INTERNSHIP PROJECT REPORT

Deep Impression for Personality Psychometric Analysis

Shrey Srivastava

B. Tech Computer Science and Engineering

Specialization in AI and ML

shrey.srivastava.08@gmail.com; shrey.srivastava2019a@vitstudent.ac.in

1. Abstract:

Apparent personality analysis from short video sequences is a challenging problem in computer vision and multimedia research. The solution to this issue I have used Deep Bimodal Regression (DBR) Framework. In DBR, conventional Convolution and Pooling Layers are used to extract visual information from the frames in the videos in the dataset. On the other hand, for audio modality, Linear regression is used to extract auditory features. These two independent analyses are clubbed and activated using sigmoid to get a quantifiable open to interpretation for the client. For the model to adapt to short and large video we need to incorporate a huge dataset for the same. The one used in this project is the one used in CVPR 2017 and is named First Impressions V2. The mean accuracy achieved is 0.9130.

Keywords: Apparent personality analysis, deep regression learning, bimodal learning, convolutional neural networks.

2. Introduction:

With time being of essence in every sector or company, putting the least effort and maximum output is of the highest priority. Such is also the case with interviews. While most of the areas involved in interview tests is covered through systematic tests like tests based on aptitude and cognitive ability, it is not only the IQ and problem solving that makes the candidate worth hiring. The other way to approach the issue of hiring an asset is to investigate their personality, which, in itself, is a difficult task. Personality Analysis is one of the major focuses of human centered video analysis. The goal of this project is to do an analysis, audio and visual, to define the personality of a person on the widely accepted standard called Big Five.

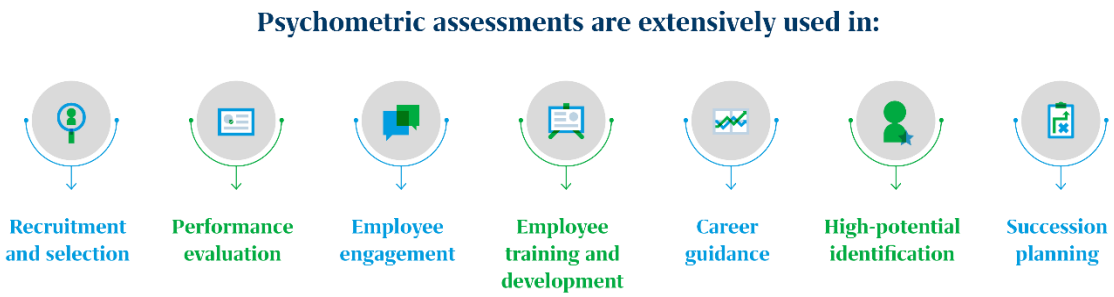
This Big Five Personality Analysis does a deep impression of the person and gives out output of levels of openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. We will discuss the aforementioned categories later. When it comes to analysis, it is very complicated as it is not just the face and the fluency to be accounted for. Speech is very different for two unique people. They might have different vocabulary, modulation, or articulation in general. Accompanied with that we must factor in noise in camera channels, infinitely many kinds of out-of-vocabulary motion, and real-time performance constraints.

The solution I have applied in my model is a widely accepted approach to human-centered video analysis. It involves the independent analysis of video and audio followed by a common analysis to determine the personality of the person. The model is so efficient that it can be pretty accurate with just 15 seconds of interview in place. The model is good for first impressions as well as a deep analysis of the personality. When this model is coupled with the already in place aptitude and technical tests, the candidates getting the green light from all these systems will have the highest probability of being immensely valuable to the company.

About Psychometric Tests:

A psychometric test or psychometric assessment is an evaluation of an individual's cognitive skills and personality traits. It helps assess whether the individual is capable of thriving in a specific professional role. Psychometric testing can help understand aspects of mental ability and behavioural style that organizations are unable to gauge during conversations and interviews.

A psychometric test is a standard and scientific method that plays an equally significant role in educational or clinical settings. It also offers an unbiased evaluation of a broad range of parameters, such as logical reasoning, industry-specific aptitude, role-specific qualities, type of personality and more.



Psychometric assessments are usually of the following two types:

- Personality tests
- Cognitive ability tests

As cognitive ability is monitorable and quantifiable by injecting a simple test, with the video input of the interview and its audio-visual analysis we will

investigate the personality part of the psychometric test in this project. But before that let's look into what exactly are the attributes on which the personality is judged.

