Evaluación y métricas

En esta parte del proyecto se analizan los resultados obtenidos por los distintos modelos desarrollados para el reconocimiento de emociones faciales. Para la evaluación se ha seguido una división clara del dataset en train, validation y test, utilizando el conjunto de test únicamente al final para obtener una medida realista del rendimiento de cada modelo.

Como métrica principal se ha utilizado la accuracy, aunque también se han tenido en cuenta métricas adicionales como precision, recall, F1-score y la matriz de confusión, ya que el dataset FER-2013 presenta cierto desbalance entre clases y algunas emociones son especialmente difíciles de diferenciar.

Evolución y métricas de la CNN

El desarrollo de la CNN se ha llevado a cabo de forma progresiva. En una primera fase se implementó una CNN sencilla desde cero, que permitió comprobar que el modelo era capaz de aprender patrones relevantes a partir de las imágenes faciales. Sin embargo, los primeros resultados mostraron rápidamente sus limitaciones.

Durante estas primeras pruebas, la accuracy en entrenamiento aumentaba de forma constante, mientras que la accuracy en validación se estabilizaba pronto en valores cercanos al 35–40%. Esto indicaba que el modelo aprendía bien los datos de entrenamiento, pero no conseguía generalizar correctamente, algo habitual en arquitecturas poco profundas aplicadas a un dataset complejo y ruidoso como FER-2013.

Antes de cambiar la arquitectura, se probaron distintas mejoras en el entrenamiento, como aumentar el número de épocas, ajustar el learning rate, introducir regularización mediante weight decay y aplicar técnicas de data augmentation. Estas modificaciones ayudaron a estabilizar el entrenamiento, pero el rendimiento en validación seguía estancándose tras varias épocas.

Para superar este problema, se diseñó una CNN más profunda basada en bloques residuales, inspirada en arquitecturas tipo ResNet. La introducción de conexiones residuales permitió entrenar una red más profunda sin problemas de gradiente y mejorar la capacidad del modelo para aprender representaciones más complejas.

Gracias a esta nueva arquitectura, la CNN mostró una mejora clara en su rendimiento, alcanzando una accuracy final en test de aproximadamente 0.77. Además, el análisis de

la matriz de confusión mostró una reducción de errores entre emociones similares, especialmente entre *Neutral* y *Sad*, lo que indica una mejor capacidad de discriminación.

Métricas del Vision Transformer (ViT)

Una vez obtenidos resultados sólidos con la CNN residual, se implementó un Vision Transformer (ViT) con el objetivo de comparar el rendimiento de modelos convolucionales con arquitecturas basadas en atención. A diferencia de la CNN, el ViT analiza la imagen dividiéndola en parches y utilizando mecanismos de self-attention para captar relaciones globales entre distintas zonas de la cara.

El modelo ViT se entrenó mediante transfer learning, partiendo de un modelo preentrenado y ajustándolo al problema concreto de clasificación de emociones. Para evitar sobreajuste y garantizar un entrenamiento estable, se utilizaron tasas de aprendizaje bajas y técnicas de regularización.

Los resultados obtenidos con el Vision Transformer fueron superiores a los de la CNN, alcanzando una accuracy en test superior al 80%. La matriz de confusión mostró una mejora generalizada en la mayoría de emociones, especialmente en aquellas donde la CNN presentaba mayores confusiones, lo que confirma la ventaja de este tipo de arquitectura para el problema planteado.

Comparativa final de modelos

A continuación se muestra una comparación resumida del rendimiento de los distintos modelos desarrollados:

Modelo	Accuracy (Test)	F1-score Macro
CNN básica	~0.40	~0.38
CNN residual	0.77	~0.74
Vision Transformer	>0.80	~0.78

Conclusión

Los resultados muestran una evolución clara desde modelos más simples hasta arquitecturas más avanzadas. La CNN básica permitió establecer un punto de partida, la CNN residual supuso una mejora importante en capacidad de generalización y el Vision Transformer se consolidó como el modelo con mejor rendimiento global. Esta progresión justifica las decisiones tomadas durante el desarrollo del proyecto y pone de manifiesto la importancia tanto del diseño de la arquitectura como de la estrategia de entrenamiento.