INTERPRETABILIDADE NA CLASSIFICAÇÃO DE ESTÁDIOS DE MATURAÇÃO PÓS-COLHEITA DE FRUTOS POR MEIO DE IMAGENS E APRENDIZADO PROFUNDO

Palavras-Chave: EXPLAINABLE ARTIFICIAL INTELLIGENCE, FOCUSED ATTENTION, VISION-TRANSFORMERS

Autores:

LÍVIA LIMA AMORIM – FEAGRI, UNICAMP

Profª. Drª. JULIANA APARECIDA FRACAROLLI (orientadora) – FEAGRI, UNICAMP

INTRODUÇÃO:

Frutos são amplamente consumidos e recomendados. Apresentam sabor agradável, benefícios nutracêuticos, vitaminas e minerais. Frutos climatéricos geralmente são colhidos antes do início do estádio climatérico para manter o processo de maturação e evitar a perecibilidade. Caso contrário, ocorre uma redução de seu tempo de prateleira. Assim, métodos de classificação de frutos baseados no amadurecimento são necessários para a comercialização e armazenamento de frutos. Em muitos lugares, sobretudo em países tropicais essa tarefa é realizada de forma manual, que é custosa, demorada, cansativa e susceptível a erros humanos. A classificação de estádios de maturação de frutos tem sido uma fonte de pesquisas, principalmente com aplicação de aprendizado profundo. Com aprendizado de máquina é possível até mesmo substituir a atividade manual. Porém, pesquisas para obter altas acurácias em modelos de aprendizado profundo têm gerado soluções complexas, cujas tomadas de decisão nem sempre são entendidas por humanos. Com o objetivo de apresentar informações para viabilizar a interpretação dos resultados, são adicionadas as técnicas de interpretabilidade. Entendendo a lógica e os recursos usados nas decisões dos modelos, é possível verificar e validar melhor os resultados, melhorar o algoritmo e os dados de treinamento e extração de informação. Assim, neste projeto propomos avaliar e estudar características de imagens oriundas de bancos de imagens disponíveis; aplicar uma técnica de interpretabilidade (Focused Attention (PLAYOUT et al., 2022)). Para classificação dos estádios de maturação, foi empregado vision-trasnformer (ViT). Foi possível identificar partes ou regiões das imagens que mais impactam nas decisões dos modelos. Com mais pesquisas pode ser validado o uso de técnicas de interpretabilidade para essa aplicação e proporcionar a visualização dos principais aspectos, características e variáveis que contribuem para a classificação de estádios de maturação de frutos, com imagens obtidas do Repositório “Fruit Ripening Process Dataset and Model” - Ripening (2022).

METODOLOGIA:

BASE DE DADOS: As imagens foram obtidas do repositório “Fruit Ripening Process Dataset and Model” (RIPENING, 2022). As imagens apresentam dimensões 416 x 416 pixels e após data augmentation é composto por um total de 18074 imagens, sendo 15792 para treino, 1525 para validação e 757 para teste. As classes são: frutos frescos, frutos maturos, maturação completa, frutos em senescência, frutos danificados e frutos imaturos.

Modelagem com Vision Transformer (ViT): Para executar este trabalho foram usados os recursos computacionais do "Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP)." O código criado implementa um Vision Transformer simples com base no trabalho de Dosovitskiy et al. (2020) com atenção focada baseada no artigo de Playout et al. (2022). No mesmo código são gerados logs de saída e erros. Foram importadas as seguintes bibliotecas: os, sys, logging, datetime, torch, torch.nn, torch.optim, pandas, torchvision, transforms, torch.utils.data, DataLoader, Dataset, PIL, sklearn.metrics, classification_report, confusion_matrix, matplotlib.pyplot. Foi executado em ambiente Python 3.10. e salva um mapa de atenção focada para interpretabilidade.

O ViT divide as imagens em patches, tratando cada uma como um token de entrada para o transformer, permitindo capturar padrões espaciais relevantes para a classificação.

Foram obtidas as métricas de desempenho, tais como: acurácia, relatório de classificação e matriz de confusão.

