Interpreting CNN Models for Apparent Personality Trait Regression

Carles Ventura, David Masip, Agata Lapedriza
Universitat Oberta de Catalunya
Barcelona, Spain
{cventuraroy,dmasipr,alapedriza}@uoc.edu

Abstract

This paper addresses the problem of automatically inferring personality traits of people talking to a camera. As in many other computer vision problems, Convolutional Neural Networks (CNN) models have shown impressive results. However, despite of the success in terms of performance, it is unknown what internal representation emerges in the CNN. This paper presents a deep study on understanding why CNN models are performing surprisingly well in this complex problem. We use current techniques on CNN model interpretability, combined with face detection and Action Unit (AUs) recognition systems, to perform our quantitative studies. Our results show that: (1) face provides most of the discriminative information for personality trait inference, and (2) the internal CNN representations mainly analyze key face regions such as eyes, nose, and mouth. Finally, we study the contribution of AUs for personality trait inference, showing the influence of certain AUs in the facial trait judgments.

1. Introduction

Humans continuously perform evaluations of personality characteristics of others. First impressions on personality traits, despite being inaccurate, play a crucial role in many essential decisions in our everyday lives, such as the results of the elections [1, 18], or court verdicts [6, 4]. These personality trait inferences are driven by informational cues with an evolutionary incentive[22].

This paper addresses the problem of automatic apparent personality trait inference. More specifically, we present a study on interpreting the representations learned by Convolutional Neural Network (CNN) models trained to regress personality trait scores from video frames.

The automated modeling of first impressions on personality has recently attracted the interest of the computer vision community (see Section 2 for an overview), specially since the publication of extensive databases and public challenges [15, 3]. In particular, we use in our work the First Impressions dataset [15], that is the most recent and large database for apparent personality trait estimation. This database is a collection of $10,000$ video clips of people facing and speaking in English. The videos are labeled according to the 'Big Five' apparent personality traits of the speakers. These five personality traits, scored in a continuous $0-1$ scale, are *Extraversion*, ranging from *friendly* (1) to *reserved* (0), *Agreeableness*, ranging from *authentic* to *self-interseted Conscientiousness*, ranging from *organized* to *sloopy*, *Neuroticism*, ranging from *comfortable* to *uneasy*, and *Openness to Experience*, ranging from *imaginative* to *practical*.

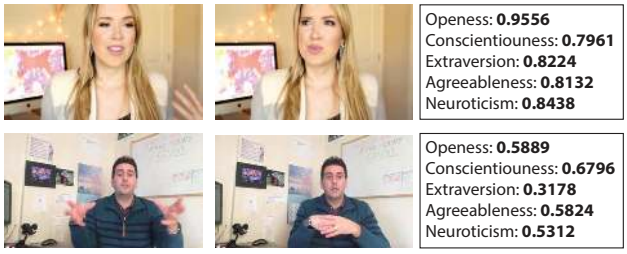

Figure 1. Apparent Personality Trait Regression (Big Five) from video frames.

In this paper we take as a baseline the work of Zhang et al. [23], which present the CNN model that a won the last edition of the ChaLearn Looking At People (LAP) 2016: First Round Challenge on First Impressions [15]. While Zhang et al. presented a CNN architecture that combined video frames and audio, we focus on video frames, discarding the audio information, and perform a deep study on the interpretability of CNN models trained exclusively from video frames to perform apparent personality trait regression. Figure 1 illustrates the problem, showing two video frames, randomly selected from two different videos, with the corresponding ground truth scores of the Big Five. More details on the Chalearn and the First Impression dataset are provided in Section 2.1.

Deep learning models are often perceived as black boxes able to learn complex non linear classification boundaries. Despite of their staggering success, the interpretability of

the resulting models is limited. Similarly, we know that humans infer personality traits in a few milliseconds [21], but the attention mechanisms that drive these decisions remain still unidentified. The knowledge of the regions of the image more influential for each inference [12] could be of extreme utility to shape these judgments.

