CIAGRO 2022
09 a 10 de Junho | Tendências e estratégias para a agroindústria do futuro

GERAÇÃO DE IMAGENS SINTÉTICAS PARA A SEGMENTAÇÃO DE FOLHAS DE CAFÉ USANDO APRENDIZADO PROFUNDO

GENERACIÓN DE IMÁGENES SINTÉTICAS PARA LA SEGMENTACIÓN DE HOJAS DE CAFÉ MEDIANTE DEEP LEARNING

SYNTHETIC IMAGE GENERATION FOR COFFEE LEAF SEGMENTATION USING DEEP LEARNING

Michel Hanzen Scheeren[1]; Mauricio Antonio Gois de Almeida[2]; Ricardo Augusto Pereira Franco[3]; Arnaldo Candido Junior[4]; Pedro Luiz de Paula Filho[5]

DOI: https://doi.org/10.31692/IIICIAGRO.0030

RESUMO

Pragas e doenças representam um risco sério para o café, podendo causar prejuízos e reduzir significativamente a produção. Nesse sentido, a identificação precoce e correta dos sintomas é uma tarefa importante para permitir um tratamento rápido e reduzir os danos. Uma área capaz de auxiliar na identificação de pragas e doenças é a inteligência artificial, mais especificamente o aprendizado profundo, responsável por melhorar o estado da arte em diversos domínios. Um problema comum nessa área é a discrepância entre o desempenho obtido com imagens de laboratório e com imagens reais de campo, o que pode ser mitigado a partir da utilização de técnicas de segmentação dos dados. No entanto, essa abordagem apresenta uma séria limitação: a necessidade de uma quantidade abundante de dados para a otimização dos parâmetros. Uma possível solução para esse problema é a utilização de procedimentos de geração de imagens sintéticas, que podem expandir a quantidade de dados disponíveis para o treinamento e melhorar consideravelmente o desempenho. Dessa forma, este trabalho estudou uma abordagem de geração de imagens sintéticas para auxiliar no treinamento de modelos de aprendizado profundo. A abordagem foi avaliada em um problema de segmentação de imagens de folhas de café obtidas em condições reais de campo a partir de dados sintéticos gerados utilizando imagens de laboratório. Como resultado, comprovou-se que a utilização das imagens sintéticas conseguiu melhorar o desempenho da segmentação em 1,3 pontos percentuais de intersecção sobre união e 0,7 pontos percentuais de F-score, demonstrando ser uma alternativa válida para o aumento de dados voltado ao treinamento de modelos de aprendizado profundo.

Palavras-Chave: imagens sintéticas, aumento de dados, aprendizado profundo.

RESUMEN

Las plagas y enfermedades representan un grave riesgo para el café y pueden causar pérdidas y reducir considerablemente la producción. En este sentido, la identificación temprana y correcta de los síntomas es una tarea importante para permitir un tratamiento rápido y reducir los daños. Un área que puede ayudar en la identificación de plagas y enfermedades es la inteligencia artificial, más específicamente el aprendizaje profundo, que es responsable de mejorar el estado del arte en varios dominios. Un problema

[1] Ciência da Computação, Universidade Tecnológica Federal do Paraná, michelscheeren@alunos.utfpr.edu.br
[2] Ciência da Computação, Universidade Tecnológica Federal do Paraná, mauricioalmeida@alunos.utfpr.edu.br
[3] Doutor, Universidade Federal de Goiás, ricardofranco@ufg.br
[4] Doutor, Universidade Tecnológica Federal do Paraná, arnaldocan@gmail.com
[5] Doutor, Universidade Tecnológica Federal do Paraná, plpf2004@gmail.com

común en este ámbito es la discrepancia entre el rendimiento obtenido con imágenes de laboratorio y con imágenes reales de campo, que puede mitigarse a partir del uso de técnicas de segmentación de datos. Sin embargo, este enfoque presenta una grave limitación: la necesidad de una abundante cantidad de datos para la optimización de los parámetros. Una posible solución a este problema es el uso de procedimientos de generación de imágenes sintéticas, que pueden ampliar la cantidad de datos disponibles para el entrenamiento y mejorar considerablemente el rendimiento. Así, este trabajo estudió un enfoque de generación de imágenes sintéticas para ayudar en el entrenamiento de modelos de aprendizaje profundo. El enfoque se evaluó en un problema de segmentación de hojas de café tomadas en condiciones reales de campo a partir de datos sintéticos generados con imágenes de laboratorio. Como resultado, se demostró que el uso del conjunto de datos sintéticos fue capaz de mejorar el rendimiento de la segmentación en 1,3 puntos porcentuales de intersección sobre unión y 0,7 puntos porcentuales de F-score, demostrando ser una alternativa válida para el aumento de datos destinados al entrenamiento de modelos de aprendizaje profundo.
Palabras Clave: imágenes sintéticas, aumento de datos, aprendizaje profundo.

ABSTRACT
Pests and diseases represent a serious risk for coffee, and can cause losses and significantly reduce production. In this sense, early and correct identification of symptoms is an important task to allow a quick treatment and reduce damage. One area that can assist in the identification of pests and diseases is artificial intelligence, more specifically deep learning, which is responsible for improving the state of the art in various fields. A common problem in this area is the discrepancy between the performance obtained with laboratory images and real field images, which can be mitigated by using data segmentation techniques. However, this approach has a serious limitation: the need for an abundant amount of data for parameter optimization. A possible solution to this problem is the use of synthetic imaging procedures, which can expand the amount of available data for training and considerably improve performance. Thus, this paper studied a synthetic imaging approach to aid in the training of deep learning models. The approach was evaluated on a problem of segmenting coffee leaves taken in real field conditions from synthetic data generated using laboratory images. As a result, it was proven that the use of the synthetic dataset was able to improve the segmentation performance by 1.3 percentage points of intersection over union and 0.7 percentage points of F-score, proving to be a valid alternative for data augmentation aimed at training deep learning models.
Keywords: synthetic images, data augmentation, deep learning.

