Prediction of Personality First Impressions With Deep Bimodal LSTM

Karen Yang
Stanford
450 Serra Mall, Stanford, CA 94305
kaiyuany03gmail.com

Noa Glaser
Stanford
450 Serra Mall, Stanford, CA 94305
noaglasr@stanford.edu

Abstract

From job interviews to first dates, a first impression can make or break an interaction. People form judgments in the first 100 ms of interaction [24]. Can an AI predict apparent personality traits given a short video?

In this work we propose a Deep Bimodal Regression LSTM model that extracts temporally ordered visual and audio features from a video clip to predict an average person 's first impression on video subject's Big-Five personality traits: Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism (OCEAN). Our model is trained and evaluated on HD Youtube videos provided by ChaLearn LAP APA2016 dataset [2].It achieves excellent performance that is comparable with the top teams in the 2016 competition on average accuracy, and outperforms these top teams on two categories: Openness and Conscientiousness.

1. Introduction

A first impression is the event when a person encounters another individual and forms a mental image about his or her personality based on apparent characteristics such as physical appearance, voices, body language, facial expression, and surrounding environment. According to psychology researchers, the first impressions are formed in limited exposure (as shorts as 100ms) to unfamiliar faces[24].One of the most well-known and commonly used personality model is the "Big-Five" model which rates the five traits of Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism (OCEAN)[7]. Evaluation of these personality has important applications in many domains such as human resourcing, computer assisted tutoring systems and user recommendation systems. An automatic analysis of personality traits can also help people to train themselves to prepare for important first interaction scenarios such as job interviews.

Estimating how personality traits are perceived by others is a complicated process. There have not been much studies on how first impression are formed, though a lot of research endeavor has explored similar computer vision tasks such as emotion recognition, face recognition, etc. However, we do not assume that facial features or emotion features are the only important factors. For robust modeling for this problem , we recognize the necessity of using a multi-modal system that utilizes general visual and audio features, as well as their temporal pattern. In this work, we pursue a novel approach that exploits CNN for visual features extraction, signal processing for audio feature extraction and LSTM for temporal modeling. Given a short HD Youtube video (about 15 sec long), the model extracts visual and audio features for temporal modeling and outputs five scores with range from 0 to 1, representing the estimated scores for Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism, respectively. Figure 1 demonstrates the input and output for our model.

In summary, the main contributions of our work are: (1) introduce the challenging problem of automatic apparent personality prediction from a short video. (2) propose a deep learning model that exploits both visual and audio modalities and their temporal patterns for robust estimation. The model achieves excellent performance that is comparable with the top teams in the ChaLearn LAP 2016 competition on the average accuracy, and outperform these top teams on two categories: Openness and Conscientiousness.

2. Related Work

Apparent personality traits prediction is an interesting yet challenging task. There have been several non-deep learning approaches that utilize different modalities such as speech[23], audio [21], [14], text [4], [15] and visual information [9][16]. To increase the robustness of predictions, multi-modal systems are also investigated [3]. Theses studies use traditional signal processing and statistical methods. For instance, Sardar et als study uses a logistic regression model with a ridge estimator for personality classification, in which they experiment with a fusion of audio-visual features, bag of word features, sentiment based and demographic features[17]. Other approached included

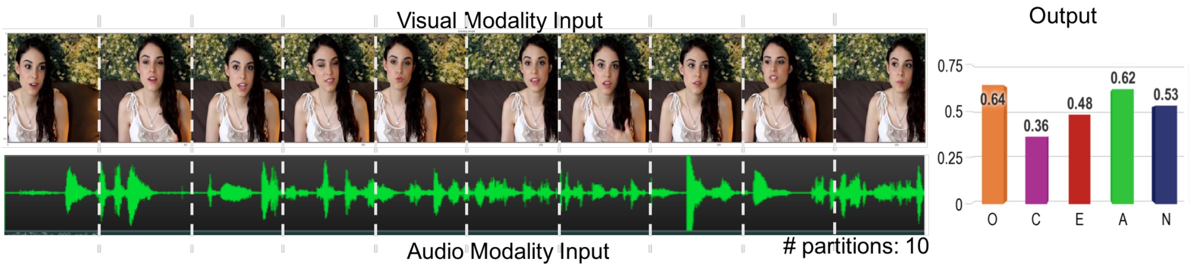

Figure 1. Left: Model input is one randomly selected frame (visual modality input) and raw audio wave (audio modality input) from each of 10 partitions ; Right: Model Output is five scores of predicted OCEAN personality traits in the range of [0,1]

variations of Support Vector Machine called SMO (Sequential Minimal Optimization for Support Vector Machine) , Bayesian Logistic Regression and Multinomial Nave Bayes sparse modeling[4].

