Trabajo 2: CNN

Hecho por Lara Loira, Aitor Eguiguren y Yago de Villoutrey

**Introducción.**

En este proyecto se desarrollará un modelo de clasificación de imágenes de rostros en un problema de clasificación multi-etiqueta. El enfoque principal ha sido implementar una red neuronal convolucional (CNN), logrando así un buen rendimiento en cuanto a la precisión.

Una vez implementados tres modelos, para encontrar cual se adaptaba mejor a las necesidades del proyecto, se ha seleccionado el tercero de ellos como el mejor modelo. Este destaca frente al resto por ser el más sencillo y el más rápido en ejecutar, sin sacrificar el rendimiento.

**Temática y características de la clasificación.**

**¿Qué objetos habéis decidido clasificar? ¿Por qué?**

Inicialmente, el objetivo del proyecto fue identificar a cada una de las celebridades presentes en el dataset. Esta tarea implicaba una clasificación multiclase con múltiples etiquetas correspondientes a las personas del conjunto de datos. Sin embargo, tras los primeros entrenamientos, se observó que el modelo no lograba un rendimiento adecuado, alcanzando un F1-score de solo 0.52 en el conjunto de prueba. Esto se atribuyó a la cantidad limitada de imágenes por celebridad, lo que dificultaba que el modelo pudiera generalizar correctamente las características individuales de cada persona.

Dado este resultado, el enfoque se redirigió hacia dos tareas más manejables, pero igualmente relevantes:

* **Uso de gafas:** Se decidió clasificar las imágenes según si las personas llevaban gafas o no. Este atributo es fácilmente distinguible en la mayoría de las imágenes y no depende tanto de la variación entre individuos, lo que lo convierte en una categoría más accesible para el modelo.
* **Género (mujer o no):** También se decidió clasificar a las personas según su género, específicamente si eran mujeres o no. Aunque esta clasificación enfrenta desafíos relacionados con la ambigüedad y la subjetividad en ciertos casos.

La decisión de cambiar el enfoque hacia estas dos categorías específicas se justificó tanto por la simplicidad relativa de las nuevas tareas como por su potencial en problemas reales de clasificación facial. Este cambio permitió obtener un rendimiento significativamente mejorado, con F1-scores cercanos a 0.965 en el conjunto de prueba, demostrando la capacidad del modelo para capturar patrones relevantes en tareas más específicas.

**¿Qué tipo de clasificación se ha aplicado? ¿Por qué?**

Se ha aplicado una clasificación de tipo multi-etiqueta, en la que cada imagen podía pertenecer simultáneamente a dos categorías independientes: “lleva gafas o no” y “es mujer o no”. Este enfoque fue seleccionado porque ambas etiquetas representan atributos específicos que no dependen entre sí y que pueden ser evaluados de manera paralela en un solo modelo.

La clasificación multi-etiqueta resulta especialmente adecuada en este caso, porque permite al modelo predecir múltiples etiquetas para una misma entrada, utilizando una capa de salida con dos neuronas. Cada una de estas neuronas, mediante una activación sigmoide, produce una probabilidad independiente para cada atributo. Esta estructura no solo simplifica el diseño del modelo, sino que también aprovecha eficientemente los datos disponibles para aprender características que contribuyen simultáneamente a ambas predicciones.

Este tipo de clasificación fue elegido porque los atributos seleccionados, como el uso de gafas y el género, son generalizables y están menos sujetos a las limitaciones del dataset. Esto permitió alcanzar un rendimiento elevado, con un F1-score de 0.965 en el conjunto de prueba, maximizando la precisión y la exhaustividad de las predicciones sin requerir un aumento significativo en la complejidad del modelo.

**¿Se han realizado modificaciones al dataset? Indica cuales y por qué.**

Se realizaron modificaciones al dataset para adaptarlo a los objetivos del proyecto y optimizar el entrenamiento del modelo. Inicialmente, las imágenes estaban organizadas en carpetas según el nombre de las celebridades, lo cual era adecuado para tareas de identificación individual pero no para las etiquetas seleccionadas.

