Detecção e Segmentação de Danos Físicos em Maçãs Utilizando Mask R-CNN

Igor da Silva Vieira Benato[1], **José Luiz Seixas Junior**[1]

[1]Ciência da Computação
Universidade Estadual do Paraná (UNESPAR)
Apucarana – Paraná

1. Introdução

Este trabalho aborda as limitações inerentes à inspeção de qualidade de maçãs, contextualizando-as frente às normas estabelecidas pelo governo brasileiro na Instrução Normativa N° 5/2006 [1]. Diante dos recentes avanços em Visão Computacional, propõe-se a utilização de Redes Neurais Convolucionais (CNNs), dada a sua robustez na extração de características. Especificamente, adota-se o modelo *Mask R-CNN*, reconhecido por sua eficácia na detecção de objetos e segmentação pixel a pixel, visando a identificação precisa de danos físicos nos frutos para otimizar o processo de controle de qualidade.

2. Fundamentação Teórica

A fundamentação teórica revisita os conceitos de Redes Neurais Artificiais (RNAs), traçando uma linha evolutiva desde o Perceptron de Rosenblatt [2] até os avanços subsequentes. Destaca-se a transição para os *Multi-Layer Perceptrons* (MLPs), que solucionaram a incapacidade de separar dados não-lineares, uma limitação crítica dos modelos iniciais, conforme discutido por Rumelhart et al. [3]. Entretanto, o trabalho ressalta as limitações das MLPs quanto à escalabilidade, visto que o aumento da dimensão dos dados de entrada acarreta um crescimento excessivo no número de pesos a serem ajustados, inviabilizando o treinamento eficiente.

Como solução para as limitações de processamento das arquiteturas anteriores, consolidaram-se as Redes Neurais Convolucionais (CNNs). Diferentemente dos modelos clássicos, as CNNs preservam a estrutura espacial dos dados de entrada (formato de matriz), o que as torna ideais para o processamento de imagens [4].

A principal inovação reside nas camadas de convolução, responsáveis pela extração automática e hierárquica de características. Isso permite uma redução drástica na quantidade de parâmetros a serem aprendidos em comparação a uma rede totalmente conectada, tornando o treinamento computacionalmente viável e mais rápido [5]. Ao final do processo, os mapas de características gerados pelas convoluções são vetorizados e inseridos em camadas densas (totalmente conectadas) para realizar a classificação final.

O trabalho aprofunda-se na análise das *backbones*, estruturas responsáveis pela extração de mapas de características robustos. Discute-se o desafio inerente ao treinamento de redes profundas, especificamente o problema do desvanecimento do gradiente (*vanishing gradient*), que degrada a precisão do modelo, e a solução proposta com as Redes Residuais.

Sequencialmente, o trabalho descreve a evolução das arquiteturas de detecção de objetos baseadas em regiões. Inicia-se com o R-CNN de Girshick et al. [6], pioneiro na

aplicação de CNNs sobre propostas de regiões para classificação. Avança-se para o *Fast R-CNN* de Girshick et al. [7], que otimizou o desempenho computacional ao processar a imagem inteira de uma única vez, gerando um mapa de características compartilhado, diferentemente de sua antecessora que processava cada região individualmente.

A evolução seguiu com o *Faster R-CNN* de Ren et al. [8], que introduziu a *Region Proposal Network* (RPN), integrando a geração de propostas à própria rede neural. Por fim, apresenta-se o *Mask R-CNN* de He et al. [9], modelo central deste estudo, que estende essa arquitetura ao adicionar um ramo paralelo para a segmentação pixel a pixel e aprimora a preservação espacial através da camada *RoI Align*.

Na revisão de trabalhos correlatos, destacam-se quatro pesquisas principais que corroboram a abordagem deste estudo. Inicialmente, analisa-se o trabalho de Hou et al. [10], que empregou a *Faster R-CNN* para a detecção de danos sutis por impacto em maçãs. O autor utilizou imagens hiperspectrais, evidenciando a complexidade da detecção desses danos em imagens convencionais (*RGB*). Contudo, o estudo limitou-se à detecção via *bounding boxes*, sem realizar a segmentação pixel a pixel da área afetada.

