



ChestAid Proyecto 2



Andre Marroquin Tarot - 22266

Sergio Orellana - 221122

Descripción del problema

En la sala de urgencias a veces el tiempo es crítico. Se necesita priorizar los estudios con mayor sospecha de neumonía para acelerar la lectura médica. El reto es contar con una herramienta simple que, a partir de una Rx de tórax, entregue:

- una clasificación binaria Neumonía vs Normal,
- un score de riesgo calibrado que podamos usar con un umbral clínico,
- y una señal visual Grad-CAM que indique las zonas que más influyeron en la decisión.

Esta herramienta no sustituye la interpretación médica, pretende apoyar el triage y el doble chequeo. Y en ocasiones poder diagnosticar al paciente mucho más rápido y con mayor certeza que la interpretación médica únicamente. Por eso el problema se resume en la velocidad que se necesita en casos de urgencias y a la incertidumbre de 1 sola opinión de un médico.

Análisis

Datos:

Conjunto público, Chest X-Ray Images Pneumonia de Kaggle.

Tamaños en este proyecto: train = 4,434, val = 782, test = 624.

El conjunto está desbalanceado y hay más neumonía que normal. Para compensarlo en el entrenamiento se usó `pos_weight = 0.34` dentro de la función de pérdida.

Contexto:

Los avances tecnológicos y estudios muestran que las CNN profundas ya logran un desempeño competitivo con radiólogos en Rx de tórax. En otras palabras, ya existen muchas investigaciones científicas sobre este tema, buenas prácticas establecidas sobre cómo entrenar y evaluar los modelos, y además bases de datos públicas como la de Kaggle que permiten a investigadores y estudiantes experimentar y comparar resultados fácilmente.

¿Qué se mide y por qué?

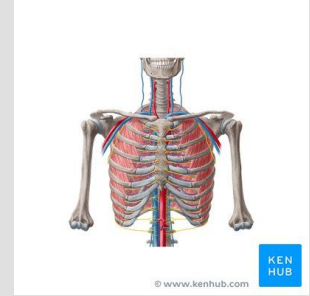
ROC-AUC: mide la capacidad general de separar positivos y negativos sin fijar un umbral.

PR-AUC: útil cuando hay desbalance, se centra en positivos.

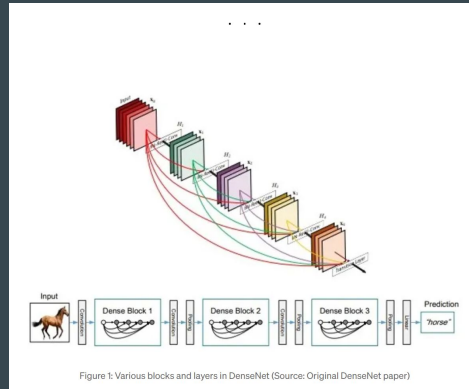
Se decide en la práctica fijar un umbral clínico para alcanzar una sensibilidad objetivo por ejemplo puede ser 0.90. Con ese umbral se reportan sensibilidad, especificidad y la matriz de confusión.

Propuesta de solución

1. Clasificador basado en transfer learning DenseNet-121 con fine-tuning parcial.
2. Entrenamiento con augmentations suaves y pos_weight para el desbalance.
3. Calibración post-hoc Temperature Scaling comparado con Platt/Isotónica para que el score sea más utilizable.
4. Explicabilidad con Grad-CAM sobre la última capa conv.
5. Demo web en Gradio para subir imágenes reales, fijar umbral y visualizar el mapa de calor.



Descripción de la solución



Flujo de datos



- **Preprocesamiento.** Redimensionado 224×224 , normalización ImageNet.
 - **Augmentations (train).** Rotación $\pm 5^\circ$ y jitter ligero de brillo/contraste.
 - **Validación/Test.** Sin augmentations para evaluar en condiciones de uso.
-

