Image and Vision Computing 78 (2018) 14–25

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Convolutional Neural Networks for Subjective Face Attributes ☆

Mel McCurrie[a,*], Fernando Beletti[a], Lucas Parzianello[a], Allen Westendorp[a], Samuel Anthony[b,c], Walter J. Scheirer[a]

[a]Department of Computer Science and Engineering, University of Notre Dame
[b]Department of Psychology, Harvard University
[c]Perceptive Automata, Inc.

ARTICLE INFO

Article history:
Received 19 October 2017
Received in revised form 15 May 2018
Accepted 21 June 2018
Available online 19 July 2018

Keywords:
Psychophysics
Face attributes
Convolutional neural networks

ABSTRACT

Describable visual facial attributes are now commonplace in human biometrics and affective computing, with existing algorithms even reaching a sufficient point of maturity for placement into commercial products. These algorithms model objective facets of facial appearance, such as hair and eye color, expression, and aspects of the geometry of the face. A natural extension, which has not been studied to any great extent thus far, is the ability to model subjective attributes that are assigned to a face based purely on visual judgments. For instance, with just a glance, our first impression of a face may lead us to believe that a person is smart, worthy of our trust, and perhaps even our admiration — regardless of the underlying truth behind such attributes. Psychologists believe that these judgments are based on a variety of factors such as emotional states, personality traits, and other physiognomic cues. But work in this direction leads to an interesting question: how do we create models for problems where there is only measurable behavior? In this paper, we introduce a convolutional neural network-based regression framework that allows us to train predictive models of crowd behavior for social attribute assignment. Over images from the AFLW face database, these models demonstrate strong correlations with human crowd ratings.

1. Introduction

In human attribute modeling, there often exists a disparity between the way humans describe humans and the way computational models describe humans. A large amount of describable attribute research in computer vision concentrates on objective traits. For example, work using the popular CelebA dataset [22,30,44,42] applies different methods to model traits such as "Male" and "Bearded" with binary annotations. Beyond objective attributes, it is possible to model more subjective traits such as expression [12,7], attractiveness [17], and humorousness [21], but research often overlooks the important interrelation between attribute modeling and social psychology. Enabling computers to make accurate predictions about objective content and enabling computers to make human-like judgments about subjective content are both necessary steps in the development of machine intelligence. Here, we focus on the latter.

Specifically, we concentrate on descriptions of the face, as an abundance of social psychology research demonstrates a human tendency to make judgments in social interactions based on the faces of fellow humans [33,40,1]. Popular human characteristics of academic interest closely related to these social interactions include emotion [24], attractiveness [1], trustworthiness [37,40,29,8], dominance [33,24], sociability, intelligence, and morality [1]. Psychologists often specifically concentrate on trustworthiness, dominance, and intelligence because they represent comprehensive abstract qualities that humans regard in each other. Alexander Todorov, one of the foremost psychologists studying these social judgments, uses dominance and trustworthiness as the basis of many in-depth studies of human judgment [36-38]. Ultimately, he finds that most other recognizable subjective traits in humans can be represented as an orthogonal function of dominance and trustworthiness [27], which suggests these two conceptual traits are ideal for computational modeling (Fig. 1).

Closely related to our work is research concentrated on the assessment of abstract traits in human faces based on the effect of facial contortions and positions. Inspired by animals' displays of dominance and submissiveness in respective head raises and bows, Mignault et al. specifically analyzed the effects of head tilt on the change in perceived dominance and emotion [24]. Not only does the study confirm the

☆ This paper has been recommended for acceptance by Alice J O'Toole.
* Corresponding author.
E-mail address: mmccurr1@nd.edu (M. McCurrie).

https://doi.org/10.1016/j.imavis.2018.06.010
0262-8856/

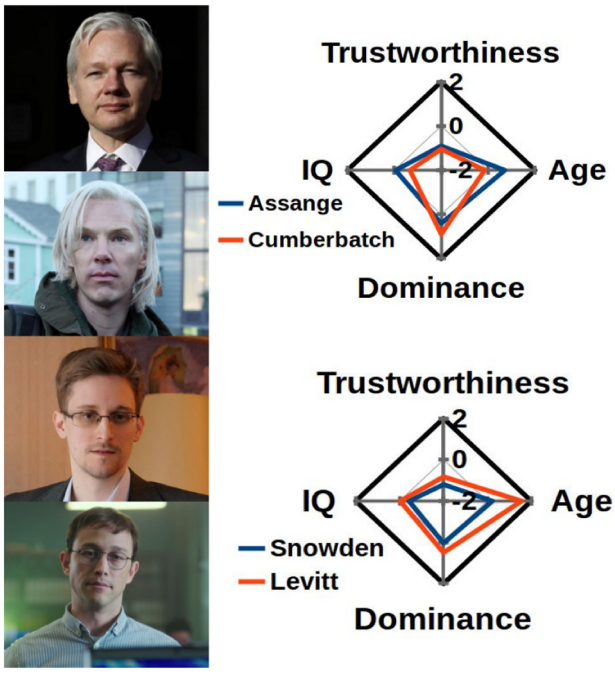

Fig. 1. Computational modeling of social attributes allows us to predict what the crowd might say about a face image. In this image we graphically compare the attribute predictions for Julian Assange and Benedict Cumberbatch, who plays Assange in the movie *The Fifth Estate*, as well as the predictions for Edward Snowden and Joseph Gordon-Levitt, who plays Snowden in the movie *Snowden*. Specifically looking at these images, our models output similar predictions between the subjects and their actors. The radar plots above reflect the output of a face processing pipeline, where faces are detected, aligned, and then processed through a deep convolutional neural network regressor that models a particular social attribute. This regression framework is the main contribution of our work. For this image we display the predictions' z-scores with respect to the training data.

hypothesized disparity in perceived traits based on head tilt, but it also finds gender has a noteworthy influence on subjects' perceptions. Keating et al. assessed the effect of eyebrow and mouth gestures on perceived dominance and happiness in a cultural context [14]. The study found smiling to be a universal indicator of happiness and showed weak associations between not smiling and dominance. It also determined the effect of a lowered-brow on perceived dominance to be generally restricted to Western subjects.

In this paper, we connect traditional machine learning and social psychology findings like those described above. We work specifically with traits that can be considered abstract representations of high-level human attributes. Additionally, we introduce a convolutional neural network-based (CNN) regression framework that allows us to train predictive models of crowd behavior for social attribute assignment. Very different from prior work, we make use of a unique visual psychophysics crowdsourcing platform, TestMyBrain.org, to gather the annotations necessary for training. As a case study, we examine four social attributes: trustworthiness, dominance, age, and IQ. We investigate each purely in the context of crowd judgments. Our models demonstrate strong correlations with crowd ratings, which we suspect are largely driven by low-level image cues.

In short, our contributions in this paper are:

- A novel dataset of over 6000 images annotated for all four traits of interest.
- The deployment of a crowd-sourced data collection regime, which collects large amounts of data on high-level social attributes from the popular psychophysics testing platform TestMyBrain.org.
- The comparison of different deep learning architectures for abstract social attribute modeling.
- A set of highly effective automatic predictors of social attributes that have not been modeled before in computer vision.

2. Related work

The related work in computer vision falls into two categories: general face attributes, and specific CNN-based approaches. We review both in this section.

2.1. Attributes in computer vision

Due to the proliferation of low-cost high performance computing resources (*e.g.*, GPUs) and web-scale image data, large-scale image classification and labeling is now commonplace in computer vision. With respect to face images from the web, Labeled Faces in the Wild [13], YouTube Faces [41], MegaFace [26], Janus Benchmark A [15], and CelebA [22] are all popular choices for a variety of facial modeling tasks beyond conventional face recognition. Attribute prediction, where the objective is to assign semantically meaningful labels to faces in order to build a human interpretable description of facial appearance, is the particular task we concentrate on in this paper.

Both Farhadi et al. [9] and Lampert et al. [19] originally conceived of visual attributes as a development supporting object recognition, rather than a primary goal in and of itself. Faces, however, are a special case where standalone analysis supports applications in biometrics and affective computing. Kumar et al. used facial attributes for face verification and image search [17]. Scheirer et al. applied the statistical extreme value theory to facial attribute search spaces to create accurate multi-dimensional representations of attribute searches [31]. Siddiquie et al. modeled the relationships between different attributes to create more accurate multi-attribute searches [34]. Lastly, Luo et al. captured the interdependencies of local face regions to increase classification accuracy [23].

