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

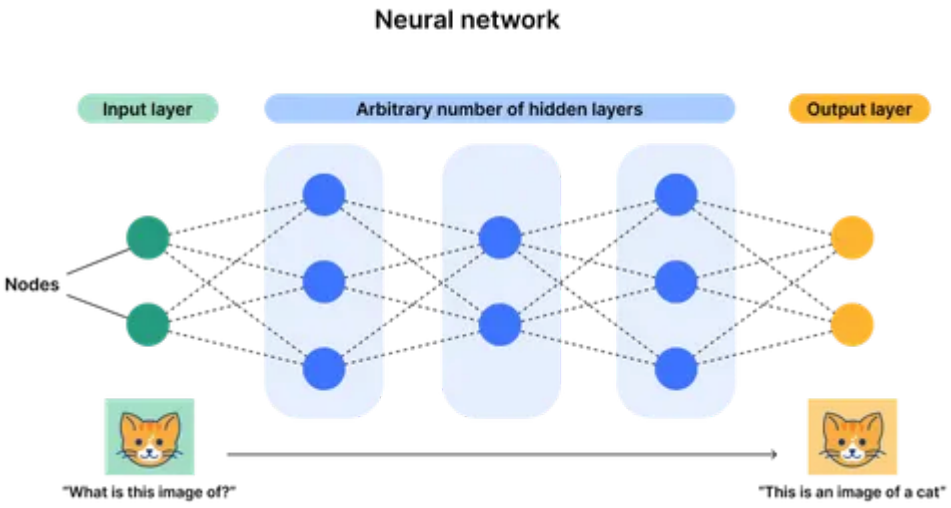
26 de mayo del 2025
Querétaro, Querétaro

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, esto con el objetivo de etiquetar cierto patrón de activación de neuronas con una etiqueta y poder predecir ejemplos similares.

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.



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.

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 eliminar imágenes.

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.

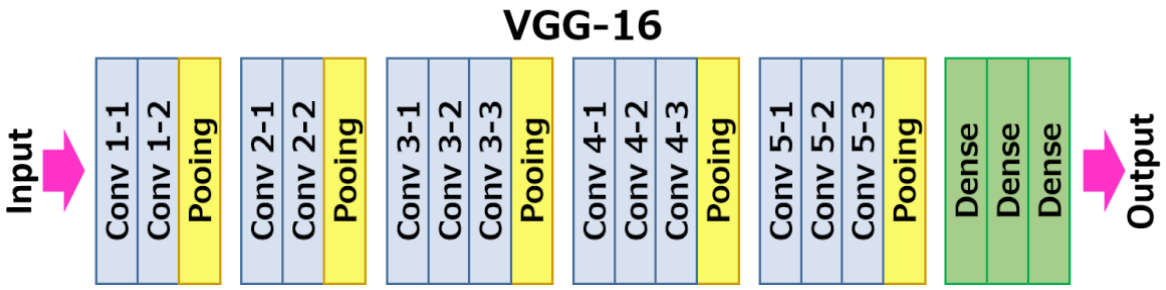
Un estándar al trabajar con imágenes es la normalización de los canales de RGB. La normalización permite trabajar con escalas manejables para el procesamiento como en las funciones de inicialización y optimización. Para cada imágen se forzó un tamaño de 224x224 y se dividieron los valores de RGB / 255.

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.

Imágenes en un sólo canal RGB

En nuestra primera iteración se realizó una versión simplificada de la arquitectura VGG 16. (descripción de VGG16)

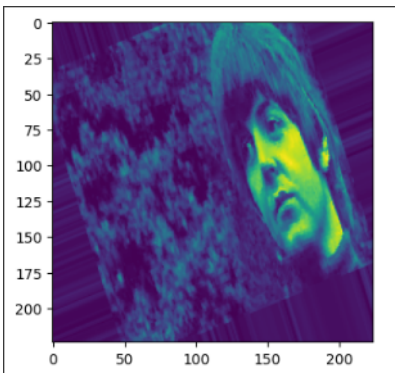


Arquitectura VGG-16

Para esta primera iteración, se tuvieron resultados prometedores donde se alcanzaba una accuracy de hasta 80%, 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.

(Matriz de confusion y otras estadísticas)

Se 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 y realizando 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 más no memorizar.



Imágen artificialmente aumentada de Paul Mcarntney

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,

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 85% , no mucha mejora con el modelo anterior sin embargo, para validation accuracy se alcanzó un porcentaje de 80%, aunque una mejora considerable la discrepancia de precisión indica de nuevo overfitting. El uso de un modelo pre entrenado de reconocimiento esperaría resultados casi perfectos, alrededor del 92.75% (referencia). Está claro entonces que nuestro problema es con la calidad de nuestro dataset.

Tercera iteración

Análisis

Referencias:

https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f
https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-3-flattening
https://pyimagesearch.com/2021/05/14/convolutional-neural-networks-cnns-and-layer-types/