UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVI SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2022

Avaliação de Políticas de Aumento de Dados de Estado-da-Arte para Reconhecimento de Espécies de Plantas em Larga Escala através de Deep Learning.

Luciano Araújo Dourado Filho[1];Rodrigo Tripodi Calumby[2];Angelo Conrado Loula[3]

1. Bolsista PIBIC/CNPq, Graduando em Engenharia da Computação, Universidade Estadual de Feira de Santana, e-mail: lucianoadfilho@ecomp.uefs.br
2. Colaborador, Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, e-mail: rtcalumby@uefs.br
3. Orientador, Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, e-mail: angelocl@uefs.br

PALAVRAS-CHAVE: Plantas; Classificação; Deep Learning.

INTRODUÇÃO

As técnicas de Aumento de Dados (AD) consistem na aplicação de transformações sobre imagens a fim de potencializar o treinamento de modelos de aprendizado de máquina através da codificação manual de invariâncias para realização de tarefas de computação visual. Exemplos de técnicas de AD envolvem a aplicação de recortes, rotações, translações, variações de contraste, transformação de estilos, entre outras. Essas abordagens permitem incrementar artificialmente o número de amostras de um conjunto de imagens e têm sido demonstradas eficazes para atenuar as adversidades como desbalanceamento de classes e risco de overfitting que afetam o processo de aprendizado de modelos como Redes Neurais Profundas (RNP) Wong et al. (2016), Pawara et al. (2017). Por conta disso, nos últimos anos, os trabalhos de reconhecimento de plantas em desafios de larga escala têm demonstrado ampla utilização desses métodos para obtenção de resultados promissores, que têm demonstrado papel decisivo para obtenção de modelos de classificação mais eficazes Goeau et al.

A realização de reconhecimento automático de espécies de plantas a partir de imagens é uma tarefa que tem beneficiado da utilização das abordagens de aumento de dados justamente por conta das adversidades provenientes de efeitos, por exemplo, do desbalanceamento na quantidade de amostras por espécie, alta variabilidade visual das imagens, similaridade intra e interclasse, entre outros que inerentes ao domínio dificultam o processo de aprendizado de características discriminativas. Apesar disso, trabalhos recentes demonstraram a capacidade de RNCPs superarem até mesmo especialistas da botânica em desafios de classificação de espécies, o que indica potencial de aprendizado desses modelos quando associados a técnicas eficazes de potencialização, como aumento de dados. Entretanto, trabalhos recentes (Dourado Filho & Calumby, 2021); (Dourado Filho & Calumby, 2022) demonstraram as dificuldades e nuances por trás da escolha dos métodos e heurísticas de aplicação de técnicas de aumento de dados e como podem ser decisivos para melhorar ou degradar a performance dos modelos de classificação. Nesse sentido, outros autores (Cubuk et al., 2019); (Cubuk et al., 2020) conduziram seus esforços em direção à investigar métodos para o descobrimento de políticas (combinações de transformações e métodos de aplicação) de aumento de dados generalizáveis para domínios específicos.

No trabalho de (Cubuk et al., 2019), os autores investigaram um método robusto para construção e exploração de um espaço de busca por políticas de aumento de dados através de Aprendizado por Reforço, denominado Auto Augment (AA). O processo de busca envolveu encontrar sequências de N transformações, suas intensidades e probabilidades de aplicação que apresentassem melhor eficácia para o treinamento de uma RNCP auxiliar destinada para realização da tarefa de classificação sobre o conjunto de dados alvo. Apesar de demonstrarem os resultados satisfatórios de um método rebuscado, o custo computacional proibitivo advindo da quantidade extensiva de hiperparâmetros e a possibilidade de sub-otimização em decorrência da necessidade de utilização de um subconjunto representativo do conjunto de dados para tornar a busca computacionalmente viável foram alguns dos principais pontos-alvo para investigação de uma abordagem alternativa pelos autores em (Cubuk et al., 2020). Por conta disso, além de demonstrarem evidências que corroboram para a hipótese de sub-otimização do AA, os autores de propuseram uma abordagem alternativa que apresentou resultados superiores, denominado Rand Augment (RA) (Cubuk et al., 2020).

Em face disso, o objetivo deste trabalho é avaliar o desempenho do RA, de forma a determinar o conjunto de parâmetros que garanta a melhor eficácia de um modelo de RNCP para realização de uma tarefa de identificação de plantas de larga escala e comparar o desempenho em relação à métodos menos rebuscados já avaliados em tarefas semelhantes. Estima-se que dadas às proporções do objeto de estudo, os parâmetros descobertos sejam concebivelmente generalizáveis para aplicações voltadas para o domínio de plantas de forma geral, no que concerne a transponibilidade para treinamento de outros modelos de classificação à nível de espécies.