Certain traits such as Age [25,18,20] and gender [21,20] have enjoyed disproportionate attention, but researchers also model numerous other facial attributes. The release of the large CelebA dataset [22] also prompted several novel studies of facial attributes on all 40 traits in the dataset [30,44,42]. For a comprehensive review of facial attribute work in practical biometric systems, see the review authored by Dantcheva et al. [6].

2.2. Convolutional neural networks for attributes

Current state-of-the-art facial attribute modeling relies on CNNs. Pioneering work in the field, Golomb et al. trained a CNN with an 8.1% error rate on gender prediction [11]. More recently, Zhang et al. used CNNs alongside conventional part-based models to predict attributes such as clothing style, gender, action, and hair style from images [43]. Wang et al. applied CNNs to an automatically generated egocentric dataset annotated for contextual information such as weather and location [39]. Levi et al. used a CNN for age and gender classification from faces [20]. Liu et al. used two cascaded CNNs and trained support vector machines to separate the processes of face localization and attribute prediction [22]. Finally, Zhong et al. extended the work of Liu et al. using off-the-shelf CNNs to build facial descriptors in a different approach to attribute prediction [44].

Most similar to our research is the recent work of Lewenberg et al. [21]. They use a CNN to predict objective traits including gender, ethnicity, age, make-up, and hair color, and subjective traits including emotional state, attractiveness, and humorousness. That research introduced a new face attributes dataset of 10,000 images annotated for these traits. To generate this dataset, Lewenberg et al. employed Amazon's Mechanical Turk raters from the US and Canada to rate

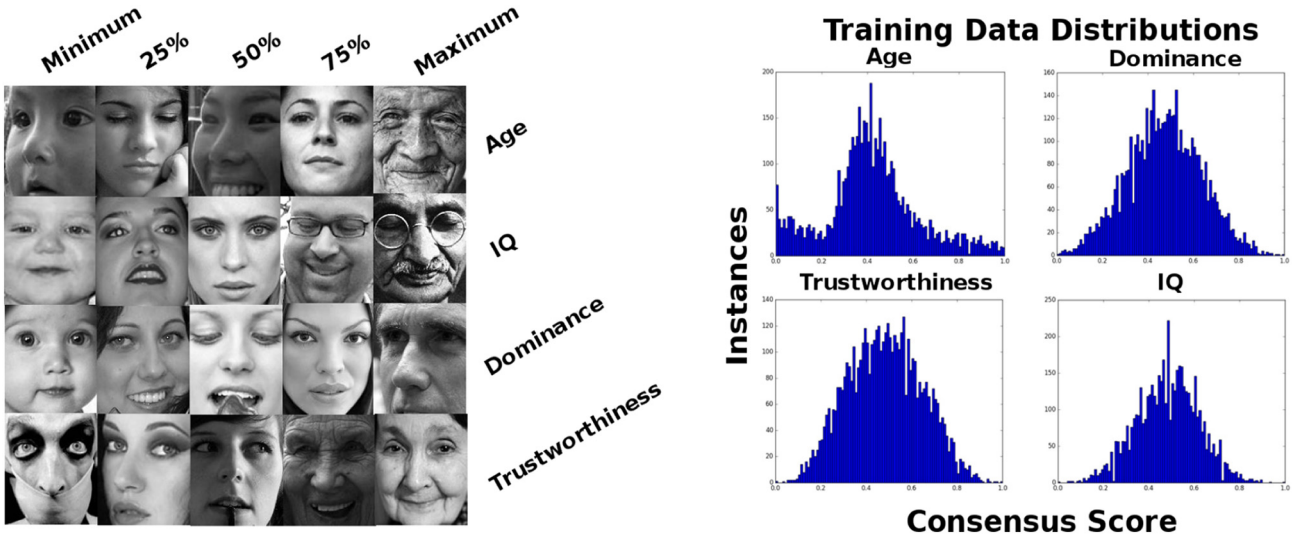

Fig. 2. Data distributions. We assert that to most accurately model humans' psychological judgments, each of these traits should be modeled on a continuous distribution. For this reason we employed the Likert Scale in our data collection and then took the average of human ratings for each image. This graphic shows faces at each quartile from the dataset (left) as well as the training data distributions (right), all of which seem to be close to normal.

a subset of the PubFig dataset, aggregating labels from three separate individuals for each image. Notably, the work only analyzes the traits with binary classification, labeling each image as "yes" or "no" with respect to a trait. Our most immediate improvement on this work is in the way in which we collect data. We use an online psychophysics testing platform, aggregating data from a larger number of raters from an arguably more reliable and geographically variable source. In addition, we model more abstract, representational traits on continuous distributions.

Also parallel to our work, and the current state-of-the-art attribute prediction, is the work of Rudd et al. [30]. Rudd et al. employ a single custom Mixed Objective Optimization Network (MOON) to multi-task facial attribute recognition, minimizing the error of their networks over all forty traits of the CelebA dataset [22]. We use our own implementation of the MOON architecture as a basis for each separate trait in our modeling.

Fig. 3. Annotation task. A sample behavioral task that a subject might see on TestMyBrain.org. All ratings collected for this work were on a Likert scale between 1 and 7, where 1 indicates the least amount of attribute presence, and the 7 indicates the most amount.

3. Crowd-sourced data collection

In this paper, we introduce a new dataset for social attribute modeling. The dataset consists of 6300 grayscale images of faces sampled from the AFLW dataset [16] and annotated for the four traits we study. Representative samples of the dataset for each trait can be seen in Fig. 2. This dataset is novel in that it captures human annotators' subjective assessment of traits with underlying truth. For traits such as Age and IQ, which are easy to record and described on well-known scales, it is of course possible to produce a dataset with verifiable measurements of the underlying traits — but this is not our objective. Rather than analyze and model actual trustworthiness, dominance, age, and IQ, we choose to study people's described perceptions of the aforementioned traits. For example, our dataset does not include actual ages, instead the images are annotated by a consensus score — aggregate statistics of what many people said about the ages of the subjects in the images.

3.1. TestMyBrain.org

For this high-level annotation, we use TestMyBrain.org [10], a crowd-sourced psychophysics testing website where users go to test and compare their mental abilities and preferences. It is one of the most popular "brain testing" sites on the web, with over 1.6 million participants since 2008. But what specific advantages does TestMyBrain.org have over a service like Amazon's Mechanical Turk?

TestMyBrain.org is a citizen science effort that facilitates psychological experiments and provides personalized feedback for the user,

Table 1
Dataset statistics: Statistics on the 5040 images used for training for all four social attribute classes (normalized to a [0, 1] range). The "Mean Std. of Ratings" refers to the average standard deviation of the human scores for each individual image.

	Trust.	Dom.	Age	IQ
Mean of ratings	0.48	0.47	0.42	0.48
Std. of ratings	0.16	0.16	.20	0.14
Mean Std. of ratings	0.34	0.32	0.13	0.27
Mean Num. of ratings	32.47	32.19	15.80	15.79

