EXTENSIONES DE MACHINE LEARNING

Fine Tunning para Clasificación de Neumonía en Imágenes de Rayos-X de Tórax (CNN vs ViT)

PRÁCTICA FINAL BLOQUE II

PABLO MARTINEZ GALINDO

ÍNDICE

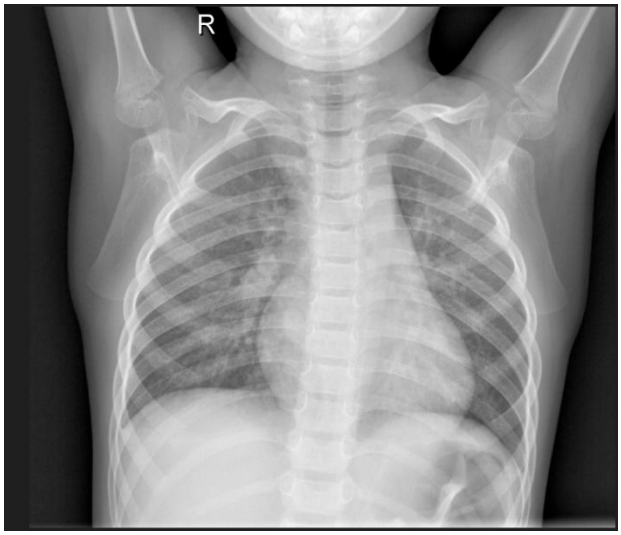
1. DEFINICIÓN DEL PROBLEMA

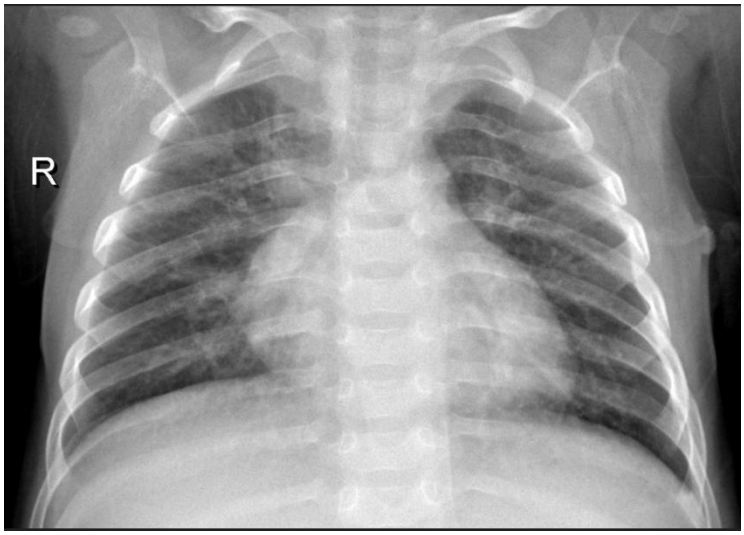
1.1. OBJETIVO

Este proyecto tiene como objetivo principal el desarrollo, reentrenamiento (fine-tunning) y comparación de modelos de aprendizaje profundo para la **clasificación binaria automática** de imágenes médicas. La tarea consiste en discriminar automáticamente entre radiografías de tórax de pacientes con **Neumonía** y pacientes **Normales** (sanos), comparando el rendimiento de arquitecturas convolucionales (CNN) frente a Transformers de Visión (ViT).

1.2. CONTEXTO Y JUSTIFICACIÓN

La neumonía es una infección respiratoria que afecta a los pulmones y representa una de las principales causas de mortalidad infantil a nivel mundial. El diagnóstico estándar se realiza mediante radiografías de tórax, un proceso que depende de la disponibilidad y capacidad de radiólogos expertos. En entornos de alta demanda asistencial o recursos limitados, la fatiga visual puede conducir a errores de diagnóstico. La implementación de un sistema de **Ayuda al Diagnóstico (CAD)** fiable permitiría actuar como una herramienta de triaje, agilizando la detección de casos positivos y reduciendo la carga de trabajo del personal clínico.





1.3. CRITERIOS DE ÉXITO

Dada la naturaleza crítica del problema médico, la métrica principal para evaluar el éxito no será la exactitud global (*Accuracy*), sino la **Sensibilidad o Recall (Tasa de Verdaderos Positivos)** para la clase *Pneumonia*. El objetivo es minimizar los falsos negativos (pacientes enfermos clasificados erróneamente como sanos), incluso si esto conlleva un ligero aumento en los falsos positivos.

