Interpreting CNN Models for Apparent Personality Trait Regression

Carles Ventura, David Masip, Agata Lapedriza

Outline

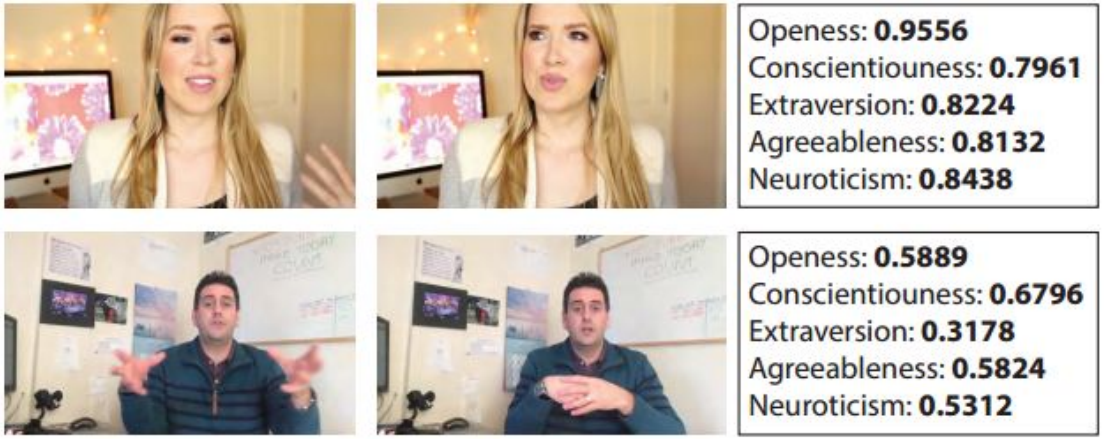
- Introduction
- Related Work
- Experiments
  - Images + audio vs Images for personality trait regression
  - Finding Discriminative Regions in video frames
  - Focusing on Faces
  - Interpretability of Face CNN
  - Action Units for Personality Traits Prediction
- Conclusions

Outline

- **Introduction**
- Related Work
- Experiments
  - Images + audio vs Images for personality trait regression
  - Finding Discriminative Regions in video frames
  - Focusing on Faces
  - Interpretability of Face CNN
  - Action Units for Personality Traits Prediction
- Conclusions

Introduction

- Problem: Automatic apparent personality trait inference
  - Big Five apparent personality traits
- Approach: Interpret CNN models
  - What internal representations emerge?
  - What image regions are more discriminative?



Introduction

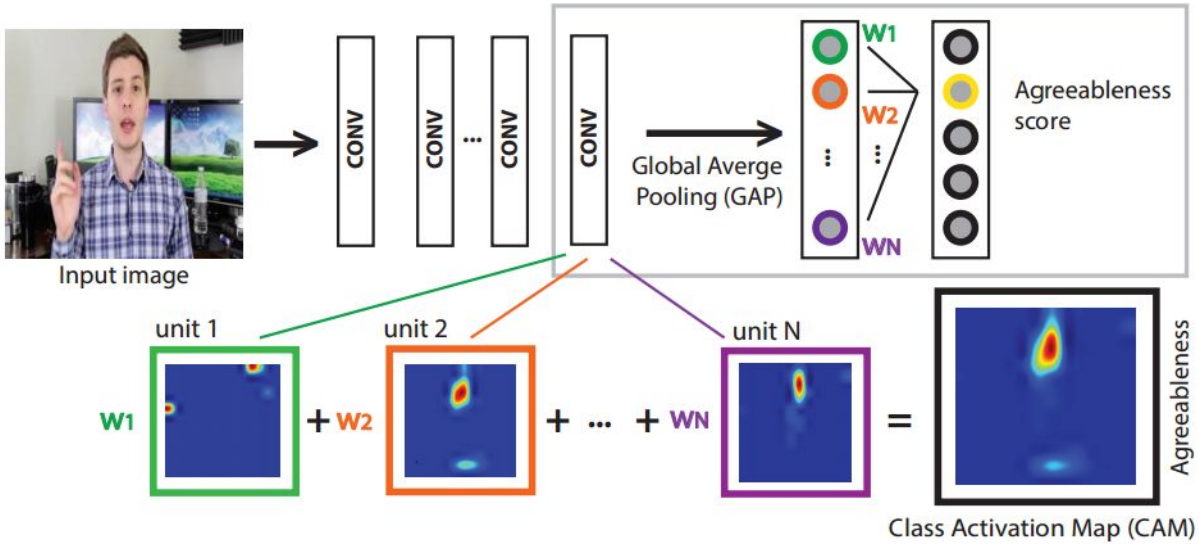
- Challenge: First Impressions dataset
  - Most recent and large database for apparent personality trait estimation
  - 10,000 video clips
  - Video frames, audio and captions available
  - Big Five personality traits annotated in a continuous 0-1 scale

