Learning to see people like people

Amanda Song
University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093
feijuejuanling@gmail.com

Chad Atalla
University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093
catalla@ucsd.edu

Linjie Li
Purdue University
610 Purdue Mall, West Lafayette, IN 47907
li2477@purdue.edu

Garrison Cottrell
University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093
gary@ucsd.edu

Abstract

The human perceptual system can make complex inferences on faces, ranging from the objective evaluations regarding gender, ethnicity, expression, age, identity, etc. to subjective judgments on facial attractiveness, trustworthiness, sociability, friendliness, etc. Whereas the objective aspects have been extensively studied, less attention has been paid to modeling the subjective perception of faces. Here, we adapt 6 state-of-the-art neural networks pretrained on various image tasks (object classification, face identification, face localization) to predict human ratings on 40 social judgments of faces in the 10k US Adult Face Database. Supervised ridge regression on PCA of the conv5_2 layer in VGG-16 network gives best predictions on the average human ratings. Human group agreement was evaluated by repeatedly randomly splitting the raters into two halves for each face, and calculating the Pearson correlation between the two sets of averaged ratings. Due to this methodology, the models correlations with the average human ratings can exceed this score. We find that 1) model performance grows as the consensus on a face trait increases, and 2) model correlations are always higher than human correlations with each other. These results illustrate the learnability of the subjective perception of faces, especially when there is consensus, and the striking versatility and transferability of representations learned for object recognition. This work has strong applications to social robotics, allowing robots to infer human judgments of each other.

1. Introduction

Recent advances in deep convolutional networks have driven tremendous progress in a variety of challenging face processing tasks including face recognition[27], face alignment[39], and face detection[25]. However, humans not only read objective properties from a face, such as gender, expression, race, age and identity, but also form subjective impressions of the social aspects of a face[31, 32], such as facial attractiveness[28], friendliness, trustworthiness[29], sociability, dominance[18], and typicality. Despite the relative less attention received by the social perception of faces, social judgment is an important part of people’s daily interactions, and it has significant impact on social outcomes, ranging from electoral success to sentencing decisions[19, 35]. Whereas current computer vision techniques exceed human abilities at recognizing a face and identifying the objective properties of a face [27, 25], awareness of human subjective judgments is important for social robotics theory-of-mind inferences. Accurate predictions of social aspects of faces can help robots better understand how humans interact with and perceive each other, and can make a robot aware of inherent human biases, as these judgments rarely correspond to reality (except, perhaps, attractiveness)[32].

In this paper, we teach a machine to infer social impressions, that match human judgments, from faces. We examine a list of 20 pairs of social features that are typically studied by social psychologists, and that are relevant to social interactions between people [33, 32, 19]. Examples are attractiveness [7, 28, 5, 10, 8], trustworthiness [6, 29], sociability, aggressiveness [18], friendliness, kindness, happiness, familiarity [20], and memorability [2, 11]). Although social perceptions of faces are subjective, there is often a consensus among human raters in how they perceive facial attractiveness, trustworthiness and dominance[6, 5]. This indicates that faces contain high-level visual cues for social interactions, and therefore it is possible to model this process with machine learning techniques. We take advantage of the state-of-the-art neural network models trained for ob-
ject recognition and face recognition tasks and use their internal representations for social perception learning. In all 40 social dimensions, our model correlates with human averaged ratings better than the humans correlate with each other.

The contributions of the paper are summarized below:

- To the best of our knowledge, this work is the first attempt to systematically examine the consistency of human social perceptions of faces, to explore the landscape of social feature semantic space, and to predict human judgments of 40 social attributes of faces;
- We adapted 6 state of the art neural network algorithms trained for various visual tasks to make social judgment predictions on faces and achieve high correlations with human ratings in all 40 dimensions;
- We evaluate the tuning properties of nodes in the best network and visualize the patterns that maximally ignite the perceptions for each specific social dimensions to facilitate a better understanding of the neural networks’ behavior in face processing.

