Universidade do Minho
Escola de Engenharia

Trabalho Prático

ViT para Classificação de Imagens de Veículos

Aprendizagem Profunda
Mestrado em Engenharia Informática

Grupo 10

Rita Costa Dantas - pg51605

Eduardo Fernando Cruz Henriques - pg54780

Marcos Paulo Pianissola de Cerqueira - pg52692

janeiro, 2024

1 Introdução

Este projeto tem como objetivo a classificação de diferentes tipos de veículos utilizando uma combinação de um modelo pré-treinado Vision Transformer (ViT) e um Perceptron Multicamadas (MLP). A classificação de veículos é uma tarefa importante em várias aplicações, desde a gestão de frotas até à vigilância em estradas e estacionamentos. Com o avanço das técnicas de visão por computador, torna-se possível automatizar este processo com elevada precisão.

A abordagem proposta neste projeto envolve dois componentes principais: a extração de características e a classificação. O ViT, um modelo de ponta baseado em transformadores, é utilizado para extrair características detalhadas e abstratas das imagens dos veículos. Este modelo foi previamente treinado num vasto conjunto de dados (ImageNet), o que lhe confere uma capacidade robusta de reconhecimento de padrões visuais.

Após a extração das características pelo ViT, estas são utilizadas como entrada para um MLP, que realiza a classificação final. O MLP é uma rede neural simples, mas eficaz, composta por camadas densas e funções de ativação, que aprende a mapear as características extraídas para as categorias específicas de veículos.

Este relatório detalha todo o processo, desde a preparação dos dados até à avaliação do desempenho do modelo.

Por fim, este projeto demonstra a eficácia da combinação de um modelo avançado de visão por computador com uma rede neural clássica para a tarefa de classificação de imagens de veículos, mostrando como técnicas modernas podem ser aplicadas para resolver problemas práticos de forma eficiente.

1.1 Caso de Estudo

O objetivo deste caso de estudo é desenvolver um sistema de classificação de imagens de veículos em diferentes categorias, utilizando técnicas avançadas de visão por computador. Para alcançar este objetivo, utilizamos uma abordagem que combina um modelo Vision Transformer (ViT) pré-treinado para extração de características e um Perceptron Multicamadas (MLP) para a classificação final.

2 Arquitetura

Nesta secção, discutiremos os conjuntos de dados utilizados para a nossa análise, abordando também os métodos de processamento, armazenamento e visualização de dados adotados. O nosso estudo foca-se em dois conjuntos de dados principais: "Vehicle Type Image Dataset".

2.1 Dataset

O dataset que escolhemos para este projeto foi então "Vehicle Type Image Dataset"é composto por 4.356 amostras de imagens que podem ser separadas em

cinco classes de tipos de veículos, da seguinte forma: 1.230 sedans, 1.240 pick-ups, 680 SUVs, 606 hatchbacks e 600 outras imagens de veículos.

2.2 Preparação dos Dados

A preparação dos dados é uma etapa crucial em qualquer projeto de aprendizagem de uma máquina, pois garante que os dados estejam no formato adequado para serem utilizados pelo modelo. Neste projeto, a preparação dos dados envolveu várias etapas importantes, que são descritas a seguir:

2.2.1 Organização e Processamentos dos Dados

O primeiro passo foi organizar as imagens no formato esperado pelo nosso pipeline. O conjunto de dados "Vehicle Type Image Dataset"foi organizado em subcategorias, onde cada corresponde a uma classe de veículo (Sedan, Pickup, SUV, Hatchback, e Outros).

2.2.2 Análise Exploratória dos Dados

Antes de realizar qualquer pré-processamento, foi realizada uma análise exploratória dos dados para entender melhor a distribuição das classes e o tamanho das imagens. Esta análise revelou que havia um total de 4.356 imagens distribuídas entre as cinco classes, com uma leve variação nos tamanhos das imagens.

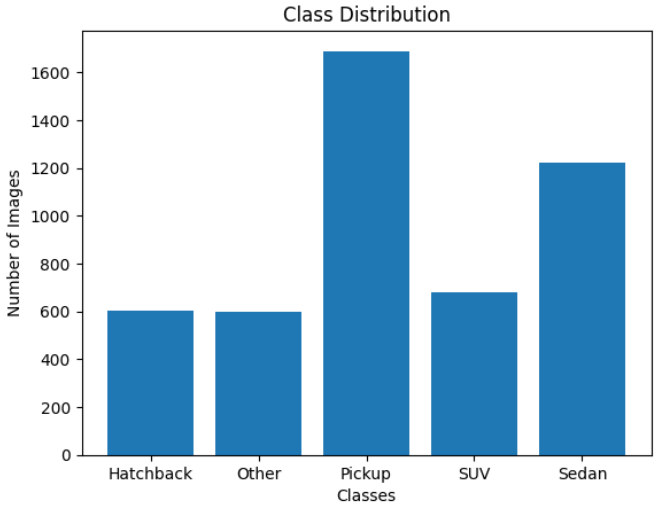

Figura 1: Distribuição das Subcategorias do Dataset

Como podemos ver, a classe pickup tem uma grande maioria das imagens presentes no dataset.

