concluiu que essa tarefa apresenta desafios a serem superados, como a dificuldade em variações climáticas, de luminosidade, escala reduzida e principalmente à generalização.

1.2 Objetivos

Diante disso, este trabalho tem como objetivo avaliar métodos de aprendizagem profunda para superar os desafios da classificação de vagas de estacionamento, como ocupada ou vazia, usando imagens.

Para alcançar o objetivo geral, alguns objetivos específicos foram definidos:

1. Revisar a literatura sobre classificação de vagas de estacionamento;
2. Construir um dataset com as características encontradas na Universidade Federal de Mato Grosso do Sul (UFMS).
3. Avaliar métodos recentes de convolução e *transformers*;
4. Avaliar o *tradeoff* entre o custo computacional e a acurácia.

1.3 Principais Resultados

Neste estudo, foram investigados métodos de aprendizagem profunda para a classificação de vagas de estacionamento como ocupadas ou vazias. Utilizamos o conjunto de dados público CNR-Park+EXT, fornecido por (Amato et al., 2016), e também criamos um novo conjunto de dados com imagens de estacionamentos da Universidade Federal de Mato Grosso do Sul (UFMS), denominado UFMS-Park. O propósito do UFMS-Park era capturar características específicas da UFMS.

Inicialmente, o treinamento e teste foram realizados no conjunto de dados CNR-Park+EXT, onde dividimos o conjunto de dados em quatro partes: uma contendo imagens sem distinção climática e três separadas de acordo com o

clima (ensolarado, nublado e chuvoso). O melhor desempenho foi alcançado utilizando a arquitetura Res2Net50 (Gao et al., 2021), resultando em uma acurácia superior a 99%. Ao avaliar o modelo para cada condição climática, observamos que em clima nublado o desempenho é superior, com precisão de 99.54%. Isso ocorre porque a iluminação constante em condições nubladas não dificulta a visualização do carro, contribuindo para uma melhor precisão.

Posteriormente, aplicamos métodos convolucionais e transformers recentes, como o ConvNeXt V2 (Sanghyun Woo and Xie, 2023) e HiViT (Zhang et al., 2023) para comparação com o modelo que obteve o melhor desempenho no conjunto público. Nessa fase, avaliamos o desempenho do ConvNeXt, HiViT e Res2Net aplicados ao UFMS-Park, e novamente o modelo Res2Net demonstrou ser o mais eficaz, alcançando uma acurácia de 100% sem erros em nenhum exemplo.

1.4 Estrutura do Trabalho

O Capítulo 2 apresenta uma revisão de literatura de trabalhos anteriores que detectaram vagas de estacionamento em imagens usando deep learning. O Capítulo 3 apresenta os materiais e métodos deste trabalho, incluindo o conjunto de dados, os métodos de aprendizado profundo e as métricas usadas. Os experimentos e resultados são apresentados no Capítulo 4 enquanto que as considerações finais são apresentadas no Capítulo 5.

CAPÍTULO

2

Revisão de Literatura

Métodos de aprendizagem profunda foram propostos para classificação de vagas de estacionamento em diferentes abordagens. (Ding and Yang, 2019) propuseram adicionar blocos no YOLO v3 para extrair características mais granulares. A rede modificada foi usada para classificação e os testes foram realizados com imagens do subconjunto PUCPR do PKLot, porém o modelo sofre influência quando há variações na iluminação e clima (ensolarado, nublado e chuvoso). O modelo alcançou uma acurácia de 93%.

(Amato et al., 2016) propuseram o mAlexNet, baseado na rede AlexNet (Krizhevsky et al., 2012). Os experimentos foram realizados no conjunto CNRPark, que possui imagens com variações de luminosidade, escala e clima. O modelo apresentou uma precisão de 90%. Da mesma forma, (Rahman et al., 2020) usou o mAlexNet para classificar vagas de estacionamento, mas empregou o mAlexNet com alteração no tamanho do *kernel* da primeira camada, o que trouxe uma acurácia de 99.12%. (Kyu Park and Young Park, 2022) propuseram um algoritmo também baseado na arquitetura AlexNet, adicionando 3 camadas a mais para evitar o aprendizado enviesado. O desempenho do algoritmo foi avaliado usando dois conjuntos, o PKLot e o CNRPark, apresentando uma precisão de 94%.

(Acharya et al., 2018) mostraram uma abordagem em que a rede VGGNet-F é treinada para extração de características e depois um SVM é usado para classificação. Os autores empregaram validação cruzada 5-*folds* nas imagens do PKLot durante a fase de teste. Os trabalhos de (Mora et al., 2018) e (Dhuri et al., 2021) usaram a VGG16 em sua forma original e (Zhang et al., 2019) em uma versão extendida, que obteve 99.2% de acurácia. Porém, (Dhuri et al., 2021) apresentaram uma solução mais completa, testando o modelo em ima-

gens de um sistema de vigilância ao vivo e fornecendo um sistema integrado para verificação da disponibilidade de vagas em tempo real. O modelo atingiu uma precisão de 95.68%.

(Oliveira et al., 2020) fizeram em sua proposta a separação das imagens de acordo com as condições climáticas. Uma vez que as imagens estão separadas por clima, modelos especialistas para classificação das vagas são gerados para cada um dos 3 tipos de clima (dia límpido, dia nublado e dia chuvoso) pela arquitetura ResNet50. Uma quarta ResNet50 foi treinada para classificar as vagas sem distinção de clima, com objetivo de avaliar se a separação das imagens por clima possui algum impacto significativo, mas o resultado mostrou um desempenho muito parecido, com cerca de 1% de diferença e 2% nas imagens de clima chuvoso, no geral 95% de acurácia.

A Tabela 2.1 apresenta a comparação dos resultados obtidos por cada proposta em termos de acurácia e mostra características consideradas em cada um, podemos observar que a classificação usando VGG16 obteve o melhor resultado. Já a Tabela 2.2 apresenta resultados obtidos de uma avaliação recente de métodos convolucionais e *transformers* em termos de F1-Score. O melhor desempenho foi alcançado pelo método EfficientNet-P, proposto por (Martynova et al., 2024), que aplica convolução, seguida por ativação *Swish* e bloqueio *Squeeze and excitation*. O modelo atingiu F1-Score de 96.83%.

Nesses trabalhos, redes convolucionais e *transformers* mais recentes, como (Sanghyun Woo and Xie, 2023) e (Zhang et al., 2023) ainda não foram avaliados.