Por ello, las imágenes fueron reorganizadas en nuevas categorías basadas en los atributos de interés:

* **Gafas:** Todas las imágenes de personas que llevaban gafas se agruparon en una carpeta específica, independientemente de otros atributos como género o identidad.
* **Mujeres:** Las imágenes de mujeres se agruparon en otra carpeta, sin considerar si llevaban gafas u otros detalles.
* **Todas:** Se creó una carpeta adicional para contener todas las imágenes del dataset, asegurando que el modelo tuviera acceso a todos los datos sin filtros adicionales.

Este cambio en la organización del dataset permitió simplificar el preprocesamiento y preparar las imágenes para las tareas de clasificación multi-etiqueta. Al agrupar las imágenes por atributos específicos, fue posible entrenar el modelo de forma más eficiente, enfocándolo en aprender patrones relevantes para las etiquetas seleccionadas, y evitando la confusión asociada con intentar clasificar directamente entre nombres de celebridades.

Estas modificaciones resultaron fundamentales para optimizar el rendimiento del modelo y garantizar un entrenamiento más directo y efectivo.

**Decisiones en cuanto a la arquitectura de la CNN.**

**¿Cuántas capas tiene vuestro modelo? ¿Por qué?**

El modelo tiene un total de 12 capas principales, organizadas en bloques de capas convolucionales, de pooling y densas. Esta arquitectura fue diseñada para equilibrar simplicidad y capacidad de aprendizaje, permitiendo al modelo capturar características relevantes del dataset sin sobrecargar los recursos computacionales ni introducir complejidad innecesaria.

Los detalles de las capas son:

* **Capas convolucionales (4)**
  + El modelo incluye cuatro capas convolucionales, con un número creciente de filtros: *neurons* en las dos primeras capas y *neurons\*2* en las dos siguientes.
  + Cada capa convolucional utiliza un tamaño de kernel de (3,3), establecido ya que este tamaño es ideal para capturar características locales importantes como bordes, texturas y detalles faciales.
  + Cada capa convolucional esta seguida por un Batch Normalization, que estabiliza y acelera el proceso de aprendizaje.
* **Capas de pooling (4)**
  + Cada capa convolucional está seguida por una capa MaxPooling2D, que reduce las dimensiones espaciales de las características. Esto no solo disminuye el número de parámetros del modelo, sino que también permite que el modelo se concentre en las características más importantes.
* **Capas densas (4)**
  + Tras aplanar las características extraídas por las capas convolucionales mediante una capa *Flatten*, el modelo incluye tres capas densas con tamaños decrecientes:
    - La primera y segunda capa tienen *neurons\*2* neuronas cada una, permitiendo combinar las características aprendidas en un espacio de representación más compacto.
    - La tercera capa tiene *neurons* neuronas, ayudando a refinar las representaciones antes de la salida.
  + La última capa densa, con dos neuronas y activación sigmoide, genera las probabilidades independientes para las etiquetas “lleva gafas” y “es mujer”.

El modelo fue diseñado para ser lo suficientemente profundo como para aprender los patrones relevantes del dataset, pero sin excederse en complejidad, lo que podría haber resultado en tiempos de entrenamiento más largos y en problemas de sobreajuste.

Esta arquitectura logró un equilibrio óptimo, proporcionando una capacidad suficiente para capturar detalles importantes sin incurrir en un consumo excesivo de recursos.

**¿Qué modificaciones habéis introducido? ¿Por qué?**

El proceso de selección del modelo demostró que las arquitecturas complejas no siempre son las más adecuadas. Durante las pruebas iniciales se diseñaron tres modelos con diferentes configuraciones.

1. El primer modelo, aunque robusto, presentaba tiempos de ejecución elevados debido a su gran cantidad de parámetros y a la utilización de filtros más grandes (5x5, 7x7). Además, mostró un ligero sobreentrenamiento, ya que su alta complejidad lo hacía propenso a memorizar los datos de entrenamiento en lugar de generalizar patrones.
2. El segundo modelo, en contraste, buscaba simplificar la arquitectura reduciendo el número de capas y filtros. Sin embargo, esta simplificación excesiva comprometió su capacidad predictiva, resultando insuficiente para abordar la complejidad del problema y logrando métricas de rendimiento significativamente más bajas.
3. El tercer modelo, finalmente, logró un equilibrio ideal. Se utilizó un diseño con filtros de 3x3, los cuales resultaron ser suficientemente pequeños para capturar detalles relevantes sin aumentar excesivamente los tiempos de computación. Además, se optó por un número moderado de neuronas, comenzando con 16 filtros en las primeras capas y duplicándolos en las posteriores. Esta configuración fue suficiente para extraer patrones relevantes del dataset sin comprometer la velocidad de entrenamiento.

