Detección y Clasificación de Patógenos Foliares mediante Visión por Computador

Gabriel Sánchez Muñoz
gabrielsm@correo.ugr.es

Germán Rodríguez Vidal
germanrv@correo.ugr.es

Pablo García Bas
pablogb@correo.ugr.es

Miguel Ángel Moreno Castro
miguelangelmc@correo.ugr.es

Abstract

La detección temprana y precisa de enfermedades vegetales es crítica para la agricultura de precisión y la seguridad alimentaria. Los enfoques actuales basados en Deep Learning, a pesar de mostrar resultados prometedores, suelen carecer de información contextual sobre el huésped, lo que limita su precisión en escenarios de múltiples cultivos. Este trabajo propone un enfoque jerárquico de dos etapas para abordar esta limitación. En primer lugar, se realiza una clasificación de la especie arbórea para establecer un contexto biológico y, posteriormente, un detector de objetos localiza y clasifica los patógenos específicos asociados a dicha especie. Los resultados muestran que incorporar la identificación previa del huésped incrementa la precisión media (mAP) en un [X]% respecto a los detectores genéricos de una sola etapa ("end-to-end"), demostrando la superioridad de los modelos condicionados biológicamente.

1. Introducción

Las enfermedades de las plantas representan una amenaza significativa para la seguridad alimentaria mundial y la estabilidad económica agrícola, causando pérdidas estimadas de hasta el 30% en el rendimiento de los cultivos cada año [**?**]. La identificación temprana y precisa de estas patologías es crucial para aplicar medidas de control eficaces y minimizar el uso de químicos. Tradicionalmente, este diagnóstico ha dependido de la inspección visual manual por parte de expertos, un proceso que resulta laborioso, subjetivo y difícil de escalar a grandes explotaciones [**?**].

En la última década, la visión por computador y, específicamente, el aprendizaje profundo, han emergido como herramientas poderosas para automatizar esta tarea. El uso de Redes Neuronales Convolucionales (CNNs) ha permitido grandes avances tanto en la clasificación de imágenes (identificar si una hoja está enferma) como en la detección de objetos (localizar la lesión exacta en la hoja) [**?**]. Sin embargo, la mayoría de las arquitecturas actuales abordan la detección de enfermedades como un problema monolítico donde los modelos se entrenan para detectar cualquier enfermedad en cualquier tipo de hoja simultáneamente, ignorando a menudo la estructura jerárquica natural de la taxonomía biológica.

Debido a que muchas patologías foliares comparten características visuales similares, independientemente de la especie, los detectores agnósticos al huésped sufren de una alta confusión inter-clase. Esto suele derivar en predicciones erróneas donde se asocian enfermedades a plantas incompatibles, reduciendo la precisión y confiabilidad del sistema en aplicaciones del mundo real.

Para resolver esto, proponemos un enfoque jerárquico que imita el diagnóstico experto donde primero se identifica la especie del árbol para simplificar la búsqueda de la enfermedad y luego se aplica un detector especializado para esa especie.

2. Dataset

Para validar nuestro enfoque jerárquico, hemos confeccionado un *dataset* que integra imágenes de tres cultivos: rosa, patata y manzana. Las imágenes provienen de la plataforma Roboflow Universe, que contiene una gran cantidad de *datasets* de código abierto, y han sido seleccionadas por su calidad y variedad de condiciones de iluminación.

- Rose Dataset: Consta de 2,725 imágenes y abarca 4 clases (Black Spot, Downy Mildew, Powdery Mildew y hojas sanas).
- Potato Dataset: Comprende 812 imágenes distribuidas en [X] clases (típicamente Early Blight, Late Blight, Healthy).
- Apple Dataset: Incluye 1582 imágenes que cubren 4 patologías (Alternaria Spot, Brown Spot, Gray Spot,

Rust).

El dataset combinado se dividió aleatoriamente en subconjuntos de Entrenamiento (70%), Validación (15%) y Prueba (15%), preservando la estratificación de clases para garantizar una evaluación equilibrada.

3. Clasificación

Para el diseño del módulo de clasificación, hemos evaluado dos paradigmas principales del estado del arte, las arquitecturas basadas en Vision Transformers (ViT) y las Redes Neuronales Convolucionales (CNNs). Nuestro objetivo fue contrastar la capacidad de modelado global de los Transformers frente a la eficiencia inductiva de las CNNs. Adicionalmente, dada la orientación práctica de este proyecto hacia una futura aplicación móvil, se incluyó en el estudio un modelo diseñado específicamente para entornos de recursos limitados.

3.1. MaxViT

Dentro de la familia de los Transformers, MaxViT (Multi-Axis Vision Transformer) combina tanto mecanismos de atención global como convoluciones locales para capturar características a múltiples escalas [**?**]. Su característica distintiva es el mecanismo de *Multi-Axis Self-Attention* (Max-SA), que descompone el cálculo de atención en dos operaciones: atención local (*Block Attention*) para capturar texturas finas, y atención global dispersa (*Grid Attention*) para relacionar partes distantes de la imagen.

3.1.1 EfficientNetV2

Esta arquitectura mejora a su predecesora mediante la introducción de bloques Fused-MBConv [**?**]. Estos bloques reemplazan las convoluciones *depthwise* separables tradicionales por convoluciones estándar 3×3 fusionadas en las primeras capas.