The rest of the paper is organized as follows. In Section 2, we review related work on social perception modeling. Section 3 and 4 summarize the methodology and the experimental framework. The experimental results and visualizations are presented in Section 5 and 6. Section 7 concludes the paper.

2. Related work

The focus of our paper is to infer as much social judgment information as possible from a face image and to predict the subjective impression of faces by learning from human group data. We review related work in terms of the visual features they use, the dataset they choose, the evaluation metric they adopt, and the social attributes they examine.

Visual features Since the early 1990s, psychologists have identified that high level visual features, such as the averageness of a face [14, 21] and the symmetry of the face [23] can explain why certain faces look more attractive.

Machine learning researchers have developed various computer vision features and models to predict social perceptions of faces, especially facial attractiveness. Yael et al. [5] used geometric ratios and distances between facial features based on facial landmarks to build an attractiveness predictor (0.65 correlation with human raters, face database size=184). End-to-end neural networks were applied to predict facial attractiveness in 2010 [8] (correlation 0.458, face database size=2056, young female faces only). Amit Kagian and his colleagues have used a combination of landmark-derived features along with global features to obtain a high correlation with human group averages on facial attractiveness [10] (0.82 Pearson correlation, face database size=91). Traditional computer vision features such as SIFT, HoG, Gabor filters have been blended to predict the relative ranking of facial attractiveness in [1] (rank order correlation 0.63, face database size=200). Rothe et al. incorporate collaborative filtering techniques with visual features extracted from pretrained VGG networks [24] to achieve individual-level prediction of facial attractiveness [22] (correlation 0.671 on female face queries, database size = 13,000). McCurrie et al. [17] build a model based on a pretrained VGG network to predict trustworthiness, dominance and IQ in faces ($R^2$ values on trustworthiness, dominance and IQ are 0.5687, 0.4601, 0.3548 respectively, face database size=6000). Previous papers have achieved correlations with human performance between 0.458 to 0.82 in attractiveness predictions, depending on the dataset and method used. However, to date, there is no standard dataset that has been used to compare these approaches.

Dataset Earlier studies employ datasets with relatively small numbers of faces (a few hundred) and most face datasets use young Caucasian faces only, as pointed out by [15]. In contrast, the MIT dataset [2] we use contains 2,222 high quality color images that vary in ethnicity, gender, age and expression, with ratings on 40 attributes. This dataset is smaller than two of the ones mentioned above. The first is collected from howhot.io, an online dating website [22] and contains 13,000 face images, but that work focused on personalized prediction of facial attractiveness, rather than average ratings. There are only binary choices (like or dislike) indicating implicit preference of facial attractiveness. The second one is collected from testmybrain.com, contains 6,000 grayscale face images [17], and includes just three social features: dominance, IQ and trustworthiness.

Evaluation metric Social perceptions of faces are collected from human participants in various ways. The most common way is to ask for a discrete rating, say from 1-9 [2], or 1-7 [5] from a number of raters, and then use the group average as the score for a face in the specified feature dimension (e.g. attractiveness). The consistency of ratings between humans is checked by repeatedly randomly splitting human participants into two subgroups and then computing the correlation between the two groups’ mean ratings. To compare model predictions with human ratings, Pearson correlation[10, 3], Spearman rank correlation[2] and R-squared values [17] are used, depending on the nature of the data. Another method is to present a pair of faces or multiple faces and ask for a relative ranking in a particular dimension (e.g., attractiveness). Prediction accuracy is measured using Kendall’s Tau and the Gamma Test [1]. In Rothe et al. [22], a person indicates his/her preference by choosing to like or dislike another user’s face photo. In this
In this section we first describe the dataset used in our experiments. Next, we introduce our method for predicting social perceptions of faces. Finally, we explain how we visualize the features that contribute most to social trait predictions.

