INVESTIGATING AUDIO, VIDEO, AND TEXT FUSION METHODS FOR END-TO-END AUTOMATIC PERSONALITY PREDICTION

Onno Kampman, Elham J. Barezi, Dario Bertero, Pascale Fung

Center for Artifcial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Introduction

- Automatic prediction of personality is important for the development of empathetic virtual agents.
- People infer personality from different cues, both behavioral and verbal. A model to predict personality should take language, speech, as well as visual cues into account.
- Personality is modeled with the Big Five personality descriptors (Goldberg, 1990). An individual's personality is defined as a collection of five scores in range [0, 1] for personality traits Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness to Experience.
- We predict personality from speech, language and video frames (facial features). We first consider the different modalities separately, to gain an understanding of how personality is expressed and which modalities contribute more. Then we analyze fusion methods to effectively combine the three modalities.

Corpus

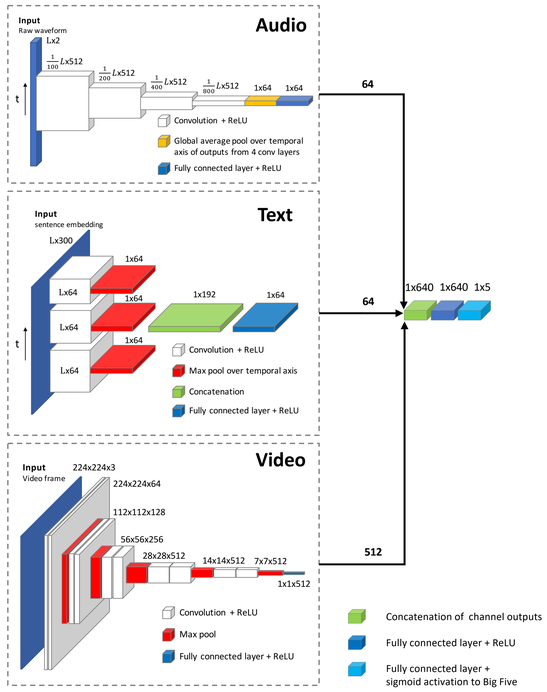
We used the ChaLearn First Impressions Dataset, which consists of YouTube vlogs clips of around 15 seconds. The speaker in each video is annotated with Big Five personality scores. The ChaLearn dataset was divided into a training set of 6,000 clips and 20% of the training set was taken as validation set during training to tune the hyperparameters, the early stopping conditions and the ensemble method training. We used pre-defined ChaLearn Validation Set of 2,000 clips as the test set.

Acknowledgement and Info

This work was partially funded by Hong Kong Research Grants Council, Innovation Technology Commission, and EMOS.AI.

Web: http://www.caire.ai
Email: ejs@ust.hk

Model Architecture



Methodology

Audio channel. The audio channel looks at acoustic and prosodic (non-verbal) information of speech. It takes raw waveforms as input instead of commonly used spectrograms or traditional feature sets. A stack of four convolutional layers is applied to the input, following by global average pooling operation.

Text channel. We extract word2vec embeddings from transcriptions and feed those into a CNN. Three convolutional windows are expected to extract compact n-grams from sentences. After this layer, a max-pooling is taken for the outcome of each of the kernels separately to get a final sentence encoding.

Video channel. We first take a random frame from each of the videos, which leads to personality recognition from only appearance, not temporal and movement information. We extract representations from the images using the VGG-face CNN model.

Multimodal fusion. We look at three different fusing methods to find how to combine the modalities best.

1. Decision-level fusion approach, a voting method: linear combination of weights for each modality, for each trait. We can read the relevance of the modalities for each of the traits , from the weights.
2. Merging the modalities by truncating each modality final fully connected layer, and concatenating the previous fully connected layers, to obtain shared representations of the input data. Finally we add two extra fully connected layers on top.

Methodology

- Limited backpropagation: all layers in the separate channels are frozen, to learn only what combination of channel outputs is optimal.
- Fully backpropagation: This enables the model to learn more complex interaction between the different channels.

Modality weights

	Big Five personality traits				
Model	E	A	C	N	O
Audio	0.44	0.32	0.27	0.45	0.54
Text	-0.03	0.22	0.13	0.03	-0.06
Video	0.59	0.46	0.60	0.52	0.52

Results

MAE		Big Five Personality Traits				
Model	Mean	E	A	C	N	O
Audio	.1059	.1080	.0953	.1160	.1077	.1024
Text	.1132	.1177	.0977	.1206	.1167	.1135
Video	.1035	.1040	.0960	.1087	.1064	.1024
DLF	.0967	.0970	**.0893**	.1049	.0979	.0947
NNLB	.0966	.0970	.0896	.1038	.0973	.0951
NNFB	**.0938**	**.0958**	.0907	**.0922**	**.0964**	**.0938**
Baseline	.1165	.1194	.1009	.1261	.1209	.1153

Conclusion

✓Language is the least relevant. Video frames (appearance) are slightly more relevant than audio (non-verbal) information.

✓Combining all modalities outperforms using individual modalities. Full backpropagation method obtains 9.4% better than the performance of the best individual modality.