Apparent Personality Trait Prediction with Audiovisual Deep Residual Networks

Yağmur Güçlütürk[1], Marc Pérez[2], Umut Güçlü[1], Xavier Baró[3,4], Baiyu Chen[9], Hugo Jair Escalante[5], Isabelle Guyon[6,7], Carlos Andujar[8], Rob van Lier[1], Marcel A. J. van Gerven[1], Julio Jacques Junior[2,4], and Sergio Escalera[2,4]

[1]Radboud University, NL. [2]University of Barcelona, ES. [3]Open University of Catalonia, ES. [4]Computer Vision Center, ES. [5]National Institute of Astrophysics, Optics and Electronics, MX. [6]University of Paris-Saclay, FR. [7]ChaLearn, US. [8]Universitat Politècnica de Catalunya, ES. [9]UC Berkeley, US.

Introduction

- We present a system that can mimic the way people form first impressions about the Big-Five personality traits of unfamiliar individuals.
- Big-Five personality traits are: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.
- The predictions that our system makes do not necessarily reflect the true personality traits of the individual, but the apparent personality traits.
- The system rather provides an opportunity for individuals to learn about what other people would think of them after a very brief interaction.

Overview of the System

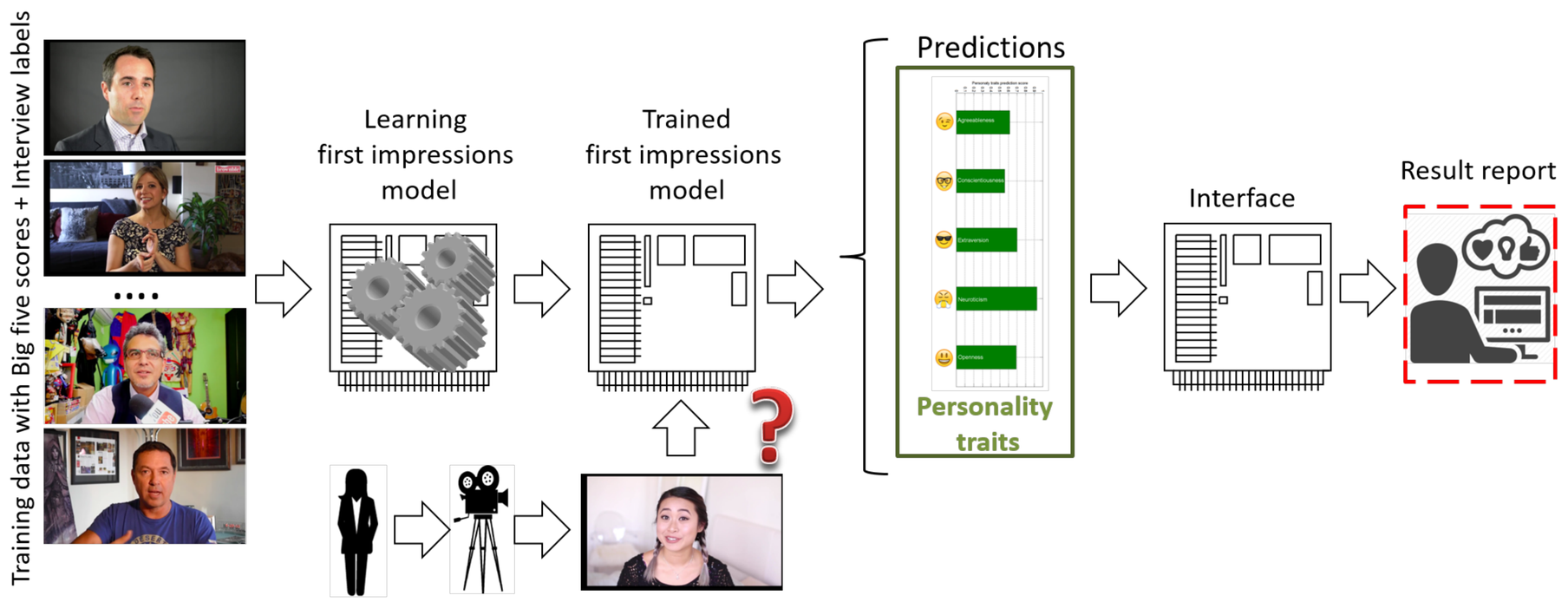

Figure 1: An on-screen avatar will guide the user throughout the demonstration. The user will briefly talk about themselves while being recorded by a camera. The video of the user will be analyzed by a deep neural network on a cloud server to predict the apparent Big Five personality traits of the user. Results will then be presented in a user friendly interface. Personality trait predictions of the user will also be compared to several job profiles as well as the apparent personality traits of well-known Machine Learning researchers.

Data

- ChaLearn First Impressions Challenge dataset [2].
- 10,000 15-second-long video clips from YouTube.
- Apparent Big Five personality traits annotations from Amazon Mechanical Turk.