3.1. Dataset

To predict how human evaluate social traits of a face at a glance, we use the dataset collected by Aude Oliva’s group [2]. The dataset consists of 2,222 images of faces sampled from the 10k US Adult Face Database and annotated for 20 pairs of social attributes. Each attribute is rated on a scale of 1-9 (1 means not at all, 9 means extremely) and each image is rated by 15 subjects. We take the average rating across all raters as a collective estimation of the social feature for every face.

The 20 pairs of social traits are: (attractive, unattractive), (happy, unhappy), (friendly, unfriendly), (sociable, introverted), (kind, mean), (caring, cold), (calm, aggressive), (trustworthy, untrustworthy), (responsible, irresponsible), (confident, uncertain), (humble, egotistical), (emotionally stable, emotionally unstable), (normal, weird), (intelligent, unintelligent), (interesting, boring), (emotional, unemotional), (memorable, forgettable), (typical, atypical), (familiar, unfamiliar) and (common, uncommon).

Clearly, some of these traits will be highly correlated, and are predictable from the others. We compute the Spearman’s rank correlation between every pair of social features and show their correlations in a heatmap (see the left figure in Figure 1). We put features together in the map based on similarity and positive/negativeness. From the figure, we can see that negative social features such as untrustworthiness, aggressiveness, cold, introverted, irresponsible form a correlated block, while most positive features such as attractive, sociable, caring, friendly, happy, intelligent, interesting, confident are highly correlated with each other. Although we chose 20 pairs of opposite features, they are not completely complementary and redundant. Principal component analysis shows that it takes 24 principal components to cover 95% of the variance.

3.2. Regression Model for Social Attributes

After we average human ratings on each face, each face receives a continuous score from 1 to 9 in all social dimensions. We model these social scores with a regression model. Our proposed algorithm is a ridge regression model on features extracted from deep convolutional neural networks (CNN). Since CNN features are usually high-dimensional, we first perform Principal Component Analysis (PCA) on the extracted features of the training set to reduce the dimensionality. The PCA dimensionality is chosen by cross-validation on a validation set, separately for each trait. The PCA weights are saved and further used in fine-tuning our CNN-regression model.
3.3. Feature Visualization

Attempting to understand what features are most helpful in social attribute prediction, we visualize the features extracted from the CNN. Two different methods were proposed in past feature visualization studies: dataset-centric methods [36, 38], and a network-centric method [36, 37].

The dataset-centric method we employed is to display image patches from the training set that cause high activation for the feature units and use the deconvolution method to highlight the portions of the image that are responsible for firing the important feature neurons [36, 38].

The network-centric approach is usually used in classification networks. This method produces an image that is based on adapting the input by maximizing the output category activation using gradient ascent, i.e., it is mainly a function of the network [36, 37]. The key idea is to optimize the input image so that the target neuron can be highly activated. We apply this idea to the output (regression) neuron as well as the top nine neurons that influence that output individually.

4. Experimental framework

In this section, we report our experimental framework using 6 CNN-based regression models with respect to two baselines, human correlation between groups of raters, and a baseline model using the geometric features.

4.1. Baseline I: Human Correlations

Since these social attributes are all subjective perceptions rated by people, it is informative to examine to what extent people agree with each other upon those social judgments. We performed the following procedure 50 times for each attribute and then averaged the results:

1. For each face, we randomly split the 15 raters evenly into two groups of 7 and 8. (Note: the raters for each face will, in general, be different sets).
2. We calculate the two group’s average ratings for each face, obtaining two vectors of length 2,222 (there are 2,222 faces in the dataset).
3. We calculate the correlation between the two vectors.

The results are shown in the second column of Table 1. For every social attribute, the averaged correlation between human subgroups serves as an index of the rating consistency.

