

Classificação de espécies de árvores em imagens de cenário urbano usando Redes Neurais Convolucionais com Triplet Loss

Filipe Correia Belfort¹, Aristófanês Corrêa Silva¹,
Anselmo Cardoso de Paiva¹, Davi Viana dos Santos¹,
Levi Carvalho dos Santos¹, Mario Pinto Freitas Filho¹,
Patrick Camara Araujo¹, Laura Giovanna Freitas Castro²,
Luan Felipe Araujo Sousa²

¹ Núcleo de Computação Aplicada, Universidade Federal do Maranhão,
Av. dos Portugueses 1966 São Luís, Maranhão, Brasil

²Instituto de Ciencia e Tecnologia Grupo Equatorial, São Luís, Maranhão, Brazil

```
{cpbelfort, ari, paiva patrick.araujo}@nca.ufma.br  
  
{davi.viana, levi.cs, mario.freitas,}@nca.ufma.br  
  
laura.castro@equatorialenergia.com.br,  
  
laura.castro@eqtlab.com.br
```

Abstract. *The cataloging of tree species is essential for biodiversity conservation and urban planning; however, traditional field-based methods are costly and have limited scalability. To address these challenges, this study investigates the use of Deep Learning for the classification of 14 tree species in urban environment images through similarity learning based on Triplet Loss and attention mechanisms. The results demonstrate that similarity-based approaches can achieve higher accuracy with fewer data and lower computational cost, with the EfficientNet-B5 network combined with Triplet Loss standing out by achieving a mean F1-score of 0.881 and greater robustness in challenging classes with limited samples, such as *Holocalyx balansae* and *Ligustrum lucidum*.*

Resumo. *A catalogação de espécies arbóreas é fundamental para a conservação da biodiversidade e o planejamento urbano, porém métodos presenciais possuem alto custo e alcance limitado. Buscando mitigar essas barreiras, o presente trabalho investiga o uso de Deep Learning para classificação de 14 espécies de árvores em imagens de contexto urbano via aprendizado de similaridade com Triplet loss e módulos de atenção. Os resultados demonstram que métodos de similaridade podem obter maior precisão com menos dados e custo computacional, destacando-se a rede EfficientNet-B5 com Triplet Loss, que alcançou F1-score médio de 0,881 e maior robustez em classes desafiadoras e com poucas amostras, como *Holocalyx balansae* e *Ligustrum lucidum*.*

1. Introdução

A gestão de ecossistemas urbanos e florestais tornou-se um desafio frente às mudanças climáticas globais e à urbanização acelerada. A catalogação precisa de espécies arbóreas

é essencial não apenas para a conservação da biodiversidade, mas também para o planejamento urbano, estimativa de estoque de carbono e manutenção de serviços ecossistêmicos [Silva et al. 2023]. Nesse cenário, os métodos convencionais de reconhecimento botânico, fundamentados em inspeções presenciais, oferecem um alto custo operacional e de alcance geográfico. Essa limitação torna necessária a adoção de soluções digitais automatizadas que assegurem um monitoramento rigoroso e escalável desses ativos biológicos.

Com o advento das imagens de alta resolução, o monitoramento de problemas relacionados a vegetação passou a integrar sistemas de informação geográfica cada vez mais complexos [Kattenborn et al. 2021], possibilitando a utilização de métodos computacionais robustos como o aprendizado profundo (*Deep Learning*) que representou uma mudança de paradigma nesse cenário. Diferentemente das abordagens clássicas de visão computacional, as Redes Neurais Convolucionais (CNNs) aprendem automaticamente representações internas a partir dos próprios pixels da imagem, sem a necessidade de extração manual de características. Isso permite que o sistema identifique nuances complexas na textura das folhas e na arquitetura das copas que seriam imperceptíveis a métodos de filtragem linear, superando a subjetividade da extração manual de características [Zhu et al. 2017].

Uma abordagem que tem apresentado bons resultados em tarefas de classificação e identificação de imagens por similaridade é o *Deep Metric Learning* (DML). Diferentemente de métodos tradicionais de classificação, essa abordagem aprende um espaço de representações no qual imagens semanticamente semelhantes são posicionadas próximas entre si, enquanto imagens de classes distintas permanecem mais distantes [Hoffer and Ailon 2015]. Essa estratégia é especialmente útil na classificação de espécies de árvores, pois permite capturar diferenças visuais sutis entre espécies semelhantes e organizar as imagens de forma mais discriminativa no espaço de características.

