

CNN explanation methods for ordinal regression tasks [2]

Alberto Fernández Merchán

Universidad de Córdoba z42ferma@uco.es

1 Introducción y Contexto

En el artículo propuesto [2] nos introducen el problema de la utilización de métodos de explicabilidad para redes neuronales convolucionales (CNN) en el ámbito de la regresión ordinal. En él nos muestran diferentes modelos de explicabilidad que se utilizan para comprender los resultados de las redes neuronales convolucionales en cuatro conjuntos de datos diferentes.

En este trabajo haremos una revisión de este artículo comenzando por explicar qué es la regresión ordinal y por qué es importante poder interpretar los resultados de las CNN en este contexto. Además, revisaremos algunos de los métodos de interpretación empleados y las aplicaciones sobre las que se utilizan.

También analizaremos las métricas que se han utilizado en el estudio para comparar los diferentes modelos y se debatirán los resultados obtenidos.

Finalmente, realizaremos un análisis sobre los algoritmos utilizados y las limitaciones identificadas en el estudio.

1.1 ¿Qué es la regresión ordinal?

La regresión ordinal es un tipo de problema de aprendizaje supervisado en el que las etiquetas de salida siguen un orden natural, pero las distancias entre estas etiquetas no son necesariamente uniformes ni cuantificables de manera precisa. A diferencia de la clasificación convencional, donde las clases son categóricas y no ordenadas, o de la regresión continua, donde las salidas son valores numéricos en un rango continuo, la regresión ordinal se sitúa en un punto intermedio. En este enfoque, se trabaja con un conjunto discreto y ordenado de categorías que representan diferentes niveles de una variable de interés.

Un ejemplo clásico de regresión ordinal es la puntuación de satisfacción del cliente en una escala de 1 a 5. Aunque las puntuaciones están claramente ordenadas (1 indica menor satisfacción y 5 mayor), la diferencia entre las puntuaciones no es necesariamente equidistante; la satisfacción percibida entre 2 y 3 podría no ser la misma que entre 4 y 5.

1.2 Diferencias con otros enfoques

La regresión ordinal comparte ciertas similitudes con otros enfoques de modelado predictivo, pero presenta características diferenciadoras que la hacen única:

- **Clasificación convencional:** En la clasificación tradicional, las clases no tienen un orden inherente; por ejemplo, clasificar imágenes de animales (gato o perro) no implica un orden natural. En cambio, la regresión ordinal trabaja con clases que sí están ordenadas, como niveles de gravedad de una enfermedad (leve, moderado, grave).
- **Regresión continua:** A diferencia de la regresión continua, donde la variable objetivo puede tomar cualquier valor dentro de un rango, la regresión ordinal se limita a un conjunto discreto de niveles ordenados, lo que permite manejar mejor ciertos tipos de problemas donde la continuidad completa no es apropiada.
- **Regresión ordinal:** Este enfoque combina lo mejor de la clasificación y la regresión, aprovechando la estructura ordenada de las clases mientras mantiene la discreción de las categorías. Además, permite explotar relaciones entre niveles cercanos para mejorar la precisión del modelo.

Entre las ventajas de la regresión ordinal se destacan:

- Facilita la identificación de patrones que distinguen entre clases ordinales, mejorando la precisión del modelo.
- Aumenta la interpretabilidad en contextos sensibles como la medicina o la evaluación de calidad, donde comprender las diferencias entre niveles es tan importante como la predicción final.
- Contribuye a la detección de errores sistemáticos o sesgos en el modelo, al aprovechar la información del orden inherente de las clases.

1.3 Importancia de la explicabilidad de las CNN

Las CNN han demostrado un gran rendimiento en tareas de visión por computador, incluidas aquellas que requieren clasificación y regresión ordinal de imágenes. Sin embargo, uno de los principales problemas asociados a estos modelos es su naturaleza de «*caja negra*», que dificulta comprender cómo se toman las decisiones dentro del modelo.

Esta falta de transparencia puede ser especialmente problemática en aplicaciones sensibles como el diagnóstico médico o la evaluación de calidad de productos, donde es de vital importancia entender no solo la predicción final, sino

también los factores que la motivan. Una mayor explicabilidad ayuda a generar confianza en el modelo, identificar posibles sesgos y validar que las decisiones se basan en características relevantes de los datos.

En el contexto específico de la regresión ordinal, la explicabilidad permite profundizar en cómo el modelo distingue entre niveles adyacentes, y qué patrones visuales influyen en la asignación de un nivel determinado. Esto resulta esencial para garantizar que las predicciones no solo sean precisas, sino también comprensibles y justificables.

