Convolutional Neural Network (CNN) for Gender Classification using Images README

Overview

Dependencies

To run the attached Python scripts, the following packages must be installed on a system with Python 3.9 or higher.

- Keras 2.15.0
- Libxml2 2.10.4
- Matplotlib 3.8.0
- Opencv 4.6.0
- Pandas 2.1.1
- Scikit-learn 1.3.0
- Tensorflow-cpu 2.15.0 (this was due to lack of GPU support on virtual machine)

Data Input

There are two main inputs of data for the model: a CSV file of profile data for users and the profile images of the users.

profile.csv

The profile.csv contains many columns of which only two pertain to the following Python scripts: userid and gender. The userid column is used to find the image of the userid since it is labeled "userid".jpg. For the training data, the gender column has either a 0 or 1 value to distinguish whether the user is male or female. Any test data will have nothing in the gender column since that is the value the model is expected to predict.

Profile Images

The profile images are usually in a separate folder and are identified by the userid. Profile images are typically 224 x 224 pixels but cannot be guaranteed to be that size. Profile images may also not be images of faces or people.

Python Scripts

There is only one script to train, validate and then test the CNN model. A separate script was created to separately evaluate test data.

tcss555_cnnopencv.py

This Python script is for the training, validation, and testing of a CNN model to predict gender based on profile images. This model achieved 68% accuracy against hidden test data and beat the class baseline.

The script takes two inputs:

Directory path where the test profile.csv is located and directory path where the test profile images are located.

xmlValidationAcc.py

This is an additional Python script to validate whether the xml files were properly classified by gender according to ground truth for new data.

The script takes two inputs:

Directory path where the xml files are located and the directory path where the ground truth data is located (a profile.csv format is expected).