Instituto Superior de Agronomia

Identificação de pragas e doenças em tomateiros com recomendação de aplicações

UC - Aprendizagem Automática Aplicada

Diana Luísa Santos Martins
20900

Índice

Introdução

Um dos desafios agronómicos que pode ser solucionado com recurso a *machine learning* passa pela identificação e classificação de pragas e doenças usando modelos de *deep learning* aplicados à classificação de imagem. Este tipo de soluções proporciona, por exemplo, um auxílio no acompanhamento técnico providenciado a agricultores, facilitando o acompanhamento das produções e o apoio técnico. Sendo uma ferramenta extremamente útil, o principal desafio associado é a obtenção de dados (imagens) de qualidade (bem identificados) que permitam o desenvolvimento deste tipo de soluções aplicadas a todas as principais culturas e pragas/doenças.

A proposta deste trabalho consiste numa destas soluções, aplicando-a a uma cultura de importante valor agronómico em Portugal - o tomateiro. Iremos, assim, implementar este sistema de aprendizagem automática a algumas pragas e doenças encontradas em tomateiro. O sistema identificará, através de imagens de sintomas, a praga/doença do tomateiro num modelo preditivo de *deep learning* com recurso à classificação de imagens. Uma segunda fase deste projecto consiste em que a predição dessa praga ou doença direccione para a gama de tratamentos possíveis de realizar na cultura. Integrando assim a identificação do problema com as soluções existentes de combate/cura.

Dados

Na realização deste tipo de trabalhos, os dados são das partes fundamentais para que o modelo tenha sucesso na identificação da praga ou doença associada. Desta forma procurámos, online, por dados de qualidade que fossem utilizáveis para este tipo de projectos.

Inicialmente obtivemos um modelo de pouco sucesso pela baixa qualidade das fotos da base de dados utilizada e que foi descartado. Por essa mesma razão acabamos por utilizar dados com origem noutra base de dados do Kaggle, que se encontravam já previamente tratados e bem identificados.

As imagens utilizadas no modelo de predição foram retiradas de uma base de dados do Kaggle com cerca de 22000 imagens de folhas de tomateiro no formato 256x256, divididas em 10 classes que incluem 9 sintomas associados a doenças e pragas e uma classe de folhas saudáveis. (https://www.kaggle.com/datasets/luisolazo/tomato-diseases/data) As classes disponíveis são: bacterial_spot, early_blight, healthy, late_blight, leaf_mold, mosaic_virus, septoria_leaf_spot, target_spot, twospotted_spider_mite, yellow_leaf_curl_virus.

Este banco de imagens já vinha tratado de forma a disponibilizar apenas imagens de qualidade, retiraram imagens que se apresentam desfocadas, e aplicaram algumas técnicas para aumentar as imagens disponíveis tais como as viragens horizontais e verticais, zooms, cisalhamentos, rotações, alterações no brilho etc.

Desta forma, para a segunda fase do projecto, retirámos da base de dados do SIFITO disponibilizado pela DGAV a lista de produtos homologados para as culturas em Portugal. Esta extracção foi depois limpa de forma a fazer a disponibilizar apenas tratamentos possíveis para as classes previamente identificadas na cultura do tomate. Esta informação foi organizada em formato .csv.

Organização dos Dados

Os dados disponíveis no Kaggle já se encontravam divididos em *Test* e *Train data*, sendo que ambas as pastas se encontravam divididas em 10 classes. O *train data set* dispõe de 17753 imagens e o *test data set* dispõe de 4440 imagens. O total de 22193 imagens permite identificar a proporção de 80% 20% normalmente usada neste tipo de soluções.

Inicialmente, os modelos desenvolvidos apenas redimensionavam as imagens de forma a serem compativeis com os formatos de entrada do ResNet50. No entanto, um novo tratamento das imagens provou que melhorava o desempenho do modelo e acabou por ser a estratégia utilizada.

Métodos

Na realização deste projecto de classificação de imagens utilizámos a CNN – Convolution Neural Network, baseada na arquitectura ResNet50 *pre-trained* em ImageNet para desenvolver este modelo de *machine learning*.