INTRODUÇÃO

Entre os setores responsáveis por sustentar a economia brasileira, a agricultura desempenha um papel de suma importância, principalmente na geração de empregos e renda para o Brasil (OLIVEIRA *et al.*, 2014). Nos anos de 2019 e 2020, o Brasil foi responsável por cerca de 32,2% de todas as exportações mundiais de café, consagrando-se como maior exportador de café do mundo (ORGANIZAÇÃO INTERNACIONAL DO CAFÉ, 2020). Além da importância econômica, o café também apresenta vários benefícios para a saúde, incluindo efeitos anti-inflamatórios, antioxidantes, auxílio no tratamento de doenças crônicas e atuação como um estimulante natural (CARVALHO *et al.*, 2018).

Existem diversos fatores capazes de prejudicar a produtividade agrícola, um risco tanto para pequenos agricultores familiares quanto para produtores de escala empresarial. Dentre esses fatores destaca-se a incidência de pragas e doenças, problema capaz de causar perdas

severas e que podem até mesmo inviabilizar a exploração da cultura (FERRÃO *et al.*, 2017). Para permitir o controle efetivo da propagação de pragas e doenças e minimizar os eventuais danos que podem ser causados, é importante realizar a identificação precisa dos sintomas, principalmente nos estágios iniciais de contaminação (XIONG *et al.*, 2020).

Uma abordagem moderna para o problema de identificar corretamente os sintomas provocados por pragas e doenças é a utilização de técnicas que combinam visão computacional e aprendizado profundo. Essa área apresentou um crescimento muito significativo, melhorando drasticamente o estado da arte nos mais diversos domínios (LECUN *et al.*, 2015) e demonstrando um desempenho de ponta se comparado às abordagens mais tradicionais de aprendizado de máquina (ALOM *et al.*, 2019).

Um obstáculo comum no treinamento de modelos para a detecção ou classificação de pragas e doenças em plantas é a discrepância entre a acurácia de reconhecimento quando são utilizadas imagens de laboratório, capturadas em condições controladas de fundo e iluminação, e quando são usadas imagens de campo em condições reais (ARSENOVIC *et al.*, 2019). É possível obter melhorias significativas nos resultados utilizando técnicas de segmentação, capazes de contribuir para a evidenciação dos principais padrões presentes nos dados e gerar imagens com características mais homogêneas, permitindo que a rede se concentre apenas nos elementos mais relevantes na cena (BARBEDO, 2019).

Apesar do desempenho de ponta que pode ser alcançado a partir da utilização dessas técnicas, elas apresentam uma séria limitação: o treinamento depende de uma quantidade abundante de dados prontamente disponíveis para a otimização dos parâmetros, até mesmo para compreender conceitos relativamente simples. Uma possível solução para esse problema é a utilização de técnicas de geração de imagens sintéticas, ou seja, exemplos artificiais que se aproximam ou imitam informações reais (MELO *et al.*, 2022). Elas podem ser utilizadas tanto isoladamente (PATKI *et al.*, 2016) quanto em conjunto com os dados reais, expandindo a quantidade de dados disponíveis para o treinamento (FAWAZ *et al.*, 2018).

Dessa forma, este artigo tem por objetivo o estudo de uma abordagem para a geração de imagens sintéticas a partir de técnicas de composição de imagens para auxiliar no treinamento de modelos de aprendizado profundo. Como estudo de caso, a abordagem foi testada na tarefa de segmentação de imagens de café tiradas em condições reais de campo utilizando imagens de laboratório.

REFERENCIAL TEÓRICO

Matematicamente, uma imagem digital pode ser definida como uma função de duas dimensões $f(x, y)$, em que x e y descrevem as coordenadas espaciais e f representa o brilho ou intensidade da imagem naquele ponto. Quando x, y e f são valores finitos e discretos, estes representam uma imagem digital (GONZALEZ; WOODS, 2018). Em termos físicos, uma imagem pode ser representada como o produto da luz que incide na cena (iluminância) e da luz refletida pelos objetos na cena (reflectância) (PEDRINI; SCHWARTZ, 2007).

Portanto, uma imagem digital bidimensional $I(m, n)$ retrata a resposta obtida por um sensor para uma série de posições fixas representadas em um sistema de coordenadas cartesianas de duas dimensões. Cada elemento individual da imagem $I(m, n)$, em que m indica a linha e n a coluna, é chamado pixel. Por convenção, a origem de uma imagem sempre está localizada no canto superior esquerdo (SOLOMON; BRECKON, 2011).

Uma representação muito comum de uma imagem digital colorida é feita utilizando um vetor triplo, em que cada posição representa a intensidade das cores primárias vermelho, verde e azul (RGB). Nesse caso, a imagem é uma combinação linear dos três canais de cores e também pode ser representada por três planos bidimensionais distintos. Alguns formatos de imagem podem ainda apresentar um quarto canal, conhecido como canal alfa, responsável por controlar a transparência da imagem (SOLOMON; BRECKON, 2011).

Por fim, a resolução de uma fonte de imagem pode ser definida de várias formas. A mais comum delas é a resolução espacial, que está fortemente relacionada com a densidade de pixels presente na imagem (PEDRINI; SCHWARTZ, 2007). Neste modelo, a quantidade de colunas (C) e linhas (L) da imagem define a quantidade de pixels usados para preencher o espaço. Normalmente, essa resolução é expressa como $C \times L$: 640×480, 1280×720, 1920×1080, entre outros (SOLOMON; BRECKON, 2011).

Processamento de Imagens Digitais

O processamento de imagens digitais utiliza uma série de abordagens também presentes em áreas correlatas, como análise de imagens e visão computacional (GONZALEZ; WOODS, 2018). Ele pode ser definido como um conjunto específico de técnicas destinadas à captura, representação, transformação e identificação das principais informações de imagens digitais (PEDRINI; SCHWARTZ, 2007).

