UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Análisis y aplicación de técnicas de explicabilidad en redes neuronales sobre problemas con imágenes

Trabajo Fin de Máster

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

AUTOR/A: Poncelas Vargas, José David

Tutor/a: Martínez Hinarejos, Carlos David

CURSO ACADÉMICO: 2023/2024

Resumen

Este trabajo se enfoca en el análisis y la implementación de técnicas de explicabilidad destinadas a mejorar la comprensión de redes neuronales aplicadas a problemas que implican el procesamiento de imágenes. En primer lugar, se realiza una revisión exhaustiva del estado actual de la inteligencia artificial explicativa, explorando las principales metodologías y enfoques utilizados en este campo. Posteriormente, se procede a llevar a cabo una serie de experimentos prácticos con el propósito de profundizar en el funcionamiento interno de las redes neuronales, estudiando principalmente el comportamiento de redes neuronales convolucionales. Estos experimentos se centran en identificar cómo estas redes procesan y clasifican imágenes, destacando las características visuales que influyen significativamente en sus decisiones. Para llevar esto a cabo se emplean diversas técnicas de explicabilidad, tales como SHAP (*SHapley Additive exPlanations*), que descompone el impacto de cada característica en la predicción, o LIME (*Local Interpretable Model-agnostic Explanations*), que explica predicciones individuales mediante ajustes locales a los datos y modelos más simples, entre otras. El análisis detallado de estas técnicas proporciona una mayor comprensión sobre el comportamiento y el razonamiento de las redes neuronales en el contexto de problemas con imágenes, tratando así de promover el desarrollo de sistemas de inteligencia artificial más transparentes, interpretables y confiables. Esto, en un futuro, puede ayudar a cumplir con las regulaciones gubernamentales en cuanto a la explicabilidad de la toma de decisiones automáticas, detectar sesgos indeseados, errores sistemáticos, fallas potenciales, mejorar la equidad de estos modelos o aumentar la confianza de los usuarios en las decisiones automáticas de estos sistemas.

Palabras clave

Explicabilidad, Interpretabilidad, Redes neuronales convolucionales, Procesamiento de imágenes, Técnicas de explicabilidad, SHAP, LIME, Transparencia en la inteligencia artificial.

Resum

Este treball se centra en l'anàlisi i la implementació de tècniques d'explicabilitat destinades a millorar la comprensió de xarxes neuronals aplicades a problemes que impliquen el processament d'imatges. En primer lloc, es realitza una revisió exhaustiva de l'estat actual de la intel·ligència artificial explicativa, explorant les principals metodologies i enfocaments utilitzats en aquest camp. Posteriorment, es procedeix a dur a terme una sèrie d'experiments pràctics amb el propòsit d'aprofundir en el funcionament intern de les xarxes neuronals, estudiant principalment el comportament de les xarxes neuronals convolucionals. Aquests experiments se centren a identificar com aquestes xarxes processen i classifiquen imatges, destacant les característiques visuals que influeixen significativament en les seves decisions. Per dur això a terme, s'empren diverses tècniques d'explicabilitat, com ara SHAP (*SHapley Additive exPlanations*), que descompon l'impacte de cada característica en la predicció, o LIME (*Local Interpretable Model-agnostic Explanations*), que explica prediccions individuals mitjançant ajustos locals a les dades i models més senzills, entre altres. L'anàlisi detallat d'aquestes tècniques proporciona una major comprensió sobre el comportament i el raonament de les xarxes neuronals en el context de problemes amb imatges, tractant així de promoure el desenvolupament de sistemes d'intel·ligència artificial més transparents, interpretables i fiables. Això, en un futur, pot ajudar a complir amb les regulacions governamentals quant a l'explicabilitat de la presa de decisions automàtiques, detectar biaixos indesitjats, errors sistemàtics, fallades potencials, millorar l'equitat d'aquests models o augmentar la confiança dels usuaris en les decisions automàtiques d'aquests sistemes.

Paraules clau

Explicabilitat, Interpretabilitat, Xarxes Neuronals Convolucionals, Processament d'Imatges, Tècniques d'Explicabilitat, SHAP, LIME, Transparència en la Intel·ligència Artificial.

Abstract

This work focuses on the analysis and implementation of explainability techniques aimed at improving the understanding of neural networks applied to problems involving image processing. First, a comprehensive review of the current state of explainable artificial intelligence is conducted, exploring the main methodologies and approaches used in this field. Subsequently, a series of practical experiments are carried out to delve into the internal workings of neural networks, primarily studying the behavior of convolutional neural networks. These experiments focus on identifying how these networks process and classify images, highlighting the visual features that significantly influence their decisions. To accomplish this, various explainability techniques are employed, such as SHAP (SHapley Additive exPlanations), which decomposes the impact of each feature on the prediction, or LIME (Local Interpretable Model-agnostic Explanations), which explains individual predictions by making local adjustments to the data and simpler models, among others. The detailed analysis of these techniques provides a greater understanding of the behavior and reasoning of neural networks in the context of image problems, thus aiming to promote the development of more transparent, interpretable, and reliable artificial intelligence systems. This, in the future, may help comply with government regulations regarding the explainability of automatic decision-making, detect undesired biases, systematic errors, potential failures, improve the fairness of these models, or increase users' confidence in the automatic decisions of these systems.

Keywords

Explainability, Interpretability, Convolutional Neural Networks, Image Processing, Explainability Techniques, SHAP, LIME, Transparency in Artificial Intelligence.

Tabla de contenidos

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En la era actual de la inteligencia artificial (IA) [1], las redes neuronales convolucionales (CNNs) [2] han demostrado ser extremadamente poderosas en la clasificación de imágenes y otras tareas complejas [3]. Sin embargo, su capacidad para tomar decisiones complejas a menudo se percibe como una 'caja negra', donde es difícil comprender cómo y por qué se realizan ciertas predicciones. Este fenómeno plantea un dilema fundamental: ¿cómo podemos confiar en los sistemas de IA si no podemos entender su razonamiento?

La explicabilidad en IA [4] aborda este desafío crucial al proporcionar métodos y técnicas para descomponer y entender las decisiones tomadas por los modelos de IA, especialmente aquellos basados en redes neuronales. La importancia de la explicabilidad radica en varios aspectos clave. Primero, mejora la transparencia, permitiendo a los usuarios y desarrolladores comprender cómo funciona realmente un sistema de IA en situaciones críticas como diagnósticos médicos automatizados o decisiones de seguridad en vehículos autónomos [4].

Además, las técnicas de explicabilidad no solo ofrecen conocimiento sobre cómo se comportan los modelos de IA, sino que también pueden identificar sesgos indeseados o errores sistemáticos, promoviendo así sistemas más justos y éticos [5]. Esto es crucial en aplicaciones donde la equidad y la no discriminación son imperativas, como en la selección de personas candidatas o en la evaluación crediticia.

Desde una perspectiva social, la explicabilidad en IA puede fomentar una mayor aceptación y confianza pública en las tecnologías de IA [5]. Al proporcionar justificaciones comprensibles y claras sobre las decisiones automatizadas, se puede facilitar una mejor integración de estas tecnologías en la sociedad, reduciendo la resistencia y la incertidumbre que a menudo rodea a la IA. De esta forma, el desarrollo y la aplicación efectiva de herramientas de explicabilidad fortalecerán la capacidad de la sociedad para utilizar y

beneficiarse de esta tecnología.

1.1. Motivación

El interés por explorar las técnicas de explicabilidad en redes neuronales aplicadas al procesamiento de imágenes surge de mi profunda fascinación por la inteligencia artificial y su impacto en nuestra sociedad. A medida que la IA se integra más en nuestra vida cotidiana, desde diagnósticos médicos hasta recomendaciones personalizadas en plataformas digitales, se vuelve cada vez más crucial entender cómo y por qué los modelos de IA toman decisiones.

Este tema no solo me parece apasionante desde un punto de vista técnico y científico, sino también debido a su importancia ética y social. La opacidad inherente de las redes neuronales plantea desafíos significativos en términos de confianza y aceptación pública de estas tecnologías. En un mundo donde la regulación de la IA está en constante evolución [6], comprender los detalles internos que explican las decisiones de los modelos IA se convierte en un factor crítico para cumplir con normativas éticas y legales cada vez más estrictas.

Además, las técnicas de explicabilidad ofrecen la oportunidad de identificar y mitigar posibles sesgos algorítmicos y errores sistemáticos en los modelos de IA, promoviendo así sistemas más justos y equitativos. Este potencial para mejorar la transparencia y la interpretabilidad de los modelos de IA no solo fortalece la confianza del usuario, sino que también abre nuevas posibilidades para aplicaciones más éticas y responsables en diversos sectores, desde la salud hasta la justicia y la seguridad [6, 7].

En este contexto, mi motivación principal radica en contribuir a la investigación y desarrollo de sistemas de IA más transparentes y confiables, capaces de ser comprendidos y utilizados de manera efectiva por profesionales, reguladores y la sociedad en general, alineándose con los principios de ética y responsabilidad tecnológica que considero fundamentales en la innovación digital del siglo XXI.

1.2. Objetivos

El principal objetivo de este trabajo es establecer una base conceptual sólida que justifique la necesidad de desarrollar técnicas de explicabilidad en inteligencia artificial, con un enfoque particular en su importancia dentro del ámbito de la clasificación de imágenes. Para ello, se busca comprender por qué es crucial que los modelos de inteligencia artificial sean no solo precisos, sino también interpretables, lo que permite un mayor grado de confianza

en sus predicciones. Además, el trabajo se propone revisar y analizar de manera exhaustiva las principales técnicas de explicabilidad disponibles en la actualidad.

Otro objetivo clave es la selección cuidadosa del conjunto de datos y la arquitectura del modelo más adecuados para la experimentación. La elección de estos elementos es crucial para garantizar que los experimentos sean representativos y que los resultados obtenidos puedan generalizarse a otros problemas similares.

Finalmente, el trabajo se centrará en la implementación y evaluación de diversas técnicas de explicabilidad en un modelo de clasificación de imágenes. Este objetivo implica aplicar técnicas de explicabilidad al modelo seleccionado, así como analizar su capacidad para generar explicaciones que sean tanto interpretables como útiles.

1.3. Estructura

El presente trabajo se estructura en cinco capítulos que abordan de manera integral el estudio y la aplicación de técnicas de explicabilidad en redes neuronales enfocadas en problemas de clasificación de imágenes. A continuación, se describe detalladamente la estructura del trabajo.

En el Capítulo 1, se establecen las bases conceptuales y motivacionales del estudio. Se inicia con la motivación del trabajo, donde se expone la importancia creciente de la explicabilidad en los sistemas de inteligencia artificial. Se identifican los retos actuales en la interpretación de modelos complejos, justificando la necesidad de desarrollar técnicas que permitan entender y confiar en las predicciones de dichos modelos. A continuación, se definen los objetivos del trabajo, estableciendo las metas que se pretenden alcanzar, como la evaluación de distintas técnicas de explicabilidad y su aplicación en redes neuronales convolucionales. Finalmente, en la sección de convenciones, se aclaran las notaciones, terminología y estilos que se utilizarán a lo largo del texto, asegurando una comprensión coherente y uniforme del contenido.

El Capítulo 2, proporciona un exhaustivo análisis de las principales técnicas de explicabilidad en la inteligencia artificial. Se aborda primero los métodos de visualización, que se dividen en dos categorías principales: métodos basados en *backpropagation*, que incluyen técnicas como Grad-CAM y *Guided Backpropagation*, y métodos basados en perturbación, como *occlusion sensitivity*. Luego, se discuten las técnicas de destilación de modelos, donde se exploran enfoques como la aproximación local y la traducción de modelos a árboles de decisión, que simplifican los modelos complejos en versiones más interpretables. El capítulo también cubre los métodos intrínsecos,

centrándose en técnicas como los mecanismos de atención y el entrenamiento conjunto para generar explicaciones textuales, que integran la explicabilidad dentro del propio proceso de aprendizaje del modelo. Finalmente, se detallan las métricas de evaluación y los métodos de evaluación de técnicas aplicadas a problemas con imágenes, proporcionando un marco para medir la efectividad de las técnicas de explicabilidad. Se concluye con una discusión sobre los problemas asociados a la explicabilidad, como la susceptibilidad a ataques adversariales.

En el Capítulo 3, se presentan los aspectos técnicos y metodológicos fundamentales del estudio. Se inicia con una descripción de los datos utilizados, donde se detallan las características del conjunto de datos y las razones detrás de la selección de este conjunto. Luego, se describe el modelo de clasificación de imágenes empleado, explicando la arquitectura de la red, en particular la elección de EfficientNet, y cómo este modelo se adapta a las necesidades del experimento.

El Capítulo 4, constituye el núcleo práctico del trabajo. Aquí se detalla el modelo de clasificación de imágenes, incluyendo el proceso de entrenamiento de la red EfficientNet. Luego, se presenta la experimentación con técnicas de explicabilidad, donde se aplican algunas de las técnicas descritas en el estado del arte al modelo entrenado, analizando su capacidad para generar explicaciones interpretables. Además, se incluye un apartado de otros experimentos, donde se exploran técnicas adicionales como *occlusion sensitivity* y LIME sin una evaluación numérica formal, destacando observaciones cualitativas sobre la aplicabilidad y resultados de estas técnicas.

Finalmente, el Capítulo 5, sintetiza los aprendizajes del trabajo. Se presentan las conclusiones principales, destacando los avances logrados en la comprensión de las técnicas de explicabilidad y su implementación en redes neuronales. Además, se analiza la relación del trabajo desarrollado con los estudios cursados, contextualizando el aprendizaje obtenido en el marco del Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital. Se discuten también los trabajos futuros, sugiriendo direcciones para la continuación de la investigación, incluyendo la evaluación numérica de técnicas como LIME y SHAP, la experimentación con otros conjuntos de datos y la implementación de un modelo que combine redes neuronales y recurrentes para generar explicaciones textuales.

1.4. Convenciones

En este trabajo se han definido las siguientes convenciones:

- Se va a utilizar la *letra cursiva* para hacer referencia a palabras escritas

en otro idioma distinto al español o para resaltar tecnicismos de la informática.

- Se van a utilizar los corchetes con un número en el interior, como por ejemplo [1], para indicar el elemento de la bibliografía que hace referencia a la información expuesta donde se utilicen.

Capítulo 2

Estado del arte

La inteligencia artificial ha avanzado rápidamente en diversas aplicaciones, desde el reconocimiento de imágenes [2] hasta la conducción autónoma [8], transformando industrias y servicios. Sin embargo, la adopción generalizada de sistemas de IA enfrenta desafíos significativos relacionados con la opacidad y la interpretabilidad de sus decisiones. En este contexto, las técnicas de inteligencia artificial explicable juegan un papel crucial en mejorar la transparencia y la interpretabilidad de las complejas redes neuronales profundas, *Deep Neural Networks (DNNs)* [9].