Autor	**Métodos**	**Cor**	**Considera clima**	**Acurácia**
Amato et al. (2016)	mAlexNet	RGB	Sim	90%
Mora et al. (2018)	VGG16	RGB	Sim	90.59%
Acharya et al. (2018)	VGG16	RGB	Sim	96.7%
Ding and Yang (2019)	YOLO v3	RGB	Não	93%
Zhang et al. (2019)	VGG16	RGB	Sim	99.2%
Rahman et al. (2020)	CmAlexNet	RGB	Sim	99.12%
Oliveira et al. (2020)	ResNet50	RGB	Sim	95%
Dhuri et al. (2021)	VGG16	RGB	Sim	95.68%
Kyu Park and Young Park (2022)	AlexNet	RGB	Sim	94%

Tabela 2.1: Resultados e características dos trabalhos citados.

Métodos	**Cor**	**Considera clima**	**F1-Score**
ResNet50	RGB	Sim	93.80%
AlexNet	RGB	Sim	95.55%
VGG16	RGB	Sim	94.96%
EfficientNet-P	RGB	Sim	96.83%
ViT	RGB	Sim	91.93%

Tabela 2.2: Resultados da revisão de (Martynova et al., 2024) em termos de F1-Score.

CAPÍTULO

3

Materiais e Métodos

3.1 Conjunto de dados

Nesta seção, são apresentados os datasets utilizados nesse trabalho, as características e configurações de cada conjunto. As imagens foram adquiridas a partir do conjunto de dados CNRPark+EXT disponibilizado por (Amato et al., 2016) e imagens capturadas na Universidade Federal de Mato Grosso do Sul (UFMS), no câmpus de Campo Grande.

CNR-Park+EXT

O conjunto é composto por imagens capturada sob diversas condições climáticas, diferentes condições de iluminação, variações de posição e inclui oclusão parcial devido a obstáculos.

A Figura 3.1 apresenta alguns exemplos, com a Figura 3.1(a) ilustrando dias ensolarados, a 3.1(b) exemplificando dias nublados e 3.1(c) mostrando imagens em dias chuvosos. O conjunto é composto por aproximadamente 150.000 imagens capturadas em um estacionamento com capacidade para 164 vagas. As imagens possuem resolução de 150x150 pixels.

Foram criados 4 datasets, um com todas condições climáticas (ALL) e outros 3 separados por clima, nublado (OVERCAST), ensolarado (SUNNY) e chuvoso (RAINY). A Figura 3.1 apresenta três exemplos para as condições climáticas.

(a) Sunny

(b) Overcast

(c) Rainy

Figura 3.1: Exemplo de imagens do conjunto de dados e seus respectivos rótulos.

UFMS-Park

Esse conjunto é composto por imagens capturadas em diferentes dias e estacionamentos da Universidade Federal de Mato Grosso do Sul (UFMS). Abrange uma variedade de condições de iluminação, variações de posição e ângulo. Também inclui oclusão parcial decorrentes de obstáculos e características específicas, tais como variações de piso no estacionamento, presença abundante de árvores e pinturas no chão para marcar vagas reservadas.

As imagens foram capturadas usando um drone DJI Mini 3, que possui uma câmera de 12MP com resolução de 1920x1080 pixels. A Figura 3.2 apresenta alguns exemplos.

Para a construção do conjunto de dados, as imagens originais foram submetidas a um processo de recorte, delimitando cada vaga de estacionamento. O conjunto resultante compreende aproximadamente 821 imagens, capturadas em estacionamentos de diferentes blocos da Universidade Federal de Mato Grosso do Sul (UFMS). A Figura 3.3 apresenta exemplos de imagens após o recorte, abrangendo condições de iluminação distintas, como dias ensolarados e nublados, além de ilustrar situações de oclusão e demonstrar variações nos

ângulos de captura. As imagens foram redimensionadas para uma resolução de 681x796 pixels.

A partir dessas imagens, foi construído um dataset que foi aleatoriamente dividido em três subconjuntos distintos: treino, teste e validação. O conjunto de treinamento é composto por 70% das imagens, enquanto 10% foram alocadas para o conjunto de teste e 20% para o conjunto de validação.

Figura 3.2: Exemplo de imagens da UFMS.

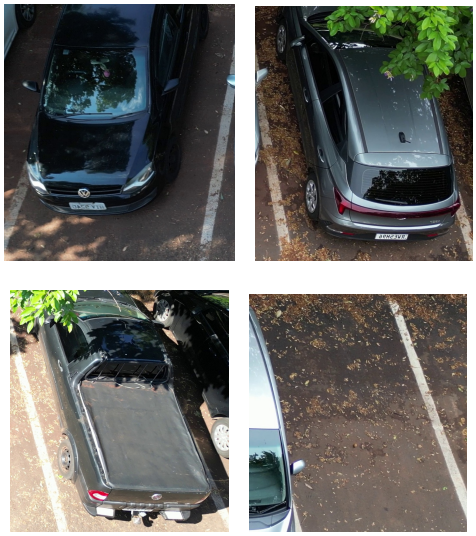


Figura 3.3: Exemplo de imagens do dataset da UFMS.

Para efeito de comparação com o conjunto de dados público (CNR-Park+EXT), a Tabela 3.1 foi criada, destacando as características de cada conjunto avaliado neste trabalho. As siglas representam:

- **DI:** Difrença de iluminação;
- **DE:** Dias ensolarado;

- **DN:** Dias nublado;
- **DC:** Dias chuvosos;
- **DME:** Difrença de marcações do estacionamento (as linhas que demarcam vagas);
- **MVR:** Marcações de vagas reservadas (pintura no chão);

DATASET	**DI**	**DE**	**DN**	**DC**	**DME**	**MVR**
CNR-Park+EXT	SIM	SIM	SIM	SIM	NÃO	NÃO
UFMS-Park	SIM	SIM	SIM	NÃO	SIM	SIM

Tabela 3.1: Difrenças entre CNR-Park e UFMS-Park.

3.2 Métodos de Aprendizado Profundo

Nos últimos anos, o aprendizado profundo obteve sucesso em uma variedade de aplicações. Os métodos têm sido propostos baseados em diferentes categorias de aprendizado, mostrando excelente desempenho quando comparado a abordagens tradicionais de aprendizado de máquina (Alom et al., 2019). Nesta seção, são apresentadas as arquiteturas avaliadas neste trabalho.

ResNet

A ResNet foi a primeira arquitetura que introduziu o bloco residual. Essa rede foi projetada para mitigar o problema de dissipação do gradiente, e possibilitar o treinamento de arquiteturas com maior profundidade. O bloco residual permite que saída de uma camada se propague como entrada da próxima e também é usada como entrada em uma camada mais profunda, ou seja, a informação é reinserida entre as etapas (He et al., 2015).

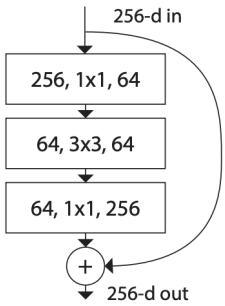


Figura 3.4: Bloco da ResNet.

