

COMP 432: Machine Learning, Concordia University

PROJECT PROPOSAL FORM

Group: 8

Student 1: Mohamad Ghanmeh

Student 2: Nigel Yong Sao Young

Date submitted: 14 Oct 2020

Propose a title for your project. If you had to guess what title you would give your project, what would it be? A good title provides a clue to the main goal and/or main method that the project focuses on. For example, the paper Intriguing properties of neural networks (Szegedy et al. 2014) has a title that, although a little vague, at least suggests that the nature project was an investigation of some kind, focusing on “neural networks,” and that the results are surprising. As another example is the paper The fastest pedestrian detector in the West (Dollar et al. 2010), whose title indicates that the goal is “pedestrian detection” and that the nature of the contribution is “speed.”

Age, Gender and Race Face Classification

Describe the goal of your project. What are you trying to achieve? What “main question” are you trying to answer, or at least to provide evidence for? Secondary goals are OK, but you should still have a clear “main goal” or “main question.” From your description, it should also be clear whether your project is about: making better predictions for some application? speeding up training and/or predictions? simply comparing predictive performance and/or speed of several methods? assessing or comparing interpretability? understanding failure modes or sensitivities of some methods? Etc.

Our main goal is to be able to make accurate predictions on a face's age, gender and ethnicity. To achieve this, we will explore ways on how Transfer Learning and splitting the training and testing data among groups affect our results and see if there are any benefits to do so. We will also use multiple models - from basic linear regression to neural networks - to achieve the best accuracy. Our main end goal then is to create a functional user interface for anyone to be able to play around with our model.

Describe the data you plan to use. One of the hardest steps for a good machine learning project is to find data that is truly suitable for your goals. Finding good data is not the most fun part, but it's one of the most important—after all, for machine learning it is “garbage in, garbage out”. Here are some things you should ideally know:

What are the ‘modalities’ that apply to the data?

The data consist of a CSV file that contains images, which have the following attributes: The age, the ethnical group, and the gender.

What does an input look like?

The input is a face image, basically a face portrait image.

For an example input, what does a desired output look like?

From the face image passed, the desired output is the prediction of the age, gender and ethnicity.

How are the training and testing data to be split? (randomly shuffled, by some grouping, by time period, etc.)

We will experiment with a basic random 80/20 split and also splitting by some grouping (for example, taking mostly young people's faces) to see how our predictive model still predicts accurately or is biased.

Will the data need preprocessing before you can feed it into a training algorithm?

No, this data is already processed and ready to be used.

Is the data small enough to train models on your computer, or is there a risk of scalability/engineering difficulties?

Yes, the size of the file is 190mb with all the images and parameters included in it. So we believe we will not experience any performance issues to train the models. If there is, we plan to use google collabs GPU and TPU capabilities.

Describe how you will measure “success.” You should explain how you will know whether you have achieved the goal(s) that you described earlier. What does “success” look like? What does “failure” look like? Keep in mind that your project can still succeed (in the sense of a good grade!) even if the experimental results are bad—what is important is that your experimental results are conclusive! A bad project is one in which you cannot even tell whether the goal was achieved or not.

Success

- Good accuracy of prediction when testing the model. A good accuracy would be above 80%.
- We would have a user-friendly interface to enable a user to interact with the model.
- Draw insightful conclusions when exploring different ways of training the data.

Failure

- Failure would look like having an accuracy below 70%, and not having a working interface for passing on new images to return their key values.

Describe how work will be divided. It is very important for everyone to have a meaningful role in the project. If one person (the most experienced person) does all the programming or writing, then everyone else in the group loses this important chance to gain experience. For example, if there is no way to “happily divide” the work because two group members want to work on the same part, that is totally OK and no one should feel guilty for wanting that; both group members can do their own version of that part of the project, and then the final report can say “two group members each implemented did this part, and their results {matched, didn’t match}” When two people attempt and come to different conclusions, that is interesting and a chance for everyone to learn!

We will both work on the code and the report, while Nigel’s main focus will be on the code and Mohamad’s main focus will be on the report. In the programming part, Nigel will work on training the neural networks, Mohamad will work on training the logistic regression model. Both Mohamad and Nigel will work on exploring how splitting our data among different groups and doing transfer learning changes our predictive model. We will actively peer-review each other in both the writing and the programming parts. The main python packages we expect to use are TensorFlow, Scikit-learn, Numpy, Matplotlib and Pandas. As for the application interface, we plan to do a simple web application where we will split work doing front-end and back-end.