Reporte del proyecto: Detector de memes de odio

Alumno: José Manuel Evangelista
Fecha: 2 de junio de 2025

Introducción

En este reporte presento un experimento sencillo para evaluar si una red neuronal basada en MobileNetV2 puede distinguir memes con discurso de odio únicamente a partir de su información visual. Para ello, se utilizó un subconjunto balanceado de 400 imágenes (200 "hate" y 200 "no_hate") extraídas del desafío *Hateful Memes*. El objetivo es demostrar, en menos de 15 minutos de entrenamiento en GPU, que la transferencia de aprendizaje en visión por computadora es capaz de alcanzar una precisión razonable sin procesar texto.

Metodología

1. **Preparación de datos:**
   - Se descargaron y descomprimieron las imágenes originales (113 MB).
   - Se leyó el archivo `dev_seen.jsonl` para extraer etiquetas, seleccionando al azar 200 ejemplos de cada clase (semilla fija en 2025).
   - Cada imagen se redimensionó a 224×224 píxeles, se normalizó en rango $[-1, 1]$ y se aplicó *data augmentation* ligero:
     - `RandomHorizontalFlip(p=0.5)` para invertir horizontalmente.
     - `ColorJitter(brightness=0.2, contrast=0.2)` para variar brillo y contraste.
   - Se dividió el conjunto en 80 % de entrenamiento y 20 % de validación (batch size = 32).
2. **Modelo:**
   - Base: *MobileNetV2* preentrenada en ImageNet, con todas las capas de `features` congeladas.
   - Cabeza final reemplazada por una capa lineal `Linear(1280, 2)` para clasificar "hate" vs. "no_hate".
   - El modelo se entrenó solo ajustando los parámetros de esta capa final.
3. **Entrenamiento:**
   - Función de pérdida: `CrossEntropyLoss`.
   - Optimizador: `Adam` sobre la cabeza final con tasa de aprendizaje 5×10^{-4}.
   - Número de épocas: 3.
   - Se registraron pérdidas y accuracies tanto en entrenamiento como en validación.

Resultados

- En la última época se obtuvo una *accuracy* de validación aproximada de 0.75 (75 %).
- La gráfica de *accuracy* (entrenamiento vs. validación) muestra que el modelo converge rápidamente; sin embargo, la pérdida de validación se estabiliza y tiende a subir ligeramente, indicando un comienzo de sobreajuste.

- La falta de procesamiento de texto implica que el modelo no reconoce sarcasmos ni juegos de palabras presentes en los memes; por lo tanto, los resultados dependen en gran medida de patrones visuales como tipografías, colores y expresiones faciales.

Conclusiones

- Una red ligera como MobileNetV2 preentrenada, con solo la última capa afinada, demuestra que es posible distinguir memes de odio con un accuracy cercano al 75 % usando exclusivamente señales visuales.
- Se completa en poco tiempo de entrenamiento en Colab con GPU y muestra claramente los pasos de transferencia de aprendizaje, manejo de DataLoaders y visualización de métricas.
- Futuras mejoras incluirían la incorporación de OCR para extraer el texto del meme (por ejemplo, con `pytesseract`), vectorizarlo con un modelo ligero de NLP (DistilBERT o `EmbeddingBag`) y concatenar las características visuales y textuales para entrenar un clasificador multimodal, aumentando así la robustez en presencia de sarcasmos y juegos de palabras.