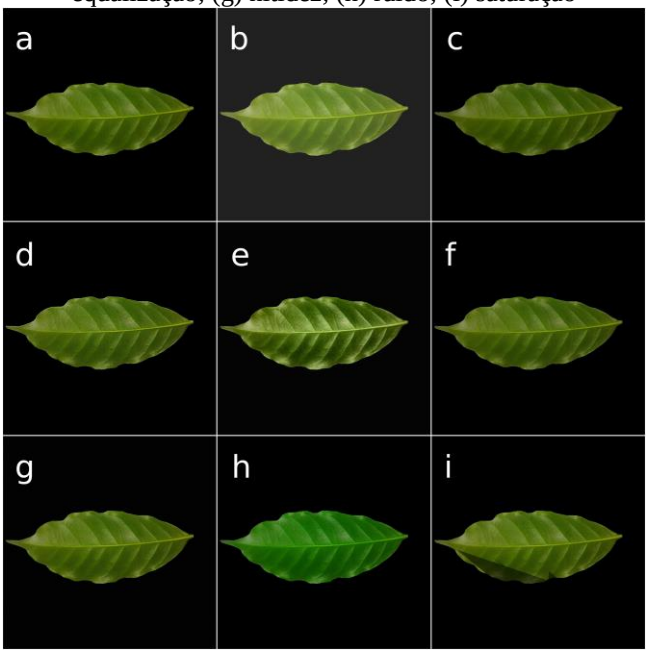
Transformações de pixel correspondem a uma série de transformações relacionadas com

variações locais dos valores dos pixels, sem comprometer a localização física dos objetos na cena, ou seja, não precisam ser aplicadas em anotações prévias. Dentre as transformações de pixel mais comuns, destacam-se (ALBUMENTATIONS, 2022):

- Brilho: variação na luminosidade da imagem, tornando-a mais clara ou escura;
- Contraste: variação na diferença de brilho entre áreas claras e escuras da imagem;
- Saturação: variação na intensidade da cor da imagem;
- Equalização: ajuste do contraste da imagem a partir de seu histograma;
- Ruído: adição de variações aleatórias de brilho e cor nos pixels da imagem;
- Realce: aplicação de técnicas que destacam as bordas dos objetos na imagem;
- Nitidez: aplicação de filtros que destacam os contornos da imagem;
- Desfoque: aplicação de borramento, reduzindo os detalhes e o ruído da imagem.

A Figura 1 apresenta alguns exemplos das transformações de pixel citadas, aplicadas a uma mesma imagem de folha de café pertencente ao *dataset* Esgario *et al.* (2020).

Figura 1: Exemplos de transformações de pixel. (a) original; (b) desfoque; (c) brilho; (d) contraste; (e) realce; (f) equalização; (g) nitidez; (h) ruído; (i) saturação



Fonte: Própria (2022).

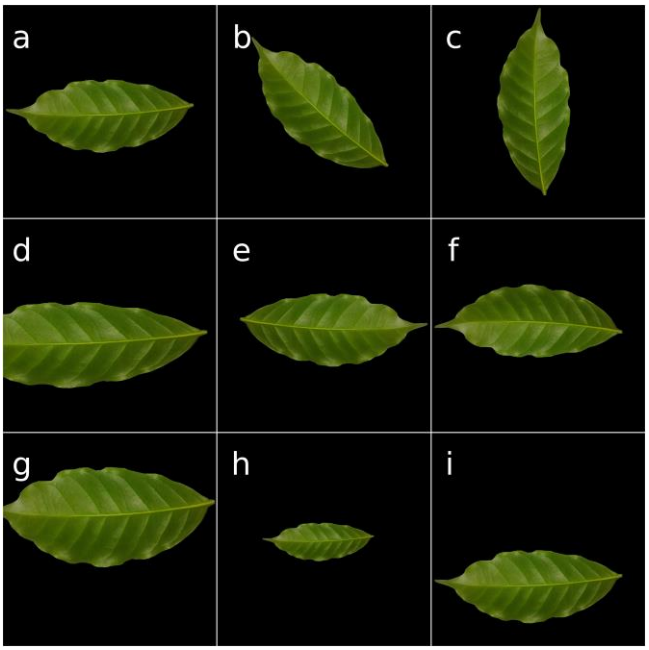
Já as transformações espaciais correspondem a um conjunto de transformações que envolvem alterações na localização dos objetos na cena, ou seja, torna necessária a transformação dos conjuntos descritores da imagem (como caixas delimitadoras e máscaras de segmentação). Dentre as transformações espaciais mais comuns, destacam-se

(ALBUMENTATIONS, 2022):

- Rotação: princípio de girar a imagem em um determinado ângulo;
- Recorte: processo de selecionar uma parte específica da imagem e remover o restante;
- Translação: movimentação da imagem em torno dos eixos *x* e *y*;
- Espelhamento: inversão horizontal ou vertical da imagem;
- Perspectiva: transformação da imagem de forma que ela pareça ter sido tirada de um ângulo diferente do original;
- Escala: aplicação de efeito de zoom à imagem.
- Distorção: deformação de regiões ou linhas da imagem.

A Figura 2 apresenta alguns exemplos das transformações espaciais citadas, aplicadas a uma mesma imagem de folha de café pertencente ao *dataset* Esgario *et al.* (2020).

Figura 2: Exemplos de transformações espaciais. (a) original; (b) recorte; (c) distorção; (d) espelhamento; (e) perspectiva; (f) rotação 90º; (g) rotação 45; (h) escala; (i) translação



Fonte: Própria (2022).

Redes Neurais Artificiais

O aprendizado profundo consiste no princípio de programar um computador de modo que ele se torne capaz de adquirir conhecimento de forma automática e de detectar padrões escondidos em um conjunto de informações (SHALEV-SHWARTZ; BEN-DAVID, 2014). As soluções de aprendizado profundo demonstram potencial na descoberta de padrões intrínsecos

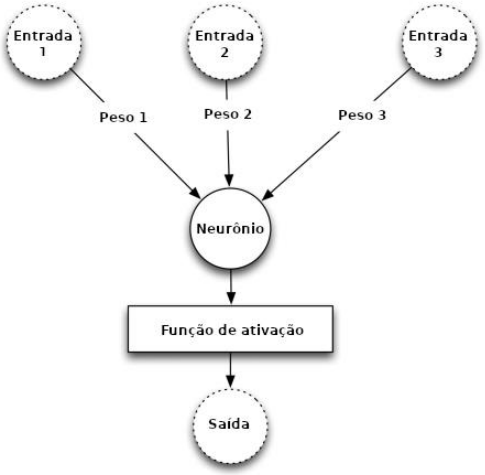
mesmo em dados complexos e de alta dimensionalidade (LECUN *et al.*, 2015).