2. RECOLECCIÓN DE DATOS Y CARACTERIZACIÓN

2.1. FUENTE DE DATOS

La elección del dataset **"Chest X-Ray Images (Pneumonia)"**, derivado de estudios pediátricos del *Guangzhou Women and Children's Medical Center*, no es trivial. Al compararlo con otras alternativas disponibles en Kaggle, su selección se fundamenta en criterios específicos que se alinean mejor con los objetivos de este proyecto. A continuación, se presenta una comparativa con otros datasets destacados en el estado del arte:

A. Dataset "RSNA Pneumonia Detection Challenge"

El dataset de la *Radiological Society of North America (RSNA)* es uno de los más populares en Kaggle para esta temática. Sin embargo, fue descartado por las siguientes razones:

- **Objetivo Diferente (Detección vs. Clasificación):** El dataset de RSNA está diseñado para tareas de detección de objetos, incluye bounding boxes para localizar las opacidades pulmonares. Para este proyecto centrado en la **clasificación binaria** (enfermo/sano), la complejidad de procesar metadatos de localización es innecesaria y desvía el foco del aprendizaje de arquitecturas de clasificación como CNN y ViT.
- **Formato de Archivo:** Originalmente se distribuye en formato DICOM, lo cual añadiría una capa extra de conversión a PNG/JPG.

B. Dataset "COVID-19 Radiography Database"

Este dataset, creado por investigadores de la Universidad de Qatar y otras instituciones. Se descartó por:

- **Problema Multiclase:** Contiene 4 clases: COVID-19, Normal, Neumonía Viral y Opacidad Pulmonar. Aunque es un buen dataset, introduce el desafío de la clasificación multiclase distinto a la idea inicial.

En conclusión, se eligió el dataset **"Chest X-Ray Images (Pneumonia)"** porque representa el **equilibrio óptimo**:

- **Complejidad Controlada:** Su naturaleza binaria y volumen (~5,800 imágenes) permiten entrenar modelos en tiempos razonables.
- **Desafío Realista:** A pesar de ser "pequeño" en comparación con otros datasets, presenta **desbalanceo de clases** y **variabilidad dimensional** (imágenes de distintos tamaños), obligando a implementar técnicas robustas de preprocesamiento y Data Augmentation que son vitales para la formación en Ingeniería de Datos.
- **Calidad:** Garantiza etiquetas validadas manualmente, eliminando el ruido que podría confundir el análisis de los resultados.

2.2. DESCRIPCIÓN DE LA VARIABLE OBJETIVO

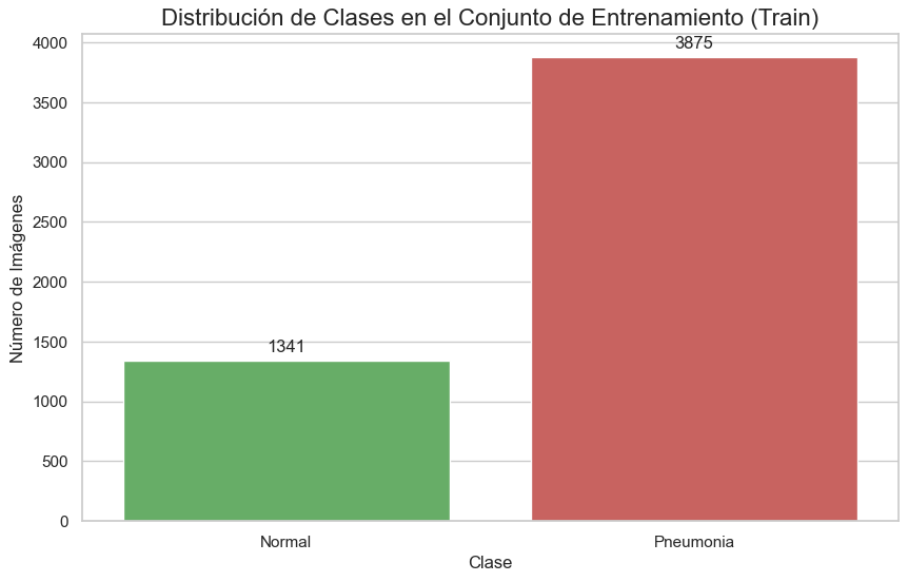
El dataset consta de imágenes no estructuradas (radiografías) organizadas en dos categorías (variable objetivo binaria):

- **Clase 0 (NORMAL):** Radiografías de pacientes sanos.
- **Clase 1 (PNEUMONIA):** Radiografías con presencia de consolidación pulmonar.

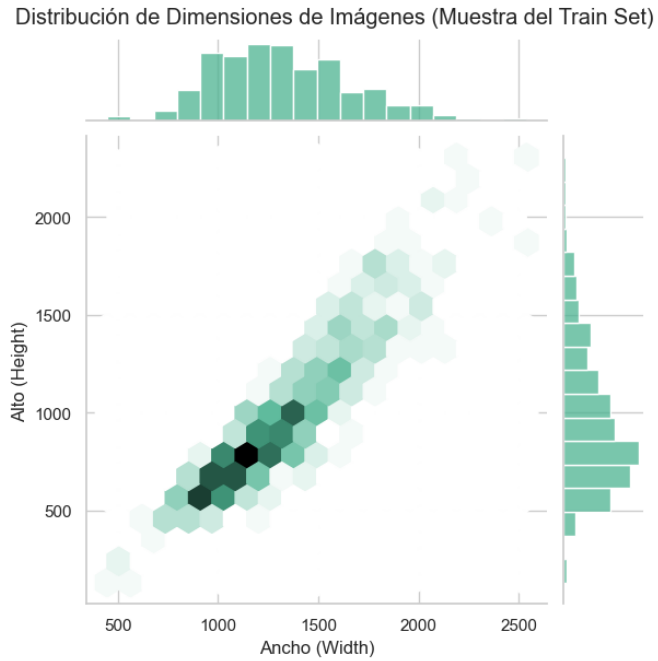
2.3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

El análisis estadístico inicial reveló las siguientes características críticas:

1. **Desbalanceo de Clases:** Se detectó una desproporción significativa en los datos originales de entrenamiento (5,216 imágenes), donde aproximadamente el **74%** corresponden a *Pneumonia* (3,875 imágenes) y solo el **26%** a *Normal* (1,341 imágenes).