En este trabajo, se analizan diferentes técnicas de explicabilidad aplicadas a CNN en tareas de regresión ordinal, con el objetivo de evaluar su capacidad para proporcionar interpretaciones útiles y transparentes en distintos dominios de aplicación.

2 Métodos Explicativos para las CNN

Uno de los principales problemas de los modelos CNN, y de las redes neuronales en general, es que son difíciles de interpretar. Una red neuronal es una caja negra que recibe unos datos de entrada y devuelve una predicción. Por el camino, realiza miles de operaciones con pesos y sesgos que no son sencillos de analizar ni de justificar. Esta falta de interpretabilidad dificulta la validación de las predicciones que se obtienen del modelo y puede hacer que no se detecten posibles sesgos o errores.

En el caso de una regresión ordinal, como la que se está trabajando en este estudio, la falta de interpretabilidad tiene aún más impacto. Al tratarse de una salida continua u ordinal, no solo interesa conocer la predicción final, sino también entender cómo las características de la entrada han influido en esa predicción y en qué medida.

Además, la alta dimensionalidad de las representaciones internas dificulta entender cómo se combinan los distintos patrones para llegar a una predicción final. Aunque algunas características iniciales, como bordes o texturas, sean interpretables, conforme se avanza en la red, estas se transforman en representaciones abstractas que no tienen una correspondencia directa con la imagen original.

2.1 ¿Por qué es difícil explicar las CNN?

La dificultad de explicar las CNN reside principalmente en su arquitectura. Estas redes están formadas por múltiples capas convolucionales, de activación y de agrupamiento que procesan la información de manera jerárquica y no lineal.

Cada capa extrae características de mayor complejidad que la anterior, desde patrones simples hasta representaciones más abstractas de la imagen. Sin embargo, esta profundidad y complejidad hacen que resulte complicado rastrear cómo una característica específica de la imagen afecta a la salida final del modelo.

En otras palabras, aunque sepamos qué filtros se aplican y qué activaciones se generan, es difícil interpretar qué significado tienen esas activaciones en relación con las características visuales de la imagen.

2.2 Métodos de explicación utilizados

Para abordar este problema de interpretabilidad, se utilizan técnicas de interpretación visual que ayudan a entender qué partes de una imagen influyen más en la predicción del modelo. Estas técnicas generan mapas de calor o mapas de relevancia que destacan las regiones importantes de la imagen para la red neuronal.

Entre las técnicas más utilizadas se encuentran Grad-CAM, que genera mapas visuales destacando las zonas de interés del modelo, y LIME, que estudia cómo pequeñas perturbaciones en la imagen afectan a la predicción. Estas herramientas permiten obtener explicaciones locales que facilitan entender decisiones concretas del modelo, en lugar de ofrecer una visión global.

Estos mapas permiten visualizar de forma clara qué zonas de la imagen han tenido mayor peso en la toma de decisiones del modelo, facilitando la interpretación de las predicciones y ayudando a identificar posibles errores o sesgos.

2.3 Matriz de explicabilidad y *Degradation Score*

En el ámbito de la regresión ordinal, una herramienta destacada es la **matriz de explicabilidad**, que representa la relación entre la salida del modelo y las características más influyentes de la entrada. Esta matriz permite analizar cómo varían las respuestas del modelo cuando se alteran los datos, facilitando la detección de patrones, sesgos y comportamientos inesperados en la red.

En problemas ordinales, la matriz proporciona detalles sobre cómo se distribuye la importancia de las características a través de las diferentes categorías, lo que facilita una mejor interpretación del modelo. En lugar de ofrecer una explicación global única, descompone la interpretación en múltiples mapas de importancia. Cada mapa corresponde a una clase ordinal específica, destacando las áreas de la imagen que contribuyen más a la predicción de cada clase.

Esto no solo ayuda a identificar por qué se seleccionó una clase determinada, sino también a comprender qué diferencia a una clase de otra, proporcionando una visión detallada del proceso de toma de decisiones. Matemáticamente, esta matriz E se define como:

$$E \in \mathbb{R}^{H \times W} \quad (1)$$

donde H y W son la altura y el ancho de la imagen, respectivamente. Los valores de E varían entre 0 y 1, indicando la relevancia de cada píxel. Esto permite visualizar de forma precisa cómo contribuye cada píxel a la predicción final.

En el estudio, se genera una matriz de explicabilidad que asigna una puntuación de relevancia a cada píxel de la imagen, indicando cuáles son los más determinantes para la predicción del modelo.

