UNIVERSIDAD DEL VALLE DE GUATEMALA

Proyecto

Andre Marroquin Tarot - 22266
Sergio Orellana - 221122

Deep Learning

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.
- Estructura de carpetas: Train/NORMAL, PNEUMONIA y Test/NORMAL, PNEUMONIA.
- 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?

- En triage lo más importante es, qué tan bien separa el modelo a los casos con y sin neumonía:
    - 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.
- La calibración del score ayuda a que el número sea más interpretable , pero no es la meta principal del prototipo de triage, aquí lo importante es la discriminación y operar con umbral.

Riesgos

- Sobre-confianza del modelo: puede dar números muy altos o bajos aunque no correspondan a una probabilidad real se ve en la ECE. Se puede interpretar tal como tratar la cifra como score de riesgo y decidir con el umbral acordado, no leer el 0.92 cómo 92% seguro, porque tiene una discrepancia no es exacto.
- Grad-CAM que es el mapa de calor es una explicación cualitativa de dónde miró la red, no una marca de lesión. En imágenes claramente normales puede aparecer un punto caliente relativo por la normalización del mapa. Se puede usar como apoyo visual, la decisión la da el score vs umbral.
- Rx con artefactos, textos o exposición atípica pueden confundir al modelo, las etiquetas públicas pueden tener ruido.

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 ±5° 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.

Explicabilidad y producto

- Grad-CAM. Se capturan activaciones y gradientes en la última conv, el mapa se normaliza por imagen y se superpone como heatmap.

- UI Gradio.

    - Subida de Rx, score de riesgo calibrado, y predicción según un umbral.

    - Slider de umbral: define la decisión no cambia el score.

    - Slider de alpha: cambia la transparencia del Grad-CAM visual, no afecta el modelo.

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

Curvas y discriminación

- **Validación** sin augmentations: ROC-AUC ≈ 0.995–0.998, PR-AUC ≈ 0.998–0.999.

- **Test externo:** ROC-AUC = 0.9605, PR-AUC = 0.9650 → discriminación alta.
  El sistema ordena bien los casos, se puede operar con sensibilidad alta manteniendo especificidad razonable.

Calibración (val_report y test)

Métricas reportadas (exactas):

- **pre-calibración | val_report**: roc-auc: **0.9984** | pr-auc: **0.9994**
- **pre-calibración | test**: roc-auc: **0.9605** | pr-auc: **0.9650**
- **val_report (pre)**: Brier **0.0067** | ECE **0.2723**
- **Temperature Scaling (TS)**: temperatura **1.1710** | val_report Brier **0.0073** | val_report ECE **0.2705**
- **Platt**: val_report Brier **0.0075** | val_report ECE **0.2718**
- **Isotónica**: val_report Brier **0.0087** | val_report ECE **0.2733**
- **Mejor calibrador según val_report**: **TS** | ECE **0.2705** | Brier **0.0073**
- **test (pre → post-TS)**: Brier **0.1469 → 0.1432**, ECE **0.3384 → 0.3325**

Discriminación del modelo (ROC-AUC y PR-AUC)

Las métricas de discriminación (ROC-AUC y PR-AUC) miden la capacidad del modelo para distinguir entre clases positivas y negativas. En este caso, el modelo muestra una discriminación casi perfecta en validación (ROC-AUC ≈ 0.9984, PR-AUC ≈ 0.9994), lo que indica que separa de manera muy buena las instancias positivas de las negativas. En el conjunto de test, los valores son algo menores (ROC-AUC = 0.9605, PR-AUC = 0.9650), lo cual sugiere ligera pérdida de generalización.

Calibración (Brier Score y ECE)

El Brier Score mide el error cuadrático medio entre las probabilidades predichas y las etiquetas verdaderas; valores bajos indican buena calibración y menor incertidumbre. El ECE evalúa la alineación entre la confianza del modelo y la frecuencia de aciertos: valores cercanos a 0 significan que las probabilidades predichas reflejan bien la realidad. Antes de calibrar, el modelo presentaba un Brier Score muy bajo (0.0067) y un ECE moderado (0.2723) en validación, lo que sugiere alta precisión pero ligera sobreconfianza, el modelo tiende a asignar probabilidades demasiado extremas cercanas a 0 o 1 respecto a la frecuencia real de los eventos.