- **Personality tests**

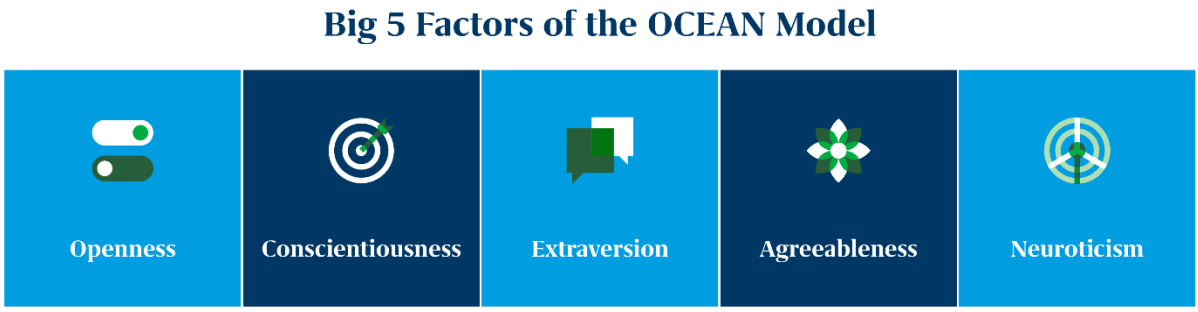
Personality Tests are a form of psychometric assessment that helps identify specific personality traits required to perform in a job role or industry. These tests offer significant insights into a candidate's key qualities, motivations, behavioural styles, etc.

The science behind testing personality

Personality tests usually follow two major schools of thought: the trait-based approach and the type approach. While the type theory categorizes personalities into introverted/extroverted, the trait theory measures the degree to which key personality traits exist in an individual.

Personality tests based on the type theory often lack objectivity and comprehensive insight into an individual's psyche. When used in a professional context, the type-based tests also lack reliability, as it is possible for an individual to get different results every time.

Trait-focused personality assessments are all based on the Big 5 Factor model, also known as the OCEAN model. Developed in the 1970s, this model enlists five major personality traits that exist among humans in varying degrees.



A good psychometric assessment can help measure the level of these traits through a series of questions and problem-solving exercises.

I. Openness

Individuals possessing this trait have a natural penchant toward adventure and art. They are curious, creative and open to change. Meanwhile, people averse to openness stick to their old routine, habits and keep new experiences at bay.

II. Conscientiousness

People high on conscientiousness are organized and have a sense of responsibility. They have the drive to achieve their goals and are highly reliable. This trait has shown marked achievement on the job. People on the opposite side, however, are spontaneous and careless.

III. Extraversion

Individuals who possess extraversion indicate various characteristics, including sociability and talkativeness. They draw their energy from day-to-day social interactions or gatherings. Such individuals are mostly cheerful and assertive in their approach. Meanwhile, introverts are a professor of ‘me time.’ While the trait often gets mixed up with being shy, that’s not the case. Individuals with a high introversion trait prefer smaller group activities when required and tend to enjoy their own company.

IV. Agreeableness

Agreeableness is indicative of a person’s kindness. Such individuals are trusting and helpful. On the other hand, disagreeable people are cold, suspicious of others and less cooperative.

V. Neuroticism

Individuals possessing this trait worry a lot and often find themselves feeling depressed and anxious. On the contrary, people low on neuroticism are emotionally stable and exhibit calmness and composure when facing problems.

Trait-focused personality assessments give a more granular picture of an individual's personality than type-focused tools. They offer people a more accurate reflection of candidates' natural preferences and behavioural styles. As a result, they tend to be more effective in making workplace decisions like recruitment, L&D, succession planning, etc.

3. Related Work:

In this section, we will briefly review the related work for visual-based deep learning, audio representations and apparent personality analysis.

3.1. Visual-Based Deep Learning:

Deep learning refers to a class of machine learning techniques, in which many information processing layers organized in a sequential structure are exploited for pattern classification and for feature or representation learning. Recently, for image-related tasks, Convolutional Neural Networks (CNNs) allow computational models that are composed of multiple processing layers to learn representations of images with multiple levels of abstraction, which have been demonstrated as an effective class of models for understanding image content, giving state-of-the-art results on image recognition, segmentation, detection and retrieval. Specifically, the CNN model consists of several convolutional layers and pooling layers, which are stacked up with one on top of another. The convolution layer shares many weights, and the pooling layer sub-samples the output of the convolution layer and reduces the data rate from the layer below. The weight sharing in the convolutional layer, together with appropriately chosen pooling schemes, endows the CNN with some invariance properties (e.g., translation invariance). In our DBR framework, we employ and modify multiple CNNs to learn the image representations for the visual modality, and then obtain the Big Five Traits predictions by end-to-end training.

