Predicting Job Hireability and Personality Traits

Harshit Malik
2017csb1078
Hersh Dhillon
2017csb1079

Indian Institute of Technology
Ropar Punjab
India

Abstract

Human beings have a tendancy to form judgements based on first impressions in the first 100ms of their interaction. This tendancy is particulary reflected in situations like first dates or job interviews where these first impressions can make or break the interaction.

We in this project seek to build and compare various Deep Neural Network models performance on prediction of Job Hirability and Big-Five personality traits: Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism (OCEAN) based on short video clips. We experimented with multiple models from all three modalities i.e. Video, Audio and Text on the First Impressions V2 dataset (CVPR '17).

1 Introduction

A first impression is the event when a person encounters an unfamiliar individual and forms a mental image about his or her personality based on apparent characteristics such as physical appearance, voices, body language and facial expression. One of the most well-known and commonly used personality model is the 'Big-Five' model which rates the five traits of Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (OCEAN). An automatic analysis of personality traits can help in automation of first interaction scenarios such as job interviews. Automation of the interviewing process can save both time and human efforts.

In this project we build some deep nerual networks to predict hireability score and five personality traits from various visual, audio and textual cues. We also compare performance of these models on hireability and personality prediction. We use models from all three modalities, Video, Audio and Text as well as models combining modalities.

A competition also took place with dataset in 2017. We also compare our performance with the competition submissions.

2 Dataset

2.1 First Impressions V2 (CVPR'17)

For the training and testing of our network for Job hirability and personality predic-

tion, First Impressions V2 (CVPR'17) (http://chalearnlap.cvc.uab.es/dataset/24/description/) is used. The first impressions V2 (CVPR'17) 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. Each clip has ground truth labels for big five personality traits and job hirabilty score represented with a value within the range [0, 1]. 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. Transcriptions of the audio is also available. 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).

2.2 Evaluation Metrics

The evaluation metric for these experiments is 1- MAE (Mean Absolute Error), referred to as accuracy. This is the evaluation metric used in the original competition also. Below is the formula used for accuracy calculation for trait t. here, N_t is total number of samples for trait 't', t_i is the ground truth score for trait t for the i th sample, p_i is the model prediction for trait t for the i th sample .

$$A = 1 - \frac{1}{N_t} \sum_{i=1}^{N_t} |t_i - p_i|$$

3 Video

3.1 Data preprocessing

We used the Open CV library to read the video files. For each video we see the total number of frames and we select 20 uniformly spaced frames from the video and we save these frames as images.

These frames are what we input to our models. We further apply transformations like resizing(according to the model requirements), normalization and random horizontal flip to the images. The targets to the models are taken from the annotation files and fed to the output of the model.

3.2 Models

3.2.1 2D CNN:- VGG 19

For 2D CNN analysis, 19 layered VGG 2D CNN model was used. Output layer with 1000 neurons was removed, and instead two hidden fully connected layer with 512, 64 neurons respectively were added along with final output layer of 6 neurons. Mean squared error (MSE) Loss was used during training on single frame (first frame of the video) with learning rate 1e-4 and batch size 64.

3.2.2 3D CNN:- ResNet18-3D

18 layered ResNet 3D model (https://arxiv.org/abs/1711.11248), with pretrained weights on Kinetics dataset was used for 3D CNN analysis. Output layer with 400 neurons was

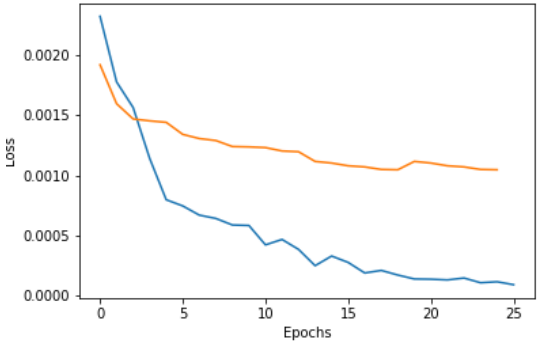

Figure 1: Loss curve for 2D CNN Model

removed, and instead two hidden fully connected layer with 128, 32 neurons respectively were added along with final output layer of 6 neurons. This model takes 16 stacked frames of size 112*112 each as input. Mean squared error (MSE) Loss was used during training on 16 uniformly spaced frames with faces, with learning rate 1e-4.