Las pruebas realizadas, terminaron demostrando que configuraciones más avanzadas, como filtros más grandes o un mayor número de neuronas, no aportaban mejoras significativas en las métricas, subrayando la importancia de un diseño ajustado a las necesidades del problema. Por tanto, el modelo final destacó tanto por su precisión (F1-score de 0.965) como por su rapidez, convirtiéndose en la mejor opción para este proyecto.

**Justificación de las métricas, función de pérdida y optimizador.**

Para el entrenamiento y evaluación del modelo, se han seleccionado las métricas, la función de pérdida y el optimizador, adaptándolos a las características del problema y los objetivos del proyecto.

* **Métrica: F1-score**

La métrica principal elegida para evaluar el rendimiento fue el F1-score. Esta métrica es particularmente adecuada en problemas de clasificación multi-etiqueta porque combina la precisión y la exhaustividad en un único valor. Dado que el dataset presentaba un equilibrio variable entre las clases (por ejemplo, más imágenes sin gafas que con gafas), el F1-score permitió evaluar de manera justa cómo el modelo manejaba ambas categorías, evitando priorizar una clase sobre otra.

El F1-score fue calculado de manera independiente para cada etiqueta (“lleva gafas” y “es mujer”) y luego combinado para obtener un promedio general. Esta decisión ayudó a evaluar el equilibrio entre ambas tareas de clasificación.

* **Función de pérdida: Binary Crossentropy**

La función de pérdida seleccionada fue la Binary Crossentropy, una elección estándar para problemas de clasificación multi-etiqueta. Esta función mide la diferencia entre las predicciones del modelo y las etiquetas reales en términos de probabilidad.

Binary Crossentropy es ideal para este proyecto porque las dos etiquetas son independientes entre sí, y esta función optimiza cada salida de manera separada. Esto asegura que el modelo pueda aprender patrones relevantes para cada categoría sin interferencias entre las predicciones.

* **Optimizador: Adam**

El optimizador seleccionado fue Adam (Adaptive Moment Estimation), conocido por su capacidad para adaptarse dinámicamente al gradiente y converger de manera eficiente. Adam combina las ventajas de los métodos de optimización SGD y RMSprop, lo que lo hace especialmente adecuado para modelos de redes neuronales profundas.

Las razones principales para su elección fueron su capacidad para manejar tasas de aprendizaje dinámicas y su eficacia en datasets donde la convergencia puede ser más lenta debido a patrones complejos o desbalanceados.

En conclusión, la combinación de F1-score, Binary Crossentropy y Adam permitió al modelo aprender eficientemente las características necesarias para las dos tareas de clasificación, alcanzando un rendimiento óptimo con un F1-score de 0.936 en el conjunto de prueba.

**Metodología y resultados.**

**Hiperparámetros y selección de éstos.**

El rendimiento del modelo dependió de la cuidadosa selección y ajuste de diversos hiperparámetros, así como de un preprocesamiento efectivo del dataset. Estas decisiones de tomaron para equilibrar precisión, eficiencia y velocidad de entrenamiento. Los aspectos más relevantes son:

1. **Tamaño de los filtros (*k\_size*):**
   * **Valor seleccionado:** 3x3
   * **Justificación:**

Este tamaño permitió capturar características locales importantes, como bordes y texturas, manteniendo un bajo costo computacional. Durante las pruebas, tamaños de filtro mayores (5x5 o 7x7) no ofrecieron mejoras significativas en las métricas y aumentaron los tiempos de entrenamiento.

1. **Número de neuronas por capa convolucional:**
   * **Valor seleccionado:** 16 filtros en las primeras capas y 32 en las posteriores.
   * **Justificación:**

Esta configuración escalonada permitió que las primeras capas aprendieran patrones generales mientras que las posteriores se enfocaran en características más abstractas.