Trabalhos recentes demonstram a eficácia do aprendizado profundo em cenários que variam do mapeamento do dossel (camada superior de uma floresta) via sensores aéreos [Osako et al. 2021] ao processamento de registros terrestres, superando a heterogeneidade morfológica das espécies. No entanto, a escassez de dados balanceados e a complexidade de ambientes urbanos permanecem como desafios centrais [Cetin et al. 2022]. Nesse contexto, abordagens baseadas em *deep learning* têm se mostrado promissoras na identificação e catalogação de espécies arbóreas, uma vez que permitem a extração automática e hierárquica de características discriminativas. O trabalho de [Passos et al. 2022] que, emprega o mesmo dataset utilizado neste estudo, o *Street Level Tree Species Classification* [Jodas et al. 2022] investiga estratégias de sobreamostragem multiclasse, devido a baixa quantidade de amostras por espécie, aplicadas a features extraídas pela arquitetura MobileNet. Nesse estudo, técnicas baseadas em Floresta de Caminhos Ótimos (OPF) são empregadas para balancear os embeddings, que servem como entrada para os classificadores Random Forest e SVM. Ao aplicar essa abordagem no conjunto de dados *Street Level Tree Species Classification*, os autores alcançaram um F1-Score médio de $0,6170 \pm 0,10$.

Este trabalho propõe uma metodologia robusta baseada em DML utilizando a triplet loss para a classificação de espécies de árvores. O estudo se utilizou dos datasets Recogna-Lab [Passos et al. 2022] em conjunto com o *Tree Dataset of Urban Street* o

mesmo utilizado no estudo de [Jodas et al. 2022], que reúne um amplo acervo de diferentes espécies de árvores em contexto urbano. O estudo busca integrar arquiteturas profundas de (CNNs) juntamente com uma abordagem de classificação baseada em DML, utilizando a função de perda *Triplet loss*. Além disso, o trabalho também apresenta um estudo comparativo com o método de classificação direta (tradicional) utilizando arquiteturas de deep learning.

O presente trabalho está organizado da seguinte forma: a Seção 2 descreve os materiais e métodos, incluindo os datasets, o método proposto, as arquiteturas utilizadas e as métricas de avaliação; a Seção 3 apresenta os resultados experimentais e suas análises; por fim, a Seção 4 apresenta as conclusões e direções para trabalhos futuros.

2. MATERIAIS E MÉTODOS

2.1. Base de Imagens

Os experimentos foram realizados utilizando a união de dois datasets. O dataset *Street Level Tree Species Classification* [Jodas et al. 2022] compreende 727 fotografias em perspectiva de rua obtidas de cinco espécies de árvore na cidade de São Paulo, Brasil, todas com diferentes resoluções com espaço de cor padrão RGB. Além disso junto ao dataset está disponível um arquivo de texto com o rótulo, informando a espécie de cada imagem. As imagens são recortes de caixas delimitadores que cercam as árvores. Todas as imagens foram redimensionadas para os tamanhos correspondentes ao padrão de entrada dos modelos de rede neural testados neste trabalho.

Tabela 1. Número de amostras disponíveis para cada espécie nos datasets de árvore.

Espécies	Amostras	Dataset
Caesalpinia peltophoroides	273	Street Level
Holocalyx balansae	48	Street Level
Ligustrum lucidum	46	Street Level
Tibouchina granulosa	109	Street Level
Tipuana tipu	251	Street Level
Cinnamomum camphora (Linn) Presl	234	Tree Dataset Urban
Flowering cherry	106	Tree Dataset Urban
Ginkgo biloba	228	Tree Dataset Urban
Lagerstroemia indica	65	Tree Dataset Urban
Magnolia grandiflora L	212	Tree Dataset Urban
Platanus	193	Tree Dataset Urban
Prunus cerasifera f. atropurpurea	99	Tree Dataset Urban
Salix babylonica	136	Tree Dataset Urban
Sapindus saponaria	224	Tree Dataset Urban

Já o *Tree Dataset of Urban Street* é um conjunto de dados público, abrangente e amplamente disponível, que contempla 50 espécies de árvores e 41.467 imagens de alta resolução para classificação, das quais 22.872 estão anotadas, provenientes de 10 diferentes cenas urbanas. O conjunto de dados é composto por subconjuntos destinados à classificação de folhas, árvores, troncos, galhos, flores e frutos, além de subconjuntos voltados à segmentação de folhas, árvores, troncos [Yang et al. 2023]. Dentre as partições disponíveis, foram utilizados os dados destinados à classificação de árvores, que contém

imagens de árvores em contexto urbano e 23 espécies catalogadas com suas respectivas labels. A partir desse conjunto foram selecionadas 9 espécies que direcionassem os experimentos para o treinamento de modelos de classificação de espécies com características mais próximas da realidade regional Brasileira. Na Tabela 1 é possível visualizar o número de exemplares por espécie no dataset. As 14 classes sobressalentes deste dataset foram utilizadas em uma etapa de pré-treinamento do modelo de classificação direta para fins de comparação do método proposto, com um modelo de classificação pré-treinado no domínio botânico.