2. **Variabilidad Dimensional:** Las imágenes presentan resoluciones muy heterogéneas, con un tamaño medio aproximado de 1319x968 píxeles, pero con una alta desviación estándar, oscilando entre anchos de 446 a 2538 píxeles.



3. **Canales:** Las imágenes originales están en escala de grises (1 canal), lo cual contrasta con la entrada esperada por los modelos pre-entrenados en ImageNet (3 canales RGB).

3. LIMPIEZA Y PREPARACIÓN DE DATOS

Para adecuar los datos a las arquitecturas de Deep Learning seleccionadas, se implementó el siguiente pipeline de pre-procesamiento:

3.1. GESTIÓN DE LA DIVISIÓN DE DATOS (SPLIT)

Se descartó la partición original de validación por ser estadísticamente insignificante (16 imágenes). Se fusionaron los conjuntos train y validación originales en un conjunto de desarrollo y se generó una nueva partición estratificada: 80% Entrenamiento (4,185 imágenes) y 20% Validación (1,047 imágenes), preservando la proporción de clases. Se mantuvo el conjunto de Test original (624 imágenes) como Hold-out para la evaluación final.

3.2. TRANSFORMACIONES Y CODIFICACIÓN

Redimensionado: Todas las imágenes fueron re-escaladas a una resolución fija de 224x224 píxeles (IMG_SIZE).

Adaptación de Canales: Se convirtió el espacio de color a RGB triplicando el canal de grises para compatibilidad con pesos de ImageNet.

Normalización: Se estandarizaron los tensores utilizando las medias (0.485, 0.456, 0.406) y desviaciones (0.229, 0.224, 0.225) de ImageNet.

3.3. TRATAMIENTO DEL DESBALANCEO

Se implementó una estrategia de Oversampling (sobremuestreo) utilizando un WeightedRandomSampler. Se asignaron pesos inversos a la frecuencia de las clases, aumentando la probabilidad de muestreo de la clase minoritaria (NORMAL) para garantizar batches equilibrados durante el entrenamiento.

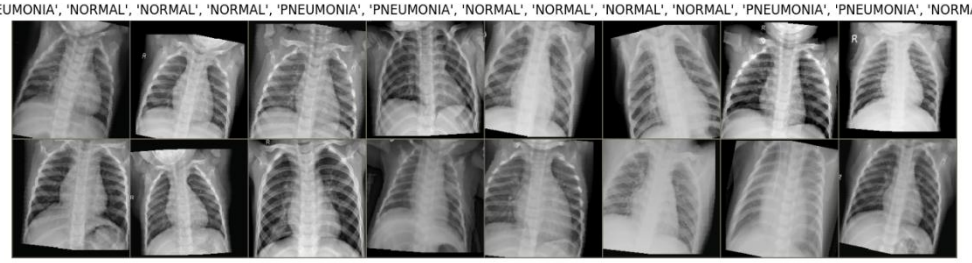
Se aplicó Data Augmentation (rotaciones aleatorias, traslaciones y ajustes de brillo/contraste) exclusivamente al conjunto de entrenamiento para evitar el sobreajuste (overfitting) derivado de la repetición de muestras.

3.4. COMPROBACIÓN DE LA PREPARACIÓN DE DATOS

Visualizamos un batch de entrenamiento para confirmar:

1. Que las imágenes tienen el tamaño correcto (224x224).
2. Que el **WeightedSampler** está funcionando (deberíamos ver una mezcla equilibrada de Normal/Neumonía, no solo Neumonía).
3. Que el **Data Augmentation** está aplicando rotaciones ligeras.

Batch de Entrenamiento (Balanceado): ['PNEUMONIA', 'NORMAL', 'NORMAL', 'NORMAL', 'PNEUMONIA', 'PNEUMONIA', 'NORMAL', 'NORMAL', 'NORMAL', 'NORMAL', 'PNEUMONIA', 'PNEUMONIA', 'NORMAL', 'PNEUMONIA', 'PNEUMONIA', 'NORMAL']



4. MODELADO Y EVALUACIÓN

4.1. SELECCIÓN DE ALGORITMOS

Se seleccionaron dos arquitecturas pre-entrenadas en ImageNet:

1. **EfficientNet-B0 (CNN):** Representante de redes convolucionales eficientes. Se modificó la última capa (classifier) por una capa lineal de 2 salidas.
2. **ViT-Base-16 (Vision Transformer):** Representante de modelos basados en atención. Se modificó la cabeza por una capa lineal de 2 salidas.