A arquitetura ResNet50 é uma rede convolucional de 50 camadas que aceita como dados de entrada imagens (224, 224, 3). O modelo é construído de forma a primeiro reduzir a dimensionalidade dos dados de entrada em vetor *GlobalAveragePooling2D*, a primeira *dense layer* com 512 unidades e a segunda com 256. A camada de saída apresenta o número de classes (10).

Inicialmente foi criado um modelo que, após avaliação das métricas, se verificou não ser adequado. Desta forma, procurámos melhorar o mesmo; aplicou-se o *optimizer Adam* com um valor baixo de *learning rate* (1e-4), para tentar melhorar a taxa de aprendizagem do modelo.

Para treinar o segundo modelo, realizámos algumas técnicas de aumento de dados para melhorar os resultados obtidos. Realizámos rotações, aumentos, cisalhamentos, zooms, flips e alterações de brilho nas imagens. Aplicámos, ainda, técnicas de *callback* como *EarlyStopping* e *checkpoint* de forma a parar o treino se o modelo não estiver a melhorar e para guardar o melhor modelo. Realizou-se, também, um compute_class_weight para garantir que as classes não estão desequilibradas. O *bach size* manteve-se em 32, mas aumentou-se o número de *epochs* para 50, considerando as alterações feitas previamente para que o treino pare no melhor modelo.

Depois de obtido o segundo modelo, versão melhorada do primeiro, voltou-se a verificar as métricas de avaliação de forma a averiguar o melhor desempenho deste modelo.

Após obtenção do modelo, passámos para a segunda fase de integração com o ficheiro .csv com as recomendações técnicas e testámos a sua executabilidade.

Resultados

Como resultado final deste trabalho, obtivemos o modelo final_tomato_disease_model.keras que foi o último modelo gravado. Este modelo apresenta métricas satisfatórias na identificação das doenças de tomateiro nestas 10 classes, mostrando um efetivo melhoramento quando comparado ao modelo inicialmente treinado.