2.2. Classificação Baseada em Similaridade

Neste tópico é apresentado as etapas do método proposto para classificação das espécies arbóreas com base na similaridade de imagem, no fluxograma da Figura 1 é apresentado o processo, em que, as imagens de entrada passam por um redimensionamento de 456x456 para se adequarem ao tamanho de entrada do *backbone* pré treinado selecionado. Durante a etapa de treinamento do modelo com a *triplet loss* é criado um espaço latente com as imagens de treino. Na etapa de inferência as imagens de teste tem sua distância calculada em relação ao restante dos dados, e a classificação é realizada com base no ranqueamento dos embeddings retornados, em que a classe da imagem de teste é inferida por meio de votação majoritária entre seus k vizinhos mais próximos no espaço de características.

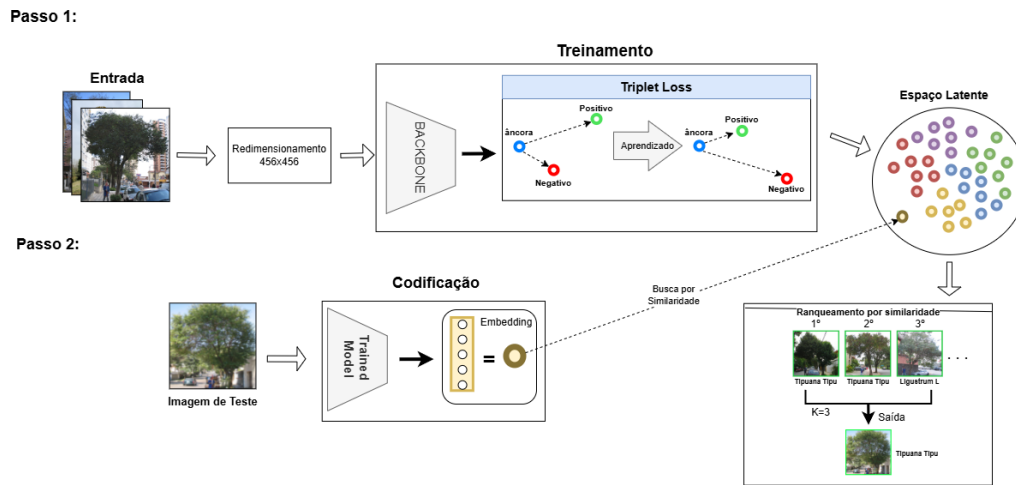


Figura 1. Fluxograma do método proposto.

O DML funciona de forma que, enquanto a classificação tradicional tenta rotular uma imagem em uma categoria fixa, o DML treina a rede neural para atuar como um mapeador de características. Em vez de uma resposta final, o modelo converte a imagem em um embedding, que é uma representação numérica das propriedades visuais daquele objeto em um espaço multidimensional. O objetivo é que a posição desses embeddings no espaço reflita a realidade: imagens do mesmo objeto devem ter coordenadas próximas, enquanto imagens de objetos diferentes devem ser mapeadas para pontos distantes [Kaya and Bilge 2019].

Para que esse mapeamento seja aprendido, utiliza-se no método proposto a *Triplet Loss*, uma função de perda que treina o modelo a partir de trios de amostras compostos por uma imagem âncora, uma imagem positiva da mesma classe e uma negativa pertencente a

uma classe diferente [Schroff et al. 2015], sendo assim uma abordagem robusta para um treinamento multi-classe onde há necessidade de aprender características sutis entre as imagens de diferentes espécies arbóreas. Durante o treinamento, cada imagem é projetada pela rede em um vetor de características no espaço de *embeddings*. A função de perda atua diretamente sobre as distâncias entre esses vetores, incentivando que a distância entre a âncora e o exemplo positivo seja menor que a distância entre a âncora e o exemplo negativo. Formalmente, a perda pode ser expressa como

$$\mathcal{L} = \max(d(f(x_a), f(x_p)) - d(f(x_a), f(x_n)) + \alpha, 0),$$

em que $f(\cdot)$ representa a função de embedding aprendida pela rede, $d(\cdot)$ é uma métrica de distância (geralmente a distância Euclidiana), x_a , x_p e x_n correspondem aos exemplos âncora, positiva e negativa, respectivamente, e α define uma margem mínima de separação entre classes [Schroff et al. 2015]. Dessa forma, o modelo é penalizado sempre que o exemplo negativo estiver mais próximo da âncora do que o positivo dentro dessa margem. Como resultado, o modelo aprende um espaço de representações no qual amostras semanticamente semelhantes tendem a se agrupar, enquanto amostras de classes distintas são afastadas no espaço de características. A implementação da *triplet loss* utilizou um método que atua na seleção de amostras durante o treinamento (hard mining). Dentro de cada *batch*, o algoritmo identifica a amostra positiva mais distante e a negativa mais próxima de cada âncora. Essa técnica direciona o modelo a reduzir a distância entre elementos da mesma classe e aumentar o intervalo entre classes distintas, o que é útil em bases de dados com alta similaridade visual entre as espécies como a utilizada neste experimento.