Recently, predicting of apparent personality traits from online and media contents has attracted more and more attention. For example, in Vernon et al's paper [22], the authors looked how facial physical attributes in online facial photographs (such as chin length, head size, posture) can affect peoples impressions of approachability, youthful-attractiveness and dominance of them. By using a linear neural network, their model outputs predictions significantly correlated to the actual impression data. Since a photograph can impact the first impression judgments about a person, and that auditory information also influences the impression formation, audio-visual data fusion seems to be a suitable medium to study first impressions. [6] demonstrates that audiovisual cues enables the best prediction performance for their regression models compared to using single cue.

Given the recent success of deep neural networks and deep residual networks [11] in particular in many computer vision tasks, a deep learning approach for apparent personality recognition was also proposed. In 2016, ChaLearn Looking at People First Impression Challenge[] released a dataset of HD Youtube videos with annotations of Big Five personality traits impression data. All three winners of this challenge extensively used deep learning in their bimodal systems, while the overall approach and the type of the network was different [25][19][10].

The first place paper [25] created an ensemble (averaging) of an augmented VGG - which they call Descriptor Aggregation Network (DAN), augmented DAN (DAN+) and ResNet-152 on 100 frames of the video. They also used hidden layer neural network (NN) for audio modality regression. Each visual network achieved roughly 90% accuracy DAN 91% and ResNet 90.8% . The researchers found that each network was sensitive to different regions of the video, with ResNet more responsive to the face and DAN networks more responsive to background.

The second runner up team's model [19] was based on Residual Networks and temporal modeling. It randomly selects 6 frames from the video, cropped and 3D aligned the faces and ran the resulting images through an 3DCNN or LSTM. This paper showed that LSTM performed better then 3DCNN for temporal modelling and achieved around 91% accuracy.

Lastly the third place paper [10] used a ResNet-152 pre-trained on image-net data fed with randomly cropped frames concatenated with a randomly selected 3.1s sample of audio. They achieved similar accuracy to the first two papers.

The above winning teams used a combination of hand-crafted features and deep learning models. Since they are mostly research teams in universities, they were able to use multiple GPUs for end-to-end training of their models. Despite of our computation, memory and time limit (for example our single GPU instance can not run ResNet152), in this paper we will deploy the similar high-level idea by selecting suitable pre-trained feature extractors for bi-modal features and using appropriate temporal modeling to tackle this task.

3. Methods

In this section, we describe the architecture of the two models we implemented for this project. First, we use a ResNet34 model which performs the regression by extracting visual features from randomly selected video frame. After confirming that the residual network is a suitable visual feature extractor, yet not a sufficient model to perform the regression, we propose a Bi-Modal LSTM model (Figure 2) that encodes both visual and audio modalities with temporal modeling.

3.1. ResNet34 model

Our first attempt fine-tunes a CNN-based model. Specifically, we fine-tuned a 34-layer residual network[20] pre-trained on ImageNet [8]. The flattened 1×512 output from the final residual block is fed to two added linear layer of dimension 512×128 and 128×5, whose weights are learned.