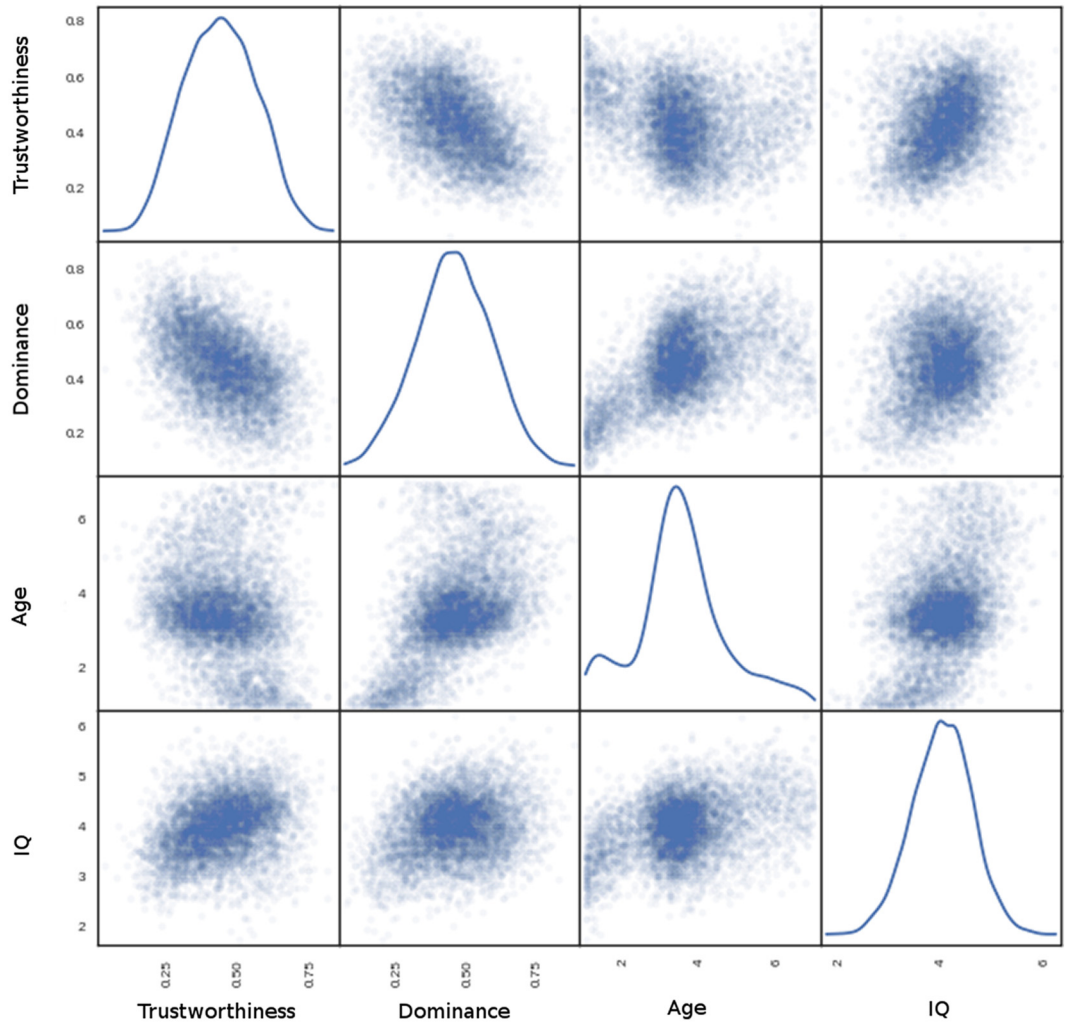

Fig. 4. Attribute relationships. Trustworthiness and Dominance have a linear negative correlation. Age tends to have a non-linear relationship with the other traits, reflecting the similar Trustworthiness, Dominance, and IQ between the elderly and children individuals. Originally we hypothesized that Trustworthiness and Dominance would provide an orthogonal basis for other traits and therefore be linearly independent. These graphs show this may not be the case.

mutually benefiting both researchers and those curious about their own mind. The subject pool is geographically diverse and provides an arguably superior psychometric testing group compared to smaller more homogeneous subject pools such as that of Mechanical Turk. In addition to being an ideal setting for aggregate, cross-cultural psychometric experiments for researchers, TestMyBrain.org provides the non-monetary incentive of detailed, personalized results for subjects. Subjects visiting the site are motivated by a desire to learn about themselves and have little incentive to respond to experiments quickly or poorly. Based on these factors, we determined that the subject pool of TestMyBrain.org is ideal for the delicate task of honestly appraising abstract attributes in faces.

Using TestMyBrain.org, we asked participants to judge faces for a select trait on a Likert Scale, a psychometric bipolar scaling method shown in Fig. 3. As can be seen in Table 1, each face has an average of about 32 judgments for Trustworthiness and Dominance and 15 for Age and IQ. We recorded the average judgment to use as the consensus score for that image and normalized the Trustworthiness and Dominance scores. In training we map all y values so that $0 \leq y \leq 1$. We calculated the coefficient of determination (R^2) of mean human ratings from two independent sets of 943 subjects for 389 random images from the AFLW set for Trustworthiness and 400 random images from the AFLW set for dominance. The Trustworthiness R^2 is 0.93 and the Dominance R^2 is 0.88. Both of these statistics are very similar to the internal reliability calculated by Oosterhof and Todorov [27]. Thus, there is indeed signal in these data that can be learned by a machine learning algorithm.

3.2. Data insights

We can learn about simple patterns in human judgment by observing correlations in the data. The Annotated Facial Landmarks in the Wild Dataset provides annotations for gender, glasses, occlusion, grayscale, facial landmarks, and bounding boxes. Comparing attribute distributions conditioned on the objective, binary traits provides interesting information. Females tend to be rated as more trustworthy and less dominant, people with glasses are generally rated as more intelligent. Occlusion has a weak correlation with trustworthy ratings. Glasses have a weak correlation with age ratings. Most likely due to our preprocessing, original image quality and grayscale are not correlated with any attributes. Some interesting relationships are shown in Fig. 5.

As discussed previously, we hypothesized that Trustworthiness and Dominance are an orthogonal basis for other human attributes,

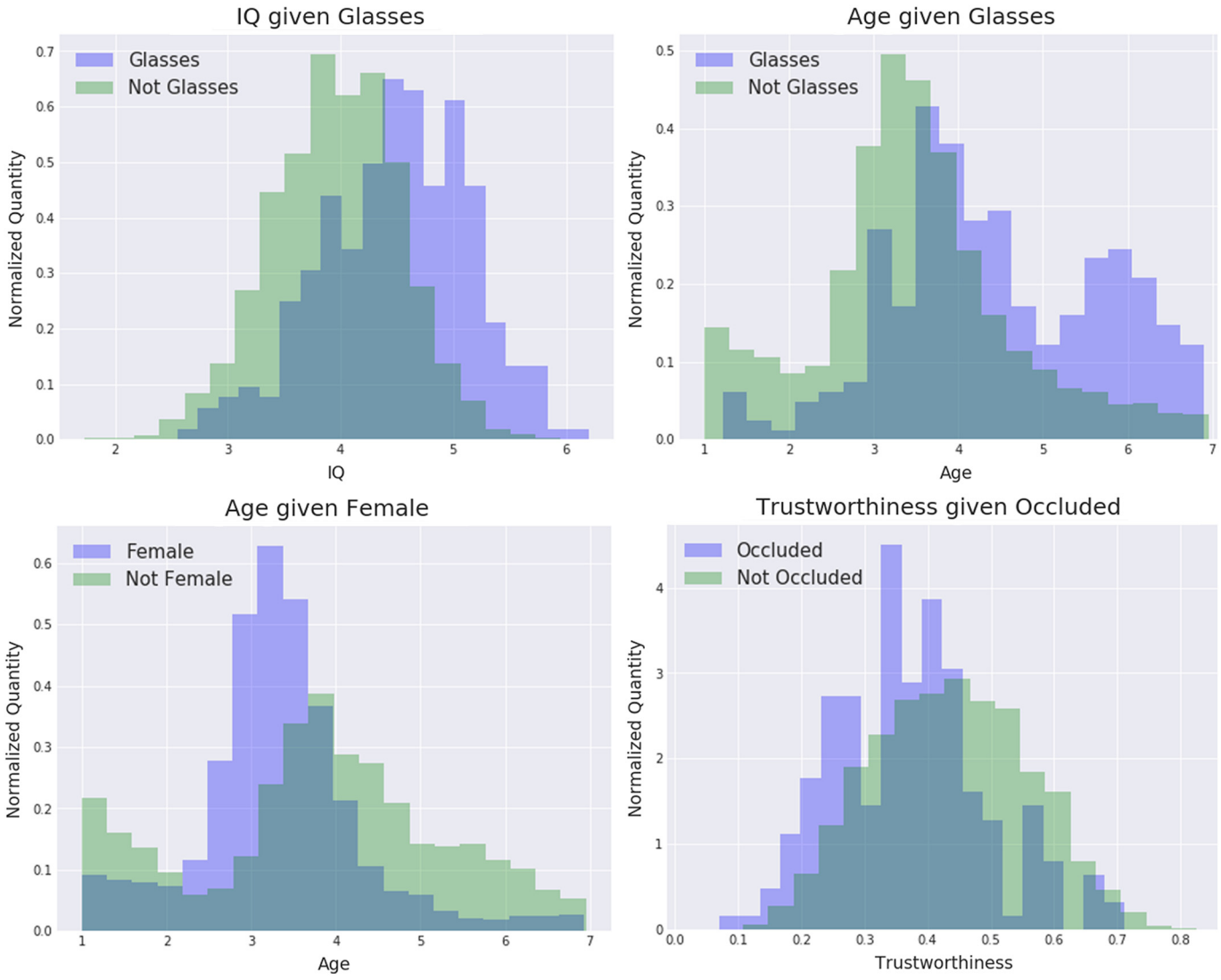

Fig. 5. Dataset patterns. A few of the interesting relationships between our annotated attributes and the AFLW attributes. From left to right, top to bottom: glasses is positively correlated with IQ, glasses is positively correlated with Age, female Age is distributed more normally than male Age, and occlusion is negatively correlated with Trustworthiness. Best viewed in color.

implying statistical independence. As seen in Fig. 4, Trustworthiness and Dominance are negatively correlated. This may be evidence against the hypothesis, but also could be a result of several factors including question wording and calculating aggregate score measurements using the mean.