2.3. EfficientNetB5 + CBAM block

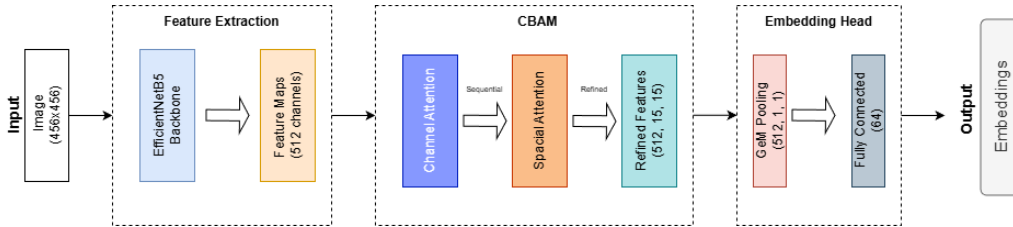


Figura 2. Fluxo da arquitetura EfficientNetB5 + CBAM, a saída são os embeddings que serão organizados pela Triplet Loss

O método proposto conta para etapa de extração de características e geração de embeddings com uso da EfficientNetB5 combinado com uso do mecanismo de atenção CBAM baseado em arquiteturas como a proposta por [Hossain et al. 2026]. A arquitetura EfficientNetB5 utilizada na etapa de extração de características conforme na Figura 2 baseia-se no método de *Compound Scaling*, que escala uniformemente a profundidade, a largura e a resolução da imagem através de um coeficiente fixo [Tan and Le 2019]. Sua unidade fundamental é o bloco MBConv, que utiliza convoluções em profundidade (*depthwise*) e expansão de canais para processar informações de forma leve e eficaz, resultando em um alto desempenho com menor custo computacional. Na implementação um bloco CBAM é inserido após o extrator de características da estrutura principal. De forma mais detalhada os dois componentes sequenciais que compõe o CBAM são: o Channel Attention, que utiliza *Global Average Pooling* e *Global Max Pooling* para identificar quais

canais contêm informações mais relevantes. A Atenção Espacial, que utiliza uma camada convolucional de núcleo 7×7 sobre os mapas agregados para destacar regiões geográficas críticas na imagem [Woo et al. 2018]. Após o refinamento multiescala, o último mapa de características é submetido à camada de *Generalized Mean Pooling* (GeM), que atua como um agregador estatístico superior à média global simples. O GeM é definido pela operação:

$$\mathbf{f} = \left[\frac{1}{N} \sum_{x \in \mathcal{X}} x^p \right]^{1/p}$$

Onde $\mathcal{X}$ representa o mapa de ativação filtrado pelo CBAM, N é o número de elementos espaciais e p é um parâmetro de norma aprendido durante o treinamento. A vantagem fundamental desta abordagem reside na flexibilidade do expoente p : ele permite que o modelo aprenda a transição ótima entre o *Average Pooling* ($p = 1$), que captura informações de contexto, e o *Max Pooling* ($p \rightarrow \infty$), que foca em características locais de maior evidência. Ao final, o vetor resultante é projetado para uma dimensão de incorporação (*embedding*) de 64 (tamanho do espaço latente).

2.4. Arquiteturas de Redes Neurais

Além da **EfficientNetB5** utilizada no método proposto, foram selecionadas arquiteturas consolidadas na literatura para extração de características em tarefas de classificação e segmentação. A **ResNet152** utiliza aprendizado residual, introduzindo *skip connections* que permitem a propagação mais eficiente do gradiente em redes profundas, fazendo com que a rede aprenda apenas refinamentos das representações já extraídas [He et al. 2016].

Já a **DenseNet201** emprega conectividade densa, em que cada camada recebe informações de todas as anteriores por concatenação, favorecendo o reaproveitamento de atributos e a preservação de características de baixo nível ao longo da profundidade da rede [Huang et al. 2017]. A **ConvNeXt Base**, por sua vez, moderniza as CNNs tradicionais ao incorporar conceitos inspirados em Transformers, como convoluções maiores, ativação GELU e *Layer Normalization*, buscando combinar eficiência convolucional com estratégias modernas de treinamento [Liu et al. 2022].