Para evaluar la fiabilidad de esta matriz de explicabilidad, se recurre a un **análisis de perturbación** [3] (o *perturbation analysis*), utilizando el *degradation score*, representado en la Figura 1. Este método cuantifica la coherencia entre las regiones destacadas por la matriz y su impacto real en la predicción ordinal del modelo. El protocolo experimental sigue tres etapas clave:

1. Preprocesamiento espacial:

- La imagen se divide en regiones cuadradas no superpuestas de 8×8 píxeles (T regiones en total), según lo descrito en el artículo [4].
- Cada región R_t se clasifica según su relevancia agregada:

$$\text{Score}(R_t) = \sum_{(i,j) \in R_t} E(i,j)$$

2. Protocolo de oclusión [1]:

- **MoRF (Most Relevant First):** Se reemplazan secuencialmente las regiones R_t con valores neutros (por ejemplo, gris medio), comenzando por las regiones con mayor valor de $\text{Score}(R_t)$.
- **LeRF (Least Relevant First):** El proceso inverso, comenzando por las regiones menos relevantes.

3. Medición de degradación:

- En cada paso k (con $1 \leq k \leq T$), se calcula la métrica ordinal (como el MAE o κ) sobre la imagen parcialmente oculta.
- Se registra la curva de degradación como $\text{Métrica}(k) = \text{Valor}(x_{\text{oculta}})$.

Interpretación del Degradation Score El artículo sugiere que el área bajo ambas curvas puede ser utilizada como una medida para evaluar la calidad de las explicaciones generadas por la matriz de explicabilidad:

$$\Delta_{\text{AUC}} = \int_0^T (\text{MoRF}(k) - \text{LeRF}(k)) dk \quad (2)$$

Esta medida cuantifica la diferencia entre la pérdida de precisión al eliminar las regiones más relevantes (*MoRF*) y las menos relevantes (*LeRF*), proporcionando una evaluación global de la efectividad de las explicaciones.

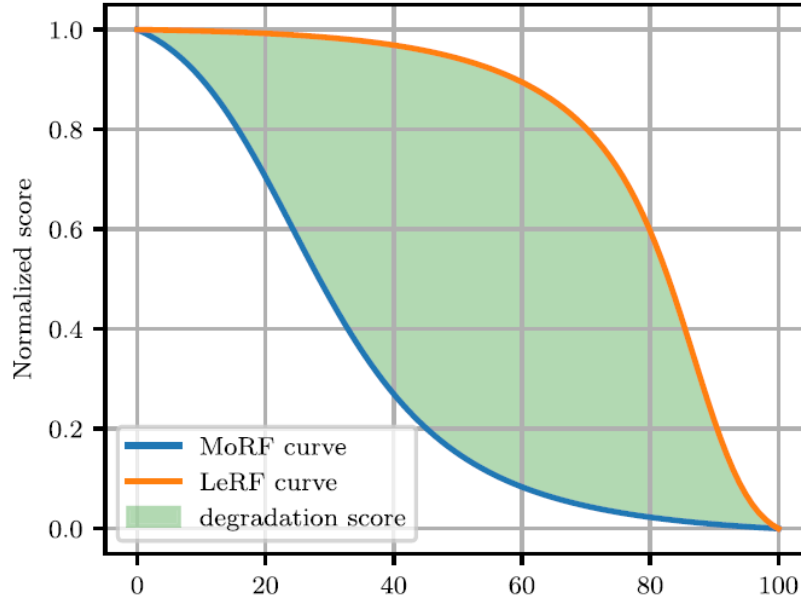


Fig. 1. Análisis de degradación (tomado del artículo)

El área entre las curvas *MoRF* y *LeRF* define el *degradation score*. Cuanto mayor sea esta área, mejor será la calidad de la explicación, ya que indica que la asignación de importancia a las características es coherente: al eliminar primero las regiones más relevantes, la capacidad predictiva del modelo disminuye drásticamente, mientras que al eliminar las regiones menos relevantes, el impacto en la predicción es mínimo al principio.

En resumen, el *degradation score* actúa como una métrica tanto gráfica como cuantitativa para validar la efectividad de la matriz de explicabilidad. Este score confirma que la matriz realmente identifica las regiones de la imagen que tienen mayor influencia en la predicción del modelo.

2.4 Modelos utilizados

En esta sección se presentan los modelos y métodos utilizados para abordar la tarea de clasificación ordinal y la explicación de las predicciones mediante mapas de activación. Los modelos se dividen en dos grupos principales: los modelos para la clasificación ordinal y los métodos para la explicación de las decisiones de los modelos.