3.2. Audio Representations:

In the past few years, many representations for audio have been proposed: some of them are time domain features, and others are frequency domain features. Among them, there are several famous and effective audio features, to name a few, Mel Frequency Cepstral Coefficients (MFCC),

Linear Prediction Cepstral Coefficient (LPCC) and Bark Frequency Cepstral Coefficient (BFCC). Particularly, the Mel Frequency Cepstral Coefficients (MFCC) features have been widely used in the speech recognition community. MFCC refers to a kind of short-term spectral-based features of a sound, which is derived from spectrum-of-a-spectrum of an audio clip. MFCC can be derived in four steps. During the four steps, the log filter bank (logfbank) features can be also obtained. In the proposed DBR framework, we extract the MFCC and logfbank features from the audios of each original human-centered video for APA. In our experiments, the results of logfbank are slightly better than the ones of MFCC. Thus, the logfbank features are used as the audio representations in DBR.

3.3. Apparent Personality Analysis

Personality analysis is a task that is specific to the psychology domain. Previous research in personality analysis usually needs psychology scientists to figure out the results or need participants to do specific tests containing large number of questions which can reflect their personalities. However, such process will cost a lot of time and funds. A similar task to personality analysis in computer vision is the emotion analysis tasks. Emotion analysis can be regarded as a multiple class classification problem, where usually four emotions (sadness, happiness, anger and neutral state) are recognized by the algorithms. However, in apparent personality analysis, it needs to predict the Big Five Traits (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) which are independent with each other and whose scores are continuous values in the range of [0, 1]. Thus, it is obvious to see the apparent personality analysis tasks is more realistic but difficult than emotion analysis.

4. DBR Framework:

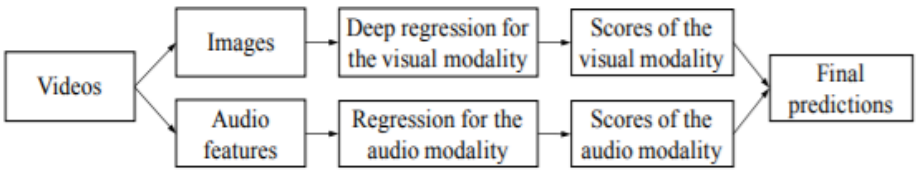


Figure 1. Framework of the proposed Deep Bimodal Regression method. In DBR, the original videos are treated as having two natural modalities, i.e., the visual modality for images and

the audio modality for speeches. After learning the (deep) regressors on these two modalities, the final predicted personality traits are obtained by late fusion.

As shown in Fig. 1, DBP has three main parts: the first part is the visual modality regression, the second part is the audio one, and the last part is the ensemble process for fusing information of the two modalities.

4.1. Deep Regression for Visual Input:

The deep regression part contains three subparts: image extraction, deep regression network training and regression score prediction.

Image Extraction:

The inputs of traditional convolutional neural networks are single images. But for the APA task, the original inputs are the human-centered videos. In order to utilize powerful CNNs to capture the visual information, it is necessary to extract images from these videos. For example, for a fifteen second length video whose frame rate is 30fps, there are 450 images/frames from each original video. However, if all the images/frames are extracted, the computational cost and memory cost will be quite large. Besides, in fact, nearby frames look extremely similar. Therefore, we down sample these images/frames to 100 images per video. That is to say, in each second, we extract 6 images from a video. After that, the extracted images/frames are labelled with the same personality traits values as the ones of their corresponding video. In consequence, based on the images, we can train the deep regressors by CNNs for APA.

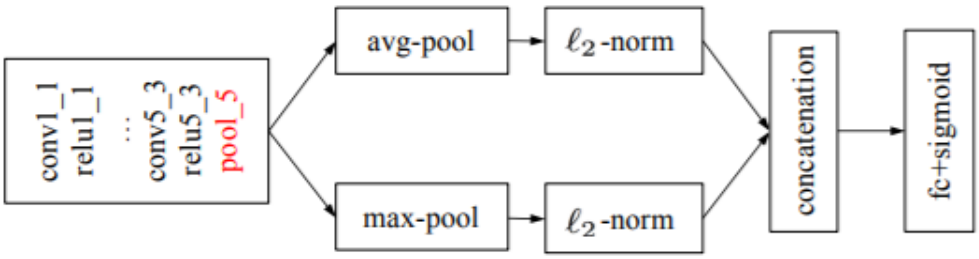


Figure 2. Architecture of the Descriptor Aggregation Network (DAN) model. Note that we removed the fully connected layers. The deep descriptors of the last convolutional layer (Pool5) are firstly aggregated by both average- and max-pooling, and then concatenated into the final image representation for regression.