Modelo

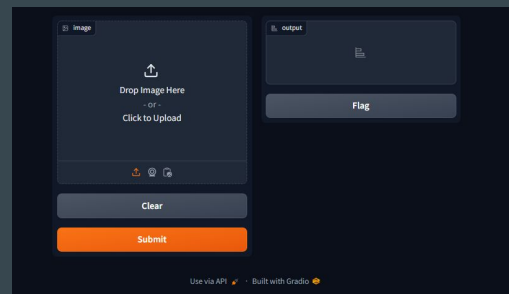
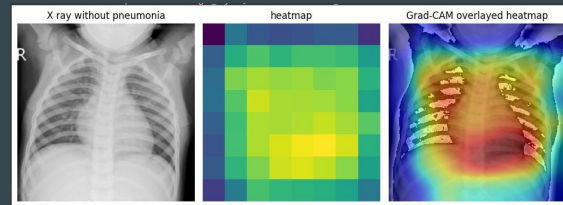
- **Backbone.** DenseNet-121 preentrenada en ImageNet; se reemplazó el head por una salida sigmoide.
- **Pérdida.** BCEWithLogitsLoss con `pos_weight = 0.34` lo que compensa el desbalance.
- **Optimización.** AdamW, ReduceLROnPlateau, entrenamiento con AMP mixta precisión.
- **Early-stopping.** Basado en ROC-AUC de validación.

Calibración

- La validación se partió internamente en dos:
 - `calib` para ajustar el calibrador y
 - `val_report` para reportar sin sesgo.
- Se comparó Temperature Scaling TS, Platt e Isotónica y se seleccionó el mejor por ECE en `val_report`.
- En este run, TS fue el mejor de los tres por poco. Temperatura aprendida aproximadamente del 1.171 el modelo era levemente sobre-confiado, TS aplasta logits y reduce extremos.

Herramientas aplicadas

- **PyTorch / TorchVision.** Definición del modelo, entrenamiento, DataLoaders, evaluación, AMP.
- **Grad-CAM.** Hooks de forward/backward, promedio de gradientes por canal y combinación con activaciones → mapa normalizado 0–1.
- Calibración.
 - **Temperature Scaling:** divide logits por T antes de la sigmoide; corrige sobre-/sub-confianza.
 - **Platt e Isotónica:** calibradores supervisados sobre logits o probabilidades, comparados y medidos por ECE y Brier.
- **Gradio.** Frontend: carga de imágenes, control de umbral y alpha, y despliegue de resultados.
- Algunas buenas prácticas implementadas, Early-stopping, separación train/val/test, split específico para calibración, control del desbalance con pos_weight.



Resultados

Etapa / Método	Conjunto	ROC-AUC	PR-AUC	Brier Score	ECE	Otros datos
Pre-calibración	val_report	0.9984	0.9994	0.0067	0.2723	—
Pre-calibración	test	0.9605	0.965	0.1469	0.3384	—
Temperature Scaling (TS)	val_report	—	—	0.0073	0.2705	Temperatura = 1.1710
Temperature Scaling (TS)	test	—	—	0.1432	0.3325	Mejora respecto a pre-calibración
Platt Scaling	val_report	—	—	0.0075	0.2718	—
Isotónica	val_report	—	—	0.0087	0.2733	—
Mejor calibrador (val_report)	—	—	—	0.0073	0.2705	Temperature Scaling (TS)

El modelo presenta una discriminación muy buena, con valores de ROC-AUC y PR-AUC cercanos a 1 en validación 0.9984 y 0.9994 y ligeramente menores en prueba 0.9605 y 0.9650, lo que demuestra una gran capacidad para distinguir correctamente entre clases positivas y negativas, con mínima pérdida de generalización. En cuanto a la calibración, el modelo muestra un Brier Score bajo 0.0067 y un ECE moderado 0.2723, indicando alta precisión, aunque con cierta sobreconfianza al asignar probabilidades muy extremas.

Tras aplicar Temperature Scaling $T=1.171$, las predicciones se suavizan y la calibración mejora de forma leve pero consistente: el ECE disminuye $0.2723 \rightarrow 0.2705$ en validación y $0.3384 \rightarrow 0.3325$ en prueba y el Brier Score apenas varía, mostrando mejor alineación entre la confianza del modelo y la realidad. Este método logra el mejor equilibrio frente a Platt e Isotónica, manteniendo la buena discriminación original y generando probabilidades más coherentes como scores de riesgo, útiles para decisiones basadas en umbrales definidos según el costo de errores.