2.5 Ordinal Binary Decomposition

La técnica *Ordinal Binary Decomposition (OBD)* propone resolver los problemas de clasificación ordinal descomponiendo la tarea original con Q clases en $Q - 1$

problemas binarios independientes. En cada uno de estos problemas, se estima la probabilidad de que la clase real de una muestra \mathbf{x} supere una clase de referencia \mathcal{C}_q , mediante una salida de la forma:

$$o_q = P(y \succ \mathcal{C}_q \mid \mathbf{x})$$

Nota: El símbolo \succ denota la relación de orden entre clases, indicando que la clase y es superior a la clase de referencia \mathcal{C}_q .

Sin embargo, esta aproximación puede conducir a combinaciones inconsistentes en las salidas. Para mantener la coherencia ordinal, se recurre al marco de *Error-Correcting Output Codes (ECOC)*, donde cada clase se representa mediante un vector ideal $\mathbf{v}(\mathcal{C}_q)$ definido como:

$$v_c^q = \mathbb{1}\{c < q\}$$

Nota: $\mathbb{1}\{\cdot\}$ es la función indicadora, que toma el valor 1 si la condición se cumple y 0 en otro caso.

La predicción final se realiza asignando la clase cuyo vector ideal minimice la distancia euclídea respecto al vector de salida del modelo:

$$\hat{y} = \arg \min_{\mathcal{C}_q} \|\mathbf{o}(\mathbf{x}) - \mathbf{v}(\mathcal{C}_q)\|_2$$

Para alinear la fase de entrenamiento con este criterio de decisión basado en distancias, se sustituye la función de pérdida de entropía cruzada por el error cuadrático medio:

$$\mathcal{L}_{SE}(\mathbf{x}_i) = \sum_{q=1}^{Q-1} (\mathbb{1}\{y_i \succ \mathcal{C}_q\} - P(y_i \succ \mathcal{C}_q \mid \mathbf{x}_i))^2$$

Este enfoque permite aprovechar la estructura ordinal de las clases, mejorando la coherencia y el rendimiento del modelo en tareas de clasificación ordinal.

Grad-CAM (Gradient-weighted Class Activation Mapping) es un método de visualización que genera mapas de activación para modelos de CNN. Grad-CAM utiliza los gradientes de la capa final del modelo para identificar las regiones de la imagen que más contribuyen a la decisión del modelo. Este método se utiliza para mejorar la interpretabilidad de los modelos de clasificación, permitiendo visualizar las áreas de una imagen que influyen más en la predicción.

Sea A_k una activación de la característica k en la capa intermedia seleccionada, y A una representación de la capa con dimensiones de altura H_A y ancho W_A . El puntaje de relevancia para cada característica k , denotado como α_k^q , se calcula con la siguiente fórmula:

$$\alpha_k^q = \frac{1}{H_A \times W_A} \sum_{i=1}^{H_A} \sum_{j=1}^{W_A} \frac{\partial f_q}{\partial A_{kij}}$$

donde f_q es la puntuación de salida para la clase de destino q .

Posteriormente, el mapa de explicación se genera como una combinación lineal de los mapas de activación A_k , ponderados por su importancia α_k^q . A continuación, se aplica una función de activación Rectified Linear Unit (ReLU) para eliminar interacciones negativas, resultando en:

$$E = \text{ReLU} \left(\sum_k \alpha_k^q A_k \right)$$

Este mapa de activación puede luego ser reescalado al tamaño de la imagen de entrada y superpuesto sobre la imagen original para facilitar la visualización de las regiones relevantes que contribuyen a la predicción.

Grad-CAM++ es una extensión de Grad-CAM que mejora la generación de mapas de activación al abordar algunas de las limitaciones de Grad-CAM, como la suavidad y la precisión de las activaciones. Mientras que Grad-CAM realiza un promedio ponderado a través de los mapas de características y un promedio no ponderado a través de las ubicaciones dentro de cada mapa de características, los autores de Grad-CAM++ argumentan que este enfoque puede ser inadecuado para localizar correctamente los objetos en imágenes que contienen múltiples ocurrencias de objetos de la misma clase. En estos casos, un mapa de características puede encontrar un objeto en una región, mientras que otro mapa de características puede encontrarlo en otra región, lo que podría llevar a una localización incorrecta.

