Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Predicting Personal Traits from Facial Images Using Convolutional Neural Networks Augmented with Facial Landmark Information

Yoad Lewenberg
The Hebrew University of Jerusalem, Israel

Yoram Bachrach
Microsoft Research Cambridge, UK

Sukrit Shankar
Machine Intelligence Lab (MIL), Cambridge University

Antonio Criminisi
Microsoft Research Cambridge, UK

Abstract

We consider the task of predicting various traits of a person given an image of their face. We aim to estimate traits such as gender, ethnicity and age, as well as more subjective traits as the emotion a person expresses or whether they are humorous or attractive. Due to the recent surge of research on Deep Convolutional Neural Networks (CNNs), we begin by using a CNN architecture, and corroborate that CNNs are promising for facial attribute prediction. To further improve performance, we propose a novel approach that incorporates facial landmark information for input images as an additional channel, helping the CNN learn face-specific features so that the landmarks across various training images hold correspondence. We empirically analyze the performance of our proposed method, showing consistent improvement over the baselines across traits. We demonstrate our system on a sizeable Face Attributes Dataset (FAD), comprising of roughly 200,000 labels, for 10 most sought-after traits, for over 10,000 facial images.

Introduction

Humans find it very easy to determine various traits of other people, simply by looking at them. Without almost any conscious effort, a glimpse at another person's face is sufficient for us to ascertain their gender, age or ethnicity. We can also easily decide whether they are attractive, look funny or are approachable; and can even figure out the emotion they display, like whether they appear sad, happy or surprised. As social creatures, making such inference is clearly important to us. Imparting commensurate capabilities to machines is bound to enable very interesting applications. However, in contrast to the relative ease in which humans infer such personal traits of an individual from their facial image, training a machine to do the same is a challenging task.

Related Work

Most prior methods for facial attribute analysis have been based on hand-crafted features such as color histograms and histogram of oriented gradients (Kumar et al. 2009). Despite their reasonable success, these solutions typically suffer from one or more of the following problems: (a) they are specifically tailored to a single task at hand; (b) they are not resistant to real-world variations in data such as multiple view-points (c) they use non-automatic pre-processing methods such as hand-labeling of key facial regions.

Deep learning based procedures can *automatically* learn a diverse set of low and high-level representations directly from the image data, and thus overcome most problems posed by approaches based on hand-designed features. Also, deep nets have been proven to build large-scale applications (Krizhevsky, Sutskever, and Hinton 2012). Given the promise of deep learning and the nature of our problem where we aim to predict attributes ranging from objective to very subjective ones, from a diverse set of facial images, we resort to CNN based models for designing our algorithm.

Face Attributes Dataset (FAD) and Analysis

We present a sizeable Face Attributes Dataset (FAD) for evaluation purposes. The dataset has been created using crowd-sourcing from Amazon Mechanical Turk (AMT), with several redundant labels aggregated using majority vote (Kosinski et al. 2012; Bachrach et al. 2012a). [1] Our dataset consists of 10,000 facial images of celebrities (public figures), where each image is tagged with various traits of the individual. We choose 10 most sought-after facial attributes on Google and Bing, as listed in Table 1. As our target variables, we have considered the traits: ender, ethnicity, age, make-up and hair color, emotional expression, attractiveness, humorousness and chubbiness. The dataset contains roughly 200,000 attribute labels.

Methodology: We begin by applying one of the most famous CNN architectures (Krizhevsky, Sutskever, and Hinton 2012) for our prediction task, widely known as AlexNet (Fig 1). Since in our problem, an image can have multiple traits (labels), training a CNN with sigmoid cross-entropy loss function (Jia et al. 2014) comes as a natural choice. However, we find that for our prediction task where the facial attributes are very subtle to detect, training with a multi-label loss does not give good performance even after repeated manual tuning of the weight regularization parameter. We thus train one net for each given trait by applying a softmax loss function, since the classes for any given trait cannot co-occur. This achieves better generalized per-

[1] For simplicity, we avoided other aggregation methods (Welinder et al. 2010; Bachrach et al. 2012b; Venanzi et al. 2014).

