Data 424 T2P4 Data Documentation

About our Data

For this project, we are working with the CelebA and MNIST datasets. The CelebA dataset was developed by the Multimedia Laboratory at the Chinese University of Hong Kong (https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html). It contains facial image data, 40 binary attributes for each image that describe the person's appearance, and 5 landmark locations on the face for each image. The MNIST dataset was developed by Yann LeCun, Corinna Cortes, and CJ Burges (https://huggingface.co/datasets/ylecun/mnist). The MNIST dataset is composed of 28x28 pixel images of handwritten digits in black and white. These digits were extracted from NIST databases. For each image, it is assigned a label that corresponds to the handwritten integer in the image (0-9, which is 10 classes).

Properties and Amounts

The CelebA dataset contains 202,599 facial images and represents 10,177 unique identities, as well as annotations of 40 binary attributes and 5 landmarks per image. The CelebA images are in color. Examples of these binary attributes include "Wearing Hat" or "Smiling" and values are represented as either 0 (false) or 1 (true) depending on the subject. The MNIST dataset consists of 70,000 black and white images of handwritten numbers, as well as a class label for each that indicates the written number in each image. For example, a black and white image of a handwritten number 9 would have a corresponding label of 9. The labels range from 0-9, which is 10 classes.

Wrangling Techniques and Storage

This image data has been thoroughly prepared and does not require typical data cleaning on our part. The overwhelming amount of our work with these datasets will be focused on splitting the data into train and test splits, which we then input into our convolutional neural networks (CNNs) to train and evaluate. Following bias evaluation with fairness metrics (which are yet to be determined pending literature review and meetings with the team), reevaluation of the train and test splits may be necessary, and these splits may be adjusted to mitigate bias in the model. For this project, our datasets will be downloaded using PyTorch and stored locally on our devices. The code to download these datasets using PyTorch can be found on our GitLab.