Estas técnicas tienen como objetivo proporcionar una perspectiva sobre los procesos de toma de decisiones de los modelos de IA, haciéndolos más comprensibles y confiables para los usuarios. En este contexto, han surgido tres enfoques principales como pilares clave de la IA explicativa: métodos de visualización, destilación de modelos y métodos intrínsecos [5, 10].

2.1. Métodos de visualización

Los métodos de visualización ofrecen una representación visual de cómo una red neuronal procesa los datos de entrada y llega a su salida. Al resaltar las características de entrada que influyen significativamente en las decisiones del modelo, estos métodos ayudan a los usuarios a entender el funcionamiento interno del sistema de IA.

Las formas comunes de visualización incluyen mapas de saliencia y mapas de calor [11], que identifican las características de entrada más relevantes que afectan a la salida del modelo. Estas visualizaciones proporcionan detalles valiosos sobre el proceso de toma de decisiones de la *DNN*, facilitando la interpretación de modelos complejos [12]. Existen dos variantes principales, los métodos basados en *backpropagation* y los basados en perturbación, los

cuales se describen a continuación.

2.1.1. Métodos basados en *backpropagation*

Los métodos basados en *backpropagation* son técnicas de explicabilidad que utilizan la retropropagación de gradientes en redes neuronales para identificar las características más relevantes de las entradas que influyen en las decisiones del modelo. Estos métodos permiten comprender mejor cómo una red neuronal procesa la información y cuáles son las partes de una entrada que tienen un mayor impacto en la predicción final [12].

A continuación se van a explicar de forma detallada los métodos basados en *backpropagation* más populares y utilizados.

Mapas de Saliencia (*Saliency Maps*)

Un mapa de saliencia es uno de los métodos más antiguos y frecuentemente utilizados para entender los modelos de aprendizaje profundo. Esencialmente, un mapa de saliencia resalta las regiones de una imagen de entrada que tienen el mayor impacto en la actividad de una capa particular o en la decisión general de la red. Los mapas de saliencia son fundamentalmente métodos de interpretación locales basados en gradientes. Aunque se utilizan principalmente para interpretar CNNs, pueden aplicarse a cualquier red neuronal debido al concepto universal de gradientes, lo que los convierte en un método de interpretación agnóstico al modelo [13].

Existen tres métodos principales para generar el mapa de saliencia de una imagen de entrada para una CNN, los cuales se explican a continuación.

Redes de deconvolución: Introducidas por Zeiler y Fergus en 2013 [14], este enfoque identifica qué características (esencialmente píxeles) en la imagen de entrada está enfocando una capa intermedia de la red. La red de deconvolución reconstruye la entrada a partir de las activaciones de esa capa, invirtiendo las operaciones realizadas entre la entrada y la capa específica. Esto implica usar deconvolución como el inverso de la convolución, *desempooling* como el inverso del *pooling*, y aplicar ReLU en reversa, ajustando los valores negativos de la señal que se mueve hacia atrás desde el espacio de activación hasta el espacio de la imagen. Dado que el *pooling* no es invertible, se utiliza un módulo llamado '*switch*' para recuperar las posiciones de los máximos en el paso hacia adelante. En la figura 2.1 se puede observar cómo se adjuntan las capas de deconvolución a las capas correspondientes de la CNN.

Retropropagación (Backpropagation): El segundo enfoque, propuesto por Simonyan en 2013 [13], consiste en utilizar el algoritmo de retropropagación

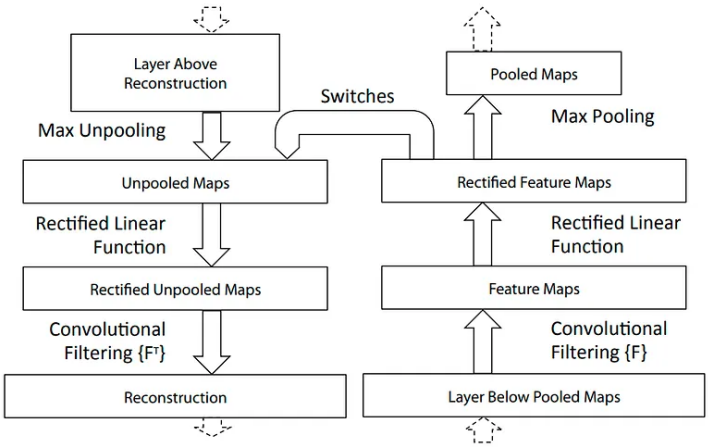

Figura 2.1: Una capa de deconvolución (izquierda) adjunta a una capa de CNN (derecha) [14]

para calcular los gradientes de las salidas de la red neuronal con respecto a su entrada. Este método resalta los píxeles de la imagen de entrada basándose en el gradiente que reciben, indicando su contribución al puntaje final. Esta técnica es directa e implica calcular gradientes de manera similar a como se calculan para entrenar la red, pero con respecto a la entrada en lugar de los parámetros.

Simonyan también demostró que este método es similar al uso de redes de deconvolución, con la diferencia clave en cómo ReLU maneja la señal de retropropagación. En una red de deconvolución, ReLU ajusta la señal de retropropagación si es negativa, mientras que en la retropropagación ajusta el gradiente de retropropagación si la entrada a ReLU en el paso hacia adelante era negativa [13].

Retropropagación guiada (Guided backpropagation, GDP): En 2014, Springenberg y sus compañeros integraron los dos enfoques comentados anteriormente y propusieron el algoritmo de retropropagación guiada como la tercera forma de obtener mapas de saliencia, ya que, a diferencia de la retropropagación estándar (que puede perder información sobre la localización espacial de características importantes) y de la deconvolución (que puede generar mapas de saliencia difusos), GDP mantiene y resalta detalles precisos al restringir los gradientes negativos [15].

Para entenderlo mejor, se va a explicar más en detalle este proceso. En una red neuronal, la entrada (por ejemplo, una imagen) se procesa a través de múltiples capas de neuronas, cada una aplicando una serie de operaciones matemáticas para transformar la entrada en una predicción. Durante el entrenamiento, el modelo ajusta sus parámetros (pesos y sesgos) para mini-

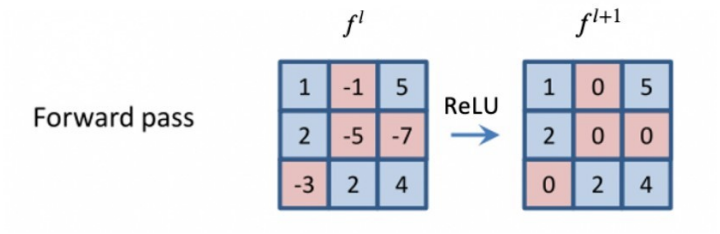

Figura 2.2: Proceso de *forward pass* en GDP

mizar el error de predicción mediante un proceso llamado *backpropagation*, donde los errores se retropropagan desde la capa de salida hasta las capas anteriores, ajustando los parámetros en el camino.

GDP modifica este proceso de retropropagación para mejorar la visualización de las características que son importantes para una predicción específica. En lugar de simplemente retropropagar los gradientes como en el *backpropagation* estándar, GDP aplica una serie de reglas para filtrar los gradientes y resaltar solo aquellas activaciones que son realmente significativas [15, 12].

Para ello, se realizan los siguientes pasos clave:

1. **Paso hacia adelante (*Forward pass*):** Primero, una imagen de entrada se pasa a través de la red neuronal como de costumbre, activando diferentes neuronas en cada capa en función de las características de la imagen. Un ejemplo de este proceso se puede apreciar en la figura 2.2.

2. **Paso hacia atrás con filtrado (*Backward pass* con filtrado):** Durante la retropropagación, GDP introduce una modificación crucial: solo se permiten pasar los gradientes positivos. Esto significa que, tanto en la dirección hacia adelante como hacia atrás, solo se consideran las activaciones que contribuyen positivamente a la salida. En términos simples, si una neurona tiene una activación negativa en la dirección hacia adelante o un gradiente negativo en la dirección hacia atrás, su gradiente se establece en cero y no se propaga más allá.

3. **Visualización de características:** El resultado de este proceso es un mapa de características más claro y menos ruidoso que destaca las partes de la imagen de entrada que son más importantes para la predicción del modelo.

Cabe destacar que este método es relativamente simple de implementar y puede integrarse fácilmente en el proceso de retropropagación estándar de

una red neuronal, sin necesidad de modificaciones significativas en la arquitectura del modelo. Además, al enfocar la retropropagación solo en activaciones positivas, GDP puede detectar características detalladas y sutiles que otros métodos podrían pasar por alto [15, 12].

Grad-CAM (*Gradient-weighted Class Activation Mapping*)

Grad-CAM es una técnica de *backpropagation* que utiliza los gradientes de cualquier clase objetivo que fluye en las últimas capas convolucionales de una red neuronal para producir un mapa de activación ponderado, el cual resalta las regiones importantes de la imagen de entrada que la red considera significativas para predecir una clase específica. Este método fue propuesto por Selvaraju en 2017 y ha sido ampliamente adoptado debido a su simplicidad y efectividad [16].

El procedimiento de Grad-CAM se puede desglosar en los siguientes pasos [16]:

1. **Propagación hacia adelante:** La imagen de entrada se pasa a través de la red neuronal para obtener las activaciones en las últimas capas convolucionales y la salida final de la red.

2. **Cálculo de gradientes:** Se calcula el gradiente de la puntuación de la clase de interés con respecto a las activaciones de la última capa convolucional. Estos gradientes representan la importancia de cada unidad en la capa convolucional para la clase objetivo.

3. **Ponderación de activaciones**: Los gradientes obtenidos se promedian sobre el ancho y alto de la activación para obtener unos pesos. Estos pesos se utilizan para realizar una combinación lineal de los mapas de características de la última capa convolucional.

4. **Generación del mapa de activación**: La combinación lineal de los mapas de características, ponderada por los gradientes, produce un mapa de activación de clase que resalta las regiones de la imagen que más influyen en la predicción de la clase específica.

5. **Interpolación y visualización**: Finalmente, el mapa de activación se interpola a la resolución de la imagen de entrada y se superpone a esta para una mejor visualización. Este es el resultado que ve el usuario.

En la figura 2.3, se muestra una imagen de ejemplo donde se aprecian los mapas de activación resultantes de este proceso para las clases ‘Perro‘ y ‘Gato‘.

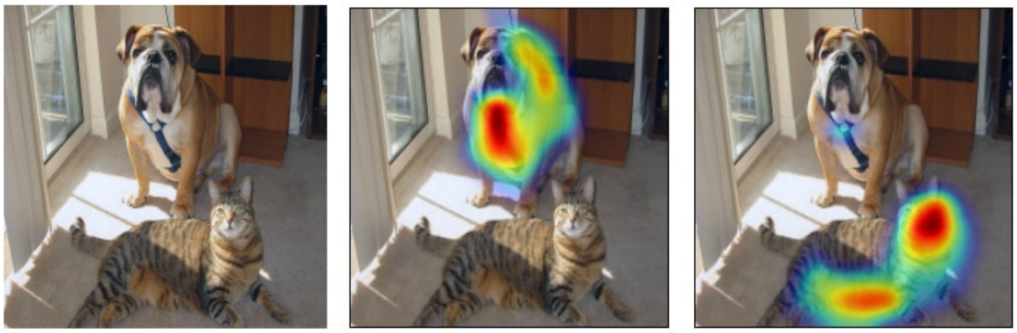

Figura 2.3: Ejemplo de los mapas de activación de Grad-CAM [16]

Grad-CAM ofrece varias ventajas. En primer lugar, proporciona explicaciones visuales intuitivas que pueden ser interpretadas fácilmente por humanos. Esto es especialmente útil en aplicaciones donde la confianza del usuario en el sistema de IA es crucial, como en diagnósticos médicos y vehículos autónomos.

A pesar de sus ventajas, Grad-CAM no está exento de limitaciones. La técnica depende de la calidad de los gradientes, que pueden ser ruidosos o poco representativos en redes muy profundas. Además, las explicaciones proporcionadas por Grad-CAM son *post-hoc*, lo que significa que es un método para producir mapas de calor que se aplica a una red neuronal ya entrenada una vez que el entrenamiento ha finalizado y los parámetros están fijos, por lo que pueden no capturar completamente el proceso de decisión de la red [16, 17].

LRP (*Layer-wise Relevance Propagation*)

Otro de los métodos basados en *backpropagation* es el *Layer-wise Relevance Propagation* (LRP) [18]. Lo que LRP hace es asignar una puntuación de relevancia a cada neurona de las capas intermedias y de entrada, proporcionando una visión detallada de cómo cada característica contribuye a la salida del modelo.

Esta técnica se basa en la idea de redistribuir la relevancia de la salida del modelo hacia las entradas, pasando por cada capa de la red. La relevancia asignada a cada neurona refleja su contribución a la decisión final del modelo. Esta redistribución se realiza siguiendo principios de conservación, donde la relevancia total se preserva a lo largo de las capas [10]. De esta forma, se garantiza que la suma de las relevancias en la capa de entrada es igual a la

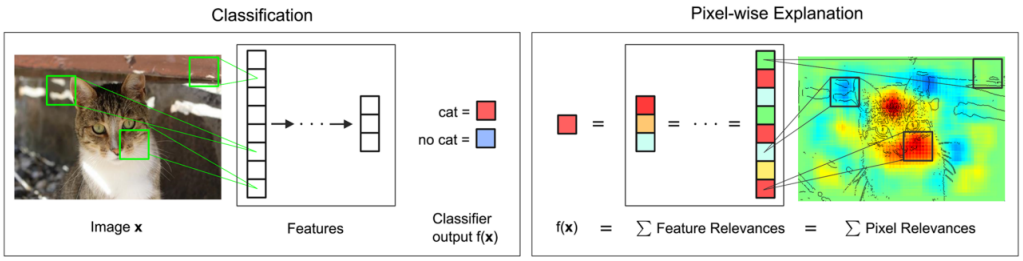

Figura 2.4: Ejemplo de salida de LRP [18]

relevancia de la salida.

El procedimiento, explicado de forma más detallada, implica los siguientes pasos [18, 19]:

- **Inicialización de relevancia**: La relevancia de la salida del modelo se inicializa según la predicción de interés. Por ejemplo, si estamos interesados en entender por qué una imagen fue clasificada como un perro, la relevancia se asigna a la puntuación de la clase 'perro'.
- **Redistribución de relevancia**: La relevancia se redistribuye desde la capa de salida hacia las capas de entrada, siguiendo reglas específicas que dependen de las funciones de activación y los pesos de las conexiones.
- **Mapeo de relevancia**: El resultado es un mapa de relevancia en la capa de entrada que destaca las características de entrada que más contribuyen a la predicción. Este mapa puede visualizarse como una imagen de calor superpuesta sobre la entrada original, proporcionando una interpretación intuitiva de las decisiones del modelo.

En la figura 2.4 se muestra un ejemplo de salida de LRP, donde se visualiza las contribuciones de los píxeles a la predicción.