In this paper we use the recent techniques presented in these works of Zhou et. al [24, 25] for visualizing the regions of the images that contain information for recognizing the apparent personality traits. The CNN architectures used in our study are described in Section 3. In a first quantitative study our results show that the informative region of the video frame clearly overlaps with the face of the speaker. We reproduce state-of-the-art results on a cropped training set, where only facial features are present. The visualization of CNN's unit responses reveal that specific facial feature detectors (i.e. eyebrows, eyes, nose, mouth) automatically emerge from the intermediate layers of the CNNs, although no training information was provided for these region's identification. Thus, even though there is a diverse set of informational cues (audio, video, text semantics and context), we experimentally show that, in the state-of-the-art model, the face accounts for the most part of the accuracy, and contains enough information to achieve the same accuracy for apparent personality trait regression. Finally, the last set of experiments present a discussion on the relation of Action Units and Personality Trait inference.

2. Related Work

Personality trait inferences based on first impressions were first studied in the field of Psychology. Their basis were stablished by several researchers performing factor analysis on textual data [5], which concluded in a model consisting on the aforementioned five traits [9]. In [13] authors focused the trait judgments on facial images. They used a data driven approach to model a basic set of traits, which were experimentally validated showing a strong correlation along two axis (dominance and valence).

These results attracted the attention of the computer vision community. Rojas et al. [16] proposed both an appearance (HOG descriptors) and a structural approach (based on distance among fiducial landmarks) to automate the inference of facial trait judgments. Biel et al. [3] introduced a first large video database where the personality traits could be identified, and showed a significant correlation effect between personality traits and facial expressions present in videos. Vernon et al. [19] proposed one of the first applications of neural networks to classify personality traits. They used 65 numeric attributes from 179 fiducial landmark points, allowing a fast identification of the key objective attributes for classifying social trait impressions. Joo et al. [11] proposed a ranking SVM modeling on intermediate features to predict facial trait judgments, and focused on the test of election outcome forecasting (strongly correlated with competence and trustworthiness traits).

Figure 2. Sample videos from First Impressions dataset that illustrates the different personality traits (figure from [15], pending permission).

2.1. ChaLearn and First Impressions Dataset

The ChaLearn First Impressions provides a large corpus of annotated videos [15, 8] and it is one of the most popular benchmarks for apparent personality trait inference. The First Impressions dataset [15, 8] consists of 10,000 clips extracted from more than $3,000$ different YouTube HD (720p) videos of people facing and speaking in English to a camera. These 10,000 clips are divided into three different subsets: $6,000$ clips for training, $2,000$ clips for validation and $2,000$ clips for testing. Each clip lasts 15 seconds. The videos are labeled according to the Big Five personality traits in a continuous $0-1$ scale. Figure 2 shows some video frames illustrating different personality traits scores.

In the past competition of this challenge, Zhang et al. [23] proposed a CNN architecture that discards fully connected layers, and aggregates convolutional layers using l2 normalized average and max-pooling. Subramaniam et al. [17] developed a bi-modal approach using both audio and image features. They used a LSTM recurrent neural network for end-to-end training. Güçlütürk et al. [10] used a Deep Residual Learning for learning the personality inferences, reaching the 3rd place in the Challenge. A more detailed review of these methods can be found in [15, 8].

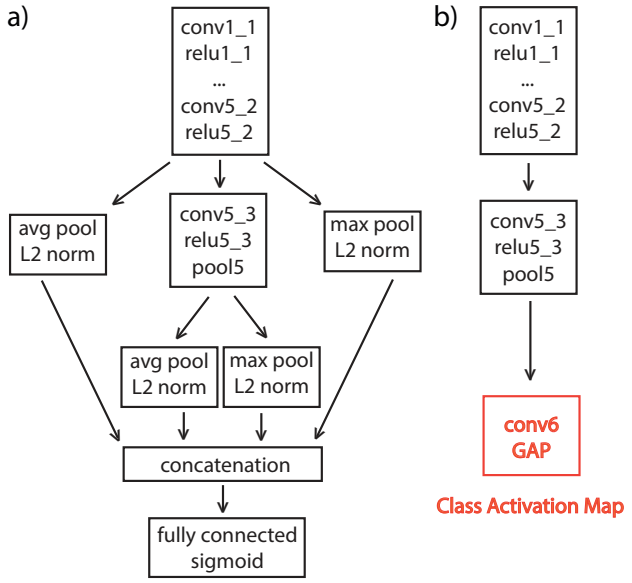

Figure 3. a) CNN architecture used for video modality in [23]; b) Our modification of the CNN architecture to add the Class Activation Map (CAM) module. More detailed information about CAM is given in Figure 4 and Section 3.1.