ResNeXt

ResNeXt é uma arquitetura de rede simples e altamente modularizada para classificação de imagens. A rede implementa a ideia de ramificações em uma única célula. Em vez de realizar convoluções sobre o mapa de características completo, a entrada é projetada em uma série de canais, onde são aplicados filtros convolucionais separadamente. Essa ideia permite que grupos separados se concentrem em diferentes características da imagem de entrada. Os número de ramificações por bloco é chamado de cardinalidade (Xie et al., 2016).

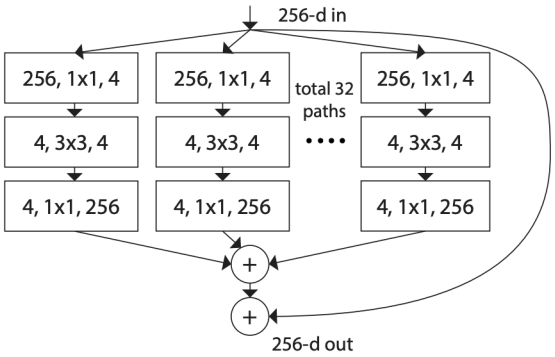


Figura 3.5: Bloco da ResNeXt com cardinalidade 32.

SE-ResNeXt

Na SE-ResNeXt uma nova unidade arquitetônica é proposta, chamada de bloco "*Squeeze-and-Excitation*"(SE). A proposta é espremer (*squeeze*) a informação global em um descritor de canal, usando o agrupamento de médias globais para gerar estatísticas por canal. E a excitação (*excitation*), tem objetivo de capturar as dependências do canal e aprender uma relação não linear e não mutuamente exclusiva entre os canais (Hu et al., 2017).

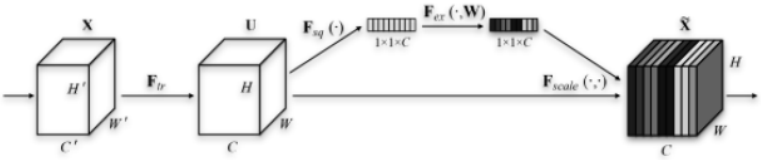


Figura 3.6: Bloco *Squeeze-and-Excitation* da SE-ResNeXt.

Res2Net

A Res2Net, ao invés de representar os recursos multiescala de maneira em camadas, constrói conexões hierárquicas semelhantes a resíduos dentro de um único bloco residual. Ele representa recursos multiescala em nível

granular e aumenta o alcance de campos receptivos para cada camada de rede.

Para atingir o objetivo, os filtros 3x3 de n canais foram substituídos por um conjunto de filtros menores, cada um com w canais. Esses grupos de filtros menores são conectados em um estilo hierárquico residual. Especificamente, os mapas de recursos de entrada foram divididos em grupos, um grupo extrai as características de um grupo de entrada e a saída é enviada para o próximo grupo. Finalmente, mapas de características de todos os grupos são concatenados e passam por um filtro de convolução 1x1 para manter o tamanho do canal deste bloco residual (Gao et al., 2021).

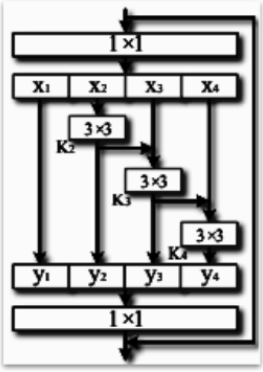


Figura 3.7: Imagem ilustrativa da Res2Net.

ConvNeXt V2

A ConvNeXt V2 (Sanghyun Woo and Xie, 2023) propõe uma estrutura de autoencoder mascarado totalmente convolucional, que é usado para aprendizado auto-supervisionado e uma nova camada de normalização de resposta global (*Global Response Normalization* - GRN) que pode ser adicionada à arquitetura ConvNeXt para aumentar a competição de recursos entre canais.

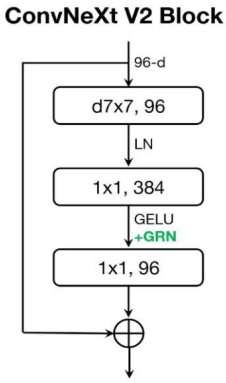


Figura 3.8: Imagem ilustrativa da ConvNeXt V2.

Vision Transformer

Transformer, aplicado pela primeira vez ao campo de processamento de linguagem natural, é um tipo de rede neural profunda baseada principalmente no mecanismo de auto-atenção. A auto-atenção é usada para fazer conexões entre locais distantes da imagem, ou seja, incorporar informações globais em toda a imagem. O *Vision Transformer* divide a imagem em uma série de patches de tamanho fixo, chamado de tokens. Os tokens são colocados em sequência e incluem a incorporação posicional, cada um deles é organizado em um sequência linear e multiplicado pela matriz de incorporação. O resultado, com a incorporação posicional, é usado como entrada para o codificador do transformador (Han et al., 2022).

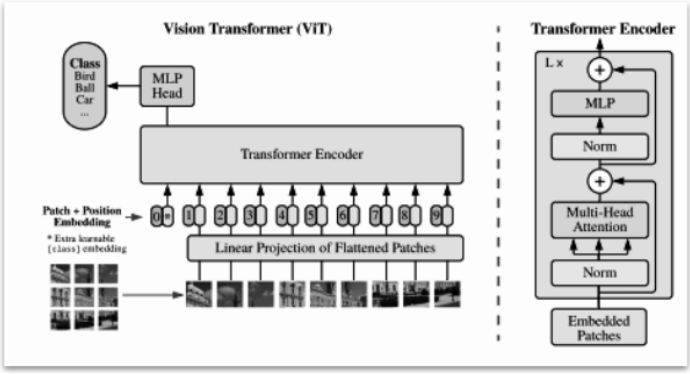


Figura 3.9: Imagem ilustrativa do *Vision Transformer*.

O codificador do transformador é crucial para extrair características, capturar relações espaciais entre os patches da imagem e incorporar informações de posição. Essas funções são essenciais para que o modelo compreenda as relações entre diferentes partes da imagem e a disposição relativa dos patches, permitindo que o modelo decodifique não apenas as características visuais individuais, mas também a estrutura global e o contexto da imagem.

HiViT

(Zhang et al., 2023) apresenta um novo projeto de transformadores de visão hierárquica denominado HiViT (abreviação de *Hierarchical* ViT) que apresenta alta eficiência e bom desempenho em modelagem de imagens mascaradas (MIM - Masked Image Modeling). A chave é a remoção de "operações locais entre unidades"desnecessárias, resultando em transformadores de visão hierárquicos estruturalmente simples nos quais as unidades de máscara podem ser serializadas como transformadores de visão simples.

