Universidad Panamericana

Maestría en Inteligencia Artificial y Ciencia de Datos

Análisis comparativo de arquitecturas CNN, Vision Transformer e híbridas para clasificación de tumores en imágenes médicas

Asignatura: Aprendizaje Profundo

Profesor: Ricardo Abel Espinosa Loera

Alumna: Miriam Elizabeth López Hernández

2025

Resumen

La clasificación de imágenes médicas mediante aprendizaje profundo ha mejorado la precisión diagnóstica y reducido tiempos de análisis. Este estudio compara el desempeño de tres arquitecturas: una CNN tradicional (ResNet-50), un Transformer puro (Vision Transformer, ViT) y un modelo híbrido (Swin Transformer). Se evaluaron métricas de clasificación y costo computacional empleando un dataset público balanceado. Los resultados mostraron que ResNet-50 destacó en precisión y eficiencia, el ViT en generalización, y el modelo híbrido combinó ventajas locales y globales. Estos hallazgos sugieren que los enfoques híbridos ofrecen un balance prometedor para aplicaciones clínicas.

1. Introducción

La detección temprana de tumores es fundamental para mejorar la tasa de supervivencia y optimizar tratamientos. Sin embargo, el análisis manual de imágenes médicas es costoso, demandante y sujeto a variabilidad del observador. El deep learning ha revolucionado el campo al permitir clasificación y segmentación automáticas, con arquitecturas que extraen características relevantes de forma jerárquica. Entre ellas destacan:

- CNNs: exitosas en visión por computadora por su capacidad de capturar patrones locales.
- Transformers (ViT): sobresalen en modelar dependencias globales mediante mecanismos de auto-atención.
- Modelos híbridos: combinan convoluciones y atención, integrando ventajas de ambos enfoques.

El objetivo de este trabajo es comparar el desempeño de estas tres arquitecturas en la clasificación binaria para el diagnóstico de Cáncer de Esófago, analizando tanto métricas de precisión como el costo computacional, para aportar evidencia en la selección de modelos aplicables en contextos clínicos.

2. Trabajos relacionados

- CNNs han sido dominantes en la última década, destacando ResNet, VGG e Inception, aplicadas en tareas de clasificación y segmentación de tumores, con resultados consistentes en precisión y eficiencia.
- Transformers (ViT) han demostrado gran capacidad en imágenes naturales y se han trasladado al ámbito médico, mostrando ventajas en contextos con suficiente data y estrategias de preentrenamiento.
- Arquitecturas híbridas (ej. Swin-UNet, TransUNet) han emergido como alternativas para combinar el modelado local (CNN) y global (self-attention), alcanzando el estado del arte en segmentación y clasificación.

3. Metodología

3.1. Datos y preprocesamiento

Para este clasificador se utilizaron solamente las imágenes de tejido sano y las imágenes de tejido con displasia/cáncer. De manera que el conjunto de datos está formado por 1,469 imágenes de tejido sano (clase 0) y 3,594 imágenes de displasia/cáncer (clase 1). Las imágenes originales fueron escaladas de 519x521 pixeles a 260x260 para reducir el tiempo y la memoria requeridos para el procesamiento. Para el desarrollo de los modelos el dataset fue dividido en: 70 % training, 15 % test, 15 % validación.

3.2. Análisis y solución para el Desbalanceo

Se observó que el dataset está altamente desbalanceado, con la clase 1 teniendo muchas más muestras que la clase 0. Un modelo entrenado en este dataset podría sobreajustarse a la clase mayoritaria y tener un mal desempeño en la clase minoritaria. Para abordar esto, se aplicó una estrategia de balanceo de clases durante el entrenamiento. En este caso, podríamos llegar a usar el sobremuestreo (oversampling) de la clase minoritaria. Esto implica duplicar o triplicar aleatoriamente las imágenes de la clase con menos muestras para igualar el número de imágenes en ambas clases. Esto se implementará directamente en la función de carga de datos para los sets de entrenamiento.

3.3. Preprocesamiento de datos

Se aplicaron técnicas de preprocesamiento: redimensionamiento uniforme, normalización y aumentación de datos (rotaciones, flips, escalados). Todas las entradas se redimensionaron a 224×224, con normalización ImageNet. Lo anterior debido a que los valores de la imagen deben estar en una escala estándar que los modelos ya "esperan", porque casi todos los backbones preentrenados fueron entrenados con esa normalización. Restar la media y dividir por la desviación estándar de ImageNet en RGB. Se aplicaron augmentaciones consistentes (flip, rotate, color jitter limitado, contraste o saturación). Para Transformers (ViT/Swin) se añadió regularización más agresiva (Mixup/ CutMix y DropPath) por su menor sesgo inductivo.

3.4. Parámetros de entrenamiento

Los modelos se entrenaron durante 5 épocas utilizando un optimizador Adam y una tasa de aprendizaje constante. La función de pérdida utilizada fue CrossEntropyLoss y un Learning Rate de 0.001

4. Resultados y discusión

Todos los modelos fueron entrenados en Google Colab Pro, empleando una GPU NVIDIA Tesla T4 (16 GB de VRAM, arquitectura Turing). El entorno de ejecución utilizó Python 3.10, PyTorch 2.0.1 y CUDA 11.8

4.1. Métricas de desempeño

Cuadro 1: Métricas de clasificación por modelo.

Modelo	Acc.	Recall	F1	Prec.
ResNet-50	0.90	0.95	0.93	0.92
ViT	0.71	1.00	0.83	0.71
Híbrido	0.71	1.00	0.83	0.71

4.2. Costo computacional

Cuadro 2: Costo computacional comparado.

Modelo	#Parámetros (M)	Tiempo (s)
ResNet-50	23	313
ViT	86	730
Híbrido	45	785

5. Conclusiones y trabajo futuro

En general, la superioridad del ResNet-50 en este estudio puede atribuirse a su capacidad para capturar las características jerárquicas y locales que son intrínsecamente importantes en imágenes médicas para la clasificación de tumores. Aunque los Transformers son excelentes para capturar dependencias globales, su falta de un sesgo inductivo fuerte para las características locales podría haber sido una desventaja en este

dataset relativamente pequeño. Es importante mencionar que, la literatura menciona que los modelos con Transformers son superiores cuando se tienen millones de datos por lo que para este proyecto donde se tenían menos de 6 mil, pudo ser una desventaja para estos modelos. Como trabajo futuro, se propone entrenamiento experimentando con una mayor cantidad de épocas, *fine-tuning* con *schedulers* y técnicas de interpretabilidad (Grad-CAM, mapas de atención).

Referencias

[1] P. K. Mall et al., "A comprehensive review of deep neural networks for medical image processing," *Healthcare Analytics*, 2023.

[2] M. E. Rayed et al., "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, 2024.

Repositorio de código

El código fuente de este trabajo está disponible en Repositorio en GitHub.