Em contrapartida, o trabalho de El Akrouchi et al. [11] enfatiza a robustez da *Mask R-CNN* em cenários complexos, aplicando o modelo para detectar, segmentar e contabilizar panículas de quinoa. Vale ressaltar que foram utilizadas fotografias comuns (*RGB*), demonstrando a viabilidade dessa modalidade de imagem. O estudo comparou três *backbones* distintos: *ResNet-50*, *ResNet-101* e *EfficientNet-B7*, obtendo bons resultados com a *ResNet-101*, arquitetura também adotada nesta pesquisa.

Na sequência, destaca-se o estudo de Osorio et al. [12], que também utiliza a *Mask R-CNN*, mas com foco na técnica de *Transfer Learning* utilizando pesos do *dataset* COCO [13]. A pesquisa comprova a eficácia do treinamento em *datasets* pequenos (sendo o maior com 327 imagens) e estabelece três conclusões vitais: o uso de *Transfer Learning* melhora os resultados finais; é possível obter boa performance mesmo sem aumento de dados (*Data Augmentation*); e a qualidade das anotações das máscaras impacta mais o desempenho do que a quantidade bruta de dados.

Por fim, examina-se o trabalho de Zhang et al. [14], que aplica a *Mask R-CNN* para fenotipagem de alfaces. Diferentemente do anterior, este estudo destaca o uso intensivo de *Data Augmentation* para contornar a escassez de dados, aplicando variadas transformações nas imagens. Além disso, utilizou-se a validação cruzada *K-fold Cross-Validation* (com $k = 5$), método que consiste em dividir o conjunto de dados em k partes, utilizando $k-1$ partes para treino e 1 parte para teste em rodadas alternadas, maximizando o uso dos dados e evitando vieses na avaliação.

3. Método de Pesquisa

A metodologia de pesquisa fundamentou-se na aquisição de imagens digitais por meio de fotografia, compondo o acervo inicial do estudo. As imagens brutas foram submetidas a uma etapa de pré-processamento, consistindo no redimensionamento das imagens.

O *dataset* original foi constituído por 300 imagens. O processo de anotação e segmentação das regiões de interesse (ground truth) foi realizado manualmente utilizando uma ferramenta própria no formato *VGG Image Annotator (VIA)*, definindo-se uma classe única para a identificação dos danos físicos.

Visando aumentar a capacidade de generalização do modelo e mitigar o risco de *overfitting*, aplicou-se a técnica de *Data Augmentation* de forma híbrida. O processo combinou variações manuais durante a captura fotográfica e uma expansão artificial gerada via algoritmo, que executou operações geométricas de rotação e inversão (espelhamento) nas imagens. Esse procedimento resultou em uma expansão de $5\times$ sobre o conjunto original, totalizando um *dataset* final de 1800 imagens. Para a avaliação rigorosa do desempenho do modelo, adotou-se o método de validação cruzada *K-fold Cross-Validation*, assegurando que todos os dados fossem utilizados tanto para treino quanto para validação em diferentes iterações.

Para a configuração do treinamento, adotou-se uma taxa de aprendizado (*learning rate*) dinâmica, estabelecida em 0.001 tanto para as etapas com congelamento de camadas (*freezing*) quanto para o ajuste fino sem congelamento. A inicialização dos modelos utilizou a técnica de *Transfer Learning*, carregando pesos pré-treinados no *dataset* COCO.

A função de perda global (L) utilizada para monitorar o desempenho é multitarefa, composta pela somatória das perdas da *Region Proposal Network* (RPN) e das cabeças da *Mask R-CNN* (classificação, regressão da caixa delimitadora e máscara), conforme expresso na Equação 1:

$$L = L_{RPN_{cls}} + L_{RPN_{box}} + L_{MRCNN_{cls}} + L_{MRCNN_{box}} + L_{MRCNN_{mask}} \tag{1}$$

Aqui está a redação final para a seção de métricas, removendo as imagens e mantendo o rigor acadêmico com as equações.

Na seção de métricas de avaliação, definiram-se três indicadores principais para mensurar a eficácia do modelo. Primeiramente, utilizou-se a Average Precision (AP) para analisar o desempenho global sobre a curva de Precisão x Revocação (Precision-Recall). Essa métrica avalia a capacidade do modelo em realizar detecções corretas, minimizando a ocorrência de falsos positivos, ao mesmo tempo em que maximiza a detecção de todos os danos presentes na imagem (alta revocação).

