Instituto Tecnológico y de Estudios Superiores de Monterrey

Reporte : Red convolucional para la identificación de personas famosas

Desarrollo de aplicaciones avanzadas de ciencias computacionales (Gpo 301)

TC3003B.201

Dirige:

Benjamín Valdés Aguirre

Presenta:

Carlos Adrián García Estrada -A01707503

26 de mayo del 2025
Querétaro, Querétaro

Introducción

Las redes convolucionales son una técnica de aprendizaje supervisado dentro del campo de Machine learning, más concretamente deep learning. Está categorizada dentro de la rama de deep learning ya que requiere de muchas capas de aprendizaje.

Una red neuronal está compuesta por capas las cuales a su vez, simulando su contraparte antropomórfica contiene neuronas. Cada capa modifica nuestro ejemplo de entrada y utiliza los patrones que estas funciones generan para "activar" las neuronas de la capa siguiente por medio de filtros convolucionales, esto con el objetivo de etiquetar cierto patrón de activación de neuronas con un identificador y poder predecir ejemplos similares. Respecto a las imágenes, los filtros convolucionales procesan cada pixel al tomar el valor de los píxeles colindantes y su suma es expresada en un solo píxel, destacando ciertas características de la imágen como por ejemplo bordes horizontales o verticales [7]. Cada capa permite distinguir entre distintas características y entre más profunda la red, estas son más detalladas y específicas.[3]

Para realizar esto, la red neuronal requiere de ajuste, como por ejemplo cambiando los pesos que ciertas neuronas tienen sobre las siguientes neuronas, incrementando o reduciendo la función de activación de cierta neurona, etc. Tomando en cuenta esto, la red neuronal constantemente toma la diferencia entre la etiqueta esperada y el resultado generado. Y realiza estas modificaciones para acercarse lo más posible al resultado esperado.

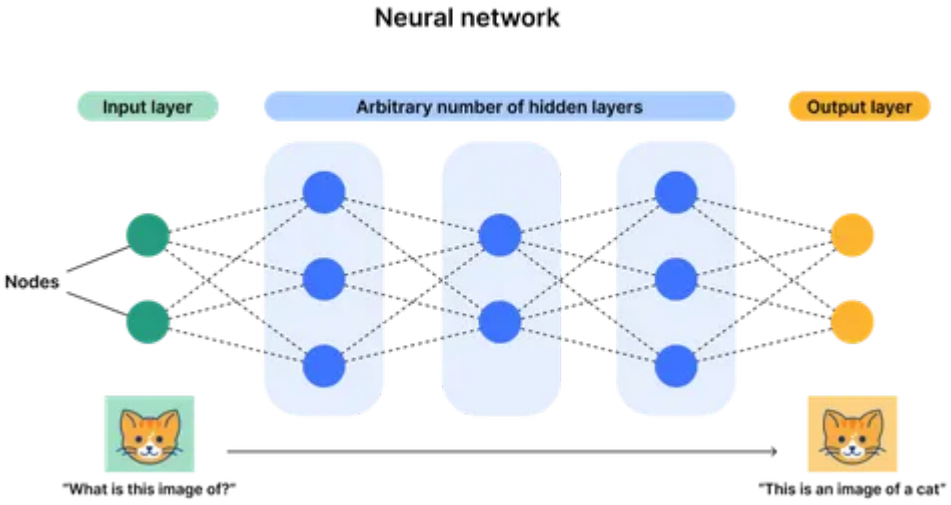


Fig 1. Ejemplo de Red Neuronal

Las redes neuronales pueden utilizarse para calcular y predecir virtualmente cualquier problemática. Sin embargo, una de las principales áreas de trabajo es tener el dataset correcto. Un dataset con pocos ejemplos no permitirá a la red neuronal identificar los patrones distintos entre nuestros clasificadores (underfitting), mientras que un dataset con mucho

ejemplos pero poca variabilidad resultará en que la red neuronal sea capaz de solo distinguir dentro de los ejemplos de entrenamiento (*overfitting)*

En esta ocasión se hará un modelo de red neuronal para distinguir entre los cuatro miembros de la banda británica más famosa de todos los tiempos. The Beatles.