4.2. Baseline II: Regression on geometric features

Past studies on facial attractiveness have found that attractiveness can be inferred from the geometric ratios and configurations of a face [5, 10]. We suggest that other social attributes can also be inferred from geometric features. We compute 29 geometric features based on definitions described in [16] and further extract a “smoothness” feature and skin color features according to the procedure in [5, 10]. The “smoothness” of a face was evaluated by applying a
Canny edge detector to windows from the cheek/forehead area [5]. The more edges detected by edge detectors within the window, the less smooth the skin is. The regions we chose to compute smoothness and skin color are highlighted in the right subplot of Figure 2). The “skin color” feature is extracted from the same window as "smoothness", converted from RGB to HSV. Regressing on these handcrafted features alone are not enough to capture the richness of geometric details about a face, we therefore use a computer vision library (dlib, C++) to automatically label 68 face landmarks (see Figure 2) for each face and compute distances and slopes between any two landmarks. Combining 29 handcrafted geometric features, smoothness, color and the distance-slope features, we obtain 4592 features in total. Since the features are highly correlated, we apply PCA to reduce dimensionality. Again, the PCA dimensionality is chosen by cross-validating on the hold out set separately for each facial attribute. Then a ridge regression model is applied to predict social attribute ratings of a face. The hyperparameter of ridge regression is selected by leave-one-out validation within the training set.

Figure 2: 68 face landmarks labeled by dlib software automatically. The gray regions are the locations used for computing smoothness and skin color.

4.3. CNN-based Regression Model

We initially compared six neural network architectures: (1) VGG16, (2) VGG-Face from the Oxford Visual Geometry Groups VGG networks [24], (3) AlexNet (the publicly available CaffeNet reference model) [13] (4) Inception from Google [26] (5) a shallow face identification Siamese neural network that we trained from scratch: Face-SNN and (6) a state of the art VGG-derived network trained for the face landmark localization task: Face-LandmarkNN. These comparisons were performed with the Caffe deep learning framework [9].

To find the best CNN to predict social attribute ratings among all six networks, we first find the best-performing feature layers of each network (with the ridge regression model), and then we compare the results among the networks to select the best network. For each layer of each network, before the ridge regression, we performed PCA and picked the PCA dimension that gave the best results on the validation set.

5. Results

Surprisingly, we found that features from conv5_2 layer of VGG16 trained for object classification slightly outperformed the AlexNet and Inception networks, while the three networks trained solely on faces, VGG-Face, Face-LandmarkNN and Face-SNN did not achieve performance as competitive as the other three for most of the social attributes. The best performing VGG16 layer was conv5_2.

We speculate that the reason for the relatively poorer performance of the face recognition networks is that they are optimized either to learn differences between faces which define identity or to learn the face landmark configurations, whereas for this task at hand, we are looking for commonalities behind certain social features which go beyond identity. The landmark network presumably should give results similar to the geometric features, but did not learn features corresponding to all of the features we used in that model. These speculations need to be checked, of course, for example by trying to predict all of our measured features, using the landmark network, but we did not do that here.

We tried fine-tuning the model as follows. We used back-propagation to fine tune the weights into the conv5_2 layer, the weights to the PCA layer from conv5_2 (initialized by the PCA weights), and the weights from the PCA layer to the output regression unit. However, this fine-tuning did not improve performance, so the results reported in Table 1 are without fine-tuning.

We evaluate the performance on 50 random train/validation/test splits of the data with a 64/16/20 percent split for training, cross-validation and testing, respectively. The prediction performance of our model is evaluated using Pearson’s correlation with the human ratings on the test set. For each social attribute, we report its human consistency as described in Section 4.1.

Table 1 summarizes the prediction performance of our model for all the social attributes compared to Baseline I and II. The table is organized in a descending order of human agreement on the putative positive attribute of the paired attributes. The three attributes where there is greater agreement among humans for the negative component of the pair are bolded.

Among all the social attributes, human subjects agree most with each other about "happy" and disagree most about "unfamiliar." For both regression models (Baseline II and our model), model performance grows as the consensus on a social trait increases and human correlations with each other are consistently lower than the models’ correlations with the average human ratings. Normally, one might consider the human correlations to be an upper bound on performance, but here they are different kinds of correlations.