4. CNN regression for social attributes

Our algorithm is a regression model that outputs a single score from an input image. A regression, rather than a binary classification, is a more realistic representation of the initial judgments humans make. For example, from our four modeled traits, both Age and IQ are already known to be described by continuous distributions and are therefore likely judged on continuous distributions. We assert the other two modeled traits, Trustworthiness and Dominance, are similarly best described by continuous distributions. For what is discussed below with respect to architectures, assume the output is always a single floating point number from a fully-connected layer after feature extraction.

4.1. Comparing architectures: what works best for social attribute modeling?

We initially compare different feature extraction algorithms, most of which are Convolutional Neural Networks with varying depths and use of regularization. We run each with similar parameters that we determined empirically. With respect to our implementation of the architectures, we made use of the Keras [5] and Theano [2] deep learning frameworks.

To test very shallow features we examine feature extraction with Histogram of Oriented of Gradients and two custom shallow CNN architectures. For the most basic feature extraction we use the Histogram of Oriented Gradients algorithm followed by several fully connected layers with dropout. In the “Shallow” network, we employ three segments of convolution and max pooling connected to fully-connected layers with dropout and Parametric ReLU activations. In the “Basic6” network, we employ four segments of a single convolution and max pooling followed by two fully-connected layers with ReLU activations and no dropout.

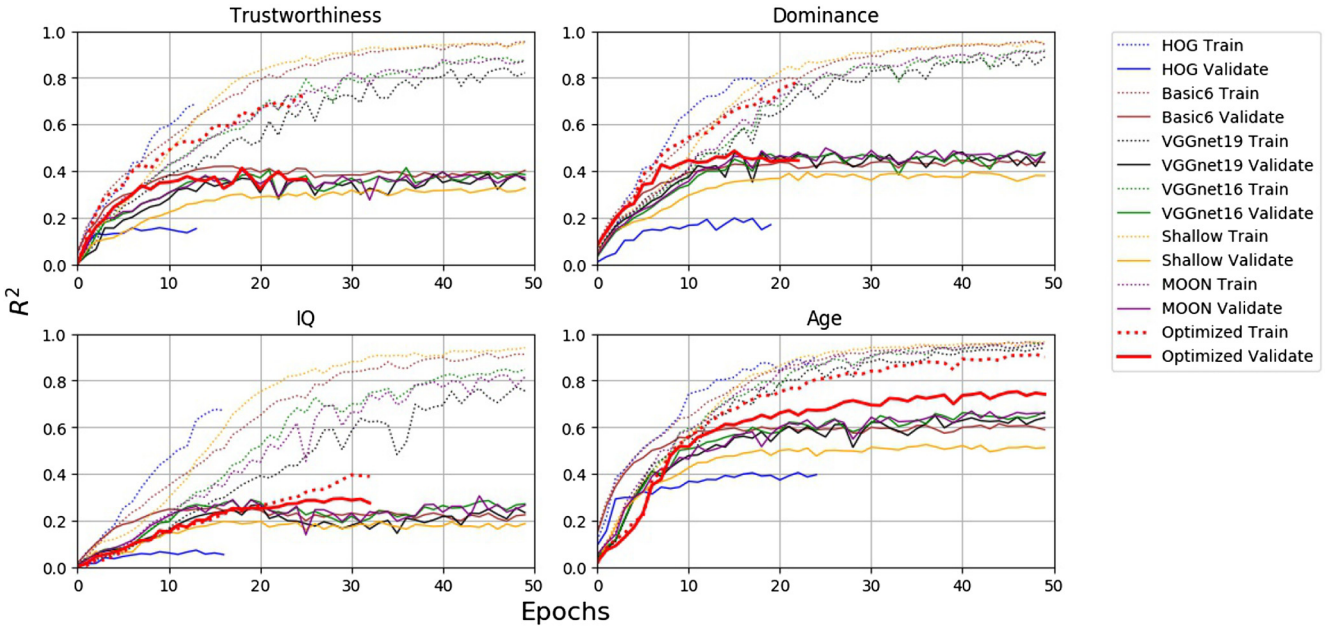

Fig. 6. Performance during training. In this image, we compare the ability of each architecture to learn the dataset and generalize to validation data. We include all four traits, training and validation scores, and all four original architectures plus our final architecture based on the hyperparameter optimization results. Of the four original architectures the MOON models generally perform better, but our optimized models consistently perform the best. (Optimized models were trained with early stopping, as can be seen in the plots.) Best viewed in color.

To test moderately deep architectures we used the Oxford Visual Geometry Group's VGG 16 and 19 networks [35], and a VGG16 variant from Rudd et al. We reproduce the convolutional architectures, modifying the shape of the input and output matrices for our smaller grayscale images and single floating point regression output. The newest architecture we analyzed is our implementation of the MOON architecture [30], which is more shallow than both of the VGGNet implementations. The convolutional feature extracting portion of the architecture is similar to the VGG networks in that it consists of several segments, where each segment has multiple

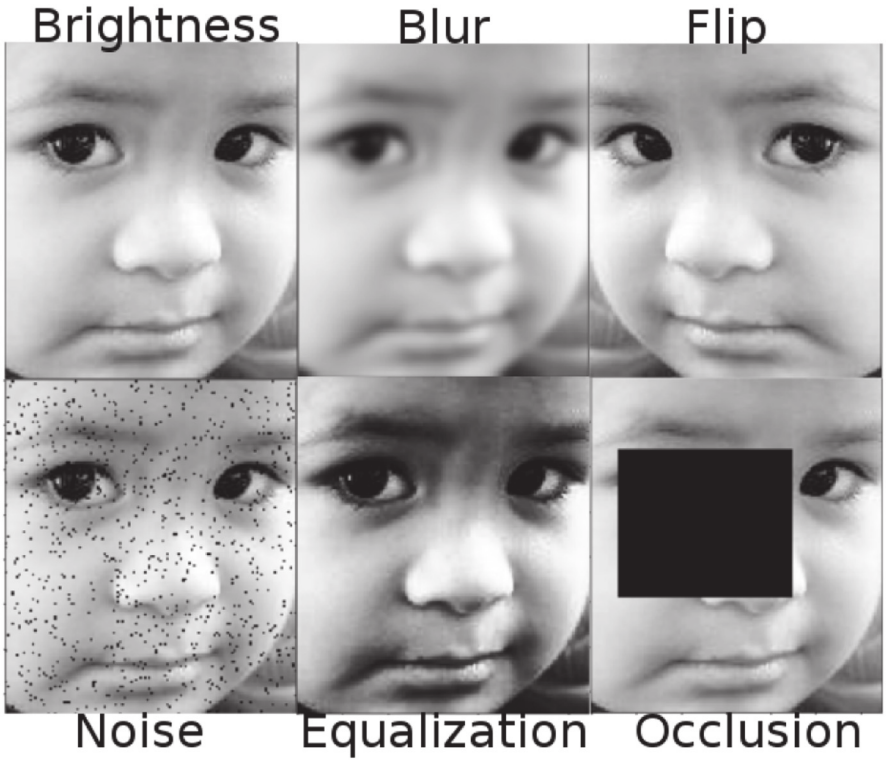

Fig. 7. Data augmentation. We explore the use of six different augmentation techniques for our small dataset, pictured here. During hyperparameter optimization, we optimize the probability that each of these augmentations occurs during training, and then the extent to which it occurs. For example, we can optimize the possible size of the occlusion, the extent of the brightness, and the possible amount of noise added to each image.