Efecto de la calibración

El Temperature Scaling ajusta la temperatura de los logits mediante un único parámetro escalar. Una temperatura $T > 1$ (1.1710) implica que el modelo original era ligeramente sobreconfiado, es decir, las probabilidades predichas eran más extremas de lo que deberían.
Al suavizar esas predicciones, TS mejora levemente la calibración:

- En validación, el ECE baja de 0.2723 a 0.2705 y el Brier sube mínimamente 0.0067→0.0073, un intercambio aceptable dado que la reducción en ECE indica mejor correspondencia entre confianza y realidad.

- En test, también hay mejora, Brier 0.1469→0.1432 y ECE 0.3384→0.3325, lo que confirma consistencia del calibrador al generalizar fuera del conjunto de entrenamiento.

Los calibradores Platt e Isotónico ofrecen resultados similares, pero TS logra el mejor equilibrio entre estabilidad, suavidad y rendimiento, evitando el sobreajuste que puede ocurrir con la calibración isotónica en conjuntos pequeños.

El modelo mantiene una excelente discriminación y una calibración buena tras aplicar Temperature Scaling. Aunque la ECE no cae de manera drástica sigue moderada, ~0.27 en validación y ~0.33 en test, las mejoras en consistencia de las probabilidades justifican su uso.

- Las probabilidades calibradas deben interpretarse como scores de riesgo, no como probabilidades clínicas absolutas.

- Las decisiones deben basarse en umbrales predefinidos según el costo de falsos positivos/negativos.

- El valor de temperatura (1.171) indica que el modelo inicial era confiado pero no excesivamente mal calibrado, y el ajuste posterior mejora la confianza sin afectar la discriminación.

Operación clínica umbral a sensibilidad objetivo

Se fijó el umbral para sensibilidad ≈ 0.90 en test:

- **Umbral:** 0.9965
- **Matriz de confusión (N=624):**
    - TP = 351, FP = 22, TN = 212, FN = 39
- **Sensibilidad:** 351 / (351 + 39) = **0.900**
- **Especificidad:** 212 / (212 + 22) = **0.906**
- **Exactitud:** (351 + 212) / 624 = **0.902**
- **PPV (Precisión):** 351 / (351 + 22) ≈ **0.941**
- **NPV:** 212 / (212 + 39) ≈ **0.845**
- **F1-score:** 2·351 / (2·351 + 22 + 39) = 702 / 763 ≈ **0.920**
- **Balanced accuracy:** (0.900 + 0.906) / 2 ≈ **0.903**

El sistema permite un triage sensible alrededor de 90% de neumonías detectadas con baja tasa de falsos positivos PPV aprox 94%. El NPV refleja la prevalencia alta de neumonía en el dataset 62.5%; en poblaciones con menor prevalencia, el NPV subiría.

Grad-CAM ofrece contexto visual de las regiones que empujan la decisión hacia neumonía.

En casos con score bajo puede aparecer un hot-spot pequeño por normalización relativa por imagen; la decisión final la define el score vs. umbral, no el mapa.

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 (coherente)
Sanity corr (≈0 ideal)	−0.004	mapas dependen de los pesos aprendidos

Tabla 1. Saliency Maps metrics.

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.

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

Tabla 2. Saliency Maps metrics.

Se aplicaron pruebas de suficiencia y eliminación sobre los píxeles más relevantes según Grad-CAM. Los resultados muestran que incluso con un 1% de píxeles, el modelo conserva casi todo el puntaje, lo que indica alta concentración de evidencia. Al eliminar esas zonas, la probabilidad baja notablemente, confirmando su influencia en la decisión. En general, Grad-CAM resulta coherente y útil como guía visual del modelo.

Interpretabilidad con Grad-CAM validación cuantitativa

Se aplicaron pruebas post-hoc de sufficiency, deletion y sanity sobre el conjunto de prueba n=64, con temperature scaling T = 1.171, sufficiency_fill = mean y deletion_fill = blur. Los resultados fueron: Sufficiency AUC = 0.989, Deletion AUC = 0.601 y Sanity corr = −0.004. La suficiencia cercana a 1 dice que la evidencia del modelo está altamente concentrada, con un porcentaje pequeño de píxeles relevantes se conserva casi todo el score. La eliminación moderada AUC de aprox 0.60, medida con relleno blur para evitar artefactos, muestra que al eliminar esas regiones el soporte del modelo disminuye en promedio, lo que respalda que los mapas señalan zonas realmente influyentes donde puede haber neumonia. La correlación cercana a cero en la prueba de sanity confirma que los mapas dependen de los pesos aprendidos.

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≈0.33 en test.