Outline

- Introduction
- **Related Work**
- Experiments
  - Images + audio vs Images for personality trait regression
  - Finding Discriminative Regions in video frames
  - Focusing on Faces
  - Interpretability of Face CNN
  - Action Units for Personality Traits Prediction
- Conclusions

Related Work

- CNN models interpretability
  - Class Activation Map (CAM) [Zhou et al, CVPR'16]
    - Visualize class-specific discriminative regions



[Zhou et al, CVPR'16] "Learning deep features for discriminative localization."

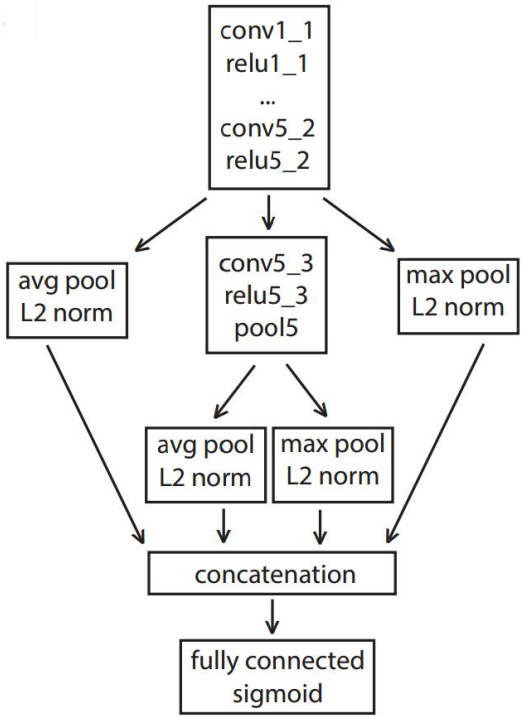
Related Work

- Deep learning architectures for personality trait regression
  - Fully Convolutional Neural Network (Zhang et al, ECCVW'16)
    - Winner last edition on First Impressions challenge
    - This architecture has been used as reference
  - LSTM Recurrent Neural Network (Subramaniam et al, ECCVW'16)
  - Deep Residual Network (Güçlütürk et al, ECCVW'16)

[Zhang et al, ECCVW'16] "Deep bimodal regression for apparent personality analysis."
[Subramaniam et al, ECCVW'16] "Bi-model first impressions recognition using temporally ordered deep audio and stochastic visual features."
[Güçlütürk et al, ECCVW'16] "Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition."

Related Work

- Fully Convolutional Neural Network (Zhang et al, ECCVW'16)
  - 2 models (images and audio) + late fusion
  - Model for images: DAN+
    - Extension of DAN (Descriptor Aggregation Networks)
    - Pre-trained VGG-face model
    - Average and max pooling at 2 different layers
  - Model for audio
    - Regression model over log filter bank features



[Zhang et al, ECCVW'16] "Deep bimodal regression for apparent personality analysis."