Deep Regression Network Training:

In the visual modality of DBR, the main deep CNN models are modified based on previous research, which are called Descriptor Aggregation Networks (DANs). What distinguishes DAN from the traditional CNN is: the fully connected layers are discarded and replaced by both average and max pooling following the last convolutional layers (Pool5). Meanwhile, each pooling operation is followed by the standard L2- normalization. After that, the obtained two 512-d feature vectors are concatenated as the final image representation. Thus, in DAN, the deep descriptors of the last convolutional layers are aggregated as a single visual feature. Finally, because APA is a regression problem, a regression (fc + sigmoid) layer is added for end-to-end training. The architecture of DAN is illustrated in Fig. 2. Because DAN has no fully connected layers, it will bring several benefits, such as reducing the model size, reducing the dimensionality of the final feature, and accelerating the model training. Moreover, the model performance of DAN is better than traditional CNNs with the fully connected layers. In the experiments of the proposed DBR framework, adopt the pre-trained VGG-Face model as the initialization of the convolutional layers in our DANs. For further improving the regression performance of DAN, the ensemble of multiple layers is employed. Specifically, the deep convolutional descriptors of ReLU5 2 are also incorporated in the similar aforementioned aggregation approach, which is shown in Fig. 3. Thus, the final image feature is a 2048-d vector. We call this end-to-end deep regression network as “DAN+”.

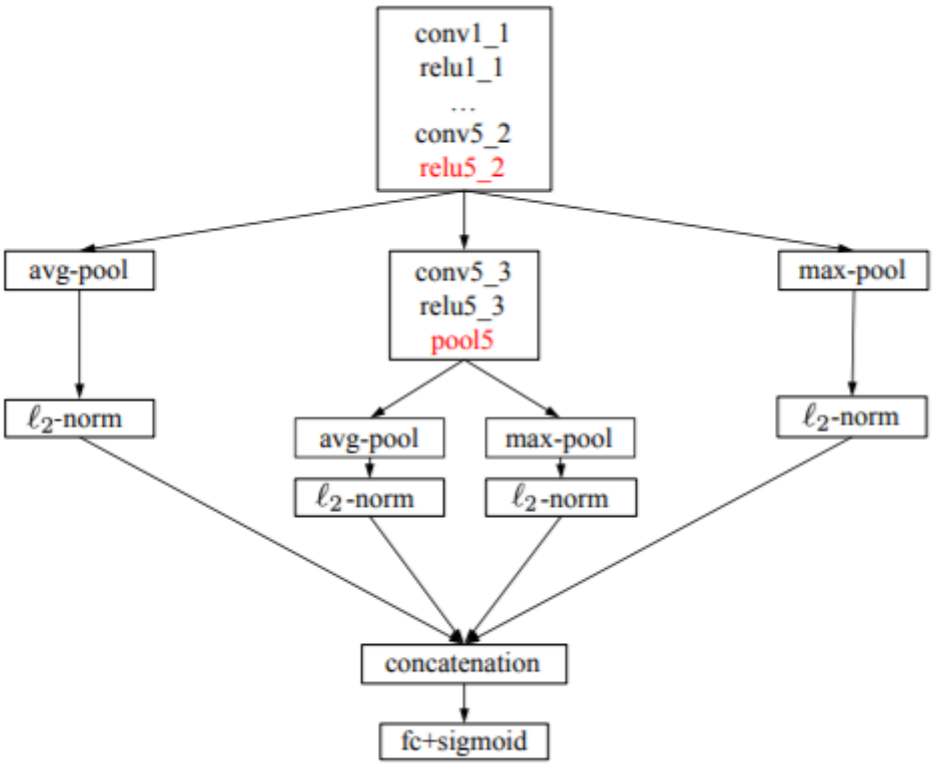


Figure 3. Architecture of the DAN+ model. In DAN+, not only the deep descriptors of the last convolutional layer (Pool5) are used, but the ones of ReLU5 2 are also aggregated. Finally, the feature vectors of multiple layers are concatenated as the final image representation for regression.

Personality Traits Prediction:

In the phase of predicting regression values, images are also extracted from each testing video. Then, the predicted regression scores of images are returned based on the trained visual models. After that, we average the scores of images from a video as the predicted scores of that video.

4.2. Regression for the Audio Modality:

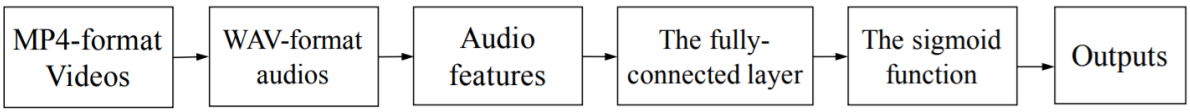


Figure 4. Pipeline of the regression for the audio modality. The log filter bank features are used as the audio representations/features. Based on the audio features, a linear regressor is trained for predictions.

