# Informe Técnico: Clasificación de Imágenes de Piel mediante Vision Transformers

Yilian Vázquez Martínez , Pablo Gómez Vidal, Francisco Préstamo Bernárdez

21 de Septiembre de 2024

**Abstract**

Este informe describe un proyecto de clasificación de imágenes de la piel para la detección de enfermedades, implementado utilizando transformers. Adicionalmente, se evaluaron técnicas de *data augmentation*, las cuales no mejoraron el rendimiento significativamente y, en algunos casos, incrementaron significativamente el tiempo de entrenamiento.

# Contents

# 1 Introducción

Las imágenes de la piel son de especial interés para identificar patologías dermatológicas de manera automatizada. En este proyecto, se utilizó la arquitectura *Vision Transformer* (`ViT`) para abordar el problema de clasificación de imágenes de la piel, dada su capacidad para capturar patrones en imágenes mediante el uso de mecanismos de atención.

# 2 Objetivos del Proyecto

El principal objetivo del proyecto es comparar el rendimiento de diferentes variaciones del modelo `ViT` en la clasificación de imágenes de la piel para la detección de enfermedades. Específicamente, se busca:

- Evaluar la precisión de los modelos `vit-base`, `vit-large` y `vit-huge`.
- Determinar el impacto de las técnicas de *data augmentation* en la precisión y el tiempo de entrenamiento.
- Identificar un modelo balanceado en términos de precisión y costo computacional.

# 3 Metodología

La metodología aplicada en este proyecto se dividió en las siguientes etapas:

## 3.1 Preparación de Datos

Las imágenes de la piel fueron preprocesadas mediante una serie de transformaciones, que incluyeron el redimensionado a 224x224 píxeles, conversión a tensores y normalización utilizando los valores de media y desviación estándar de `ImageNet`:

```
transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor(),
    transforms.Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225])
])
```

## 3.2 Modelos de `Vision Transformer`

Se utilizaron tres variaciones del modelo ViT:

- `google/vit-base-patch16-224-in21k`
- `google/vit-large-patch16-224-in21k`
- `google/vit-huge-patch14-224-in21k`

Debido a limitaciones en recursos computacionales, el modelo `vit-huge` no pudo ser evaluado. Las pruebas principales se realizaron con los modelos `vit-base` y `vit-large`.

### 3.3 Entrenamiento del Modelo

El entrenamiento se realizó utilizando un conjunto de datos preprocesado y con el uso de `PyTorch` para implementar las redes. Se utilizaron GPUs cuando estuvieron disponibles, lo que permitió acelerar considerablemente el tiempo de entrenamiento.

### 3.4 Evaluación

La evaluación se realizó mediante la precisión total (%) y el análisis de la matriz de confusión. La métrica clave fue la precisión del modelo en el conjunto de datos de prueba.

## 4 Resultados

### 4.1 Precisión de los Modelos

El rendimiento de los modelos se resume en la Tabla 1:

Table 1: Comparación de la precisión de los modelos ViT

| **Modelo** | **Precisión (%)** | **Tiempo de Entrenamiento** |
|---|---|---|
| `vit-base` | 87.00% | 2h 20min |
| `vit-large` | 89.02% | 3h |

Como se puede observar, el modelo `vit-large` mejoró en un 2.02% con respecto al modelo `vit-base`, aunque a costa de un incremento significativo en el tiempo de entrenamiento.

### 4.2 Impacto de las Técnicas de *Data Augmentation*

Se probaron diversas técnicas de *data augmentation* para mejorar el rendimiento del modelo. Los resultados se muestran en la Tabla 2:

Table 2: Resultados de *Data Augmentation*

| **Técnica** | **Precisión (%)** | **Tiempo de Entrenamiento** |
|---|---|---|
| Sin Augmentation | 89.02% | 3h |
| Rotación y Mirror | 89.14% | 8h 40min |
| Generación Avanzada | 87.96% | 12h |

A pesar de la ligera mejora con rotación y *mirror*, el tiempo de entrenamiento se incrementó drásticamente. Las técnicas avanzadas de generación de imágenes, por su parte, no solo no mejoraron los resultados, sino que redujeron la precisión.