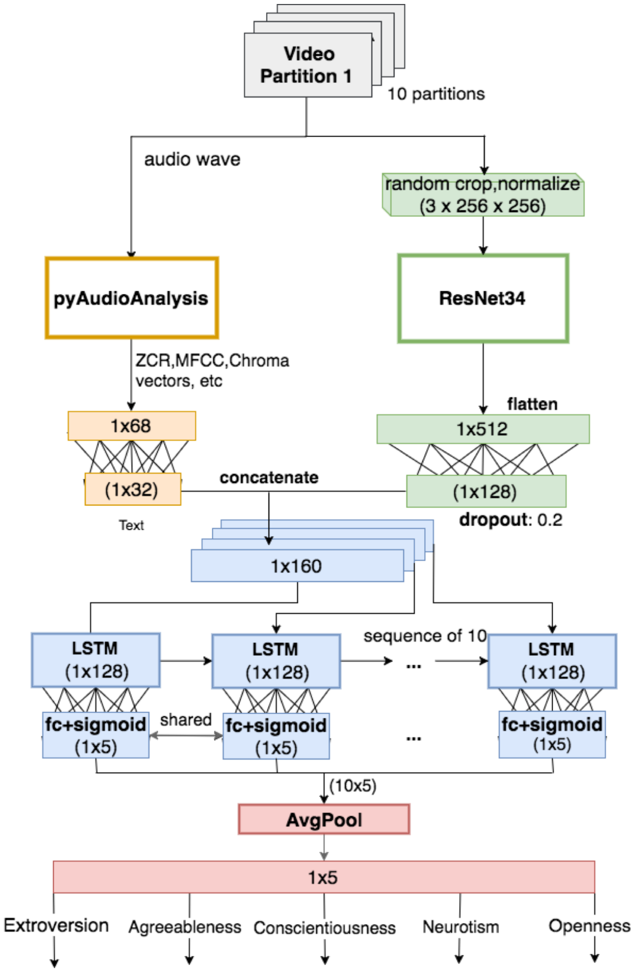

Figure 2. Deep Bi-modal LSTM Model Architecture.

We generate the $[0,1]$ regression scores by passing the output of the fully connected layers through a sigmoid layer.

For our ResNet34 model, each training batch was preprocessed as following:

- Load random videos frame from a mini-batch of size 32 images with pixel value whitened into a range of $[0,1]$
- Extract a random 256×256 pixels spatial crop of the video frame
- Randomly flip the crop horizontally with probability .5
- Normalize using $mean = [0.485, 0.456, 0.406]$ and $std = [0.229, 0.224, 0.225]$ as required by the pre-trained model

The ResNet34 pre-trained model[20] we used is adapted from He et. al's paper[11], obtained from Pytorch model zoo[20]. It contains one convolutional layer and 16 residual blocks of two convolutional layers. All convolutional layers are followed by batch normalization [[13]] and rectified linear units. The first layer is additionally followed by max pooling while the last also includes global average pooling. In the residual blocks that do not change the dimensionality of their inputs, identity shortcut connections are used. We used the 34-layer ResNet as apposed to deeper models due to memory and computation limits on a single GPU instance. We also had to limit batch size, which incurred some concerns in training.

This ResNet34 model is straightforward and acts as a proof-of-concept that, through tranfer learning, a residual network pre-trained on ImageNet is useful for our problem. As the ResNet34 is only imputed a single video frame out of entire video, it loses information potentially embedded in the audio attributes and the temporal connections between frames. Thus we presume it has weaker model generalizability. Next, we propose a Bi-modal LSTM model, in which the ResNet34 is used as the visual stream for visual feature extraction. In the second model, the ResNet is the same except that we truncate the last 128×5 and sigmoid layers in order to use the second last linear layer output as model the visual feature representation.

3.2. Bi-modal LSTM model

Our deep Bi-modal LSTM network for automatic apparent personality estimation has two branches, one for extracting audio features and the other for extracting visual features. The output from these two branches'output are concatenated in later stage as the input for each time step of the LSTM. Figure 2 demonstrates our model architecture. Inputs (See Section 4 for more details) to both the audio and visual branches of the model are generated after pre-processing the raw video data.

We split the video into 10 partitions and aligning the visual and audio data to build up the temporal relationship with a co-occurring frame and the audio profile. The choice of 10 partitions is arbitrary and is proven to work well. We think it is worthwhile experimenting with different input granularities as future work.

For our Bi-modal LSTM model, each training batch was pre-processed as following:

- For each video in a mini-batch of size 8, split it into 10 sequential partitions.
- Visual Modality
  - Load a random images from each video's 10 partitions (thus in total 80 images in a batch) with pixel value fit into a range of $[0,1]$
  - Random 256×256 pixels spatial crop, we did not apply random flip here to preserve the relative similarity between videos frames.

  - Normalize using $mean = [0.485, 0.456, 0.406]$ and $std = [0.229, 0.224, 0.225]$ as required by the pre-trained model
  - Generate a $80 \times 3 \times 256 \times 256$ as input batch to the ResNet34 to extract visual features.

- Audio Modality
  - Load the pre-processed raw .wav for each video's 10 partitions (thus in total 80 .wav file input in a batch) as the input for pyAudioAnalysis[1] to extract audio features. (See 4.3)

The visual branch contains the similar ResNet34 model described in Section 3.1. In the LSTM model, we only fine-tune one linear layer of dimension 512×128 and apply dropout of $p_{drop} = 0.2$ after generating video features to prevent over-fitting[18]. This visual stream outputs a 1×128 visual representation vector for each video frame.