METODOLOGIA

Através do Rand Augment, os autores de (Cubuk et al., 2020) possibilitaram a criação de uma abordagem de aumento de dados em que a busca por políticas se delimita apenas à quantidade de operações (transformações de AD) aplicadas em sequência e à magnitude (M), que corresponde à intensidade de aplicação da operação. O método consiste em, dado um conjunto de K transformações de aumento de dados, obter aleatoriamente (com probabilidade uniforme $p=1/K$) um subconjunto de N transformações para serem aplicadas consecutivamente sobre cada imagem que compõe um lote apresentado à rede. Com isso, a tarefa de encontrar uma política é reduzida a apenas encontrar o par {M, N} que permita a obtenção do melhor desempenho para o modelo de classificação avaliado sobre a tarefa desejada, tornando-se computacionalmente viável para uma diversidade muito maior de aplicações.

Tendo em vista não apenas a avaliação do método RA sobre um conjunto de dados suficientemente representativo em termos de quantidade de espécies e de estado-da-arte, mas também realizar um comparativo de desempenho com políticas de DA previamente avaliadas na literatura em conjuntos proporcionalmente menores (quantidade de amostras e espécies), esse trabalho foi realizado sobre o conjunto de dados PlantCLEF2022. O conjunto apresenta em torno de 4 milhões de imagens referentes a 80 mil espécies de plantas e em três subconjuntos: "trusted" (imagens confiáveis), "web" (advinda de *web crawling*) e "test" (destinada apenas à avaliação de modelos).

Apesar da vasta cobertura vegetal e grande quantidade de amostras, o conjunto apresenta adversidades significativas, como desbalanceamento de classes (interno e externo), baixa representatividade, heterogeneidade visual (fundo não homogêneo),

entre outros, como imagens que não correspondem às classes as quais foram associadas. Esses efeitos decorrem principalmente do processo de aquisição de amostras e podem ser atenuados durante o processo de aprendizado através da utilização de AD. Por conta disso, para construção e validação do modelo proposto foi realizada uma análise prévia do conjunto que possibilitou identificar que haviam espécies (cerca de 10%) cuja a quantidade de amostras eram inferiores a 2 imagens confiáveis, enquanto que em torno de 40% do conjunto apresentava menos de 10 imagens confiáveis por espécie. Em consequência disso, a fim de reduzir os impactos provenientes da utilização de possíveis imagens ruidosas, mas também aumentar a quantidade de amostras por espécies a ponto de ao menos garantir uma maior cobertura de classes para obtenção de um particionamento de validação minimamente representativo, juntou-se ao conjunto de imagens confiáveis, as imagens possivelmente ruidosas correspondentes às classes de menor representatividade. Em seguida realizou-se um particionamento aleatoriamente estratificado a nível de espécie, com divisão entre treino e validação (70-30).

Para o processo de aprendizado e reconhecimento ajustamos modelos de arquitetura EfficientNetB0 para treinamento (*fine-tuning*) de todas as camadas, a partir do pré-treinamento sobre a base de dados ImageNet. Para avaliar as configurações sugeridas pelos autores em (Cubuk et al., 2020) (N={1,2} e M={6,10,14}) treinou-se um modelo para cada par de M, N a fim de se comparar o desempenho em relação ao obtido por meio do treinamento do modelo *baseline* com a política Translate + Crop (Dourado Filho & Calumby, 2022). Todas as redes foram treinadas sob as mesmas configurações para fins de comparabilidade, para isso treinou-se até a convergência da função de custo ou até atingir 40 épocas. As demais configurações utilizadas foram: taxa de aprendizado=0.035 e otimizador Stochastic Gradient Descent (SGD) com batch_size=128 com a função de custo Categorical Crossentropy, além da acurácia de validação. Para implementação dos métodos utilizamos as bibliotecas TensorFlow e Imgaug (https://imgaug.readthedocs.io/en/latest/).