Table 2
Hyperparameters: Some important hyperparameter optimization results per trait for our optimized MOON architectures.

	Trust.	Dom.	Age	IQ
Learning rate	$10^{-4.2}$	$10^{-4.4}$	$10^{-4.8}$	$10^{-4.6}$
Dropout	55%	31%	45%	38%
2x Convolution 0	64	32	32	64
2x Convolution 1	64	64	128	32
2x Convolution 2	128	–	–	–
3x Convolution 3	256	256	256	256
3x Convolution 4	256	512	512	256
3x Convolution 5	256	512	512	–
FC layers	1	3	4	3
FC outputs	2079	2227	2187	1244

Table 3
Results: R^2 values of validation and testing results from our optimized MOON architectures for each trait. Best results on the test set are shown in bold.

	Trust.	Dom.	Age	IQ
Validation	0.41	0.49	0.75	0.29
Validation ensemble	0.43	0.51	0.74	0.30
Test	0.38	**0.46**	0.72	0.24
Test ensemble	**0.43**	**0.46**	**0.74**	**0.27**

convolutional layers followed by a max pooling layer. We modify the architecture for our smaller grayscale images and connect the convolutional layers to fully-connected layers that output a single score. To test a very deep network we use Resnet-50. We find that Resnet-50 with random weight initializations does not converge well when trained on so few images.

As will be discussed below in Section 6, the differences in model performances on the validation sets during training are not very large, suggesting the architecture choice may not make a significant difference. The newer MOON architecture performs slightly better on most of the traits, however, so we chose to use it as a basis for our final optimized models. Note that the earlier work of Lewenberg et al. [21] used an AlexNet block structure augmented with supervised features (facial landmark information) and a custom loss function, while MOON is a more straightforward VGG [35] modification.

4.2. Augmentation

To make the most of our small dataset, we augment the images with occlusion, brightness, blur, histogram equalization, horizontal flipping, and random noise at optimization time (Fig. 7). We predict that some of these augmentation techniques will help, such as flipping the image, but given our findings that features of the image such as occlusion are correlated with human ratings, we suspect that overuse of certain augmentation will harm our models ability to generalize. We explore this further in our hyperperameter optimization.

4.3. Under and over sampling

As seen in Fig. 2, the aggregate ratings for each trait tend toward a normal distribution. Thus, the extreme values are rarely observed by the model in training and values that tend toward the mean score are observed frequently. We chose to treat our dataset as imbalanced and adapt class under and over sampling for a regression situation. We first bin all values, normalize the discrete distribution, and assign a default probability of $1 - P(bin)$, where $P(bin)$ is the relative probability of a bin in the training data. We then choose a prior function over the probability of the binned values:

$$P(p) = p^{\alpha} + 10^{-\beta}$$

where $p \in [0, 1]$ is the original bin probability, $\alpha \in [2, 20]$ controls how often extreme values are chosen and $\beta \in [1.2, 3]$ controls how often the most common values are chosen. Then we optimize α and β during hyperperamater optimization.

4.4. Hyperparameter optimization

Rudd et al. train their MOON models on RGB images that are larger than our grayscale images and model hypothetically less

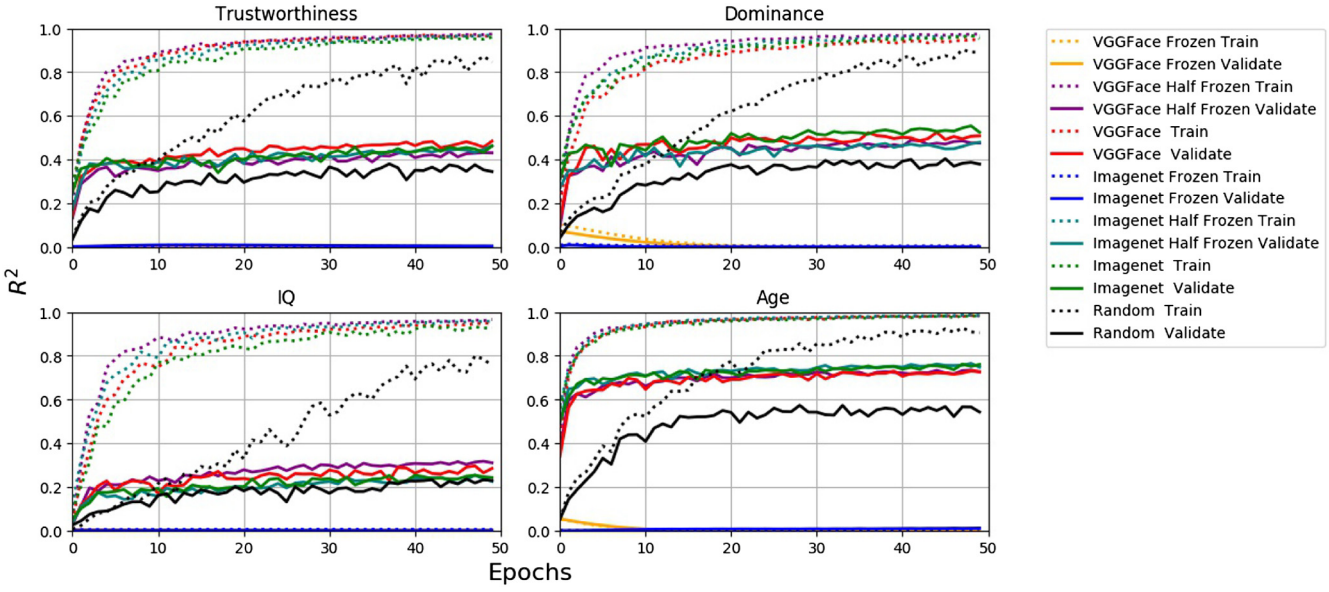

Fig. 8. Transfer learning performance during training. In this image, we compare different weight initialization methods, namely random, Imagenet pretrained, and VGG-Face pretrained. We also observe the effects of freezing some layers. Pre-trained initializations provide a clear advantage in both speed and performance The difference between VGG-Face and Imagenet pretrained weight initializations seems to be negligible. Using the original pretrained representations by freezing all the layers, however, hinders performance. Best viewed in color.

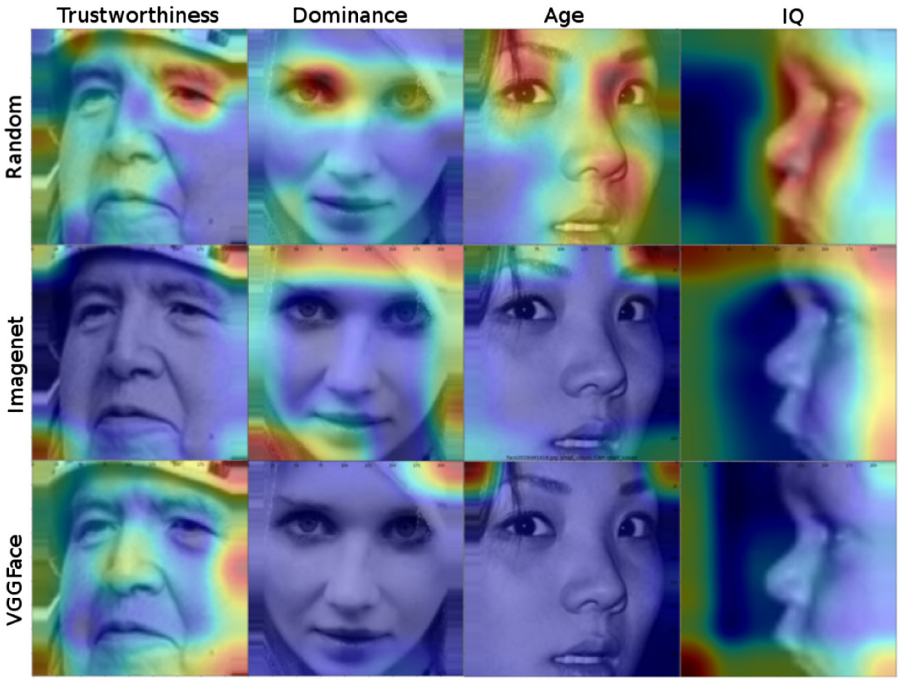

Fig. 9. Transfer learning saliency. Here, we display the different localization abilities of models trained with different initialization schemes on images from our validation sets. Random initializations appear to be the best at localizing obvious features, but perform the worst.

abstract objective attributes from the CelebA dataset, which is annotated for binary classification. This suggests that our very different dataset and features could benefit from some deviations in parameter choices.