Também foram avaliadas arquiteturas baseadas em mecanismos de atenção. A **Vision Transformer (ViT)** representa a imagem como uma sequência de *patches* convertidos em *tokens*, processados por mecanismos de autoatenção global capazes de modelar dependências entre regiões distantes da imagem desde as primeiras camadas [Dosovitskiy et al. 2020]. Já a mais recente **Swin Transformer** utiliza uma estrutura hierárquica com atenção local em janelas deslocadas (*Shifted Windows*), permitindo capturar relações espaciais locais e globais de forma escalonada [Liu et al. 2021].

2.5. Métricas Utilizadas

O desempenho dos modelos foi avaliado por meio de acurácia (*Accuracy*), precisão (*Precision*), sensibilidade (*Recall*) e F1-score. A acurácia representa a proporção de classificações corretas, enquanto a precisão e o recall medem, respectivamente, a confiabilidade das predições positivas e a capacidade do modelo em identificar corretamente as amostras de cada classe. O F1-score, utilizado como principal métrica de comparação, combina precisão e recall em uma única medida. Além disso, foram considerados o *Macro F1-score*, que atribui o mesmo peso a todas as classes ao calcular a média dos

F1-scores individuais, e o *Weighted F1-score*, que pondera a contribuição de cada classe de acordo com sua frequência no conjunto de dados, tornando-se mais adequado para cenários com desbalanceamento de classes.

3. RESULTADOS

3.1. Configurações do Experimento

Para a realização dos experimentos, o conjunto de dados foi dividido em conjuntos de treino e teste. A divisão dos dados para os experimentos Hold-Out que utilizam a *triplet loss* para realizar a classificação através da similaridade da imagem e identificar o *backbone* que possui melhores métricas foi a de 70% para treino e 30% para teste totalizando 1556 imagens para treino e 668 para teste. Além disso foi realizada uma validação cruzada de 3 folds no âmbito de um estudo comparativo entre a melhor rede mediante os experimentos da Subseção 3.2, representando o modelo de classificação por similaridade e as demais arquiteturas de classificação direta.

Para a classificação direta foi realizada a aplicação de um *data augmentation* online, com o objetivo de mitigar o desbalanceamento entre as classes do conjunto de dados. Dessa forma, a cada época, as imagens são apresentadas ao modelo sob diferentes variações durante o treinamento. As transformações utilizadas incluem rotação aleatória de até 25°, *flip* horizontal e *crop* aleatório, com fator de escala variando entre 0.8 e 1.0. Neste método proposto, o pré-treinamento das redes utilizou 14 espécies excedentes do conjunto de dados *Tree Dataset of Urban Street* ou seja, aquelas que não foram selecionadas entre as classes mais comuns da região brasileira relatadas na Seção 2.2. Visto que o pré-treinamento com este dataset de domínio específico apresentou resultados superiores ao realizado com o amplamente utilizado conjunto de dados ImageNet [Deng et al. 2009], ele foi o escolhido para os demais experimentos no treinamento ponta a ponta.

Para a classificação por similaridade foi realizado um Grid Search entre as métricas de distância Euclidiana, Manhattan e Cosseno, além dos espaços latentes gerados pela *triplet loss* com representações de tamanho 32, 64, 128. Desta forma os melhores parâmetros foram: distância euclidiana e representações de tamanho 64. Além de que nestes experimentos não foram aplicadas técnicas de aumento de dados.

Os experimentos foram conduzidos por até 50 épocas para a comparação entre *backbones* no modelo de similaridade, e 100 épocas no estudo comparativo entre o melhor *backbone*, o *backbone* + atenção e o modelo de classificação direta. Além disso foi utilizado *batch size* 32. A função de perda adotada para a classificação direta foi a *CrossEntropyLoss*, em conjunto com o otimizador Adam [Adam et al. 2014], utilizando *learning rate* fixo de até 0.0001, de forma que para classificação por similaridade foi utilizado o mesmo otimizador. O modelo de classificação por similaridade apresentou um tempo total de inferência de até 39.13 segundos nas 668 imagens de teste, 1.86 segundos por inferência por *batch* de tamanho 32 respectivamente. O ambiente de treinamento foi implementado em Python 3.9, utilizando a biblioteca PyTorch para a implementação das arquiteturas e do loop de treinamento. Os experimentos foram executados com o auxílio de uma placa de vídeo NVIDIA RTX 3060 com 12 GB de memória, por meio do driver CUDA.

Tabela 2. Resultados de classificação por espécie utilizando diferentes *backbones* de redes neurais.