The audio branch comprises of a signal processing stage. We used an open-sourced package pyAudioAnalysis[1] to extract a 1×68 audio attributes (See 4.3 for details) and then pass these features into a 68×32 linear layer. This audio stream outputs a 1×32 audio representation vector for each video partition.

Outputs of the audio branch and the visual branch are merged to a 1×160 vector representation for each of the video partition. From here, we trained a Long Short-term Memory network(LSTM)[12].The LSTM has one layer of 128 long short-term memory units and one layer of five linear units. Dropout was used to regularize the hidden layers. At each time step the LSTM will output a prediction for each of the five score based on the vector representation for the partition . The final prediction is the average of predictions over the 10 time steps .

3.3. Evaluation Metric

Following the criterion used in the ChaLearn LAP 2016 competition, for validation and test phrase, we evaluate our model in a supervised fashion using the average L1 loss:

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

4. Dataset and Features

4.1. Dataset

The first impressions data set comprises of HD YouTube videos (average duration 15s) of people speaking in English facing a camera. Videos subjects are vary in gender, age, nationality, and ethnicity [5]. Amazon Mechanical Turk (AMT) was used to label first impressions of the videos on the five personality traits on the range [0,1]. These five traits, Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness are known as the Five Factor Model (or Big Five) and is the dominant paradigm in personality research. AMT workers were also asked whether they would invite the video subject to an interview. 6,000 videos were used for training, split at a ratio of .8 for 4,800 training videos and 1,200 validation videos.

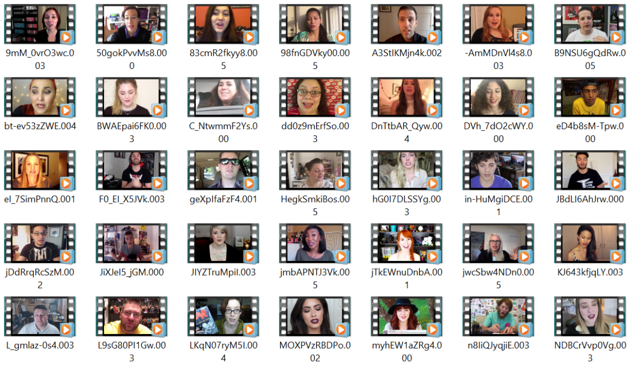

Figure 3. A look at a tiny sample of the First Impressions Challenge Data

4.2. Visual Data

We extracted frames for each video by splitting the video into 10 non-overlapping partitions of 1.5 seconds each and extracting a random frame from each of these partitions. Before the visual data is fed into the model, it is cropped to a resolution of 256x256 pixels. In training this is done by first scaling so that the smaller dimension is 256 pixels long, and then random cropping the other dimension. During validation and test, the scaling step is maintained but the crop is center-aligned. Each image channel is independently whitened.

4.3. Audio Data

Using the same splitting mechanism as described above, we extract the audio components from each of 10 individual partitions. Specifically, we extract the mean and standard deviation of audio signals as listed in Table 4.3.The time-domain features (ZCR, Energy and Entropy) are directly extracted from the raw signal samples. The frequency-domain features (all other features apart from the MFCCs) are based on the magnitude of the Discrete Fourier Transform (DFT). The cepstral domain (e.g. used by the MFCCs) results after applying the Inverse DFT on the logarithmic spectrum[1].

Attribute Name	Description
ZCR	Zero crossing rate of sign-changes of the signal during the duration of a particular frame
Energy	Sum of squares of the signal values, normalized by the respective frame length
Entropy	The entropy of sub-frames normalized energies, and can be interpreted as a measure of abrupt changes.
MFCCs	Mel Frequency Cepstral Coefficients form a cepstral rep- resentation where the frequency bands are not linear but distributed according to the mel-scale.
Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
Chroma Deviation	The standard deviation of the 12 chroma coefficients.
Spectral Centroid	The center of gravity of the spectrum.
Spectral Spread	The second central moment of the spectrum.
Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
Spectral Flux	Squared difference between the normalized magnitudes of the spectra of the two successive frames.
Spectral Rolloff	The frequency below which 90% of the magnitude distribu- tion of the spectrum is concentrated.

Table 1. Audio attributes extracted using pyAudioAnalysis [1]. The mean and standard deviation of each of the attributes are used, resulting in a audio modality input vector of 1×68 dimension.