Saliency Maps

Métrica	Valor	Lectura en 1 línea
Sufficiency AUC	0.989	evidencia muy concentrada con top-k% pequeño
Deletion AUC	0.601	al borrar top-k% se pierde soporte lo que es coherente
Sanity corr (≈ 0 ideal)	-0.004	mapas dependen de los pesos aprendidos

top-k%

Es el porcentaje k de píxeles más importantes según el mapa de saliencia. Importantes quiere decir aquellos con los valores más altos en el mapa, es decir, las zonas más calientes.



Saliency Maps

k_percent	prob_keep	prob_delete
1	0.998	0.32
10	0.994	0.429
30	0.981	0.62
50	0.987	0.89

prob_keep = probabilidad tras conservar solo ese top-k%. prob_delete = probabilidad tras eliminar ese top-k%.

Con k pequeño, el modelo mantiene casi todo el score lo que dice evidencia concentrada, al borrar esas zonas, el score cae con, lo que indica que sí son influyentes.

Con el 1–10% más relevante, el modelo mantiene casi todo su score, si se elimina justo esas zonas, el score cae. Eso muestra que Grad-CAM no solo dibuja un heatmap, sino que marca evidencia que realmente sostiene la predicción.

Discusión, riesgos y consideraciones

- El uso previsto apoyo al triage y priorización de lectura, no diagnóstico autónomo.
 - Al llevarlo a otro hospital/escáner, es recomendable recalibrar con una cohorte local 100–300 casos sin re-entrenar el backbone.
 - El número mostrado es un score de riesgo calibrado, no una probabilidad clínica exacta $ECE \approx 0.33$ en test.
 - Operar con umbral gobernado , monitorear los errores (FP/FN) y registrar feedback.
-

DEMO

chest-aid — triage asistido por ia

herramienta para apoyar el triage radiológico inicial en rx de tórax (neumonía vs normal). sube una imagen, ajusta el umbral y observa la probabilidad calibrada y el grad-cam.

radiografía de tórax






imagen original



grad-cam (superposición)



umbral de decisión

0

1

0.5

alpha del grad-cam (transparencia)

0

1

0.35

evaluar

temperatura calibrada detectada: 1.171 (archivo encontrado)

resultados

```
1  {
2    "probabilidad_calibrada": 0.9968,
3    "umbral": 0.5,
4    "prediccion": "NEUMONÍA",
5    "temperatura_usada": 1.171,
6    "nota":
7      "la probabilidad está calibrada con temperature scaling; grad-cam resalta
       regiones atencionales."
8  }
```

Conclusión

- Chest-Aid alcanza alta discriminación en test (ROC-AUC \approx 0.96, PR-AUC \approx 0.97), permitiendo priorizar estudios por riesgo con sensibilidad \approx 0.90, especificidad \approx 0.91, F1 \approx 0.92 y PPV \approx 0.94 al usar un umbral clínico de 0.9965.
- Temperature Scaling ($T\approx 1.171$) reduce la sobreconfianza, el modelo ofrece scores de riesgo más consistentes, aunque la ECE en test sigue moderada-alta.
- El valor mostrado debe interpretarse como score de riesgo calibrado, no como probabilidad clínica absoluta; las decisiones se basan en un umbral clínico definido.
- Grad-CAM ofrece interpretabilidad al mostrar regiones relevantes del tórax, aumentando la confianza del usuario, aunque no sustituye criterios diagnósticos ni marca lesiones directamente.

Referencias

- *Chest X-Ray images (Pneumonia)*. (2018, 24 marzo). Kaggle.
<https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia?resource=download>
- *PyTorch Transfer learning with Densenet*. (2018, 27 marzo). PyTorch Forums.
<https://discuss.pytorch.org/t/pytorch-transfer-learning-with-densenet/15579>
- Huang, G., Liu, Z., Laurens, V. D. M., & Weinberger, K. Q. (2016, 25 agosto). *Densely connected convolutional networks*. arXiv.org.
<https://arxiv.org/abs/1608.06993>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal Of Computer Vision*, 128(2), 336-359. <https://doi.org/10.1007/s11263-019-01228-7>
- *Understanding Model Calibration - A gentle introduction and visual exploration of calibration and the expected calibration error (ECE)*. (2025). <https://arxiv.org/html/2501.19047v2>