Backbone	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
ResNet152	0.8832	0.8206	0.8014	0.8090	0.8826
DenseNet201	0.9162	0.8671	0.8539	0.8588	0.9150
EfficientNet-B5	0.9237	0.8822	0.8736	0.8763	0.9233
Swin Transformer Base	0.3593	0.3357	0.3266	0.3292	0.3592
ViT Tranformer 224	0.2246	0.1857	0.1893	0.1854	0.2234
ConvNeXt Base	0.3743	0.3546	0.3434	0.3466	0.3752

3.2. Classificação por Similaridade (Múltiplos Backbones)

Os resultados apresentados na Tabela 2 revelam uma superioridade das arquiteturas convolucionais profundas em comparação aos modelos baseados em Transformers para a tarefa de classificação de espécies de árvores. O modelo EfficientNet-B5 apresentou o melhor desempenho geral, alcançando uma acurácia de 0.9237 e Macro F1 de 0.8763, o modelo DenseNet201 alcançou resultados próximos com macro F1 de 0.8539, o que evidencia a força desses modelos baseados em processos convolucionais para extração de características e de técnicas de compartilhamento de informações entre features e alta profundidade encontradas nessas redes. Por outro lado, os modelos baseados em atenção pura ou mecanismos híbridos modernos, como o ViT Transformer, Swin transformer e ConvNeXt apresentaram os resultados mais baixos do experimento, com acurácia de 0.2246, 0.3593 e 0.3743, respectivamente.

Nesse cenário, o desempenho superior da família EfficientNet fundamenta-se em sua arquitetura composta por blocos MBConv. Esses blocos empregam convoluções em profundidade, que desmembram a operação convolucional padrão em duas etapas, reduzindo drasticamente o número de parâmetros enquanto mantêm a capacidade de filtrar características espaciais complexas. Somado a isso, o uso de gargalos invertidos (inverted residuals) e mecanismos de Squeeze-and-Excitation permite que a rede recalibre a importância de cada canal de informação de forma dinâmica, focando nos detalhes mais discriminativos das imagens de árvores. Complementarmente, a DenseNet destaca-se pelo reuso de atributos via conexões densas, garantindo a persistência de detalhes de baixo nível e a estabilidade dos gradientes. Em contrapartida, os transformers apresentaram desempenho inferior muito devido a uma necessidade de alta quantidade de dados para esse tipo de arquitetura e configuração multiclasse que adiciona maior complexidade ao problema.

3.3. Resultado da classificação geral

No estudo comparativo entre o método de classificação Direto e por similaridade, temos que, durante a fase de inferência, as imagens de teste são processadas pela arquitetura treinada, que produz um vetor de probabilidades correspondente às quantidade de classes disponíveis no problema. Na camada final da rede é aplicada a função de ativação *softmax*, responsável por normalizar as saídas do modelo e estimar a probabilidade de cada imagem pertencer a uma das 14 classes. A classe associada à maior probabilidade é então selecionada como a predição final do modelo para cada imagem do conjunto de teste.

Os resultados da validação cruzada com 3 *folds*, apresentados na Tabela 3, de-

Tabela 3. Comparação dos modelos de classificação direta (CD) e *Triplet loss* utilizando F1-Score por classe.

Species	EfficientNetB5 (CD)	EfficientNet Triplet + CBAM	EfficientNetB5 + CBAM (CD)	Triplet EfficientNetB5
Caesalpinia peltophoroides	0.711±0.040	0.866±0.019	0.760±0.027	0.888±0.024
Holocalyx balansae	0.199±0.076	0.462±0.048	0.213±0.075	0.666±0.056
Ligustrum lucidum	0.153±0.108	0.537±0.161	0.585±0.119	0.558±0.045
Tibouchina granulosa	0.586±0.030	0.723±0.065	0.693±0.093	0.787±0.053
Tipuana tipu	0.772±0.067	0.902±0.025	0.841±0.036	0.902±0.025
Magnolia grandiflora L.	0.775±0.068	0.972±0.012	0.800±0.078	0.998±0.003
Platanus	0.671±0.095	0.932±0.010	0.718±0.026	0.984±0.007
Prunus cerasifera f. atropurpurea	0.849±0.020	0.969±0.017	0.836±0.040	0.990±0.014
Salix babylonica	0.838±0.041	0.952±0.034	0.894±0.082	1.000±0.000
Sapindus saponaria	0.936±0.009	0.991±0.009	0.960±0.015	0.971±0.009
Cinnamomum camphora	0.762±0.048	0.971±0.008	0.751±0.073	0.983±0.008
Flowering cherry	0.971±0.020	0.986±0.021	0.971±0.023	0.971±0.012
Ginkgo biloba	0.947±0.051	0.982±0.005	0.966±0.010	0.985±0.008
Lagerstroemia indica	0.849±0.012	0.947±0.011	0.776±0.103	0.976±0.019
Mean F1-score	0.720±0.047	0.849±0.033	0.783±0.060	0.881±0.024