### 4.3 Análisis de la Matriz de Confusión

La matriz de confusión para el modelo `vit-large` se muestra en la Figura 1:

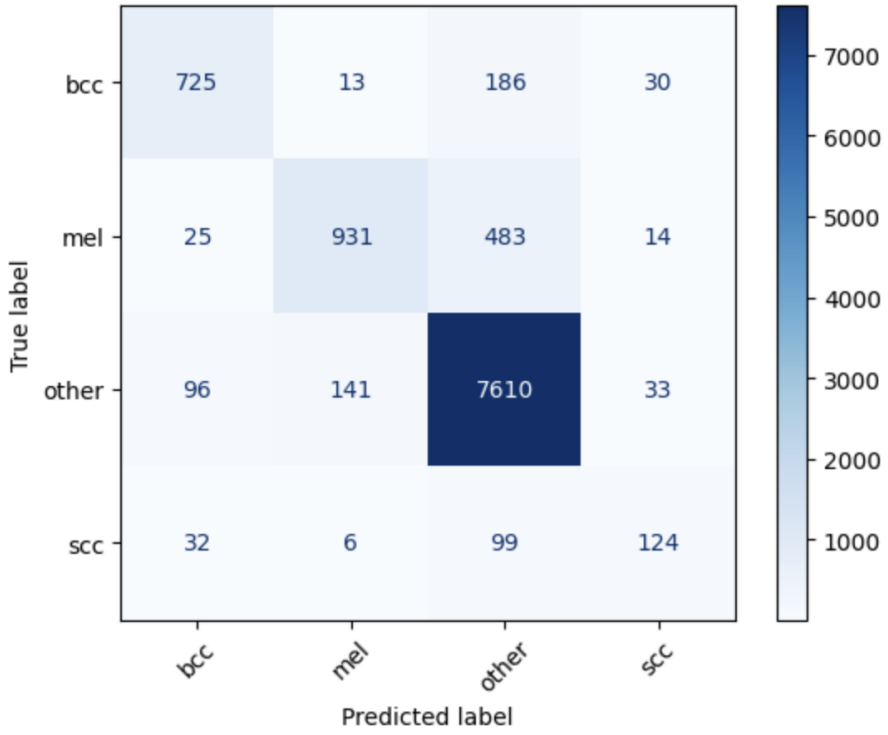


Figure 1: Matriz de confusión para el modelo `vit-large`.

El modelo presentó una alta tasa de aciertos en la mayoría de las clases, con una leve confusión en aquellas condiciones dermatológicas con características visuales similares.

# 5 Discusión

Los resultados muestran que el modelo `vit-large` ofrece una mejora significativa en la precisión respecto al modelo `vit-base`, aunque esto implica un costo computacional mucho mayor. Además, las técnicas de *data augmentation* no lograron mejorar el rendimiento de manera sustancial. Esto sugiere que los transformers, en su configuración actual, ya están extrayendo características robustas de las imágenes sin la necesidad de aumentar los datos.

# 6 Conclusiones

En conclusión, el modelo `vit-large` demostró ser la mejor opción para la tarea de clasificación de imágenes de la piel, alcanzando una precisión del 89.02%. A pesar de los intentos de mejorar el modelo mediante técnicas de *data augmentation*, estas no resultaron efectivas, y en algunos casos incrementaron innecesariamente el tiempo de entrenamiento.

Para futuras investigaciones, se recomienda explorar modelos aún más avanzados o combinar transformers con redes convolucionales para optimizar tanto el tiempo como la precisión en tareas de clasificación médica.

# 7 Referencias


- Dosovitskiy, A., et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021. https://arxiv.org/abs/2010.11929

- Hugging Face, google/vit-large-patch16-224-in21k, https://huggingface.co/google/vit-large-patch16-224-in21k