3. CNN Architectures

This work takes as a reference the CNN architecture proposed in [23]. This model, which was proposed by the team NJU-LAMDA, won the last edition of the ChaLearn Looking At People (LAP) 2016: First Round Challenge on First Impressions [15].

The architecture proposed in [23] consists of two separate models for still images and audio, processing multiple frames from the video and employing a two-step late fusion of the frame and audio predictions. However, since our work is focused on the interpretability of the video modality, the audio modality will not be considered and we use in this work only the model for still images.

The scheme of this model is shown in Figure 3.a. This architectures in named DAN+, and it is an extension of the Descriptor Aggregation Networks (DAN) [20]. In contrast to DAN, DAN+ applies max pooling and average pooling at two different layers of the CNN. These poolings are followed by L2 normalizations and the outputs are concatenated before feeding them to a fully connected layer. A pretrained VGG-face model [14] is used, replacing the fully-connected layers and fine-tuning the model with the First Impressions dataset.

To understand how DAN+ works to predict the personality traits, we make a modification at the last layers of the network, that allows to visualize the regions of the image that support the decision of the network. The scheme of this modified CNN is shown in Figure 3.b. Specifically, we substitute all the layers after the average pooling and the max pooling after pool5 by the Class Activation Map module proposed in [25] (summarized in next subsection). We also removed the average pooling and the max pooling after relu5_2 to avoid feature concatenation.

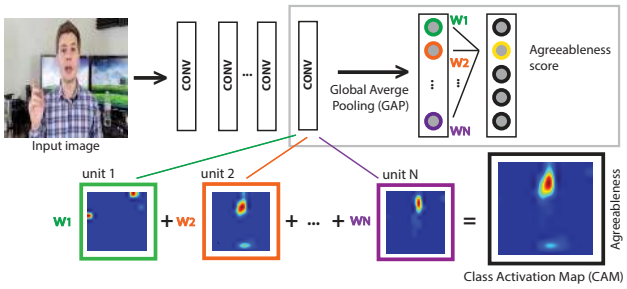

Figure 4. Class Activation Map [25] scheme.

3.1. Class Activation Map

In [25], the authors propose a simple technique, Class Activation Map (CAM), to visualize class-specific discriminative regions. The architecture of the CAM module is shown in Figure 4. This module follows the last convolutional layer of a CNN. It includes a Global Average Pooling (GAP) right before to the classification or regression layer.

As shown in Figure 4 the output of the last convolutional layer is a collection of 2D feature maps. Each map is composed of the response of a specific unit of the convolutional layer at the different locations. We denote by coordinate indexes $\{x, y\}$ the different 2D locations of these feature maps. Thus, given the i-th map, denoted in the figure by $unit_i$, the Global Average Pooling of this map results in the following feature

$$f_i = \sum_{x,y} unit_i(x, y) \tag{1}$$

With this set of $\{f_i\}_{i=1:N}$ features, the final classification or regression is done by

$$F(\sum_{i=1:N} w_i f_i) \tag{2}$$

where F is the classification or regression function and w_i is the learned weight corresponding to the feature f_i. For example, in Figure 4, F would be the regression function for $Agreeableness$. The interesting aspect of this architecture is that, additionally to the classification or regression response of the CNN, we can also have a visualization of the image region that highly supported the CNN decision. Specifically, this visualization corresponds to the weighted sum of the unit maps of the last convolutional layer, as shown in the bottom part of Figure 4.

4. Experiments

As previously stated, the winner CNN model for the last ChaLearn challenge on First Impressions combined video frames and audio. However, since our expertise is on image analysis, in this work we focus on the performance and interpretability of CNN models that deal just with visual information. In particular, the motivation of our experiments is to address the following questions: *(1)* Is it possible to obtain state-of-the-art results using just the video frames?, *(2)* Taking as input just video frames, what region of the image provides most of the information to the model?, and *(3)* Given a CNN model trained with the relevant visual information, can we interpret its internal representation? The following subsections address, sequentially, these three questions.