Since the change in expression would produce a change in landmark locations, it is not surprising that landmark-
based geometric features (Baseline II) achieve comparable or slightly higher correlation as our model for predicting those social attributes that are highly related to expressions (such as "happy", "unhappy", "cold" and "friendly" etc.). While for other social attributes, our model slightly outperforms landmark-based geometric features by about 0.04 correlation on average and significantly outperforms human correlation by about 0.12 correlation on average. This implies that CNN features encode much more information than just landmark-based features. It is essential to visualize those features and understand what features extracted from CNN make our model powerful enough to predict social attributes.

To quantitatively compare the face social features perceived by humans and those predicted by our best performing model, we take the model predictions on all social features, and compute the Spearman correlation between every pair in the set (see the right figure in Figure 1). Not surprisingly, this has very similar patterns compared the heatmap generated from human ratings (see the right panel in Figure 1). Pearson Correlation between the upper triangle of the two similarity matrices (human and model prediction) is 0.9836. However, note that each predictor was trained independently.

6. Feature Visualization

In this section, we visualize the features that are of importance to social perceptions. We choose facial attractiveness as an example. The same method can be applied to the other social features. We employ the two methods described in section 3.3 to visualize features learned by our model.

6.1. Data-centric Visualization

To identify visual features that ignite attractiveness perception, we find the top 9 units of highest influence on attractiveness at conv5_2 as follows. First, we compute a product of three terms: (1) A unit’s activation from conv5_2, (2) that unit’s weight to the following fc_pca layer, (3) the fc_pca unit’s weight to the output unit. We then sort all conv5_2 units’ average products of the three terms and identify the top 9 neurons as the ones that contribute most to the output neuron for the corresponding social feature. Then we employ the method described in [36, 38] to find top-9 input images that cause high activations in each of the top-9 conv5_2 neurons. Also we further produce the deconvolutional images by projecting each activation separately down to pixel space.

Figure 3 captures the features that are important to predict the attractiveness of a face. The feature importance descends from left to right and top to bottom. The important features identified by our model are related to eyes, hair with bangs, high nose-bridge, high cheeks, dark eyebrows, strong commanding jawline, chin and red lips. Note that among the 9 cropped input image patches, not all the faces are perceived as attractive or rated as attractive. An attractive face needs to activate more than one feature in order to be considered attractive. This observation agrees with our intuition that attractiveness is a kind of holistic judgment, requiring a combination of multiple features.

It also seems to be the case that several of the features

| Social Attributes | Baseline I | Baseline II | Our Model |
|-------------------|-----------|------------|-----------|
| happy             | 0.84      | 0.86       | 0.84      |
| unhappy           | 0.75      | 0.81       | 0.80      |
| friendly          | 0.78      | 0.83       | 0.82      |
| unfriendly        | 0.72      | 0.80       | 0.79      |
| sociable          | 0.74      | 0.78       | 0.78      |
| introverted       | 0.50      | 0.64       | 0.65      |
| attractive        | 0.72      | 0.66       | 0.75      |
| unattractive      | 0.62      | 0.62       | 0.70      |
| kind              | 0.72      | 0.79       | 0.79      |
| mean              | 0.69      | 0.75       | 0.73      |
| caring            | 0.72      | 0.78       | 0.79      |
| cold              | 0.71      | 0.81       | 0.79      |
| trustworthy       | 0.62      | 0.72       | 0.73      |
| untrustworthy     | 0.60      | 0.69       | 0.70      |
| responsible       | 0.58      | 0.65       | 0.70      |
| irresponsible     | 0.55      | 0.64       | 0.67      |
| confident         | 0.55      | 0.55       | 0.61      |
| uncertain         | 0.45      | 0.62       | 0.63      |
| humble            | 0.55      | 0.64       | 0.63      |
| egotistic         | 0.52      | 0.62       | 0.62      |
| emotionally stable| 0.53      | 0.64       | 0.67      |
| emotionally unstable| 0.50    | 0.62       | 0.64      |
| normal            | 0.49      | 0.58       | 0.61      |
| weird             | 0.52      | 0.50       | 0.56      |
| intelligent       | 0.49      | 0.53       | 0.62      |
| unintelligent     | 0.43      | 0.53       | 0.58      |
| interesting       | 0.42      | 0.64       | 0.67      |
| boring            | 0.39      | 0.54       | 0.60      |
| calm              | 0.41      | 0.47       | 0.50      |
| aggressive        | 0.65      | 0.72       | 0.72      |
| emotional         | 0.33      | 0.60       | 0.60      |
| unemotional       | 0.56      | 0.76       | 0.75      |
| memorable         | 0.30      | 0.38       | 0.48      |
| forgettable       | 0.27      | 0.40       | 0.48      |
| typical           | 0.28      | 0.41       | 0.43      |
| atypical          | 0.24      | 0.40       | 0.43      |
| common            | 0.25      | 0.37       | 0.40      |
| uncommon          | 0.27      | 0.38       | 0.40      |
| familiar          | 0.24      | 0.42       | 0.44      |
| unfamiliar        | 0.18      | 0.40       | 0.44      |

