Prediction of personality traits based on facial features extracted from videos using Deep Learning

INDIANA UNIVERSITY

Ankit Saxena

Abstract

"According to psychology researchers, the first impressions are formed in limited exposure (100ms) to unfamiliar faces."[1] During interviews, the mindset of the interviewer or perception of the interviewer can affect the selection of an individual. It is possible to train Machine Learning models to classify the personality traits of an individual based on facial expressions, body language, or speech of the individual in the video. One of the most commonly used personality model is the Big-Five model which rates the five traits of Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN).

1. "Prediction of Personality First Impressions With Deep Bimodal LSTM" by Karen Yang, Stanford and Noa Glaser, Stanford

Background

There are many organizations in the market like Affectiva, Tilr, Emotion Research Lab that are working on detection of human emotions and cognitive states from facial cues or physiological responses. Most of them model the personality traits based on the Big-Five traits.

The existing non-deep learning approaches use speech, audio, text, and visual information utilizing variations of Support Vector Machines, Logistic Regression, Convolutional Neural Networks, and Hidden Markov Models. Many of the top performers using the deep learning approaches are from the ChaLearn Looking at People challenge. The top performers in the competition achieved around 90% accuracy.[1, 2]

Most of the existing approaches use both video and audio temporal data to predict the OCEAN traits. The overall movement of the individual in the video might suggest their interest in the topic, excitement level, or their kinetic expressiveness.

Objective

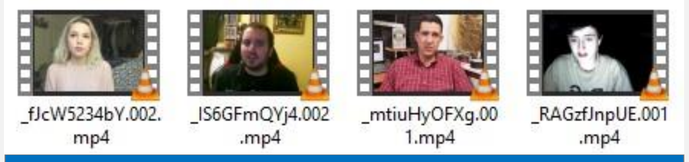
- To replicate the results of a few of the deep learning models and try to improve on those implementations.
- The dataset used is taken from ChaLearn Looking at People 'First Impression 2016' challenge. It consists of 11,000 15-second videos that are collected from YouTube, and annotated with OCEAN personality traits by Amazon Mechanical Turk workers.[2]

2. http://chalearnlap.cvc.uab.es/dataset/20/description/

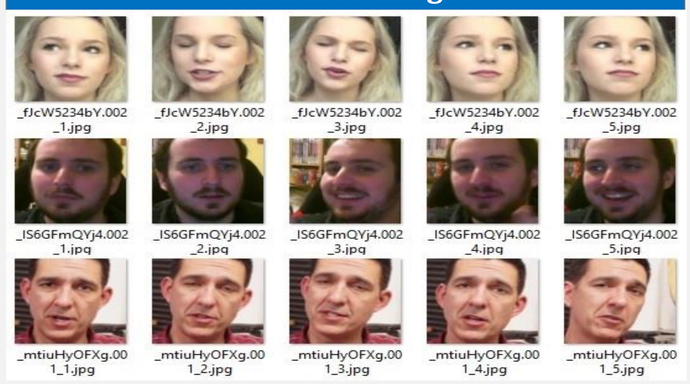
Dataset

Training dataset consists of 6000 short videos of average duration 15 seconds. Validation dataset consists of 3000 videos and the test dataset consists of another 2000 videos.

A tiny sample of the First Impression



Generated Images



Preparation of the training dataset

Each video in training and validation dataset has associated OCEAN trait values in [0, 1]. These values are stored in a csv file with the video name and numerical values for 5 features, namely, openness, conscientiousness, extroversion, agreeableness, and neuroticism. From each video in the training dataset, 5 random frames were chosen as training images.

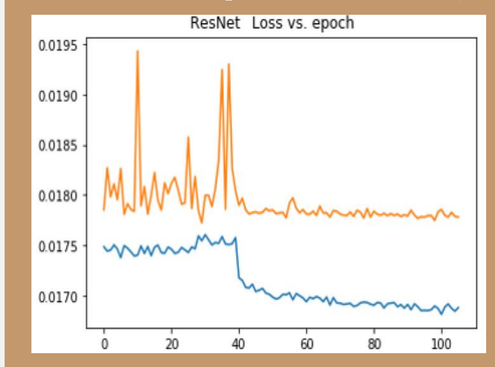
Experiment

ResNet34

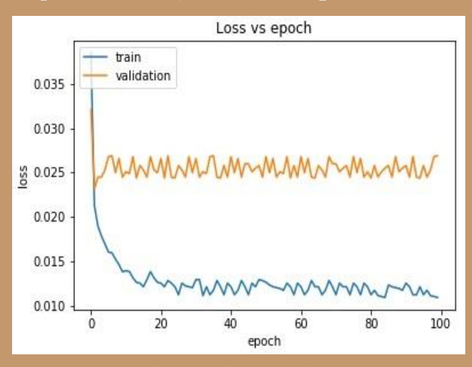
- As the per the experiment referenced[1], the model was trained with a batch size of 32 using stochastic gradient descent with a learning rate of 5e-3, momentum of 0.9, and weight decay of 5e-4. Images were normalized using mean = [0.485, 0.456, 0.406] and sd = [0.229, 0.224, 0.225].

