Prácticas: Gender Recognition

Ana Valentina López Chacón

Visión por Computadora
MIARFID, UPV

Mayo, 2025

1. Objetivos

Desarrollar un sistema de reconocimiento de género a partir de imágenes faciales, utilizando el dataset *Labeled Faces in the Wild* (LFW) [1]. Este conjunto de datos contiene imágenes reales de personas en diversas condiciones de pose, expresión y luminosidad, lo que lo convierte en un desafío representativo para sistemas de clasificación robustos. Se dos *benchmarks*, el primero de lograr un 98 % de accuracy en test y el segundo es lograr un 95 % de accuracy en test para una red con menos de 100k de parámetros.

2. Dataset y Preprocesamiento

El conjunto de entrenamiento contiene 10,585 imágenes, mientras que el conjunto de prueba cuenta con 2,648 imágenes. Cada imagen está recortada y centrada en la cara, con una resolución de 100×100 píxeles y 3 canales (RGB). Con el fin de mejorar la generalización, se aplicaron las siguientes técnicas de aumento de datos:

- `Resize` a 64×64 píxeles
- `RandomHorizontalFlip`
- `RandomRotation` con ángulo máximo de 15°
- `ColorJitter` (variación de contraste de hasta 10 %)
- Conversión a tensores con `ToTensor`
- Normalización con media `[0.485, 0.456, 0.406]` y desviación estándar `[0.229, 0.224, 0.225]`

Al conjunto de prueba solo se le aplicó redimensionamiento y normalización, sin aumentos de datos.

3. Configuración de Experimentos

Los siguientes hiperparámetros se mantuvieron constantes en todos los experimentos con arquitecturas convolucionales:

- Optimizador: Adam con tasa de aprendizaje de 0.01
- Regularización L2 (weight decay): 1^{-4}
- Tamaño de lote: 64
- Épocas: 50
- Scheduler: `ReduceLROnPlateau` con factor 0.1, paciencia 10, LR mínimo de 1^{-5}
- Función de pérdida: Entropía cruzada (CrossEntropy)

Todos los modelos se entrenaron con la GPU de Kaggle y se fijaron semillas para garantizar reproducibilidad. Para los experimentos con modelos preentrenados se redujo la tasa de aprendizaje a 10^{-4} y el minimo en `ReduceLROnPlateau` paso a ser de 1×10^{-5}, haciendo el entrenamiento unicamente por 5 épocas con un tamaño de lote de 32 y una imagen de tamaño inicial de 224×224.

4. Bloques Convolucionales

La arquitectura de los bloques convolucionales (ConvBlock) se da de la forma:

- `Conv2D` 3×3, con `padding=1` y `stride=1`.
- `BatchNorm2D`.
- `ReLU`.
- `MaxPool2D` 2×2 y `stride=2`.

Luego de pasar por los bloques convolucionales, las representaciones son aplanadas mediante una capa `Flatten` y alimentadas a una capa totalmente conectada (`Linear`) de salida con dos neuronas, correspondiente a las dos clases (masculino y femenino).

5. Resultados

La siguiente tabla resume los resultados obtenidos con distintas configuraciones de red. Se indica si se utilizó aumento de datos, el número total de parámetros entrenables y la precisión alcanzada en el conjunto de prueba.

Modelo	Arquitectura	DA	Parámetros	Precisión
ConvBlock-100k	3 bloques (16,32,64) + FC	Sí	32,002	**96.37 %**
ConvBlock	4 bloques (16,32,64,128) + FC	No	102,018	96.56 %
ConvBlock2	5 bloques (16,32,64,128) + FC	Sí	102,018	96.71 %
ConvBlock3	5 bloques (16,32,64,128,256) + FC	Sí	395,650	97.05 %
EfficientNet	EfficientNet (preentrenada) + FC	Sí	51,087,748	**98.22 %**

Cuadro 1: Comparación de modelos CNN para reconocimiento de género. DA: Aumento de datos.

Se resaltaron los valores de precisión acordes a los *benchmarks* establecidos en los objetivos. La variación entre modelos de `ConvBlocks` radica exclusivamente en la profundidad de la red y la cantidad de filtros utilizados en cada uno de ellos. Estas decisiones estructurales impactan directamente en la capacidad expresiva del modelo y en el número total de parámetros entrenables.

6. Conclusiones

Para finalizar este informe podemos destacar las siguientes conclusiones:

- A pesar de su simplicidad, los modelos basados en `ConvBlocks` demostraron ser altamente efectivos. Con solo tres bloques y una capa completamente conectada, se alcanzó un 96.37 % de precisión, lo que demuestra que es posible construir modelos livianos y precisos con arquitecturas cuidadosamente diseñadas.
- El uso de técnicas de aumento de datos resultó fundamental para mejorar la capacidad de generalización del modelo, especialmente en configuraciones más profundas donde se evitó el sobreajuste.
- El modelo basado en `EfficientNet` preentrenado alcanzó el mejor desempeño con un 98.22 % de precisión, lo que demuestra la ventaja de transferir conocimiento desde redes previamente entrenadas en grandes conjuntos de datos como ImageNet. Sin embargo, este rendimiento viene acompañado de un costo computacional significativamente mayor.

Referencias

[1] Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). *Labeled faces in the wild: A database forstudying face recognition in unconstrained environments*. In Workshop on faces in'Real-Life'Images: detection, alignment, and recognition