El reconocimiento facial es una área de estudio de gran utilidad como en áreas de seguridad y videovigilancia. Sin embargo tiene varias capas de complejidad, en el análisis inicial se consideraron distintas situaciones que elevan la complejidad de entrenar una red neuronal desde cero para identificar y distinguir caras. Como lo es la diferencia de estilo de cabello, vello facial, accesorios y ropa a lo largo del tiempo, cambios físicos relacionados a la edad, peso u otros. Además de los problemas naturales del procesamiento de imágenes como diferencias en iluminación, ángulos, calidad y tamaño de las fotografías.

Implementación del modelo

Preprocesado

Siguiendo la temática de un correcto dataset, más allá de su obtención es necesario preprocesar los datos. Para la obtención del dataset, se utilizó un script que tomaba el query de google y guardaba los resultados en las carpetas. En promedio se obtuvieron alrededor de 600 imágenes por Beatle. Sin embargo, los querys utilizados resultaban muy generales por lo que se optó por mejorar la calidad del dataset por medio de selección.

El criterio de selección era el siguiente:

- El Beatle debe aparecer solo
- No debe ser obstruido por un objeto
- Debe representar distintas épocas de la vida del Beatle, desde Adulto Joven - Vejez
- La imágen no debe estar severamente editada
- La imágen debe ser real, no IA o ilustración
- No debe haber añadidos prominentes como títulos u otras imágenes superpuestas

Después del criterio de selección resultaron aproximadamente 340 imágenes por Beatle.

Un estándar al trabajar con imágenes es la normalización de los canales de RGB. La normalización permite trabajar con escalas manejables para el procesamiento como en las funciones de inicialización y optimización. Para cada imágen se forzó un tamaño de 224x224 y se dividieron los valores de RGB / 255. Las imágenes fueron divididas en 80% train y 20% test.

El siguiente paso era generar un *one hot encoding*. Esta técnica modifica nuestras etiquetas string [John, Paul, George, Ringo] y les asigna un valor único dentro de un vector del tamaño de nuestras etiquetas. Esta normalización permite que los datos de entrada y salida se encuentren en formato consistente (únicamente números) y la red neuronal puede realizar

cálculos de minimización de gradiente. La minimización del gradiente es lo que entendemos por "aprender", para la red neuronal significa sólo reducir la diferencia entre la etiqueta esperada y la predicción generada.

Finalmente todas las imágenes fueron reducidas a un solo canal RGB. Esto porque los beatles vivieron su fama en la transición de las fotos y video de blanco y negro a color, esto aunado con que John y George fallecieron antes mientras que Paul y Ringo siguen vivos, es posible que los primeros tengan una mayor cantidad de fotos en blanco y negro que a color y viceversa. Eliminando un posible sesgo

Fig 2. Imágenes en un sólo canal RGB

En la primera iteración se realizó una versión simplificada de la arquitectura VGG [3]. Esta arquitectura puede tener de dieciséis a diecinueve capas. Toma como parámetros una imágen RGB de 224 x 224. Para cada capa se tiene grupos de filtros 3x3, de manera incremental. Así mismo se tiene una capa de Max Pooling, esta capa reduce la dimensionalidad de la imagen y toman solo aquellos detalles más importantes. Al final se tiene 3 capas densas completamente conectadas, y puesto que está diseñada para problemas categóricos, una activación softmax. Se eligió ya que esta arquitectura fue revolucionaria al superar otras arquitecturas existentes como GoogleLeNet, MSRA o Clarifai.