Essencialmente, o aprendizado profundo se baseia no uso de redes neurais profundas, que podem ser definidas como redes neurais que apresentam em sua arquitetura mais de duas camadas ocultas (HEATON, 2015). É justamente a organização da rede em várias camadas hierárquicas interconectadas que permite a compreensão e extração dos recursos dos dados em altos níveis de complexidade (VASILEV *et al.*, 2019).

Uma rede neural artificial (RNA) é um modelo computacional inspirado na maneira como o cérebro humano funciona. O cérebro é um sistema de processamento de dados complexo, paralelo e não linear. Ele consegue organizar suas estruturas básicas, os neurônios, para processar certos tipos de informação bem mais rapidamente do que qualquer computador já construído (HAYKIN, 2005). Partindo de um ponto de vista matemático, uma rede neural é um grafo direcionado em que os nós correspondem aos neurônios e as arestas correspondem às ligações entre eles (SHALEV-SHWARTZ; BEN-DAVID, 2014).

O neurônio artificial é a unidade básica de processamento de informações dentro de uma RNA (HAYKIN, 2005). Ele multiplica cada valor de entrada que recebe por um peso sináptico, responsável por atribuir diferentes níveis de importância a cada atributo de entrada. Posteriormente, esses valores são somados e aplicados a uma função de ativação, responsável por restringir a amplitude de saída do neurônio (HEATON, 2015). A Figura 3 apresenta a estrutura básica de um neurônio artificial junto de seus principais componentes.

Figura 3: Estrutura básica e componentes principais de um neurônio artificial



Fonte: Adaptado de Heaton (2015).

Uma RNA pode apresentar uma vasta quantidade de neurônios em sua arquitetura, organizados ao longo de uma rede formada por várias camadas conectadas. A camada de entrada é a que recebe os dados de treinamento, ou seja, o estado inicial do sistema. A camada de saída é a responsável por apresentar os resultados obtidos durante o processamento. As camadas extras localizadas entre a entrada e a saída do modelo são chamadas camadas ocultas (VASILEV *et al.*, 2019).

Pode-se simplificar uma RNA como uma aproximação de função matemática com certo nível de erro. Treinar a rede significa fazer pequenas alterações nos pesos das conexões entre os neurônios visando minimizar o erro (VASILEV *et al.*, 2019). Se os recursos computacionais disponíveis para o treinamento dos modelos fossem infinitos, bastaria esgotar todas as possibilidades existentes para as configurações de pesos da rede até encontrar aquela que apresente o menor erro. No entanto, como os recursos são limitados, é necessário utilizar uma abordagem mais inteligente para o problema, já que mesmo redes extremamente simples podem apresentar uma grande quantidade de combinações de pesos (HEATON, 2015).

Um dos métodos mais utilizados para o treinamento de redes neurais é o *backpropagation*, um tipo especial de algoritmo de gradiente descendente (HEATON, 2015). De forma geral, o algoritmo pode ser dividido em dois momentos: a fase *forward*, que consiste na multiplicação dos pesos de cada neurônio pelos valores da entrada e aplicação da função de ativação, da primeira até a última camada da rede; e a fase *backward*, que utiliza o erro gerado durante a fase anterior para fazer pequenas alterações nos pesos de modo a reduzir o erro geral da aproximação da função (FACELI *et al.*, 2011).

Redes Neurais Convolucionais

As Redes Neurais Convolucionais (CNNs), são um tipo especial de rede neural profunda. Sua aplicação está normalmente associada com tarefas que envolvem processamento de imagens digitais, tais como reconhecimento e detecção de objetos, já que esse tipo de rede funciona muito bem com dados organizados em uma topologia de grade, ainda mais quando existe uma forte dependência espacial entre seus componentes (AGGARWAL, 2018). As CNNs utilizam um tipo especial de estrutura conhecida como camada de convolução, que possibilita a extração de recursos da imagem, tais como bordas, cores, manchas e outros elementos visuais (HEATON, 2015).

A U-Net é uma rede neural convolucional profunda proposta por Ronneberger et al.

(2015). Foi inicialmente projetada para a tarefa de segmentação semântica para aplicação em imagens biomédicas, embora seja atualmente aplicada nas mais diversas áreas de segmentação. A arquitetura geral da U-Net é baseada em uma estrutura em formato de "U" com duas fases principais: contração, responsável por extrair as características e recursos da imagem, enquanto simplifica cada vez mais sua representação; e expansão, encarregada de reconstruir a imagem de forma segmentada (RONNEBERGER *et al.*, 2015).

Métricas de Desempenho

Diversas métricas para avaliação da acurácia dos modelos de segmentação foram propostas pelos pesquisadores da área ao longo dos anos. A intersecção sobre união (IoU) mede a quantidade de pixels comuns entre a segmentação correta e o resultado do modelo em relação ao total de pixels compartilhados entre ambos os casos. É uma das métricas de avaliação de desempenho que mais se destaca na literatura devido a sua capacidade de representatividade aliada a simplicidade (GARCIA-GARCIA *et al.*, 2017).

A precisão é uma métrica responsável por descrever o quão bom um modelo é na tarefa de identificar corretamente os objetos em uma imagem. Já a cobertura é uma boa métrica para indicar a capacidade do modelo de identificar todos os casos relevantes. Por conta da natureza complementar dessas duas métricas, elas costumam ser combinadas em uma única métrica chamada de *F-score*, que nada mais é do que a média harmônica entre a precisão e a cobertura (ATIENZA, 2020).