4.2. ENTRENAMIENTO Y ESTRATEGIAS

Se realizaron dos experimentos secuenciales para cada modelo:

- **Fase 1: Fine-Tunning:** Congelación del resto de capas y entrenamiento exclusivo del clasificador.
- **Fase 2: Full-Fine-Tuning:** Descongelación de todos los pesos para re-entrenar la red completa con una tasa de aprendizaje baja (1e-5).

Se ha definido una función global para entrenar los modelos en la que se usa como función de pérdida **CrossEntropyLoss**.

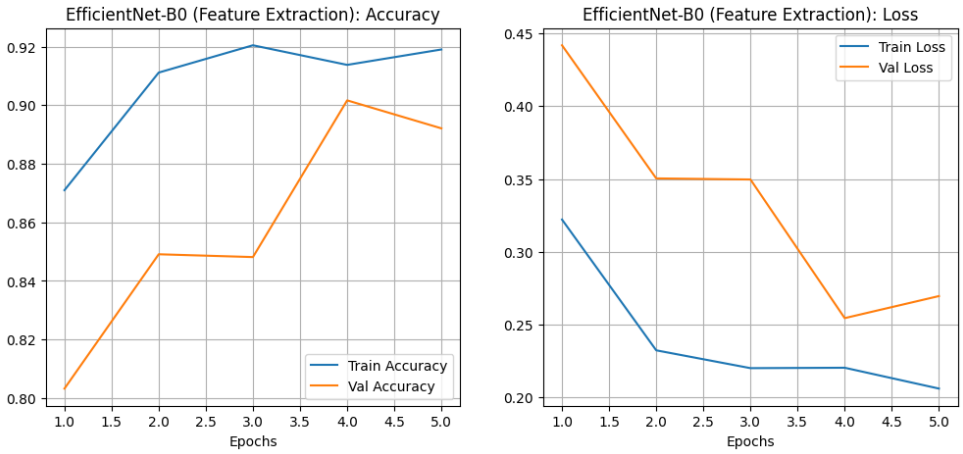
4.2.1. FINE-TUNNING

4.2.1.1. ENTRENAMIENTO DE LOS MODELOS

Para realizar el entrenamiento del clasificador de la CNN (**EfficientNet-B0**) y del Vision Transformer (**ViT-Base-16**) se aplicó la siguiente configuración de **hiperparámetros:**

- **Optimizador**: AdamW (Learning Rate = 0.001).
- **Épocas:** 5 (Suficiente para ver convergencia rápida en la cabeza).
- **Balanceo:** Gestionado por el **WeightedRandomSampler** definido en el DataLoader.

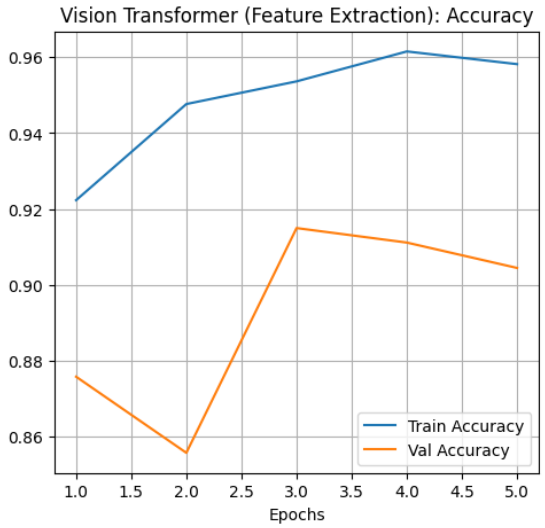
En el caso de la **EfficientNet-B0** las curvas de aprendizaje fueron las siguientes:

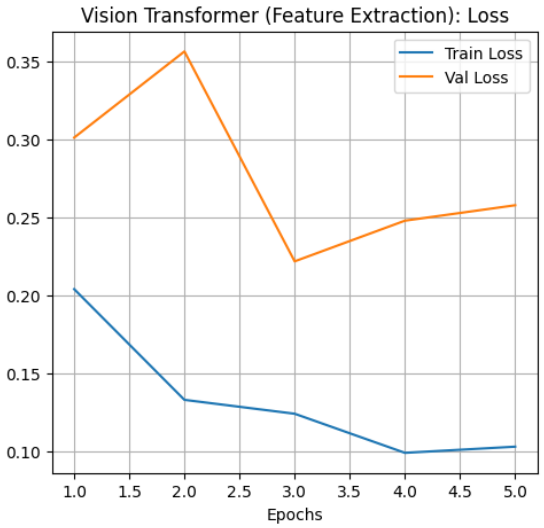


Las gráficas muestran una rápida convergencia de la pérdida y una ausencia total de **overfitting**, consecuencia directa de mantener congelado el cuerpo de la red. Sin embargo, la exactitud parece se estanca en una **~89%**, representando **underfitting** leve. Este comportamiento confirma que,

aunque la inicialización es robusta, es imprescindible avanzar a la fase de **Full-Fine-Tuning** para adaptar los filtros profundos a las características de las radiografías y superar este techo de rendimiento.

