UNIVERSIDAD CENTRAL DE VENEZUELA 08 DE AGOSTO DE 2025

FACULTAD DE CIENCIAS

ESCUELA DE COMPUTACIÓN

DEEP LEARNING

Juan Fernández 28447019

TAREA #3 DEEP LEARNING

ViT: El Visual Transformer

1. **Adaptación del Codebase al Dataset BRICS:**

- Se reemplazó el dataset MNIST (blanco y negro, 28x28) por el dataset médico BRICS (imágenes RGB, variadas).
- Se utilizó ImageFolder porque BRICS ya viene organizado en carpetas por clase.
- Se implementó un DataLoader con transformaciones específicas:
    - Resize(224, 224): para que todas las imágenes tengan el tamaño esperado por ViT.
    - RandomHorizontalFlip, ColorJitter, RandomRotation: aumentan el dataset (data augmentation).
    - ToTensor() y Normalize(): convierten la imagen a tensor y normalizan los valores entre -1 y 1 (normalización centrada).

Esto ya que el ViT necesita entradas de tamaño fijo y canales RGB. Además, un modelo más robusto necesita **más datos distintos**, por eso el aumento de datos.

2. **Entendiendo Vision Transformer (ViT):**

ViT (Vision Transformer) aplica la arquitectura de transformers (originalmente usada en NLP) a imágenes.

En la implementación podemos ver que funciona de la siguiente manera:

- **Patches**: Divide la imagen 224x224 en bloques más pequeños (por ej. 16x16). Resultado: 14x14 = **196 patches**.
- **Patch embedding**: Cada patch se aplana (flatten) y se transforma en un vector
- **[CLS] token**: Se agrega un token especial al principio que resume toda la imagen.
- **Positional Encoding**: Se le suma información de posición a cada patch (porque los transformers no "ven" el orden).
- **Transformer encoder**: Aplica self-attention para que cada patch "vea" a los otros y sepa qué es y qué no es relevante.
- **MLP Head**: Solo el token [CLS] se pasa por una red neuronal para predecir la clase (Benigno, Maligno, …)

Esto facilita a entender relaciones espaciales complejas mejor a como lo hacen los CNN.

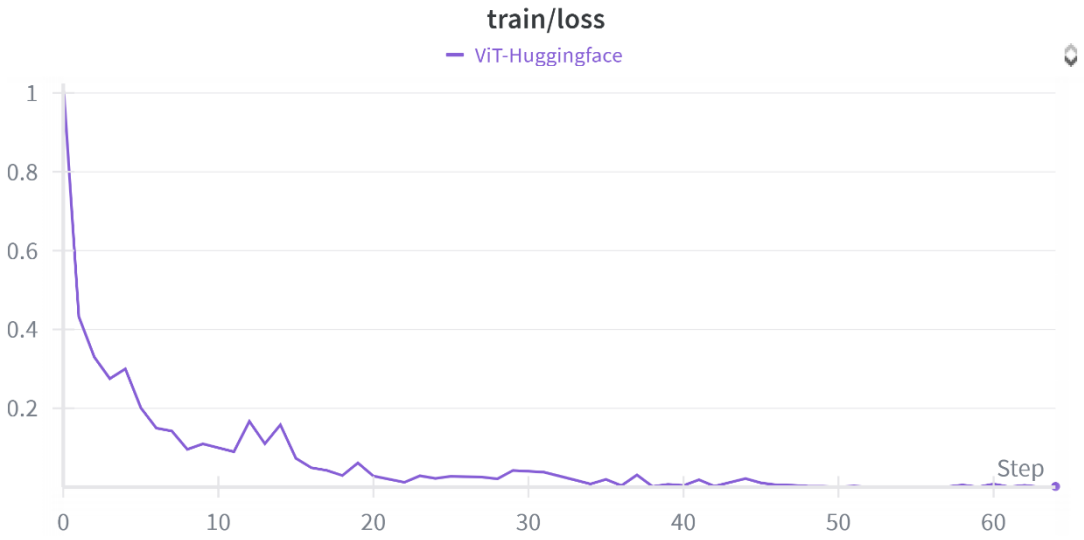
3. **Atención en ViT y su utilidad:**

   Es una representación de "qué partes" de la imagen miró el modelo para tomar su decisión. Se genera a partir del token [CLS] que se conecta con todos los patches de la imagen. El valor de atención mide cuánto peso le dio el modelo a cada región.

   ¿Por qué es útil? Ej. Medicina por BRISC

   - Proporciona explicabilidad visual: permite saber dónde "miró" el modelo para detectar un tumor.
   - Ayuda a validar si el modelo tomó en cuenta la zona correcta (lesión o masa).
   - Es un paso clave hacia modelos clínicamente confiables y auditables.
   - Evita que el modelo clasifique por *artefactos irrelevantes* (texto, marcas, bordes).