5. Experiments

5.1. ResNet32 experiment

The pre-trained ResNet32 was obstained from Pytorch model zoo[20]. The weights for all the residual blocks are freezed. We truncate the last linear layer and add two linear layers of dimension 512×128 and 128×5 respectively, as well as a final sigmoid layer. This modified model is trained with batch size of 32 using SGD with a learning rate of 5e-3, momentum of 0.9 and weight decay of 5e-4.

5.2. Bi-modal LSTM experiment

The Bimodal-LSTM data was batch size of 8 sample videos, where each training sample in the batch includes 10 extracted frames from the video as well as 10 pre-extracted audio attribute data, thus there are in total 80 images and audio attribute input in a batch. After merging of the audio and visual feature outputs, the LSTM takes 1×160 vectors for 10 time points as its input. The linear layers at the end of visual and audio branches, as well as the LSTM were trained with SGD with a learning rate of 5e-5, momentum of 0.9, and weight decay of 5e-4. Because [19] trained using L2 loss, we tried to train the models using both L1 and L2 (followed by L1) loss. We reasoned that the different functions weigh different magnitudes of error differently - L1 is the final score for the competition but L2 can be less sensitive to smaller errors and could over-fit less in this case, given that our output and target values are within range $[0, 1]$.

6. Results

Table 2 shows the performance of our two models, as measured by the competition standard of $1 - \frac{1}{5N}\sum|\hat{y} - y|$, and compared to the top 5 teams in the ChaLearn challenge.

Of the models we trained, the Bimodal LSTM trained on L2 loss achieved hte best results, .908 overall accuracy as well as outperforming competitors in predicting Conscientiousness and Openness scores. The ResNet failed to appreciably train on the data. This is likely due to to two factors: firstly, it is a much simpler model and therefore has less explanatory power. Second, and is given much less data about each video - only 1 frame as opposed to 10, and without any audio information.

The training curve for the ResNet is shown in figure 4. From the validation curve it is clear that ResNet over-fit the training data. Adding dropout of .35 between the the last two layers of the ResNet, dropping the learning rate and several re-starts provided comparable results. We concluded that the ResNet34 model does not have sufficient explanatory power to capture the complicated model for apparent personality recognition, but can still be used as a good feature extractor.

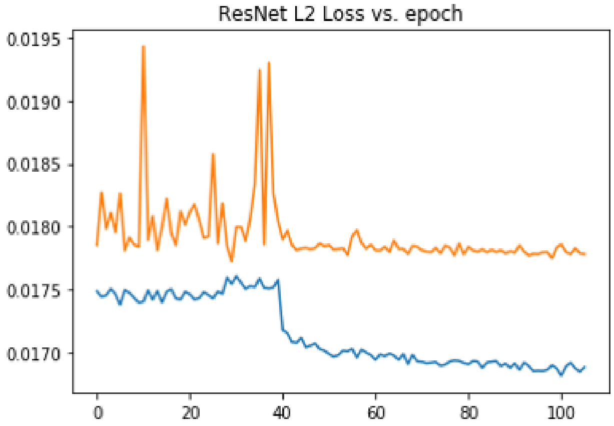

Figure 4. L2 loss from training (blue) and validation (yellow) at each epoch. This graph would suggest that the model lacks the explanatory power to be able to predict apparent personality.

Training the LSTM using an L1 criterion and L2 criterion produced similar results. The L2 trained model was able to perform slightly better, this could be due to a better starting point or the fact that the same learning rate was used for both (and L2 loss is about one order of magnitude smaller than L1 on this data.) The L2 training curve for the LSTM is shown in figure 5. Because the models were measured on L1 performance, we experimented with fine tuning the trained L2 checkpoints on an L1 criterion. We save the checkpoint of the trained L2 model and passed it through a short round of training using an L1 criterion and learn-

Evaluation Result						
	Total	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
LSTM L2	0.9083	0.9110	0.8944	**0.9220**	0.9005	**0.9136**
LSTM L1	0.8963	0.8977	0.8977	0.8941	0.9033	0.8888
ResNet	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

Table 2. Accuracy of our three models and the top 5 competitors in the ChaLearn challenge evaluated on the competition standard L1 loss. Accuracy is shown for each of the five traits as well as total accuracy over all of the data. From our models,the Bimodal LSTM trained using L2 loss model performed the best, outperforming all of the competition models on Conscientiousness and Openness.