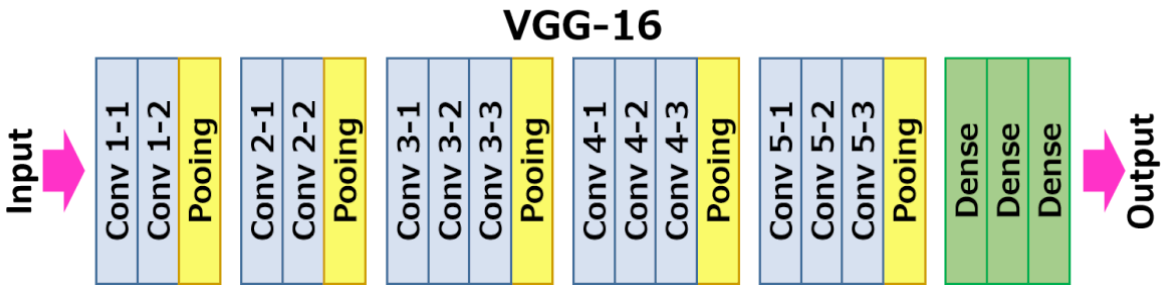
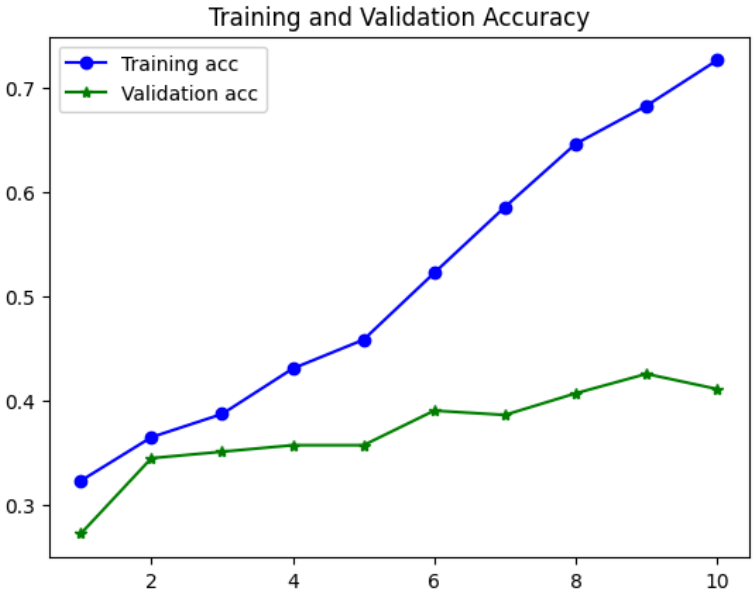
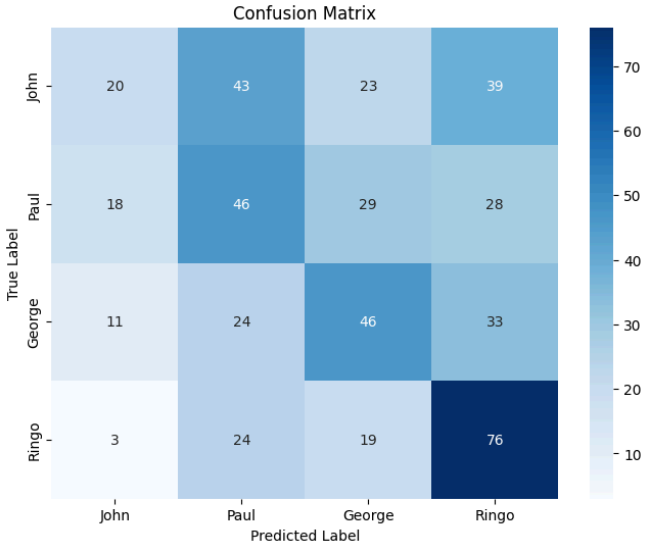


Fig 3. Arquitectura VGG-16

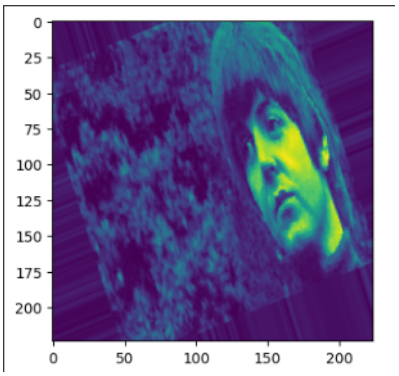
Para esta primera iteración, se tuvieron resultados prometedores donde se alcanzaba una accuracy de hasta 72%, sin embargo revisando la validation accuracy esta se encontraba en los 40% una clara señal de overfitting. El modelo actúa de manera correcta los datos de entrenamiento, sin embargo al ver nuevos datos los patrones aprendidos no son suficientemente generalizados para realizar una predicción correcta.