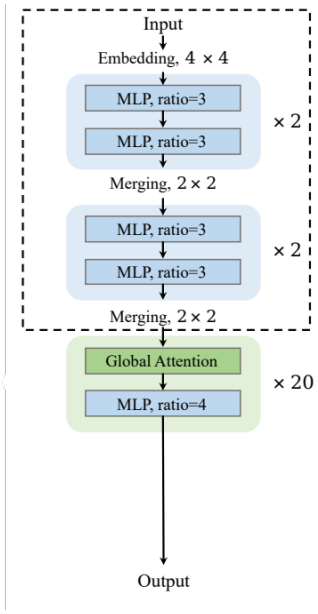


Figura 3.10: Imagem ilustrativa do *HiViT*.

3.3 Métricas de Avaliação

Para avaliar o desempenho dos modelos, utilizou-se o cálculo da acurácia, precisão, revocação e F1-Score. A acurácia (Equação 3.1) é calculada pela razão entre a quantidade de vagas ocupadas e vazias classificadas corretamente (TP + TN) e o número real de vagas ocupadas e vazias (P + N). Já a precisão (Equação 3.2), é dada pela razão entre as vagas ocupadas devidamente classificadas (TP) e a quantidade de exemplos preditos como ocupado (TP + FP). A revocação (Equação 3.3) é a razão entre exemplos de vagas ocupadas classificados corretamente (TP) e a quantidade de exemplos com vagas realmente ocupadas (P). Por último, F1-Score (Equação 3.4) é calculada pela razão entre a multiplicação da precisão, revocação por dois (2 x P x R) e a soma da precisão e revocação (P + R). A terminologia básica usada é dada por:

- **Condição Positiva (P):** Número de casos reais positivos nos dados;
- **Condição Negativa (N):** Número de casos reais negativos nos dados;
- **Verdadeiro Positivo (TP):** Condição positiva detectada como positiva;
- **Verdadeiro Negativo (TN):** Condição negativa detectada como negativa;

$$A = \frac{TP + TN}{P + N} \tag{3.1}$$

$$Precisão = \frac{TP}{TP + FP} \tag{3.2}$$

$$Revocação = \frac{TP}{P} \tag{3.3}$$

$$F1 - Score = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação} \tag{3.4}$$

A acurácia é a métrica usada para avaliar a quantidade de acertos. A precisão avalia quão preciso o modelo foi para classificar as condições positivas. A revocação (do inglês, *Recall*), avalia a capacidade do modelo detectar com sucesso resultados classificados como positivos. Já o F1-Score, é uma média calculada com base na precisão e revocação (Mariano, 2021).

CAPÍTULO 4

Experimentos e Resultados

4.1 Delineamento Experimental

Para os experimentos, um dataset contendo todas as imagens foi criado usando os dados disponibilizadas por (Amato et al., 2016), esse conjunto foi nomeado como ALL. Em seguida, outros três datasets foram criados para realização dos testes, os novos conjuntos separaram as imagens de acordo com o clima (chuvoso, nublado ou ensolarado). Com essa configuração, conseguimos avaliar o desempenho dos modelos de acordo com cada clima e suas respectivas variações de aparência, como iluminação. Além disso, um novo conjunto de dados foi construído utilizando imagens capturadas na Universidade Federal de Mato Grosso do Sul (UFMS), câmpus de Campo Grande, permitindo assim a avaliação de desempenho dos modelos em características específicas da instituição (UFMS-Park). Os conjuntos ALL e UFMS-Park, foram posteriormente empregado no treinamento dos modelos e nesta etapa, as imagens foram redimencionadas para resolução de 224 x 224 pixels.

Os primeiros métodos avaliados compreenderam técnicas de aprendizado residual, especificamente as arquiteturas ResNet50, ResNeXt50, SE-ResNeXt50 e Res2Net50. Todos esses modelos foram treinados por um máximo de 30 épocas, utilizando o otimizador SGD (Ruder, 2016), com uma taxa de aprendizado estabelecida em 0,1 e um tamanho de lote configurado como 64.

Posteriormente, foram examinadas abordagens mais recentes de convolução e *transformers*. O modelo ConvNeXt V2 (Sanghyun Woo and Xie, 2023), foi submetido a um treinamento de até 100 épocas, empregando o otimizador AdamW com uma taxa de aprendizado de 0,0025 e mantendo o tamanho do

lote em 64. O *Vision Transformer* (Han et al., 2022) e o HiViT (Zhang et al., 2023) foram treinados com o mesmo otimizador e taxas de aprendizado 0,003 para o *Vision Transformer* e 0,001 para o HiViT. Os tamanhos de lote foram configurados como 16 e 64, respectivamente, e o treinamento foi conduzido por até 30 épocas para o *Vision Transformer* e 300 épocas para o HiViT. Os parâmetros foram usados conforme as sugestões dos autores. A Tabela 4.1 apresenta as configurações citadas.

As redes foram implementadas utilizando o *MMClassification* (Contributors, 2020) em um sistema com a seguinte configuração: processador Intel(R) Xeon(R) E3-1270, 32GB de RAM e GPU NVIDIA GeForce GTX TITAN X.

Métodos	**Épocas**	**Otimizador**	**Taxa de aprendizado**	**Tamanho de lote**
ResNet	30	SGD	0.1	64
Res2Net	30	SGD	0.1	64
ResNeXt	30	SGD	0.1	64
SE-ResNeXt	30	SGD	0.1	64
Vision Transformer	30	AdamW	0.003	64
ConvNeXt V2	100	AdamW	0.0025	64
HiViT	300	AdamW	0.001	64

Tabela 4.1: Configurações para treinamento.

4.2 Resultados e Discussão

Nesta seção, são apresentados os resultados da avaliação experimental dos métodos usados para classificação de vagas de estacionamento em termos de acurácia, precisão, revocação e F1 (Seção 3.3), bem como uma análise visual dos resultados.

4.2.1 Resultados Quantitativos

Para realizar a análise quantitativa, utilizamos as métricas de acurácia, precisão, revocação e F1-Score com objetivo de medir o desempenho dos modelos. Os resultados quantitativos são apresentados a seguir, com base nos conjuntos de dados específicos, primeiramente para o CNR-Park e, em seguida, para o UFMS-Park.

CNR-Park

Inicialmente foram avaliados os resultados dos testes realizados com o dataset ALL, ou seja, o que contém todas as imagens. A Tabela 4.2 mostra que todos os modelos atingiram bons resultados, com taxa de acerto superior a

98%. (Amato et al., 2016) usando o mesmo conjunto de imagens e a rede mAlexNet, alcançou um resultado inferior, com taxa de acerto de 90%. O melhor desempenho foi obtido pela Res2Net50 com acurácia superior a 99%, alcançando o mesmo desempenho do modelo proposto por (Zhang et al., 2019), que usou imagens panorâmicas e a VGG16 para classificação.