Dados Sintéticos

O aprendizado profundo possibilitou uma revolução de alto nível nas mais diversas áreas, com destaque para a visão computacional e o processamento de linguagem natural. Apesar do potencial, os modelos de aprendizado profundo apresentam uma séria limitação: dependem de uma enorme quantidade de dados anotados para a otimização de seus parâmetros, até mesmo para compreender conceitos relativamente simples (MELO *et al.*, 2022). Segundo Nikolenko (2021), as fases de obtenção e anotação dos dados para o treinamento de modelos de aprendizado profundo podem representar cerca de 80% do tempo gasto com qualquer projeto real da área.

Uma possível solução para o problema é a geração de conjuntos de dados sintéticos. A técnica consiste na criação de exemplos artificiais que buscam imitar ou, pelo menos, se

aproximar de conjuntos de dados reais. Exemplos sintéticos costumam ser muito mais rápidos de serem obtidos do que exemplos reais, além de serem inesgotáveis e pré-anotados. O uso de dados sintéticos também pode ajudar a prevenir dilemas éticos, como no caso de dados sigilosos, ou ser útil em situações em que a coleta pode ser impraticável ou envolver questões de segurança (MELO *et al.*, 2022).

Fawaz *et al.* (2018) comprovaram a partir de uma série de experimentos realizados que a utilização de conjuntos de dados sintéticos como uma forma de *data augmentation*, ou seja, de aumentar a quantidade de exemplares disponíveis para o treinamento dos modelos de aprendizado profundo, consegue proporcionar uma melhoria significativa dos resultados obtidos, principalmente considerando o desempenho com dados que ainda não foram vistos pela rede.

Uma abordagem adequada para a geração de conjuntos de dados sintéticos é a fusão de diferentes fontes de dados, a partir da sobreposição de objetos na cena (composição de imagem). A técnica baseia-se na combinação de objetos, como pessoas ou animais, sob diferentes configurações e fundos, garantindo uma maior variabilidade das amostras (MELO et al., 2022).

METODOLOGIA

Ao longo de todo o desenvolvimento do projeto, optou-se pela utilização do *Python*, uma linguagem de programação de alto nível, interpretada, dinâmica, multiplataforma e de código aberto, amplamente utilizada para a construção de soluções de inteligência artificial, aprendizado de máquina e ciência de dados. O ambiente principal adotado foi o *Google Colaboratory*, mais conhecido como *Colab*, um ambiente on-line e interativo que permite escrever código Python direto no navegador. O Colab também possibilita o uso de *Graphics Processing Unit* (GPUs) e *Tensor Processing Unit* (TPUs), capazes de acelerar bastante o processo de treinamento de redes neurais profundas.

O *Tensorflow* é um ecossistema flexível, abrangente e *open source* de bibliotecas para a área de aprendizado de máquina que permite a criação e treinamento de redes neurais profundas de forma facilitada e com poucas linhas de código. *Segmentation Models*, desenvolvida por Yakubovskiy (2019), é uma biblioteca *Python* baseada em *Tensorflow* que conta com diversas implementações de redes neurais profundas para a segmentação de imagens, incluindo o modelo U-Net. *Albumentations* é uma biblioteca *Python* eficiente e flexível que oferece extensa variedade de operações de transformação de imagem, muito utilizada para o

aumento de dados para tarefas de aprendizado profundo.

Bases de Dados

Para a construção do conjunto de dados sintéticos, foi necessário buscar por bases de dados públicas com imagens que pudessem ser utilizadas como novo plano de fundo para as imagens sintéticas geradas. Dessa forma, diversos bancos de imagens disponíveis na Internet foram utilizados, tais como Pexels, Unsplash, IStock, Pixabay, FreeImages, Burst e o Repositório Digipathos. Desses conjuntos, foram selecionadas 200 imagens que apresentam características relacionadas com imagens de campo, tais como terra, grama, árvores, plantações e outras. A Figura 4 apresenta alguns exemplos das imagens utilizadas.

Figura 4: Exemplos usados como plano de fundo das imagens sintéticas

Fonte: Adaptado de Barbedo et al. (2018).

Quanto às imagens de folhas de café, o primeiro conjunto de dados utilizado foi produzido por Parraga-Alava et al. (2019) e possui 1.560 imagens de café da espécie conilon com resoluções variando de 1280×720 a 2048×1152 pixels. O conjunto contém imagens de folhas saudáveis e outras com um ou mais sintomas de ferrugem e ácaro vermelho. As fotos foram capturadas no campo em condições reais, incluindo variações de luminosidade (manhã e tarde, dias ensolarados e nublados) e diferentes planos de fundo (outras plantas, terra, ervas daninhas). A Figura 5a apresenta um exemplo de imagem desse conjunto de dados.

A segunda base de dados usada foi desenvolvida por Esgario *et al.* (2020) e conta com 1.747 imagens de café da espécie arábica com resolução de 2048×1024 pixels. O *dataset* contém folhas saudáveis e outras apresentando um ou mais sintomas de bicho mineiro, cercospora, phoma e ferrugem. As imagens foram capturadas em laboratório e com uma variedade de câmeras e em diferentes épocas do ano para garantir maior variabilidade dos dados. A Figura 5b apresenta um exemplo de imagem desse conjunto de dados.

O terceiro *dataset* utilizado foi criado por Silva *et al.* (2020) e conta com um total de 539 imagens de café da espécie arábica com resolução de 4000×2250 pixels. As folhas

apresentam sintomas de bicho mineiro e de ferrugem em diversos estágios de contaminação. Todas as imagens foram coletadas com uma câmera de smartphone e em ambiente de laboratório. A Figura 5c apresenta um exemplo de imagem desse conjunto de dados.

Figura 5: Exemplos dos conjuntos de dados de folhas de café

Fonte: Adaptado de Parraga-Alava et al. (2019), Esgario et al. (2020) e Silva et al. (2020).