Para quantificar a qualidade da segmentação espacial, adotou-se o Intersection over Union (IoU). Esta métrica determina a porcentagem de sobreposição correta entre a máscara predita pelo modelo e a anotação real (ground truth). O cálculo é dado pela razão entre a área de interseção e a área de união das duas regiões, conforme a Equação 2:

$$IoU = \frac{\text{Área da Interseção}}{\text{Área da União}} = \frac{A_{pred} \cap A_{real}}{A_{pred} \cup A_{real}} \tag{2}$$

Por fim, aplicou-se o F1-Score, uma métrica que fornece a média harmônica entre a Precisão e a Revocação. O F1-Score é essencial para indicar o equilíbrio do sistema, penalizando modelos que possuem uma dessas métricas muito baixa em detrimento da outra, conforme expresso na Equação 3:

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3}$$

4. Experimentos

Os experimentos foram estruturados em três etapas distintas para avaliar diferentes estratégias de treinamento. No primeiro experimento, aplicou-se a validação cruzada *K-fold Cross-Validation* com $k = 5$, utilizando pesos pré-treinados no *dataset* COCO. Nesta etapa, aplicou-se o congelamento (*freezing*) de todas as camadas de extração de características da rede, mantendo apenas a última camada (cabeça de predição) treinável. O objetivo foi verificar se o ajuste restrito exclusivamente à camada final seria suficiente para que o modelo detectasse os danos físicos com precisão.

O segundo experimento replicou a metodologia anterior, com a exceção de que o treinamento foi realizado com o descongelamento total da rede (*unfreezing*). Essa etapa visou validar se o retreinamento de todas as camadas apresentaria ganho ou degradação de desempenho em comparação à abordagem com *freezing*.

Por fim, o terceiro experimento foi conduzido de forma isolada utilizando o *dataset* expandido (aumentado), com todas as camadas descongeladas (*unfreezing*). Manteve-se a taxa de aprendizado (*learning rate*) em 0.001, considerando que as etapas por época foram configuradas para percorrer a totalidade das imagens. Essa estratégia foi adotada para adequar o tempo de convergência ao volume superior de dados. Ressalta-se que, diferentemente dos experimentos anteriores, neste experimento não se aplicou a validação cruzada *K-fold Cross-Validation*, devido ao custo computacional inviável para o escopo desse trabalho.

5. Resultados

Os resultados experimentais demonstraram comportamentos distintos entre as estratégias adotadas. No primeiro experimento (*freezing*), observou-se uma boa convergência durante o treinamento, contudo, a etapa de validação apresentou valores de perda total relativamente altos (1.5). Ao realizar a decomposição da função de perda, identificou-se que a principal responsável por essa elevação foi a perda da máscara (L_{mask}), que aumentou gradativamente ao longo das épocas.

Em contrapartida, as demais perdas (classificação e bounding box) não convergiram totalmente, mas estabilizaram-se. Isso indica que o modelo manteve a capacidade de localizar a região do dano corretamente, porém sua capacidade de segmentação pixel a pixel degradou-se conforme o avanço do treinamento.

O segundo experimento (*unfreezing*) apresentou um comportamento de curva semelhante. Embora os valores absolutos de perda na validação, tenham se mostrado ligeiramente inferiores (1.25) em comparação ao primeiro cenário, a tendência de degradação na segmentação persistiu. Por fim, o terceiro experimento (com *dataset* aumentado) apresentou resultados significativamente superiores. A convergência no treinamento ocorreu de forma gradual, similar às etapas anteriores, porém o destaque residiu na perda de validação, que se manteve abaixo de 0.8. Comparativamente, isso representa uma melhoria de aproximadamente 53,3% em relação ao experimento com *freezing* (1.5) e uma redução considerável frente ao experimento sem *freezing* (1.25). Além da redução global, o comportamento dos componentes da perda alterou-se: pela primeira vez, a perda da máscara (L_{mask}) convergiu gradualmente e permaneceu reduzida até as épocas finais, evidenciando que o aumento de dados foi crucial para que o modelo generalizasse corretamente a tarefa de segmentação.

