Gender prediction research methodology

Research Questions

This project is focused on understanding and practicing the implementation of different machine learning algorithms. As such, the following Research Questions (RQs) revolving around implementation and performance of the models will be answered.

RQ1 – How accurate are the machine learning models at predicting age and gender?

This question aims to explore the accuracy of the models used during the project. The models will be trained and tested on the same training and testing sets from the dataset. Each model will be evaluated and a direct comparison between the models will be provided.

RQ2 – How do the models perform when detecting male faces compared to female faces?

This question aims to find if the models perform better at detecting either gender. Two important metrics of binary classification will be used to answer this research question. Namely, the metrics are specificity and precision which can be calculated as shown below.

Specificity = TN/TN+FP

Precision = TP/TP+FP

Where:

True positive (TP) is the number of males correctly classified
True negative (TN) is the number of females correctly classified
False positive (FP) is the number of females incorrectly classified as male
False negative (FN) is the number of males incorrectly classified as female

Data collection and preparation

For the project, the IMDB-WIKI face images dataset will be used. Within the data set, the group will focus on the images of the cropped faces. The entire data set contains 523,051 images which were gathered from IMDB and Wikipedia. The creators of the data set used a pretrained face detector to find faces within the images and create cropped images of the faces. The data set also contains the gender labels which will be utilized for our project. The group will use the cropped face images to train and test our machine learning models. The image data set will be uploaded to the University of Calgary Spark cluster where the data manipulation and the model training will be performed.

As the project is based on image processing, there is minimal data clean up required before training the models. The data set is labelled, meaning it can be used as is to train the models. For preliminary training, the unique data sets will be created by randomly selecting 1%, 5%, 10%, 20% of the original data set. Since the project uses a large data set, even a 1% sample will contain a quantity large enough to train our models.

Analysis

Using the IMDB face images data set, the group aims to create machine learning models which will be able to predict the gender of the person; from the created models, our goal is to answer the above-mentioned research questions. The following supervised machine learning classification techniques along with one artificial neural network technique will be used to create the models:

- Convolutional Neural Network (CNN)
- Support Vector Machines (SVM)
- Trees
- Linear regression classifier
- Logistic regression
- Naïve Bayes classifier

With the smaller sample data sets that will be created as explained in the data collection and preparation section, multiple models will be trained to determine the best model for each of the above categories. With the selected models, the group aims to answer the following research questions.

Expected Results

RQ1 – The accuracy of each technique is expected to vary significantly. The linear regression classifier, Naïve Bayes classifier, trees, and the logistic regression models are expected to perform poorly given the nature of the features used in image processing. It is expected that the CNN and SVM models will outperform the other models given the group is able implement a robust model.

RQ2 – It is expected that the models will perform with similar accuracy for predicting either gender. An initial exploratory data analysis will provide more information about the distribution of the genders within the data set. With a skewed data set, the models may perform better for one gender compared to the other. The expectation is based upon the data set being equally distributed within the two genders; any difference in the accuracies will be further explored should they occur.

Overall workflow of the proposed methodology