Geração dos Dados Sintéticos

A geração das imagens sintéticas foi feita utilizando a biblioteca *Albumentations* e as imagens de laboratório dos *datasets* de Esgario *et al*. (2020) e Silva *et al.* (2020). O objetivo foi, a partir de imagens tiradas em condições controladas de fundo e iluminação, obter exemplares com características mais próximas das encontradas em imagens reais de campo, para aumentar a quantidade de dados relevantes disponíveis para o treinamento do modelo.

Primeiramente, foi realizado o ajuste das 2.286 imagens de folhas de café, processo que consistiu na adição de borda às imagens para que sua resolução corresponda ao tamanho alvo do *dataset* sintético (2.048×2.048), e aplicação de transparência ao fundo utilizando o canal alfa e as máscaras de segmentação. Depois, foi realizado o ajuste das 200 imagens do fundo, o que consistiu na padronização da resolução para o intervalo de 2.048×2.048 até 3.072×3.072, visando dar flexibilidade para que o fundo apresente pequenas mudanças mesmo quando a mesma imagem é utilizada mais de uma vez.

Por fim, foi feita a geração das imagens sintéticas a partir dos dois conjuntos pré-processados de folhas e fundos. As etapas de geração incluem:

- Etapa 1: uma imagem de fundo é escolhida aleatoriamente, lida e removida do conjunto, adicionada novamente quando todas as imagens foram escolhidas;
- Etapa 2: aplicam-se transformações espaciais na imagem selecionada na Etapa 1;
- Etapa 2.1: recorte aleatório para o tamanho alvo (*2048 × 2.048*);
- Etapa 2.2: espelhamento horizontal e vertical, distorção (até 5%) e desfoque gaussiano (tamanho do *kernel* entre 5 e 7), com 50% de chance de aplicar cada transformação;
- Etapa 3: uma quantidade aleatória de imagens no intervalo de 2 a 6 é escolhida, lida e

removida do conjunto total de 2.286 folhas, adicionada novamente quando todas as imagens foram escolhidas;

- Etapa 4: aplicam-se transformações espaciais em cada uma das imagens da Etapa 3, incluindo rotação (-180º a 180º), escala (60% a 90% do tamanho original), translação (até 50% do tamanho da imagem), perspectiva (-30º a 30º), espelhamento horizontal e vertical, cada transformação com 50% de chance de ser aplicada;
- Etapa 5: as imagens resultantes da Etapa 4 são combinadas com a imagem de fundo resultante da Etapa 2, gerando uma nova imagem sintética.

A Figura 6 apresenta um exemplo de imagem sintética gerada.

Figura 6: Exemplo de imagem gerada (esquerda) e sua máscara de segmentação (direita)

Fonte: Própria (2022).

Para garantir uma representatividade ainda maior do conjunto de imagens sintéticas, foram definidas mais uma série de transformações espaciais e de pixel aplicadas às imagens resultantes da geração do conjunto sintético. Cada operação teve uma chance de 40% de ser aplicada. As transformações incluem:

- Brilho, contraste e saturação (variação de até 20% do valor original);
- Equalização adaptativa;
- Ruído Gaussiano (variação no intervalo de 10 a 50 dos valores originais);
- Realce de contornos (alfa 20% a 50%);
- Espelhamento horizontal e vertical;
- Escala (aumento ou redução de até 10% do tamanho original);
- Perspectiva (até 10º).

Experimentos Propostos

O modelo de aprendizado profundo selecionado foi o U-Net, por tratar-se de uma rede

consolidada e que apresenta um treinamento mais rápido do que outros modelos citados na literatura. Como CNN pré-treinada, optou-se pela utilização da ResNet-50, principalmente por conta de seu bom desempenho e baixo consumo de memória. A utilização de CNNs pré-treinadas é uma prática muito comum no treinamento de redes neurais profundas, relacionada com o conceito de transferência de aprendizado, em que uma rede treinada em um conjunto de dados A é utilizada para agilizar e melhorar a generalização em um conjunto de dados B.

A função de perda utilizada foi a *focal loss*, proposta por Lin et al. (2017), uma generalização da entropia cruzada binária que busca reduzir o efeito de desbalanceamento dos dados de treinamento. Além disso, as demais configurações de treinamento implementadas incluem a resolução das imagens em 512×512, taxa de aprendizado de 10^{-4}, tamanho do lote igual a 16 e 100 épocas de treinamento. Optou-se ainda pelo descongelamento de todos os pesos treináveis da rede, incluindo tanto os parâmetros da CNN pré-treinada (ResNet-50) quanto os do modelo de segmentação (U-Net). Foram utilizadas 64 imagens para validação e 64 para teste do dataset de Parraga-Alava et al. (2019) divididas igualmente para os conjuntos de teste e validação.

Visando testar a eficiência do conjunto de dados sintéticos e comparar seu desempenho com a utilização de imagens reais, foram propostos 3 experimentos:

- Experimento 1: treinamento utilizando 510 imagens reais de folhas de café do *dataset* Parraga-Alava *et al.* (2019);
- Experimento 2: treinamento utilizando 571 imagens sintéticas geradas conforme o descrito na seção anterior;
- Experimento 3: treinamento utilizando a junção dos dois conjuntos de dados dos Experimentos 1 e 2, totalizando 1.081 imagens.

RESULTADOS E DISCUSSÃO

Após a definição dos materiais e da metodologia adotada, a etapa seguinte consistiu na implementação dos experimentos propostos. A Tabela 1 ilustra os resultados obtidos. O Experimento 3, que reúne tanto as imagens reais de campo quanto às imagens sintéticas geradas utilizando as definições mencionadas no capítulo de metodologia, obteve os melhores resultados dos testes, com 1,3 pontos percentuais (pp) de IoU e 0,7 pp de *F-score* a mais do que o Experimento 1. Analisando as métricas de precisão e cobertura, percebe-se que, embora o Experimento 3 tenha ficado ligeiramente menos preciso (0,2 pp) do que o Experimento 1, ele

ficou mais sensível na detecção das folhas presentes nas imagens (1,78 pp).