Para **ViT-Base-16** tenemos los siguientes resultados del entrenamiento:

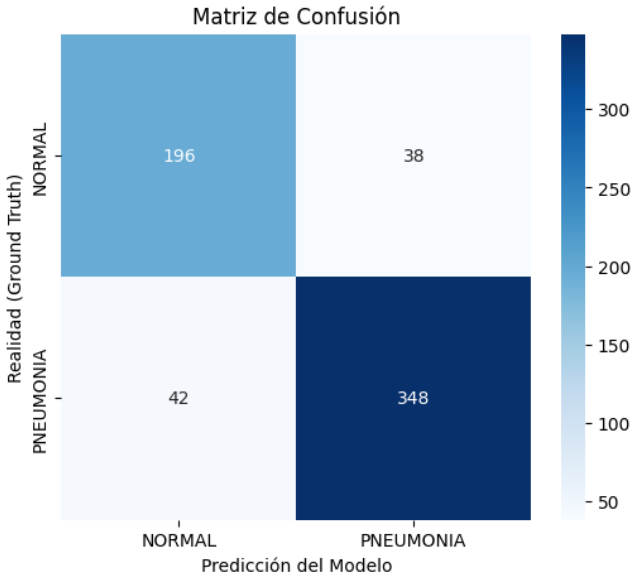




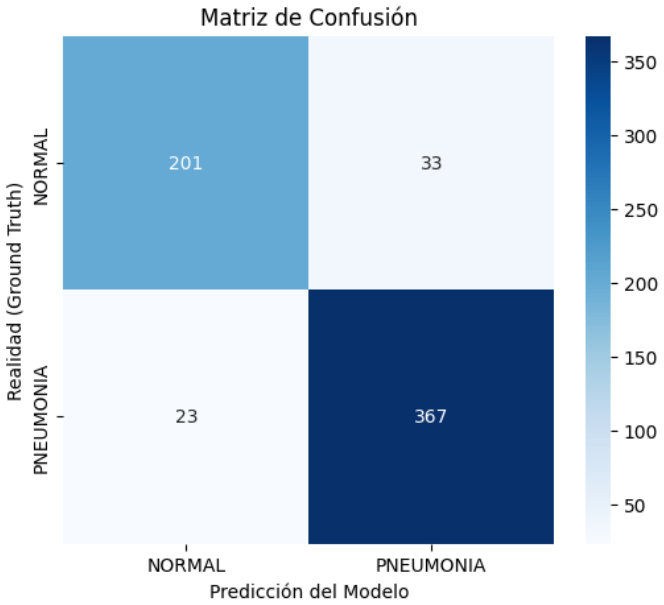
El ViT muestra falta de ajuste en validación, con la exactitud entre (85% - 91%) y la función de pérdida estabilizándose a partir de la tercera época en aproximadamente en **~0.25**. A diferencia de la CNN, aquí se aprecia un **overfitting** claro: el modelo aprende muy rápido en *train* (96%) pero le cuesta transferir ese conocimiento. Esto confirma que los mapas de atención pre-entrenados en ImageNet no son directamente extrapolables a radiografías, haciendo que el paso de **Full-Fine-Tuning** sea obligatorio para alinear los embeddings y estabilizar el modelo.

4.2.1.2. EVALUACIÓN DE LOS RESULTADOS

En cuanto a los resultados obtenidos de la **CNN,** muestran un desempeño base aceptable (**Accuracy 87%**), pero insuficiente para el objetivo clínico debido a un **Recall de Neumonía del 89.2%**. Esto implica una tasa de **falsos negativos cercana al 11%**, dejando sin diagnosticar a una fracción significativa de pacientes enfermos. Esta limitación confirma que el clasificador lineal, por sí solo, es incapaz de capturar las sutilezas radiológicas necesarias para un cribado seguro, haciendo imprescindible el **Full-Fine-Tuning** para maximizar la sensibilidad.



El modelo **ViT** sin embargo, demuestra una capacidad de generalización inicial superior a la CNN, alcanzando una **Exactitud del 91%** y elevando el **Recall de Neumonía al 94.1%** con los pesos congelados. Aunque el mecanismo de atención global identifica mejor los patrones patológicos de base, todavía persiste un **6% de falsos negativos** (pacientes enfermos no detectados). Este margen de error confirma que, pese al buen rendimiento basal, se debe aplicar **Full-Fine-Tuning** para perfeccionar la sensibilidad y alcanzar la fiabilidad clínica requerida.