4.1. Video plus Audio vs. just Video

The original technique from [23] achieves an accuracy of 0.913 by combining video and audio modalities. This final accuracy is computed as the mean accuracy along each personality trait (for all input videos) between the predicted continuous values and the continuous ground truth values. Per each predicted continuous value, the accuracy is computed as $1 - d$, where d is the absolute distance among the predicted value and the ground truth value.

In this first experiment we use the same architecture from [23] but only over the video modality (see Figure 3.a) and with a lighter downsampling (only 10 frames per video have been used both in training and testing instead of the 100 frames per video from [23]). The obtained mean accuracy slightly decreases to 0.909. Notice that using just video frames we can achieve results dramatically close to the results obtained with the video and audio combination.

The results of the original technique and this first experiment are collected in Table 1 (first and second rows, respectively). Since the decrease in performance is very low, we performed the downsampling in all of our experiments to avoid computational cost.

4.2. Finding the Discriminative Regions in the Video Frames

We use the architecture of Figure 3.b to obtain the Class Activation Maps (CAM) for each personality trait. These CAMs are generated for the 50 images that give the highest predicted value for each personality trait. Figure 5 shows the CAMs obtained for the agreeableness personality trait. Notice that the higher support regions (in red) clearly overlap with the face of the person. After observing CAMs for the different personality traits, we noted that the network mainly focuses on the faces region areas to discriminate among the different personality traits and predict their values.

Figure 5. Discriminative localization (Class Activation Maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

We evaluated this observation quantitatively by applying a face detection algorithm [2] and computing the overlap between the bounding box of the face detected and the highest activated area of the CAM. More precisely, let us denote M_{face} the mask that corresponds to the bounding box of the face detected and M_{CAM} the mask that results from binaryzing the CAM using as threshold $0.8N$, where N is the maximum value that CAM takes. Then, the overlap between M_{face} and M_{CAM} is computed as:

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

The reason why we compute the overlap as a recall measure instead of as a Jaccard measure (intersection over union) is that we want to penalyze the region areas of M_{CAM} that are outside the M_{face} but without penalyzing the fact that M_{CAM} may not cover the whole M_{face}.

As a result of computing the overlap between the bounding boxes of the detected faces and the CAMs we obtain that 72.80% of the CAMs have at least an overlap of 0.9 with the detected face. The average overlap is 0.76. With this, we conclude that the face provides most of the discriminant information. This result motivates the experiments presented in the next subsections, that focus on the face.

4.3. Focusing on Faces

Once the CAMs revealed that faces are the most discriminative regions of the video frames, we trained the same CNN architecture just on the cropped faces, instead that training on the whole video frame. The faces are cropped using the same face detector as in previous section [2]. The estimated location of the eyes (given by the same detector) is used to align and normalize all the images from the dataset, crop the faces and resize them to 224×224 pixels. The average image is recomputed over the train and validation subsets before retraining the model.

The result of retraining the model focusing only on faces and discarding all the context (i.e. the background) is a mean accuracy of 0.912, which slightly outperforms the

Meth	MA	O	C	E	A	N
[23]	0.913	0.912	0.917	0.913	0.913	0.910
img	0.909	0.909	0.911	0.909	0.910	0.905
face	0.912	0.910	0.914	0.915	0.912	0.907

Table 1. Results for our image baseline approach (img, see Section 4.2), our face approach (face, see Section 4.3) and original approach from state-of-the-art [23]. MA refers to Mean Accuracy and O, C, E, A and N to accuracy for each personality trait: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism, respectively.

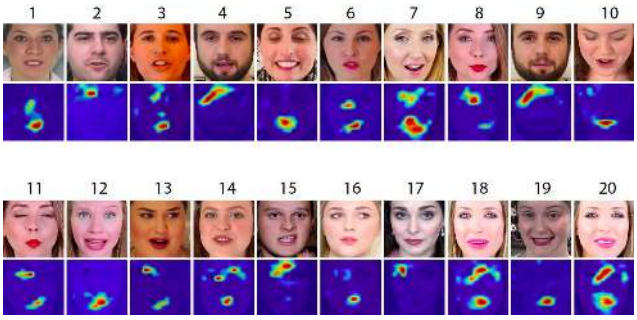

Figure 6. Discriminative localization (class activation maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

mean accuracy over the model trained using the whole image (i.e. with both faces and context). Table 1 (third row) shows the results for each personality trait and compares it with our baseline approach from Section 4.2 as well as the state-of-the-art technique [23]. From this table, we can also observe that the improvement in the prediction is along all five personality traits. Furthermore, the mean accuracy is almost as good as the original approach from [23], which uses both video and audio modalities.