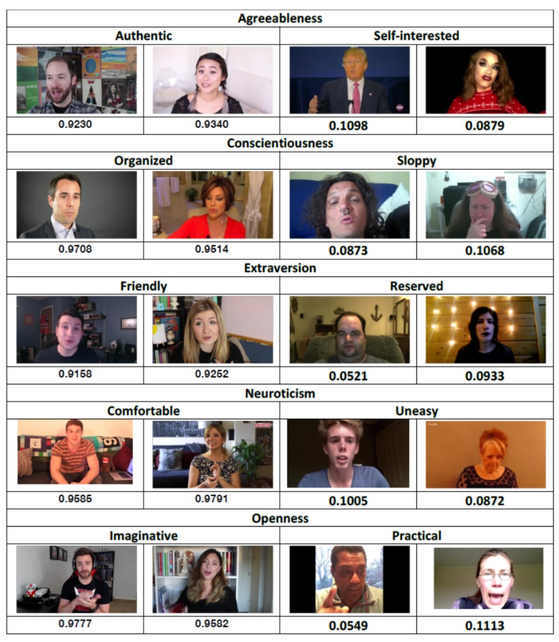

Figure 2: Screenshots from the videos of people perceived to have the highest and lowest levels of each trait.

Model

- Predicting the apparent Big Five personality traits of people from their short video clips.
- End-to-end training: No feature engineering, auditory analysis or visual analysis.
- Test accuracy (1 - MAE) of 0.9109.

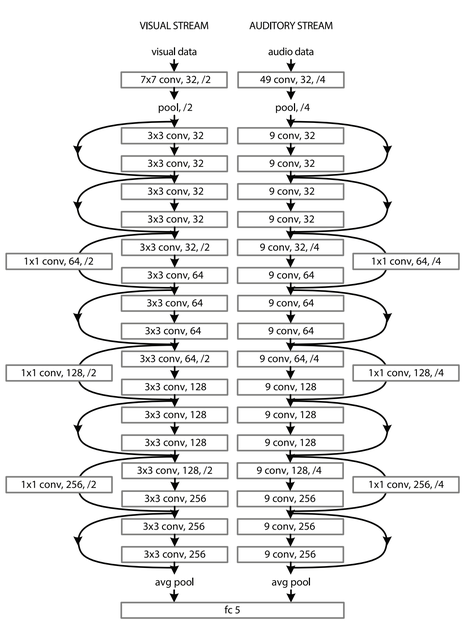

Figure 3: Audiovisual deep residual network comprising 17 layer auditory stream, 17 layer visual stream and one layer late fusion stream [1].

Visualization

- *Question*: What is driving model predictions?
- *One solution*: Occlusion analysis - systematically masking the inputs to the network and measuring the changes in predictions as a function of location or predefined region.

Pixel-level occlusion analysis

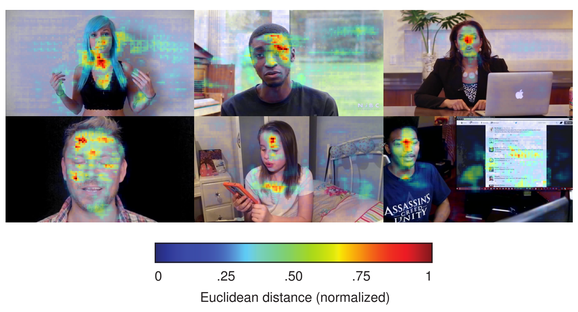

Figure 4: Visual pixel-level occlusion analysis. Each image shows the changes in trait predictions as a function of location resulting from systematically masking a representative example video overlaid on the input itself. Masks are defined as 10×10 pixels centered on every fourth point in the spatial axes. Change is defined as the Euclidean distance between the predictions before and after masking the videos.

Visualization

- *Rationale*: If a certain location or predefined region was driving the predictions, then masking it would either increase or decrease these predictions, enabling us to visualize the regions that had the most effect for the predictions of each trait.

Segment-level occlusion analysis

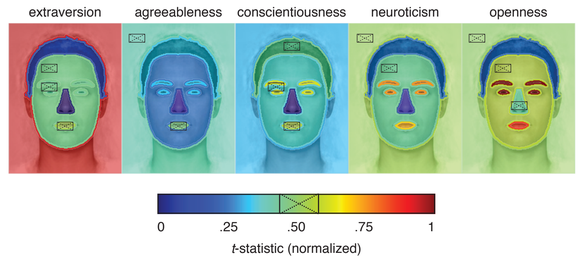

Figure 5: Segment-level occlusion analysis. Each image shows the changes in the prediction of the corresponding trait as a function of a predefined region resulting from systematically masking all videos overlaid on an average face. Masks are estimated with a separate deep neural network trained for segmenting faces to six regions. Change is defined as the effect size of the difference between the predictions before and after masking the videos.

Five factors

Additional Information

- Implementation is in Chainer with CUDA and cuDNN.
- Processing takes ~50 milliseconds per training example and 2.7 seconds per validation/test example on a single chip of an Nvidia Tesla K80 GPU accelerator.
- Implementation is available at github.com/yagguc/deep_impression

References

[1] Y. Güçlütürk et al. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. *Computer Vision – ECCV 2016 Workshops*, 2016.

[2] V. Ponce-López et al. ChaLearn LAP 2016: First round challenge on first impressions - dataset and results. *Computer Vision – ECCV 2016 Workshops*, 2016.

Contact Information

- http://demo.see4c.eu/traits
- http://chalearnlap.cvc.uab.es/
- http://www.ccnlab.net/
- http://www.socsci.ru.nl/robvl/