Table 1: Prediction performance of all the social attributes. The reported performance is averaged on 50 random train/validation/test splits of the data.
include relationships between the parts. For example, while
the first feature in the upper left of the figure emphasizes
the eye, it also includes the nose. This is also true of the
upper right feature. Smiling is also important in order to
be perceived attractive, as emphasized by the feature in the
lower left of the figure.

6.2. Network-centric Visualization

In section 6.1, we have identified the top-9 units and their
feature maps from the conv5_2 layer that maximally activates
the attractiveness neuron. Here, we use the gradient-ascent
method to optimize the input image that would highly ac-
tivate a specific neuron of the network. This method is
also performed on the pretrained-VGG16 regression model,
which is trained to predict attractiveness.

Figure 4a shows the optimized image corresponding to
the output neuron from a random input image. Optimiz-
ing the input image for the output neuron of a regression
model does not result in a particularly interpretable figure,
although it does appear to emphasize the eyes. Our sec-
ond approach is to optimize the input image with respect to
the top-9 contributing neurons from conv5_2 layer that have
been identified in section 6.1. Figure 4b presents 9 opti-
mized images with respect to the corresponding top-9 fea-
ture maps of the top-9 neurons from conv5_2 layer. Since
we use a pretrained-VGG16 network for visualization, it is
not surprising that the corresponding top-9 feature maps at
conv5_2 layer are not particularly encoding facial patterns.

We also present the optimized image initialize with a
face image, along with the original face image for compari-
son in Figure 5. The optimized image tends to highlight the
eyes, nose, cheeks and the contour of the face, which is con-
sistent with the features identified by data-centric method.

7. Conclusion

We have shown that a deep network can be used to pre-
dict human social judgments with high correlation with the
average human ratings. As far as we know, this is the
widest exploration of social judgment predictions, showing
human-like perceptions on 40 social dimensions. Unsur-
prisingly, given previous work recognizing facial expres-
sions, where happiness is the easiest to recognize, our high-
est correlation is on the happy feature. However, previous
work in this area tended to classify a face as happy or not,
rather than the degree of rated happiness.

We find that for attributes that correspond to elements
of the face that require muscle movement, or a lack of it
(such as happy, unhappy, cold, aggressive, unemotional)
a simple regression model based on the placement of facial
landmarks works well. For ones that don’t appear to suggest
emotions, such as friendly, note that friendly and happy are
highly correlated (see Figure 1, and the red block indexed
by happy and friendly). Similarly, aggressive and mean are
highly correlated, which presumably requires not smiling.

Perhaps of more significance are the correlations with
judgments of traits, such as trustworthiness, responsi-

bility, confidence, and intelligence, which would correspond
to more static features of the face. In this area, the deep
network, which responds to facial textures as well as shape,
has superior performance. While these judgments do not
correspond to a notion of "ground truth," they are things for
which humans have a fair amount of agreement, suggesting
that there is a signal to be recognized.