We also use CAM to visualize, in this case, the regions of the face that provide the most discriminative information. Figure 6 shows again the CAMs obtained for the agreeableness personality trait, but now just on the face of the person.

From the visualization of the CAMs for the different personality traits, we can observe that the network mainly focuses on the faces region areas corresponding to the eyes and the mouth to discriminate among the different personality traits and predict their values. However, there are no significant differences between the CAMs resulting from the different personality traits. Therefore, the same region areas are considered to discriminate among them.

4.4. Interpretability of the Face CNN for apparent personality trait regression

In [24], it is proposed to visualize the images that produce the highest activation given a unit of a specific layer and segment them using the estimated receptive field to

Figure 7. Visualization of the 10 images that produce the highest activations for 10 different units in $conv5\_3$ layer (one unit per row).

view which part of the image each unit of the convolutional neural network pays attention to.

One of the goals of our work is to understand better the model trained for the prediction of the personality traits and visualize whether some semantic detectors also emerge from this network. This is done following the same methodology as in [24]. Therefore, for each unit and for each image processed by the network, the activations are obtained. Then, most confident images for each unit (the ones that produce the highest activations) are segmented and visualized. Figure 7 shows the 10 most confident images for the first 10 units of the last convolutional layer, i.e. $conv5\_3$.

We observe that the units seem to be related with semantic regions such as eyes, nose, mouth etc. Since the faces are now aligned, we can use the average face to create masks of the different face parts, and perform a quantitative study on what regions of the face the units show higher responses. Figure 8, at the left, shows the average face computed using all the training set in a in 14×14 grid, that corresponds to the resolution for the last convolutional layer of our network. In Figure 8 (right) we mark the regions used to generate the masks for mouth (in purple), nose (in red), eyes (in blue) and eyebrows (in orange).

To obtain the semantics of the internal units automatically, the following process is performed for each unit. First, the N most confident images, i.e. the ones that produce the N highest activation values, are identified. Then, the location with the highest activation value is detected for each one of these N images. Finally, these locations are aggregated into a spatial histogram that represents the spa-

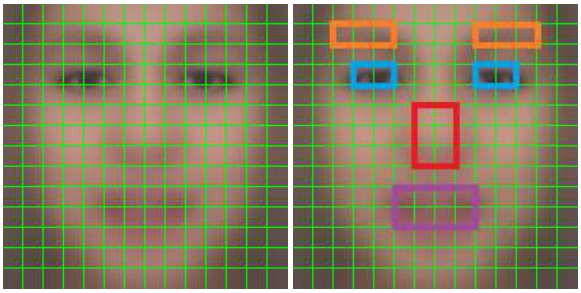

Figure 8. Left: Division of the average image into 14×14 cells. Right: Some semantic masks define to identify the face part detectors that emerge from the network automatically.

tial distribution of the highest activation values for that unit. This process is repeated for each unit of the layer.

Once the spatial histograms have been obtained, we can automatically connect semantic face regions and areas where units show their highest response. Thus, for a given semantic region, e.g. the mouth, the values of the spatial histograms are added over the region defined by the mouth mask (purple region from Figure 8-right). This way, for each unit, we obtain the number of images N' out of the N most confident images that have their highest values located within the region of interest. As a result, for a given semantic region, if the units are decreasingly sorted according to their values of N', the units that best represent such semantic region are automatically obtained as well as how good the connection between the semantic region and the unit is (the bigger N' the better).

We used in our experiments $N = 50$ and we show in Figure 9 the 5 most confident units for eyes, nose, and mouth detection.

More generally, we performed a quantitative visualization that shows what are the regions of the image that mostly activate the different units of each layer. To obtain this distribution we proceed as follows. First, for each unit, the N most confident images are obtained, i.e. the ones that produce the highest activation for that unit. Second, the position of the highest activation is stored as a matrix with all 0s except for such position, which has value 1. Third, the matrices for the N most confident images are added so that a spatial histogram of the highest activation locations is obtained for that unit. Then, the spatial histograms are added along all the units from the same layer. Finally, they are pooled to a resolution of 14x14 by adding the values belonging to the same cell so that all spatial histograms have the same resolution independently of the layer from which they have been obtained.

