



UNIVERSIDAD
DE
CÓRDOBA

Escuela Técnica Superior de Ingeniería

Universidad de Córdoba

Máster en Inteligencia Computacional e Internet de las Cosas

Aprendizaje Profundo



Métodos de explicabilidad de CNN en regresión ordinal

Alumno: Alberto Fernández Merchán

Profesor: Pedro Antonio Gutierrez Peña



Índice

1. Introducción	pág. 1
2. Importancia de la Explicabilidad	pág. 3
3. Métodos y Modelos	pág. 8
4. Experimentos	pág. 11
5. Conclusiones	pág. 14



1. Introducción

Regresión Ordinal

- Clases ordenadas (niveles de gravedad de una enfermedad)
- Conjunto discreto de niveles ordenados

Clasificación Convencional

- Las clases no tienen un orden inherente (gatos o perros)

Regresión Continua

- La variable objetivo puede tomar cualquier valor en un rango



Regresión Ordinal

- Facilita la **identificación de patrones** que distinguen entre clases ordinales (+ **precisión**).
- **Aumenta la interpretabilidad** en contextos como la medicina
- Permite **detectar errores y sesgos** al aprovechar el orden de las clases

Es importante **conocer la diferencia entre niveles**



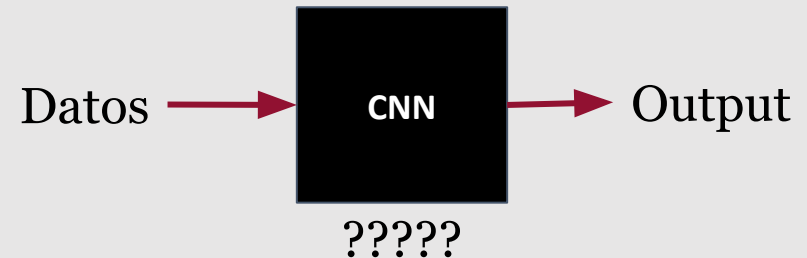


2. Importancia de la explicabilidad

Naturaleza de **caja negra**

¿Te fiarías de extirpar un tumor si te lo dice una IA sin que te explique qué ha “visto”?

¿Por qué ha decidido que es grave o es leve? ¿Qué patrones ha identificado?



Predicciones:

- ✓ Precisas
- ✓ Comprensibles
- ✓ Justificables



2. Importancia de la explicabilidad

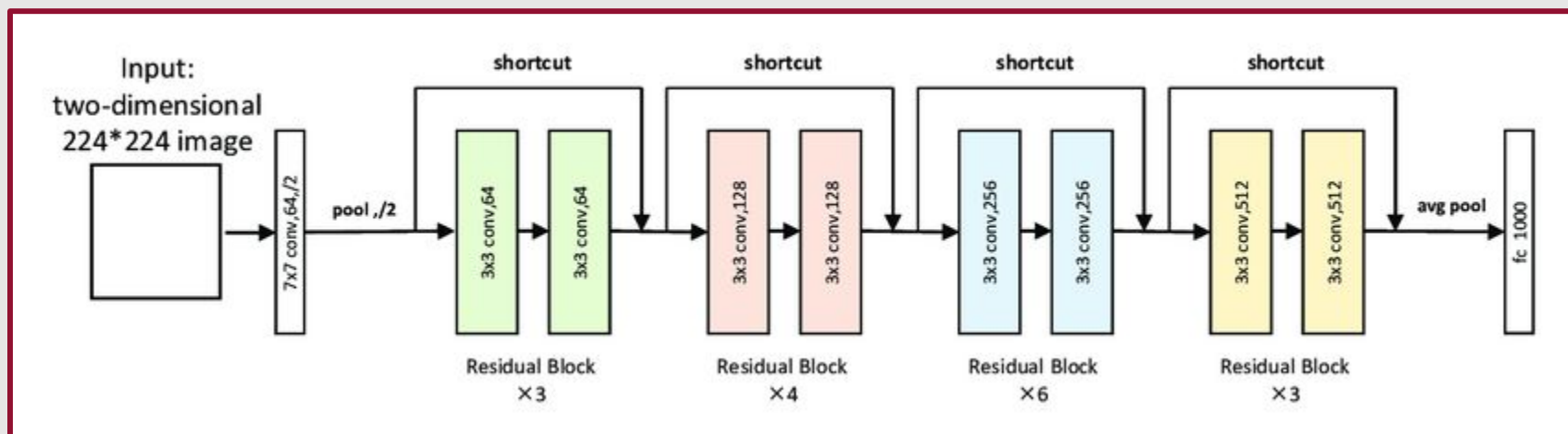
Capas Convolucionales

Capas Activación

Capas Agrupamiento

Extraen
**características
complejas**

Procesamiento
No Lineal de
la Información



Arquitectura de la ResNet34 (A Multi-Dimensional Covert Transaction Recognition Scheme for Blockchain)

Métodos de explicación

Ordinary Binary
Decomposition

Grad-CAM

Grad-CAM++

Score-CAM

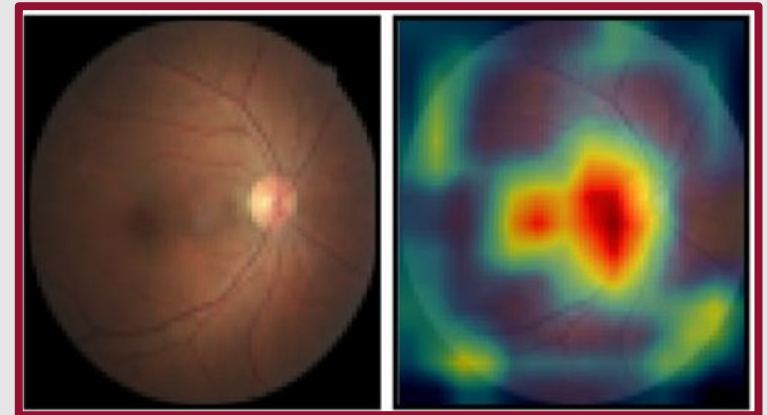
Importance-Based
Attribution (IBA)

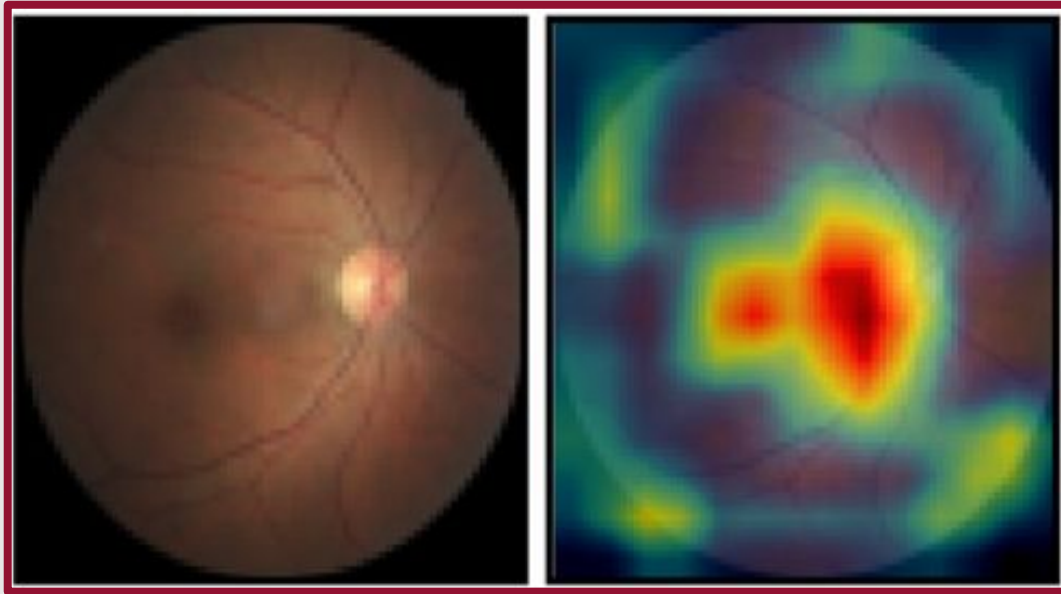
GradOBD-CAM

OIBA

Modelos que
generan **mapas
de calor**

Ayudan a
entender qué
partes de la
imagen son
más influyentes





¿Cómo sabemos si es fiable?