2.2.3 Normalização e Redimensionamento das Imagens

Para garantir que todas as imagens tivessem o mesmo tamanho e estivessem normalizadas, utilizamos transformações padrão. As imagens foram redimensionadas para 224x224 pixels, o que é adequado para a entrada do modelo Vision Transformer (ViT). Além disso, aplicamos normalização usando as médias e desvios padrão das três cores (RGB) do ImageNet, conforme esperado pelo modelo pré-treinado.

2.2.4 Aumento dos Dados

Por outro lado, para melhorar a robustez e a generalização do modelo, aplicamos técnicas de aumento de dados (data augmentation). Essas técnicas incluem rotações aleatórias, cortes e ajustes de brilho/contraste, que ajudam o modelo a aprender invariâncias e se tornar mais robusto a variações nas imagens de entrada.

2.2.5 Divisão dos Dados

Po fim, dividimos o conjunto de dados em 80 PORCENTO para treinamento e 20 PORCENTO para teste. Desta forma, esta divisão garante que o modelo possa ser avaliado de forma justa em dados que não foram vistos durante o treino.

Com todas estas etapas, garantimos então que os dados estivessem prontos para serem utilizados pelo modelo ViT para extração de características e pelo MLP para classificação. A preparação cuidadosa dos dados é fundamental para o sucesso de qualquer projeto de aprendizagem, garantindo que o modelo receba entradas consistentes e representativas do problema a ser resolvido.

2.3 Extração de Características com ViT

A extração de características é uma etapa fundamental no pipeline de classificação de imagens, pois transforma as imagens brutas em representações mais abstratas e discriminativas, que são mais adequadas para a tarefa de classificação. Neste projeto, utilizamos o Vision Transformer (ViT), um modelo avançado baseado em transformadores, para realizar essa tarefa. A seguir, detalhamos o processo de extração de características utilizando o ViT.

2.3.1 Vision Transformers (ViT)

Para este projeto, utilizamos um modelo ViT pré-treinado no conjunto de dados ImageNet. O ImageNet é um vasto banco de dados com milhões de imagens rotuladas em milhares de categorias, o que torna os modelos treinados nele extremamente capazes de reconhecer uma ampla gama de padrões visuais.

Utilizando um modelo pré-treinado, aproveitamos esse conhecimento prévio, permitindo que o ViT extraia características relevantes das imagens de veículos com alta precisão.

2.3.2 Extração de Características

A extração de características envolve passar cada imagem pelo ViT e recolher as representações abstratas (embeddings) geradas pelo modelo. No nosso caso, utilizamos um DataLoader para processar as imagens em lotes e extrair as características correspondentes.

2.3.3 Armazenamento e Organização das Características

Após a extração, as características (embeddings) e os rótulos (targets) são armazenados em arrays numpy. Desta forma, estas características serão utilizadas como entrada para o Perceptron Multicamadas (MLP) na etapa de classificação. Organizar os embeddings desta maneira permite um processamento eficiente e facilita a integração com o MLP.

2.4 Classificação com MLP

Após a extração de características das imagens utilizando o Vision Transformer (ViT), a próxima etapa é a classificação dessas características nas diferentes categorias de veículos. Para essa tarefa, utilizamos um Perceptron Multicamadas (MLP), uma rede neural clássica composta por camadas densas e funções de ativação.

2.4.1 Estrutura do MLP

O MLP é uma rede neural composta por uma camada de entrada, neste caso duas camadas ocultas e uma camada de saída. Cada camada é então formada por neurônios totalmente conectados que aplicam uma transformação linear seguida por uma função de ativação não-linear.

2.4.2 Configuração do Treino

Para treinar o MLP, utilizamos a função de perda CrossEntropyLoss e o otimizador Adam. A função de perda mede a diferença entre as previsões do modelo e os rótulos reais, enquanto o otimizador ajusta os pesos do modelo para minimizar essa perda.

2.4.3 Loop de Treinamento

Neste projeto, o loop de treino envolve as seguintes etapas:

- Passagem direta: As características extraídas pelo ViT são passadas pelo MLP.
- Cálculo da perda: A perda é calculada comparando as previsões do MLP com os rótulos reais.
- Retropropagação: O erro é propagado de volta pelo modelo para ajustar os pesos.

- Otimização: O otimizador atualiza os pesos para minimizar a perda.