To determine the best network size and deviation in parameters from the original MOON architecture, we optimize the network for each trait using hyperopt [3], a python library for hyperparameter optimization. Our search space includes learning rate, dropout, the number of filters in each layer, the number of layers, the amount of data augmentation, and the parameters of a sampling function. Employing hyperopt with the Tree of Parzen Estimators (TPE) algorithm allowed us to test a multitude of different parameter and architecture combinations. After a very wide parameter search, we perform a refined search with early stopping, and use the best models.

We maximize the model's performance with respect to the coefficient of determination (R^2) from the regression of $\hat{y}$, the model's predicted scores, on y, the average human annotations. We use the coefficient of determination as the measure of performance because it represents the percentage of prediction variation explained by the regression model. Thus our measure of performance does not reflect accuracy with respect to the underlying trait, but rather captures agreement between our model and the aggregate assessments of human annotators.

As seen in Fig. 6, each trait trains very differently. Following this trend, each trait's coefficient of determination is optimized by slightly different hyperparameters and deviations from the MOON architecture as seen in Table 2. However, the improvements are only modest, suggesting that deeper architectures and data augmentation are not helpful for this task.

4.5. Smoothing errors with an ensemble

Another simple improvement is to train an ensemble of optimized networks on our data and take the average of each model's regression output at inference time. Using the optimized parameters we train five models holding out a different 20% of the training data for each model. Despite each model using less training data, correlations with crowd judgments improve for all attributes but Dominance. The final R^2 values are seen in Table 3.

5. Transfer learning

Given the dataset only contains 6300 images, training may be improved by initializing model weights with those previously trained on other datasets. We concentrate on VGG16 and observe the effects of different pretrained weight initializations.

We choose VGG16 weights trained on the Imagenet dataset for the general task of object recognition, as well as VGG16 weights trained for Face Recognition [28]. Weights from a face recognition task should encode features specific to the face, which may help our understanding of the features necessary for this task. We train iterations of each of these models with all layers frozen, half the layers frozen, and no layers frozen. For each network, we only use the convolutional layers and add a single fully connected layer after.

Both pretrained networks were trained on color images of size 224×224 so we expand our tightly cropped images exemplified in Figs. 2 and 3 by repeating the last row or column and adding two channels. Doing so adds considerably more useless information to the input, affecting our ability to compare these models to our optimized models. Therefore, we also train a network with random weight initializations for fair comparison.

The models' training over time can be seen in Fig. 8. Transfer Learning has two clear benefits over random initializations. First the pretrained models perform better, reaching a minimum that generalizes better on the validation set. Second the pretrained models approach this superior minimum very quickly, allowing for significantly shorter training time. As shown in Fig. 8 there is not an obvious difference between VGG-face and Imagenet initializations, and there is not an obvious difference between training all layers and training only the last half of the layers.

We can also compare the ability of each network to localize facial features to understand why pretrained models perform better. We take the gradient with respect to the output of our model and weight

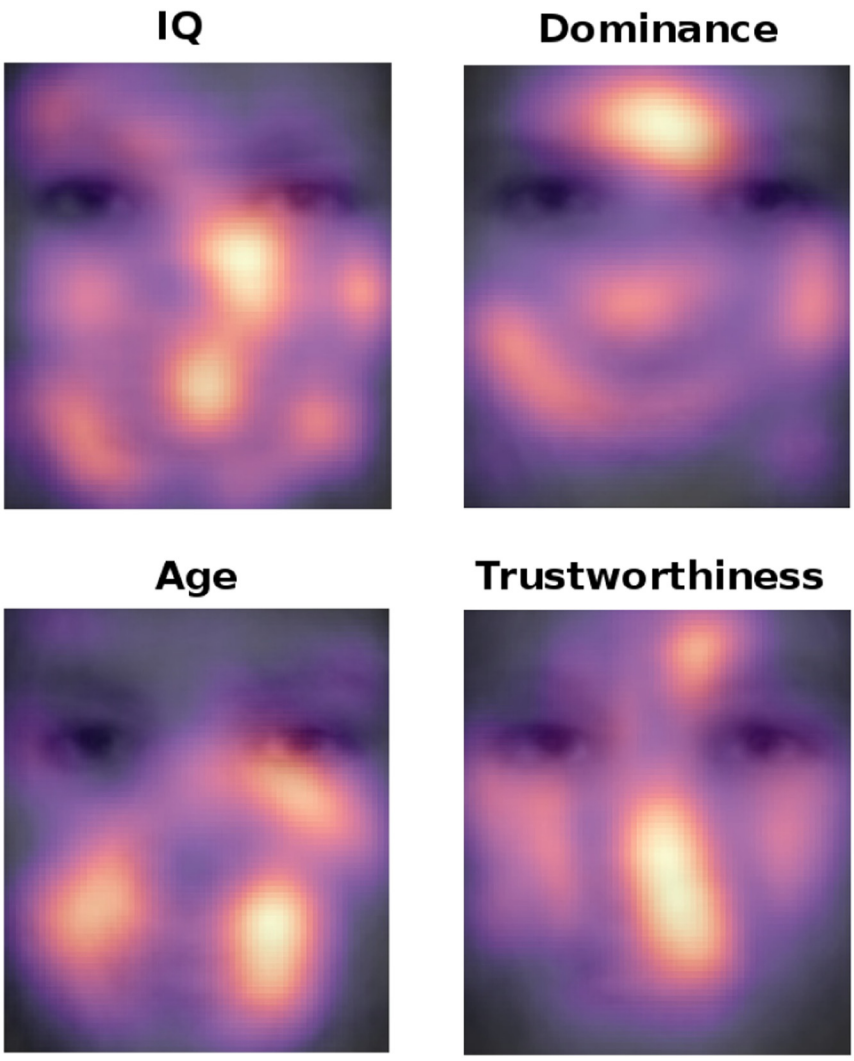

Fig. 10. Saliency from occlusion. We can visualize regions of the face that are most important to the trained models by systematically covering parts of the face and recording the absolute differences. Here, we separately analyze 100 images of the validation set and display the average differences as a heatmap on top of the averaged faces. Best viewed in color.

convolutional features by their relative significance as in Ref. [32]. We upscale and overlay the given heatmap on the input image, giving a rough idea of which areas are most important for an input. Fig. 9 compares the localization between networks with and without transfer learning, showing immense differences in feature localization. It is interesting to see that the randomly initialized networks that generalize considerably worse than the pretrained networks seem to localize features such as the eyes and mouth considerably better than the pretrained models. We do not use pretrained networks further in our evaluation.

6. Experimental evaluation

There are two important facets of evaluation with respect to our social attribute models: (1) model correlation with human crowd ratings of images, and (2) feature importance for social attribute models. Each of these facets is explored in this section. After data collection, our dataset consisted of 6300 grayscale images of faces, aligned to correct for in-plane rotation using the CSU Toolkit [4] and annotated for Trustworthiness, Dominance, Age, and IQ. We randomly separated 80% of the original dataset into a training set, and split the remaining 1260 images into a validation and test set (630 images each). The test set is held out during training, while the validation set is used to tune the hyperparameters.

6.1. Correlations with crowd judgments

We employ the R^2 value from a regression of $\hat{y}$, the model's predicted scores, on y, the original human annotations, as a measure of our model's performance. This is a reasonable metric given the linear relationship of y and $\hat{y}$. To properly compare architectures and assess the training performance of our optimized model, we record the R^2 at each epoch and graph them in Fig. 6.

Looking at the graphs, the validation R^2 values are ultimately very similar between architectures. There is some variability in training speed, and randomly good weight initializations seem to help, however the depth of the architecture does not seem to explain any improvement in scores.