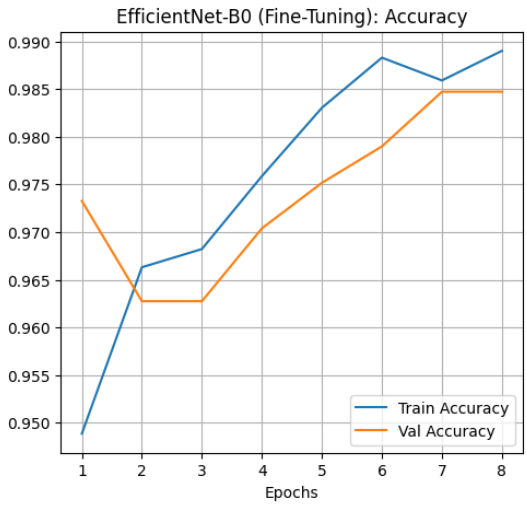
4.2.2. FULL-FINE-TUNNING

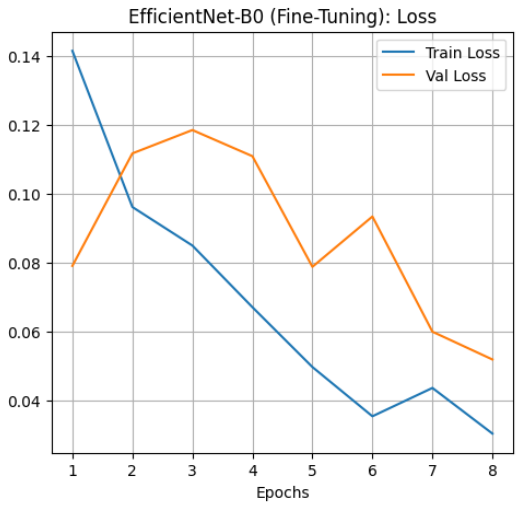
4.2.2.1. ENTRENAMIENTO DE LOS MODELOS

Ahora a la hora de realizar el entrenamiento realizamos los siguientes cambios para intentar mejorar los resultados de nuestros modelos:

- **Descongelar**: Liberamos todos los pesos del modelo.
- **Learning Rate Bajo** (1e-4): Para refinar suavemente los pesos sin destruir el pre-entrenamiento.
- Aumentamos a **8 épocas** de entrenamiento**.**

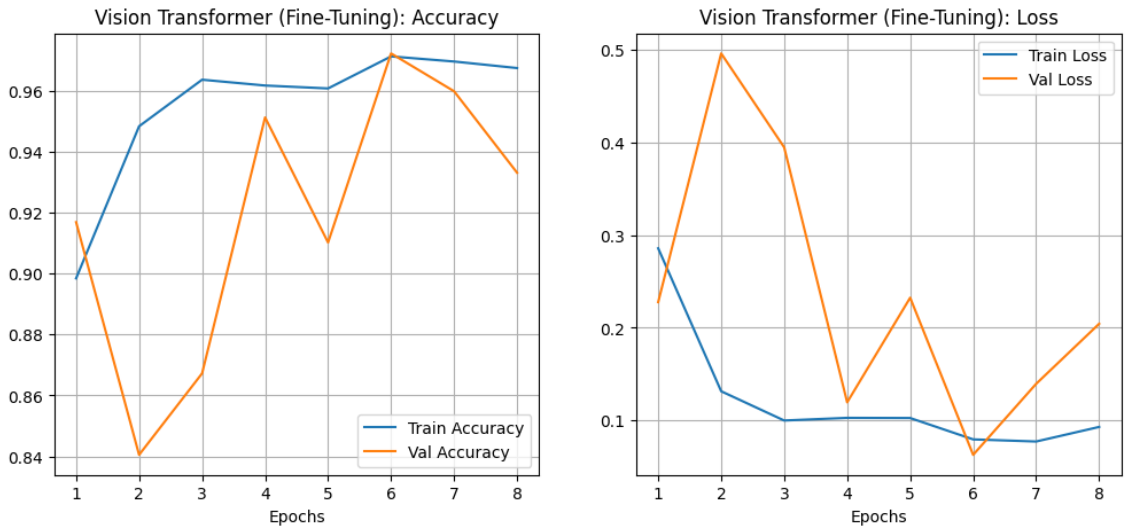
Al volver a entrenar la **CNN** con esta configuración obtenemos las siguientes curvas de aprendizaje:





El descongelado de pesos desbloquea el potencial de la CNN, reduciendo la función de pérdida de validación hasta un mínimo de **0.0518** y elevando la exactitud al **98.47%**. Se puede decir que modelo converge muy bien superando la limitación de la fase anterior. La relación entre las métricas de entrenamiento y validación confirma una **buena capacidad de generalización**, demostrando que la red ha aprendido patrones específicos sin caer en sobreajuste.

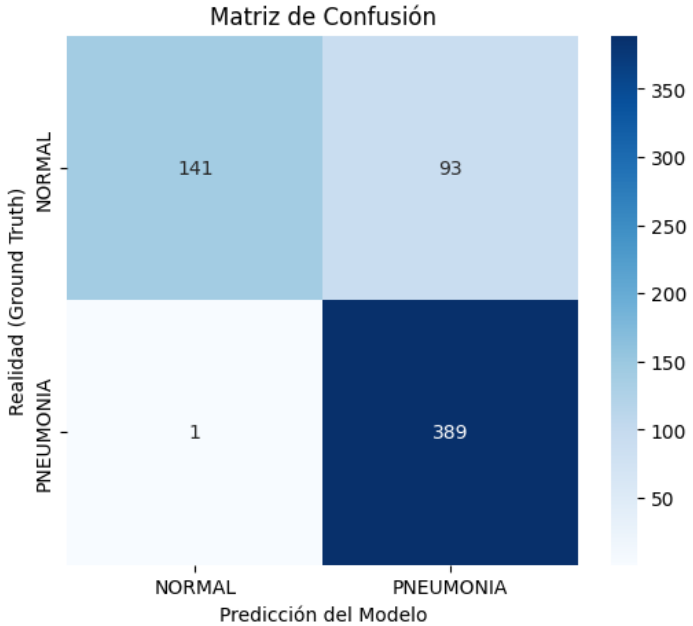
Por otro lado, el ViT nos devuelve la siguiente curva de aprendizaje:



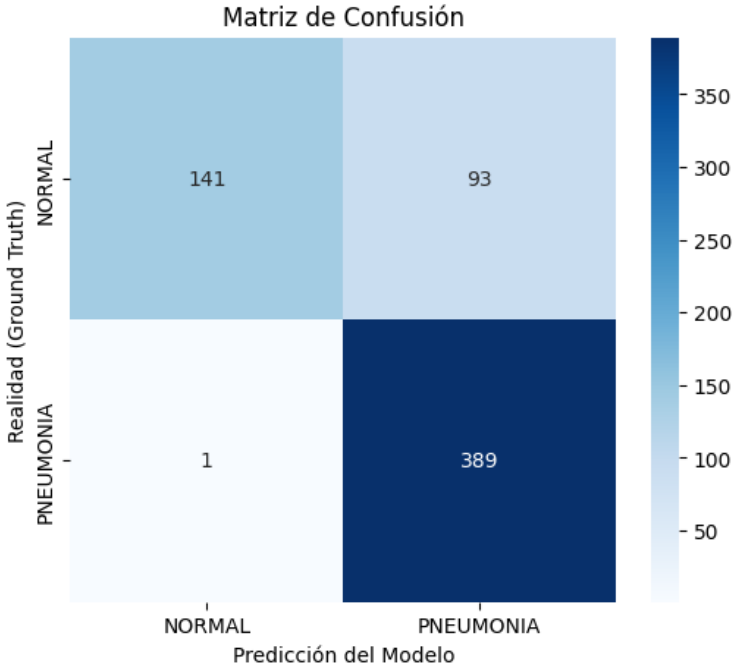
Este otro muestra una **alta inestabilidad**, con cambios drásticos en la validación (la exactitud rebota bruscamente) sin aparentar estabilidad. Aunque en la **época 6** logra un pico de rendimiento muy bueno (**Loss 0.06**, **Acc 97.2%**) equiparable a la CNN, su incapacidad para sostener la convergencia (cayendo al 93% posteriormente) indica la dificultad de estabilizar Transformers en datasets limitados. Esta volatilidad confirma que, aunque el ViT tiene un alto potencial, carece de la robustez y consistencia de entrenamiento que mostró la EfficientNet.

4.2.2.2. EVALUACIÓN DE LOS RESULTADOS

A la hora de hablar de los resultados obtenidos en la clasificación de la CNN tras aplicarle un full-fine-tunning, el modelo alcanza el objetivo clínico del proyecto con una **Sensibilidad (Recall) del 99.74%** para Neumonía, lo que implica una detección casi infalible de casos positivos. Este comportamiento agresivo genera un trade-off claro: aumenta los falsos positivos (reduciendo el Recall de la clase NORMAL al 60%) y sitúa la Exactitud global en un 85%. Sin embargo, este resultado valida al sistema como una herramienta de **triaje de alta seguridad**, priorizando la detección de patologías sobre la especificidad, tal como se definió en los objetivos.

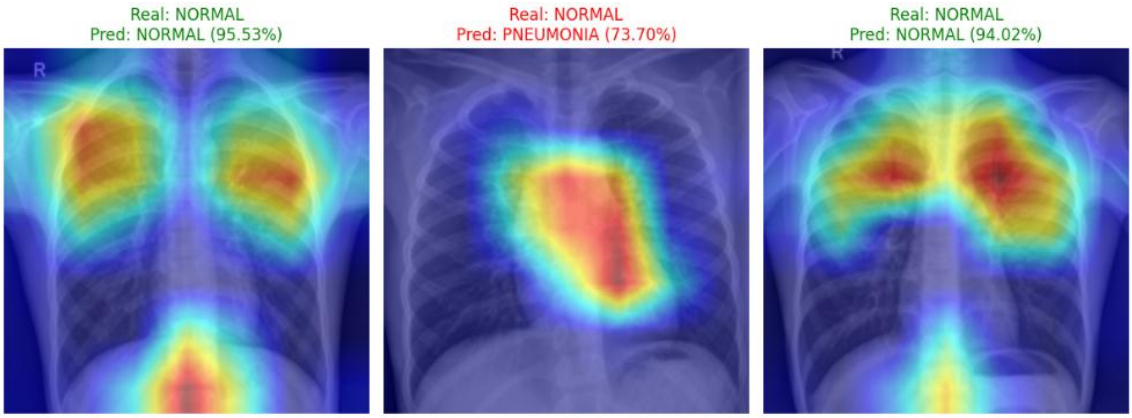


El ViT consolida su adaptación logrando un **Recall de Neumonía del 99.49%**, comportamiento prácticamente idéntico al de la CNN en cuanto a seguridad clínica. Con una **Exactitud del 87%**, este modelo se valida como una herramienta robusta que combina una sensibilidad extrema para detectar enfermos con una relativamente buena fiabilidad en el descarte de sanos.