Para solucionar este problema, Grad-CAM++ redefine el cálculo de la importancia α_k^q de cada mapa de características k introduciendo un peso adicional γ_{kij}^q para cada píxel de la activación, que se combina con la derivada del puntaje de salida f_q respecto a la activación. La fórmula de α_k^q en Grad-CAM++ se expresa como:

$$\alpha_k^q = \sum_{i=1}^{H_A} \sum_{j=1}^{W_A} \gamma_{kij}^q \text{ReLU} \left(\frac{\partial f_q}{\partial A_{kij}} \right)$$

Este ajuste mejora la localización y resolución del mapa de activación, proporcionando una visualización más precisa de las regiones relevantes que contribuyen a la predicción del modelo.

Score-CAM es un método alternativo de visualización que calcula los mapas de activación sin necesidad de utilizar gradientes, lo que permite evitar algunos

problemas comunes relacionados con la retropropagación, como la saturación o desaparición de gradientes debido a las no linealidades de las funciones sigmoide y ReLU. A diferencia de Grad-CAM, que utiliza un promedio ponderado de los gradientes, Score-CAM se basa en la importancia de cada mapa de características k , la cual se calcula mediante la siguiente fórmula:

$$\alpha_k^q = f_q(\mathbf{x} \circ B_k) - f_q(\mathbf{x}_b)$$

donde B_k es la versión normalizada y redimensionada del mapa de características A_k con dimensiones $H \times W$, cuyos valores están en el intervalo $[0, 1]$, y \mathbf{x}_b es una imagen base (como una imagen negra). La operación \circ denota el producto elemento a elemento entre las activaciones de la capa final y el mapa de características.

La intuición detrás de Score-CAM es que la importancia de cada mapa de características debe ser evaluada en función de su impacto directo en la predicción del modelo, lo que mejora la interpretación visual al eliminar la dependencia de los gradientes y la retropropagación.

IBA El *Importance-based Attribution* (IBA) es un método de atribución basado en perturbaciones que inserta un cuello de botella en una capa intermedia de la red para medir la relevancia de las regiones de la entrada. Introduce una variable aleatoria Z que combina la activación original A y ruido ε , regulado por un parámetro $\lambda(x)$:

$$Z = \lambda(x)A + (1 - \lambda(x))\varepsilon$$

El objetivo es maximizar la información compartida entre Z y la predicción del modelo, mientras se minimiza la dependencia de la entrada, mediante:

$$\max I[s_q(x); Z] - \beta I[x; Z]$$

La función de pérdida que se optimiza combina la entropía cruzada con una medida de la información retenida:

$$\mathcal{L}_\lambda = \mathcal{L}_{CE} + \beta \mathcal{L}_I$$

Así, $\lambda(x)$, ajustado mediante descenso de gradiente, actúa como mapa de importancia, donde valores cercanos a 1 indican regiones relevantes de la entrada.

Tanto Grad-CAM como IBA han demostrado su validez y rendimiento en tareas de clasificación nominal tradicional. Sin embargo, es importante destacar que ninguno de estos métodos considera la relación de orden inherente a las etiquetas de clase en tareas de regresión ordinal.

En particular, mientras Grad-CAM y IBA ofrecen interpretaciones útiles de las decisiones del modelo, su diseño no incorpora la estructura ordinal de las clases, lo que limita su capacidad para capturar las relaciones entre categorías ordenadas.

GradOBD-CAM es una extensión de Grad-CAM específicamente diseñada para tareas de clasificación ordinal con el modelo OBD. En lugar de centrarse únicamente en la activación de una neurona de salida, utiliza las derivadas de todas las neuronas de salida del modelo OBD, donde cada una representa la probabilidad de superar un determinado umbral de clase $P(y \succ \mathcal{C}_p \mid \mathbf{x})$.

Para calcular la importancia de cada mapa de características k , GradOBD-CAM introduce un nuevo coeficiente α_k^q , definido como:

$$\alpha_k^q = \sum_{c=1}^{Q-1} \delta_c^q \frac{1}{W_A \times H_A} \sum_i^{H_A} \sum_j^{W_A} \frac{\partial o_c}{\partial A_{ij}^k} \quad (3)$$

Donde el parámetro δ_c^q considera la relación ordinal entre la clase c y la clase objetivo q , tal que:

$$\delta_c^q = \begin{cases} +1 & \text{si } c < q \\ -1 & \text{si } c \geq q \end{cases} \quad (4)$$

Esta formulación permite que GradOBD-CAM asigne mayor importancia a las características que contribuyen positivamente a las probabilidades de salida para clases inferiores a la actual y penalice las que contribuyen a clases superiores, respetando la naturaleza ordinal de las etiquetas.