Outline

- Introduction
- Related Work
- **Experiments**
  - Images + audio vs Images for personality trait regression
  - Finding Discriminative Regions in video frames
  - Focusing on Faces
  - Interpretability of Face CNN
  - Action Units for Personality Traits Prediction
- Conclusions

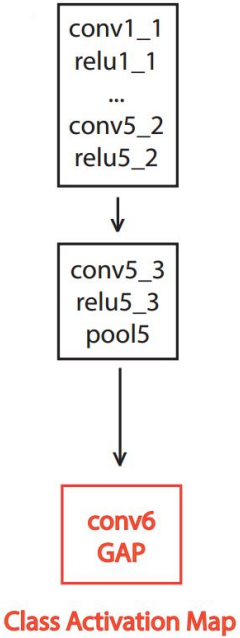
Experiments

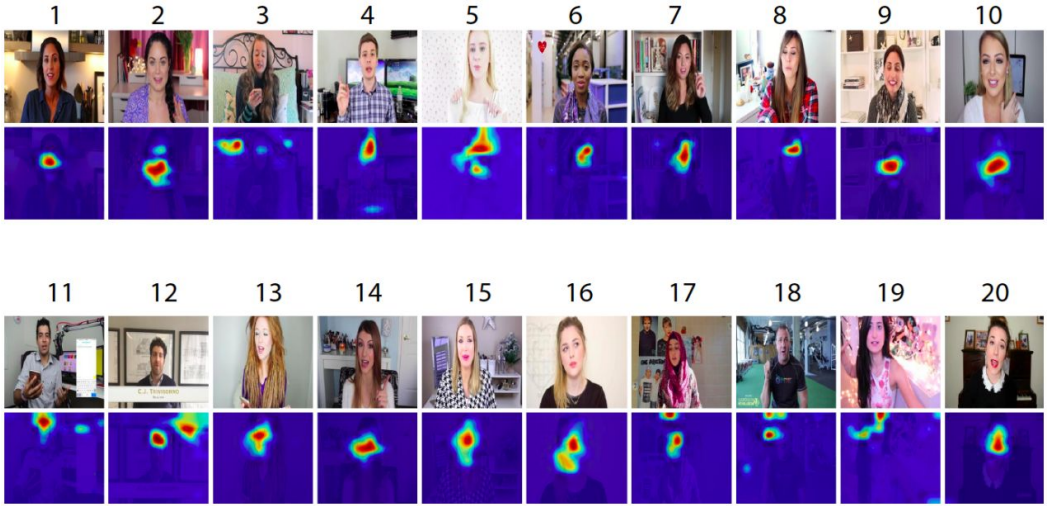
- 1. Images + audio vs Images for personality trait regression
  - Objective: Focusing only on image model interpretation
  - Accuracy of the models
    - Images (100 frames per video) + audio: 0.913
    - Only images (10 frames per video): 0.909

	Mean accuracy	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
img+audio	91.3	91.2	91.7	91.3	91.3	91.0
img	90.9	90.9	91.1	90.9	91.0	90.5

Experiments

- 2. Finding Discriminative Regions in video frames
  - CAM (Class Activation Maps) is applied to the image model

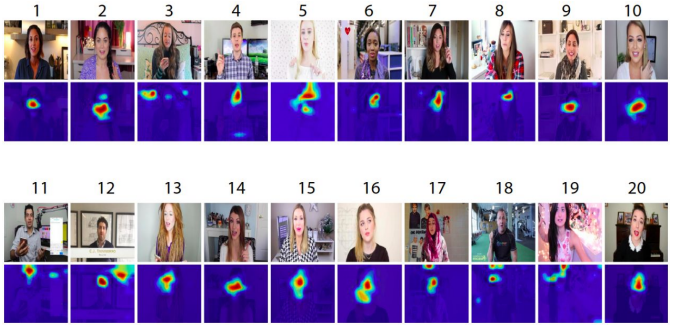




Discriminative localization for 20 images with highest predicted value for agreeableness