monstram a superioridade dos métodos baseados em aprendizado por similaridade em relação à classificação direta. O modelo *EfficientNetB5 (Triplet)* obteve o melhor desempenho global, alcançando F1-score médio de 0.881 ± 0.024 , seguido pelo *Triplet EfficientNet CBAM* com 0.849 ± 0.033 . Os valores superiores aos obtidos anteriormente são justificados pelo treinamento mais longo (100 épocas) e pela avaliação em múltiplos *folds*. Em nível de classe, o modelo *EfficientNetB5 (Triplet)* apresentou os maiores F1-scores na maioria das espécies, destacando-se em *Magnolia grandiflora L.* (0.998 ± 0.003), *Platanus* (0.984 ± 0.007), *Prunus cerasifera f. atropurpurea* (0.990 ± 0.014) e *Salix babylonica* (1.000 ± 0.000). Já os modelos de classificação direta apresentaram maior dificuldade em espécies com menor quantidade de amostras ou elevada similaridade visual, como *Holocalyx balansae* e *Ligustrum lucidum*. Embora a inclusão do mecanismo de atenção CBAM tenha proporcionado melhorias em algumas classes específicas, seus ganhos não foram consistentes, sendo frequentemente superado pelo modelo *Triplet* sem atenção. De forma geral, os resultados indicam que a abordagem baseada em similaridade é mais robusta para a classificação de espécies arbóreas, especialmente em cenários com maior variabilidade intra-classe e menor disponibilidade de amostras.

3.4. Estudo de caso

Esta Seção tem como objetivo analisar, por meio do *Grad-CAM (Gradient-weighted Class Activation Mapping)*, as principais regiões da imagem que contribuíram para a tomada de decisão do modelo durante o processo de inferência. O *Grad-CAM* gera mapas de calor a partir dos gradientes das últimas camadas convolucionais, permitindo visualizar as áreas que exerceram maior influência na predição da rede. Para esta análise, foi utilizado o modelo baseado em *Triplet Loss* com o *backbone EfficientNet-B5*, selecionado por apresentar os melhores resultados nos experimentos realizados.

Conforme ilustrado na Figura 3, observa-se que o modelo concentra sua atenção principalmente na copa e em regiões do tronco das árvores, características relevantes para a identificação das espécies. Em alguns casos, como no exemplo B, o modelo também considera árvores localizadas ao fundo da imagem, o que pode ocasionar classificações incorretas. Entretanto, como frequentemente múltiplas árvores da mesma espécie estão presentes em uma mesma cena, esse comportamento tende a não comprometer signifi-

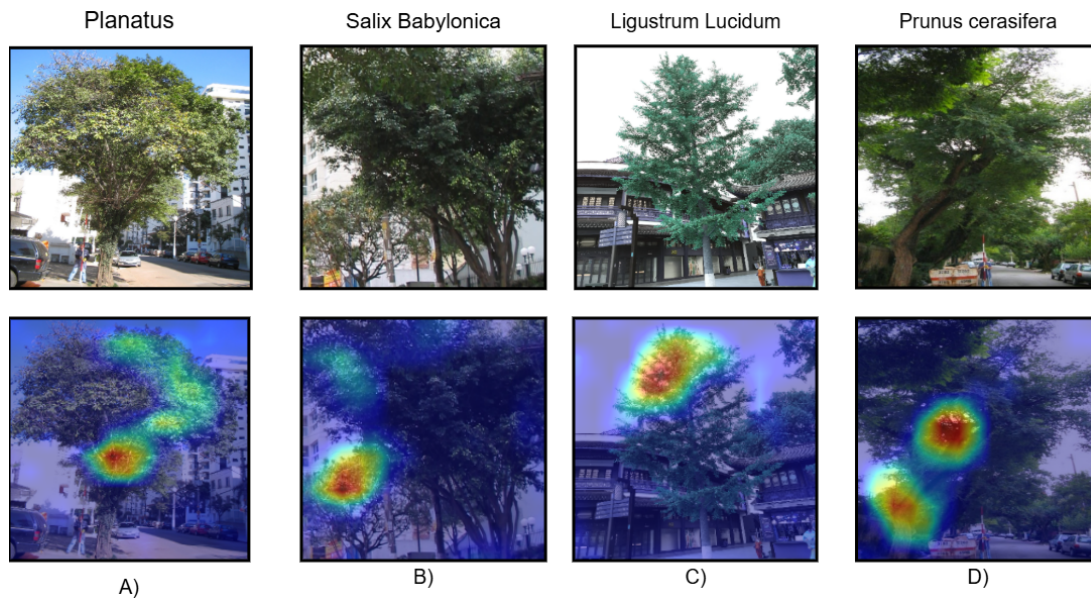


Figura 3. Exemplos de imagens de teste com GradCAM aplicado.

cativamente o aprendizado. De modo geral, os mapas de ativação indicam que o modelo aprendeu a focar em regiões visualmente relevantes para a classificação das espécies arbóreas.

