Face\_Recognition

**Activation:**

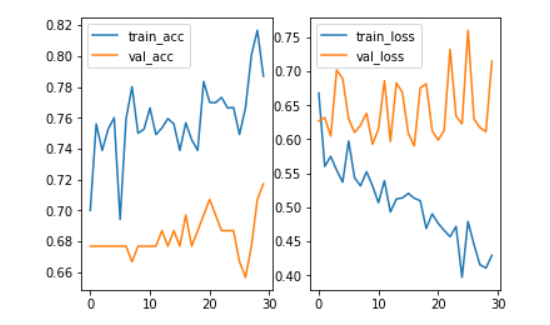
As a IT person, we usually have a nickname called “homeboy” or heard you have a “flavor of IT”. I think it is not fair for us because we just a cluster of person loving technology. Therefore, I want to research if it is truth.

**Design Process:**

1. **Define question and evaluation metric**
   1. Let data from crawler be training data and validation data.
   2. Let manual data be test data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Time per photo** | **Advantage** | **Disadvantage** |
| **Manual** | **1 min** | **Label without error** | **Slow** |
| **Crawler** | **1 second** | **Fast** | **Label problem** |

1. **Collect data**
   1. Using **selenium** to avoid being detected by Linkedin.
   2. Get 1000 images of engineers and 1000 images of salesmen.
   3. Collect people’s name, job, and image from the Linkedin.
   4. Decode url of images.
2. **Model**
   1. Split 80% for training data and 20% for validation data.
   2. Detect face through **CascadeClassifier** package.
   3. Use **Cnn** to predict.

****

**Problem**

According to above image, the result is very bad. Therefore, I organize the following points which is probably the reason resulting in failure.

1. Labeling data exists potential problem such as that the people will change another kinds of job.
2. There are some factors including gender or age probably will influence the result.
3. Need more data due to this question is not a obvious classifiaction.

**Improvement**

According to above problems, I have some ideas to improve. However, I still don’t ensure if It will be successful unless I try.

1. Labeling data exists potential problem such as that the people will change another kinds of job.
2. Define more explicitly question such as that I should focus on certain gender and range of age.
3. Collect more data through other platforms because Linkedin have detected me one times.

**Conclusion**

The final accuracy of validation data is 70% and accuracy of testing data is 40%. I think the performance is terrible due to above reasons.

To be honest, this is not a successful project but interesting. I will stop temporarily this project due to other plan. I will improve it again in the future according to above points.

To sum up, I think the most important in the data pipeline is defining problem. Our final purpose is to solve problem, so we need to ensure and focus on our problem.