Modelo para la detección de imágenes generadas por IA

Instituto Tecnológico de Estudios Superiores de Occidente

(ITESO)

Roberto Garrido Hernández
Marcos Antonio Fierros Estrada
Rodrigo Emmanuel Macias Pantoja

29 de Noviembre del 2025

1 Introducción

El rápido avance de los modelos generativos ha incrementado la necesidad de desarrollar sistemas capaces de detectar si una imagen fue creada por inteligencia artificial o capturada del mundo real. La facilidad con la que estas imágenes pueden manipular información visual plantea retos en ámbitos como la verificación digital, la seguridad, el periodismo y la prevención de desinformación.

El presente trabajo tiene como objetivo desarrollar, entrenar y evaluar dos modelos distintos para la detección de imágenes generadas por IA, comparando su desempeño y seleccionando el mejor. El primer enfoque emplea un modelo híbrido que combina representaciones profundas extraídas con una CNN (embeddings CNN) junto con características diseñadas a mano (features handccrafted), las cuales se integran mediante un perceptrón multicapa (MLP). El segundo enfoque emplea un Vision Transformer (ViT) preentrenado, aprovechando técnicas modernas de transfer learning. La evaluación de los modelos se realizará utilizando y analizando métricas estándar de clasificación.

2 Marco Teórico

2.1 Detección de imágenes generadas por IA

La detección de imágenes sintéticas busca identificar patrones sutiles introducidos por modelos generativos. Aunque estas imágenes pueden verse realistas al ojo humano, suelen contener artefactos estadísticos o estructurales que pueden ser detectados mediante técnicas de visión computacional. Los métodos de

detección pueden utilizar características manuales, modelos de aprendizaje profundo o enfoques híbridos.

2.2 Redes Neuronales Convolucionales (CNN)

Una CNN es un tipo de red neuronal diseñada para procesar datos con estructura espacial, como imágenes. Sus capas convolucionales aprenden filtros que detectan bordes, texturas y patrones complejos. Las CNN suelen emplearse como extractores de características: en lugar de entrenarlas desde cero, se utilizan sus capas finales para obtener un embedding o vector que representa el contenido visual (por ejemplo, uno de 512 dimensiones). Este embedding puede luego combinarse con otros descriptores para entrenar un clasificador adicional.

2.3 Handcrafted features

Las características diseñadas a mano consisten en descriptores creados manualmente, basados en intuiciones o propiedades estadísticas de la imagen. Ejemplos comunes incluyen Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), o métricas de ruido. En la detección de contenido sintético, estas características pueden capturar irregularidades que no siempre son evidentes para un modelo profundo por sí solo. Su combinación con embeddings de CNN permite crear modelos híbridos que integran conocimiento explícito con aprendizaje profundo.

A continuación, se describen las características diseñadas que se utilizarán en el modelo híbrido:

- **LBP:**
  - Captura textura local
  - Es robusto a cambios de iluminación

2.4 Perceptrón Multicapa (MLP)

El MLP es una red neuronal totalmente conectada compuesta por capas densas. En este trabajo se usa como clasificador final: recibe la concatenación del embedding de la CNN y el vector de características manuales, y aprende una función que asigna cada imagen a una de dos clases: real o generada por IA. Arquitecturas típicas incluyen capas como $256 \to 64 \to 1$ con activaciones ReLU y salida sigmoidal para clasificación binaria.

2.5 Vision Transformer (ViT)

Los Vision Transformers son una arquitectura basada en transformers, originalmente diseñada para texto, pero adaptada para imágenes dividiéndolas en pequeños parches que funcionan como "tokens". El ViT aprende relaciones globales entre estos parches mediante mecanismos de self-attention.

El modelo seleccionado, google/vit-base-patch16-224-in21k, es un ViT entrenado en un conjunto de más de 14 millones de imágenes (ImageNet-21k). Mediante transfer learning, se ajustan sus pesos para la tarea específica de detección de imágenes generadas por IA, permitiendo obtener resultados robustos incluso con conjuntos de datos limitados.

2.6 Transfer learning

El transfer learning consiste en reutilizar modelos ya entrenados en tareas generales para resolver tareas nuevas con menos datos y menor tiempo de entrenamiento. En visión computacional, esto es común con CNNs o ViTs preentrenados: se aprovecha el conocimiento previo del modelo (capacidad para detectar formas, texturas y estructuras) y solo se ajustan sus capas finales. Esto mejora el desempeño y evita tener que entrenar desde cero.

3 Metodología

ToDo

4 Resultados

ToDo+

5 Conclusiones

ToDo

Referencias

ToDo

Anexos

ToDo