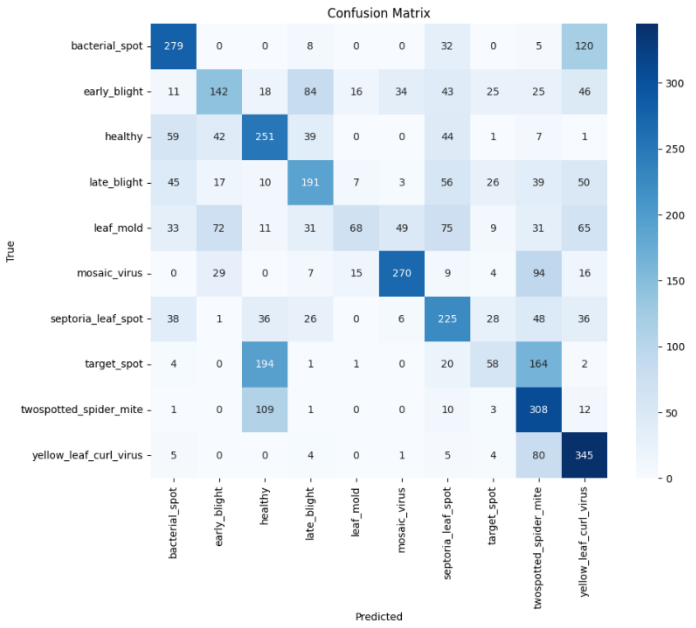

Figura 1 Confusion Matrix do modelo inicial

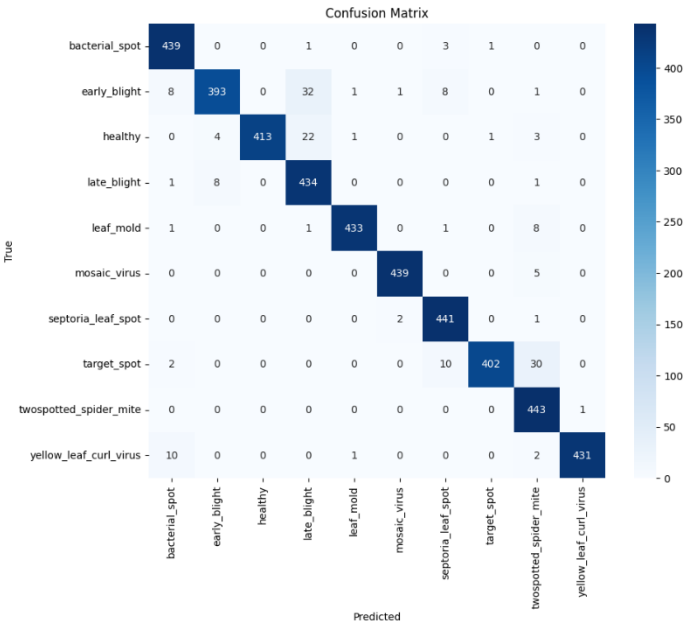

Figura 2 Confusion Matrix do modelo final

Ao nível das métricas de *performance* do modelo inicial obtivemos os seguintes valores:

```
Accuracy: 0.4813
Precision: 0.5005
Recall: 0.4813
F1-Score: 0.4590
                        precision    recall  f1-score   support

        bacterial_spot       0.59      0.63      0.61       444
          early_blight       0.47      0.32      0.38       444
               healthy       0.40      0.57      0.47       444
           late_blight       0.49      0.43      0.46       444
             leaf_mold       0.64      0.15      0.25       444
         mosaic_virus        0.74      0.61      0.67       444
    septoria_leaf_spot       0.43      0.51      0.47       444
           target_spot       0.37      0.13      0.19       444
twospotted_spider_mite       0.38      0.69      0.49       444
yellow_leaf_curl_virus       0.50      0.78      0.61       444

              accuracy                           0.48      4440
             macro avg       0.50      0.48      0.46      4440
          weighted avg       0.50      0.48      0.46      4440
```

Enquanto que os apresentados no modelo final foram:

```
                        precision    recall  f1-score   support

        bacterial_spot       0.95      0.99      0.97       444
          early_blight       0.97      0.89      0.93       444
               healthy       1.00      0.93      0.96       444
           late_blight       0.89      0.98      0.93       444
             leaf_mold       0.99      0.98      0.98       444
         mosaic_virus        0.99      0.99      0.99       444
    septoria_leaf_spot       0.95      0.99      0.97       444
           target_spot       1.00      0.91      0.95       444
twospotted_spider_mite       0.90      1.00      0.94       444
yellow_leaf_curl_virus       1.00      0.97      0.98       444

              accuracy                           0.96      4440
             macro avg       0.96      0.96      0.96      4440
          weighted avg       0.96      0.96      0.96      4440

Accuracy: 0.9613
Precision: 0.9637
Recall: 0.9613
F1-Score: 0.9613
```

Com um modelo a fazer predicções correctamente, foi possivel verificar a correcta leitura dos dados do ficheiro .CSV criado a partir dos dados extraídos do SIFITO e confirmar a correcta apresentação dos resultados:

```
Disease: leaf_mold
Treatments:
Product: SCORE 250 EC, Dose:  - 500 mL/ha, IS: 7 / -
Product: ZANOL, Dose:  - 500 mL/ha, IS: 7 / -
Product: MAVITA 250 EC, Dose:  - 500 mL/ha, IS: 7 / -
Product: DAGONIS, Dose: 1 L/ha, IS: 3 / -
Product: GALAVIO, Dose:  - 500 mL/ha, IS: 7 / -
Product: DIZOLE, Dose:  - 500 mL/ha, IS: 7 / -
Product: BLIN 25 EC, Dose:  - 500 mL/ha, IS: 7 / -
Product: DIFENOFIN, Dose: 250 - 500 mL/ha, IS: 7 / -
```

Análise

O desenvolvimento deste projeto foi de acordo com o proposto. Nesta medida, foi possível realizar um modelo de predicção que consegue, com uma boa taxa de sucesso, identificar as doenças em tomate com base na sintomatologia das folhas.