Of further note is that we have shown, yet again, that a
machine can recognize attractiveness, presumably without
any hormonal influences. For this dataset, our deep net-
work correlates with human ratings at 0.75. This provides a
benchmark for this dataset.

Finally, it is of note that we can see that some of the
traits considered to be "opposite" in this list are not simply
the reverse of one another. For example, there is a large
difference in human agreements on "sociable" (0.74) versus
"introverted" (0.50), suggesting they are not opposites.

These results are significant for the field of social
robotics. While a robot should not judge a human based
completely on their appearance, it can be useful knowledge
that humans might judge a person to be trustworthy, while
the robot can be more objective. Similarly, a robot need
do not treat an attractive and unattractive person differently,
but this knowledge could affect how the robot interacts with
the unattractive person, knowing in advance that this person
may have had many negative experiences interacting with
people.

In this paper, we train each social feature separately, due
to their varied consistency and reliability. In the future, it
is worth trying to train one single convnet to learn multiple
tasks simultaneously and evaluate whether shared represen-
tation may further improve the model performance.

In summary, we have provided the first machine learn-
ing system to learn subjective human judgments of a wide
spectrum of traits. We found that the more humans agree on
such subjective judgments, the more the system could pick
up on the features driving those judgments. It will be of in-

terest to investigate further what those features are, beyond
the attractiveness features we displayed here.

One step further from predicting the value in a certain
social feature is to move faces on the social manifold and
to increase a face’s elicited social perceptions in positive
ways (e.g. to make a face look more sociable/ trustworthy/
attractive). Although the images generated by our current
visualization method are still far away from being photoreal-

istic, it may be a fruitful area in the future to develop
generative models that can achieve this goal.
Figure 3: Visualization of features in the pretrained-VGG16 regression network. For conv5_2 layer, we show the top 9 activations of the top 9 neurons that maximally activate the attractiveness neuron across the training data, projected down to pixel space using the deconvolutional network approach [38] and their corresponding cropped image patches. Best viewed in electronic form, and zoomed in.

References

[1] H. Altwaijry and S. Belongie. Relative ranking of facial attractiveness. In Applications of Computer Vision (WACV), 2013 IEEE Workshop on, pages 117–124. IEEE, 2013. 2
[2] W. A. Bainbridge, P. Isola, and A. Oliva. The intrinsic memorability of face photographs. Journal of Experimental Psychology: General, 142(4):1323, 2013. 1, 2, 3
[3] A. Dantcheva and J. Dugelay. Female facial aesthetics based on soft biometrics and photo-quality. In Proceedings of ICME, 2011. 2
[4] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1657–1664. IEEE, 2011. 3
[5] Y. Eisenthal, G. Dror, and E. Ruppin. Facial attractiveness: Beauty and the machine. Neural Computation, 18(1):119–142, 2006. 1, 2, 4, 5
[6] V. Falvello, M. Vinson, C. Ferrari, and A. Todorov. The robustness of learning about the trustworthiness of other people. Social Cognition, 33(5):368, 2015. 1
[7] K. Grammer and R. Thornhill. Human (homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness. Journal of comparative psychology, 108(3):233, 1994. 1
[8] D. Gray, K. Yu, W. Xu, and Y. Gong. Predicting facial beauty without landmarks. In Computer Vision–ECCV 2010, pages 434–447. Springer, 2010. 1, 2
[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014. 5
[10] A. Kagian, G. Dror, T. Leyvand, I. Meilijson, D. Cohen-Or, and E. Ruppin. A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. Vision research, 48(2):235–243, 2008. 1, 2, 4
[11] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva. Modifying the memorability of face photographs. In Computer Vision (ICCV), 2013 IEEE International Conference on, pages 3200–3207. IEEE, 2013. 1
[12] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In European conference on computer vision, pages 472–488. Springer, 2014. 3
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012. 5
[14] J. H. Langlois and L. A. Roggman. Attractive faces are only average. Psychological science, 1(2):115–121, 1990. 2
[15] A. Laurentini and A. Bottino. Computer analysis of face beauty: A survey. Computer Vision and Image Understanding, 125:184–199, 2014. 2, 3
[16] D. S. Ma, J. Correll, and B. Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. Behavior research methods, 47(4):1122–1135, 2015. 4
Figure 4: Visualization of features using network-centric method. To produce Fig 4a, we use gradient ascent to optimize the output neuron. Fig 4b shows 9 optimized images for the feature maps corresponding to top-9 contributor neurons from conv5_2 layer.