Análisis de
Perturbación

mapas de calor = matriz de explicabilidad

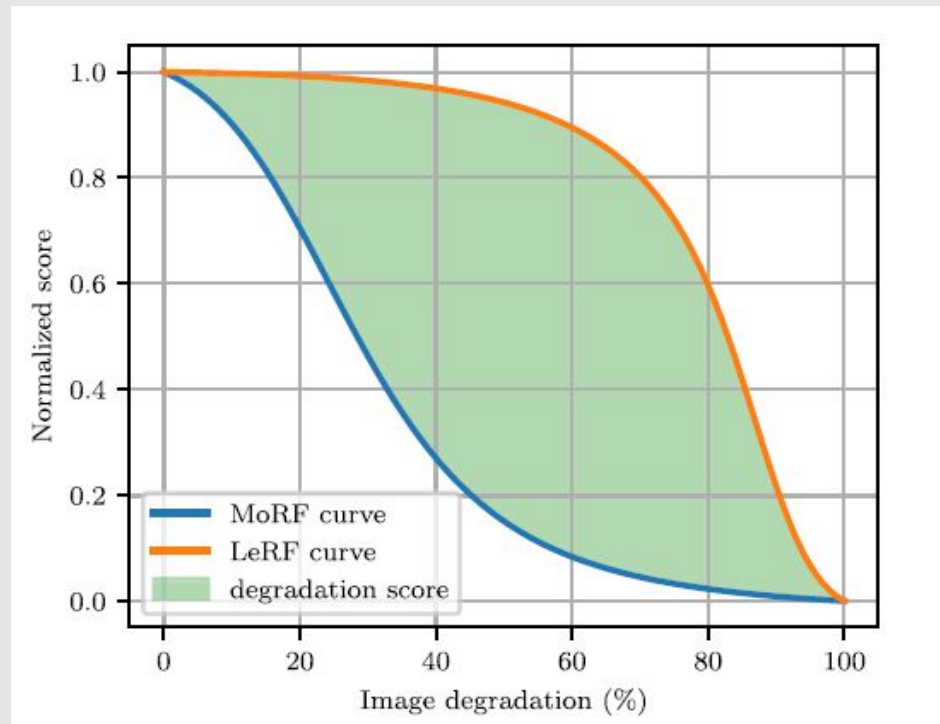
$$E \in \mathbb{R}^{H \times W}$$

- **H**: Altura de la imagen
- **W**: Ancho de la imagen
- **E**: contiene valores de 0 a 1 que indican la importancia



Análisis de Perturbación

Cuantifica la coherencia entre las regiones destacadas por la matriz y su impacto real en la predicción ordinal del modelo



- ◆ **MoRF**: Se reemplazan las regiones más relevantes con valores neutros
- ◆ **LeRF**: Se reemplazan las regiones menos relevantes

Cuanto **mayor** área, **mejor** explicabilidad.

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



3. Métodos y Modelos

Ninguno de estos métodos considera la relación de orden en tareas de regresión ordinal

Ordinary Binary Decomposition

Descompone el problema en **Nº de clases - 1** problemas binarios

Grad-CAM

Genera mapas de activación utilizando los gradientes de la capa final

Grad-CAM++

Redefine cálculos de importancia mejorando la localización y resolución del mapa de activación

Score-CAM

No utiliza gradientes para generar los mapas

Importance-Based Attribution (IBA)

Se basa en su impacto directo sobre la predicción, no en las activaciones intermedias del modelo