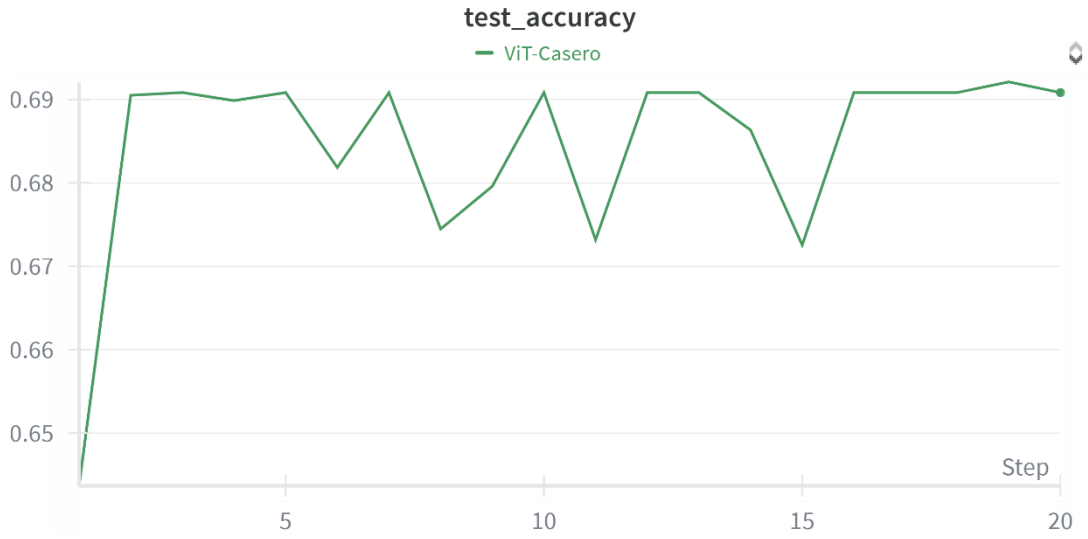
4. Comparación: ViT Casero vs Huggingface:

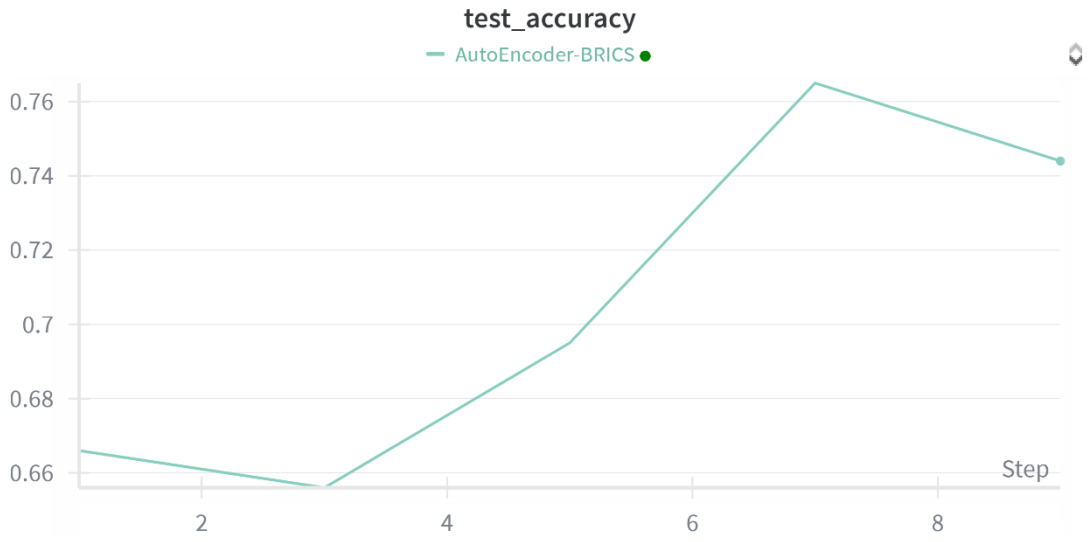


La diferencia de rendimiento entre ambos modelos se explica porque el ViT de Hugging Face aprovecha el conocimiento previo adquirido en datasets masivos mediante preentrenamiento, mientras que nuestro ViT casero debe aprender desde cero con un dataset limitado. Esto lo pone en clara desventaja inicial, aunque ofrece más flexibilidad y comprensión interna de la arquitectura. Además, BRISC no es tan grande. Para un modelo desde cero, necesitarías muchos más datos. Hugging Face ya tiene una

representación general del mundo visual, y solo necesita ajustar los pesos al nuevo problema (**fine-tuning**)

5. **Comparar con Stacked AutoEncoders (SAE):**


test_accuracy
ViT-Casero
0.69
0.68
0.67
0.66
0.65
Step
5
10
15
20


test_accuracy
AutoEncoder-BRICS
0.76
0.74
0.72
0.7
0.68
0.66
Step
2
4
6
8

El modelo ViT aplicado al dataset BRISC demuestra una alta capacidad de detección de tumores, destacándose no solo por su precisión, sino por su capacidad de **explicabilidad visual** gracias a las matrices de atención.

En contraste, los modelos de HuggingFace ofrecen una solución rápida y potente mediante **transfer learning**, aunque con menor control sobre su arquitectura y funcionamiento interno.

Por otro lado, los Stacked AutoEncoders (SAE) funcionan como un **baseline eficiente y liviano**, pero carecen de interpretabilidad clínica, lo que limita su aplicabilidad en entornos sensibles como el diagnóstico médico.

La integración de herramientas como **Weights & Biases (wandb)** permitió un seguimiento detallado del entrenamiento, la comparación visual de modelos y la generación de reportes reproducibles, fundamentales para una investigación rigurosa.

Finalmente, el uso de la **atención como mapa visual de decisión** no solo aumenta la confianza en el modelo, sino que representa un puente entre la inteligencia artificial y la medicina explicable, acercando la tecnología a contextos clínicos reales.