Figure 10 shows these spatial histograms for each convolutional layer represented as heatmaps. We can observe that the deeper the layer in the CNN the spikier the spatial histogram of the highest activation locations. This can be interpreted as follows: the higher the layer is, the units are more specialized in analyzing the eyes, nose, and mouth.

Additionally to the semantic regions of the face, we also observed that there are other semantic concepts that are detected by some units. For example, we manually identified some units that respond to eyes with glasses, as shown in Figure 11. Despite being a concept easy to locate in the aligned images (a similar mask as the one used for eyes detection could be used), the presence or absence of glasses should be manually checked along the units that have their highest activation values distributed over the eyes region area.

4.5. Action Units for Personality Traits Prediction

Our last set of experients are focused on evaluating the influence of shown emotion for the problem of personality trait inferences. We used the same Openface library [2] applied to crop the faces and generate the face dataset used in Sections 4.3 and 4.4 to also predict a subset of 17 Action Units (AUs) from the Facial Action Coding System (FACS) [7] (namely: AU1-Inner brow raiser, AU2-Outer brow raiser, AU4-Brow lowerer , AU5-Upper lid raiser, AU6-Cheek raiser, AU7-Lid tightener, AU9-Nose wrinkler, AU10-Upper lip raiser, AU12-Lip corner puller, AU14-Dimpler, AU15-Lip corner depressor, AU17-Chin raiser, AU20-Lip stretched AU23-Lip tightener, AU25-Lips part, AU26-Jaw drop, AU28-Lip suck and AU45-Blink)

We computed the intensity of each AU, and performed a correlation analysis between personality traits and automatically detected AUs.

The first experiment consists in checking if AU detectors also emerge from the internal units of the network as done with some semantic regions in Section 4.4. Given an AU, the N frames $\{F_{AU}\}$ that have the highest predicted intensity value for such an AU are identified. Then, for each internal unit, the N frames $\{F_{unit}\}$ that have the highest activation for such unit are obtained. Finally, the internal unit with the highest intersection between $\{F_{AU}\}$ and $\{F_{unit}\}$ is identified. Table 4.5 shows the results and the significance levels obtained for $N = 50$.

The best identification between AUs and internal units has been found for AU12 (Lip Corner Puller) and AU15 (Lip Corner Depressor) with units 108 and 220 respectively, both with an intersection value of 20%. A total of 3 out of 17 AUs show significantly above chance frames intersection between AU presence and trait prediction. Statistical significance has been computed according to the following formula:

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

where F is the number of frames from the whole set, N is the number of frames selected (N from $\{F_{AU}\}$ and N