As aforementioned, in the audio modality, we choose the log filter bank (logfbank) features as the audio representations. The logfbank features can be extracted directly from the original audios from videos. After that, we use a model composed of a fully connected layer followed by a sigmoid function layer to train the audio regressor. The L2 distance is used as the

loss function to calculate the regression loss. The whole pipeline of the audio modality can be seen in Fig. 4.

4.3. Modality Ensemble:

After the training of both the visual and the audio modalities, modality ensemble is used as the late fusion approach for getting the final regression scores. The ensemble method we used in DBR is the simple yet effective simple averaging method. In APA, the predicted result of a trained regressor is a five-dimensional vector which represents the Big Five Traits values, i.e., si = (si1, si2, si3, si4, si5). We treat each predicted result of these two modalities equally. For example, the predicted results of the visual modality are s1, s2 and s3, and the results of the audio modality are s4 and s5. The final ensemble results are calculated as the mean of all the trait values.

5. Experiments:

Following is the analysis of the framework regarding accuracy, performance, comparison and involves a detailed outlook into the Dataset used for training the model.

5.1. Dataset and Evaluation Metric:

The first impressions data set comprises 10000 clips (average duration 15s) extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing and speaking in English to a camera. The videos are split into training, validation, and test sets with a 3:1:1 ratio. People in videos show different gender, age, nationality, and ethnicity.

Videos are labelled with personality traits variables. Amazon Mechanical Turk (AMT) was used for generating the labels. A principled procedure was adopted to guarantee the reliability of labels. The considered personality traits were those from the Five Factor Model (also known as the Big Five), which is the dominant paradigm in personality research. It models human personality along five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. Thus, each clip has ground truth labels for these five traits represented with a value within the range [0, 1].

Here, we also introduce an extension of this dataset. Specifically, we supplement the dataset with new language data (transcriptions), which complements the existing sensory data (videos) as well as a new job-interview variable (interview annotations), which complements the existing personality trait variables (trait annotations).

Transcriptions. All words in the video clips were transcribed by the professional transcription service Rev. In total, 435984 words were transcribed (183861 non-stopwords), which corresponds to 43 words per video on average (18 non-stopwords). Among these words, 14535 were unique (14386 non-stopwords).

Interview annotations. In addition to labelling the apparent personality traits, AMT workers labelled each video with a variable indicating whether the person should be invited or not to a job interview (the "job-interview variable"). This variable is also represented with a value within the range [0, 1].

But in this rendition of the dataset, interview annotations are not taken into consideration. The parameter whether the applicant is to be considered for further correspondence is kept liquid and user based.

For evaluation, given a video and the corresponding traits values, the accuracy is computed simply as one minus the absolute distance among the predicted values and the ground truth values. The mean accuracy among all the Big Five traits' values is calculated as the principal quantitative measure:

$$\text{Mean accuracy} = \frac{1}{5N}\sum_{j=1}^{5}\sum_{i=1}^{N} 1 - \left|\text{ground\_truth}_{i,j} - \text{predicted\_value}_{i,j}\right|$$

Where N is the number of predicted videos

The first work studying the big-5 traits from audio-visual cues in thin-slices of YouTube personal videos using MTurk annotations and automatic inference of personality traits was presented in:

- J.-I. Biel, O. Aran, and D. Gatica-Perez, You Are Known by How You Vlog: Personality Impressions and Nonverbal Behaviour in YouTube in Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM), Barcelona, Jul. 2011

- J.-I. Biel and D. Gatica-Perez, The YouTube Lens: Crowdsourced Personality Impressions and Audio-visual Analysis of Vlogs, IEEE Trans. on Multimedia, Vol. 15, No. 1, pp. 41-55, Jan. 2013

5.2. Implementation:

Details of the visual modality

As aforementioned, in the visual modality, we firstly extract about 100 images from each video. Specifically, for most videos, 92 images are extracted (about 6.1fps). After that, we resize these images into the 224 x 224 image resolution. In consequence, there are 560,393 images extracted from the training videos, 188,561 images from the validation ones, and 188,575 images from testing. Fig. 5 illustrates three examples of extracting image from videos. In our experiments, the visual DAN models in the proposed DBR framework are implemented using the open-source library MatConvNet. Beyond the DAN models, we also employ a popular deep convolutional network, i.e., Residual Network, as another regression network for boosting the visual regression performance. In the training stage, the learning rate is 10^{-3}. The weight decay is 5×10^{-4}, and the momentum is 0.9 for all the visual models.