Métodos	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
ResNet	98,96%	98,92%	98,98%	98,95%
Res2Net	99,09%	99,04%	99,13%	99,08%
ResNeXt	99,03%	99,05%	98,99%	99,02%
SE-ResNeXt	98,82%	98,90%	98,73%	98,81%
Vision Transformer	98,38%	98,40%	98,32%	98,36%

Tabela 4.2: Resultados dos modelos para o dataset ALL.

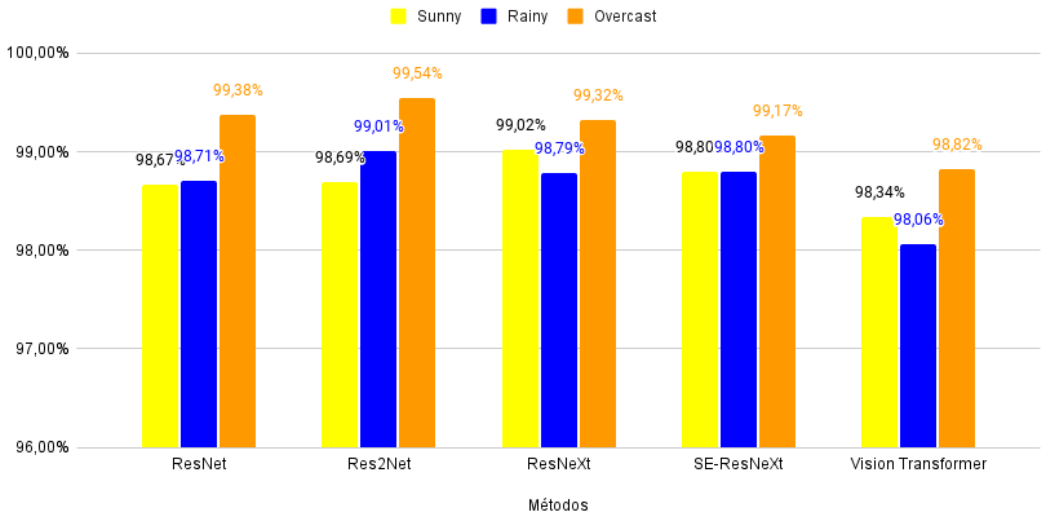


Figura 4.1: Precisão dos modelos de acordo com a condição climática.

Em seguida, os testes foram realizados em datasets específicos para cada clima, neste contexto conseguimos avaliar em quais condições os modelos possuem maior dificuldade para classificação. A Figura 4.1 mostra que a maior precisão é atingida em dias nublado (Overcast). Dessa forma, podemos observar que em dias nublados, a iluminação é constante e não dificulta a visualização do carro. Em dias chuvosos, a visualização é prejudicada caso a chuva seja intensa. Da mesma forma, a visualização dos carros é dificultada em dias ensolarados, pois a imagem pode apresentar luz excessiva com partes saturadas e sem detalhes.

O mesmo cenário é encontrado no trabalho de (Oliveira et al., 2020) cuja a proposta apresentada foi combinar duas redes, a DenseNet (Huang et al.,

2016) e MobileNet (Howard et al., 2017) para separar as imagens de acordo com as codições climáticas e depois, usar a ResNet50 (He et al., 2015) para classificar. Apesar disso, eles atingiram uma precisão de 96% para classificação em dias nublado. Logo, nossos modelos alcançaram melhores desempenho, atingimos alta precisão em dias nublados (Overcast), todos acima de 99% para redes residuais e um destaque para o *Vision Transformer*, que é uma proposta nova e obteve precisão de 98.82%.

Na Figura 4.2 podemos observar em quais condições climáticas cada modelo obtém melhor precisão. Em dias ensolarados (Sunny), o ResNeXt se destacou, alcançando uma precisão de 99.02%. Para dias chuvosos e nublados, o Res2Net foi o melhor modelo, atingindo precisão de 99.01% e 99.54% respectivamente.

UFMS-Park

Após a conclusão dos testes realizados no conjunto de dados público, procedemos à avaliação dos modelos utilizando o conjunto de dados constituído pelas imagens da Universidade Federal de Mato Grosso do Sul (UFMS). Ao analisar os resultados, identificamos que o melhor desempenho foi alcançado pelo modelo Res2Net50 (Gao et al., 2021). Este modelo, quando aplicado ao novo dataset, exibiu uma excelente performance, atingindo uma acurácia de 100%.

Dado o escopo do estudo, onde um dos objetivos era a avaliação de métodos de convolução e *transformers* recentes, testamos os métodos ConvNeXt V2 (Sanghyun Woo and Xie, 2023) e HiViT (Zhang et al., 2023) para fins comparativos. Os resultados desses testes revelaram que o método ConvNeXt V2 alcançou uma acurácia de 95.88%, enquanto o método HiViT atingiu uma acurácia ligeiramente inferior, registrando 94.85%, conforme ilustrado na Tabela 4.3.

Métodos	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
Res2Net	100,00%	100,00%	100,00%	100,00%
ConvNeXt V2	95,88%	96,43%	95,56%	95,82%
HiViT	94,85%	94,90%	94,74%	94,81%

Tabela 4.3: Resultados dos modelos para o dataset UFMS-Park.

O trabalho de (Martynova et al., 2024) também investigou o uso de *transformers* para a mesma tarefa. No entanto, os modelos avaliados neste estudo apresentaram resultados superiores. Enquanto no estudo anterior o modelo ViT alcançou um F1-Score de 91.93%, neste trabalho, obteve-se um F1-Score de 98.36% (Tabela 4.2). Além disso, o modelo HiViT atingiu um F1-Score de

94.81%, conforme detalhado na Tabela 4.3. Esses resultados indicam uma melhoria no desempenho em comparação com o estudo anterior.

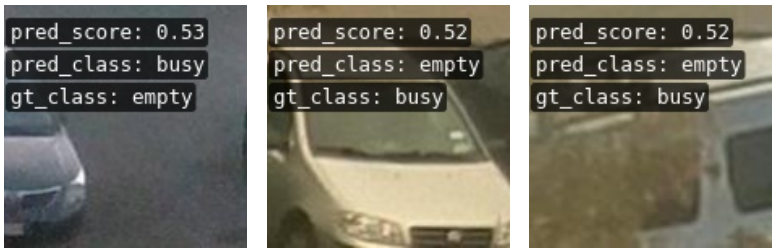
Com base nos resultados obtidos, é evidente que os métodos convolucionais superam os *transformers* na tarefa de classificação de vagas de estacionamento, o mesmo cenário é observado no trabalho de (Martynova et al., 2024). As Tabelas 4.2 e 4.3 demonstraram que o desempenho dos *transformers* em cada conjunto de dados (CNR-Park - Tabela 4.2 e UFMS-Park - Tabela 4.3), é inferior para essa tarefa específica.