OIBA OIBA extiende el método IBA incorporando la pérdida ordinal propia del modelo OBD en el proceso de optimización de la máscara de perturbación $\mathbb{E} = \lambda(\mathbf{x})$.

En lugar de utilizar la pérdida de entropía cruzada, que no es la más adecuada para tareas ordinales, OIBA sustituye esta por la pérdida de error cuadrático medio:

$$\mathcal{L}_\lambda = \mathcal{L}_{SE} + \beta \mathcal{L}_I \quad (5)$$

Esta modificación (Ec. 5) permite integrar la estructura ordinal directamente en la construcción del mapa explicativo, aprovechando mejor la información de todas las salidas del modelo en su representación nativa.

3 Experimentos y Resultados

3.1 Datasets utilizados



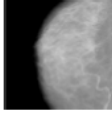
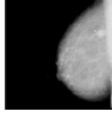
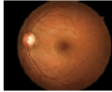
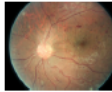
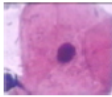
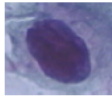
En los experimentos se utilizaron cuatro conjuntos de datos de imágenes (ver Figura 2), cada uno asociado con tareas de clasificación ordinal. Estos conjuntos fueron seleccionados por su diversidad en tipos de imágenes y la presencia de etiquetas ordinales de predicción. Los datasets utilizados son los siguientes:

- **Adience**: Conjunto de 17,702 fotos de personas extraídas de la web, con rostros prealineados. Está dividido en 8 grupos de edad, que van desde 0-2 años hasta 60 años o más.

- **CBIS-DDSM**: Base de datos que contiene 2,620 estudios de mamografías escaneadas, cada uno evaluado con el sistema BI-RADS por un radiólogo entrenado. Las imágenes fueron recortadas y redimensionadas a un tamaño de 224×224 píxeles.
- **Diabetic Retinopathy**: Conjunto de 53,569 imágenes de retina de alta resolución, etiquetadas con 5 clases de la enfermedad retinopatía diabética.
- **Herlev Pap Smear**: Conjunto de 917 imágenes de células de frotis cervicales clasificadas en 7 clases, de las cuales 3 son normales y 4 anormales. Las clases fueron condensadas en 4 categorías ordinales.

Todas las imágenes fueron redimensionadas a una resolución de 224×224 píxeles y normalizadas según la media y desviación estándar de ImageNet por canal de entrada.

Table 1
Datasets used for the experiments.

Dataset	Number of observations	Number of classes	Illustration first class	Illustration last class
Adience [9]	17 702	8		
CBIS-DDSM [21]	2620	6		
Retinopathy ¹	53 569	5		
Herlev Pap-Smear [10]	917	4		

¹ <https://www.kaggle.com/c/diabetic-retinopathy-detection>

Fig. 2. Datasets utilizados en el artículo

3.2 Métricas de evaluación

Para evaluar el desempeño de los métodos de explicación, se utilizaron las siguientes métricas:

- **CCR** (Tasa de Clasificación Correcta): Es la proporción de predicciones correctas realizadas por el modelo, definida como:

$$\text{CCR} = \frac{\text{Número de predicciones correctas}}{\text{Número total de predicciones}} \quad (6)$$

- **AvAUC** (Área Promedio Bajo la Curva): Es la media del área bajo la curva ROC (Receiver Operating Characteristic) para las diferentes clases del modelo. La fórmula es:

$$\text{AvAUC} = \frac{1}{n} \sum_{i=1}^n \text{AUC}_i \quad (7)$$

donde n es el número de clases y AUC_i es el área bajo la curva para la clase i .

- **MAE** (Error Absoluto Medio): Es la media de las diferencias absolutas entre las predicciones del modelo y los valores reales, dada por:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (8)$$

donde y_i es el valor real y \hat{y}_i es la predicción del modelo.

- κ (Cohen's Kappa): Mide la concordancia entre dos observadores o el modelo y la verdad de terreno, teniendo en cuenta la posibilidad de coincidencias aleatorias. La fórmula es:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (9)$$

donde p_o es la proporción de observaciones en las que ambos están de acuerdo, y p_e es la proporción de coincidencias esperadas por azar.

- r_s (Coeficiente de Correlación por Rangos de Spearman): Mide la correlación entre dos variables, basándose en los rangos de los valores, y se define como:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (10)$$

donde d_i es la diferencia entre los rangos de las dos variables para el i -ésimo par de observaciones, y n es el número de observaciones.

La evaluación se centró en la degradación de estas métricas: una mayor degradación indicó un mejor rendimiento de los métodos de explicación.

