Interpretabilidad y Métodos de Entrenamiento de Modelos de Deep Learning

Facundo Quiroga[1,2], Franco Ronchetti[1,2], Oscar Stanchi[1,6], Gastón Rios[1,3], Pedro Dal Bianco[1,3], Santiago Ponte Ahon[1], Juan Seery[1], Tatiana Badaracco[5], Federico Rabinovich[5], Saif Khalid[4], Hatem Rashwan[4], Domenec Puig Valls[4], Laura Lanzarini[1], Waldo Hasperué[1]

1 Instituto de Investigación en Informática LIDI, Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina.*
2 Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CICPBA)
3 Becario postgrado UNLP
4 Universitat Rovira i Virgili, Tarragona, España
5 Universidad de Buenos Aires, Buenos Aires, Argentina
6 Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
* Centro asociado de la Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CIC)

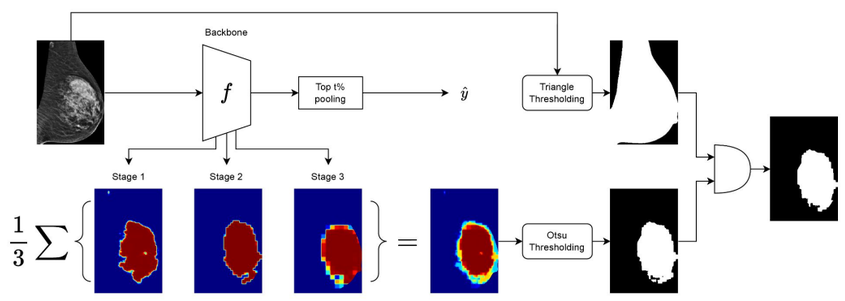
{fquiroga}@lidi.info.unlp.edu.ar

Contexto

Esta presentación corresponde a algunas de las tareas de investigación que se llevan a cabo en el III-LIDI en el marco del proyecto “Inteligencia de Datos. Técnicas y Modelos de Machine Learning” perteneciente al Programa de Incentivos (2023-2026).

Líneas de Investigación y Desarrollo

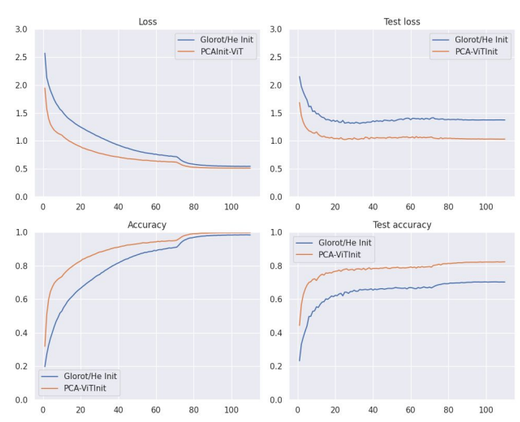
- Redes neuronales profundas, convolucionales y transformers.
- Herramientas de interpretabilidad para visión por computadora
- Interpretabilidad y entrenamiento no supervisado de Redes Convolucionales y ViT.
- Invarianza en redes neuronales, en particular de modelos ampliamente difundidos.



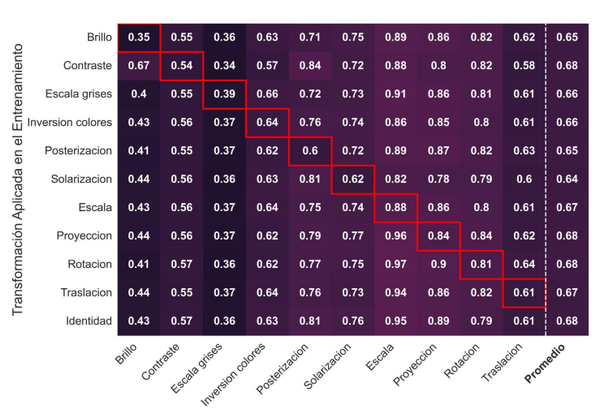
Método propuesto de generación de segmentaciones con etiquetas débiles. Se entrena un clasificador con etiquetas generales de clase (A vs D de la clasificación BI-RADS). Luego se generan máscaras mediante GradCAM, que se utilizan como etiquetas débiles

Resultados Esperados y Obtenidos

- Interpretabilidad mediante invarianza de modelos EfficientNet y ViT
- Entrenamiento no supervisado de modelos EfficientNet y ViT para mejorar desempeño e interpretabilidad.
- Validación mediante métodos de interpretabilidad de un modelo de predicción de calidad de imágenes de retinopatía diabética basado en UNet.
- Validación de un método de entrenamiento semi supervisado para modelos de segmentación de mamografías basado en técnicas de interpretabilidad.



Los filtros de un modelo ViT inicializados mediante el método propuesto mejoran la velocidad de convergencia en algunos escenarios



Invarianza promedio de un modelo FasterViT variando la transformación aplicada en el entrenamiento (filas) y en la evaluación (columnas)

Formación de Recursos Humanos

- 3 Profesores y 1 JTP DE
- 2 Investigadores CIC-PBA
- 4 becarios de posgrado de la UNLP
- 3 Tesis de Doctorado y 2 de Maestría