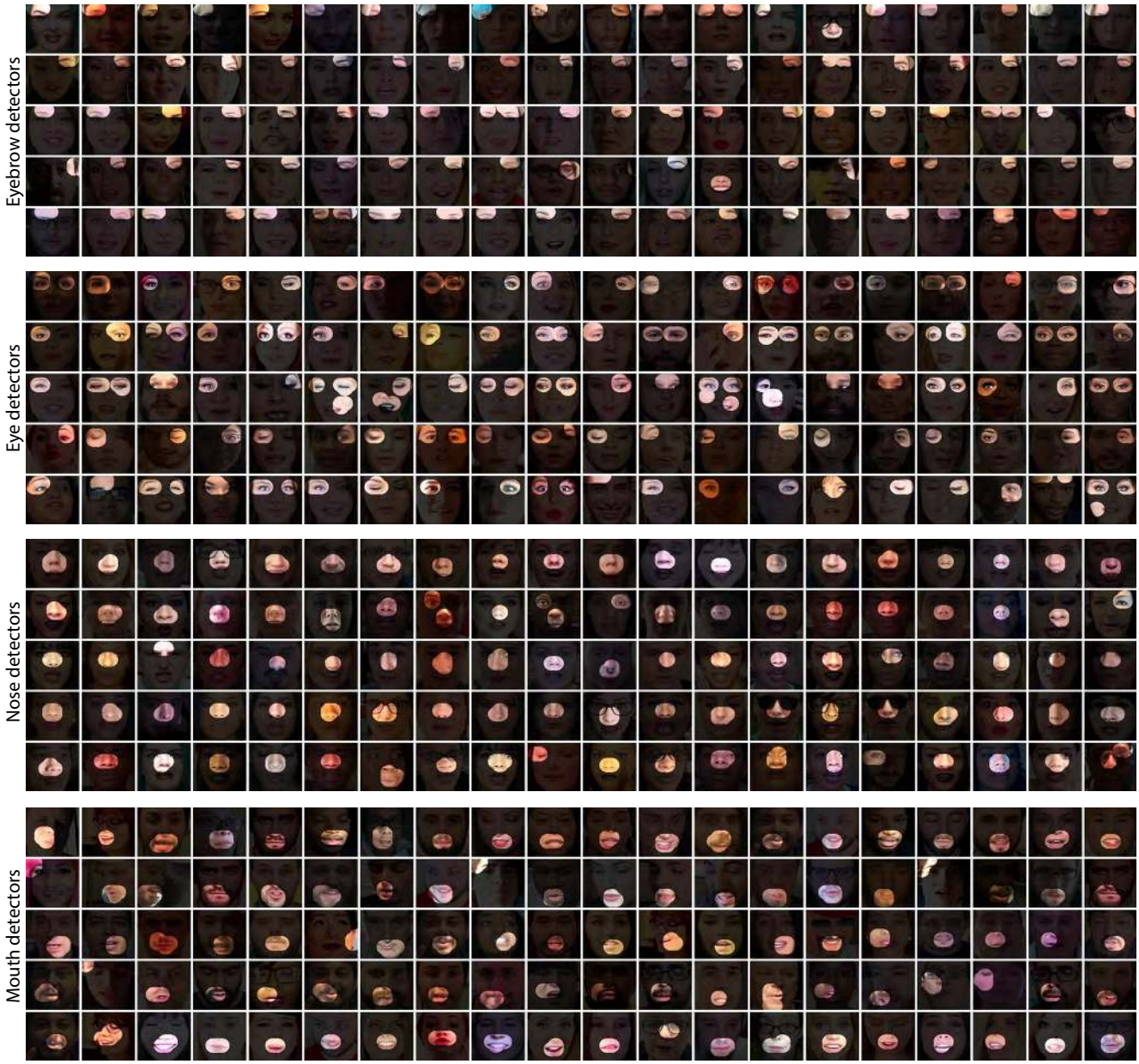

Figure 9. 5 most confident units for detecting eyebrow, eye, nose, and mouth.

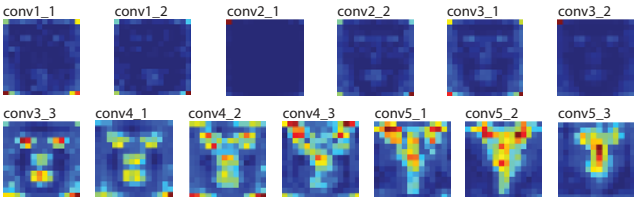

Figure 10. Spatial histograms of the most frequent activation locations for each convolutional layer. The deeper the layer in the CNN the spikier the spatial histogram of the highest activation locations.

Figure 11. Manual identification of units responding to eyes with glasses.

Table 2. Results for automatic identification of AU detectors emerging from internal units, and statistical significance (parenthesis).

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)

from $\{F_{unit}\}$), k is the minimum number of frames from the intersection of $\{F_{AU}\}$ and $\{F_{unit}\}$ for at least one of the units, and U is the number of units from the CNN layer being analyzed. The statistical significance values from Table 4.5 have been obtained from layer conv5_3 (U=512), using the test subset (F=2000) and selecting 50 frames (N=50).

We also further explored the trait predictive capabilities of AUs in images. We used the AU activation as a 17-dimensional feature vector, and trained a simple linear classifier on this data. This simple model yields an accuracy close to 0.886 with this reduced set of features. This result suggests that there is a dual informational cue when inferring social traits from facial images. Published results from the state-of-the-art show that a single still image can predict with high accuracy the trait inference, which is consistent with the Psychological literature that suggests that trait inferences are performed fastly, in milliseconds, before facial dynamics take place [1]. Nevertheless, mid-term temporal cues such as facial action units involved in emotion exposition have also influence on trait inferences, despite the low dimensionality of the high level signal.