Experiments

- 2. Finding Discriminative Regions in video frames
  - CAM (Class Activation Maps) is applied to the image model
  - Discriminative regions mainly on faces regions
  - Quantitative evaluation
    - Face detection algorithm
    - Overlap of face bbox and CAM regions



$$overlap = \frac{M_{face} \cap M_{CAM}}{M_{CAM}}$$

  - Result: 72.80% of CAM regions have at least an overlap of 0.9 with the detected face

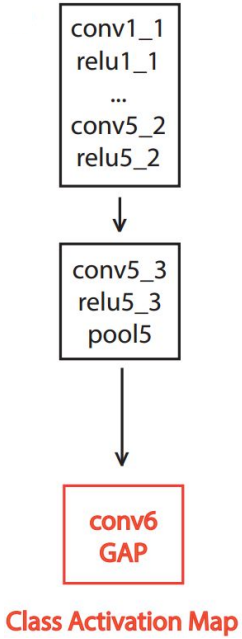
Experiments

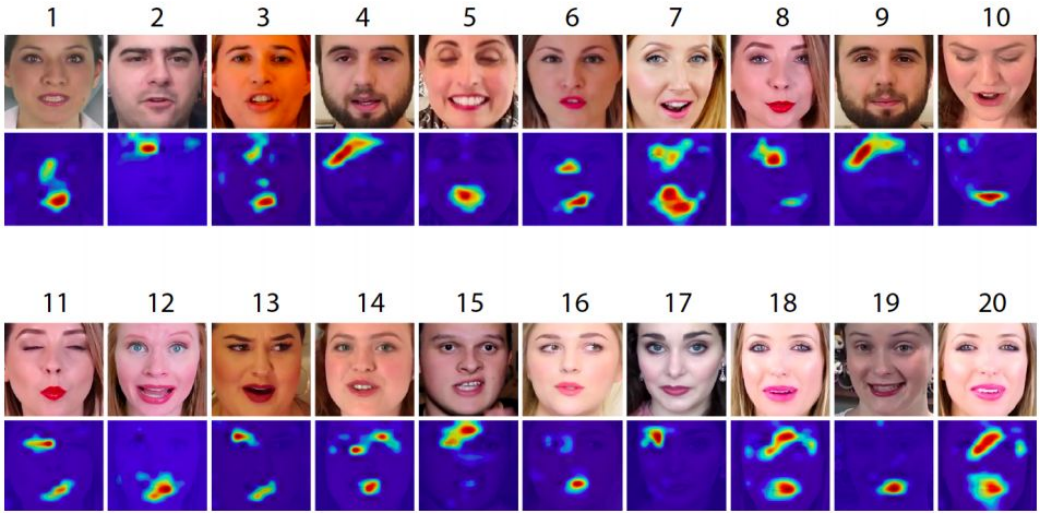
- 3. Focusing on Faces
  - Idea: Training the same architecture on cropped faces
  - Pre-processing:
    - Face region cropping
    - Eyes estimated localization for alignment
    - Image resize
  - Results:

	Mean accuracy	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
img	90.9	90.9	91.1	90.9	91.0	90.5
face	91.2	91.0	91.4	91.5	91.2	90.7

Experiments

- 3. Focusing on Faces: Finding Discriminative Regions
  - CAM (Class Activation Maps) is applied to the image model

