Instituto Tecnológico y de Estudios Superiores de Monterrey

Evidencia: Avance: Generación o selección del set de datos y Preprocesado de los datos

Desarrollo de aplicaciones avanzadas de ciencias computacionales (Gpo 301)

TC3003B.301

Presenta:

Carlos Adrián García Estrada -A01707503

18 de mayo del 2025

Querétaro, Querétaro

Obtención del dataset

Mi modelo de aprendizaje será respecto a la banda británica más famosa, The Beatles. Formada en los 60´s fue integrada por cuatro miembros principales: John Lennon, Paul Mccartney, George Harrison y Ringo Starr. El objetivo del modelo es distinguir entre los distintos miembros con la precisión más alta posible.

El dataset fue generado por medio de un script (con autorización) de mi compañero Osvaldo. Este script toma como input un query web y obtiene las imágenes jpeg y las guarda en una dirección de carpeta personalizada. Para que el script funcione es necesario añadir *chromedriver* para windows.

Al utilizar el script con el query del nombre de cada integrante se obtuvo en promedio 600 imágenes por integrantes.

George - 657
John - 621
Paul - 654
Ringo - 672

Preprocesado de datos

Una vez obtenida las imágenes, se realizó una limpieza del data set. Puesto que nuestro query es bastante general, este devuelve todo lo relacionado con el input. Por lo que se tuvo que purgar imágenes donde nuestro integrante no se encontrara solo (bastante común), imágenes severamente editadas como covers de álbumes, fanart, etc. Videos de youtube con títulos, y en ocasiones fotos de la infancia.

Las imágenes restantes tienen alta variabilidad en iluminación, cercanía, colores y épocas de la vida del integrante. Después de la limpieza para cada Beatle quedaron.

George 162
John 158
Paul 161
Ringo 163

Dentro del código, se siguió las recomendaciones dentro de clase el preprocesado, todas las imágenes fueron escaladas en sus valores RGB de 0 a 1. Se asignó un tamaño estándar de 160x160. Seguidamente se realizó un split 80 - 20 con la librería sklearn utils.

Código:

```
#IMPORTS
import os
import numpy as np
from tensorflow.keras.preprocessing.image import load_img, img_to_array
from sklearn.utils import shuffle
from google.colab import drive
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split

drive.mount('/content/drive')
%cd "/content/drive/MyDrive/Beatles"
!ls

#LOAD IMAGES
john_folder = 'john'
paul_folder = 'paul'
george_folder = 'george'
ringo_folder = 'ringo'
image_size = (160, 160)

John_images = []
Paul_images = []
George_images = []
Ringo_images = []


def preprocess(folder_name):
  temp = []
  for filename in os.listdir(folder_name):
      if filename.lower().endswith(('.png', '.jpg', '.jpeg')):
          img_path = os.path.join(folder_name, filename)
          img = load_img(img_path, target_size=image_size)
          img_array = img_to_array(img) / 255
          temp.append(img_array)

  temp = np.array(temp)
  return shuffle(temp)

John_images= preprocess(john_folder)
Paul_images = preprocess(paul_folder)
George_images = preprocess(george_folder)
```

```
Ringo_images = preprocess(ringo_folder)
```

```
#SPLIT 80 - 20
ringo_train, ringo_test, = train_test_split(Ringo_images
,test_size=0.20, random_state=42)
george_train, george_test, = train_test_split(George_images
,test_size=0.20, random_state=42)
paul_train, paul_test, = train_test_split(Paul_images ,test_size=0.20,
random_state=42)
john_train, john_test, = train_test_split(John_images ,test_size=0.20,
random_state=42)

f, axarr = plt.subplots(4, 5, figsize=(10, 10))

for i in range(5) :
  axarr[0][i].imshow(paul_train[i])
  axarr[1][i].imshow(ringo_train[i])
  axarr[2][i].imshow(john_train[i])
  axarr[3][i].imshow(george_train[i])
  axarr[0][i].axis('off')
  axarr[1][i].axis('off')
  axarr[2][i].axis('off')
  axarr[3][i].axis('off')
plt.show()
```