3.3 Resultados

Los resultados obtenidos (ver Figura 3 para ver los mapas de calor) se presentan mediante diagramas de caja como los de la Figura 4 que muestran la degradación de las métricas para cada conjunto de datos. Los mapas de explicación generados por GradOBD-CAM mostraron un rendimiento generalmente superior en comparación con Grad-CAM++ y Score-CAM, siendo comparable con Grad-CAM en varios casos. La diferencia más destacada se observó en el conjunto de datos Adience, aunque también se evidenciaron mejoras en otros conjuntos.

En el caso de los métodos IBA, OIBA superó a IBA en todos los conjuntos de datos, excepto en CBIS-DDSM y Herlev, donde no se observó una diferencia

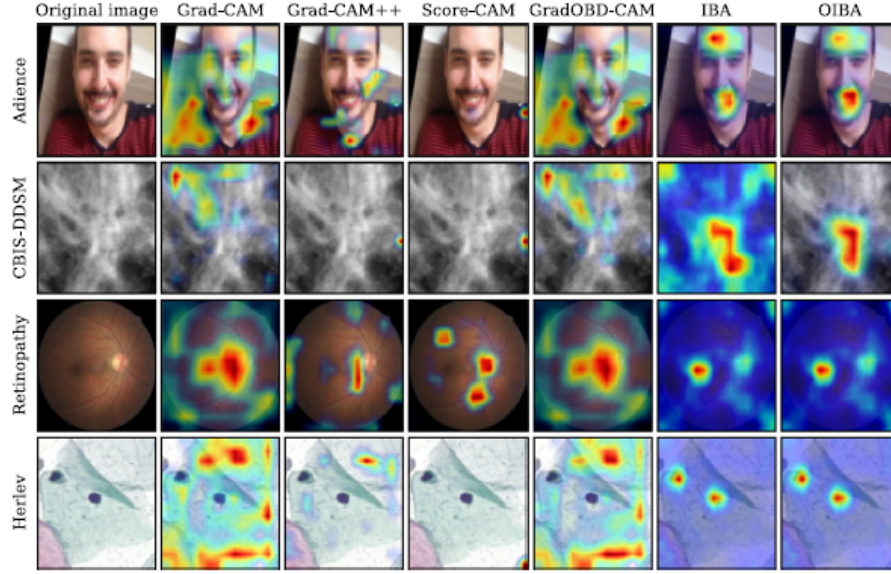


Fig. 6. Example explanation maps from each dataset generated by each one of the compared explanation methods.

Fig. 3. Resultados de los algoritmos estudiados

significativa.

Los resultados fueron validados mediante pruebas de hipótesis, específicamente el test de rangos con signo de Wilcoxon. Este análisis reveló que GradOBD-CAM presentó un mejor rendimiento en términos de mediana para los conjuntos de datos Adience y Herlev en todas las métricas. En el conjunto CBIS-DDSM, GradOBD-CAM también superó a Grad-CAM++ y Score-CAM, pero mostró un rendimiento similar al de Grad-CAM. En el conjunto de Retinopatía Diabética, GradOBD-CAM mostró un mejor desempeño en todas las métricas, excepto en el coeficiente de correlación de Spearman.

4 Análisis y Aplicaciones

4.1 Fortalezas

Las principales fortalezas de GradOBD-CAM y OIBA son su capacidad para mejorar la interpretabilidad en modelos de regresión ordinal. En particular, GradOBD-CAM demostró un rendimiento superior en tres de los cuatro conjuntos de datos, lo que resalta su efectividad para tareas de clasificación ordinal complejas. Además, OIBA mostró un desempeño más alto que IBA en dos de los conjuntos de datos evaluados.