Se intuyó que no había suficientes ejemplos por lo que introdujo variabilidad al dataset con aumentación de imágenes. La aumentación de imágenes nos permite introducir una mayor cantidad de datos utilizando el dataset ya obtenido. De esta manera podemos resolver problemas como deficiencia en la calidad y cantidad de datos. Para nuestro problema sobre reconocimiento de imágenes se realizaron transformaciones a las imágenes como rotación, estiramiento y girar. El objetivo es simular imágenes con diferente ángulo y nivel de zoom,

suficientemente diferente para que la red neuronal pueda aprender patrones generales faciales más no memorizar.



Imágen artificialmente aumentada de Paul Mcartney

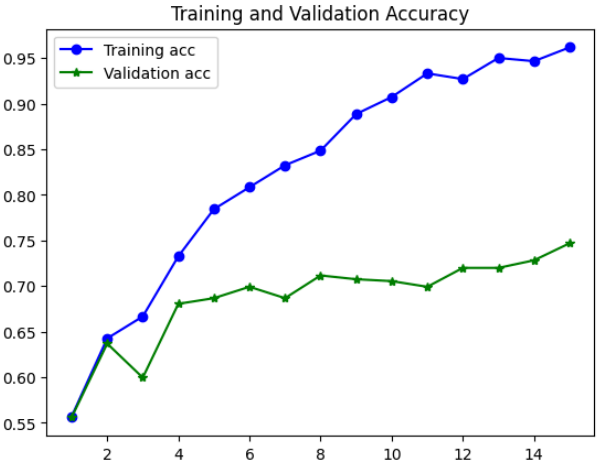
Con train generator se alcanzó una accuracy tanto en validation como en train de 28%, siendo cuatro clases básicamente el modelo parece estar ahora "adivinando" y no realmente aprendiendo, sin embargo se ha evitado el overfitting.

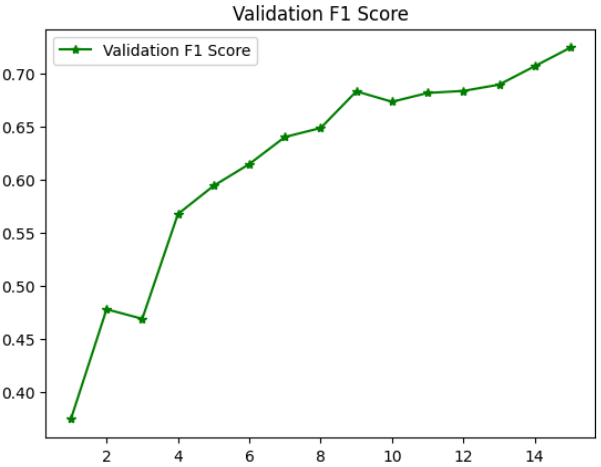
Modelo VGG16

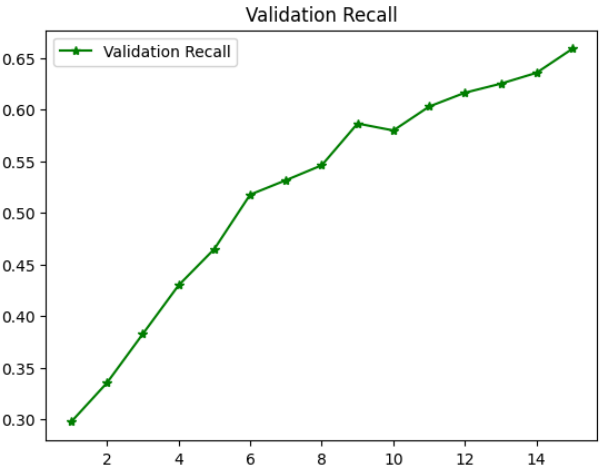
Para un segunda iteración se importar el modelo VGG 16 completo desde Keras, Después de la base se aplana todas las capas en un vector, se añadió una capa densa de 256 y un dropout de neuronas del 0.5, es decir la mitad de las neuronas fueron eliminadas para prevenir overfitting.

