Deep Impression

July 14, 2016

1 Team details

- Team name - *DCC*
- Team leader name - *Umut Güçlü**
- Team leader address, phone number and email - *Address: Montessorilaan 3, 6525 HR, Nijmegen, the Netherlands; Phone number: +31243611158; Email: u.guclu@donders.ru.nl.*
- Rest of the team members - *Yağmur Güçlütürk*, Marcel, A. J. van Gerven, Rob van Lier*
- Team website URL (if any) - *None.*
- Affiliation - *Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands*

  **equal contribution*

2 Contribution details

- Title of the contribution - *Deep Impression*
- Final score - *Final score was not released at the time of writing.*
- General method description - *A multimodal personality trait recognition model comprising an auditory stream (17 layer deep residual network) and a visual stream (17 layer deep residual network), followed by an audiovisual stream (one fully-connected layer) that is trained end-to-end to predict the big five personality traits of people from their short videos.*
- References - *1. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: "Deep Residual Learning for Image Recognition", 2015; arXiv:1512.03385. 2. Umut Güçlü, Jordy Thielen, Michael Hanke, Marcel A. J. van Gerven: "Brains on Beats", 2016; arXiv:1606.02627.*

- Representative image / diagram of the method - *See Figure 1.*
- Describe data preprocessing techniques applied (if any) - *Audio data were temporally resampled. Visual data were spatially and temporally resampled. No other preprocessing was performed.*

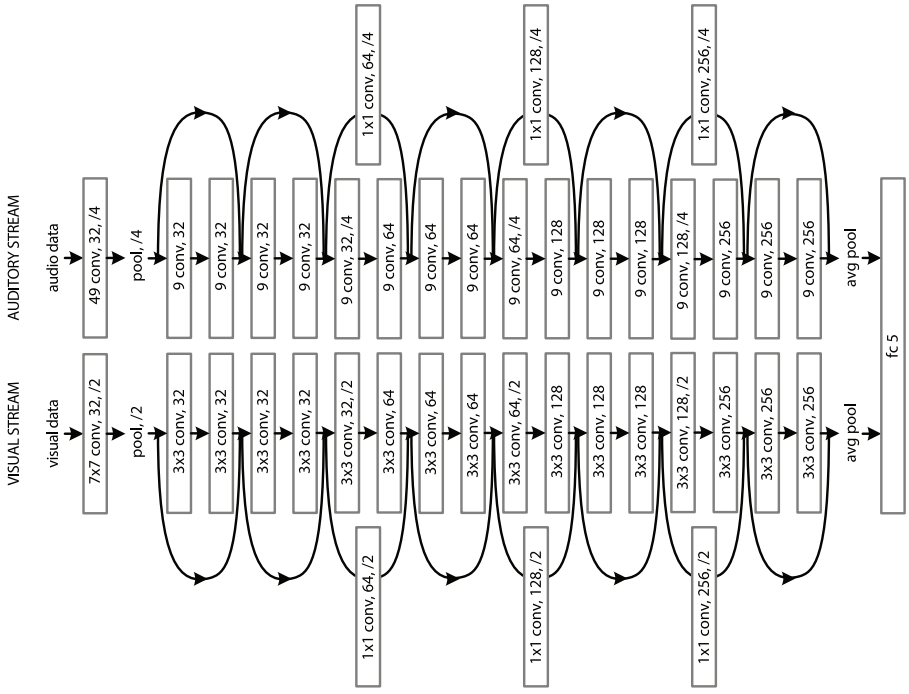

Figure 1: Illustration of the deep impression model.

3 Visual Analysis

None of the mentioned visual analyses (face detection, face landmark alignment or facial expression recognition) were performed.

4 Personality Trait recognition from Visual data

4.1 Features / Data representation

Raw visual data.

4.2 Dimensionality reduction

No dimensionality reduction was performed.

4.3 Compositional model

No compositional model was used.

4.4 Learning strategy

The visual stream and auditory stream of the model were jointly trained with mini-batch stochastic gradient descent by iteratively minimizing the mean absolute error between the target traits and the predicted traits.

4.5 Other techniques

Deep residual network with 2D kernels.

4.6 Method complexity

Method complexity is reported in terms of the time it takes to process a single example: approximately 25 milliseconds per training example and 1.35 seconds per validation/test example on a single GPU device of an Nvidia Tesla K80.

5 Personality Trait recognition from Audio data

5.1 Features / Data representation

Raw audio data.

5.2 Dimensionality reduction

No dimensionality reduction was performed.

5.3 Learning strategy

The auditory stream and visual stream of the model were jointly trained with mini-batch stochastic gradient descent by iteratively minimizing the mean absolute error between the target traits and the predicted traits.

5.4 Other techniques

Deep residual network with 1D kernels.

5.5 Method complexity

Method complexity is reported in terms of the time it takes to process a single example: approximately 25 milliseconds per training example and 1.35 seconds per validation/test example on a single GPU device of an Nvidia Tesla K80.

6 Multimodal Personality Trait recognition

6.1 Data Fusion Strategies

As described above, the multimodal personality trait recognition model comprised an auditory stream (17 layer deep residual network) and a visual stream (17 layer deep residual network), followed by an audiovisual stream (one fully-connected layer). The model processed a video as follows:

- The audio data and the visual data of the video was extracted.
- A random temporal crop of the audio data and the entire audio data were fed into the auditory stream in the training phase and the validation/test phase, respectively. The activities of the penultimate layer of the auditory stream were temporally pooled.
- A random spatial crop of a random frame of the visual data and the entire visual data were fed into the visual stream in the training phase and the validation/test phase, respectively. The activities of the penultimate layer of the visual stream were spatially pooled.
- The pooled activities of the auditory stream and the visual stream were concatenated and fed into the fully-connected layer.
- The fully-connected layer output five continuous prediction values between the range [0, 1] corresponding to each trait for the video.

6.2 Global Method Description

- Total method complexity: all stages - *Total method complexity is reported in terms of the time it takes to process a single example: approximately 50 milliseconds per training example and 2.7 seconds per validation/test example on a single GPU device of an Nvidia Tesla K80.*
- Which pre-trained or external methods have been used (for any stage, if any) - *None.*
- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any) - *None.*
- Qualitative advantages of the proposed solution - *End-to-end training, i.e. the model does not require any feature engineering or visual analysis such as face detection, face landmark alignment or facial expression recognition.*

- Results of the comparison to other approaches (if any) - *None.*
- Novelty degree of the solution and if is has been previously published - *To the best of our knowledge, the use of audiovisual deep residual networks for multimodal personality trait recognition is novel and has not been published.*

7 Other details

- Language and implementation details (including platform, memory, parallelization requirements) - *Language: Python in general and Chainer with CUDA and cuDNN in particular; Platform: Ubuntu 14.04.4; Memory: 64 GB main memory and 12 GB GPU memory; Parallelization requirements: None.*
- Human effort required for implementation, training and validation? - *Two persons for one week.*
- Training/testing expended time? - *Training and test on the provided training and test sets take 2.5 days and 1.5 hours, respectively.*
- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis? - *Apart from some technical issues, we were very satisfied with the challenge. We hope that there will be similar challenges in the future.*