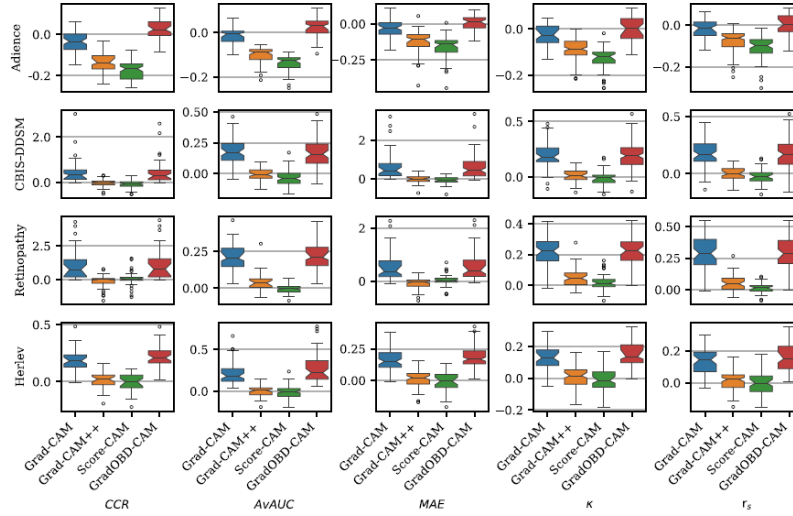


Fig. 4. Diagramas de caja de la degradación de las métricas para los métodos CAM en cada métrica y conjunto de datos (cuanto mayor, mejor). Las muescas de la caja representan el intervalo de confianza alrededor de la mediana. Los puntos fuera de las líneas de los bigotes representan valores atípicos.

Ambos métodos ofrecen una mejor comprensión de los resultados de las redes neuronales, lo que es útil en aplicaciones médicas y en áreas donde la explicabilidad es clave para la toma de decisiones.

4.2 Limitaciones

Una limitación de este estudio es que los métodos fueron probados solo en un conjunto específico de modelos y datasets. Los métodos propuestos podrían no generalizar tan bien a otros tipos de datos o arquitecturas de redes neuronales. Además, los datasets utilizados son relativamente pequeños en comparación con otros datasets de imágenes más grandes, lo que podría limitar la capacidad de los métodos para adaptarse a una mayor diversidad de datos.

Otra limitación es que la técnica de interpretación utilizada se basa en gradientes, lo que puede no ser siempre suficiente para capturar toda la complejidad de los modelos de redes neuronales más grandes y profundos.

4.3 Aplicaciones

Las aplicaciones más relevantes de los métodos GradOBD-CAM y OIBA se encuentran en áreas donde se requiere una interpretación precisa de las decisiones del modelo. Por ejemplo:

- **Diagnóstico médico:** En la clasificación de imágenes médicas, como mamografías o imágenes de retina, la interpretabilidad es crucial para que los profesionales de la salud comprendan las decisiones del modelo y puedan confiar en sus resultados.
- **Sistemas de diagnóstico por imagen:** Para tareas como la clasificación de la edad en imágenes faciales o la detección de anomalías en células de frotis cervical, la capacidad de explicar el proceso de toma de decisiones mejora la transparencia y la confianza en el sistema.
- **Seguridad:** En aplicaciones donde las redes neuronales se usan para identificar patrones en imágenes de vigilancia, la interpretación de los resultados puede ser importante para validar la fiabilidad de los sistemas de visión artificial.

5 Conclusión

En este estudio, se presentaron dos métodos de explicación innovadores para redes neuronales convolucionales aplicadas a la regresión ordinal: **GradOBD-CAM** y **OIBA**. Ambos métodos mostraron mejoras significativas en términos de interpretabilidad en comparación con otros enfoques como Grad-CAM++ y Score-CAM, destacándose especialmente en conjuntos de datos con un mayor número de clases ordinales.

Los experimentos mostraron que GradOBD-CAM tenía un rendimiento superior en tres de los cuatro datasets evaluados, con una diferencia notable en los conjuntos de datos Adience y Retinopatía Diabética. OIBA, por su parte, ofreció mejores resultados que IBA en dos de los conjuntos de datos. Estos resultados respaldan la validez de los métodos propuestos para mejorar la transparencia en modelos de clasificación ordinal.

Para futuros trabajos, se plantea explorar nuevas mejoras en los métodos de explicación, incluyendo la posibilidad de integrarlos con otros modelos de regresión ordinal o incluso con redes neuronales que utilicen distribuciones probabilísticas. También sería interesante investigar cómo la ordinalidad podría integrarse en otras familias de métodos de interpretación, como DeepLift, SHAP o LRP, lo que ampliaría las posibilidades de aplicar estos métodos en una variedad más amplia de tareas y modelos.

References

1. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks (2018), <https://arxiv.org/abs/1711.06104>
2. Barbero-Gómez, J., Cruz, R.P., Cardoso, J.S., Gutiérrez, P.A., Hervás-Martínez, C.: Cnn explanation methods for ordinal regression tasks. *Neurocomputing* **615**, 128878 (2025)
3. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (2018). <https://doi.org/10.1109/WACV.2018.00097>
4. Schulz, K., Sixt, L., Tombari, F., Landgraf, T.: Restricting the flow: Information bottlenecks for attribution (2020), <https://arxiv.org/abs/2001.00396>