Project Proposal: Deep Learning for Automatic Personality Analysis

Jingxi Xu jx2324

Qiangeng Xu qx2128

Xuefeng Hu xh2348

Abstract

We propose a project to develop deep learning models to automatically analyze candidates' big five personality traits from YouTube videos of them speaking. In particular, we intend to use the cutting-edge video and audio processing libraries to preprocess the video and adopt *long-short-term-memory* (LSTM) techniques to exploit temporal patterns among image frames, as inspired by the work of Subramaniam et al. (2016). This project will have potential applications to automatic evaluation of job applicants from interview videos, and even lie-detecting of crime suspects.

1 Introduction

The ability of machines to interpret and analyze images and videos have made great strides in recent years as a result of the development of computer vision and deep learning techniques. While the state-of-the-art deep learning models have already made superhuman performance in image classification problems, the analysis of videos to generate insightful interpretation (clustering, classification, description, etc.) remains to be a harder problem and is attracting huge attention from computer vision and deep learning research communities.

In this project, we aim to leverage the cutting-edge video and audio processing techniques and deep learning models to automatically analyze videos of persons speaking in front of the camera to reveal their personalities, with regards to the *big five personality traits*.

Many contemporary personality psychologists believe that there are five basic dimensions of personality, often referred to as the "big five personality traits" (contributors, 2018). The five personalities are listed as follows:

- Openess: appreciation for art, emotion, adventure, unusual ideas, curiosity, and variety of experience.
- Conscientiousness: a tendency to be organized and dependable, show self-discipline, act dutifully, aim for achievement, and prefer planned rather than spontaneous behavior.
- Extraversion: energy, positive emotions, surgency, assertiveness, sociability and the tendency to seek stimulation in the company of others, and talkativeness.
- Agreeableness: a tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others.
- Neuroticism: neuroticism identifies certain people who are more prone to psychological stress.

Our goal is to output five scores in the range of $[0, 1]$ (closed interval), one for each personality of the person speaking in the video. We want a system to automatically do the job with deep learning techniques. This project can easily be adapted to other various applications. We can apply the learned scores from interview videos to help HR choose proper candidates, or even develop a learning system

to make the decision from the five scores directly with the labels given by HR. In addition, from public security's perspective and with the help of psychologists, we might be able to apply similar technologies to tell whether a crime suspect is lying or not based on his facial expressions.

2 Related Work

The most related work in literature would be the *ECCV 2016 workshop on automatic personality analysis and first impressions challenge* (Ponce-López et al., 2016). In this challenge, various teams have developed different deep learning models to predict five personality scores for a 15-second video of a person speaking.

Wei et al. (2017) proposes a Deep Bimodal Regression (DBR) framework. For the visual modality, they modify the traditional convolutional neural networks for exploiting important visual clues. For the audio modality, they extract audio representations and build a linear regressor. To combine these complementary information, they ensemble these predicted regression scores by both early fusion and late fusion.

Subramaniam et al. (2016) proposes two end-to-end trained deep learning models that use audio features and face images for recognizing first impressions. The first model uses Volumetric (3D) convolution based deep neural network for determining personality traits. The second model formulates an LSTM based deep neural network for learning temporal patterns in the audio and visual features. The LSTM model slightly outperforms the 3D convolution based model.

3 Plan

3.1 Data Set

The data set can be obtained from the ChaLearn Looking at People website (`http://chalearnlap.cvc.uab.es/dataset/20/description/`).

There is a newly collected data set of 10000 15-second videos of people speaking to the camera collected from YouTube by Microsoft, annotated with personality trait scores by *Amazon Mechanical Turk* (AMT) workers. The groundtruth for each video is a vector of 5 real numbers in the range of $[0, 1]$, each representing a personality trait. Thus, this project can also be viewed as a clustering problem of videos.

3.2 Model

Videos can be viewed as collections of images in a temporal order. Therefore, the most significant task in video clustering/classification is to take advantage of the temporal patterns among image frames of a video. LSTM (Hochreiter and Schmidhuber, 1997) is a very efficient technique in learning the temporal sequence, so we are particularly interested in the work by Subramaniam et al. (2016).

We will divide the video separately into a visual part and a audio part. For each part, we further divide them into non-overlapping partitions. We adopt the state-of-the-art libraries (OpenFace and pyAudioAnalysis in particular) to extract useful features as the preprocessing steps for each partition. These features from separate parts will then be concatenated into a single vector for each partition, and the multiple vectors (one for each partition) will be fed together (maintaining the temporal order) to a LSTM model where the temporal patterns are exploited.

4 Objectives

The existing work of Subramaniam et al. (2016) was written in Torch framework, so our first objective is to re-implement this model in Keras and Tensorflow, and be able to obtain similar results on the test data.

Then we want to modify the model in one of the following directions (note that we might not be able to get better results on test data after these modifications, but what matters here are the ideas):

Body language Since the existing model only analyzes the personalities from subjects' facial landmarks, head positions and audio signals, we might want to exploit their body languages as well, which means we might need to get their limb landmarks as well.

Different preprosessing The current audio features obtained by pyAudioAnalysis might not be good enough for the task, we might want to try other libraries.

Prior knowledge Modify the models to incorporate advice from psychological experts such as that girls tend to show more extraversion, smiles indicate openness, etc.

Bias reduction We want our model to focus on personality and is not biased towards other factors like gender. Thus the features learned from these videos will be good at predicting personalities but for predicting gender, they should be as poor as random guess.

References

Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European Conference on Computer Vision*, pages 337–348. Springer, 2016.

Wikipedia contributors. Big five personality traits — wikipedia, the free encyclopedia, 2018. URL `https://en.wikipedia.org/w/index.php?title=Big_Five_personality_traits&oldid=831425438`. [Online; accessed 22-March-2018].

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision*, pages 400–418. Springer, 2016.

Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.