Incrementar este número (por ejemplo, a 64 filtros) aumentó los tiempos de entrenamiento sin una mejora sustancial en el F1-score.

1. **Batch size:**
   * **Valor seleccionado:** 32
   * **Justificación:**

Este tamaño de batch logró un balance adecuado entre eficiencia computacional y estabilidad del gradiente. Batch sizes más grandes redujeron el tiempo por época, pero afectaron la convergencia, mientras que valores menores alargaron el entrenamiento sin mejorar el rendimiento.

1. **Learning rate:**
   * **Valor predeterminado de Adam:** 0.001
   * **Justificación:**

Este valor fue suficiente para asegurar una convergencia rápida y estable al optimizar el modelo con Adam, sin requerir ajustes adicionales.

1. **Data augmentation:**
   * **Configuración utilizada:**
     + **Rotación:** 10º
     + **Zoom:** 15%
     + **Desplazamientos horizontales y verticales:** 20%
     + **Flip horizontal:** Activado
   * **Justificación:**

Estas técnicas incrementaron la variabilidad del dataset de entrenamiento, ayudando al modelo a generalizar mejor y a reducir el riesgo de sobreajuste. Los valores se mantuvieron en rangos moderados para evitar introducir distorsiones excesivas.

1. **Número de épocas:**
   * **Valor ajustado dinámicamente:** Early stopping con paciencia de 5 épocas.
   * **Justificación:**

Early stopping evitó el sobreentrenamiento restaurando los mejores pesos alcanzados cuando la pérdida en validación no mejoraba después de 5 épocas consecutivas.

1. **Eliminación de imágenes duplicadas:**

Durante la preparación del dataset, se identificó la presencia de imágenes duplicadas que podían afectar negativamente al entrenamiento. Para abordar este problema, se implementó un método basado en hashes únicos. Cada imagen fue convertida a un array y, posteriormente, se calculó su hash, lo que permitió identificar duplicados con rapidez y precisión.

Una vez identificadas las imágenes duplicadas, se eliminaron, asegurando que cada entrada del dataset fuera única. Esta limpieza no solo redujo el tamaño del dataset sin pérdida de información relevante, sino que también previno que el modelo se sesgara hacia patrones redundantes. Como resultado, se mejoró la calidad general del entrenamiento y la evaluación, garantizando que los resultados reflejaran la capacidad del modelo para aprender patrones generalizables en lugar de detalles repetitivos.

Estas decisiones en la selección de hiperparámetros y el preprocesamiento del dataset han sido claves a la hora de conseguir alcanzar un rendimiento óptimo, logrando así un buen F1-score y manteniendo un entrenamiento eficiente.

**Resultado de las métricas seleccionadas.**

El modelo fue evaluado utilizando el F1-score como la métrica principal, dado que se trataba de un problema de clasificación multi-etiqueta. Esta métrica combina tanto la precisión como la exhaustividad, siendo especialmente útil cuando se busca un equilibrio entre ambas, como en el caso propuesto.

Los resultados que se han obtenido son:

* **F1-score en el conjunto de prueba:** 0.965

El modelo alcanzó un F1-score muy alto, lo que indica que es capaz de realizar predicciones precisas y completas en ambas categorías (“lleva gafas”, “es mujer”). Esto sugiere que el modelo ha aprendido correctamente las características diferenciadoras para cada tarea sin sobreajustarse.

* **Precisión en el conjunto de prueba:** 95%

La precisión indica que el modelo predijo correctamente el 93% de las etiquetas positivas.

* **Exhaustividad en el conjunto de prueba:** 94%

La exhaustividad muestra que el modelo identificó correctamente el 94% de las instancias positivas de cada categoría.

Los resultados del F1-score y las métricas asociadas demuestran que el modelo tiene un rendimiento robusto, equilibrando adecuadamente precisión y exhaustividad, lo que lo hace adecuado para el propósito de clasificar si una persona lleva gafas y si es mujer, con una alta tasa de aciertos en ambas categorías.

**Gráficas de resultados.**