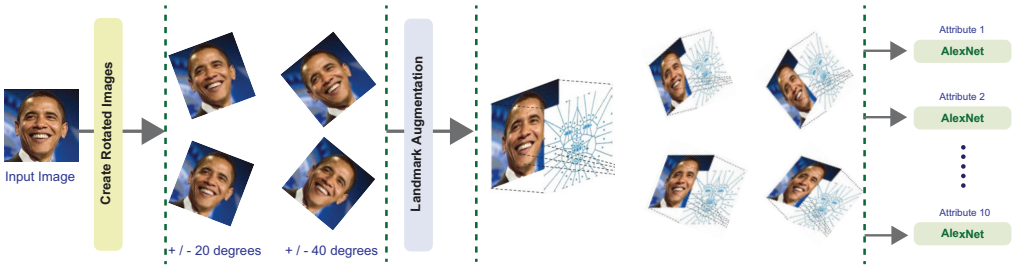

Figure 1: *Illustration of our proposed Landmark Augmented CNN:* Each input image is first data-supplemented, i.e. its four rotated versions are created. Next, the RGB channels of all input images (including the additional rotated ones) are augmented with one more channel which encodes the facial landmark information. The 4-channel input images are fed to AlexNets for training, one for each trait. Landmark augmentation improves the prediction accuracy of facial attributes as shown in Table 1.

Trait	Data distribution	Baseline	LACNN
Gender (Male)	Y (0.51), N (0.49)	**98.46%**	98.33%
Ethnicity (White)	Y (0.79), N (0.21)	82.7%	**83.35%**
Hair Color (Dark)	Y(0.60), N (0.40)	91%	**91.69%**
Makeup	Y (0.39), N (0.61)	92.5%	**92.87%**
Age (Young)	Y (0.68), N (0.32)	88.42%	**88.83%**
Emotions (Joy)	Y (0.64), N (0.36)	**88.93%**	88.33%
Attractive	Y (0.66), N (0.34)	78.44%	**78.85%**
Humorous	Y (0.56), N (0.44)	66.8%	**69.06%**
Chubby	Y (0.57), N (0.43)	60.6%	**61.38%**

Table 1: *Attributes / Traits in FAD and Comparison of Prediction Accuracy:* Personal traits in FAD along with the corresponding classes are listed (Y = Yes, N = No), along the data class distribution. Some traits have more skewness across their classes as compared to others. LACNN improves the prediction accuracy as compared to the AlexNet baseline for most traits.

formance, and hence, we establish our baseline and perform all our experiments with this choice.

To account for pose variation, for each training image, we create 4 new training images as its rotated versions. Each training image is rotated by $\{-40, -20, 20, 40\}$ degrees (Fig 1). Data augmentation is done both for baselines and our proposed approach in order to draw a fair comparison.

Our proposed approach is based on incorporating facial landmark information in the input data. Facial landmark localization algorithms are designed to find the location of several key "landmarks" in an image, such as the location of the center of the eyes, parts of the nose or the sides of the mouth. Consider a list $L = (l_1, \ldots, l_k)$ of facial landmarks. Facial landmark localization algorithms receive a facial image I as an input, and output the coordinates in the image for each of the landmarks $C^I = (c_1^I, \ldots, c_k^I)$ where $c_j^I = (x_j^I, y_j^I)$ are the coordinates of landmark l_j in the image I.

Our approach operates by associating each pixel in the facial image with the *closest* facial landmark for that image. We then add this association as an additional channel to each input image. Since we use facial landmark localization algorithm just as a subroutine, so any such algorithm could be used. In our experiments, we have used TCDCN (Zhang et al. 2014) as the facial landmark detector.

Experiments: We used FAD for all our experiments with a 80/20 train/test split, evaluating our method on roughly 36,000 labels. For training and doing inference with CNNs, we have used the Caffe Library (Jia et al. 2014). Table 1 shows the prediction accuracy of the baseline CNN method and LACNN on FAD. It is clear that CNN provides an overall impressive baseline performance, and the proposed LACNN depicts consistent improvements over it. Note that LACNN has the capability to improve performance for both the objective as well as the subjective traits. This substantiates our intuition that face alignment information can prove efficient in predicting facial attributes using deep nets.

Conclusions: We have proposed a method for predicting personal attributes from facial images, based on a CNN architecture augmented with face alignment information. Experiments on a new dataset, called FAD, show consistent improvement across various traits with our method.

References

Bachrach, Y.; Graepel, T.; Kasneci, G.; Kosinski, M.; and Van Gael, J. 2012a. Crowd iq: aggregating opinions to boost performance. In *AAMAS*.

Bachrach, Y.; Graepel, T.; Minka, T.; and Guiver, J. 2012b. How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. *ICML*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ICM*, 675–678. ACM.

Kosinski, M.; Bachrach, Y.; Kasneci, G.; Van-Gael, J.; and Graepel, T. 2012. Crowd iq: Measuring the intelligence of crowdsourcing platforms. In *ACM Web Sciences*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and Simile Classifiers for Face Verification. In *ICCV*.

Venanzi, M.; Guiver, J.; Kazai, G.; Kohli, P.; and Shokouhi, M. 2014. Community-based bayesian aggregation models for crowdsourcing. In *WWW*.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. J. 2010. The multidimensional wisdom of crowds. In *NIPS*.

Zhang, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Facial landmark detection by deep multi-task learning. In *ECCV*.