4. CONCLUSÃO

Este trabalho apresenta métodos de deep learning para a classificação de imagens de espécies de árvores em contextos urbanos. Foram comparadas duas abordagens: um método de classificação por similaridade baseado em *Triplet loss*, que aprende um espaço de embeddings onde imagens da mesma classe apresentam menor distância entre si, e um método de classificação direta utilizando a arquitetura EfficientNetB5 [Tan and Le 2019]. Ambos os métodos foram avaliados com uma versão modificada da rede incorporando o módulo de atenção CBAM e a camada de pooling GeM. Os resultados mostram que os métodos baseados em similaridade apresentaram melhor desempenho geral em comparação com as abordagens end-to-end. Em particular, o modelo *EfficientNetB5* + *TripletLoss* obteve os maiores valores de F1-score na maioria das classes, com resultados elevados em espécies como *Salix babylonica* (1.000), *Magnolia grandiflora L* (0.982). Além disso, os métodos com *Triplet loss* apresentaram ganhos relevantes em classes mais desafiadoras, como *Holocalyx balansae* (até 0.666) e *Ligustrum lucidum* (até 0.585), enquanto os modelos de classificação direta obtiveram desempenho inferior nessas mesmas classes. Além de que os modelos de classificação por similaridade têm a vantagem de não precisar de aumento de dados e pré-treinamento no domínio botânico para alcançar melhor desempenho. Como trabalhos futuros, pretende-se avaliar o método em bases de dados com maior diversidade de espécies e explorar arquiteturas com mecanismos adicionais de atenção, como o Coordinate Attention (CA) [Hou et al. 2021] e Multi-Head Self-Attention [Vaswani et al. 2017] além de outros métodos conhecidos de DML e aprendizado com poucas amostras como a N-Pair Loss [Sohn 2016] e Prototype Networks [Snell et al. 2016].

5. Agradecimentos

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil — Código de Financiamento 001; ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil; e à Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA), Brasil. Este trabalho foi apoiado pelo Instituto de Ciência e Tecnologia Grupo Equatorial e pelo Grupo Equatorial, por meio do programa de PDI ANEEL, sob o projeto nº 00037-0046/2022.

Referências

- Adam, K. D. B. J. et al. (2014). A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6).
- Cetin, Z., Yilmaz, A., and Ozcan, G. (2022). Deep learning-based tree species classification using street view images in urban environments. *Urban Forestry Urban Greening*, 74:127638.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- 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*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Hoffer, E. and Ailon, N. (2015). Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.
- Hossain, M. A., Tayef, G., and Subhan, N. (2026). Hybridsolarnet: A lightweight and explainable efficientnet-cbam architecture for real-time solar panel fault detection. *arXiv preprint arXiv:2601.02928*.
- Hou, Q., Zhou, D., and Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- Jodas, D. S., Yojo, T., Brazolin, S., Velasco, G. D. N., and Papa, J. P. (2022). Detection of trees on street-view images using a convolutional neural network. *International Journal of Neural Systems*, 32(01):2150042.
- Kattenborn, T., Leitold, V., Schiefer, F., and Feilhauer, H. (2021). Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:24–49.
- Kaya, M. and Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry*, 11(9):1066.

- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986.
- Osako, L. S., Barbosa, W. N., et al. (2021). Individual tree crown segmentation and species classification in subtropical forest using airborne lidar data and convnets. *Remote Sensing*, 13(18):3714.
- Passos, L. A., Jodas, D. S., Akaki, A. S., and Papa, J. P. (2022). Multiclass oversampling via optimum-path forest for tree species classification from street-view perspectives. *Computers and Electronics in Agriculture*.
- Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- Silva, A. et al. (2023). Classificação de coberturas e espécies de árvores em ambientes urbanos tropicais utilizando redes neurais convolucionais e conceitos de conjuntos fechados e abertos: Revisão bibliográfica. *Revista de Teledetecção*.
- Snell, J., Swersky, K., and Zemel, R. (2016). Prototypical networks for few-shot learning.
- Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.
- Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.
- Yang, T., Zhou, S., Huang, Z., Xu, A., Ye, J., and Yin, J. (2023). Urban street tree dataset for image classification and instance segmentation. *Computers and Electronics in Agriculture*, 209:107852.
- Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36.