As expected, our optimized architectures outperform the other four architectures. We display our final results from the optimized networks in Table 3, which shows R^2 values from regressions of our model's predicted values on human annotated consensus scores for both the validation and testing sets. Each trait has a slightly different coefficient of determination, however all scores are strong for a

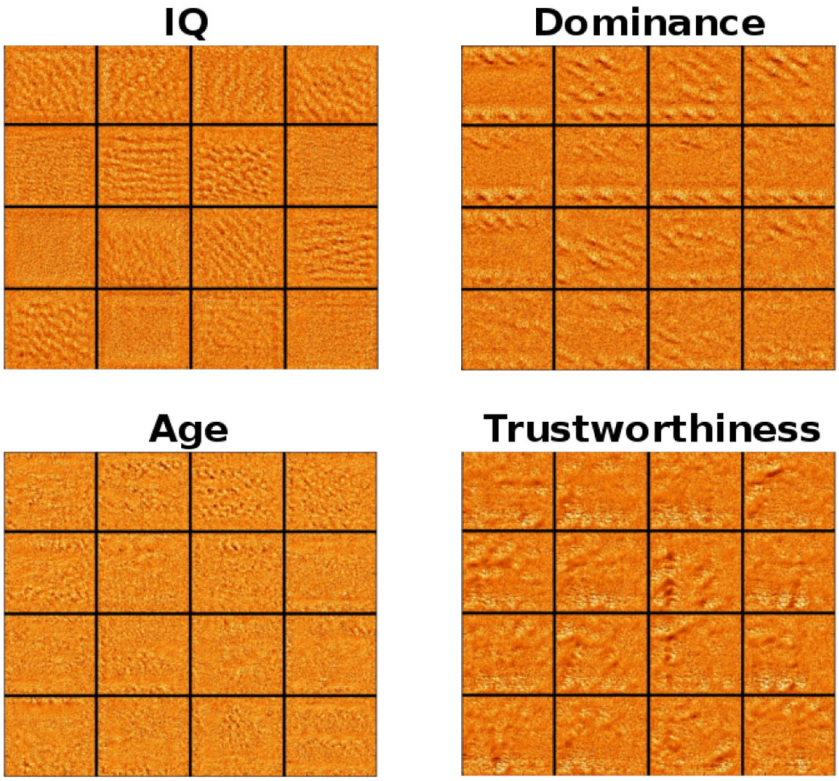

Fig. 11. Final filters. Visualizing a sample of the filters from the last convolutional layer of each optimized model, we can observe the resemblance of the output to a low-level feature extractor, consistent with our observation that deeper architectures add little to no improvement. (Color added to improve contrast.)

psychology-oriented experiment incorporating noisy human measurements. Our models for Age are the strongest, IQ are the weakest, and Trustworthiness and Dominance perform similarly to each other.

6.2. Visualizations of feature importance

Visualizations of the hyperparameter optimized CNN models show localized areas of importance on the face for each trait. As an example, we overlay average heatmaps for each trait on the averaged faces of 100 images from the validation data in Fig. 10. To produce these graphics we systematically moved a gray box over an image, iteratively scaling the box down after each pass. We then recorded the absolute difference in total score at each point. This visualization is intriguing because it allows us to view, in a certain image, or over an average of images, what areas of the face have the most or least significant effect on the final prediction.

The performance of a model trained on subjective human assessments is much less intuitive than the performance of a model trained on an objective task. Although interpretation of performance within the realm of supervised learning remains the same, validating the results is not as simple as looking at the outputs. Referring back to previous social psychology research [24,14] both Trustworthiness and Dominance are expected to rely on the mouth. Our models indicate a heavy reliance on areas near the mouth and chin. Similarly, Keating et al. [14] determined that a lowered brow should affect the (mostly Western) perception of Dominance. Both our Dominance and Trustworthiness models approximately locate the brow midsections. These observations indicate that our models have learned to look in the same places that humans do, possibly replicating the way we judge high-level attributes in each other.

Another method of analyzing our models is a visualization of the filters. Our visualizations of the filters from the final convolutional layer of each network in Fig. 11 are intriguing because they resemble the output of a low-level feature extractor. This indicates that despite the high-level abstract quality of these traits, low-level features might be enough for humans to make their immediate judgments. This is consistent with our observation that deeper architectures add little to no improvement.

6.3. Processing faces in video

A very good litmus test for our models is video processing. For each frame from a video, we can apply face detection and face alignment, and then use our optimized models to predict the score of each trait. We can even do this in real time — displaying the predicted change in crowd assessment as subjects alter facial position and expression in the video. Fig. 12 shows several frames from a couple of example videos being processed. In Fig. 12, all scores are mapped to a standard normal distribution and shown over time on both a line plot and a histogram. A selection of processed videos are provided as supplemental material.

7. Discussion

Current state-of-the-art visual recognition algorithms in computer vision, and more specifically algorithms for facial attribute prediction [21,30], show accuracy that promises new applications in the near future. It is in the best interest of both researchers and developers in industry to promote research that focuses on the interrelation of machine learning, computer vision, and social psychology.

A model is only as good as its data. The dataset and its annotations will ultimately have the most significant effect on the psychological validity and usefulness of the models. When annotating a dataset

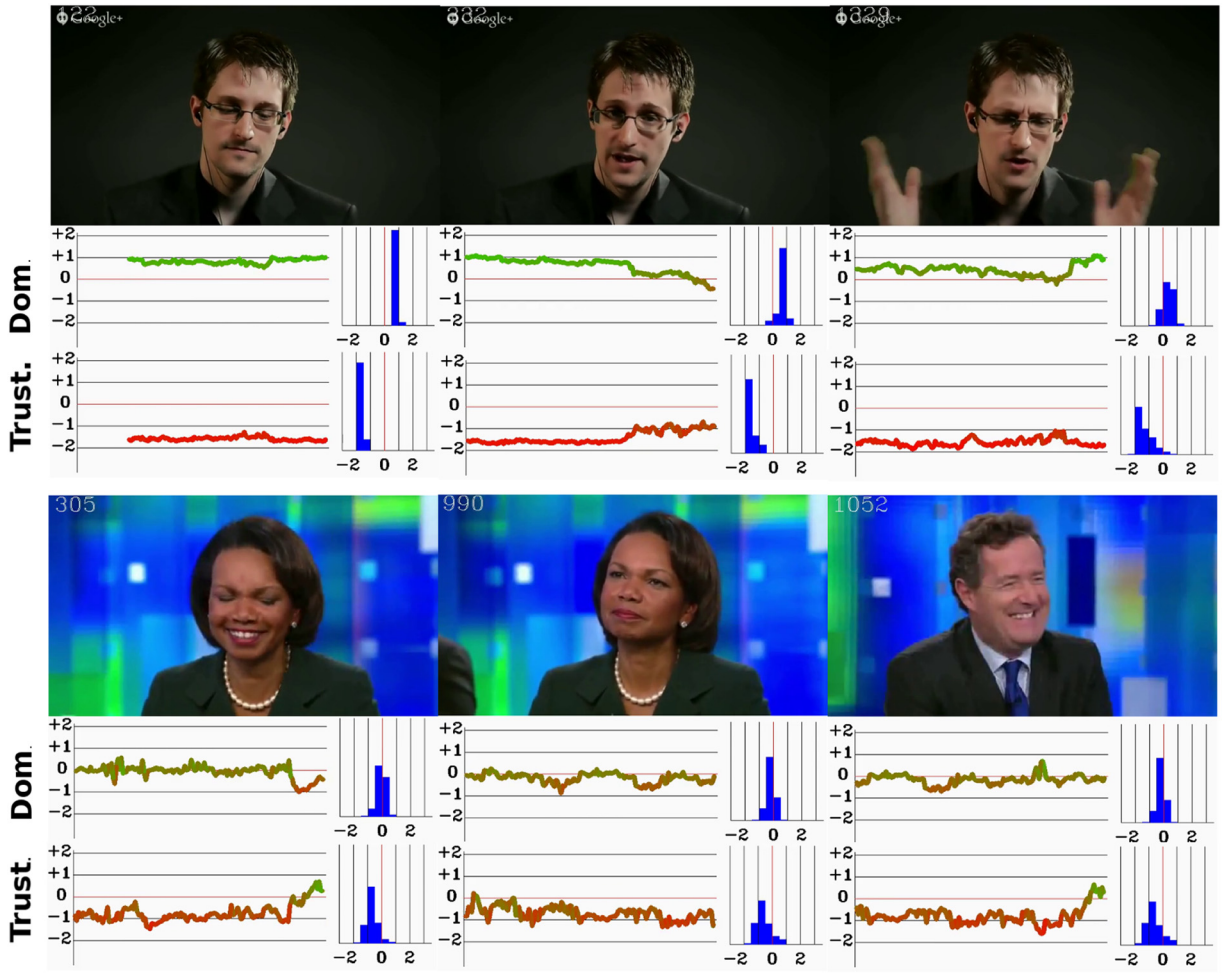

Fig. 12. Video processing. Frames taken from real time video processing examples. The scores are normalized with respect to the training data statistics and then displayed over time on a line plot and a histogram. These frames exemplify changes in predictions based on facial expression and head movement.

for subjective traits, small differences such as the number of annotations and the geographic and cultural differences of the annotators must be taken into consideration. Different cultures and languages affect the way people interpret traits, or the description of traits. Just as intriguing as the generalizations about people that we made in our work is the study of different cultures and focus groups. Models trained only on the annotations of a focus group could generalize to new data, enabling cross-culture comparisons — useful in research, marketing, political campaigning and more.