En la medicina, por ejemplo, LRP permite identificar qué partes de una imagen de rayos X son más relevantes para el diagnóstico de una enfermedad, facilitando la interpretación por parte de los profesionales de la salud [20].

Entre las ventajas de LRP se incluyen su capacidad para proporcionar explicaciones detalladas y su aplicabilidad a una amplia gama de modelos y tareas. Sin embargo, también presenta limitaciones, como la necesidad de un conocimiento profundo de la estructura del modelo y la posible complejidad computacional en redes muy profundas [18, 19, 20].

2.1.2. Métodos basados en perturbación

Los métodos basados en perturbación implican modificar de manera sistemática las entradas del modelo y observar los cambios en las salidas para identificar qué partes de las entradas son más relevantes para las decisiones del modelo. La idea central es que al alterar una característica y notar un cambio significativo en la predicción, se puede inferir que dicha característica es importante para el modelo [5, 21, 22, 23].

Los métodos de perturbación suelen diferenciarse en tres tipos:

- ***Occlusion sensitivity***: Este método consiste en ocultar o perturbar partes específicas de la entrada para ver cómo afecta la salida del modelo.
- ***Perturbation analysis***: Se introducen perturbaciones aleatorias en diferentes partes de la entrada y se observa cómo varían las predicciones.
- ***Feature ablation***: Similar al *occlusion sensitivity*, pero implica la eliminación de características específicas de la entrada para estudiar su impacto en la predicción.

A continuación, se describe más en detalle el método de *occlusion sensitivity*, ya que es uno de los basados en perturbación más utilizados. El procedimiento de este método se puede desglosar en los siguientes pasos [24, 25, 26]:

1. **Selección de la imagen de entrada**: Se selecciona la imagen que se quiere analizar. Esta imagen será perturbada sistemáticamente para evaluar la sensibilidad del modelo a diferentes partes de la misma.
2. **Definición de la región de oclusión**: Se define un tamaño fijo para la región que será ocluida. Esta región puede ser un cuadrado de tamaño $k \times k$ píxeles, por ejemplo.
3. **Oclusión**: Se recorre la imagen desplazando la región de oclusión de manera sistemática. En cada posición, se reemplaza la región con un valor constante, como el promedio de los píxeles de la imagen o un valor de gris neutro.
4. **Evaluación del modelo**: Para cada posición de la región de oclusión, se evalúa el modelo con la imagen perturbada y se registra la predicción. Esto permite medir cómo la oclusión de esa parte específica afecta la salida del modelo.

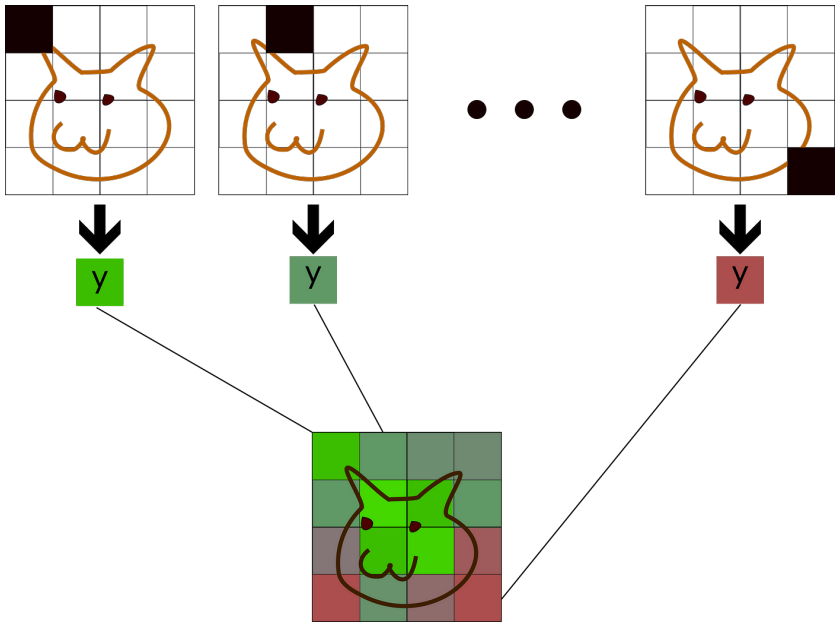

Figura 2.5: *Occlusion sensitivity* aplicado a una imagen [27]

5. **Generación del mapa de sensibilidad**: Se crea un mapa de sensibilidad que muestra la importancia de cada región de la imagen en la predicción del modelo. Las áreas donde la oclusión causa una gran variación en la predicción se consideran más importantes.

Para ilustrar el método de sensibilidad a la oclusión (*occlusion sensitivity*), se muestra la imagen de la figura 2.5, donde se considera una red neuronal convolucional entrenada para clasificar imágenes tal que queremos entender qué partes de una imagen son más importantes para que el modelo la clasifique correctamente.

Este método cuenta con ciertas ventajas y limitaciones, que se comentan a continuación:

Ventajas:

- **Simplicidad**: Es un método intuitivo y fácil de implementar.
- **Modelo agnóstico**: No requiere acceso a los parámetros internos del modelo, lo que lo hace aplicable a una amplia variedad de modelos de caja negra.
- **Visualización clara**: Proporciona un mapa visual que es fácil de interpretar, mostrando las áreas más relevantes de la entrada.

Limitaciones:

- **Coste computacional**: Puede ser computacionalmente costoso, especialmente para imágenes grandes o modelos complejos, ya que requiere evaluar el modelo múltiples veces.
- **Resolución**: La precisión del mapa de sensibilidad depende del tamaño de la región de oclusión, ya que regiones muy grandes pueden perder detalles importantes, mientras que regiones muy pequeñas aumentan el coste computacional.

2.2. Destilación de modelos

La destilación de modelos implica entrenar un modelo separado, un modelo más sencillo 'caja blanca', es decir, interpretable de forma intrínseca, que imita el comportamiento de una *DNN* ('caja negra'). Este modelo de caja blanca está diseñado para capturar las reglas de decisión y las características de entrada que influencian las salidas de la red neuronal, haciéndolo inherentemente explicable.

Al destilar el conocimiento de una *DNN* compleja en un modelo más interpretable, los usuarios pueden obtener una comprensión más clara de cómo opera el sistema de IA y toma decisiones. La destilación de modelos es particularmente útil para usuarios que requieren una representación más transparente y comprensible del proceso de toma de decisiones de la red neuronal. Este método presenta varias aproximaciones que se presentan a continuación.

2.2.1. Aproximación local

Los métodos basados en aproximación local buscan interpretar el comportamiento de modelos complejos mediante la aproximación de su funcionamiento en regiones específicas del espacio de características [38]. A diferencia de los enfoques globales, que intentan explicar el modelo en su totalidad, los métodos de aproximación local se centran en entender cómo el modelo toma decisiones en torno a un punto de interés particular.

Entre los métodos más destacados de aproximación local se encuentran LIME (*Local Interpretable Model-agnostic Explanations*) [37] y SHAP (*SHapley Additive exPlanations*) [41], los cuales se describen a continuación.

LIME

El procedimiento de LIME se puede desglosar en los siguientes pasos:

1. **Selección de la instancia de interés**: Se elige la instancia (entrada) para la cual se desea obtener una explicación.

2. **Generación de datos sintéticos**: Se generan datos sintéticos perturbando las características de la instancia de interés. Estos datos sintetizados se utilizan para entender cómo el modelo responde a variaciones en las características.

3. **Evaluación del modelo complejo**: Se evalúa el modelo complejo en los datos sintéticos generados para obtener las predicciones correspondientes.

4. **Ajuste del modelo interpretable**: Se ajusta un modelo interpretable (como una regresión lineal) utilizando las características perturbadas y las predicciones del modelo complejo. Este modelo local captura la relación entre las características y la predicción en la vecindad de la instancia de interés.

5. **Generación de la explicación**: Se interpretan los coeficientes del modelo interpretable para entender la importancia de cada característica en la predicción del modelo complejo.

En la figura 2.6 se muestra un ejemplo ilustrativo para representar cómo funciona LIME. La función de decisión modelo complejo (red neuronal) es desconocida para LIME y está representada por el fondo azul/rosa, que no puede ser bien aproximada por un modelo lineal. La cruz roja en negrita es la instancia que se está explicando. LIME toma muestras de instancias, obtiene predicciones usando la función de decisión de la red neuronal y las pondera por la proximidad a la instancia que se está explicando (representada aquí por el tamaño). Por último, la línea discontinua de la imagen es la explicación aprendida por LIME, que es localmente (pero no globalmente) fiel. Es decir, sirve para explicar cómo se comporta el modelo con la muestra escogida pero no se puede generalizar al resto de muestras [37, 38].

En la figura 2.7 se muestra un modelo que predice que un paciente tiene gripe, y LIME destaca los síntomas en el historial del paciente que llevaron a esta predicción. 'Estornudo' y 'dolor de cabeza' apoyan la predicción de gripe, mientras que 'no fatiga' va en contra. Esto permite al médico decidir si confiar en la predicción del modelo.

Una de las principales ventajas de LIME es que puede generar explicaciones para datos tabulares, textuales e imágenes; sin embargo, generar y evaluar muchas perturbaciones puede ser computacionalmente costoso.

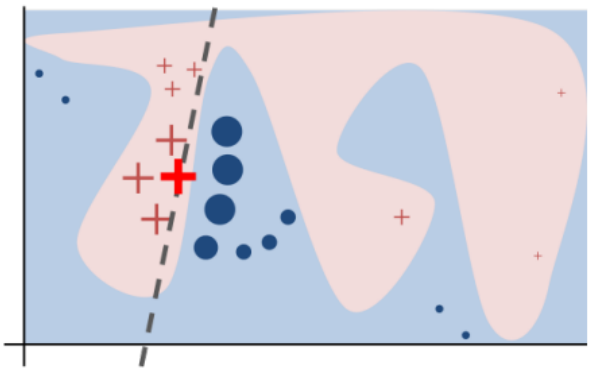

Figura 2.6: Primer ejemplo del funcionamiento de LIME [37]

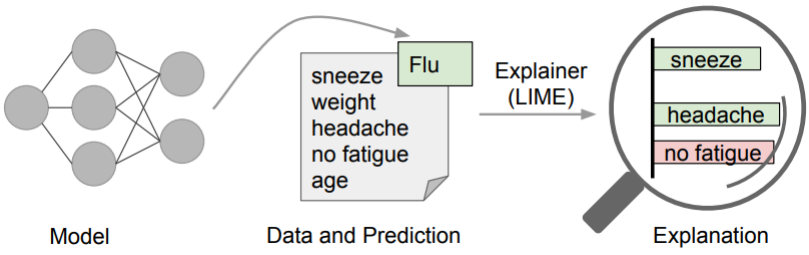

Figura 2.7: Segundo ejemplo del funcionamiento de LIME [37]

SHAP (*SHapley Additive exPlanations*)

SHAP es un método basado en la teoría de valores de Shapley, una solución de la teoría de juegos para distribuir de manera justa la contribución de cada característica en la predicción del modelo [40]. Propuesto por Lundberg y Lee en 2017, SHAP combina propiedades deseables de consistencia y equidad, proporcionando una explicación aditiva que suma las contribuciones de cada característica para igualar la predicción del modelo [41].

El procedimiento de SHAP se puede resumir en los siguientes pasos:

1. **Modelado de las contribuciones**: Para cada característica de la instancia de interés se calcula su contribución a la predicción basada en la teoría de valores de Shapley. Esto implica evaluar cómo cambia la predicción del modelo al añadir cada característica, considerando todas las posibles combinaciones de características.

2. **Generación de valores de Shapley**: Se generan los valores de Shapley para cada característica, que representan la contribución media marginal de esa característica a la predicción del modelo.

3. **Sumatoria de las contribuciones**: La suma de los valores de Shapley de todas las características debe igualar la diferencia entre la predicción del modelo y la media base.

4. **Visualización de la explicación**: Se presentan los valores de Shapley para cada característica, permitiendo una interpretación clara de la importancia relativa de cada una.

Una de las propiedades esenciales de los valores de Shapley es que siempre suman la diferencia entre el resultado del modelo cuando se consideran todas las características y el resultado cuando no se considera ninguna. Por tanto, en el contexto de los modelos de aprendizaje automático esto significa que los valores SHAP de todas las características de entrada suman la diferencia entre lo que el modelo esperaría por defecto y lo que realmente predice para una determinada predicción. Un modo sencillo de visualizar esto es a través de la imagen de la figura 2.8, que comienza con la expectativa inicial del modelo para el precio de una casa y luego muestra cómo cada característica contribuye al cambio en la predicción final del modelo.

Por último, cabe destacar que SHAP es un método agnóstico, por lo que puede ser aplicado a cualquier modelo independientemente de su tipo, pero también tiene la desventaja de que calcular valores de Shapley puede ser computacionalmente costoso, especialmente para modelos complejos y conjuntos de datos grandes [41, 42].

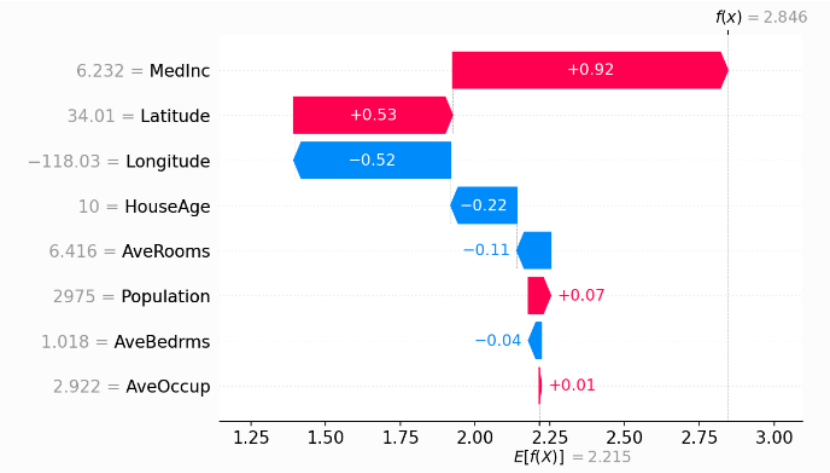

Figura 2.8: Contribución de cada característica a la predicción final [39]

2.2.2. Traducción del modelo a un árbol de decisión

Explicar por qué una red neuronal entrenada toma una decisión de clasificación específica en un caso de prueba concreto es complicado. Por tanto, si pudiéramos transformar el conocimiento adquirido por la red neuronal en un modelo que se base en decisiones jerárquicas sería mucho más sencillo explicar una decisión particular. Es por eso que este método busca utilizar una red neuronal entrenada para crear un tipo de árbol de decisión que generaliza mejor que uno derivado directamente de los datos de entrenamiento, facilitando así la explicación de las decisiones de la red neuronal a través del árbol de decisión [28].