- The experiment referenced (R) only takes one frame per video, performs training on all the generated images, and while testing predicts the OCEAN traits of an individual using only one test frame from a video.
- The experiment performed (P) takes five frames per video, performs training on all the generated images as independent samples, and while testing predicts the OCEAN traits as the average of predictions for the five test frames from a video.

Here is a comparison of the loss (mean squared error) for two experiments



Referenced Experiment
Loss from training (blue) and validation (orange) at each epoch



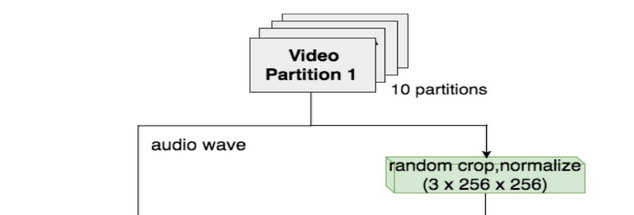
Performed Experiment
Loss from training (blue) and validation (orange) at each epoch

These graphs suggest that the models lack the explanatory power to predict apparent personality

Note: The experiment was performed on only 40% of the training and validation dataset.

Bi-modal LSTM

- As the per the experiment referenced[1], the Bi-modal LSTM network has two branches, one for extraction of visual features and one for audio features. The output of these two branches are later concatenated as input for each time step of the LSTM. The video is partitioned into 10 sequence partitions, and a mini-batch size of 8 is used for training.

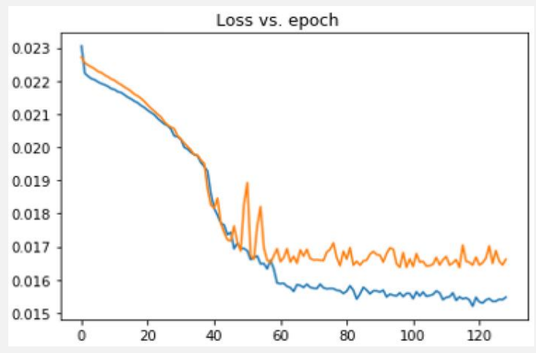


Deep Bi-modal LSTM Architecture from the referenced experiment

- In the experiment performed, each video is split into 5 sequential partitions, and a random frame with the face is extracted from each partition to build up the temporal relationship between the video frames.
- Videos are processed in a batch of 10 videos. Therefore, each batch consists of 50 images with pixel values in range [0, 1].
- The ResNet34[3] model is same as used earlier. There is only one linear layer of dimension 512 * 128 followed by dropout with probability of 0.2 to avoid overfitting.
- After this an LSTM is trained with 128 long short-term units and one layer of five linear units. The final prediction is the average of predictions over 5 time steps, instead of 10 in referenced experiment[1].

3. Deep Residual Learning for Image Recognition by Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Microsoft Research

Referenced Experiment
Loss from training (blue) and validation (orange) at each epoch



Results

- Evaluation Results: Accuracy of the referenced models, the top 5 competitors in the ChaLearn challenge, and the experiments performed are given below. The scores were calculated using the criterion used in the ChaLearn LAP 2016 competition[1, 2] :

$$accuracy = 1 - \frac{1}{5N}\sum_{j=1}^{5}\sum_{i=1}^{N} \|\text{groundTruth}_{ij} - \text{predicted}_{ij}\|$$

	Total	E	A	C	N	O
LSTM (R)	0.9083	0.9110	0.8944	0.9220	0.9005	0.9136
ResNet (R)	0.8935	0.8942	0.8952	0.8901	0.9012	0.8867
cNJU-LAMBDA	0.9130	0.9133	0.9126	0.9166	0.9100	0.9123
evolgen	0.9121	0.915	0.9119	0.9119	0.9099	0.9117
DCC	0.9100	0.9107	0.9102	0.9138	0.9089	0.9111
ucas	0.9098	0.9129	0.9091	0.9107	0.9064	0.9099
BU NKU	0.9094	0.9161	0.907	0.9133	0.9021	0.9084
LSTM (P)	0.8931	0.8927	0.8964	0.9022	0.8821	0.8923
ResNet34 (P)	0.8518	0.9027	0.9014	0.9570	0.840	0.9337

Conclusion and Future Work

- ResNet34[3] implementation performs better than other models for Openness, Conscientiousness traits. It fails to generalize well for the Neuroticism trait.
- LSTM implementation performs better for all the traits on average. The results are comparable to the referenced experiment. This could be a good model to predict first impressions of apparent personality.
- One of the major assumption in the experiments performed is that the face is the most salient element in a frame.
- For future work, I would like to include the transcript of the audio as one of the inputs in the model. Also, I would like to repeat the same experiments without making the assumption mentioned above. Other factors like background, clothes, or body could be important factors.

Note: All the images marked as referenced experiment were reprinted from the paper "Prediction of Personality First Impressions With Deep Bimodal LSTM by Karen Yang, Stanford and Noa Glaser, Stanford"

Acknowledgment

I would like to thank professor David Crandall and the associate instructors for their valuable guidance through every phase of this project.

Done as a part of CSCI-B657 Computer Vision course at Indiana University, Bloomington.