In systematically analyzing human judgments, it is also important to choose traits that best fulfill a purpose. In our case, Trustworthiness and Dominance are the best representations of the abstract judgments humans make about each other. IQ and Age, while not as fundamental in a psychological sense, still have conceivable applications, including the assessment of preconceived notions of intelligence and seniority — subtle social cues we often take for granted.

Code, data, and supplemental material for this paper can be found at: http://github.com/mel-2445/Predicting-First-Impressions.

Acknowledgments

M. McCurrie was supported by a gift from the Boeing Company. F. Beletti and L. Parzianello were supported by the Brazil Scientific Mobility Program. A. Westendorp was supported by NSF CNS RET Award #1609394. S. Anthony was supported in part by NSF SBIR Award #IIP-1621689. Hardware support was generously provided by the NVIDIA Corporation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imavis.2018.06.010.

References

[1] M.D. Alicke, R.H. Smith, M.L. Klotz, Judgments of physical attractiveness: the role of faces and bodies., Personal. Soc. Psychol. Bull. 12 (4) (1986) 381–389.
[2] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, Theano: a CPU and GPU math compiler in Python, SciPy, 2010. pp. 1–7.
[3] J. Bergstra, D. Yamins, D.D. Cox, Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms, SciPy, 2013. pp. 13–20.
[4] D.S. Bolme, J.R. Beveridge, M. Teixeira, B.A. Draper, The CSU face identification evaluation system: its purpose, features, and structure, International Conference on Computer Vision Systems, Springer. 2003, pp. 304–313.
[5] F. Chollet, Keras, 2015, https://github.com/fchollet/keras.
[6] A. Dantcheva, P. Elia, A. Ross, What else does your biometric data reveal? A survey on soft biometrics, IEEE T-IFS 11 (3) (2016) 441–467.
[7] M. Dumas, Emotional expression recognition using support vector machines, International Conference on Multimodal Interfaces, 2001.
[8] V. Falvello, M. Vinson, C. Ferrari, A. Todorov, The robustness of learning about the trustworthiness of other people, Soc. Cogn. 33 (5) (2015) 368.
[9] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, IEEE CVPR, 2009.

[10] L. Germine, K. Nakayama, B.C. Duchaine, C.F. Chabris, G. Chatterjee, J.B. Wilmer, Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments, Psychon. Bull. Rev. 19 (5) (2012) 847–857.
[11] B.A. Golomb, D.T. Lawrence, T.J. Sejnowski, SEXNET: A Neural Network Identifies Sex From Human Faces, NIPS, 1990.
[12] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber, B. Radig, Facial expression recognition with recurrent neural networks, International Workshop on Cognition for Technical Systems, 2008.
[13] G.B. Huang, M. Mattar, H. Lee, E. Learned-Miller, Learning to Align from Scratch, NIPS, 2012.
[14] C.F. Keating, A. Mazur, M.H. Segall, P.G. Cysneiros, J.E. Kilbride, P. Leahy, W.T. Divale, S. Komin, B. Thurman, R. Wirsing, Culture and the perception of social dominance from facial expression., J. Pers. Soc. Psychol. 40 (4) (1981) 615.
[15] B.F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, A.K. Jain, Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A, IEEE CVPR, 2015.
[16] M. Koestinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization, First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011.
[17] N. Kumar, A. Berg, P.N. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, IEEE T-PAMI 33 (10) (2011) 1962–1977.
[18] Y.H. Kwon, N. da Vitoria Lobo, Age classification from facial images, Comput. Vis. Image Underst. 74 (1) (1999) 1–21.
[19] C.H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, IEEE CVPR, 2009.
[20] G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, IEEE CVPR Workshops, 2015.
[21] Y. Lewenberg, Y. Bachrach, S. Shankar, A. Criminisi, Predicting Personal Traits from Facial Images using Convolutional Neural Networks Augmented with Facial Landmark Information, 2016.arXiv preprint arXiv:1605.09062.
[22] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, IEEE ICCV, 2015.
[23] P. Luo, X. Wang, X. Tang, A deep sum-product architecture for robust facial attributes analysis, IEEE ICCV, 2013.
[24] A. Mignault, A. Chaudhuri, The many faces of a neutral face: head tilt and perception of dominance and emotion, J. Nonverbal Behav. 27 (2) (2003) 111–132.
[25] A. Montillo, H. Ling, Age regression from faces using random forests, IEEE ICIP, 2009.
[26] A. Nech, I. Kemelmacher-Shlizerman, Megaface 2: 672,057 Identities for Face Recognition, 2016.
[27] N.N. Oosterhof, A. Todorov, The functional basis of face evaluation, PNAS 105 (32) (2008) 11087–11092.
[28] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep Face Recognition., BMVC, vol. 1, 2015. pp. 6.
[29] A.E. Pinkham, J.B. Hopfinger, K. Ruparel, D.L. Penn, An investigation of the relationship between activation of a social cognitive neural network and social functioning, Schizophr. Bull. 34 (4) (2008) 688–697.
[30] E. Rudd, M. Günther, T. Boult, MOON: A Mixed Objective Optimization Network for the Recognition of Facial Attributes, ECCV, 2016.
[31] W.J. Scheirer, N. Kumar, P.N. Belhumeur, T.E. Boult, Multi-attribute spaces: calibration for attribute fusion and similarity search, IEEE CVPR, 2012.
[32] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad–cam: Why Did You Say That? Visual Explanations from Deep Networks Via Gradient-based Localization, 2016.arXiv preprint arXiv:1610.02391.
[33] C. Senior, M. Phillips, J. Barnes, A. David, An investigation into the perception of dominance from schematic faces: a study using the World-Wide Web, Behav. Res. Methods Instrum. Comput. 31 (2) (1999) 341–346.
[34] B. Siddiquie, R.S. Feris, L.S. Davis, Image ranking and retrieval based on multi--attribute queries, IEEE CVPR, 2011.
[35] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, 2014.arXiv preprint arXiv:1409.1556.
[36] A. Todorov, S.G. Baron, N.N. Oosterhof, Evaluating face trustworthiness: a model based approach, Soc. Cogn. Affect. Neurosci. 3 (2) (2008) 119–127.
[37] A. Todorov, B. Duchaine, Reading trustworthiness in faces without recognizing faces, Cogn. Neuropsychol. 25 (3) (2008) 395–410.
[38] A. Todorov, M. Pakrashi, N.N. Oosterhof, Evaluating faces on trustworthiness after minimal time exposure, Soc. Cogn. 27 (6) (2009) 813–833.
[39] J. Wang, Y. Cheng, R.S. Feris, Walk and Learn: Facial Attribute Representation Learning from Egocentric Video and Contextual Data, 2016.arXiv preprint arXiv:1604.06433.
[40] J.S. Winston, B.A. Strange, J. O'Doherty, R.J. Dolan, Automatic and intentional brain responses during evaluation of trustworthiness of faces, Nat. Neurosci. 5 (3) (2002) 277–283.
[41] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, IEEE CVPR, 2011.
[42] K. Zhang, L. Tan, Z. Li, Y. Qiao, Gender and Smile Classification Using Deep Convolutional Neural Networks, IEEE CVPR Workshops, 2016.
[43] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, Panda: pose aligned networks for deep attribute modeling, IEEE CVPR, 2014. pp. 1637–1644.
[44] Y. Zhong, J. Sullivan, H. Li, Face attribute prediction using off-the-shelf CNN features, IAPR Int. Conf. on Biometrics, 2016.