Integración de diferentes técnicas para visualizar la influencia de regiones de una imagen en su clasificación por una red neuronal

Andrés Gardella Ruiz[1], Gabriela Pérez[1,2], Claudia Pons[1,3,4]

[1] LIFIA, Facultad de informática - Universidad Nacional de La Plata

[2]UNAJ - Universidad Nacional Arturo Jauretche, Florencio Varela, Bs As, Argentina

[3] CIC, Comisión de Investigaciones Científicas, Bs As, Argentina

[4] UAI. Universidad Abierta Interamericana, Ciudad de Buenos Aires, Argentina

andresmgr@gmail.com, gabriela.perez@gmail.com, cpons@lifia.info.unlp.edu.ar

Resumen. Hoy en día es común la utilización, en múltiples ámbitos, de redes neuronales que permiten realizar actividades complejas como clasificación de imágenes. Si bien es una tecnología muy útil debido a la información que provee, en contextos sensibles como el de la salud pública, es necesario poder entender y confiar en dicha información, ya que la falta de precisión puede acarrear consecuencias negativas significativas. Esta necesidad de comprender el funcionamiento y la toma de decisiones de las redes neuronales, ha dado lugar al surgimiento de métodos y técnicas de visualización, que permiten comprender mejor las decisiones tomadas por éstas en base a la información ingresada.

Este trabajo tiene como propósito analizar algunos de estos métodos de visualización para luego desarrollar una herramienta que simplifique su uso y la visualización de las explicaciones. La herramienta permitirá comparar los resultados y facilitará la interpretación de las decisiones de la red, haciendo que estos métodos sean más accesibles.

Palabras clave: Red Neuronal Convolucional, Análisis de Imágenes, Visualización.

1 Introducción

La inteligencia artificial y el aprendizaje profundo han experimentado avances notables en los últimos años, dando lugar a la creación de modelos capaces de realizar con éxito tareas sumamente complejas. Entre los modelos más destacados se encuentran las redes neuronales convolucionales (CNNs), que han demostrado un gran éxito en el campo de la visión artificial, especialmente en tareas como la clasificación de imágenes y la detección de objetos [1][2][3]. Estas redes suelen utilizarse para una variedad de aplicaciones, que van desde la inspección de componentes en la industria, el control de calidad de alimentos en el sector alimenticio hasta el apoyo en diagnósticos médicos, como la identificación de tumores o detección de enfermedades a partir de imágenes. Sin embargo, carecen de interpretabilidad debido a su complejidad.

Dado que algunos de estos usos son sensibles y críticos, es fundamental tener la capacidad de comprender y explicar las decisiones tomadas por estos sistemas. La transparencia y la explicabilidad en el funcionamiento de estos modelos son esenciales para garantizar la confianza (o desconfianza) en su uso, así como para abordar preocupaciones éticas y legales asociadas con su implementación.

Afortunadamente, en la actualidad existen diversas técnicas y métodos que proporcionan cierto grado de interpretabilidad a estos modelos, permitiendo entender qué partes específicas de la imagen procesada influyen en su decisión.

En este estudio, se realizará un análisis de los métodos de visualización de decisiones tomadas por una CNN. Posteriormente, se seleccionarán algunos de estos métodos en función de criterios como la facilidad de uso, el tiempo de procesamiento y su adaptabilidad a diversas arquitecturas de CNN. Estos métodos se integrarán en una herramienta con el objetivo de facilitar su acceso y simplificar su utilización por parte de los usuarios. El propósito es hacer que los resultados individuales de una red neuronal sean comprensibles y accesibles, permitiendo una interpretación informada sin que se perciban como respuestas inexplicables de una caja negra.

2 Background

Las redes convolucionales (CNN) son un tipo de red neuronal profunda que se utilizan principalmente en tareas de reconocimiento de imágenes debido a su capacidad para capturar y aprender patrones espaciales.

Existen diversas propuestas para interpretar y explicar las clasificaciones realizadas por estas redes. En este trabajo se analizaron cuatro de las técnicas más destacadas como CAM[4], Grad-CAM[5], LIME[6] y SHAP[7].

CAM y Grad-CAM generan mapas de calor que resaltan las regiones de imágenes que más influyen en la clasificación realizada, combinando linealmente las activaciones ponderadas de la última capa convolucional.

LIME propone explicar las predicciones de cualquier clasificado mediante aproximaciones locales con un modelo interpretable. Divide una imagen en subregiones (superpixeles) y se resaltan aquellos que más influyen en la decisión tomada por el clasificador.

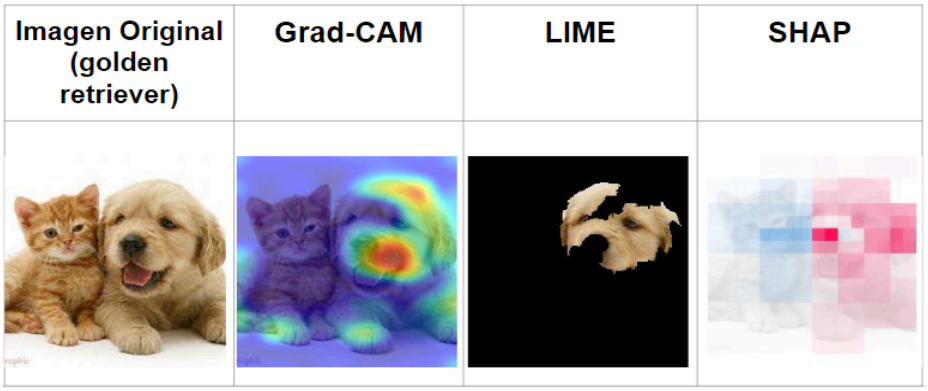

Fig. 1. Resultados de aplicar los métodos de visualización Grad-CAM, LIME y SHAP sobre la clasificación de una imagen mediante una red neuronal convolucional

SHAP es un enfoque de teoría de juegos para explicar la salida de cualquier modelo de aprendizaje automático, proveyendo una medida de la contribución de cada participante (pixel) respecto a la ganancia obtenida (la clasificación).