Corroborando a análise do comportamento da função de perda, os resultados das métricas de avaliação quantitativa evidenciam a evolução da capacidade do modelo, conforme detalhado na Tabela 1. O experimento com *Freezing* estabeleceu a linha de base do estudo. Ao aplicar o descongelamento total da rede (*Unfreezing*), observou-se um ganho de desempenho geral nas métricas avaliadas em comparação à etapa inicial.

Contudo, foi o modelo treinado com o *Dataset* Expandido que apresentou os resultados mais expressivos, confirmando a hipótese de que o aumento de dados é crucial para a generalização da rede. Diferentemente dos cenários anteriores, nesta etapa todas as métricas atingiram seus valores máximos. Destaca-se o salto significativo na Average Precision (AP) e no F1-Score, indicando que a rede tornou-se muito mais robusta e equilibrada na detecção dos danos. O mIoU também apresentou uma evolução consistente em relação aos modelos de *freezing* e *unfreezing*, validando que a segmentação da área do dano foi refinada concomitantemente à melhora na detecção global.

Tabela 1. Comparativo dos resultados máximos obtidos nas métricas de avaliação (AP, F1 e mIoU) para os três cenários experimentais.

Experimento	AP@50	F1-Score@50	mIoU
Freezing	0.6953	0.7227	0.6129
Unfreezing	0.7737	0.7899	0.6441
Expanded	**0.8766**	**0.8767**	**0.7059**

Nota: Os valores em negrito indicam o melhor desempenho em cada métrica.

6. Conclusão

Este estudo cumpriu o objetivo de avaliar o desempenho técnico e a utilidade prática da arquitetura Mask R-CNN na segmentação de instâncias de danos físicos. A investigação confirmou que o modelo é capaz não apenas de localizar, mas de delinear a área das avarias, requisito essencial para a mensuração de área exigida pelas normas regulatórias.

Os experimentos demonstraram que a aplicação exclusiva de Transfer Learning em datasets reduzidos foi insuficiente para a precisão da máscara, resultando em divergência na função de perda de segmentação. A estabilização do modelo ocorreu apenas com a introdução do aumento de dados (Data Augmentation). Com a variabilidade espacial ampliada, a rede aprendeu a definir as bordas dos defeitos com exatidão, alcançando métricas superiores a 0.87 em precisão (AP) e F1-Score.

Portanto, a combinação da arquitetura Mask R-CNN com a expansão de dados consolida-se como uma ferramenta eficaz para a segmentação de danos em maçãs, oferecendo uma alternativa robusta para auxiliar no controle de qualidade.

Referências

[1] Brasil. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n.º 5, de 22 de fevereiro de 2006. Aprova o Regulamento Técnico de Identidade e Qualidade da Maçã. Diário Oficial da União: seção 1, Brasília, DF, p. 3, 2006. URL https://sistemasweb.agricultura.gov.br/sislegis/

action/detalhaAto.do?method=visualizarAtoPortalMapa&chave=805793610. Acesso em: 19 nov. 2025.

[2] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

[3] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

[4] Rikiya Yamashita, Masaki Nishio, Robert K Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9:611–629, 2018. doi: 10.1007/s13244-018-0639-9.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016. ISBN 9780262035613. URL http://www.deeplearningbook.org.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[10] Jingli Hou, Yuhang Che, Yanru Fang, Hongyi Bai, and Laijun Sun. Early bruise detection in apple based on an improved faster rcnn model. *Horticulturae*, 10(1): 100, 2024. doi: 10.3390/horticulturae10010100.

[11] Manal El Akrouchi, Manal Mhada, Dachena Romain Gracia, Malcolm J. Hawkesford, and Bruno Gérard. Optimizing mask r-cnn for enhanced quinoa panicle detection and segmentation in precision agriculture. *Frontiers in Plant Science*, 16: 1472688, 2025. doi: 10.3389/fpls.2025.1472688.

[12] Kelly Osorio, Andrés Puerto, Cesar Pedraza, David Jamaica, and David Rodríguez. Mask R-CNN refitting strategy for plant counting and sizing in UAV imagery. *Remote Sensing*, 12(18):3015, 2020. doi: 10.3390/rs12183015.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] Yali Zhang, Zhaolong Wang, Yi Liu, and Jianhua Zhang. Study on utilizing mask R-CNN for phenotypic estimation of lettuce's growth status and optimal harvest timing. *Agronomy*, 14(6):1271, 2024. doi: 10.3390/agronomy14061271.