En el trabajo de Frosst y Hinton (2017) [29], se propone el concepto de ‘árboles de decisión suaves‘, los cuales emplean el descenso por gradiente estocástico para su entrenamiento, basándose en las predicciones y filtros aprendidos de una red neuronal específica. Además, se ha podido comprobar que los árboles de decisión suaves superan en rendimiento a los árboles convencionales entrenados directamente sobre el conjunto de datos dado, aunque su desempeño sigue siendo inferior al de las redes neuronales preentrenadas.

En la figura 2.9 se muestra una imagen de un árbol de decisión suave de profundidad 4 entrenado con el conjunto de datos MNIST. Los nodos internos contienen los filtros aprendidos, mientras que en las hojas se representan las distribuciones de probabilidad sobre las clases. Se indican las clasificaciones más probables en cada hoja, así como las predicciones probables en cada conexión entre nodos. Por ejemplo, en el nodo interno más a la derecha, las únicas clasificaciones posibles son 3 u 8, lo que demuestra que el filtro

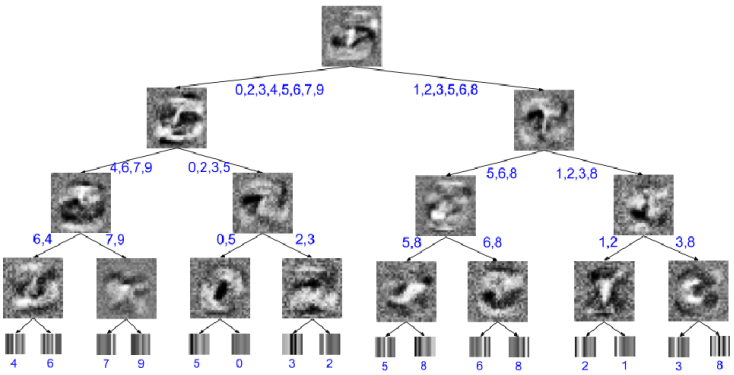

Figura 2.9: Ejemplo de árbol de decisión suave [29]

aprendido se especializa en distinguir entre estos dos dígitos. Esto se logra detectando áreas que podrían conectar los extremos del 3 para formar un 8.

2.3. Métodos intrínsecos

Los métodos intrínsecos se centran en desarrollar redes neuronales que sean inherentemente explicables, es decir, que proporcionen explicaciones junto con sus salidas [30]. Estos modelos están diseñados para optimizar tanto el rendimiento como la calidad de las explicaciones que producen.

A diferencia de las explicaciones *post hoc*, los métodos intrínsecos integran las explicaciones en la arquitectura o el proceso de entrenamiento del modelo, proporcionando a los usuarios información adicional durante el entrenamiento del modelo. Si bien los métodos intrínsecos requieren conocimientos especializados y recursos computacionales adicionales, ofrecen una ventaja única al proporcionar explicaciones en tiempo real, lo que los hace valiosos para aplicaciones críticas para la toma de decisiones [31].

A continuación se va a explicar en detalle las dos ramas principales de este tipo de métodos.

2.3.1. Mecanismos de atención

Los mecanismos de atención [31] permiten que un modelo de IA se enfoque en partes específicas de la entrada al tomar decisiones, similar a cómo los humanos se concentran en ciertas partes de una escena visual o un texto mientras ignoran otras. En términos técnicos, un mecanismo de atención asigna pesos a diferentes partes de la entrada, indicando la importancia relativa de cada parte en la decisión final del modelo. Estos pesos se calculan

durante el proceso de entrenamiento y se ajustan de manera que el modelo pueda enfocarse en las características más relevantes para la tarea específica.

En un modelo de procesamiento del lenguaje natural, *Natural Language Processing (NLP)* [32], por ejemplo, un mecanismo de atención puede asignar pesos a diferentes palabras en una oración para determinar cuáles son más relevantes para predecir la siguiente palabra o clasificar la oración. De manera similar, en visión por computador puede enfocarse en regiones específicas de una imagen que son cruciales para la clasificación o detección de objetos.Al enfocarse en las características más relevantes, los mecanismos de atención pueden mejorar significativamente el rendimiento del modelo, permitiéndole hacer predicciones más precisas y eficientes. Además, los pesos de atención pueden ser visualizados para mostrar qué partes de la entrada influyeron más en la decisión del modelo, proporcionando una explicación intuitiva y fácil de entender para los usuarios humanos [31].

En el campo de la visión por computador, una aplicación popular de los mecanismos de atención es la generación de mapas de atención que resaltan las regiones de una imagen que son más importantes para una tarea de clasificación o detección. Por ejemplo, al clasificar una imagen de un gato, un mecanismo de atención puede enfocarse en las características distintivas del gato, como sus orejas y ojos, mientras ignora el fondo, por lo que facilitan en gran medida el poder entender las decisiones de estos modelos.

A pesar de sus ventajas, los mecanismos de atención no están exentos de limitaciones y desafíos. Uno de los principales desafíos es que, aunque los pesos de atención proporcionan una indicación de qué partes de la entrada son importantes, no siempre explican completamente por qué esas partes son relevantes. Además, en algunos casos, los mecanismos de atención pueden ser difíciles de interpretar si los pesos asignados no corresponden claramente a características comprensibles por humanos.

Otra limitación es el coste computacional adicional asociado con la implementación de mecanismos de atención, especialmente en modelos grandes y complejos, lo que puede suponer un obstáculo en aplicaciones donde los recursos computacionales son limitados [33, 34].

En la figura 2.10 se muestra una imagen de cómo actuaría el mecanismo de atención en el ámbito del procesamiento de lenguaje natural, concretamente en un problema de traducción.

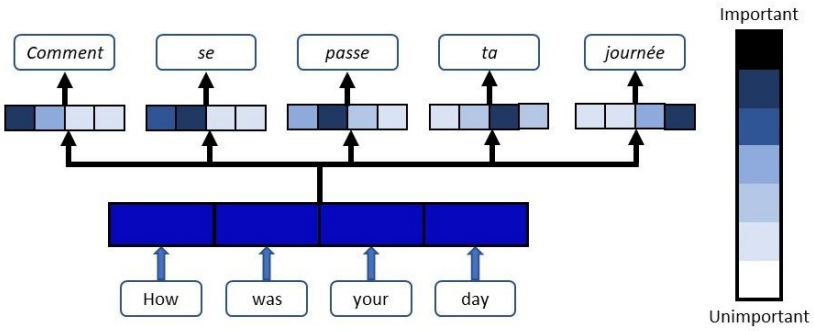

Figura 2.10: Mecanismo de atención en una tarea de *NLP*

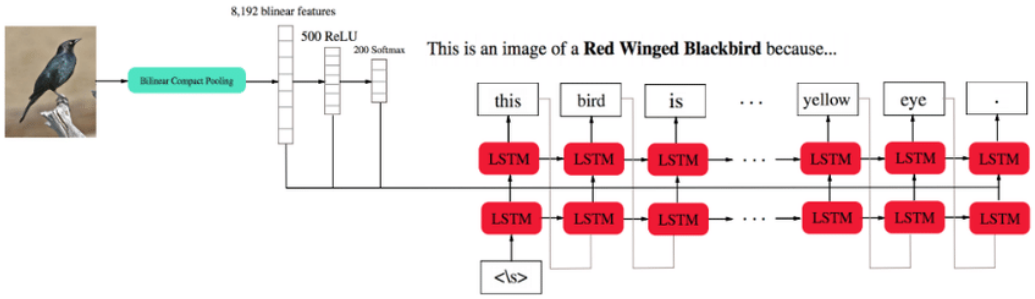

Figura 2.11: Arquitectura de red neuronal con entrenamiento conjunto [36]

2.3.2. Entrenamiento conjunto para generar explicaciones textuales

El entrenamiento conjunto implica la modificación de la arquitectura original de una red neuronal para incluir un componente generador de explicaciones. Este componente se entrena simultáneamente con el modelo principal, de manera que ambos aprenden a la par. La generación de explicaciones en lenguaje natural es uno de los resultados más destacados de este enfoque, ya que presenta las razones detrás de las decisiones del modelo de una manera comprensible para los usuarios, sin requerir conocimientos técnicos avanzados. Las explicaciones pueden generarse palabra por palabra, similar a las tareas de generación de secuencias, o seleccionarse entre múltiples opciones candidatas [35].

En la figura 2.11 se muestra una imagen con un ejemplo de arquitectura de red neuronal donde se aplica el entrenamiento conjunto para generar explicaciones.

2.4. Evaluación de las técnicas de explicabilidad

2.4.1. Métricas de evaluación

La evaluación de las técnicas de explicabilidad en IA se centra en medir la capacidad de un modelo para proporcionar explicaciones comprensibles y útiles sobre sus predicciones. Esta evaluación es esencial para determinar si las técnicas de explicabilidad cumplen con su objetivo de hacer las decisiones de la IA más transparentes y accesibles. Existen varios criterios y enfoques para evaluar estas técnicas, siendo las más populares y utilizadas las que se comentan a continuación.

Causalidad: La causalidad es una métrica cuantitativa que busca identificar las razones detrás de una predicción. Por ejemplo, se puede generar un texto que explique por qué el método predice un resultado específico a partir de ciertos insumos. Además, los ejemplos contrafactuales ayudan a entender por qué no se produjo un resultado particular [43].

Infidelidad: La infidelidad mide qué tan bien la explicación captura los cambios en la predicción del modelo cuando se perturba la entrada. Evalúa el grado en que la explicación se alinea con el impacto de las perturbaciones en la salida del modelo [44].

Una menor infidelidad indica que la explicación refleja con precisión cómo la entrada influye en la predicción del modelo, es decir, cambia de manera coherente con el comportamiento del modelo.

Esta métrica es útil porque asegura que las explicaciones proporcionadas por técnicas como GradCAM o Saliency correspondan bien con el proceso de toma de decisiones del modelo. Una alta fidelidad (baja infidelidad) significa que la explicación es confiable y sigue de cerca la lógica del modelo. De esta forma, la infidelidad asegura que las explicaciones sean fieles al comportamiento del modelo, haciéndolas fiables para comprender y depurar [44, 45].

Sensibilidad: La sensibilidad mide cómo cambia la explicación cuando se altera ligeramente la entrada. Evalúa si los pequeños cambios en la entrada conducen a cambios proporcionales y consistentes en la explicación [46]. Una mayor sensibilidad indica que la explicación responde a los cambios en la entrada, lo que significa que captura con precisión la reacción del modelo a estos cambios.

Esta métrica es importante porque prueba la estabilidad y la fiabilidad de las explicaciones. Una explicación sensible refleja los matices en las predicciones del modelo, asegurando que las variaciones menores en la entrada se destaquen correctamente en la explicación. De esta forma, la sensibilidad asegura que las explicaciones sean dinámicas y respondan, capturando efectivamente las complejidades del modelo [45, 46].

Por último, cabe destacar que bajos niveles de infidelidad y niveles apropiados de sensibilidad aumentan la confianza en los métodos de explicabilidad, proporcionando seguridad de que las explicaciones son significativas y precisas

2.4.2. Métodos de evaluación de técnicas aplicadas a problemas con imágenes

Estos métodos se dividen en dos categorías: métodos de evaluación manuales, que utilizan encuestas y cuestionarios, y métodos de evaluación automatizados, que analizan los resultados de los métodos de explicabilidad.

Comparación con explicaciones humanas: Un enfoque más cualitativo para evaluar la explicabilidad en problemas de imágenes es comparar las explicaciones generadas por el modelo con las explicaciones proporcionadas por expertos humanos. Este enfoque puede incluir estudios de usuarios en los que se pide a los expertos que evalúen la relevancia y la claridad de las explicaciones generadas por el modelo en comparación con sus propias interpretaciones. La coincidencia entre las explicaciones del modelo y las explicaciones humanas puede ser un indicador de la calidad y la utilidad de las técnicas de explicabilidad. Por ejemplo, se pueden analizar aspectos como la bondad, la satisfacción, la confianza, la validez y la curiosidad a partir de las respuestas. Cabe destacar que esto tiene ciertos inconvenientes, ya que se necesita evaluadores humanos, lo cual es costoso temporal y económicamente [47].

Métodos de evaluación automatizada: Algunos estudios emplean un método basado en perturbaciones para evaluar la calidad de los mapas de calor. Este enfoque consiste en comparar la calidad de los mapas de calor generados por diferentes técnicas mediante la sustitución de las regiones correspondientes por datos aleatorios y la posterior verificación de cuánto cambia la puntuación de clasificación. Según este método, cuanto mayor es el cambio en la puntuación, mejor corresponde el mapa de calor a las características discriminantes de la clase.

Por ejemplo, existe una métrica llamada c-Eval que evalúa los mapas de calor perturbando regiones no indicadas como importantes. Esta métrica mide indirectamente la precisión de los mapas de calor: a mayor valor de c, más robusto es el clasificador y más precisa es la identificación de características discriminantes de clase [48].

2.5. Problemas asociados a la explicabilidad

Uno de los problemas más grandes asociados a la explicabilidad de modelos de inteligencia artificial son los ataques adversariales (conocidos comúnmente como '*Adversarial attacks*'), los cuales se refieren a técnicas maliciosas empleadas con el propósito de engañar, manipular o confundir deliberadamente a los modelos de inteligencia artificial, con el fin de alterar sus decisiones o predicciones. Estas estrategias pueden manifestarse de diversas maneras, como la inserción de ruido imperceptible en los datos de entrada, la modificación intencionada de características clave o la manipulación de las salidas del modelo [49].

Dentro del ámbito de la explicabilidad de la inteligencia artificial, los ataques adversariales representan un desafío significativo debido a que comprometen la confianza y la seguridad de los sistemas de IA. Si un modelo de explicabilidad puede ser fácilmente engañado o manipulado, las explicaciones que proporciona pueden resultar inexactas o engañosas, lo que conlleva a decisiones erróneas en contextos críticos como la medicina, la justicia o la seguridad.

La presencia de este tipo de ataques plantea importantes interrogantes para la comunidad académica y los profesionales del campo, por lo que se deben idear estrategias para proteger los modelos que se intentan explicar contra estas amenazas, además de diseñar métodos de explicación que sean resistentes a este tipo de ataques [49].

Capítulo 3

Descripción de los datos y del modelo

3.1. Descripción de los datos

Los datos empleados para la experimentación llevada a cabo con el modelo de clasificación de imágenes provienen de un conjunto de datos específico de razas de perros, conocido como el 'Stanford Dogs Dataset' [50]. Este conjunto de datos se compone de un total de 20,580 imágenes, organizadas en 120 clases distintas de perros, con aproximadamente 150 imágenes por cada clase. La elección de este conjunto de datos no fue arbitraria, sino que obedeció a una serie de consideraciones tanto prácticas como teóricas.

El tema de las particiones de los conjuntos de entrenamiento, validación y test se abordará en detalle más adelante en el capítulo 4, dedicado al entrenamiento de la red y a la experimentación, y se discutirán las pruebas realizadas con diferentes tamaños del conjunto de datos para el entrenamiento. Es importante destacar que, en todas las pruebas que se llevarán a cabo, se utilizará el conjunto de datos de test completo como conjunto de validación.

