Milestone 2 AC209B

Allison Yao, Isabella Bossa, Elie Attias, Isidora Diaz, Luis Henrique Simplicio Ribeiro

March 2023

1 Exploratory Data Analysis and Description

CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter. CelebA has large diversities, large quantities, and rich annotations.

1.1 Dataset

- Number of celebrities: 10,177
- Number of images: 202,599

1.2 Data Files

- img_align_celeba.zip: All the face images, cropped and aligned.
- list_eval_partition.csv: Recommended partitioning of images into training, validation, testing sets. Images 1-162770 are training, 162771-182637 are validation, 182638-202599 are testing.
- list_bbox_celeba.csv: Bounding box information for each image. "x_1" and "y_1" represent the upper left point coordinate of bounding box. "width" and "height" represent the width and height of bounding box.
- list_landmarks_align_celeba.csv: Image landmarks and their respective coordinates. There are 5 landmarks: left eye, right eye, nose, left mouth, right mouth.

2 Potential Data Issues

- One of the biggest issues was that the dataset did not contain the names of celebrities, so they could not be identified.
- We also observed class imbalances in the dataset:
  - Some celebrities were present in as many as 35 images, while others only had 1 picture that included them.
  - Slightly over 58% of the observations belong to female celebrities, while the remaining 42% are from male celebrities. Although there is a slight class imbalance here, the proportions are relatively close to those of the general population and the difference between the two ratios is not too large, so we believe this will not be a problem.
  - Some attributes are more present than others. For example, less than 5% of the images contain bald celebrities. This potential domain shift would only be an issue if we were planning to use the model with a subset of celebrities that had a different distribution of the target attribute.
  - We will use contrastive learning to train the model, which implies to use positive/negative pairs of images, a pair composed of two different pictures of the same celebrity would be a positive pair, and a pair of two pictures of two different celebrities would be a negative. Some celebrities have a single image on the dataset.

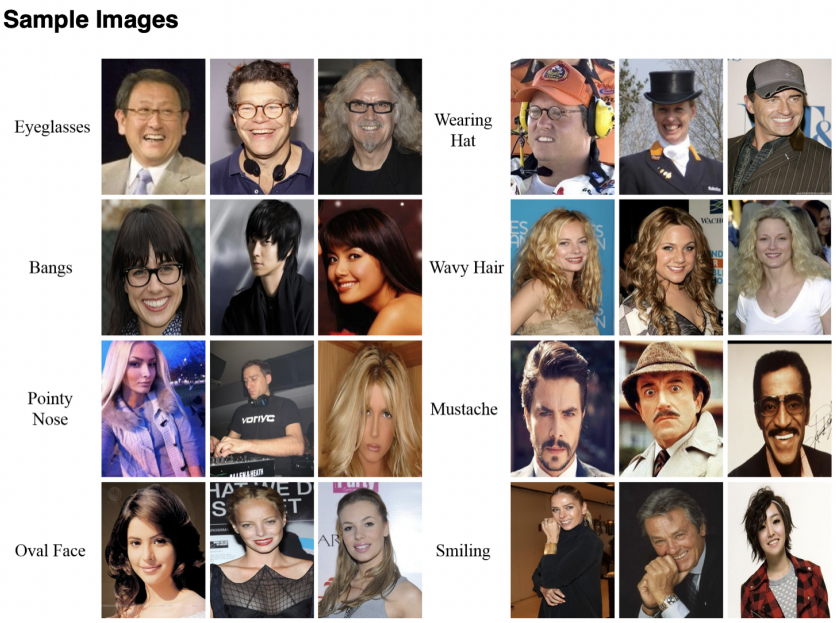

Figure 1: Sample Images from the CelebA data sets.

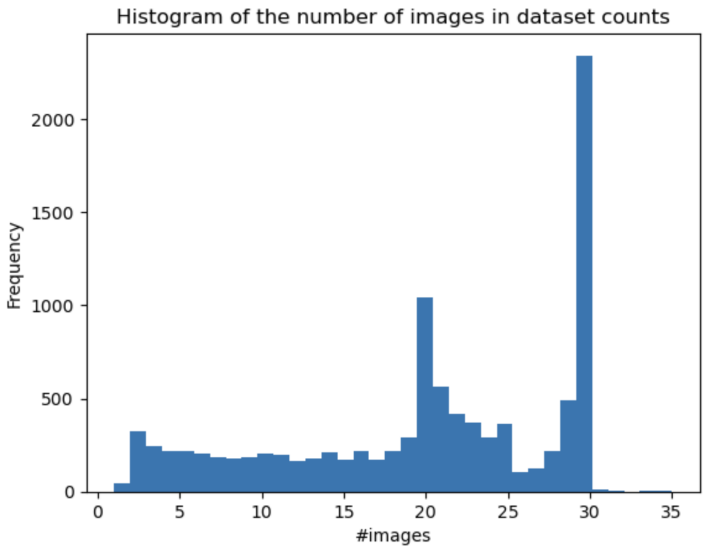

Figure 2: Histogram of image counts per celebrity

3 Addressing the Issues

- After contacting the authors of the celebA dataset, we were able to successfully retrieve the face identities for each image.
- To deal with class imbalance, we plan on using the following strategies:
  - There are 44 celebrities that have only 1 image in the dataset. We plan on including more data for these celebrities by searching them on Google and adding more of their images.
  - We will evaluate if data augmentation helps us reach a higher accuracy, eventually augmenting only for celebrities with a low image count in the data set.