Nossos testes revelaram que a Res2Net50 (Gao et al., 2021) alcançou o melhor resultado, demonstrando uma acurácia de 99.09% no conjunto de dados CNR-Park e 100% no conjunto da UFMS. Esses resultados destacam a eficácia desse modelo na tarefa de classificação de vagas de estacionamento.

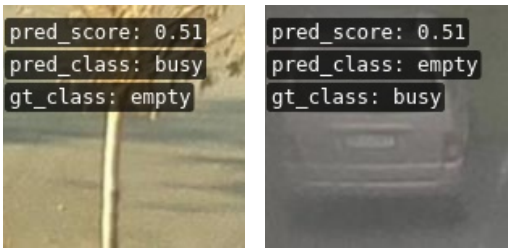
4.2.2 Resultados Qualitativos

Para avaliar as dificuldades e entender a razão dos erros, foram salvas as imagens com erros mais consideráveis. A Figura 4.3 apresenta alguns exemplos. O primeiro fato a ser observado é que a iluminação traz dificuldade para o modelo, como os resultados apresentados na Seção 4.2.1. Os exemplos mostram uma quantidade maior de erros para dias ensolarados e erros com relevância para um sistema de classificação de vagas de estacionamento, pois diz que a vaga está vazia, mas na verdade está ocupada.

Os resultados também mostraram dificuldade para classificação em dias chuvosos (RAINY), pois as imagens ficam embaçadas e a luminosidade é baixa, apresentando mais sombra. Dependendo da cor do carro, ele pode ser confundido com asfalto e então a vaga ser marcada como disponível, como o exemplo exibido na Figura 4.3(e).



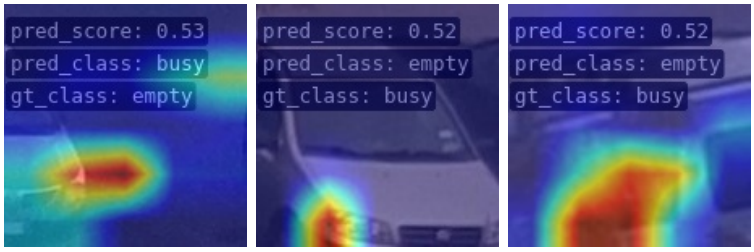
(a) ResNet (b) Res2Net (c) ResNeXt



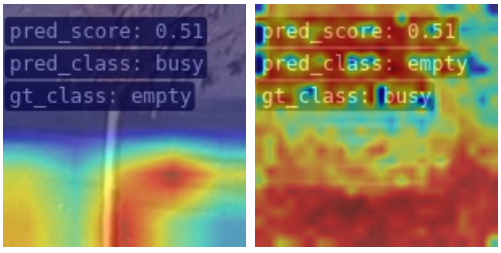
(d) SE-ResNeXt (e) Vision Transformer

Figura 4.3: Exemplos de erros dos modelos.

Para entender o que levou os métodos a cometerem os erros, nós usamos o método GradCAM para visualizar o mapa de ativação da classe. Essa ferramenta é fornecida pelo (Contributors, 2020). A Figura 4.4 exibe alguns exemplos de visualização, por exemplo, a ResNet (Figura 4.4(a)) considerou uma parte do carro para classificar a imagem, mas a vaga em questão estava desocupada. Neste contexto, outra situação relevante é a presença de obstáculos nas imagens, como pessoas, árvores e partes de carro.



(a) ResNet (b) Res2Net (c) ResNeXt



(d) SE-ResNeXt (e) Vision Transformer

Figura 4.4: Mapa de ativação das imagens erradas.

Nos testes conduzidos com o novo conjunto de dados (UFMS), foram identificadas características das imagens incorretamente classificadas por cada modelo avaliado. O modelo HiViT demonstrou imprecisão na classificação de cinco imagens, como mostra a Figura 4.5(a), sendo três delas relacionadas às vagas ocupadas e duas às vagas vazias. A análise revelou uma correlação entre as três imagens classificadas como vagas vazias, uma vez que compartilham o mesmo ângulo de captura. Essa observação sugere uma dificuldade do modelo em lidar com imagens capturadas de cima do veículo. Para compreender as falhas relacionadas às vagas vazias classificadas como ocupadas, procedemos com a análise do mapa de ativação, conforme ilustrado na Figura 4.6. O método identificou erroneamente a pintura do chão como sendo um veículo ocupando a vaga em questão. Em outra instância, a classificação foi confundida com asfalto e um pedaço de carro da vaga ao lado.

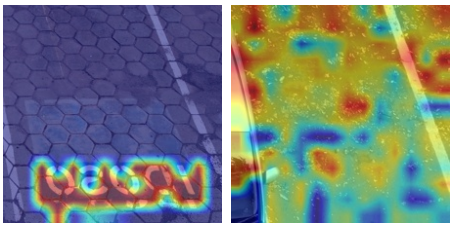


Figura 4.6: Mapa de ativação das imagens erradas pelo HiViT (UFMS-Park).

A Figura 4.5(b) exibe imagens que foram incorretamente classificadas pelo método ConvNeXt V2. Observa-se a presença de variações significativas na iluminação, com duas dessas imagens capturadas em dias ensolarados, enquanto as outras duas exibem veículos de coloração branca intensa. A possível dificuldade enfrentada pelo modelo pode estar relacionada a essas características de iluminação.

4.2.3 Complexidade Computacional

Nesta seção, nós comparamos os métodos em termos de eficiência computacional. A Tabela 4.4 apresenta o tempo médio obtido por cada método para classificação de 13.050 imagens (CNR-Park) e a Tabela 4.5, para classificação de 97 imagens (UFMS - Park). Com esses resultados, calculamos o FPS, que é a quantidade de imagens processadas em 1 segundo. Neste contexto, conseguimos discutir a viabilidade de execução em tempo real de cada modelo.

Métodos	**Tempo Médio**	**Desvio Padrão**	**FPS**
ResNet	0,008s	0,001s	125
Res2Net	0,020s	0,002s	50
ResNeXt	0,008s	0,001s	125
SE-ResNeXt	0,010s	0,001s	100
Vision Transformer	0,455s	0,442s	2.20

Tabela 4.4: Tempo para médio para classificar um conjunto de imagens e FPS.

Considerando o primeiro experimento, a Tabela 4.4 mostra que o ResNet e ResNeXt obtiveram o melhor tempo, ambos com tempo médio de 0.008s, desvio padrão 0.001s e 125 FPS. O pior tempo foi o do *Vision Transformer* com tempo médio de 0,455s, desvio padrão 0,442s e 2.20 FPS.