Inicialmente, se evaluaron otros conjuntos de imágenes como posibles alternativas, tales como conjuntos de imágenes de especies de pájaros [51] o de especies de peces y algunos animales marinos [52]. Sin embargo, desde la perspectiva de un observador no especializado en fauna, la tarea de distinguir entre diferentes especies de pájaros o peces resulta considerablemente más compleja. Este factor de complejidad podría interferir en la interpretación y comprensión de los resultados obtenidos. Por lo tanto, teniendo en cuenta que el objetivo primordial de este trabajo es experimentar y evaluar técnicas de explicabilidad aplicadas a modelos de clasificación de imágenes, se consideró más conveniente utilizar un conjunto de datos que, además de ser adecuado

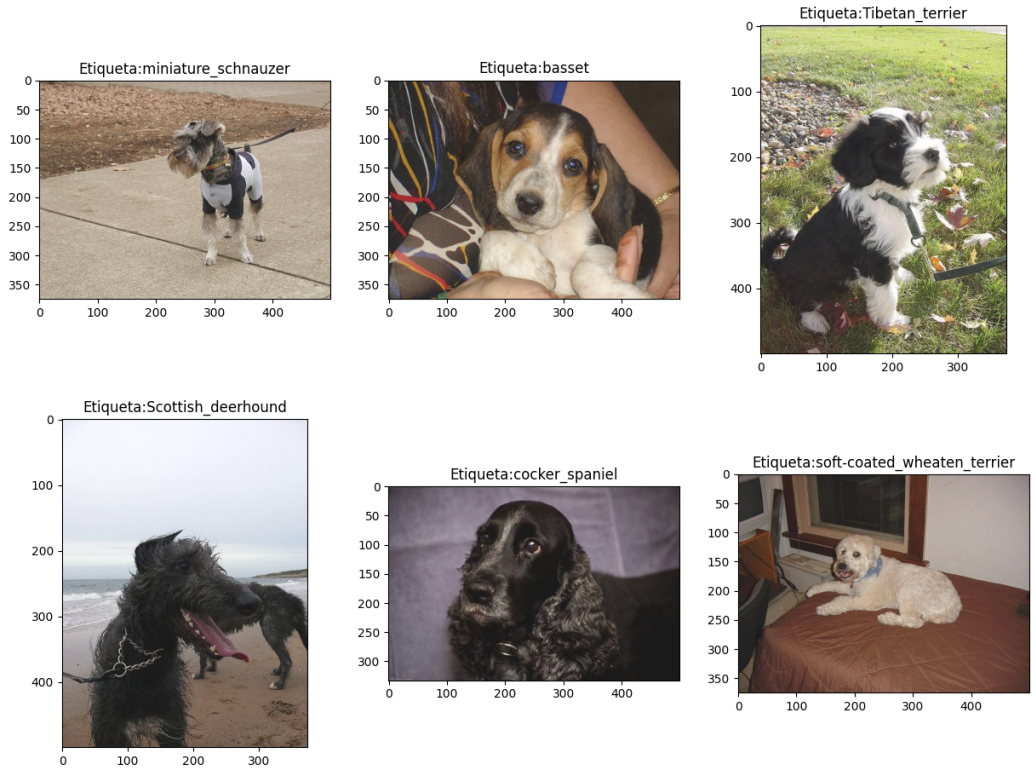

Figura 3.1: Imágenes de ejemplo del *dataset* 'Stanford Dogs'

para las métricas de evaluación utilizadas, también pudiera ser visualmente interpretado por el usuario sin necesidad de un conocimiento profundo en zoología.

El 'Stanford Dogs Dataset' ofrece una ventaja significativa en este contexto, debido a la familiaridad general que las personas tienen con las diversas razas de perros en comparación con la fauna aviar o acuática. Esto facilita no solo la implementación y prueba del modelo de clasificación, sino también la interpretación de las técnicas de explicabilidad que se aplican.

En la figura 3.1 se muestran varias imágenes de ejemplo de algunas de las clases del conjunto de datos utilizado.

En la figura 3.2 se muestra la distribución del conjunto de datos utilizado, donde se puede observar que la distribución de las imágenes en el conjunto de datos de razas de perros no es uniforme, lo que revela un cierto desbalance en el número de muestras por clase. Este desequilibrio es evidente al observar que algunas razas cuentan con un mayor número de imágenes en comparación con otras. Si bien la diferencia en el número de imágenes no es

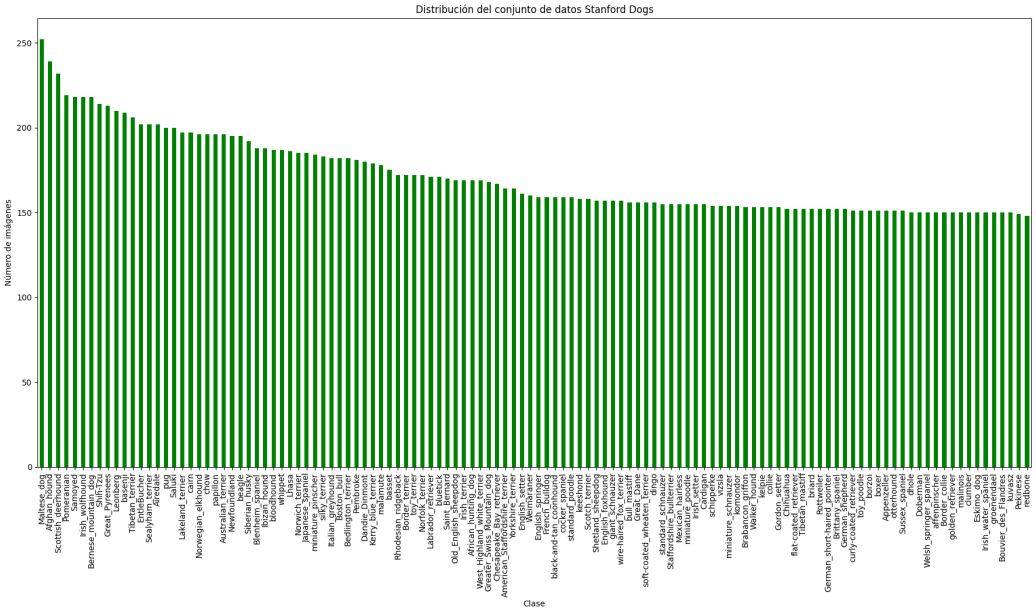

Figura 3.2: Distribución del conjunto de datos 'Stanford Dogs Dataset'

extremadamente pronunciada, se observa que la clase con más imágenes tiene alrededor de 250 muestras, mientras que la mayoría de las clases cuentan con aproximadamente 150-200 muestras.

3.2. Descripción del modelo de clasificación de imágenes

Se ha utilizado el modelo EfficientNet [53] para la clasificación de imágenes de razas de perros, específicamente el EfficientNetB0 [54], debido a sus características que lo hacen ideal para esta tarea, por lo que se comenta a continuación.

EfficientNetB0 se destaca por encontrar un equilibrio óptimo entre varias métricas clave. En primer lugar, ofrece un rendimiento sobresaliente en tareas de clasificación de imágenes, demostrando capacidades superiores incluso en comparación con modelos más grandes y complejos. Esta ventaja en el rendimiento se complementa con una precisión robusta, asegurando que las predicciones del modelo sean altamente confiables y precisas [55].

Además de su rendimiento y precisión, EfficientNetB0 se distingue por su tamaño compacto en términos de parámetros. Con solo 5.3 millones de

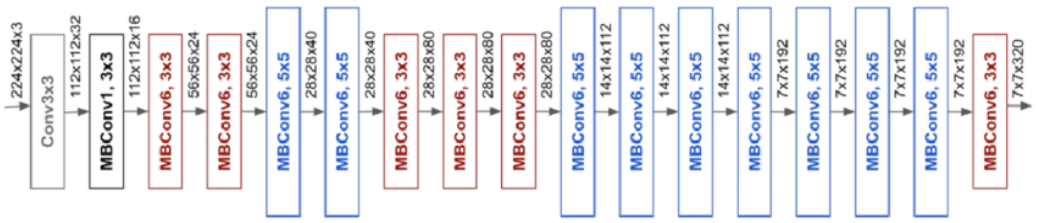

Figura 3.3: Arquitectura de la red EfficientNetB0 [56].

parámetros es considerablemente más pequeño en comparación con otros modelos avanzados como ResNet50, que contiene 25.6 millones de parámetros [53, 55]. Estas características hacen que EfficientNetB0 sea especialmente adecuado para este trabajo, con un buen equilibrio entre precisión y eficiencia computacional. En la imagen de la figura 3.3 se muestra la arquitectura de esta red.

Capítulo 4

Experimentación y resultados

En el contexto de la experimentación con el modelo EfficientNet se han aplicado diversas técnicas de explicabilidad para comprender mejor su funcionamiento. Se han utilizado los métodos GradCAM, mapas de saliencia, capas de deconvolución y retropropagación guiada. Además, se han empleado otros métodos como *occlusion sensitivity* y LIME.

4.1. Entrenamiento de la red EfficientNet

En primer lugar, antes de realizar la experimentación con las diferentes técnicas de explicabilidad utilizadas se ha entrenado la red neuronal con el conjunto de datos de perros comentado en la sección 3.1 para la clasificación. Para ello, se han utilizado diferentes tamaños del conjunto de datos para el entrenamiento, desde el 100 % hasta el 20 %, para analizar cómo varían los resultados dependiendo del tamaño. Cabe destacar que en todos los casos el tamaño del conjunto de test ha sido el 20 % del tamaño del conjunto de entrenamiento utilizado. Estos resultados se pueden observar en al tabla 4.1.

% entrenamiento	***Accuracy***	**F1**
100 %	0.832	0.823
80 %	0.835	0.832
60 %	0.823	0.819
40 %	0.804	0.797
20 %	0.772	0.753

Tabla 4.1: Resultados de diferentes porcentajes del *dataset* para entrenamiento usando EfficientNet

Los resultados muestran que la precisión (*Accuracy*) y la puntuación F1

disminuyen conforme se reduce el porcentaje del *dataset* utilizado para el entrenamiento. Aunque el uso del 100 % del *dataset* produce una precisión del 83.2 %, una reducción al 40 % aún mantiene una precisión razonable del 80.4 %. Este tamaño del conjunto es el más adecuado porque logra un buen equilibrio entre rendimiento y eficiencia computacional, reduciendo significativamente el tiempo y los recursos necesarios para el entrenamiento sin sacrificar demasiado la precisión del modelo.

4.2. Experimentación con técnicas de explicabilidad

Para la experimentación con las diferentes técnicas de explicabilidad se ha utilizado el modelo entrenado con el 40 % del conjunto de datos, ya que, como se ha comentado, es el que porporciona unos resultados más óptimos y equilibrados.

Además, se han evaluado la técnicas utilizando las métricas de infidelidad y sensibilidad, para entender qué tan bien explica este modelo. Estas métricas han sido escogidas sobre la causalidad, también explicada en el capítulo 2, debido a su enfoque específico en la coherencia y la capacidad de respuesta de las explicaciones generadas.

Inicialmente se planteó utilizar la misma imagen para presentar los resultados en cada experimentación con las diferentes técnicas y ver así cómo varían los resultados. Sin embargo, decidí realizar la experimentación utilizando diferentes imágenes, ya que considero que esto permite apreciar mejor los resultados específicos de cada técnica aplicada a imágenes concretas. Utilizando imágenes diferentes, puedo obtener conclusiones más precisas y relevantes que las que obtendría al aplicar todas las técnicas a una única imagen. Además, este enfoque demuestra que he experimentado con una variedad de imágenes, en lugar de limitarme a una sola imagen para todas las técnicas, lo que creo refuerza la validez y amplitud de el análisis.

4.2.1. GradCAM

En primer lugar, se ha experimentado y evaluado la técnica de GradCAM, la cual se ha comentado en el capítulo 2. Aplicando esta técnica a imágenes concretas del conjunto de datos podemos observar algunos resultados de los mapas de calor obtenidos en las figuras 4.1 y 4.2.

Por ejemplo, en la figura 4.1 se puede observar que el mapa de calor generado a la derecha muestra las áreas de la imagen que el modelo considera más importantes para hacer la clasificación. Las áreas más oscuras en el mapa

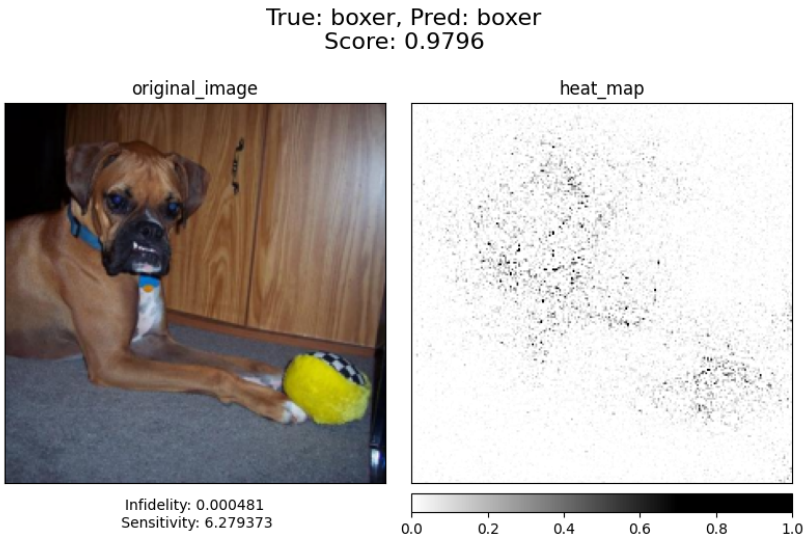

Figura 4.1: Mapa de calor obtenido con GradCAM para un ejemplo de la raza de perro *Boxer*

representan las regiones de la imagen que contribuyeron más a la decisión del modelo. Aunque el mapa de calor es algo disperso y las áreas más destacadas no parecen concentrarse en características claras del perro, hay algunas zonas que coinciden con la cabeza del perro, lo cual es esperable, ya que la cabeza y el rostro son características clave para identificar la raza. La infidelidad es baja (0.000481), lo que sugiere que la explicación proporcionada por el mapa de calor es consistente con la predicción del modelo. Por otro lado, la sensibilidad es alta (6.279373), lo cual indica que el modelo es bastante sensible a cambios en las regiones importantes identificadas en la imagen.

En la figura 4.2 se puede observar que el modelo no ha sido capaz de clasificar correctamente dicha imagen, ya que ha predicho que la raza es un *American Staffordshire terrier*, pero con una confianza bastante baja. Esto es coherente con el mapa de calor resultante, ya que no se están destacando partes clave de la raza *Boxer* que puedan llevar a una clasificación correcta.

Tras esto, se ha evaluado la propia técnica GradCAM para analizar cómo de bien explica este modelo, cuyos resultados se muestran en la tabla 4.2.