4.3. EXPLICABILIDAD DEL MODELO (GRAD-CAM)

Para mitigar el efecto de "caja negra" de a las redes neuronales profundas y validar la relevancia clínica de las predicciones, se ha utilizado la técnica **Grad-CAM (Gradient-weighted Class Activation Mapping)** sobre la última capa convolucional de la arquitectura EfficientNet-B0.



Como se observa en la figura adjunta, esta técnica genera mapas de calor superpuestos a las radiografías originales, donde las zonas rojas/cálidas indican las regiones que más influyeron en la decisión del modelo:

- **Validación Anatómica:** Los mapas de activación confirman que la red centra su atención principalmente en la zona de los **pulmones** y el diafragma, ignorando correctamente el ruido de fondo, etiquetas, huesos u órganos irrelevantes.

- **Detección de Patología:** En los casos clasificados como PNEUMONIA (Verdaderos Positivos), el modelo resalta las áreas de **consolidación, opacidad o infiltrados**, coincidiendo con los signos radiológicos típicos de la enfermedad. Mientras que en los clasificados como PNEUMONIA que en realidad eran NORMAL (Falso Positivo) se muestra que el modelo se ha fijado en la zona del corazón en lugar de fijarse en la cavidad pulmonar.

Este análisis cualitativo aporta un grado de confianza adicional al sistema, demostrando que el alto rendimiento numérico (Recall >99%) está respaldado por el aprendizaje de características visuales.

5. INTERPRETACIÓN DE RESULTADOS Y CONCLUSIONES

5.1. HALLAZGOS PRINCIPALES

1. **Impacto del Full-Fine-Tuning:** La estrategia de Full-Fine-Tuning ha sido decisiva. En la fase de extracción de características, la CNN tenía un Recall de 0.89. Tras el ajuste fino, el Recall subió a **0.9974**, equivocándose únicamente en 1 de los 390 casos positivos de test.
2. **Comparativa CNN vs ViT:** Aunque el Vision Transformer (ViT) es una arquitectura más moderna, la red convolucional (**EfficientNet-B0**) demostró un rendimiento ligeramente superior en sensibilidad y eficiencia computacional para este dataset específico, probablemente debido a que las CNNs requieren menos datos para generalizar correctamente en comparación con los Transformers.

5.2. CONCLUSIÓN

El modelo final basado en **EfficientNet-B0 con Fine-Tuning** cumple exitosamente con el objetivo clínico planteado. Al alcanzar un **Recall del 99.7%** para la clase Pneumonia, el sistema minimiza excelentemente el riesgo de falsos negativos, validándose como una herramienta de triaje segura y efectiva.

5.3. LIMITACIONES Y TRABAJO FUTURO

- **Falsos Positivos**: Para maximizar la sensibilidad, el modelo ha sacrificado precisión en la clase NORMAL (Precision ~0.85), lo que implica una tasa moderada de falsas alarmas (pacientes sanos marcados como enfermos).
- **Datos Pediátricos**: El dataset está restringido a pacientes de 1 a 5 años. El modelo podría no generalizar correctamente en radiografías de adultos sin un re-entrenamiento previo con datos demográficos más amplios.

6. HERRAMIENTAS UTILIZADAS

El desarrollo se ha realizado en **Python 3**, utilizando las siguientes librerías especializadas para cada etapa del flujo de datos:

- **Ingestión y Estructuración:**
  - **glob / os:** Lectura de archivos y gestión de rutas.
  - **pandas:** Organización de metadatos en DataFrames.
  - **numpy:** Operaciones matriciales y cálculo estadístico base.

- **Procesamiento de Imágenes:**
  - **PIL / OpenCV:** Carga y lectura de imágenes.
  - **torchvision.transforms:** Pipeline de preprocesamiento (Redimensionamiento, Conversión a Tensor y Normalización con medias de ImageNet).

- **Visualización y Análisis (EDA):**
    - **matplotlib / seaborn:** Generación de gráficos estadísticos y visualización de predicciones.

- **Modelado (Deep Learning):**
    - **PyTorch (torch, torch.nn):** Framework para cálculo de tensores, autograd y construcción de capas neuronales.
    - **torchvision.models:** Importación de arquitecturas pre-entrenadas para CNN y ViT.
    - **DataLoader / WeightedRandomSampler:** Gestión de batches y balanceo de clases durante el entrenamiento.

- **Evaluación:**
    - **sklearn.metrics:** Cálculo de métricas objetivas (Matriz de Confusión, Recall, F1).
    - **Grad-CAM:** Herramienta de explicabilidad para visualizar mapas de calor sobre las radiografías.