Importa referir que, neste caso particular, as doenças iniciais classificadas provocam um real desafio a este trabalho por serem doenças de dificil identificação apenas com base em observação da sintomatologia em folhas, havendo bastante dificuldade na real identificação das doenças, pois apresentam sintomatologia muito semelhante.

Com estas considerações importa referir o real melhoramento do modelo onde foi possível verificar um aumento significativo das métricas de avaliação. Este último modelo apresenta então uma *accuracy* de 96.13%, uma *precision* de 96.37%, um *recall* de 96.13%, e um F1-Score de 96.13% valores bastante satisfatórios indicando robustez e precisão.

A segunda função neste projeto, ligada às recomendações, prova-se eficiente com a correcta identificação da praga/doença pelo modelo e a ligação com as recolmendações para a praga/doença em causa. Desta forma, a ligação simplificada dos dados em formato .csv permite que a leitura dos tratamentos seja fácil e rápida. Esta ligação foi particularmente importante para a escalabilidade prática deste projecto, pois a identificação de pragas e doenças faz sentido para a aplicabilidade de soluções e curas. Esta pequena diferença permite que esta projecto tenha valor agronómico.

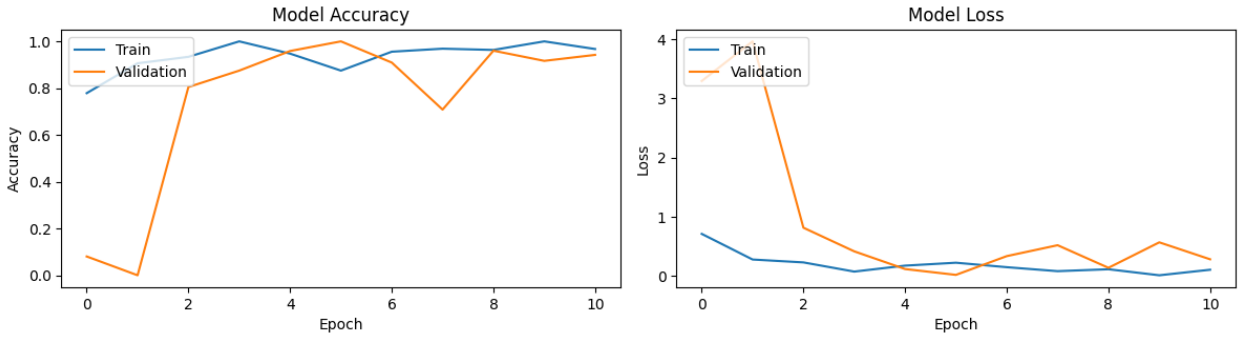

Figura 3 Gráficos de Accuracy e Loss do modelo final

Conclusão

O desenvolvimento deste projecto provou ser enriquecedor e demonstrou a possibilidade na contrução destes modelos com recurso a dados de qualidade. A maior complexidade associada reside na obtenção de dados de qualidade e bem identificados. Importa considerar que, apesar das dificuldades na separação de algumas doenças pelas parecenças dos sintomas, não estão a ser consideradas sintomatologias do fruto (tomate) ou de outras partes da planta, bem como especificidades varietais. Todos estes detalhes aumentariam a robustez deste tipo de modelos e acrescentariam valor. No entanto, a obtenção ou criação destas bases de dados será desafiante.

Outra oportunidade de melhoria seria a integração das recomendações técnicas ligada através de API diretamente ao SIFITO de forma a manter a constante actualização do sistema. Este tipo de API não se encontra ainda disponível.

Concluindo, apesar dos desafios e oportunidades de melhoria, a realização deste trabalho foi de encontro ao objectivo estabelecido.

Referências

https://www.kaggle.com/datasets/luisolazo/tomato-diseases

https://sifito.dgav.pt/divulgacao/usos