3.2.3 LRCN

Long Term Recurrent Convolutional Neural network (https://arxiv.org/abs/1411.4389) with a pretrained ResNet-50 (Imagenet) encoder and a single layer LSTM decoder, was used for the analysis of the video data. The model takes 20 images and the encoder CNN creates features for the frames of the video of size 512, for each frame which is fed to the LSTM decoder across different time frames. The output of the LSTM of size 512 is fed into a linear layer of size 256 which then then connected to the final layer which gives 6 output values.

L1 Loss was used for training the model, with learning rate for the pretrained ResNet being 1e-6 and the other untrained layers being between 1e-4 to 1e-5 (depending upon whether being learnt from scratch or fine-tuned) for different experiments. The optimizer used was Adam.

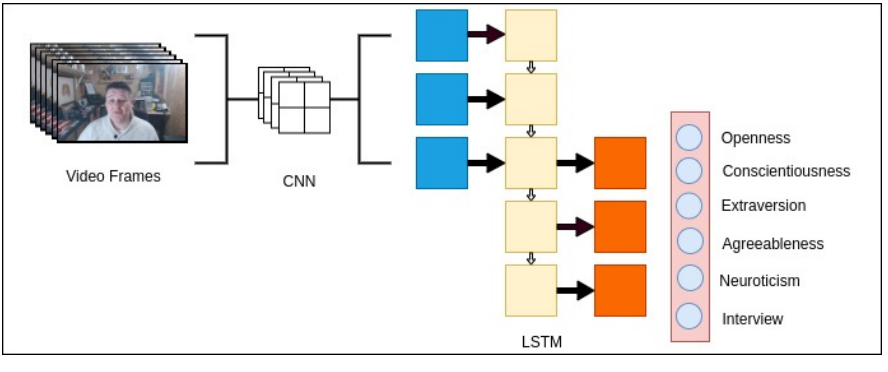

Figure 2: LRCN model representation

3.3 Results

Models	I	O	C	E	A	N
2D VGG19	0.9106	0.9048	0.9070	0.9044	0.9078	0.9019
3D ResNet18	**0.9140**	**0.9089**	**0.9148**	**0.9086**	**0.9096**	**0.9057**
LRCN	0.9075	0.9041	0.9032	0.9003	0.9070	0.8992

Table 1: Comparision of various models based on visual modality

(a) (b) (c) (d)

Personality	(a)		(b)		(c)		(d)	
traits	actual	pred	actual	pred	actual	pred	actual	pred
Interview	0.503	0.514	0.457	0.486	0.496	0.561	0.358	0.439
Extraversion	0.457	0.533	0.377	0.570	0.450	0.495	0.357	0.421
Agreeableness	0.534	0.549	0.503	0.615	0.525	0.703	0.426	0.615
Conscientiousness	0.543	0.563	0.532	0.369	0.542	0.641	0.367	0.544
Neuroticism	0.511	0.417	0.457	0.542	0.505	0.594	0.383	0.458
Openness	0.532	0.655	0.458	0.733	0.531	0.567	0.449	0.611

Figure 3: 2D CNN output for 4 different instances selected randomly

3.4 Inferences

- Video modality performs the best out of all three single modalities.
- We see that 2D CNN performed better than LRCN, given that 2D CNN was just used as a baseline model to compare other more complex models on to observe the contribution of the temporal aspect. We do get a temporal aspect of the data in 3D CNN and that performs better than 2D CNN but the results are close. Hence we can attribute this behaviour to two things. Firstly, people might form their first impressions very early, that is from the very first look at a person and secondly the LRCN model given it complexity might not be able to capture the differences well. They have been useful in tasks like activity recognition where the change between the frames is considerable, unlike the First Impressions Dataset.
- Interview score and Conscientiousness have the best scores among the 6 traits

4 Audio

Experimented with two approaches - audio features and audio spectrograms.

4.1 Features Extraction

For audio features, we used librosa library to extract 56 features that include mean and standard deviation values for mfcc(20 in number), energy, zero-crossing rate, tempo,spectral flatness, spectral bandwidth, spectral rolloff, contrast and tonnetz. For spectrogram generation from the audio files, we used matplotlib library.

MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale
Energy	Sum of squares of the signal values, normalized by the respective frame length
ZCR	Zero crossing rate of sign-changes of the signal during the duration of a particular frame
tempo	Beats per minute
spectral flatness	Spectral flatness is a measure to quantify how much noise-like a sound is, as opposed to being tone-like
spectral bandwidth	p'th-order spectral bandwidth, default p = 2
spectral rolloff	The frequency below which 90distribution of the spectrum is concentrated
spectral contrast	For each sub-band, the energy contrast is estimated by comparing the mean energy in the top quantile (peak energy) to that of the bottom quantile (valley energy).
tonnetz	Tonal centroid features

Table 2: Extracted audio features and their description

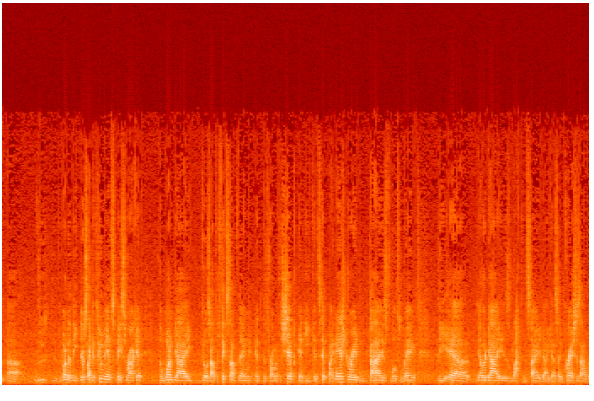

Figure 4: A sample generated spectrogram

4.2 Models

For audio feature analysis, random forest regressor was used as it gave better performance than other regressor. For the spectrogram analysis, 11 layered VGG 2D CNN model was used. Output layer with 1000 neurons was removed, and instead two hidden fully connected

layer with 512, 64 neurons respectively were added along with final output layer of 6 neurons. Mean squared error (MSE) Loss was used during training with learning rate 1e-4.

4.3 Results

Models	I	O	C	E	A	N
Audio Feature	**0.9044**	**0.9039**	**0.8986**	**0.9000**	**0.9055**	**0.9001**
Audio Spectrogram	0.8954	0.8963	0.8870	0.8917	0.8984	0.8915

Table 3: Comparison of audio on features and spectrogram based models

4.4 Inferences

- Audio model performed significantly better than text based models and close to visual cues based model.
- Model based on extracted audio features performed slightly better than model based on audio spectrogram.
- Openness and agreeableness seemed to be predicted slightly better than other personality traits in both the audio features and spectograms based models. It can be inferred that audio cues are good indicators of these traits as compared other personality traits.

5 Text

5.1 Data preprocessing

Data preprocessing was done in two ways

5.1.1 Bag of Words(BOW)

From the transcripts the stopwords were removed and we used 4 categories of words :- Adjectives, Adverbs, Verbs and Nouns to make up our vocabulary. From the the top 5000 most frequently appearing words were selected as feature vectors. This way each transcript is represented by a 5000 length vector and various algorithms are applied to those vectors.

5.1.2 Glove embeddings

We form a vocabulary of 15000 most used words in the transcriptions (among all transcription words, not limited to the word types used by Bag of Words approach) and use the glove embeddings for those words to make an embedding layer which will be used to feed into the LSTM Neural Network pipeline. We use Glove embeddings trained on 6 billion words with a 50 dim embedding for each word(glove6B50d).

5.2 Models

5.2.1 Random Forest Regression

For Random Forest Regression we use the features provided by the BOW approach, 5.1.1. We fit the training set multiple times while in order to fine tune the parameters using the validation set, in terms of max depth and criterion used. For the parameters that perform the best on validation set, we use them to predict the output for test set.

5.2.2 Support Vector Regression

For Support Vector Regression we again use the features provided by the BOW approach, 5.1.1. We fit the training set multiple times while in order to fine tune the parameters using the validation set, in terms of kernel and regularization constant used. For the parameters that perform the best on validation set, we use them to predict the output for test set.

5.2.3 LSTM Regression

For the LSTM regression we used the Glove embeddings as mentioned in 5.1.2. The LSTM model consists of a single layer LSTM decoder followed by 2 linear layers and a sigmoid layer to give an output between 0 and 1. Since we use 50 dim glove embeddings, that feeds the LSTM encoder and the embedding layer is frozen and not trained.

The learning rate used was 1e-6, with the loss function as L1 loss. The optimizer used was Adam.