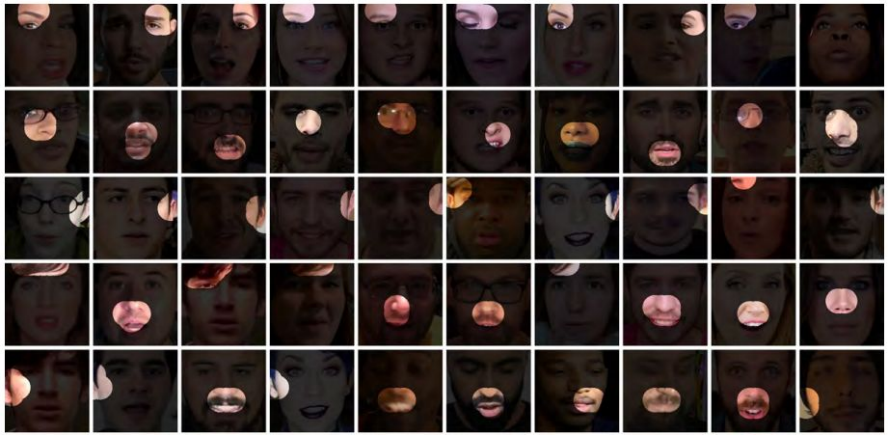




Discriminative localization for 20 images with highest predicted value for agreeableness

Experiments

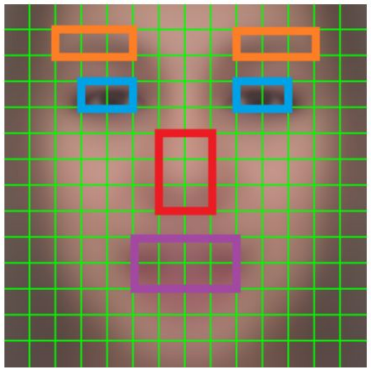
- 4. Interpretability of Face CNN
  - Goal: Visualize whether semantic detectors emerge from the network
  - Methodology (based on Zhou et al, ICLR'15)
    - Visualization of images that produce the highest activation given a unit of a layer
    - Images are segmented using an estimated receptive field



[Zhou et al, ICLR'15] "Object detectors emerge in deep scene CNNs."

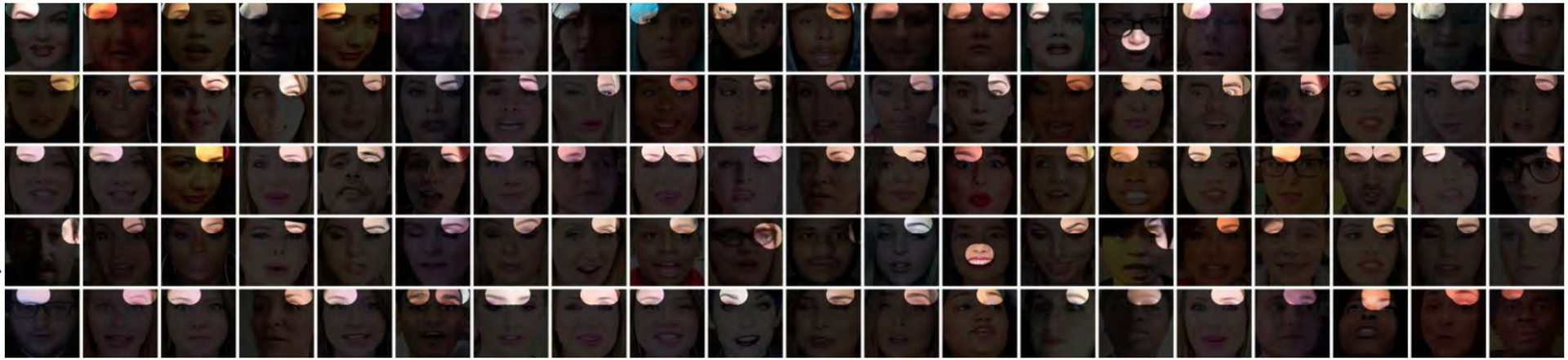
Experiments

- 4. Interpretability of Face CNN
  - Result: Semantic regions such as eyes, nose and mouth emerge
  - Previous methodology: manual inspection
  - New approach: automatic identification of emerging semantic detectors
    - Images are aligned
    - Semantic regions are defined
    - Spatial histograms from highest activations localization are computed for each unit of the CNN architecture
    - The addition of the spatial histogram values for a specific semantic region is applied to identify semantic detectors