A Tabela 4.5 exibe os tempos médios registrados para a classificação, utilizando dados composto por imagens da UFMS. Observou-se que tanto o modelo ConvNext V2 quanto o modelo HiViT alcançaram o melhor desempenho temporal, com tempos médios de 0.028s e um desvio padrão de 0.008s. Esses resultados correspondem aproximadamente 35.7 FPS, ou seja, 35.7 imagens processadas por segundo.

Métodos	**Tempo Médio**	**Desvio Padrão**	**FPS**
Res2Net	0,035s	0,007s	28.57
ConvNeXt V2	0,028s	0,008s	35.71
HiViT	0,028s	0,008s	35.71

Tabela 4.5: Tempo para médio para classificar imagens do segundo experimento e FPS.

A Figura 4.7 ilustra uma comparação entre a quantidade de imagens processadas por segundo (FPS) e a taxa de acerto de cada método. Os círculos representam os métodos avaliados utilizando o conjunto de dados CNR-Park, enquanto os quadrados representam os métodos avaliados utilizando o conjunto de dados UFMS-Park. No contexto do conjunto CNR-Park, observou-se que o modelo ResNeXt demonstrou o melhor desempenho, exibindo maior viabilidade para aplicação em tempo real, com uma taxa de acerto superior a 99% e 125 FPS. Em relação ao UFMS-Park, tanto o ConvNext V2 quanto o HiViT alcançaram resultados temporais equivalentes, entretanto, o modelo Res2Net demonstrou maior eficácia na classificação, alcançando uma taxa de acerto de 100%.

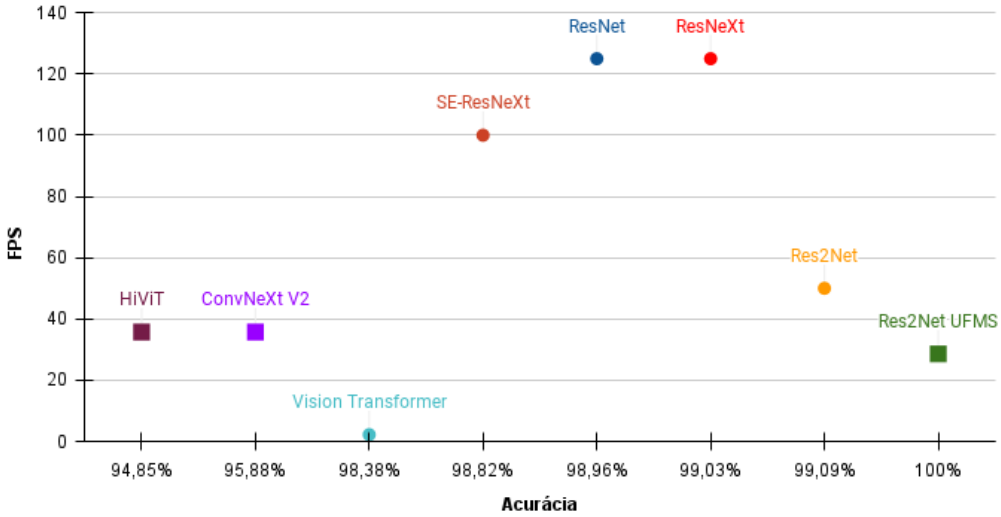


Figura 4.7: FPS x Taxa de Acerto.

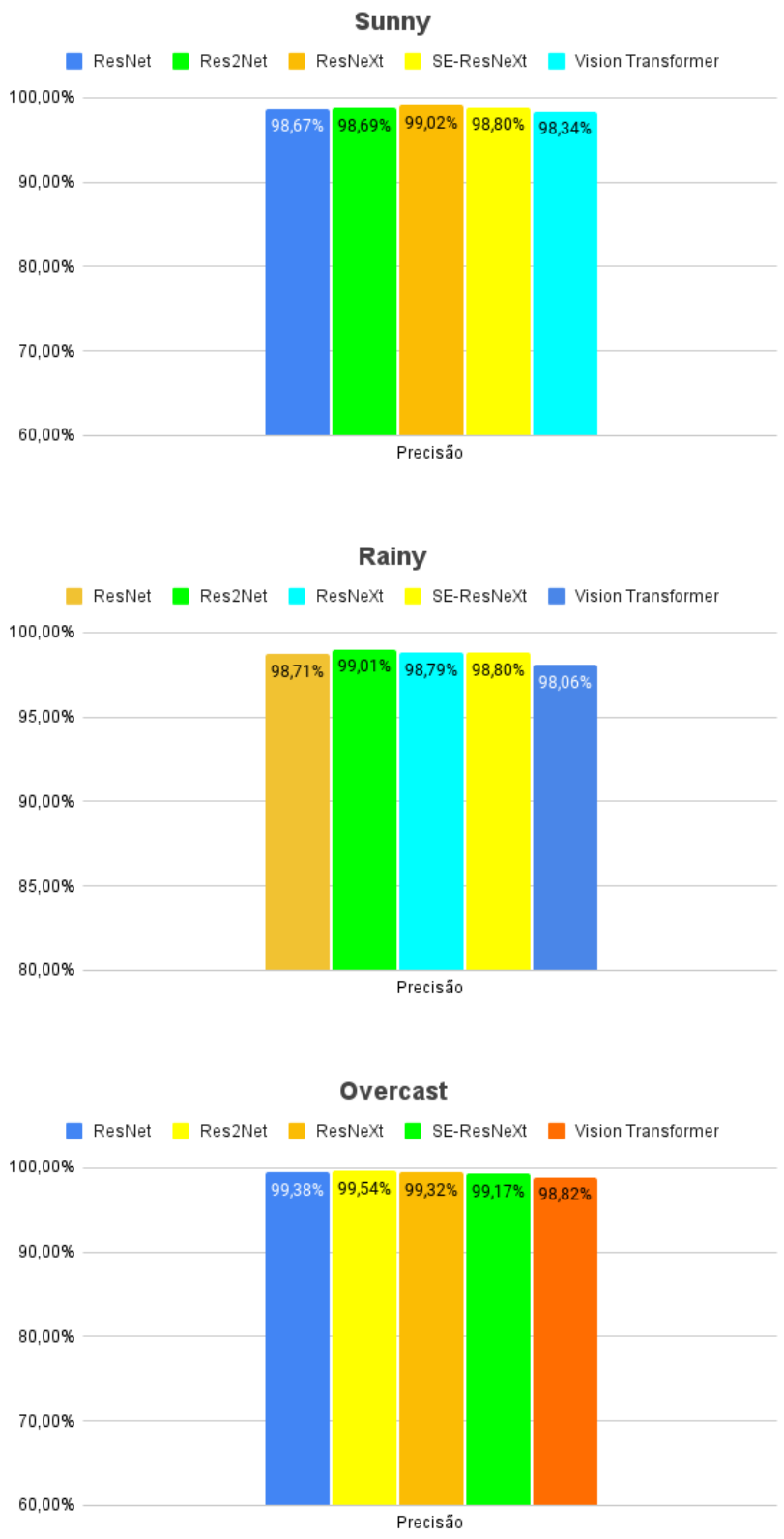
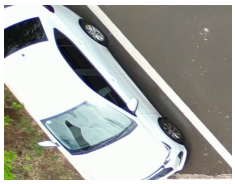


Figura 4.2: Precisão de cada modelo para cada clima.