La arquitectura VGG es soportada por ImageNet[1]. Es una enorme base de datos, el propósito es generar un cúmulo de imágenes similar a google pero específicamente para el reconocimiento de imágenes con calidad y propiamente etiquetadas. Las imágenes, que son más de 3.2 millones, están etiquetadas en más de 5 mil clases, doce subárboles y aumentando. ImageNet cuenta con imágenes de personas no específicas y se ha utilizado anteriormente para reconocimiento facial, especialmente en la ofuscación de imágenes, donde se alcanzó un accuracy con arquitectura VGG del 90% [6]

Para esta iteración se revirtieron las imágenes a tres canales RGB ya que ese es el input esperado de la arquitectura.







La compilación de este modelo como era de esperarse, trajo resultados mucho mejores. Se alcanzó una accuracy de alrededor 90% , que es cierta mejora con el modelo anterior sin embargo, para validation accuracy se alcanzó un porcentaje de 70%, aunque una mejora considerable la discrepancia de precisión indica de nuevo overfitting. Pero comparándolo con los resultados anteriores, podemos ver que aún estamos por debajo del objetivo esperado. Una mejor medida de la efectividad son las métricas de F1 y Recall. Recall indica cuántas instancias de verdaderos positivos el módelo identificó correctamente, mientras que FI es la medida promedio de Recall y precisión, obteniendo una medida relativamente baja (70%) aún

incluso con un modelo pre entrenado está claro entonces que nuestro problema es con la calidad de nuestro dataset.

Tercera iteración

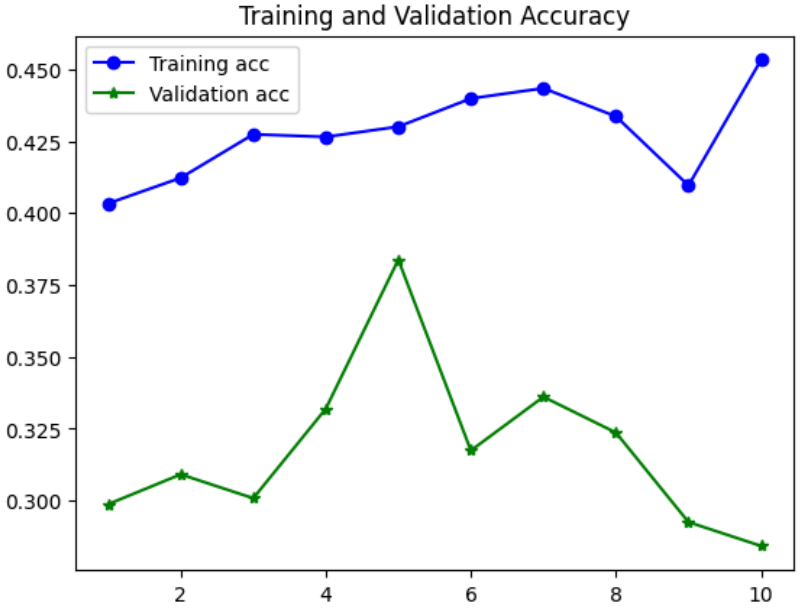
En nuestra última y tercera iteración, se obtuvieron una mayor cantidad de imágenes, alcanzando más de 400 imágenes por Beatle, esta ocasión las imágenes fueron seleccionadas manualmente para el aseguramiento de la calidad de las mismas.