CLASSIFICATION REPORT:

- **Precisão:** Proporção de classificações positivas corretas para aquela classe.
- **Recall:** Proporção de exemplos daquela classe corretamente identificados pelo modelo.
- **F1-score:** Média harmônica entre precisão e recall, fornecendo uma visão equilibrada entre ambos.
- **Support:** Número de amostras reais daquela classe no conjunto de teste.
- **Accuracy:** Proporção total de acertos.
- **Macro avg:** Média simples de cada métrica entre as classes (todas as classes têm peso igual).
- **Weighted avg:** Média ponderada de cada métrica, de acordo com a quantidade de amostras por classe.

MATRIZ DE CONFUSÃO: A matriz de confusão é uma ferramenta que ajuda a entender como o modelo de classificação está se comportando em relação a cada classe. Ela mostra não só os acertos, mas também os erros do modelo, evidenciando como as classes estão sendo confundidas entre si. As linhas correspondem às classes reais e as colunas correspondem às classes preditas. Cada célula da matriz mostra o número de exemplos da classe real ‘x’ que foram classificados como classe ‘x’ pelo modelo. A diagonal principal mostra os acertos (quando classe real = classe predita).

Técnica de Interpretabilidade: Foi estudada e aplicada a técnica Focused Attention (Playout et al., 2022) para avaliar a interpretabilidade das decisões do modelo ViT, destacando regiões das imagens mais influentes nas predições.

RESULTADOS E DISCUSSÃO:

DESEMPENHO DO MODELO VISION TRANSFORMER: O modelo Vision Transformer alcançou acurácia de 64%. O desempenho do modelo foi avaliado utilizando as métricas de precisão (precision), revocação (recall) e f1-score para cada classe, conforme apresentado no relatório de classificação (Tabela 1). Essas métricas permitem analisar não apenas a acurácia global, mas também a capacidade do modelo em identificar corretamente cada estádio de maturação do fruto. Os valores de precisão indicam a proporção de previsões corretas para cada classe, enquanto a revocação avalia a capacidade do modelo em recuperar todos os exemplos reais de cada classe. O f1-score combina essas duas métricas, oferecendo uma avaliação equilibrada. A acurácia geral do modelo sobre o conjunto de teste foi de 64%. É necessário empreender trabalho para aumento da acurácia.

Tabela 1 – Relatório de classificação do modelo.

Relatório de classificação				
	Precisão	Recall	f1-score	support
freshripe	0,58	0,60	0,59	156
freshunripe	0,74	0,57	0,64	212
overripe	0,51	0,78	0,62	178
ripe	0,62	0,67	0,64	351
rotten	0,72	0,63	0,67	448
unripe	0,70	0,59	0,64	174
Acurácia			0,64	1519
macro avg	0,64	0,64	0,63	1519
weighted avg	0,66	0,64	0,64	1519

MATRIZ DE CONFUSÃO:

Tabela 2 – Matriz de confusão

Real/Predito	Freshripe	Freshunripe	Overripe	Ripe	Rotten	Unripe
Freshripe	93	0	1	56	6	0
Freshunripe	0	121	0	37	17	37
Overripe	1	0	139	19	19	0
Ripe	60	0	13	234	44	0
Rotten	6	3	117	31	284	7
Unripe	0	40	2	3	27	102

A Tabela 2 apresenta a matriz de confusão do modelo para o conjunto de teste. Observa-se que a classe “overripe” foi frequentemente confundida com “ripe”, indicando que o modelo encontra

dificuldades para diferenciar esses estádios de maturação, possivelmente devido à semelhança visual entre eles. Esse banco de imagens apresenta grande complexidade, principalmente quando se tenta segmentar os frutos. O fundo usado acaba por confundir e tem as mesmas cores das frutas. Por outro lado, a classe “rotten” apresentou menos confusões, sugerindo que suas características são mais distintivas para o modelo.