RESULTADOS E DISCUSSÃO

A Figura 1 apresenta os resultados de acurácia de validação para as 10 últimas épocas de treinamento dos modelos. Observa-se que o modelo de melhor desempenho foi obtido por meio do treinamento com a política Translate + Crop (Dourado Filho & Calumby, 2021); (Dourado Filho & Calumby, 2022), com acurácia final=0.5277, enquanto a melhor política obtida pelo rand augment obteve acurácia de 0.5235, para os valores de N=1 e M=6. Em contrapartida observou-se também que o modelo que obteve pior desempenho foi obtido através da configuração de aumento mais intensa dentre as avaliadas, com N=2 e M=14. Apesar disso, a política Translate + Crop também apresentou a melhor eficiência, em termos de tempo de processamento possibilitou levar em média aproximadamente 51% do tempo de processamento por época que o método RandAugment de N=1 e aproximadamente 37% do tempo que os modelos treinados com as políticas de N=2.

Entretanto, a baixa expressividade entre os resultados da maior parte das políticas avaliadas chamou atenção para a possibilidade de que a abordagem de aumento de dados não tenha sido capaz de atenuar tão significativamente as adversidades inerentes ao conjunto, entretanto, uma análise mais aprofundada de desempenho, a nível de desbalanceamento, como realizado em (Dourado Filho & Calumby, 2022), pode ser

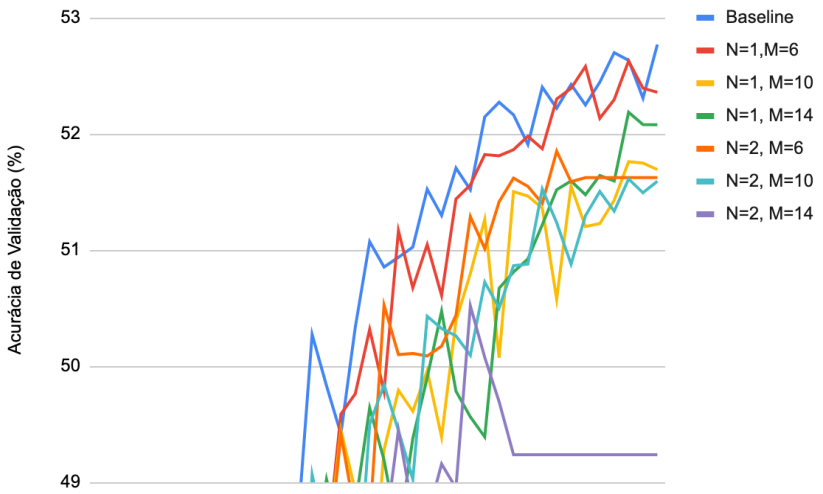

Figura 1: Acurácia de validação ao longo das épocas.

realizada a fim de obter mais conclusões. Apesar disso o desempenho final dos modelos avaliados foram considerados satisfatórios de forma geral, tendo em vista as adversidades de desbalanceamento de classes, imagens possivelmente ruidosas, etc.

CONCLUSÃO

Neste trabalho foi realizado um estudo comparativo de desempenho envolvendo eficácia e eficiência de métodos de estado-da-arte de aumento de dados para treinamento de modelos de RNCP para classificação de espécies de plantas. Os resultados demonstraram a prevalência da superioridade da política Translate + Crop em relação ao método RandAugment para as configurações avaliadas. Como trabalhos futuros visa-se a avaliação de outras configurações paramétricas para o RA, além da condução de análises mais aprofundadas i.e., desbalanceamento, a fim de obter conclusões adicionais a respeito do desempenho dos métodos de estado-da-arte para conjuntos de dados de grandes proporções.

REFERÊNCIAS

WONG, S. C. et al. Understanding data augmentation for classification: When to warp? In: Proceedings of the ICDICTA [S.l.: s.n.], 2016. p. 1–6.

PAWARA, P. et al. M. Data augmentation for plant classification. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, p. 615-626, 2017.

GOËAU, H. et al. Overview of expert life clef 2018. Working Notes of CLEF, 2018. [S.l.: s.n.].

CUBUK, E.D. et al. Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF. 2019. p. 113-123.

CUBUK, E.D. et al. Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF. p. 702-703.

DOURADO FILHO, L.A; CALUMBY, R.T. Experimental evaluation of Data Augmentation heuristics for plant identification systems based on Deep Learning. In: Anais do XIII Congresso Brasileiro de Agroinformática. SBC, 2021. p. 136-143.

DOURADO FILHO, L.A; CALUMBY, R.T. Data Augmentation policies and heuristics effects over dataset imbalance for developing plant identification systems based on Deep Learning: A case study. Revista Brasileira de Computação Aplicada, v. 14, n. 2, p. 85-94, 2022.