Los resultados muestran una baja infidelidad, al ser muy cercana a cero, indicando que las explicaciones proporcionadas por GradCAM son coherentes con los cambios en las predicciones del modelo. Sin embargo, la sensibilidad, al ser mayor a 0, sugiere que estas explicaciones pueden ser afectadas por

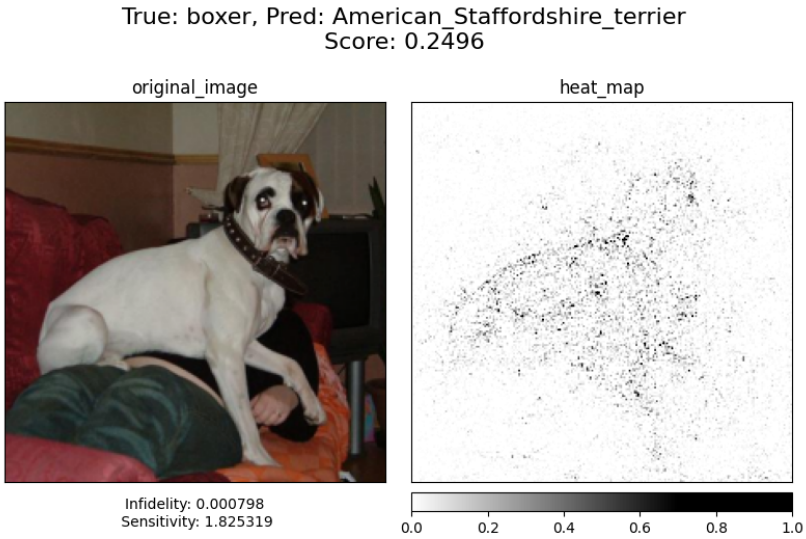

Figura 4.2: Mapa de calor obtenido con GradCAM para un ejemplo de la raza de perro *Boxer*

Infidelidad	**Sensibilidad**
0.00085	2.03509

Tabla 4.2: Valores de infidelidad y sensibilidad para la técnica GradCAM

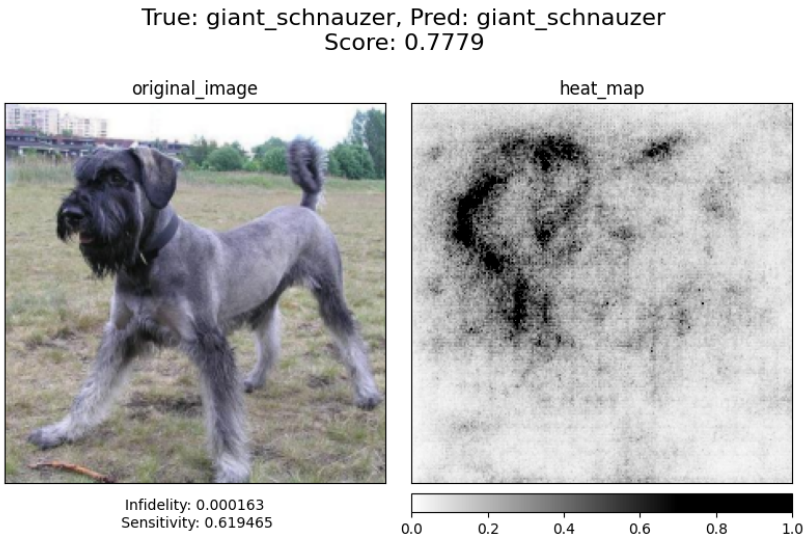

Figura 4.3: Primer ejemplo del mapa de calor obtenido con la técnica de mapas de saliencia

pequeñas variaciones en los datos de entrada, lo cual es crucial para capturar la complejidad del modelo de manera precisa.

4.2.2. Mapas de saliencia

En segundo lugar, se ha experimentado y evaluado la técnica de mapas de saliencia, la cual se ha comentado en el capítulo 2. Aplicando esta técnica a imágenes concretas del conjunto de datos podemos observar algunos resultados de los mapas de calor obtenidos en las figuras 4.3 y 4.4.

En el caso de la figura 4.3 se observa que en el mapa de saliencia las zonas de interés están más concentradas en la región de la cabeza del perro, lo cual es coherente dado que la cabeza es una característica distintiva clave para identificar la raza. La infidelidad es extremadamente baja, con un valor de 0.000163, lo que indica que el mapa de saliencia es muy consistente con la predicción realizada por el modelo. Además, la sensibilidad del modelo es de 0.619465, lo que sugiere que el modelo es moderadamente sensible a los cambios en las regiones que ha identificado como importantes.

En el caso de la figura 4.4 el modelo es capaz de identificar las zonas más características de la razas *Giant Schnauzer*, que casualmente es una raza muy similar a *Newfoundland*, por lo que el modelo, a pesar de conocer

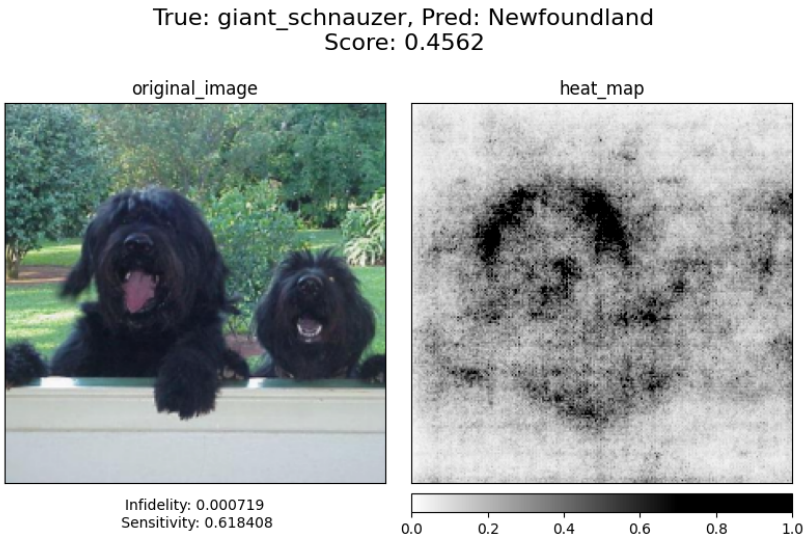

Figura 4.4: Segundo ejemplo del mapa de calor obtenido con la técnica de mapas de saliencia

cuales son las regiones más importantes de la imagen, como la zona de las orejas, es incapaz de asegurar con una alta confianza que se trate de la clase *Newfoundland*.

A continuación, se ha evaluado la técnica de saliencia para analizar qué tan bien explica este modelo. Los resultados se muestran en la tabla 4.3.

Infidelidad	**Sensibilidad**
0.00088	0.71150

Tabla 4.3: Valores de infidelidad y sensibilidad para la técnica de saliencia

Los resultados indican que la técnica de saliencia tiene una infidelidad ligeramente mayor en comparación con GradCAM (0.00088 frente a 0.00085), es decir, un aumento del 3.5 %, lo que sugiere que las explicaciones pueden no ser tan precisas en reflejar completamente los cambios en las predicciones del modelo. Además, la sensibilidad es considerablemente más baja (0.71150 frente a 2.03509), es decir, una disminución del 65 %, lo que indica que la técnica de saliencia podría ser menos sensible a las variaciones sutiles en los datos de entrada.

Dados estos resultados, si priorizamos la alineación precisa con las predicciones del modelo, GradCAM parece ofrecer mejores resultados debido a

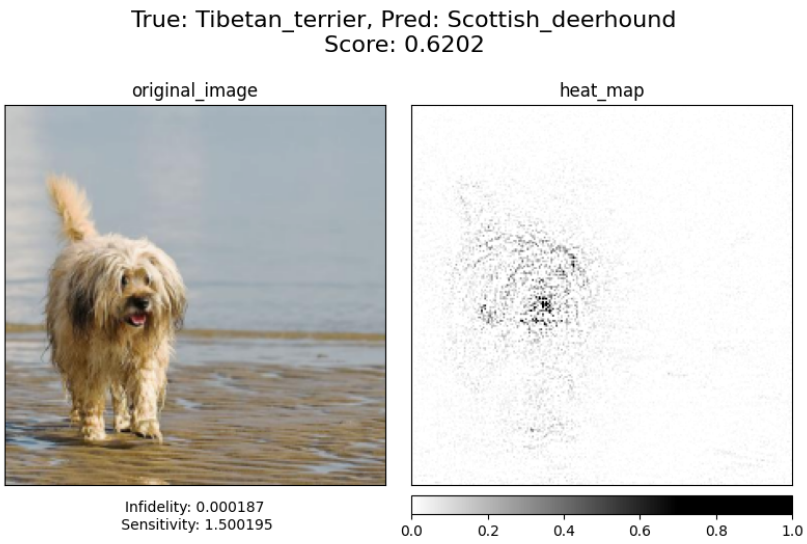

Figura 4.5: Primer ejemplo del mapa de calor obtenido con la técnica de capas de deconvolución

su menor infidelidad. Sin embargo, si la estabilidad y la robustez frente a pequeñas variaciones en los datos son más importantes, entonces la técnica de saliencia podría ser preferible debido a su menor sensibilidad.

4.2.3. Capas de deconvolución

En tercer lugar, se ha experimentado y evaluado la técnica de capas de deconvolución, la cual se ha comentado en el capítulo 2. Aplicando esta técnica a imágenes concretas del conjunto de datos podemos observar algunos resultados de los mapas de calor obtenidos en las figuras 4.5 y 4.6.

En el caso de la figura 4.5 se observa que las zonas de interés están más concentradas en la región de la cabeza del perro, lo cual es coherente dado que la cabeza es una característica distintiva clave para identificar la raza. La infidelidad es extremadamente baja, lo que indica que las explicaciones proporcionadas por la técnica son muy consistentes con las características que el modelo utilizó para tomar su decisión, incluso aunque esta decisión no sea la correcta. Además, la sensibilidad del modelo sugiere que el modelo es ciertamente sensible a los cambios en las regiones que ha identificado como importantes. Esto mismo también se observa en la imagen de la figura 4.6.

A continuación, se ha evaluado la técnica de capas de deconvolución para

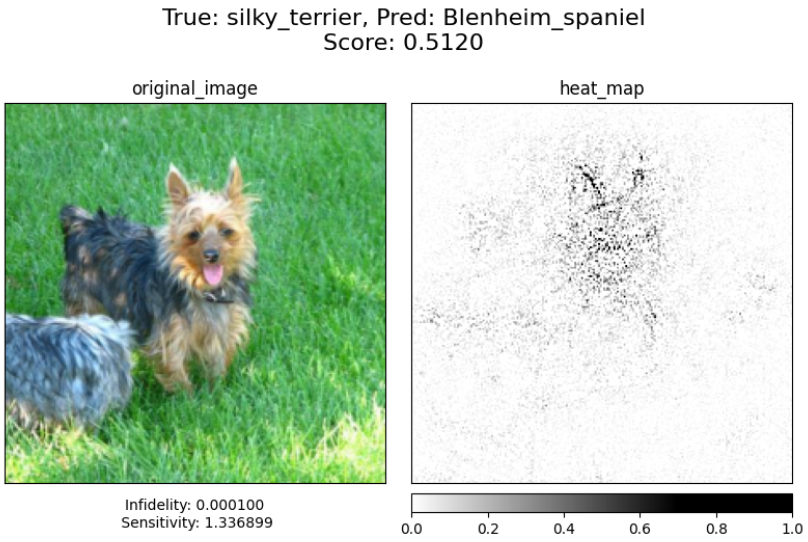

Figura 4.6: Segundo ejemplo del mapa de calor obtenido con la técnica de capas de deconvolución

analizar qué tan bien explica este modelo. Los resultados se muestran en la tabla 4.4.

Infidelidad	**Sensibilidad**
0.000572	1.466

Tabla 4.4: Valores de infidelidad y sensibilidad para la técnica de deconvolución

Los resultados obtenidos, que representan el promedio de las distintas iteraciones o perturbaciones aplicadas, indican que la técnica de capas de deconvolución muestra una infidelidad de 0.000572. Este valor sugiere que las explicaciones proporcionadas por la técnica reflejan bien las variaciones en las predicciones del modelo.

Por otro lado, la sensibilidad media de la técnica se sitúa en 1.466, lo que indica que tiene una respuesta moderada a las variaciones sutiles en los datos de entrada, en comparación con las otras técnicas analizadas. Aunque este nivel de sensibilidad podría considerarse adecuado, no necesariamente asegura que las explicaciones sean estables o robustas frente a pequeñas alteraciones en la imagen.

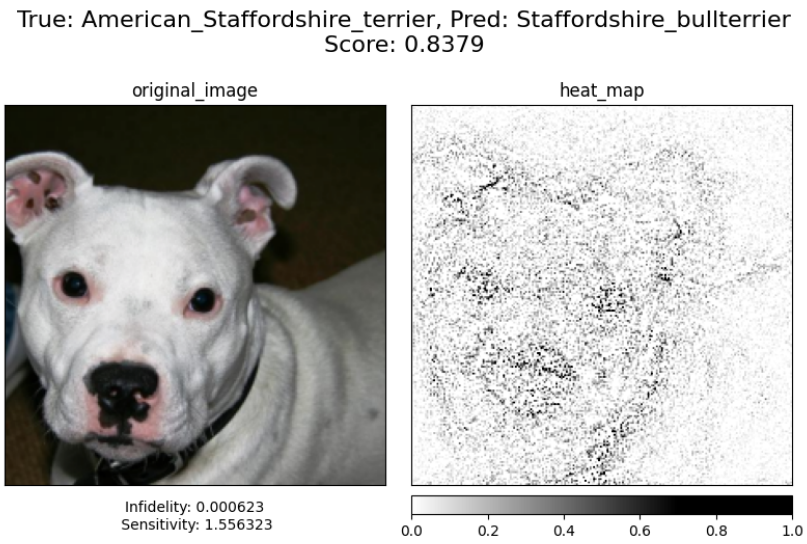

Figura 4.7: Primer ejemplo del mapa de calor obtenido con la técnica de retropropagación guiada

4.2.4. Retropropagación guiada (*Guided Backpropagation, GDP*)

En cuarto lugar, se ha experimentado y evaluado la técnica de capas de convolución, que se explicó en detalle en el capítulo 2. Aplicando esta técnica a imágenes concretas obtenemos los resultados de las figuras 4.7 y 4.8.

Como se puede apreciar en la figura 4.7, esta técnica es capaz de destacar las áreas más importantes de la raza *Staffordshire terrier* (la predicha por el modelo), aunque la clase verdadera resulta ser *American Staffordshire terrier*, que físicamente apenas es distinguible de la raza predicha, ya que son practicamente identicos, por lo que el modelo no ha sido capaz de diferenciarlas completamente.

Sin embargo, en la imagen de la figura 4.8 sí que se puede apreciar que el modelo ha conseguido predecir correctamente la clase, siendo el hocico y la zona de los ojos las más características en este caso.