Também implementamos a técnica de paragem antecipada (early stopping) para evitar overfitting. O treino é interrompido se a perda de validação não melhorar após um certo número de épocas consecutivas.

2.4.4 Avaliação do Modelo

Após o treino, avaliamos o desempenho do modelo utilizando métricas como precisão, recall e F1-score. Estas métricas ajudam a entender a eficácia do modelo na classificação das diferentes categorias de veículos.

Histórico de Perda

Durante o treino, monitoramos a perda de treino e validação ao longo das étapas. O gráfico a seguir mostra como as perdas mudaram:

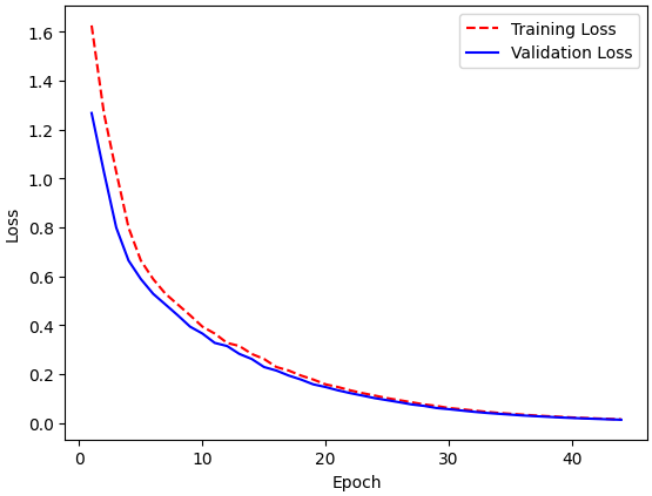

Figura 2: Validação

O gráfico de perda indica que tanto a perda de treino quanto a perda de validação diminuíram significativamente ao longo das épocas. A consistência entre as perdas de treino e validação sugere que o modelo estava a aprender de maneira eficaz sem overfitting.

Avaliação Final

Após o treino, reavaliamos o modelo para determinar o seu desempenho final. Sendo estes:

- Precisão do Teste Final: 98,33
- Precisão Final: 0,98
- Recall Final: 0,98
- Pontuação F1 Final: 0,98

A matriz de correlação final ilustra a precisão do modelo treinado:

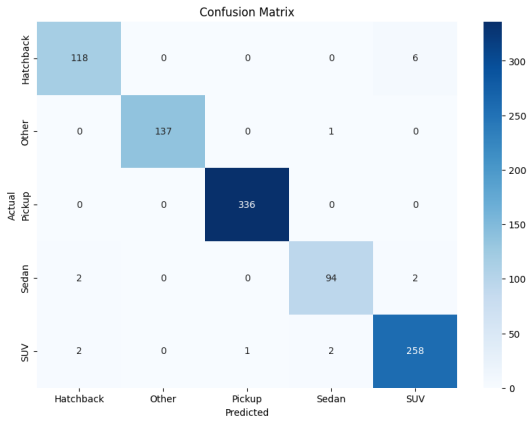

Figura 3: Matriz correlação pós-treino

A matriz de confusão final mostra que o modelo treinado foi capaz de classificar corretamente a maioria das imagens nas categorias corretas. Por exemplo, todas as pickups foram corretamente classificadas, e a maioria dos SUVs e Sedans também foram corretamente identificados. As poucas confusões restantes são mínimas e não afetam significativamente o desempenho geral do modelo.

Por fim, os resultados demonstram que a combinação de características extraídas pelo Vision Transformer (ViT) com um Perceptron Multicamadas (MLP) resulta em um classificador altamente eficaz para a tarefa de classificação de veículos. O modelo inicial apresentou um desempenho fraco, mas após o treino, o modelo final alcançou uma precisão, recall e F1-score altos, mostrando a robustez e eficácia da abordagem proposta.

3 Resultados

Para a análise de resultados fizemos uma análise das imagens classificadas incorretamente e uma comparação entre os vários modelos ViT.

3.1 Imagens Classificadas Incorretamente

A análise das imagens classificadas incorretamente fornece insights adicionais sobre os desafios enfrentados pelo modelo. Abaixo estão exemplos de imagens

que foram classificadas incorretamente pelo modelo final, juntamente com as suas classificações verdadeiras e preditas:

Figura 4: Imagens mal classificadas

Estes exemplos mostram casos em que o modelo final teve dificuldades, como confundir sedans com hatchbacks ou SUVs. Esta análise visual ajuda a identificar possíveis melhorias no pré-processamento ou no modelo.