En la Figura 1 se pueden observar los resultados de los métodos aplicados a una red cuando clasifica una imagen.

3 Propuesta

Grad-CAM, LIME y SHAP son técnicas modelo-agnósticas que pueden aplicarse a cualquier arquitectura de CNN, ya que no dependen de la estructura interna del modelo. Aunque cada técnica puede aplicarse individualmente, integrar su visualización simultánea permitiría una comparación de los resultados, lo que podría facilitar una mejor evaluación de la explicación proporcionada por cada método.

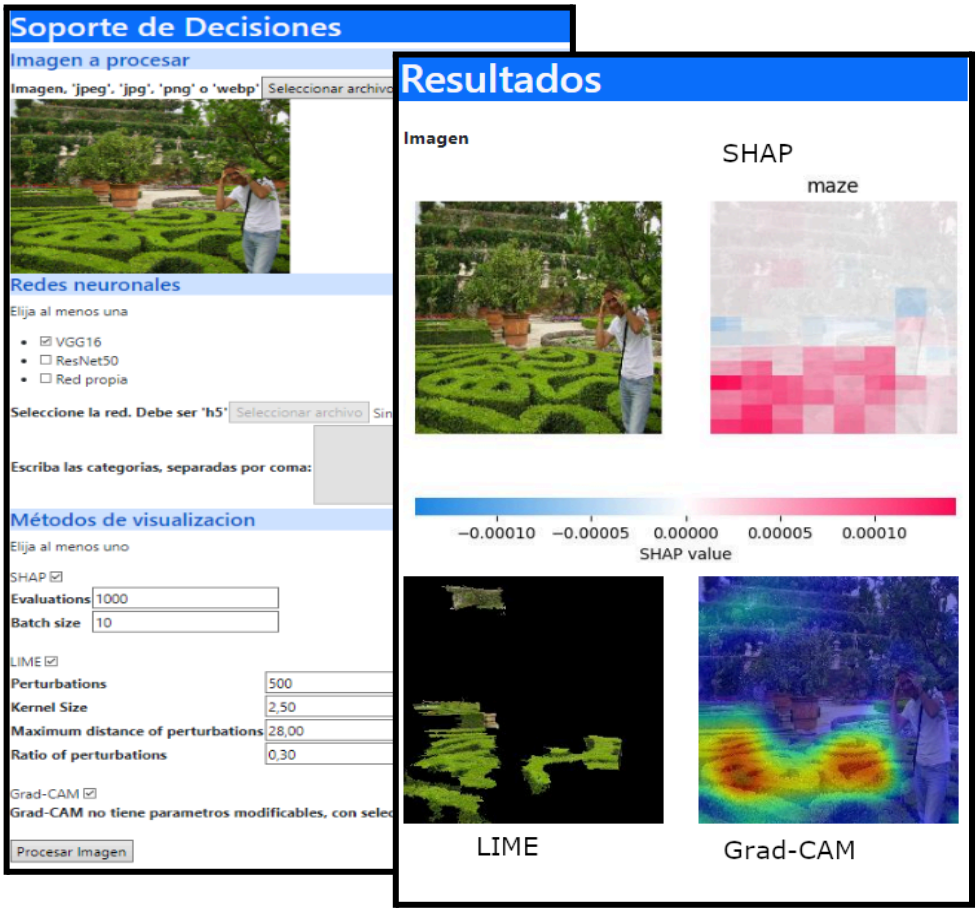

Fig. 2. A la izquierda puede verse el formulario de la aplicación desarrollada, y a la derecha se presentan los resultados de los métodos utilizados para la red seleccionada.

Por esta razón nos propusimos integrarlas en una herramienta que facilita la aplicación de estas técnicas y permite la visualización conjunta de las explicaciones generadas por cada una de ellas.

3.1 Herramienta desarrollada

Como resultado de este trabajo se implementó una herramienta que integra los métodos de visualización mencionados y pueden aplicarse sobre varias arquitecturas de CNN, como VGG16, Resnet50 o una red CNN personalizada. En la Figura 2 pueden observarse dos capturas de pantalla de la herramienta. En la parte izquierda se muestra la sección de configuración para cada método, mientras que en la parte derecha se presentan los resultados.

4 Conclusiones y trabajos futuros

La herramienta facilita el uso y el acceso a los métodos de visualización, permitiendo realizar pruebas y comparaciones de manera eficiente. Esto se traduce en una mejora significativa en la comprensión de las decisiones tomadas por redes neuronales en la clasificación de imágenes.

Para trabajos futuros, se prevé la incorporación de nuevas arquitecturas de CNN, incluyendo redes personalizadas. También se plantea incluir más métodos de visualización.

Referencias

1. Russell, S. J., Norvig, P., Davis, E. ; Genesereth, M. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
2. Simonyan, Karen and Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.
3. Krizhevsky, A., Sutskever, I, Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386. (2012)
4. Zhou, Bolei & Khosla, Aditya & Lapedriza, Àgata & Oliva, Aude & Torralba, Antonio. (2016). Learning Deep Features for Discriminative Localization. 10.1109/CVPR.2016.319.
5. Rs, Ramprasaath & Cogswell, Michael & Das, Abhishek & Vedantam, Ramakrishna & Parikh, Devi & Batra, Dhruv. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 618-626. 10.1109/ICCV.2017.74.
6. Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
7. Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.