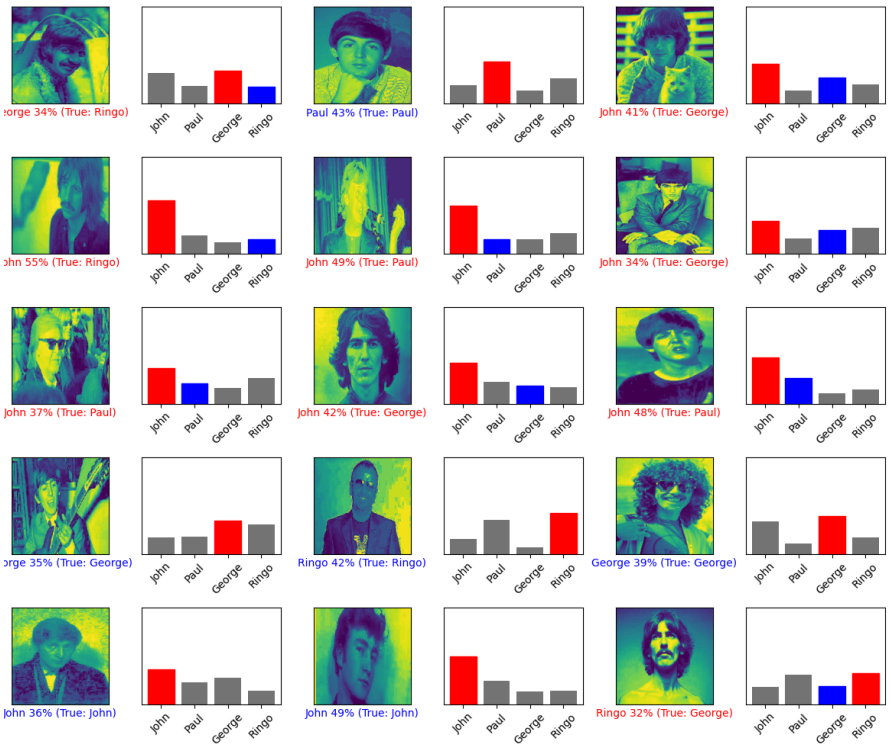
La arquitectura final es la siguiente:

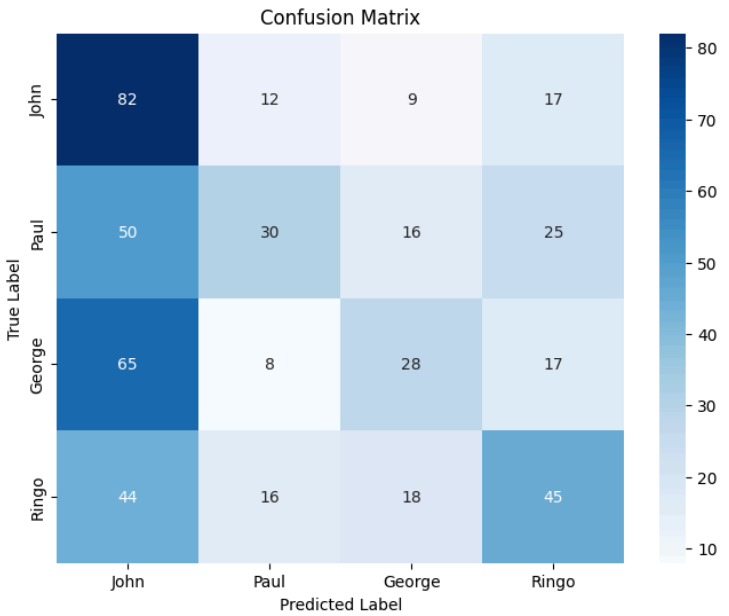
Layer (type)	Output Shape	Param #
conv2d_60 (Conv2D)	(None, 224, 224, 64)	640
batch_normalization_60 (BatchNormalization)	(None, 224, 224, 64)	256
conv2d_61 (Conv2D)	(None, 224, 224, 64)	36,928
batch_normalization_61 (BatchNormalization)	(None, 224, 224, 64)	256
max_pooling2d_30 (MaxPooling2D)	(None, 112, 112, 64)	0
dropout_34 (Dropout)	(None, 112, 112, 64)	0
conv2d_62 (Conv2D)	(None, 112, 112, 128)	73,856
batch_normalization_62 (BatchNormalization)	(None, 112, 112, 128)	512
conv2d_63 (Conv2D)	(None, 112, 112, 128)	147,584
batch_normalization_63 (BatchNormalization)	(None, 112, 112, 128)	512
max_pooling2d_31 (MaxPooling2D)	(None, 56, 56, 128)	0
dropout_35 (Dropout)	(None, 56, 56, 128)	0
conv2d_64 (Conv2D)	(None, 56, 56, 256)	295,168
batch_normalization_64 (BatchNormalization)	(None, 56, 56, 256)	1,024
conv2d_65 (Conv2D)	(None, 56, 56, 256)	590,080
batch_normalization_65 (BatchNormalization)	(None, 56, 56, 256)	1,024
max_pooling2d_32 (MaxPooling2D)	(None, 28, 28, 256)	0
dropout_36 (Dropout)	(None, 28, 28, 256)	0
global_average_pooling2d_8 (GlobalAveragePooling2D)	(None, 256)	0
dense_20 (Dense)	(None, 256)	65,792
dropout_37 (Dropout)	(None, 256)	0
dense_21 (Dense)	(None, 4)	1,028