3.2 Comparação de Diferentes Modelos ViT

Posteriormente, foi realizada uma comparação entre diferentes variantes do modelo ViT. O gráfico abaixo mostra a precisão alcançada por cada modelo com uma e três transformações de dados:

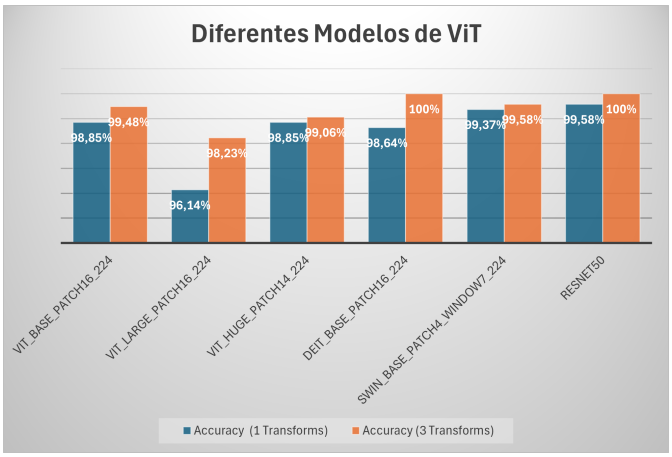

Figura 5: Diferentes modelos

Desta forma, os resultados indicam que o modelo "ViT_BASE_PATCH16_224"com três transformações obteve a melhor precisão, alcançando 99,48 %. Essa comparação é útil para escolher a melhor arquitetura para aplicações futuras.

É interessante observar que, ao comparar as redes ViT com as redes convolucionais, no caso utilizando a ResNet50, a rede convolucional, apesar de ter obtido um desempenho ligeiramente superior, apresenta um custo computacional excessivamente elevado quando comparado com os modelos ViT e MLP.

Concluindo, os resultados finais demonstram que a combinação do Vision Transformer (ViT) para extração de características e do Perceptron Multicamadas (MLP) para classificação é altamente eficaz para a tarefa de classificação de veículos. O modelo final alcançou uma alta precisão, recall e F1-score, mostrando-se robusto e eficiente para a aplicação proposta. As análises adicionais das classificações incorretas e a comparação entre diferentes modelos ViT fornecem também valiosos insights para melhorias e futuras implementações.

4 Limitações

Apesar dos resultados promissores, o modelo apresenta algumas limitações:

4.1 Dependência de Dados Rotulados

O desempenho do modelo depende da qualidade e quantidade de dados rotulados. Dados insuficientes ou mal anotados podem prejudicar a precisão do modelo.

4.2 Generalização Limitada

O modelo foi testado em um conjunto específico de dados e pode não generalizar bem para novos cenários com diferentes condições de iluminação, ângulos de

visão e tipos de veículos não presentes no treinamento.

4.3 Erros de Classificação

Algumas categorias de veículos, como sedans e hatchbacks, são frequentemente confundidas. Melhorias no pré-processamento e mais dados de treinamento podem ajudar a reduzir esses erros.

4.4 Complexidade Computacional

O uso do Vision Transformer (ViT) exige algum poder computacional, especialmente GPUs, o que pode limitar a sua aplicação em ambientes com recursos restritos ou que necessitam de inferência em tempo real.

4.5 Sensibilidade a Transformações

O modelo mostrou sensibilidade a diferentes transformações de dados. É necessário garantir que o modelo mantenha desempenho consistente em diversas condições.

Enquanto a combinação do Vision Transformer (ViT) e do Perceptron Multicamadas (MLP) demonstrou ser eficaz na tarefa de classificação de veículos, as limitações mencionadas devem ser abordadas para melhorar ainda mais a robustez e a aplicabilidade do modelo. Trabalhos futuros podem focar em expandir o conjunto de dados, melhorar a generalização do modelo e otimizar a eficiência computacional para superar essas limitações.

5 Conclusão

Os resultados deste projeto demonstram a eficácia da combinação do Vision Transformer (ViT) para extração de características e do Perceptron Multicamadas (MLP) para classificação de veículos. O modelo final alcançou alta precisão, recall e F1-score, destacando-se como uma solução eficiente para a classificação de veículos.

A análise das imagens classificadas incorretamente revelou desafios como a confusão entre sedans e hatchbacks ou SUVs, indicando áreas de melhoria no pré-processamento e no modelo.

A comparação entre variantes do ViT mostrou que o modelo ViT BASE PATCH 16 224 com três transformações de dados alcançou a melhor precisão, o que é relevante para futuras implementações.

Apesar dos bons resultados, o modelo tem limitações, como a dependência de dados rotulados de qualidade, dificuldades de generalização para novos cenários, e alta complexidade computacional. Futuras melhorias podem envolver a ampliação do conjunto de dados e a optimização da eficiência computacional.

Este projeto demonstra como técnicas avançadas de visão por computador podem resolver problemas práticos de classificação de veículos, fornecendo uma base para o desenvolvimento de soluções mais robustas em cenários reais.