Tabela 1: Resultados obtidos no conjunto de experimentos propostos

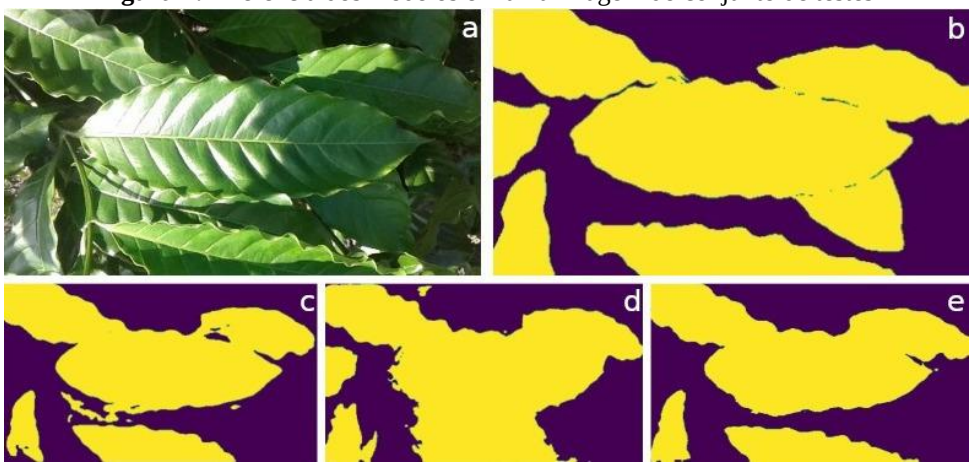
Experimento	IoU	Precisão	Cobertura	*F-score*
1	0,85995	0,94533	0,90563	0,92204
2	0,81735	0,90977	0,89173	0,89618
3	0,87371	0,94295	0,92343	0,92993

Fonte: Própria (2022).

O Experimento 2 apresentou os piores resultados entre os experimentos propostos (4,2 pp de IoU e 2,5 pp de *F-score* a menos que o Experimento 1). Entretanto, considerando que seu treinamento foi feito unicamente usando imagens sintéticas geradas com *datasets* de laboratório, tiradas sob condições controladas de fundo e iluminação, e que seu desempenho foi avaliado em imagens de campo em condições reais, seu resultado pode até ser considerado quando a quantidade de dados reais disponíveis não for suficiente para o treinamento.

A Figura 7 apresenta uma imagem do conjunto de testes (a), sua máscara verdadeira (b) e as máscaras geradas a partir dos modelos treinados nos experimentos 1 (c), 2 (d) e 3 (e). Por ela, é possível perceber que a inferência que mais se aproxima da máscara verdadeira (b) é a gerada no Experimento 3 (e), seguida pelo Experimento 1 (c) e pelo Experimento 2 (d).

Figura 7: Inferência dos modelos em uma imagem do conjunto de testes



Fonte: Própria (2022).

CONSIDERAÇÕES FINAIS

Este trabalho teve por objetivo principal desenvolver e testar uma abordagem de geração de imagens sintéticas utilizando a composição de objetos na cena, com a finalidade de expandir a quantidade de dados representativos disponíveis para o treinamento de modelos de

aprendizado profundo a partir de conjuntos de imagens mais simples. Como estudo de caso, foram utilizadas imagens de laboratório de folhas de café para o treinamento de um modelo de segmentação testado em imagens reais de campo.

O conjunto sintético desenvolvido se mostrou capaz de melhorar os resultados do modelo de segmentação quando utilizado com o conjunto de imagens reais, incrementando o desempenho do modelo em 1,3 pp de IoU e 0,7 pp de *F-score*, em comparação com o modelo treinado apenas com imagens reais. A utilização das imagens sintéticas, além de colaborar para o aumento na quantidade de dados disponíveis para o treinamento, fornece mais exemplos e situações únicas de combinações de folhas e fundos, contribuindo particularmente com a sensibilidade do modelo em localizar todas as folhas presentes nas imagens.

Por outro lado, a utilização apenas do conjunto sintético para o treinamento não conseguiu resultados tão satisfatórios, ficando 4,2 pp de IoU e 2,5 pp de *F-score* atrás do modelo treinado com as imagens reais. Isso pode indicar que a abordagem utilizada para a geração das imagens sintéticas não conseguiu representar todas as características de exemplares reais de folhas de café, indicando que mais testes e experimentos são necessários, visando melhorar a representatividade do conjunto.

Dessa forma, a utilização de imagens sintéticas geradas a partir das técnicas de combinação de objetos na cena apresenta o potencial de melhorar o desempenho de modelos de aprendizado profundo, principalmente quando a quantidade de dados reais disponíveis para o treinamento é limitada. As folhas segmentadas geradas a partir do modelo de aprendizado profundo treinado neste trabalho também podem contribuir para a melhoria dos resultados em outras tarefas de visão computacional, como na classificação ou detecção de pragas e doenças.

REFERÊNCIAS

AGGARWAL, Charu C. Neural Networks and Deep Learning. [S.l.]: Springer, 2018. ISBN 978-3-319-94463-0.

ALBUMENTATIONS. Welcome to Albumentations documentation. 2022. Disponível em: https://albumentations.ai/docs/.

ALOM, Md Zahangir; TAHA, Tarek M.; YAKOPCIC, Chris; WESTBERG, Stefan; SIDIKE, Paheding; NASRIN, Mst Shamima; HASAN, Mahmudul; ESSEN, Brian C. Van; AWWAL, Abdul A.S.; ASARI, Vijayan K. A state-of-the-art survey on deep learning theory and architectures. [S.l.]: MDPI AG, 2019.

ARSENOVIC, Marko; KARANOVIC, Mirjana; SLADOJEVIC, Srdjan; ANDERLA, Andras;

STEFANOVIC, Darko. Solving current limitations of deep learning based approaches for plant disease detection. Symmetry, v. 11, 2019. ISSN 20738994.

ATIENZA, Rowel. Advanced Deep Learning with TensorFlow 2 and Keras. 2. ed. Birmingham, UK: Packt Publishing, 2020. 512 p. ISBN 9781838821654.