GradOBD-CAM

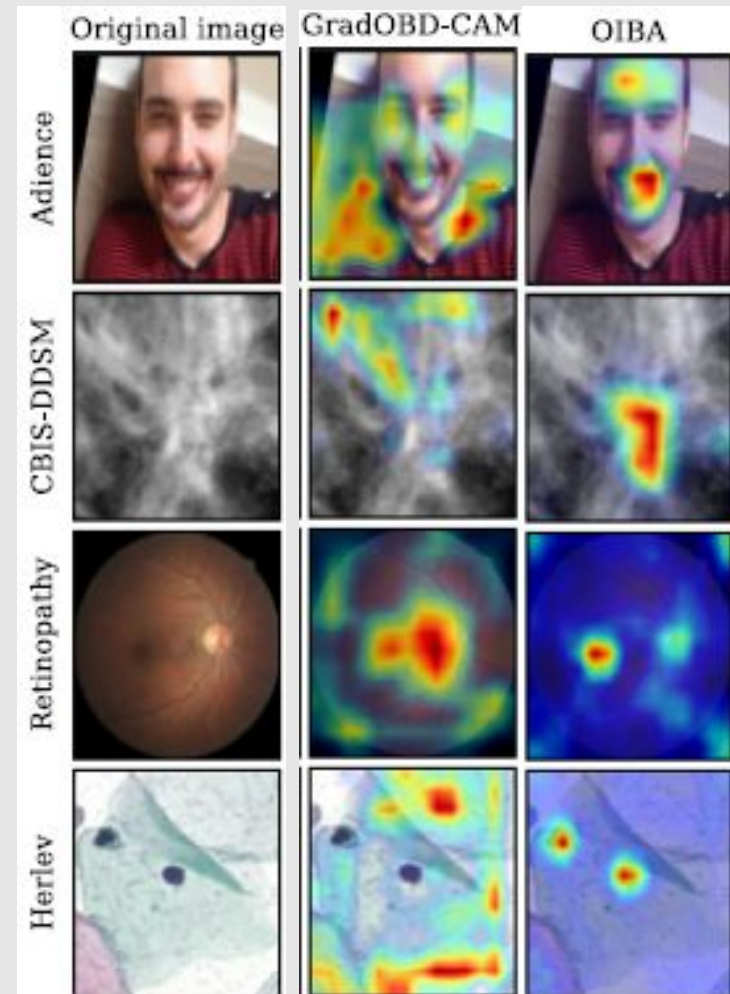
En lugar de centrarse únicamente en la activación de una neurona de salida, utiliza **las derivadas de todas las neuronas de salida** del modelo OBD, donde cada una representa la probabilidad de superar un determinado umbral de clase

Ofrece **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 (Ordinal Importance Based Attribution)



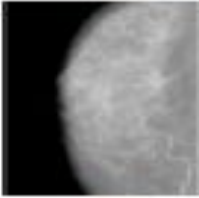
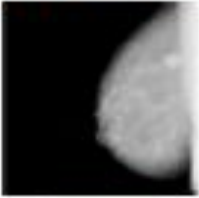

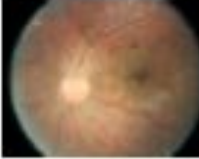


Incorpora la pérdida ordinal propia del modelo OBD en el proceso de optimización de la máscara de perturbación.

Sustituye la **función de pérdida** de la entropía cruzada por la del **error cuadrático**