5. Conclusions

In this paper we focused on the interpretability of deep learning models for apparent personality trait inferences. Taking as a reference the state-of-the-art model, and focusing only on the video content, i.e. discarding the audio signal, we found that facial information plays the key role in the trait prediction. In the light of this result, we retrained the model obtaining improved accuracies using only the facial region.

A recursive application of the visualization tool proposed showed that specific facial regions play a key role in the trait predictions. A set of facial part detectors automatically emerged from the last layers of the CNN with no supervision provided on this task. Specific units in the last convolutional layer ($conv5\_3$) unsupervisedly specialized on detecting mouth, nose, eyes and eye-brows. This methodology can be easily exported to other network architectures, and provide an explainable visualization to the results of the CNNs.

Finally we explored the influence of the emotional information on the trait prediction. Psychological studies suggest that humans infer trait judgments from a single image in a few milliseconds. Although the use of still images on CNNs suffices to obtain state-of-the-art accuracies, we hypothesized that dynamic information from the emotions portrayed also influence the trait predictions. We automatically annotated a set of 17 action units found in videos, and used this information to correlate AU presence and unit activation in the network predicting the personality traits. We found above chance relationships between certain units and AUs activation. In addition we trained a classifier using only the AU activation as features on the trait judgments prediction task. These simple attributes yielded fair regression results denoting the influence of AUs in facial trait judgments.

6. Acknowledgements

This research was supported by TIN2015-66951-C2-2-R grant from the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE) and NVIDIA Hardware grant program.

References

[1] C. C. Ballew and A. Todorov. Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, 104(46):17948–17953, 2007. 1, 8

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016. 4, 6

[3] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 53–56. ACM, 2012. 1, 2

[4] I. V. Blair, C. M. Judd, and K. M. Chapleau. The influence of afrocentric facial features in criminal sentencing. *Psychological science*, 15(10):674–679, 2004. 1

[5] J. M. Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990. 2

[6] J. L. Eberhardt, P. G. Davies, V. J. Purdie-Vaughns, and S. L. Johnson. Looking deathworthy: Perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, 17(5):383–386, 2006. 1

[7] P. Ekman and W. V. Friesen. Facial action coding system. 1977. 6

[8] H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen, et al. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. *Proceedings of ICPRW*, 2016. 2

[9] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003. 2

[10] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier. Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition. In *Computer Vision–ECCV 2016 Workshops*, pages 349–358. Springer, 2016. 2

[11] J. Joo, F. F. Steen, and S.-C. Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3712–3720, 2015. 2

[12] D. Masip Rodo, A. Todorov, and J. Vitrià Marca. The role of facial regions in evaluating social dimensions. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 210–219. Springer, 2012. 2

[13] N. N. Oosterhof and A. Todorov. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087–11092, 2008. 2

[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015. 3

[15] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops*, pages 400–418. Springer, 2016. 1, 2, 3

[16] M. Rojas, D. Masip, A. Todorov, and J. Vitria. Automatic prediction of facial trait judgments: Appearance vs. structural models. *PloS one*, 6(8):e23323, 2011. 2

[17] 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. In *Computer Vision–ECCV 2016 Workshops*, pages 337–348. Springer, 2016. 2

[18] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall. Inferences of competence from faces predict election outcomes. *Science*, 308(5728):1623–1626, 2005. 1

[19] 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. 2

[20] X.-S. Wei, J.-H. Luo, and J. Wu. Selective convolutional descriptor aggregation for fine-grained image retrieval. *arXiv preprint arXiv:1604.04994*, 2016. 3

[21] J. Willis and A. Todorov. First impressions making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598, 2006. 2

[22] L. A. Zebrowitz. The origin of first impressions. *Journal of Cultural and Evolutionary Psychology*, 2(1-2):93–108, 2004. 1

[23] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu. Deep bimodal regression for apparent personality analysis. In *Computer Vision–ECCV 2016 Workshops*, pages 311–324. Springer, 2016. 1, 2, 3, 4, 5

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 2, 5

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 3