Face Attributes in the Wild

Hongzhi Liu

May 29, 2018

Abstract

As we all know, predicting face attributes in the wild is a challenge due to complex face variations such as different poses, bright or dark lightings and occlusions. Today, I read a thesis written by Ziwei Liu, who studys at Department of Information Engineering in the Chinese University of Hong Kong. His team propose a novel deep learning framework for attribute prediction in the wild. This framework not only outperforms the state-of-the-art with a large margin, but also reveals valuable facts on learning face representation.

1. Overview of Deep Learning Framework

Deep learning recently achieved great success in attribute prediction due to their ability to learn compact and discriminative features. Zhang *et al.*[6] showed that better performance can be achieved by ensembling learned features of multiple pose-normalized CNNs. The main drawback of the method is that it relies on accurate landmark detection and pose estimation in both training and testing steps. Recent research shows that face localization and alignment are still not well solved problems especially in the wild condition, although much progress has been achieved in the past decade.

In this paper, Dr. Liu proposes a novel deep learning framework, which combines *massive objects* and *massive identities* to pre-train two CNNs for face localization and attribute prediction respectively [5]. Besides, A novel fast feed-forward algorithm for CNN with *locally shared filters* is devised. They also contribute a large facial attribute database with more than eight million attribute labels and it is 20 times larger than the largest publicly available dataset.

2. Framework for the Approach

This work revisits global methods by proposing a novel deep learning framework, which integrates two CNNs, LNet and ANet, where LNet locates the *entire face region* and ANet extracts high-level face representation from the located region. Fig. 1 illustrates our pipeline where LNet locates the entire face region in a coarse-tofine manner as shown in (a) and (b), while ANet extracts features for attribute recognition as shown in (c).

Different from existing works that rely on accurate face and landmark annotations, LNet is trained in a weakly supervised manner with only image-level annotations. Specifically, it is pre-trained with one thousand object categories of ImageNet [2] and fine-tuned by image-level attribute tags. The former step accounts for background clutters, while the latter step learns features robust to complex face variations. Learning LNet in this way not only significantly reduces data labeling, but also improves the accuracy of face localization. Both $LNet_o$ and $LNet_s$ have network structures similar to AlexNet, whose hyper parameters are specified in Fig. 1 (a) and (b) respectively. The fifth convolutional layer (C5) of $LNet_o$ indicates headshoulders while C5 of $LNet_s$ indicates faces, with their highly responsed regions in their averaged response maps. Moreover, the input x_o of $LNet_o$ is a $m \times n$ image, while the input x_s of $LNet_s$ is the head-shoulder region.

As illustrated in Fig. 1 (c), ANet is learned to predict attributes y by providing the input face region x_f , which is detected by $LNet_s$ and properly resized. Specifically, multiview versions of x_f are utilized to train ANet. Furthermore, ANet contains four convolutional layers, where the filters of C1 and C2 are globally shared and the filters of C3 and C4 are locally shared. The effectiveness of local filters have been demonstrated in many face related tasks. To handle complex face variations, ANet is pre-trained by distinguishing massive face identities, which facilitates the learning of discriminative features.

Fig. 1 (d) outlines the procedure of attribute recognition. ANet extracts a set of feature vectors (FCs) by cropping overlapping patches on x_f. An efficient feed-forward algorithm is developed to reduce redundant computation in the feature extraction stage. SVMs [3] are trained to predict attribute values given each FC. The final prediction is obtained by averaging all these values to cope with small misalignment of face localization.