5.3 Results

Models	I	O	C	E	A	N
RF	0.8824	0.8850	0.8816	0.8805	0.8941	0.8785
SVR	**0.8862**	**0.8874**	**0.8842**	**0.8831**	**0.8970**	**0.8815**
LSTM	0.8818	0.8834	0.8746	0.8779	0.8936	0.8774

Table 4: Regression accuracy of the Text Models

5.4 Inferences

- We see that the models on the text modality under performs considerably than it's Video and Audio counterparts, which is somewhat expected as the videos are really short for the content of the speeches to make much of a difference
- We also see that the LSTM model under performs the Random Forest Regressor and Support Vector Regressor. This is somewhat against what is usually seen. Even simple neural network models like the single layer LSTM outperform the more traditional ML approaches like BOW is most cases. This can probably be attributed to the lack of text data to train on as generally text datasets comprise of huge corpora of data.
- We see that out of all the traits, agreeableness seems to the most accurate for text.

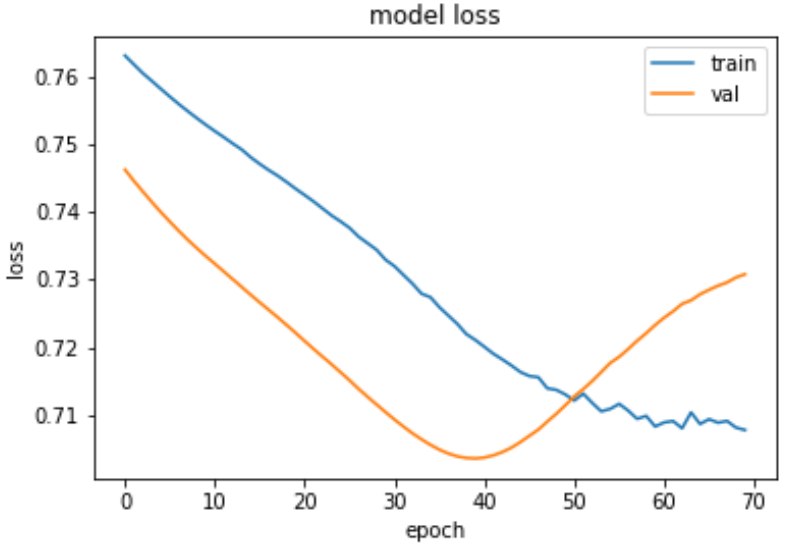

Figure 5: Loss curve for LSTM model

6 Video + audio (Feature Fusion)

6.1 Model

18 layered ResNet 3D model, same from 3D CNN analysis was used for extracting visual cues. Output layer with 400 neurons was removed to get 512 dimension feature vector. 11 layered 2D VGG model from audio spectrogram analysis was used for extracting features from audio spectrograms. First two fully connected layers were retained and modified to get 512 dimension feature vector. These two feature vectors are combined to get shared representation of dimension 1024. This 1024 shared representation vector is feed into two fully connected layers with 128 and 32 neurons each, along with final output layer of 6 neurons. Mean squared error (MSE) Loss was used during training and 16 uniformly spaced frames with faces were selected for ResNet 3D model, with learning rate 1e-4 and batch size 32.