- Operar con umbral gobernado , monitorear los errores (FP/FN) y registrar feedback.

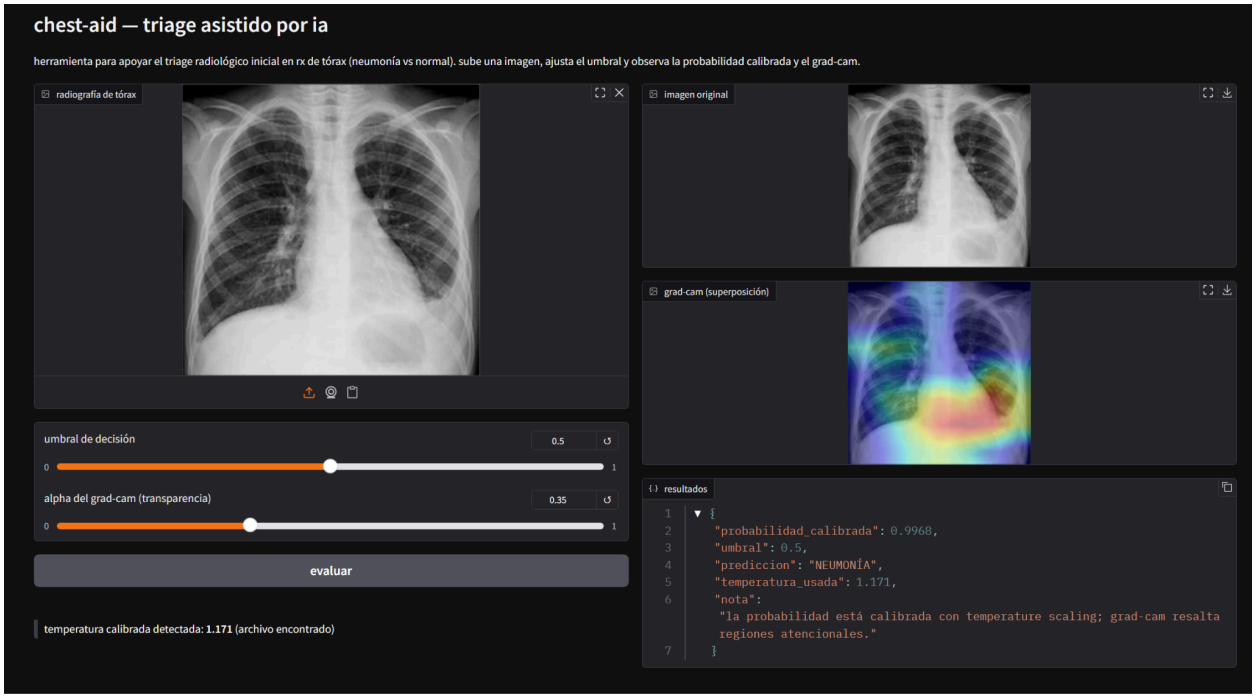
Conclusión

Chest-Aid demuestra que, con un pipeline cuidadoso y transfer learning, es posible construir un asistente de triage para Rx de tórax con discriminación sólida en test ROC-AUC≈0.96, PR-AUC≈0.97. Este nivel de desempeño permite ordenar eficazmente los estudios por riesgo y operar con sensibilidad alta al fijar un umbral clínico, con el umbral seleccionado 0.9965 se obtuvo sensibilidad ≈0.90, especificidad≈0.91, F1≈0.92 y PPV≈0.94, cifras adecuadas para priorización en entornos de urgencia.

La calibración mejoró con Temperature Scaling (T≈1.171): el modelo redujo su sobre-confianza Brier y ECE bajan levemente, pero la ECE en test se mantiene moderada-alta. Por ello, el valor mostrado al usuario debe comunicarse como score de riesgo calibrado y no como una probabilidad clínica literal, la decisión se toma con un umbral clínico gobernado. Esta distinción favorece un uso responsable del sistema en triage, donde el objetivo principal es no pasar por alto casos potencialmente positivos.

La señal visual con Grad-CAM aporta explicabilidad cualitativa, ayuda a entender dónde miró la red al elevar el score, reforzando la confianza del usuario cuando el patrón es anatómicamente plausible. Aun así, el mapa de calor no es una máscara de lesión ni criterio diagnóstico por sí mismo; la etiqueta final depende del score vs. umbral.

Demo



Se puede visualizar el umbral de decisión la probabilidad calibrada entre otros datos importantes como la predicción si tiene o no Neumonía.

Bibliografía

- *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