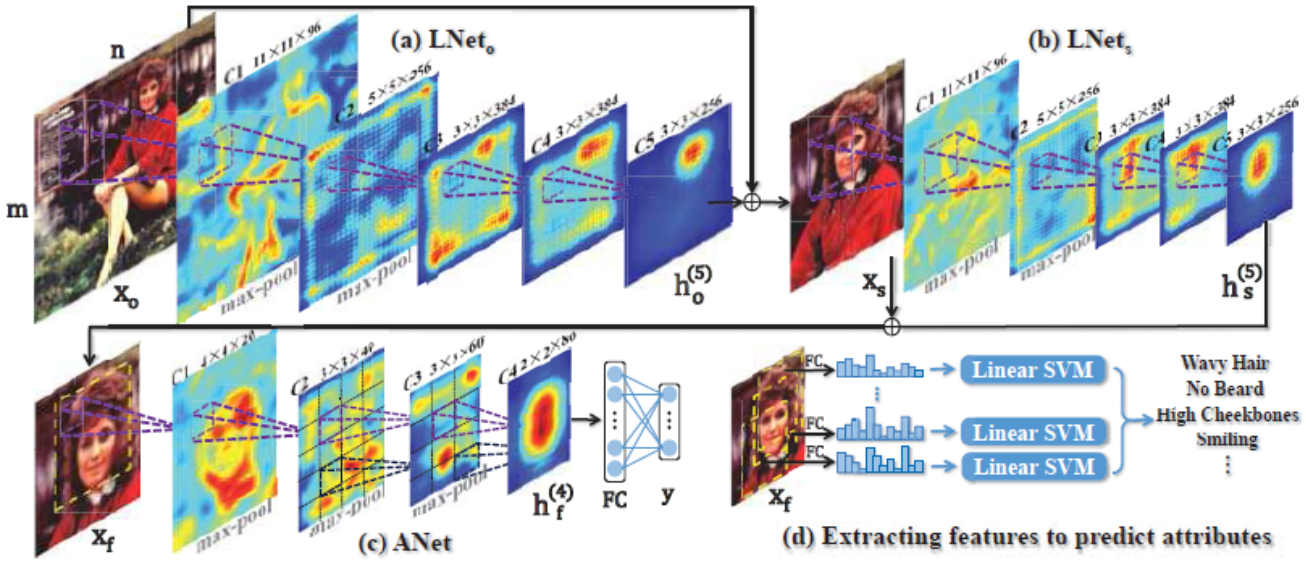

Figure 1. The proposed pipeline of attribute prediction (**Best viewed in color**)

2.1. Face Localization

The cascade of $LNet_o$ and $LNet_s$ accurately localizes face regions by being trained on image-level attribute tags. Both $LNet_o$ and $LNet_s$ are pretrained with 1, 000 general object categories containing 1.2 million training images and 50 thousands validation images. Dr. Liu adopts softmax for object classification, which is optimized by stochastic gradient descent (SGD) with back-propagation (BP). As shown in Fig. 2 (a.2), the averaged response map in C5 of LNeto already indicates locations of objects including human faces after pre-training.

The cross-entropy loss is used for attribute classification, *i.e.* $L = \sum_{i=1} y_i \log p(y_i|x) + (1 - y_i) \log(1 - p(y_i|x))$, where $p(y_i = 1|x) = \frac{1}{1+\exp(-f(x))}$ is the probability of the i-th attribute given image x. As shown in Fig. 2 (a.3), the response maps after fine-tuning become much more clean and smooth, indicating that the filters learned by attribute tags can detect face patterns with complex variations. To appreciate the effectiveness of pretraining, we also include the averaged response map in C5 of being directly trained from scratch with attribute tags but without pre-training in Fig. 2 (a.4). It cannot separate face regions from background and other body parts well.

2.2. Attribute Prediction

ANet is learned to extract features and SVM classifiers are used to predict attributes as shown in Fig. 1 (c) and (d). Specifically, in the pre-training stage, ANet is trained by classifying massive face identities. In the finetuning stage, Dr. Liu first extends the localized face region, which is

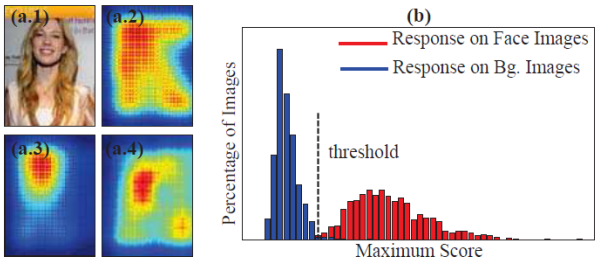

Figure 2. (a.1) Original image. (a.2)-(a.4) are averaged response maps in C5 of LNeto after pre-training (a.2), fine-tuning (a.3) and directly training from scratch with attribute tags but without pre-training (a.4). (b)Determine threshold.

properly resized, with a small factor to incorporate more context information. Then, multiple patches are cropped from the enlarged face region and utilized as inputs of ANet. ANet is fine-tuned by attributes to learn the highlevel feature FC. Furthermore, as shown in Fig. 1 (d), each feature vector is adopted to train SVM classifier for attribute prediction. The above strategy is similar to the multiview data augmentation, increasing the robustness of attribute recognition. In the testing stage, attributes are predicted by averaging the SVM scores over all the patches.

To this end, we propose an *interweaved operation*, which is a fast feed-forward method for CNN with locally-shared filters. They suppose have four local filters in the next locally convolutional layer. Each local filter is then apply on its corresponding interweaved map. Since the interweaved map capturing the entire image, each local filter is turned

into a global filter such that its computation can be shared across different patches. It enables extracting multiple feature vectors with only onepass of feed-forward evaluation. This operation can be repeated when more locally convolutional layers are added. The proposed feature extraction scheme has achieved 6 times speedup empirically when compared with patch-by-patch scanning. It is applicable to CNNs with local filters and compatible to all existing CNN operations.

3. Experimental Results of Framework

The team demonstrates the effectiveness of the framework in experiments. They report performance on several extended attributes and compare their result with FaceTracer and POOF as shown in Table 1.

Table 1. Performance comparison on extended attributes.

	Gender	Youth	M. Aged	Senior
FaceTracer [4]	91	66	54	70
POOF [1]	92	71	60	80
LNets+ANet	94	80	77	81

References

[1] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013. 3

[2] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[3] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. Liblinear: A library for large linear classification. *JMLR*, 2008. 1

[4] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008. 3

[5] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1

[6] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014. 1