A continuación, se presentan los resultados de evaluar esta técnica para comprobar qué tan bien explica el modelo de clasificación. Los resultados se muestran en la tabla 4.5.

La infidelidad media de la retropropagación guiada es de 0.000822, ligeramente superior a la de la deconvolución, que presenta un valor de 0.000572.

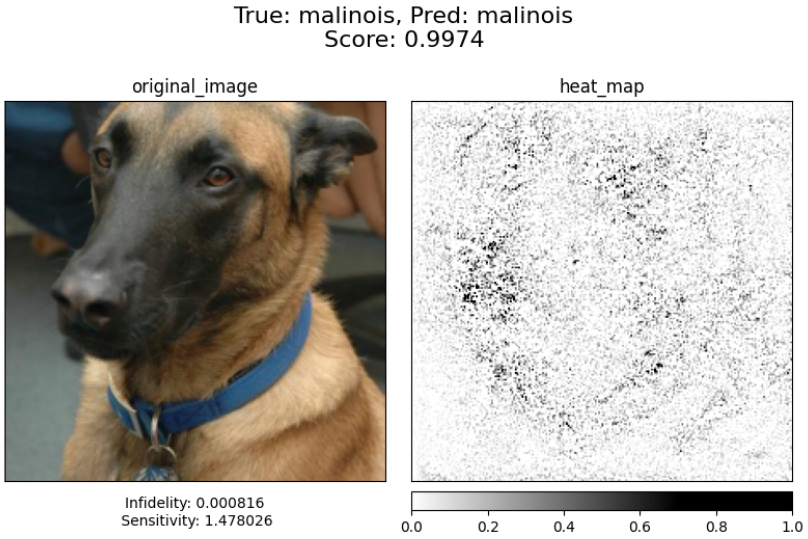

Figura 4.8: Segundo ejemplo del mapa de calor obtenido con la técnica de retropropagación guiada

Infidelidad	**Sensibilidad**
0.000822	1.457

Tabla 4.5: Valores de infidelidad y sensibilidad para la técnica de retropropagación guiada

Esto sugiere que, aunque ambas técnicas proporcionan explicaciones bastante fieles a las predicciones del modelo, la deconvolución podría tener una leve ventaja en términos de precisión y correspondencia entre las características explicadas y las decisiones del modelo. En cuanto a la sensibilidad, la retropropagación guiada muestra un valor promedio de 1.457, ligeramente inferior al 1.466 obtenido con la técnica de deconvolución. Este resultado indica que la retropropagación guiada es un poco menos sensible a pequeñas variaciones en los datos de entrada, lo que podría hacerla ligeramente más robusta en ciertos contextos en comparación con la deconvolución. Por tanto, mientras que la técnica de capas de deconvolución parece ofrecer una explicación un poco más precisa según el valor de infidelidad, la retropropagación guiada podría proporcionar una mayor robustez debido a su sensibilidad ligeramente inferior.

En la tabla 4.6 se presenta un resumen con los valores de infidelidad y sensibilidad para las diferentes técnicas de explicabilidad evaluadas, lo que permite una fácil comparativa entre las distintas técnicas en términos de ambas métricas.

Técnica	**Infidelidad**	**Sensibilidad**
GradCAM	0.00085	2.035
Saliencia	0.00088	0.711
Deconvolución	0.00057	1.466
Retropropagación guiada	0.00082	1.457

Tabla 4.6: Valores de infidelidad y sensibilidad para diferentes técnicas de explicabilidad

4.3. Otros experimentos (sin evaluación numérica de las técnicas de explicabilidad)

En esta sección, se presentan y analizan los resultados obtenidos mediante la aplicación de dos técnicas de interpretación de modelos ampliamente utilizadas en el ámbito de la inteligencia artificial: *Occlusion sensitivity* y LIME, que se explicaron en detalle en el capítulo 2.

Cabe destacar que no se ha llevado a cabo una evaluación numérica de las explicaciones generadas por estas técnicas para determinar cuán bien estas técnicas explican las predicciones del modelo. Esto se debe a que el campo de la inteligencia artificial explicable, especialmente en el contexto de imágenes, se encuentra en una etapa de investigación bastante prematura. Actualmente,

Figura 4.9: Imagen con la técnica *occlusion sensitivity* aplicada

no se han encontrado métodos de evaluación estandarizados y universalmente aceptados que sean adecuados para este tipo de problemas, lo que limita la capacidad de cuantificar de manera precisa la calidad de las explicaciones en problemas con imágenes [57].

4.3.1. *Occlusion sensitivity*

Al analizar la imagen de la figura 4.9, en la cual se ha aplicado la técnica de *occlusion sensitivity*, se pueden extraer conclusiones significativas sobre el comportamiento del modelo. *Occlusion sensitivity*, como se describe en el capítulo 2, es una técnica que consiste en modificar gradualmente la imagen original, ocultando pequeñas porciones de la misma, y observando cómo estos cambios afectan la predicción del modelo. Esto permite identificar qué partes de la imagen son más importantes para la toma de decisiones del modelo.

En la imagen de la figura 4.9, las áreas resaltadas en tonos amarillos son las que el modelo considera más importantes para realizar su predicción. Estas áreas se concentran predominantemente en los alrededores del perro, particularmente en las zonas alrededor de las piernas.

El *stride* (o paso) en el contexto de la técnica de *occlusion sensitivity* se refiere a la distancia que se desplaza el parche de oclusión sobre la imagen entre cada evaluación. En otras palabras, es el número de píxeles que se mueve el parche desde una posición a la siguiente. Un *stride* mayor significa que el parche se desplaza más lejos entre cada evaluación, cubriendo la imagen con menos superposiciones, mientras que un *stride* menor implica un movimiento

Figura 4.10: Imagen con la técnica *occlusion sensitivity* aplicada con la mitad de paso

más pequeño y, por tanto, más evaluaciones y superposiciones, lo que permite un análisis más detallado de la imagen.

Por ejemplo, en la imagen de la figura 4.10 se puede apreciar cómo varían los resultados al aplicar un *stride* (o paso) menor, concretamente la mitad del de la figura 4.9. Como se puede apreciar, los rasgos más característicos para la predicción del modelo son la zona de la cabeza y las patas, lo que es coherente, ya que son los atributos más distintivos de esta raza con respecto a cualquier otra raza de perro del conjunto de datos. Es decir, es coherente con lo que se esperaría de un modelo bien entrenado para clasificar razas de perros.

4.3.2. LIME

LIME, como se explicó en el capítulo 2, es una técnica de interpretabilidad que permite comprender las decisiones de los modelos de aprendizaje automático, proporcionando explicaciones locales que destacan las características más relevantes de una instancia específica para la predicción realizada.

Al aplicar LIME a la misma imagen utilizada con la técnica de *occlusion sensitivity*, se observaron resultados insatisfactorios que se pueden apreciar en la figura 4.11. En este caso particular, LIME no logra destacar adecuadamente las características más distintivas de la raza "Giant Schnauzer", tales como la cabeza o el pelaje específico. En lugar de ello, la técnica resalta áreas irrelevantes para la clasificación de la raza, y se enfoca en zonas de la imagen

Figura 4.11: Imagen con la técnica LIME aplicada

que no deberían influir significativamente en la predicción, a excepción de la zona de las patas, que más o menos detecta bien. Este resultado sugiere que LIME, al menos en esta instancia, no es capaz de explicar correctamente la decisión del modelo para esta imagen en particular, lo que podría suponer una limitación de la técnica en su capacidad para interpretar adecuadamente las predicciones en problemas de visión por computador.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

Este trabajo ha abordado de manera integral el estudio y la aplicación de técnicas de explicabilidad en inteligencia artificial, específicamente en el contexto de problemas relacionados con el procesamiento de imágenes. En primer lugar, se llevó a cabo una exhaustiva investigación sobre las técnicas de explicabilidad en inteligencia artificial. Durante esta fase, se exploraron los diferentes tipos de técnicas que se agrupan principalmente en métodos de visualización, destilación de modelos y métodos intrínsecos, lo que proporcionó una comprensión amplia del panorama actual en el campo.

A lo largo del proceso, se adquirió un profundo entendimiento sobre el funcionamiento de diversas técnicas de explicabilidad, lo que resultó fundamental para tomar decisiones informadas en la implementación de dichas técnicas durante la fase de experimentación. Este conocimiento permitió enfrentar de manera más efectiva los desafíos prácticos asociados con la aplicación de estas técnicas en un entorno real.

El estudio también incluyó una revisión de las métricas de evaluación utilizadas para medir la eficacia de las técnicas de explicabilidad, con el objetivo de determinar en qué medida estas técnicas explican adecuadamente las predicciones realizadas por los modelos. Este análisis fue bastante útil para evaluar la validez y la utilidad de las explicaciones generadas por las técnicas implementadas.

Además, se investigó sobre uno de los desafíos más críticos en el campo de la explicabilidad, los ataques adversariales. Se discutió cómo estos ataques pueden comprometer la integridad de las explicaciones generadas, destacando la necesidad de desarrollar técnicas más robustas y seguras frente a estas amenazas.

Posteriormente, se describió el conjunto de datos utilizado en la experimentación y se detalló el modelo seleccionado para la aplicación de las técnicas de explicabilidad, 'EfficientNetB0'. Durante esta fase, surgieron algunos desafíos, particularmente en la evaluación de técnicas como LIME y *occlusion sensitivity*, lo que subraya la complejidad y las limitaciones inherentes a estas metodologías.

En conclusión, el desarrollo de este trabajo ha resultado en la adquisición de nuevos conocimientos y habilidades en el ámbito de la inteligencia artificial explicable. Estos avances permitieron superar las dificultades encontradas y cumplir con los objetivos propuestos. No obstante, las limitaciones identificadas durante la experimentación resaltan la necesidad de continuar investigando y desarrollando nuevas técnicas en este campo, con el fin de mejorar la transparencia y la confiabilidad de los modelos de inteligencia artificial en aplicaciones reales.

5.2. Relación del trabajo desarrollado con los estudios cursados

El desarrollo de este trabajo ha demostrado una relación directa y profunda con los conocimientos y habilidades adquiridos durante el Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital. Los estudios cursados han proporcionado una base sólida en varias áreas clave que han sido fundamentales para la comprensión y ejecución del proyecto.

En primer lugar, las asignaturas relacionadas con redes neuronales y aprendizaje profundo han sido cruciales para llevar a cabo el análisis y la implementación de técnicas de explicabilidad en redes neuronales convolucionales. La que más influencia ha tenido ha sido Aplicaciones de Reconocimiento de Formas. El conocimiento teórico sobre cómo funcionan estas redes, adquirido durante el máster, ha sido esencial para entender técnicas como la retropropagación guiada y LIME.

Por otro lado, la habilidad para programar y aplicar técnicas avanzadas de análisis, fortalecida a lo largo del máster, ha sido vital para implementar y experimentar con diversas metodologías de explicabilidad. La programación, particularmente en lenguajes como Python, ha permitido la implementación eficiente de los algoritmos y técnicas estudiados, facilitando la realización de los experimentos y el análisis de resultados.

Por tanto, este trabajo ha sido una oportunidad para aplicar de manera integral los conocimientos adquiridos durante el máster, consolidando así la conexión entre la teoría y la práctica en el campo de la inteligencia artificial.

5.3. Trabajos futuros

Este trabajo ha sentado una base sólida para el estudio y la aplicación de técnicas de explicabilidad en redes neuronales aplicadas al procesamiento de imágenes. Sin embargo, existen varias direcciones en las que se podría continuar la investigación para ampliar y profundizar los conocimientos obtenidos.

En primer lugar, se propone el desarrollo de un modelo híbrido que combine redes neuronales convolucionales con redes neuronales recurrentes para generar explicaciones textuales de las clasificaciones realizadas por el modelo. Este enfoque permitiría no solo identificar las regiones o características específicas de la imagen que más han influido en la predicción, sino también articular dichas influencias en un formato textual comprensible. La integración de ambos tipos de redes neuronales tiene el potencial de mejorar la interpretabilidad del modelo, proporcionando explicaciones más accesibles y detalladas, que podrían ser especialmente útiles en contextos donde la transparencia y la justificación de las decisiones en un lenguaje natural son cruciales.

Además, aunque este trabajo ha utilizado técnicas como GradCAM, mapas de saliencia, LIME y *occlusion sensitivity*, entre otros, para evaluar la explicabilidad de las redes neuronales, se plantea como futuro trabajo realizar una evaluación numérica más exhaustiva de estas técnicas. Esto implicaría no solo una medición más precisa de su efectividad, sino también la experimentación con otras técnicas como SHAP. SHAP, en particular, ofrece un marco teórico sólido basado en teoría de juegos, que podría aportar una nueva perspectiva sobre la importancia relativa de diferentes características en las predicciones del modelo. Comparar estas técnicas en términos de robustez, fidelidad y consistencia contribuiría significativamente al entendimiento de sus fortalezas y debilidades en distintos escenarios.

Por último, se considera fundamental experimentar con otros conjuntos de datos de imágenes para evaluar cómo se comportan las técnicas de explicabilidad en diferentes contextos. Los resultados obtenidos en este trabajo están basados en un conjunto de datos específico, por lo que extender la investigación a otras bases de datos permitiría comprobar la generalizabilidad de las técnicas de explicabilidad utilizadas. Este enfoque ayudaría a identificar si las conclusiones alcanzadas son aplicables a una variedad más amplia de problemas de clasificación de imágenes.

Estos trabajos futuros propuestos tienen como objetivo no solo mejorar la capacidad de los modelos para explicar sus decisiones, sino también evaluar y validar estas técnicas en una gama más amplia de escenarios y con diversas metodologías, lo que podría llevar a desarrollos significativos en el campo de la inteligencia artificial explicable.

Agradecimientos

Me gustaría expresar mi agradecimiento a ValgrAI - *Valencian Graduate School and Research Network for Artificial Intelligence* y a la Generalitat Valenciana, por su apoyo económico para mis estudios del Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital.

Figura 5.1: Emblema de ValgrAI

Figura 5.2: Emblema de la Generalitat Valenciana

Bibliografía

[1] Sheikh, H., Prins, C., Schrijvers, E. (2023). Artificial Intelligence: Definition and Background. In: Mission AI. Research for Policy. Springer, Cham. pp 15-41. https://doi.org/10.1007/978-3-031-21448-6_2

[2] S. Albawi, T. A. Mohammed and S. Al-Zawi, 'Understanding of a convolutional neural network' 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 2017, pp 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[3] Lorente, O., Riera, I., & Rana, A. (2021). Image Classification with Classic and Deep Learning Techniques. *ArXiv*, abs/2105.04895.

[4] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J.D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2023). Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. *ArXiv*, abs/2310.19775.

[5] Xiong, Haoyi, Xuhong Li, Xiaofei Zhang, Jiamin Chen, Xinhao Sun, Yuchen Li, Zeyi Sun, and Mengnan Du. 'Towards Explainable Artificial Intelligence (XAI): A Data Mining Perspective.' arXiv preprint arXiv:2401.04374 (2024).