3.2. MobileNet

La base de su eficiencia radica en las convoluciones *depthwise* separables, que factorizan la operación de convolución estándar en dos capas más ligeras (profundidad y punto a punto), reduciendo drásticamente la cantidad de parámetros y operaciones [**?**]. Aunque su capacidad de representación es menor que los modelos anteriores, su inclusión es crítica para evaluar el compromiso (*trade-off*) entre precisión y latencia en una aplicación real para agricultores.

4. Detección de Objetos

El objetivo fundamental en detección de objetos es localizar y clasificar regiones de interés (*bounding boxes*). Históricamente, métodos dominantes como *Deformable Parts Model* (DPM) abordaban este problema mediante un enfoque de ventana deslizante (*sliding window*) que resultaban inviables computacionalmente al procesar todas las posibles sub-ventanas [**?**].

La evolución comenzó con **R-CNN**, un modelo de 3 etapas donde una búsqueda selectiva genera regiones potencialmente interesantes, una CNN extrae características y un SVM las clasifica [**?**]. Sus limitaciones principales son el alto coste de procesar regiones por separado y la falta de aprendizaje en la búsqueda selectiva.

4.1. Faster R-CNN

Faster R-CNN integra la propuesta de regiones dentro de la red neuronal, reemplazando el algoritmo fijo por una *Region Proposal Network* (RPN) entrenable y más rápida [**?**].

En primer lugar se definen *anchor points* sobre el mapa de características de la imagen, actuando como centros de posibles regiones con diferentes escalas y ratios. Seguidamente, la RPN predice simultáneamente la probabilidad de objeto y las coordenadas del *bounding box*.

A pesar de su alta precisión, el hecho de ser un modelo de dos etapas penaliza su velocidad de inferencia.

4.2. YOLO

Como respuesta a la latencia, *You Only Look Once* (YOLO) redefinió la detección no como una clasificación de regiones, sino como un problema único de regresión, descartando por completo el pipeline disjunto de DPM y R-CNN.

YOLO utiliza una única red neuronal convolucional que procesa la imagen completa de una sola vez. Divide la imagen en una cuadrícula ($S \times S$) donde cada celda predice B bounding boxes y sus respectivas puntuaciones de confianza.

La arquitectura unificada de YOLO permite una inferencia en tiempo real inalcanzable para los métodos de dos etapas [**?**]. Mientras que Faster R-CNN procesa aproximadamente 0.5 FPS, YOLO alcanza los 45 FPS, ofreciendo la inmediatez necesaria para aplicaciones agrícolas prácticas.

5. Métodos

Descripción detallada de los métodos utilizados y/o propuestos, y justificación clara de por qué se usan estos métodos y no otros.

6. Experimentos

Aquí se presentan los datos utilizados, el protocolo de validación experimental, las métricas usadas, los experimentos realizados, los resultados obtenidos y su discusión.

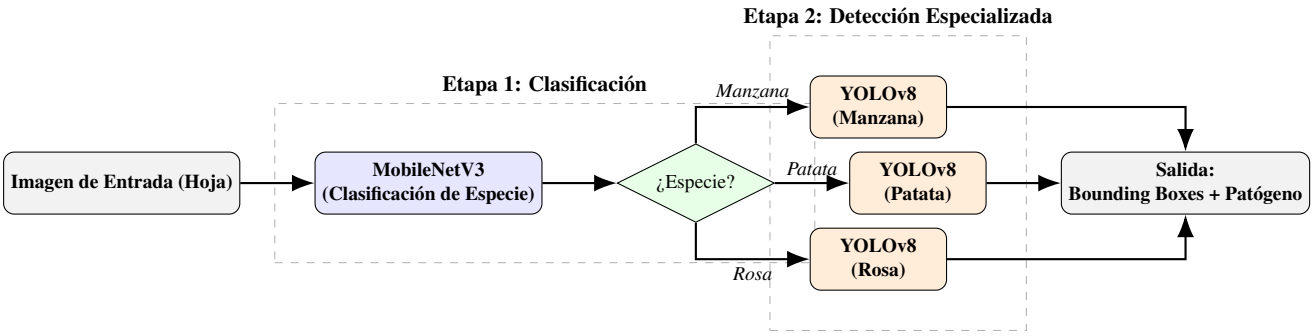

Figure 1. Visión general de la arquitectura jerárquica propuesta. Una imagen de entrada es procesada primero por una red MobileNetV3 para identificar el cultivo huésped (Etapa 1). En base a esta predicción, se selecciona dinámicamente un modelo de detección YOLOv8 especializado (Etapa 2) para localizar y clasificar las patologías foliares correspondientes, evitando la confusión inter-especie.

6.1. Conjunto de Datos

7. Conclusiones

Sección que presenta, brevemente y a modo de resumen, las principales conclusiones del trabajo realizado. También suele incluir posibles trabajos futuros. Es decir, cuáles son las líneas más prometedoras para continuar con este trabajo, así como posibles propuestas de mejora. IMPORTANTE: estas son las conclusiones científicas alcanzadas en el proyecto; ¡no tus conclusiones personales sobre el trabajo que has realizado!