6.2 Results

Model	I	O	C	E	A	N
Aud + Vid	0.9180	0.9102	0.9153	0.9150	0.9111	0.9100

Table 5: Video + Audio model output accuracy

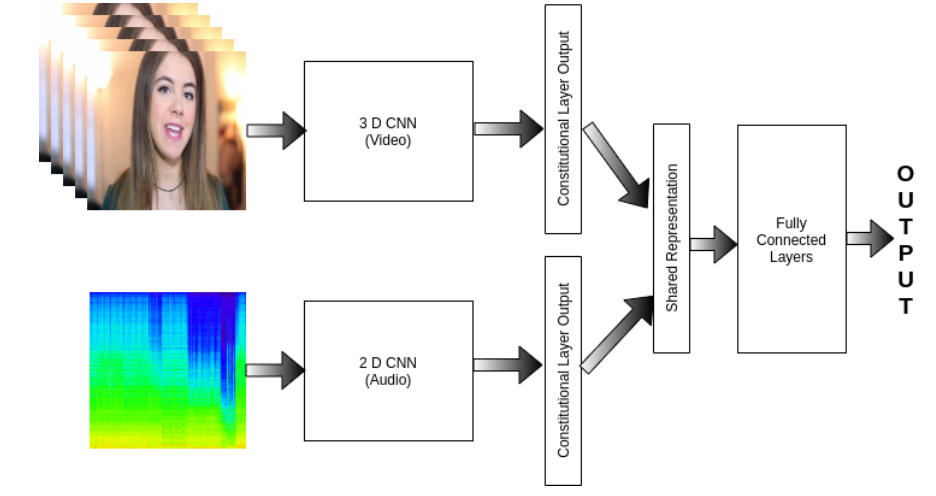

Figure 6: Video + Audio feature fusion model representation

6.3 Inferences

- Multi modality model, i.e video and audio feature fusion based model performed better than individual modalities namely visual, audio and text, based models. Here, 3D ResNet and VGG11 feature fusion model achieved better performance than their individual performances as stated in previous sections.

7 Comparison with the contest entries

Model	I	O	C	E	A	N
Heysky (Winner)	0.9209	0.9170	0.9197	0.9213	0.9137	0.9146
Aud + Vid (Ours)	0.9180	0.9102	0.9153	0.9150	0.9111	0.9100
Bekhouche(2nd Place)	0.9157	0.9100	0.9137	0.9155	0.9103	0.9083
3D ResNet18 (Ours)	0.9140	0.9089	0.9148	0.9086	0.9096	0.9057
Go2Chayan(3rd Place)	0.9019	0.9047	0.8950	0.9026	0.9032	0.9011

Table 6: Comparison with the competition entries

- We see that out best model beats the 2nd placed model in all but one category.
- Even our best performing visual cues based model does well, beating the third placed model in all categories
- The first placed system consists of two image models and one audio model and decision fusion using decision tree at the end[11]. They use faced cropped frames as well as the normal video frames along with a audio CNN to make their results. We show that even without the overhead of cropping faces, we can achieve a good accuracy score on the dataset.

8 Conclusion

- We compared various modalities (visual, audio, text) and found visual modality to be superior to other modalities with audio modality as the next most important.
- We show that that combination of modalities (Visual and audio) perform better than the single modalities themselves hence, visual and audio cues combined together are good indicators of hireabilty and personality traits scores.
- We also note an interesting observation that the 2D CNN performs really well for the dataset, almost comparable to the 2nd and 3rd place models in the contest. This suggests that people can frame their opinion at a glance and/or the similarity between the frames mean that it is difficult for models to capture the very subtle changes in the frames. This might even be an inherent issue with the dataset which is interesting and can be explored more.

9 Future Work

- Even though our models achieved appreciable accuracy scores, but they are not inherently explainable. Tasks like hireability and personality traits prediction need to be explainable in order to be accepted by people. Explainability of the models can be explored in future works which was also the second part of the competition.
- We observed combination of modalities achieved better performance than individual modalities.As a future work, text modalities and other annotations (age, gender and ethinicity) available with the dataset can be combined with visual and audio modalities.

References

[1] First Impressions V2 (CVPR'17), ChaLearn LAP
http://chalearnlap.cvc.uab.es/dataset/24/description/

[2] Prediction of Personality First Impressions With Deep Bimodal LSTM
http://cs231n.stanford.edu/reports/2017/pdfs/713.pdf

[3] Long-term Recurrent Convolutional Networks for Visual Recognition and Description.
https://arxiv.org/abs/1411.4389

[4] An End-to-end 3D Convolutional Neural Network for Action Detection and Segmentation in Videos.
https://arxiv.org/abs/1712.01111

[5] Torchvision Models
https://pytorch.org/docs/stable/torchvision/models.html

[6] Keras Models
https://github.com/keras-team/keras/blob/master/keras

[7] VGGFace2 Dataset for Face Recognition
https://github.com/ox-vgg/vgg_face2

[8] First impressions: making up your mind after a 100-ms exposure to a face.
https://www.ncbi.nlm.nih.gov/pubmed/16866745

[9] LONG SHORT-TERM MEMORY
https://www.bioinf.jku.at/publications/older/2604.pdf

[10] Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
https://arxiv.org/abs/1502.03167

[11] Explaining First Impressions: Modeling, Recognizing, and Explaining Apparent Personality from Videos
https://arxiv.org/abs/1802.00745