(a) HiViT



(b) ConvNeXt V2

Figura 4.5: Imagens erradas do conjunto de dados da UFMS.

CAPÍTULO

5

Considerações Finais

Neste capítulo são apresentadas as conclusões deste trabalho. Na Seção 5.1 é realizado um paralelo entre os objetivos do trabalho e os resultados obtidos. Na Seção 5.2 são apresentadas algumas direções de trabalhos futuros.

5.1 Resumo dos Objetivos e Principais Resultados

Neste trabalho, nós avaliamos as redes convolucionais e os *Transformers* para classificação de vagas de estacionamento usando imagens. Os resultados mostraram excelente desempenho para a tarefa, com destaque para a Res2Net Gao et al. (2021) e para o *Vision Transformer* Phuong and Hutter (2022), que atingiram taxa de acerto superior a 99% e 98%, respectivamente. O Res2Net obteve excelente resultado em todos experimentos, vale destacar a taxa de acerto 100% no UFMS-Park. Ambas redes são propostas recentes e em nossa avaliação se mostraram adequadas para o problema, mesmo em condições climáticas adversas como chuva. Por outro lado, o custo computacional mostrou que o Transformer possui um custo computacional acima dos demais, alcançando 2.2 imagens por segundo. A ResNet e ResNeXt alcançaram 125 de fps, mostrando-se mais adequada para um processamento em tempo real.

5.2 Trabalhos Futuros

Para trabalhos futuros, deixamos como sugestão:

- Ampliar a avaliação dos modelos em diferentes conjuntos de dados para

alcançar maior generalização;

- Incluir informação temporal na abordagem de classificação.

Referências Bibliográficas

Acharya, D., Weilin, Y., e Khoshelham, K. (2018). Real-time image-based parking occupancy detection using deep learning. Citado nas páginas 7 e 8.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., e Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8(3). Citado na página 14.

Amato, G., Carrara, F., Falchi, F., Genaro, C., Meghini, C., e Vairo, C. (2016). Deep learning for decentralized parking lot ocupancy detection. Citado nas páginas 4, 7, 8, 11, 21, e 23.

Contributors, M. (2020). Openmmlab's image classification toolbox and benchmark. https://github.com/open-mmlab/mmclassification. Citado nas páginas 22 e 26.

Derísio, J. C. (2017). *Introdução ao controle de poluição ambiental*. São Paulo: Oficina de Textos, 5th edition. Citado na página 3.

Dhuri, V., Khan, A., Kamtekar, Y., Patel, D., e Jaiswal, I. (2021). Real-time parking lot occupancy detection system with vgg16 deep neural network using decentralized processing for public, private parking facilities. Citado nas páginas 7 e 8.

Ding, X. e Yang, R. (2019). Vehicle and parking space detection based on improved yolo network model. Citado nas páginas 4, 7, e 8.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., e Torr, P. (2021). Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662. Citado nas páginas 5, 16, 24, 25, e 33.

Gomes, E. P. (2009). Levantamento das principais fontes de emissões atmosféricas na cidade de manaus. Citado na página 3.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., e Tao, D. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, páginas 1–1. Citado nas páginas 17 e 22.

He, K., Zhang, X., Ren, S., e Sun, J. (2015). Deep residual learning for image recognition. Citado nas páginas 14 e 24.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., e Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. Citado na página 24.

Hu, J., Shen, L., Albanie, S., Sun, G., e Wu, E. (2017). Squeeze-and-excitation networks. Citado na página 15.

Huang, G., Liu, Z., van der Maaten, L., e Weinberger, K. Q. (2016). Densely connected convolutional networks. Citado na página 23.

Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., e Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. Citado na página 7.

Kyu Park, J. e Young Park, E. (2022). Classification of parking lot occupancy using deep learning. Citado nas páginas 4, 7, e 8.

Mahmud, R., Saif, A., e Gomes, D. (2020). A comprehensive study of real-time vacant parking space detection towards the need of a robust model. Citado na página 4.

Mariano, D. (2021). Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e f-score. Citado na página 19.

Martynova, A., Kuznetsov, M., Porvatov, V., Tishin, V., Kuznetsov, A., Semenova, N., e Kuznetsova, K. (2024). Revising deep learning methods in parking lot occupancy detection. Citado nas páginas 1, 8, 9, 24, e 25.

Mora, J., Lopera, J., e Cortes, D. (2018). Automatic visual classification of parking lot spaces: A comparison between bof and cnn approaches. Citado nas páginas 7 e 8.

Oliveira, L., Milena, P., Araújo, G., Dias, J., e Haddad, D. (2020). Detecção inteligente de vagas de estacionamento baseado em climas usando imagens aéreas e aprendizado profundo. In *Xxxviii Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. Sbrt. Citado nas páginas 3, 8, e 23.

Phuong, M. e Hutter, M. (2022). Formal algorithms for transformers. Citado na página 33.

Rahman, S., Ramli, M., Arnia, F., Sembiring, A., e Muharar, R. (2020). Convolutional neural network customization for parking occupancy detection. In *2020 International Conference on Electrical Engineering and Informatics (ICELTICs)*, páginas 1–6. Citado nas páginas 7 e 8.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. Citado na página 21.

Sanghyun Woo, Shoubhik Debnath, R. H. X. C. Z. L. I. S. K. e Xie, S. (2023). Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*. Citado nas páginas 5, 8, 16, 21, e 24.

Xie, S., Girshick, R., Dollár, P., Tu, Z., e He, K. (2016). Aggregated residual transformations for deep neural networks. Citado na página 15.

Zhang, W., Yan, J., e Yu, C. (2019). Smart parking system based on convolutional neural network models. Citado nas páginas 7, 8, e 23.

Zhang, X., Tian, Y., Xie, L., Huang, W., Dai, Q., Ye, Q., e Tian, Q. (2023). Hivit: A simpler and more efficient design of hierarchical vision transformer. In *International Conference on Learning Representations*. Citado nas páginas 5, 8, 17, 22, e 24.