[6] European Parliament: Artificial Intelligence Act: Deal on comprehensive rules for trustworthy AI.

[7] European Commission: Regulatory framework for AI.
`https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai`

[8] Dong, X., & Cappuccio, M.L. (2023). Applications of Computer Vision in Autonomous Vehicles: Methods, Challenges and Future Directions. *ArXiv*, abs/2311.09093.

[9] Camburu, O. (2020). Explaining Deep Neural Networks. *ArXiv*, abs/2010.01496.

[10] Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. *arXiv preprint*, arXiv:2107.07045.

[11] Samuel, S.Z., Kamakshi, V., Lodhi, N., & Krishnan, N.C. (2021). Evaluation of Saliency-based Explainability Method. *ArXiv*, abs/2106.12773.

[12] Bhat, Ashwin, Adou Sangbone Assoa and Arijit Raychowdhury. "Gradient Backpropagation based Feature Attribution to Enable Explainable-AI on the Edge." 2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC) (2022): 1-6.

[13] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. CoRR, abs/1312.6034.

[14] Zeiler, M.D., Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, vol 8689, pg 818-833. Springer, Cham. `https://doi.org/10.1007/978-3-319-10590-1_53`

[15] Springenberg, J.T., Dosovitskiy, A., Brox, T., & Riedmiller, M.A. (2014). Striving for Simplicity: The All Convolutional Net. CoRR, abs/1412.6806.

[16] Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., \& Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128, 336-359.

[17] Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that? *ArXiv*, abs/1611.07450.

[18] Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, 10, e0130140. `https://doi.org/10.1371/journal.pone.0130140`

[19] Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-Wise Relevance Propagation: An Overview. En *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 193-209). Springer. `https://doi.org/10.1007/978-3-030-28954-6_10`

[20] Böhle M, Eitel F, Weygandt M, Ritter K. Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. Front Aging Neurosci. 2019 Jul 31;11:194. doi: 10.3389/fnagi.2019.00194. PMID: 31417397; PMCID: PMC6685087.

[21] Schlegel, U., & Keim, D. A. (2023). A Deep Dive into Perturbations as Evaluation Technique for Time Series XAI. En *xAI*. Recuperado de `https://api.semanticscholar.org/CorpusID:259766456`

[22] Qiu, Luyu, Yi Yang, Caleb Chen Cao, Jing Liu, Yueyuan Zheng, Hilary Hei Ting Ngai, Janet Hsiao and Lei Chen. "Resisting Out-of-Distribution Data Problem in Perturbation of XAI." ArXiv abs/2107.14000 (2021).

[23] Maksims Ivanovs, Roberts Kadikis, Kaspars Ozols, *Perturbation-based methods for explaining deep neural networks: A survey*, Pattern Recognition Letters, vol. 150, pp. 228-234, 2021. doi: `https://doi.org/10.1016/j.patrec.2021.06.030`,

[24] Benedikt Limmer, *Implementation and Comparison of Occlusion-based Explainable Artificial Intelligence Methods*, Bachelorarbeit, Universität Augsburg, Fakultät für Angewandte Informatik, Lehrstuhl für Multimodale Mensch-Technik Interaktion, Sommersemester 2020, Betreut von Prof. Dr. Elisabeth André.

[25] Blücher, Stefan, Johanna Vielhaben and Nils Strodthoff. "Decoupling Pixel Flipping and Occlusion Strategy for Consistent XAI Benchmarks." ArXiv abs/2401.06654 (2024).

[26] Resta, Michele, Anna Monreale, and Davide Bacciu. 2021. 'Occlusion-Based Explanations in Deep Recurrent Models for Biomedical Signals' Entropy 23, no. 8: 1064. https://doi.org/10.3390/e23081064

[27] InShort, *Occlusion Analysis for Explaining DNNs*, Towards Data Science, Web Article. `https://towardsdatascience.com/inshort-occlusion-analysis-for-explaining-dnns-d0ad3af9aeb6`

[28] Zhang, Quanshi, Yu Yang, Ying Nian Wu and Song-Chun Zhu. "Interpreting CNNs via Decision Trees." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018): 6254-6263.

[29] Frosst, Nicholas and Geoffrey E. Hinton. "Distilling a Neural Network Into a Soft Decision Tree." ArXiv abs/1711.09784 (2017).

[30] Swamy, V., Frej, J., & Käser, T. (2023). The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations. *ArXiv*, abs/2307.00364.

[31] Bhargav Kotipalli, *The Role of Attention Mechanisms in Enhancing Transparency and Interpretability of Neural Network Models in Explainable AI*, Harrisburg University Dissertations and Theses, Harrisburg University of Science and Technology, Spring 4-14-2024.

[32] Khurana, Diksha, Aditya Koli, Kiran Khatter and Sukhdev Singh. "Natural language processing: state of the art, current trends and challenges." Multimedia Tools and Applications 82 (2017): 3713 - 3744.

[33] Lifu Chen, Xingmin Cai, Zhenhong Li, Jin Xing, Jiaqiu Ai, *Where is my attention? An explainable AI exploration in water detection from SAR imagery*, International Journal of Applied Earth Observation and Geoinformation, vol. 130, pp. 103878, 2024. doi: `https://doi.org/10.1016/j.jag.2024.103878`.

[34] N. El Houda Dehimi and Z. Tolba, 'Attention Mechanisms in Deep Learning : Towards Explainable Artificial Intelligence' 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), EL OUED, Algeria, 2024, pp. 1-7, doi: 10.1109/PAIS62114.2024.10541203.

[35] J. -P. Poli, W. Ouerdane and R. Pierrard, "Generation of Textual Explanations in XAI: the Case of Semantic Annotation,"2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg, Luxembourg, 2021, pp. 1-6, doi: 10.1109/FUZZ45933.2021.9494589.

[36] Vilone, Giulia, Longo, Luca. (2021). Classification of Explainable Artificial Intelligence Methods through Their Output Formats. Machine Learning and Knowledge Extraction. 3. 615. 10.3390.

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[38] Rao, S., Mehta, S., Kulkarni, S., Dalvi, H., Katre, N., & Narvekar, M. (2022). A Study of LIME and SHAP Model Explainers for Autonomous Disease Predictions. *2022 IEEE Bombay Sec-*

tion Signature Conference (IBSSC), Mumbai, India, pp. 1-6. doi: 10.1109/IBSSC56953.2022.10037324.

[39] SHAP Documentation. An introduction to explainable AI with Shapley values.
`https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html`

[40] Rozemberczki, B., Watson, L., Bayer, P., Yang, H., Kiss, O., Nilsson, S., & Sarkar, R. (2022). The Shapley Value in Machine Learning. *ArXiv*, abs/2202.05594.

[41] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[42] Santos, Mailson Ribeiro, Affonso Guedes, and Ignacio Sanchez-Gendriz. 2024. "SHapley Additive exPlanations (SHAP) for Efficient Feature Selection in Rolling Bearing Fault Diagnosis"Machine Learning and Knowledge Extraction 6, no. 1: 316-341. https://doi.org/10.3390/make6010016.

[43] Carloni, Gianluca, Andrea Berti and Sara Colantonio. "The role of causality in explainable artificial intelligence." ArXiv abs/2309.09901 (2023).

[44] Kemal Erdem, (Mar 2021). "Measuring XAI methods with Infidelity and Sensitivity". https://erdem.pl/2021/03/measuring-xai-methods-with-infidelity-and-sensitivity

[45] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. 2019. On the (in)fidelity and sensitivity of explanations. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 984, 10967–10978.

[46] Coroama, L., Groza, A. (2022). Evaluation Metrics in Explainable Artificial Intelligence (XAI). In: Guarda, T., Portela, F., Augusto, M.F. (eds) *Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2022. Communications in Computer and Information Science*, vol 1675, pg 401-413. Springer, Cham. `https://doi.org/10.1007/978-3-031-20319-0_30`

[47] Hase, Peter and Mohit Bansal. "Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" Annual Meeting of the Association for Computational Linguistics (2020), 5540–5552.

[48] M. N. Vu, T. D. Nguyen, N. Phan, R. Gera and M. T. Thai, 'c-Eval: A Unified Metric to Evaluate Feature-based Explanations via Perturbation' 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 927-937, doi: 10.1109/BigData52589.2021.9671895.

[49] Baniecki, Hubert and P. Biecek. "Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey." ArXiv abs/2306.06123 (2023).

[50] Stanford University, *Stanford Dogs Dataset*, Accedido el 10 de julio de 2024. `http://vision.stanford.edu/aditya86/ImageNetDogs/`

[51] Piosenka, G. BIRDS 525 SPECIES - IMAGE CLASSIFICATION. Kaggle. https://www.kaggle.com/datasets/gpiosenka/100-bird-species. 525 species, images 224x224x3 JPEG.

[52] Marine Animal Images Dataset. Kaggle. https://www.kaggle.com/datasets/mikoajfish99/marine-animal-images.

[53] Keras. *Keras Documentation: Models*. Accedido el 14 de julio de 2024. Recuperado de `https://keras.io/api/applications/`.

[54] Keras. *Keras Documentation: EfficientNet*. Accedido el 14 de julio de 2024. Recuperado de `https://keras.io/api/applications/efficientnet/`.

[55] Tan, M., & Le, Q.V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. ArXiv, abs/1905.11946.

[56] Montalbo, F. J., & Alon, S. A. D. (2021). Empirical Analysis of a Fine-Tuned Deep Convolutional Model in Classifying and Detecting Malaria Parasites from Blood Smears. *KSII Transactions on Internet and Information Systems*, 15(1), 147-165. `https://doi.org/10.3837/tiis.2021.01.009`.

[57] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805. ISSN 1566-2535.

Anexo ODS

Objetivos de Desarrollo Sostenible	Alto	Medio	Bajo	No procede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**	X			
ODS 5. **Igualdad de género.**		X		
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**		X		
ODS 9. **Industria, innovación e infraestructuras.**	X			
ODS 10. **Reducción de las desigualdades.**	X			
ODS 11. **Ciudades y comunidades sostenibles.**		X		
ODS 12. **Producción y consumo responsables.**		X		
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**	X			
ODS 17. **Alianzas para lograr objetivos.**		X		

Este trabajo, al estar enfocado en la explicabilidad de redes neuronales aplicadas al procesamiento de imágenes, presenta una serie de contribuciones significativas a varios de los Objetivos de Desarrollo Sostenible (ODS). En primer lugar, se observa una clara relación con el *ODS 3: Salud y bienestar*, ya que las técnicas de inteligencia artificial, y en particular las redes neuronales explicables, pueden tener un impacto directo en la mejora de sistemas de diagnóstico médico y otras aplicaciones en el ámbito sanitario. La capacidad de comprender cómo y por qué las redes neuronales toman decisiones es crucial para garantizar diagnósticos precisos y confiables, lo que no solo contribuye al bienestar general, sino que también fomenta un mayor nivel de confianza en la adopción de estas tecnologías en el sector salud.

Por otro lado, este trabajo tiene una implicación relevante en el *ODS 4: Educación de calidad*. La IA se ha convertido en un componente esencial en la educación moderna, tanto en términos de contenido formativo como de herramientas tecnológicas. Al estudiar y mejorar la explicabilidad de los modelos de IA, se promueve una mejor comprensión de estos sistemas, lo cual es vital para educar a futuras generaciones de profesionales que estarán a

cargo de su desarrollo y regulación. Además, la transparencia en los modelos de IA contribuye a crear herramientas educativas más justas e inclusivas, reforzando la equidad en los procesos de enseñanza y aprendizaje.

Asimismo, la investigación aborda temas vinculados al *ODS 5: Igualdad de género*, ya que uno de los objetivos clave de la explicabilidad en IA es identificar y mitigar los sesgos inherentes en los sistemas automáticos. Los modelos de redes neuronales pueden perpetuar, sin intención, sesgos relacionados con el género u otros factores sociales. Al hacer visibles estos sesgos y proponer formas de corregirlos, se puede avanzar hacia una IA más equitativa y justa, que no reproduzca desigualdades preexistentes.

En términos de desarrollo económico, este trabajo también guarda relación con el *ODS 8: Trabajo decente y crecimiento económico*. El uso de modelos de IA explicables en entornos laborales permite asegurar que las decisiones automáticas, como las relacionadas con la contratación o el desempeño, se basen en criterios transparentes y no perpetúen discriminaciones inadvertidas. La implementación de IA más interpretables puede mejorar tanto la equidad en el lugar de trabajo como el crecimiento económico a través de la innovación responsable.

El *ODS 9: Industria, innovación e infraestructuras* también se ve impactado positivamente por este estudio, dado que la IA explicable es esencial para impulsar la innovación tecnológica de manera sostenible. Las redes neuronales son componentes clave de muchas infraestructuras tecnológicas, y garantizar su transparencia y fiabilidad contribuye a desarrollar sistemas más robustos, lo cual es fundamental para mantener la confianza en el uso de la IA en la industria.

Otra dimensión importante de este trabajo se refleja en su alineación con el *ODS 10: Reducción de las desigualdades*. Las técnicas de explicabilidad permiten detectar desigualdades en los sistemas de IA, lo que facilita la creación de soluciones tecnológicas más inclusivas y justas. En este sentido, se busca evitar que los modelos automáticos exacerben desigualdades existentes en áreas como el acceso a servicios públicos o el mercado laboral.

De igual forma, este trabajo tiene implicaciones en el *ODS 11: Ciudades y comunidades sostenibles*, ya que la IA explicable puede ser aplicada en la gestión de ciudades inteligentes. Al hacer más transparentes los algoritmos que toman decisiones en aspectos como la movilidad urbana o la distribución de recursos, se asegura una mayor confianza de los ciudadanos en los sistemas automatizados que impactan directamente en sus vidas cotidianas.

El *ODS 12: Producción y consumo responsables* también es relevante, ya que las técnicas de explicabilidad en IA pueden ser fundamentales para hacer más eficientes y transparentes los procesos industriales y de consumo, lo que lleva a un uso más responsable de los recursos y una mayor sostenibilidad.

Finalmente, se puede afirmar que este trabajo tiene un fuerte vínculo con el *ODS 16: Paz, justicia e instituciones sólidas*, dado que la transparencia en los sistemas de IA es crucial para garantizar la justicia en la toma de decisiones automatizadas. La explicabilidad de estos modelos permite a los reguladores y legisladores evaluar mejor su equidad y confiabilidad, asegurando que las instituciones que emplean IA lo hagan de manera ética y responsable.

En resumen, este trabajo contribuye de manera significativa a varios Objetivos de Desarrollo Sostenible, al promover el uso de técnicas de inteligencia artificial explicable que mejoran tanto la transparencia como la equidad en diversas áreas, desde la salud y la educación hasta la industria y la reducción de desigualdades. Si bien no todos los ODS están directamente relacionados con el tema tratado, los que sí lo están reflejan el potencial de continuar investigando en esta área para influir positivamente en sectores clave de la sociedad.