Details of the audio modality

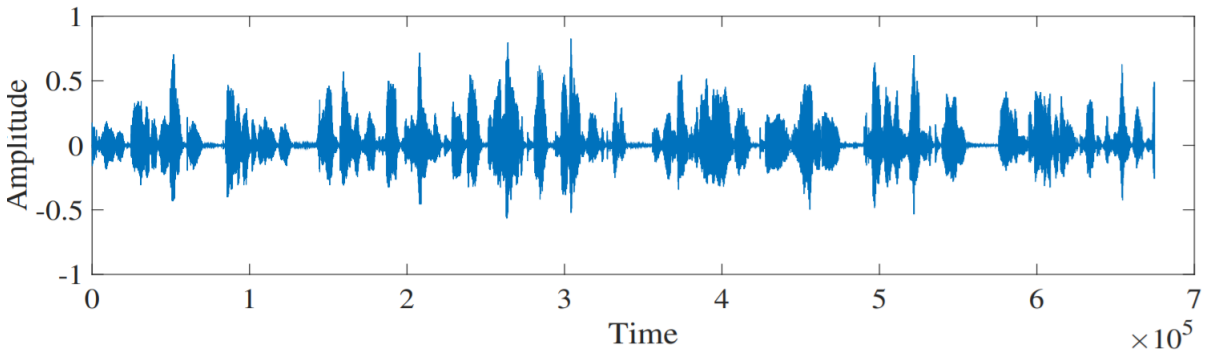


Figure 6. The wave forms of a sampled audio. The horizontal axis stands for the time. Because we set the sampling frequency as 44,100Hz, the unit of the horizontal axis is 1/44100s. The vertical axis is the amplitude.

In the audio modality, we firstly extract the audio features from the original videos, and then learn a linear regressor based on these audio features. In the APA computation the open-source library FFmpeg1 is employed for extracting audios from the original videos. Regarding the parameters of

FFmpeg, we choose two channels for the WAV format audio outputs, 44,100Hz for the sampling frequency, and 320kbps for the audio quality. The average memory cost of each audio file is about 2.7MB in Disk. Fig. 6 presents the wave forms of one sampled audio from its corresponding video.

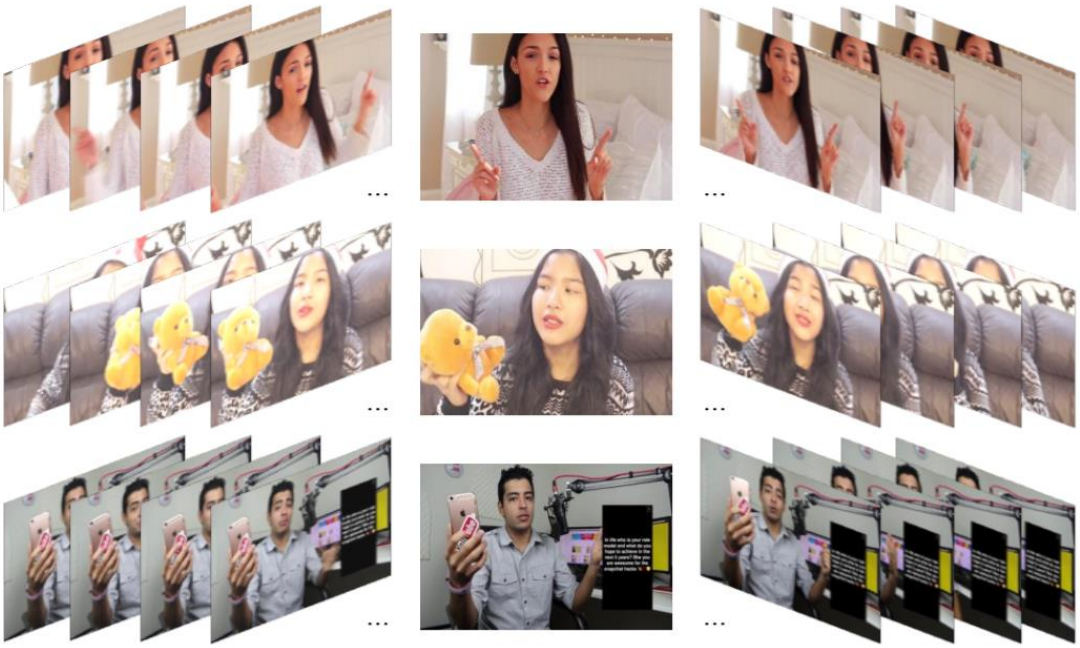


Figure 5. Examples of extracting images from videos. For each video, we extract about 100 images