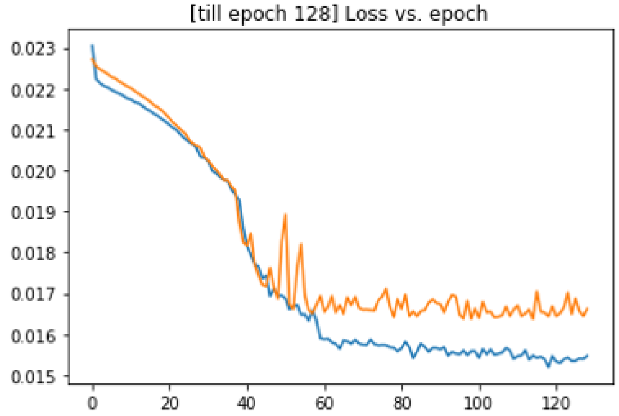

Figure 5. The LSTM L2 loss from training (blue) and validation (yellow) at each epoch. Learning rate was increased at epoch 38 to 5e-4, and decreased at epoch 59 to 5e-5. We see that the

ing rate of 5e-6. Figure 6 shows the corresponding test and validation loss curves. This fine tuning did not appreciably improve performance, suggesting that using an L2 criterion trained a model that performs well with both loss functions and that the two procedures are, in effect, comparable .

After training, we wanted to visualize what the model has learned. We examined saliency heat maps on the visual data from the two models.

The saliency heat maps for ResNet34 (figure 8) seems to indicate that the model is highly influenced by the subject's face.

As for the LSTM, from figure 7 we see that the LSTM is also highly influenced by the subject's faces - even in adversarial situations such as cropped or uncentered faces, yet also examines more of the background. The dimming effect in the last two rows in the figure is likely due to using average pooling of all of the LSTM outputs through time to determine the final score. Therefore, if frames at earlier time steps were impactful on proceeding scores, their influence magnified in determining the scores. This temporal relationship is consistent with modeling human first impressions and the idea that people form judgments withing one-tenth of a second.

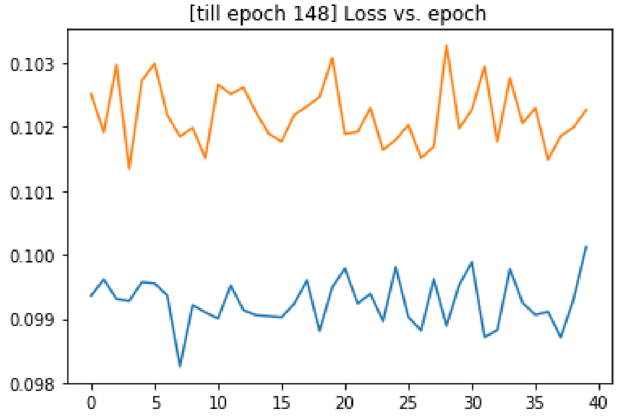

Figure 6. The last checkpoint of the LSTM model trained on L2 criterion was then fine tuned using training on L1 loss. We see from the resulting loss plots that the model had already arrived at state hat performed well for L1 loss and so no improvement occurred. Loss curves are shown are training (blue) and validation (yellow) at each epoch.

7. Conclusion and Future Work

We have achieved results comparable to the top five competitors in the international ChaLearn competition. It appears that the bi-modal LSTM is a good model for predicting first impressions of apparent personality.

It is interesting that both the ResNet34 and bimodal LSTM models learn were influenced mostly by the faces in the subjects video. Unlike [19] we did not make the assumption that the face is the most salient element in the frame. We reasoned that background, clothes or body could all be important factors in the scores. This data seems to suggest that it is indeed the face that is most significant.

It is also interesting that the top performers in the com-

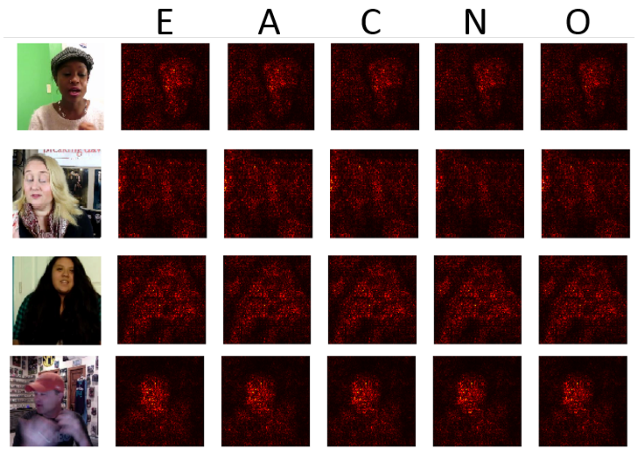

Figure 7. Saliency heat maps - magnitudes of gradients of each of the five scores predicted by the trained ResNet with respect to pixel data. Each row corresponds to an input image and each column corresponds to one of the five scores. We see that the ResNet trained to mostly be influenced by the subjects' faces.