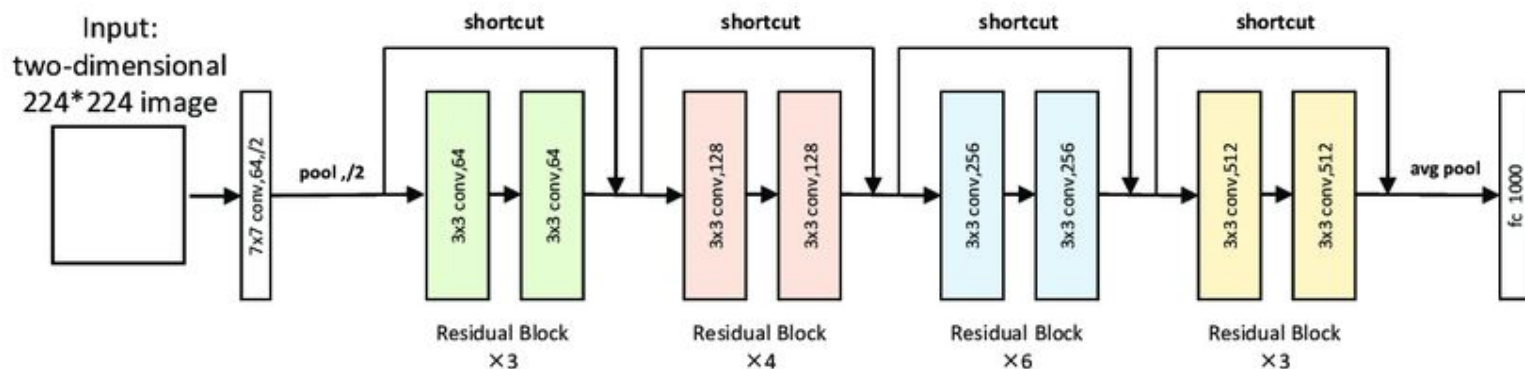
4. Experimentos

DATASETS

Dataset	Number of observations	Number of classes	Illustration first class	Illustration last class
Adience [9]	17702	8		
CBIS-DDSM [21]	2620	6		
Retinopathy ¹	53 569	5		
Herlev Pap-Smear [10]	917	4		

Modelo

RESNET34



✓ Pre-entrenada con ImageNet-1k

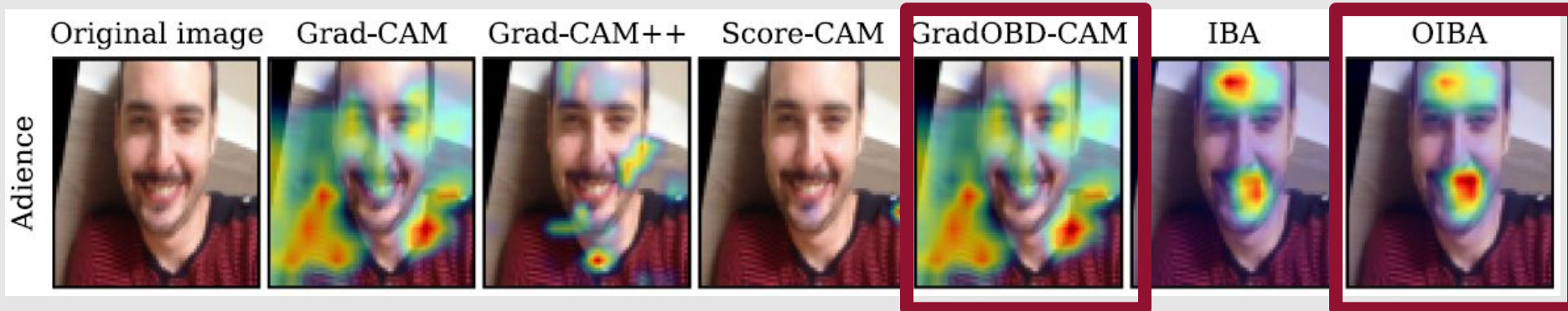
Batches: 64

Epochs: 200

Tr / Val / Test : 80 / 10 / 10

Resultados

Los mapas de explicación generados por **GradOBD-CAM** mostraron un rendimiento superior en comparación con Grad-CAM++ y Score-CAM.



En el caso de los métodos IBA, **OIBA superó a IBA** en todos los conjuntos de datos



5. Conclusiones

Limitaciones

✗ Limitación de modelos y datasets.

Podrían no generalizar bien en otras arquitecturas o tipos de datos



5. Conclusiones

- Métodos para explicar CNNs en regresión ordinal.
- Mejoran significativamente a otros modelos.

→ Integrar con otros métodos de interpretación como DeepLift, SHAP, o LRP.



¿Alguna pregunta?



UNIVERSIDAD
DE
CÓRDOBA

Escuela Técnica Superior de Ingeniería

Universidad de Córdoba

Máster en Inteligencia Computacional e Internet de las Cosas

Aprendizaje Profundo



Métodos de explicabilidad de CNN en regresión ordinal

Alumno: Alberto Fernández Merchán

Profesor: Pedro Antonio Gutierrez Peña