Figure 5: Visualization of the optimized image with a input face image: Figure 5a is the original face image before optimization. Figure 5b is produced by performing optimization with respect to the output unit.

[17] M. McCurrie, F. Beletti, L. Parzianello, A. Westendorp, S. Anthony, and W. Scheirer. Predicting first impressions with deep learning. arXiv preprint arXiv:1610.08119, 2016.
[18] A. Mignault and A. Chaudhuri. The many faces of a neutral face: Head tilt and perception of dominance and emotion. Journal of nonverbal behavior, 27(2):111–132, 2003.
[19] N. N. Oosterhof and A. Todorov. The functional basis of face evaluation. Proceedings of the National Academy of Sciences, 105(32):11087–11092, 2008.
[20] M. Peskin and F. N. Newell. Familiarity breeds attraction: Effects of exposure on the attractiveness of typical and distinctive faces. PERCEPTION-LONDON-, 33(2):147–158, 2004.
[21] G. Rhodes, L. Jeffery, T. L. Watson, C. W. Clifford, and K. Nakayama. Fitting the mind to the world face adaptation and attractiveness aftereffects. Psychological science, 14(6):558–566, 2003.
[22] R. Rothe, R. Timofte, and L. Van Gool. Some like it hot: visual guidance for preference prediction. arXiv preprint arXiv:1510.07867, 2015.
[23] J. E. Scheib, S. W. Gangestad, and R. Thornhill. Facial attractiveness, symmetry and cues of good genes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 266(1431):1913–1917, 1999.
[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[25] R. Stewart, M. Andriluka, and A. Y. Ng. End-to-end people detection in crowded scenes. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.
[27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1701–1708, 2014.
[28] R. Thornhill and S. W. Gangestad. Facial attractiveness. Trends in cognitive sciences, 3(12):452–460, 1999.
[29] A. Todorov, S. G. Baron, and N. N. Oosterhof. Evaluating face trustworthiness: a model based approach. Social cognitive and affective neuroscience, 3(2):119–127, 2008.
[30] A. Todorov, R. Dotsch, J. M. Porter, N. N. Oosterhof, and V. B. Fulvello. Validation of data-driven computational models of social perception of faces. Emotion, 13(4):724, 2013.
[31] A. Todorov, P. Mende-Siedlecki, and R. Dotsch. Social judgments from faces. *Current opinion in neurobiology*, 23(3):373–380, 2013. 1, 3

[32] A. Todorov, C. Y. Olivola, R. Dotsch, and P. Mende-Siedlecki. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Psychology*, 66(1):519, 2015. 1, 3

[33] A. Todorov, C. P. Said, A. D. Engell, and N. N. Oosterhof. Understanding evaluation of faces on social dimensions. *Trends in cognitive sciences*, 12(12):455–460, 2008. 1

[34] Y. Wang and G. W. Cottrell. Bikers are like tobacco shops, formal dressers are like suits: Recognizing urban tribes with caffe. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 876–883. IEEE, 2015. 3

[35] 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. 1

[36] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 4, 6

[37] S. Yu and K. Zipser. A deep neural net trained for person categorization develops both detailed local features and broad contextual specificities. *Journal of Vision*, 16(12):411–411, 2016. 4

[38] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014. 4, 6, 8

[39] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. *arXiv preprint arXiv:1511.07212*, 2015. 1