5.3. Final Evaluation:

Software and Hardware Configuration:

Hardware

The hardware comprised of 8 GB DDR5 Random Access Memory, 1 TB Hard Disk Drive, 256 GB Solid State Drive and Intel Core processor i5 8th Generation which clocks at a speed 1.8Ghz

Software

The project is based on Python programming language. The interpreter is Python 3.9.0 run on Microsoft Visual Code with the following list of appended libraries for this project:

- absl-py==0.8.1
- astor==0.8.0
- ffmpeg==1.4

- gast==0.3.2
- google-pasta==0.1.8
- grpcio==1.25.0
- h5py==2.10.0
- Keras-Applications==1.0.8
- Keras-Preprocessing==1.1.0
- Markdown==3.1.1
- NumPy==1.16.4
- OpenCV-python==4.1.2.30
- pandas==0.25.3
- Pillow==6.2.1
- protobuf==3.11.1
- python-dateutil==2.8.1
- python-speech-features==0.6
- pytz==2019.3
- scipy==1.3.3
- six==1.13.0
- tensorboard==1.14.0
- tensorflow==1.14.0
- tensorflow-estimator==1.14.0
- termcolor==1.1.0
- Werkzeug==0.16.0
- wrapt==1.11.2

Results:

The final accuracy for the model is 0.9130 and accuracy for every trait is as follows:

- Extroversion: 0.9133
- Agreeableness: 0.9126
- Consciousness: 0.9166
- Neuroticism: 0.9100
- Openness: 0.9123

In the Final Evaluation phase, we directly employ the optimal models in the Development phase to predict the Big Five Traits values on the testing set. These optimal modes are logfbank features for audio modality and use of DAN+ architecture instead of alternative like ResNet.

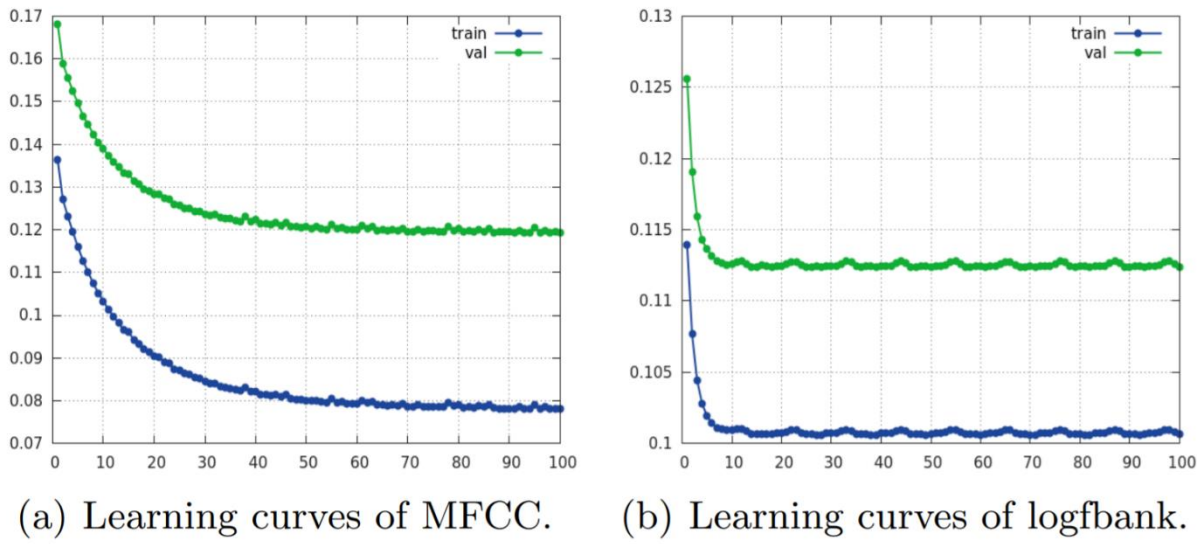


Figure 7. Learning curves of two different audio features, i.e., MFCC and logfbank. The horizontal axis is the training epoch, and the vertical axis is the regression error.

The architecture analysis in research across the decade shows that ResNet has trouble paying attention on the human and starts prioritizing the surroundings. DAN+ on the other hand manages both of the regions optimally. Moreover, for the regression accuracy of each Big Five Trait value, the proposed DBR framework achieved the best result in four traits. Since we just use the simple average method to do the late fusion, for further improving regression performance of the proposed method, advanced ensemble methods, e.g., stacking, can be used to learn the appropriate weights for the late fusion. Additionally, the deep audio networks should be tried to learn the more discriminative audio representations. The ensemble of multiple audio models can be also applied into the DBR framework to achieve better apparent personality analysis performance.