BARBEDO, Jayme Garcia Arnal. Plant disease identification from individual lesions and spots using deep learning. Biosystems Engineering, v. 180, 2019. ISSN 15375110.

BARBEDO, J. G. A.; KOENIGKAN, L. V.; HALFELD-VIEIRA, B. A.; COSTA, R. V.; NECHET, K. L.; GODOY, C. V. Annotated plant pathology databases for image-based detection and recognition of diseases. IEEE Latin America Transactions, v. 16, p. 1749–1757, 6 2018. ISSN 1548-0992.

CARNEIRO Álvaro Leandro Cavalcante; SILVA, Lucas de Brito; FAULIN, Marisa Silveira Almeida Renaud. Rust (hemileia vastatrix) and leaf miner (leucoptera coffeella) in coffee crop (coffea arabica). Mendeley Data, v. 5, 2020.

CARVALHO, Cleidisson Nunes de; OLIVEIRA, Ykaro Richard; SILVA, Paulo Henrique da; ABREU, Maria Carolina de. Coffea arabica l.: potencialidades e ações medicinais. Revista Intertox de Toxicologia, Risco Ambiental e Sociedade, v. 11, 2018. ISSN 1984-3577.

ESGARIO, José G.M.; KROHLING, Renato A.; VENTURA, José A. Deep learning for classification and severity estimation of coffee leaf biotic stress. Computers and Electronics in Agriculture, Elsevier B.V., v. 169, 2 2020. ISSN 01681699.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André C. P. L. F. De. Inteligência Artificial: uma abordagem de aprendizado de máquina. [S.l.]: LTC, 2011. ISBN 978-85-216-1880-5.

FAWAZ, Hassan Ismail; FORESTIER, Germain; WEBER, Jonathan; MULLER, Pierre-Alain. Data augmentation using synthetic data for time series classification with deep residual networks. In: . [S.l.]: arXiv, 2018.

FERRÃO, Romário Gava; FONSECA, Aymbiré Francisco Almeida da; FERRãO, Maria Amélia Gava; MUNER, Lúcio Herzog de. Café Conilon. 2. ed. [S.l.]: Incaper, 2017. ISBN 978-85-89274-26-5.

GARCIA-GARCIA, Alberto; ORTS-ESCOLANO, Sergio; OPREA, Sergiu; VILLENA-MARTINEZ, Victor; RODRÍGUEZ, José García. A review on deep learning techniques applied to semantic segmentation. CoRR, abs/1704.06857, 2017. Disponível em: http://arxiv.org/abs/1704.06857.

GONZALEZ, Rafael C; WOODS, Richard E. Digital Image Processing. 4. ed. Pearson Education, 2018. ISBN 9780133356724. Disponível em: www.pearsoned.com/.

HEATON, Jeff. Artificial intelligence for humans. 1. ed. [S.l.]: Heaton Research, 2015. v. 3. ISBN 978-1505714340.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. Nature, Nature Publishing Group, v. 521, p. 436–444, 5 2015. ISSN 14764687.

LIN, Tsung-Yi; GOYAL, Priya; GIRSHICK, Ross; HE, Kaiming; DOLLáR, Piotr. Focal loss for dense object detection. 8 2017. Disponível em: http://arxiv.org/abs/1708.02002.

MELO, Celso M. de; TORRALBA, Antonio; GUIBAS, Leonidas; DICARLO, James; CHELLAPPA, Rama; HODGINS, Jessica. Next-generation deep learning based on simulators and synthetic data. Trends in Cognitive Sciences, Elsevier Ltd, v. 26, p. 174–187, 2 2022. ISSN 1879307X.

NIKOLENKO, Sergey I. Synthetic Data for Deep Learning. Springer Optimization and Its Applications, 2021. Disponível em: http://www.springer.com/series/7393.

OLIVEIRA, C. M.; AUAD, A. M.; MENDES, S. M.; FRIZZAS, M. R. Crop losses and the economic impact of insect pests on brazilian agriculture. Crop Protection, v. 56, p. 50–54, 2 2014. ISSN 02612194.

ORGANIZATION, International Coffee. Annual review: coffee year 2019/2020. 2020. Disponível em: www.ico.org/documents/cy2020-21/annual-review-2019-2020-e.pdf.

PARRAGA-ALAVA, Jorge; CUSME, Kevin; LOOR, Ang Elica; SANTANDER, Esneider. Rocole: A robusta coffee leaf images dataset for evaluation of machine learning based methods in plant diseases recognition. Mendeley Data, v. 2, 2019.

PATKI, Neha; WEDGE, Roy; VEERAMACHANENI, Kalyan. The synthetic data vault. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2016. p. 399–410. ISBN 9781509052066.

PEDRINI, Hélio; SCHWARTZ, William. R. Análise de imagens digitais: princípios, algoritmos e aplicações. [S.l.]: Cengage Learning, 2007. ISBN 9788522128365.

RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-net: Convolutional networks for biomedical image segmentation. In: . [S.l.]: Springer Verlag, 2015. v. 9351, p. 234–241. ISBN 9783319245737. ISSN 16113349.

SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. Understanding Machine Learning. [S.l.]: Cambridge University, 2014. ISBN 978-1-107-05713-5.

SOLOMON, Chris; BRECKON, Toby. Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab. 1. ed. [S.l.]: Wiley-Blackwell, 2011. ISBN 9780470689776.

VASILEV, Ivan; SLATER, Daniel; SPACAGNA, Gianmario; ROELANTES, Peter; ZOCCA, Valentino. Python deep learning: exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow. 2. ed. [S.l.]: Packt Publishing, 2019. ISBN 978-1-78934-846-0.

XIONG, Yonghua; LIANG, Longfei; WANG, Lin; SHE, Jinhua; WU, Min. Identification of cash crop diseases using automatic image segmentation algorithm and deep learning with expanded dataset. Computers and Electronics in Agriculture, v. 177, 2020. ISSN 01681699.

YAKUBOVSKIY, Pavel. Segmentation Models. 2019. Disponível em: https://github.com/qubvel/segmentation_models.