Figura 1: Imagem usada para gerar o mapa de atenção

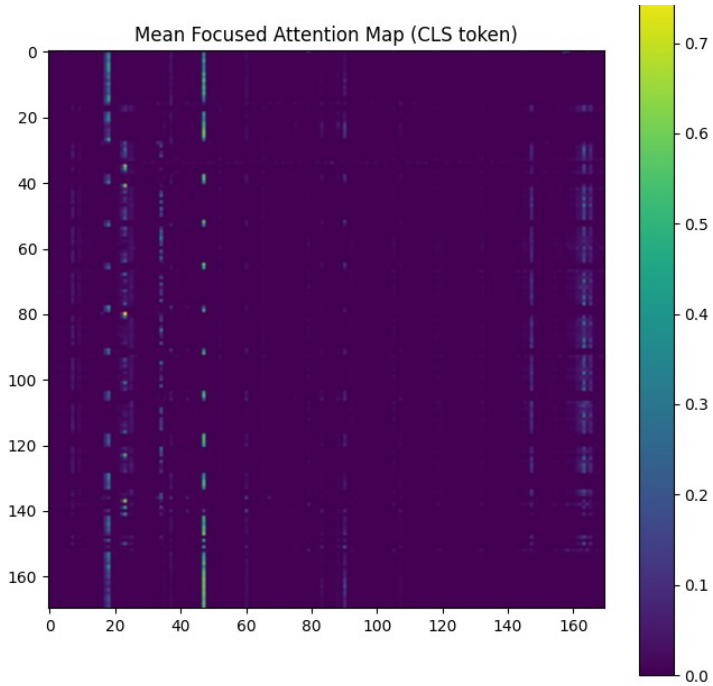

Figura 2: Mean Focused Attention Map (CLS token)

A Figura 2 "Mean Focused Attention Map (CLS token)" ilustra o mapa médio de atenção focada da camada final do Vision Transformer (ViT) para uma imagem de teste (Figura 1). No ViT, a imagem é dividida em pequenos patches, que são processados como tokens. O CLS token é inserido antes desses tokens de patch e, após o processamento pelas camadas do Transformer. O CLS token ("Classification Token") é utilizado para agregar informações globais e realizar a classificação da imagem. Sua saída é usada para prever a classe da imagem e permite que o modelo tome decisões de classificação a partir dele.

A imagem original (Figura 1) de 416×416 pixels é dividida no processamento em 13×13 patches de 32×32 pixels cada, totalizando 169 patches. Cada posição no mapa de atenção corresponde a um desses patches, indicando o grau de importância atribuído pelo modelo àquela região da imagem para a tarefa de classificação.

Embora as dimensões do mapa de atenção (Figura 2) (170×170, incluindo o token CLS) não correspondam diretamente à resolução da imagem original, é possível relacioná-las considerando que cada célula do mapa cobre uma região retangular da imagem.

O mapa evidencia as regiões da imagem que receberam maior atenção do modelo durante a classificação, destacando as áreas consideradas mais relevantes para a identificação do estádio de

maturação do fruto. As regiões mais claras no mapa indicam maior importância atribuída pelo modelo, enquanto regiões escuras ou azuladas tiveram menor influência na predição.

A escala de cores na Figura 2 indica a intensidade da atenção (valores mais altos, tendendo a 0,7 indicam regiões mais relevantes para o modelo). Assim, as áreas mais claras (valores próximos a 0.7) mostram onde o modelo teve maior atenção.

Isso permite uma análise interpretável do modelo, mostrando se ele está focando nas regiões corretas do fruto, como manchas, textura, cor ou sinais de maturação, ao invés de áreas irrelevantes do fundo.

Em trabalhos futuros, para facilitar a interpretação visual, pode-se redimensionar o mapa de atenção para o tamanho da imagem original e sobrepor, evidenciando graficamente as regiões de maior interesse para o modelo.

CONCLUSÕES:

É necessário realizar alterações na metodologia para que o modelo alcance melhor acurácia. Após garantir que o Vision Transformer tenha bom desempenho no banco de imagens, aplica-se a técnica de interpretabilidade para a visualização de quais regiões da imagem contribuíram para a decisão do modelo. Com essas informações é possível saber se o modelo tomou por base características relevantes do fruto ou se foi baseado em regiões fora do fruto, que não estão relacionadas com o processo de maturação.

BIBLIOGRAFIA

DOSOVITSKIY, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020.

PLAYOUT, C. et al. Focused Attention in Transformers for interpretable classification of retinal images. **Medical Image Analysis**, v. 82, p. 102608, nov. 2022.

RIPENING, F. **Fruit Ripening Process Dataset**. **Roboflow Universe**Roboflow, , out. 2022. Disponível em: <https://universe.roboflow.com/fruit-ripening/fruit-ripening-process>