petition all achieved 90-91% accuracy within a very narrow range. The tightness of the spread, and extent of the competition could suggest that there is a factor in the data limiting achievable accuracy. Scoring on apparent personalities is an inherently subjective task which is prone to personal bias and estimation. One could model this personal bias as random noise in the data. Flipside metrics could also be interesting, future work could explore how the accuracy achieved by large ensembles of models relates to levels of personal bias in data. It is also possible that the trained models learn latent prejudices in the data. The videos in this dataset are not labeled with gender or ethnicity, therefore this was not explored within the scope of this project. However, future work could examine these concerns.

Additional future work could also be targeted towards improving the model by experimenting with other visual feature extractors, such as VGG, or Inception Net as opposed to ResNet34. In addition, it is possible that more partitions in the RNN, i.e. improved time granularity, could yield interesting results. Our current framework of a a random frame every 1.5 seconds does not capture movement which could be modeled using other RNN design choices. In addition, an important factor in first impressions is what the subject is saying. We could add encodings of the video transcripts, extracted by speech recognition, in an effort to improve prediction accuracy.

We believe that this research has impact in generating hypotheses and a better understanding of how personality is perceived through analyzing the explanatory power of different models and their saliency region.

Besides being academically interesting, these insights could prove very valuable to many people. For example, many people do not have career help or interview preparation resources. Often, the technical information can be learned on-line, but soft skills require personal feedback and coaching. A model similar to ours could serve in a democratized tool which helps people train for job interviews by giving them feedback about how their soft skills might be perceived.

We are excited about the potential of this project in apparent personality trait recognition and look forward to upcoming future work.

8. Acknowledgement

We would like to thank CS231N instructors Fei-Fei Li, Justin Johnson and Serena Yeung , as well as TAs for their general guidance and support (Figure9). We would also thank Google Cloud for granting us education credits and GPU instances for our training use.