Experiments

- 4. Interpretability of Face CNN
  - Eyebrow detectors



Experiments

- 4. Interpretability of Face CNN
  - Eye detectors

Experiments

- 4. Interpretability of Face CNN
  - Nose detectors

Experiments

- 4. Interpretability of Face CNN
  - Mouth detectors

Experiments

- 5. Action Units in Personality Traits Regression
  - Influence of shown emotion for personality trait
  - 17 Action Units (AU) from Facial Action Coding Systems
  - AU as 17-dimensional feature vector
  - Linear regressor trained on these feature vectors
  - Mean Accuracy: 88.6

	Mean accuracy
img	90.9
face	91.2
AUs	88.6

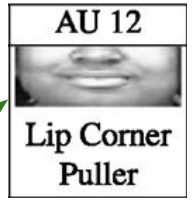
Experiments

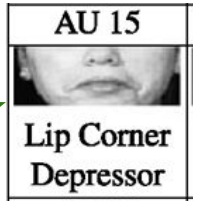
- 5. Emergence of Action Unit Detectors in Personality Traits Regression
  - Do AU detectors emerge from internal units of CNN model?
    - N frames with highest predicted intensity value for a given AU: $\{F_{AU}\}$
    - N frames with highest activation for a given internal unit: $\{F_{unit}\}$
    - Internal unit with highest intersection I_{max} between $\{F_{AU}\}$ and $\{F_{unit}\}$ is identified
    - Probability p to obtain I_{max} by chance is computed

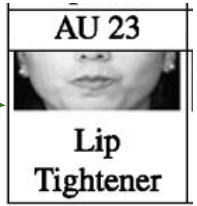
Experiments

- 5. Emergence of Action Unit Detectors in Personality Traits Regression

Action Unit	Most confident unit	intersection
AU1	159	6/50 (p<0.439)
AU2	344	7/50 (p<0.086)
AU4	35	6/50 (p<0.439)
AU5	261	7/50 (p<0.086)
AU6	380	6/50 (p<0.439)
AU7	7	5/50 (p<0.974)
AU9	350	6/50 (p<0.439)
AU10	397	7/50 (p<0.086)
AU12	108	10/50 (p<9.32e-5)
AU14	254	6/50 (p<0.439)
AU15	220	10/50 (p<9.32e-5)
AU17	77	6/50 (p<0.439)
AU20	74	7/50 (p<0.086)
AU23	475	9/50 (p<1.10e-3)
AU25	146	6/50 (p<0.439)
AU26	55	7/50 (p<0.086)
AU45	302	6/50 (p<0.439)







Outline

- Introduction
- Related Work
- Experiments
  - Images + audio vs Images for personality trait regression
  - Finding Discriminative Regions in video frames
  - Focusing on Faces
  - Interpretability of Face CNN
  - Action Units for Personality Traits Prediction
- **Conclusions**

Conclusions

- Interpretability of deep learning models for apparent personality trait inference
- Facial information was found to play a key role from discriminative region visualization
- Facial part detectors automatically emerged from last layers with no supervision provided on this task
- Influence of emotional information on trait prediction with the use of Action Units was explored


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
PALDIES
BOLZÏN
MERCI
BÏYAN
SHUKRIA
TINGKI

Experiments

- Action Units for Personality Traits Prediction
  - Influence of shown emotion for personality trait inference
  - 17 Action Units (AU) from Facial Action Coding Systems
  - Do AU detectors emerge from internal units of CNN model?
    - N frames with highest predicted intensity value for a given AU: $\{F_{AU}\}$
    - N frames with highest activation for a given internal unit: $\{F_{unit}\}$
    - Internal unit with highest intersection I_{max} between $\{F_{AU}\}$ and $\{F_{unit}\}$ is identified
    - Probability p to obtain I_{max} by chance is computed

$$p = P(intersection \geq k) = 1 - \left(1 - \sum_{i=k}^{N} \frac{\binom{F-N}{N-i}\binom{N}{i}}{\binom{F}{N}}\right)^{U}$$

Experiments

- Interpretability of Face CNN
  - Spatial histograms of the most frequent activation locations for each convolutional layer