Una mejora evidente no solo es la cantidad de capas, que permitirían captura mucho más detalle, sino también la implementación de Batch Normalization, al utilizar esta técnica directamente después de una capa convolucional se reduce el tiempo de computación e incrementa la cantidad de memoria utilizada al normalizar los inputs de cada capa para una distribución constante. Así mismo, se incrementó el contraste y se redujo la agresividad de la

aumentación de imágenes. Aún con estas mejoras, los resultados comparados con la primera no mejoraron significativamente, con tan solo un incremento de 10% tanto en accuracy como validation accuracy.







Conclusión

El reconocimiento facial es un problema complejo dentro del reconocimiento de imágenes. Para alcanzar una alta precisión en la identificación es necesario no solo con una alta cantidad de imágenes sino con una calidad consistente entre ellas. Existe una relación lineal entre la cantidad de imágenes y la precisión pronósticada. Aunque se considera que de 150 a 500 imágenes son suficientes para realizar modelados en áreas específicas [2]. Sin embargo, para un problema como reconocimiento facial donde la complejidad es mucho mayor, esta cantidad no es suficiente.

Sin embargo, un factor que se dejó por alto fue la gran variedad en la calidad de las imágenes y la cercanía a la cara. En nuestro dataset, había una alta variabilidad respecto a este factor y debió ser necesario hacer un recorte de las imágenes para la mayor cantidad de información de la cara posible. Además, la cantidad de capas de nuestro modelo no son la suficientes para poder aprender detalles más específicos, una característica esencial para la distinción facial. Por esta misma razón es que el modelo tuvo una mucha más alta mejora en la implementación de VGG. La implementación de ImageNet está pre entrenada para reconocer rostros, por lo que la distinción entre ellos es mucho más fácil.

La resultados no satisfactorios del problema viene del hecho que nuestro modelo tiene que distinguir principalmente no sólo el cuerpo / cara del fondo, que en sí tiene un cierto nivel de complejidad, sino también aprender las diferentes características únicas de cuatro individuos Aunado con que nuestros cuatro sujetos compartieron una alta variedad de estilos a lo largo de su carrera y diferencias en la calidad de sus fotos debido a la extensión de vida, la complejidad requiere de modelos pre entrenados de reconocimiento facial y una mucha más alta cantidad de imágenes de ejemplo para tener una mucho más alta precisión.

Referencias:

1. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.Cite Bay+6BibSonomy+6BibSonomy+6

2. M. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, Y. Yang, and Y. Zhou, "How many images do I need? Understanding how sample size per class affects deep learning model performance," *Pattern Recognition Letters*, vol. 133, pp. 98–104, 2020, doi: 10.1016/j.patrec.2020.03.012.ScienceDirect

3. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 2015.

4. M. Coskun, A. Uçar, and Y. Demir, "Face recognition based on convolutional neural network," in *Proc. Int. Conf. Comput. Sci. Eng. (UBMK)*, Antalya, Turkey, Oct. 2017, pp. 719–723, doi: 10.1109/UBMK.2017.8093521.Semantic Scholar

5. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, Nov. 2022, doi: 10.1016/j.array.2022.100258.ScienceDirect+5arXiv+5SCIRP+5

6. K. Yang, J. H. Yau, L. Fei-Fei, J. Deng, and O. Russakovsky, "A Study of Face Obfuscation in ImageNet," in *Proc. 39th Int. Conf. Machine Learning (ICML)*, Baltimore, MD, USA, Jul. 2022, vol. 162, pp. 25313–25330.

7. R. Keshari, M. Vatsa, R. Singh, and A. Noore, "Learning Structure and Strength of CNN Filters for Small Sample Size Training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 9349–9358, doi: 10.1109/CVPR.2018.00974.