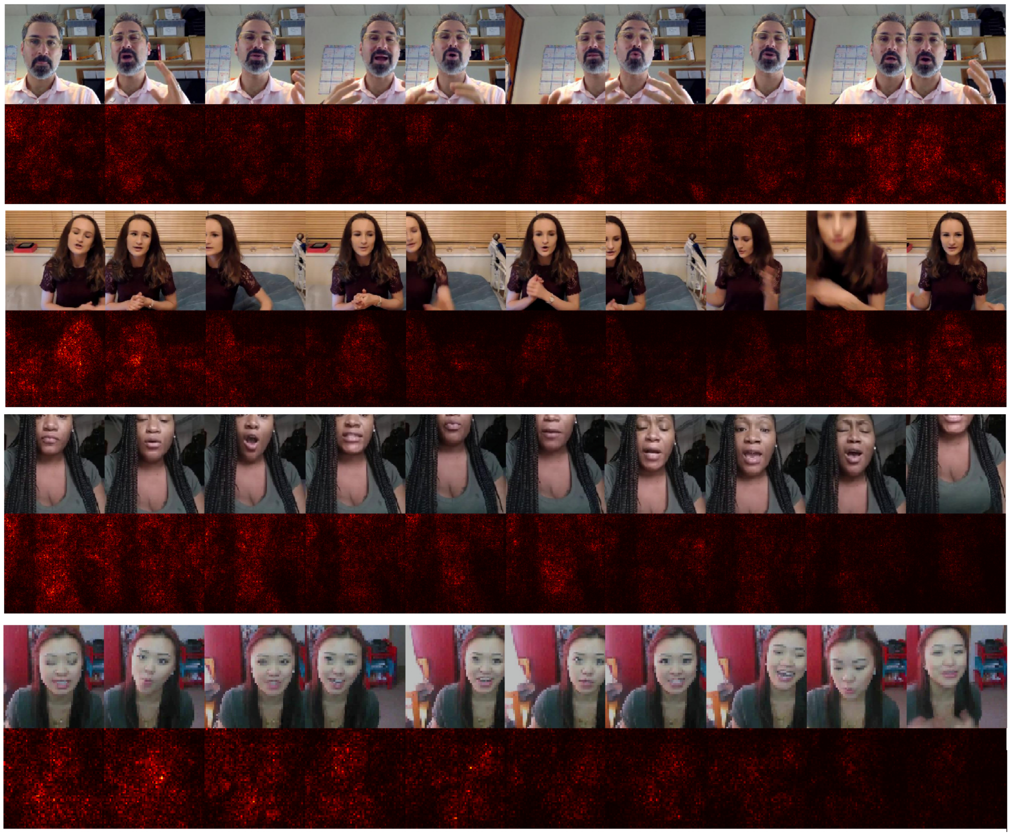

Figure 8. Saliency heat maps - magnitudes of gradients the extroversion score as predicted by the trained LSTM model with respect to pixel data. Each row of images corresponds the 10 frames from a videos.

References

[1] pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS ONE*, 2015.

[2] Open face dlib output, 2017. [Online; accessed May 10, 2017].

[3] F. Alam and G. Riccardi. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 15–18. ACM, 2014.

[4] F. Alam, E. A. Stepanov, and G. Riccardi. Personality traits recognition on social network-facebook. *WCPR (ICWSM-13), Cambridge, MA, USA*, 2013.

[5] J.-I. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on*, 15(1):41–55, 2013. Data for First Impressions V2 (CVPR'17) Dataset available at `http://chalearnlap.cvc.uab.es/dataset/24/description/#`.

[6] O. Celiktutan and H. Gunes. Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. *IEEE Transactions on Affective Computing*, 2015.

[7] E. R. Dahlen and R. P. White. The big five factors, sensation seeking, and driving anger in the prediction of unsafe driving. *Personality and Individual Differences*, 41(5):903–915, 2006.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[9] T. Fernando et al. Persons personality traits recognition using machine learning algorithms and image processing techniques. *Advances in Computer Science: an International Journal*, 5(1):40–44, 2016.

[10] Y. Gucluturk, U. Guclu, M. A. J. van Gerven, and R. van Lier. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. 2016.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] N. Madzlan, J. Han, F. Bonin, and N. Campbell. Towards automatic recognition of attitudes: Prosodic analysis of video blogs. *Speech Prosody, Dublin, Ireland*, pages 91–94, 2014.

[15] S. Nowson and A. J. Gill. Look! who's talking?: Projection of extraversion across different social contexts. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 23–26. ACM, 2014.

[16] R. Qin, W. Gao, H. Xu, and Z. Hu. Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face. *arXiv preprint arXiv:1604.07499*, 2016.

[17] C. Sarkar, S. Bhatia, A. Agarwal, and J. Li. Feature analysis for computational personality recognition using youtube personality data set. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 11–14. ACM, 2014.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[19] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. 2016.

[20] torchvision. Resnet pytorch pretrained model. `https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py`, 2017.

[21] F. Valente, S. Kim, and P. Motlicek. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus.

[22] R. J. Vernon, C. A. Sutherland, A. W. Young, and T. Hartley. Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, 111(32):E3353–E3361, 2014.

[23] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[24] J. Willis and A. Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, (17):592–598, 2006.

[25] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu. Deep bimodal regression for apparent personality analysis. 2016.

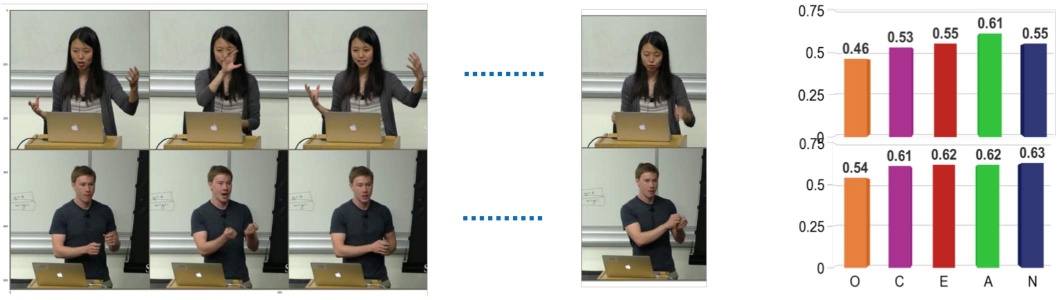

Figure 9. our mini unlabeled test set on our favorite lecture videos. We enjoyed the lectures and the class a lot more than this model! (Plus lecture videos are not really represented in the training set that consists of selfie videos.)