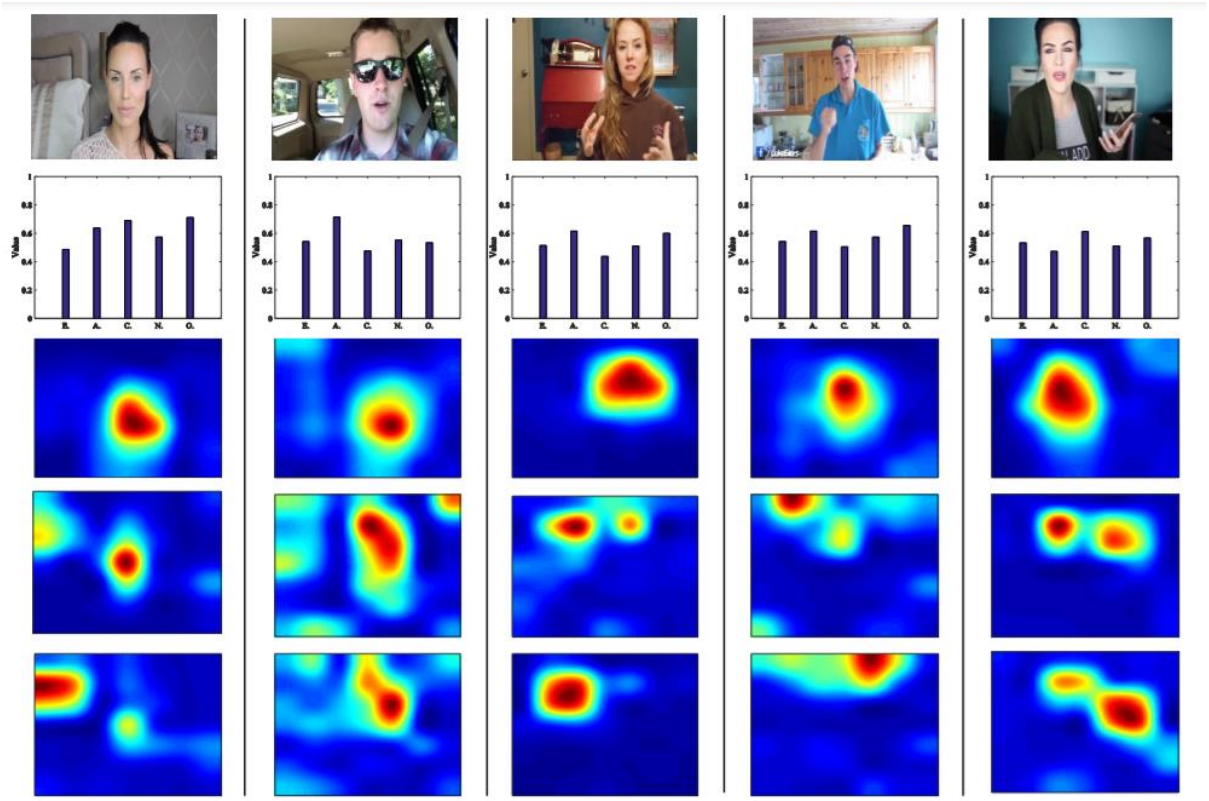


Figure 8. Feature maps of five sampled images in the visual modality of DBR. The first row shows the images, and the second row presents their corresponding Big Five Traits values. The third, fourth and fifth rows show the heatmaps of ResNet, DAN and DAN+, respectively. For each feature map, we sum the responses values of all the channels in the final pooling layer for each deep network. These figures are best viewed in colour.

6. Conclusions:

This solution creates a pipeline which emphasizes on independent analysis of the two modes in a video. The Deep Bimodal Regression Framework comes in very handy when encountering human-centred video analysis. The problem of modulation, vocabulary and tempo are efficiently dealt with.

7. Future Work:

In the future, we can introduce advanced ensemble methods into our framework and incorporating more discriminative deep audio representations for apparent personality analysis. Also, we can establish a neural network to determine the weightage of video and audio in analysis. Is direct dependency correct or should the effects of the input modes be modulated is a matter of some concern. Additionally, we can also incorporate NLP such that we can

words of importance efficiently. The best way to do so would be extract text from speech and run it through NLP pipeline with vectorizers to assign weightage to the speech. We can then add this to the model as a Trimodal Framework.