Collaborative Learning Network for Face Attribute Prediction

Shiyao Wang, Zhidong Deng, Zhenyang Wang

State Key Laboratory of Intelligent Technology and Systems Tsinghua National Laboratory for Information Science and Technology Department of Computer Science, Tsinghua University, Beijing 100084, China

Abstract. This paper proposes a facial attributes learning algorithm with deep convolutional neural networks (CNN). Instead of jointly predicting all the facial attributes (40 attributes in our case) with a shared CNN feature extraction hierarchy, we cluster the facial attributes into groups and the CNN only shares features within each group in later feature extraction stages to jointly predicts the attributes in each group respectively. This paper also proposes a simple yet effective attribute clustering algorithm, based on the observation that some attributes are more collaborated (their prediction accuracy improve more when jointly learned) than others, and the proposed deep network is referred to as the collaborative learning network. Contrary to the previous state-of-the-art facial attribute recognition methods which require pre-training on external datasets, the proposed collaborative learning network is trained for attribute recognition from scratch without external data while achieving the best attribute recognition accuracy on the challenging CelebA dataset and the second best on the LFW dataset.

1 Introduction

Face attributes are deemed to be middle-level concepts which human use to describe a certain person. Empirically, when we describe a person, we may choose some characteristics on his/her attributes, such as gender, age, color, hair style(also see Fig.1). Apart from many entertainment applications inspired by attribute prediction, it contributes to improve the performance of relative tasks as well, including face verification [1], [2], [3], and [4], face retrieval [5], [6] and [3], suspect search according to eyewitness descriptions [7], and some relevant work about object recognition or classification [8], [9], and [10]. All above previous work has proved that attributes act as powerful representations of images, extracting discriminative features benefit the following process and boost the final performance.

In order to accurately and robustly predict attributes from images, a general process is divided into four standard steps: 1) detect faces in a picture [11], [12], and [13]; 2) face alignment [12], [14], and [15]; 3) feature extraction; 4) classification about the presence/absence of each attribute. In this paper, we will focus on the latter two steps that given a face after alignment, and then predict the

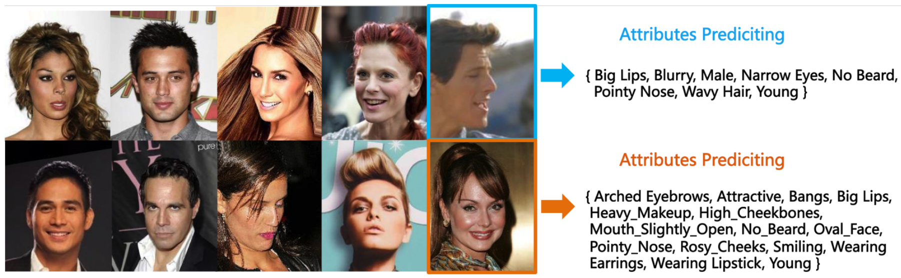


Fig. 1. Illustration of the facial attribute prediction on CelebA.

presence or absence of each required attributes which is similar to describ aspects of visual appearance. Some papers suggest extracting hand-crafted features from the entire images or several specified local parts of the subject. Although well-designed features could sometime achieve appreciable results, it may lack of robustness in images with large variations or deformations of objects. On the other hand, the deep learning framework especially deep convolutional neural network (CNN) [16], [17], and [18] is widely used to extract the features recently due to its superb ability to learn effective feature representations. It is a promising idea to use this deep network, nevertheless, given millions of labeled persons, the existing methods still need extra labels such as identity labels or other labeled data to pre-train or fine-tuning their nets which seems quite unefficient. Therefore, we instead present a new framework to address this issue by excavating valuable information behind the given data and providing priori knowledge benefits for designing a most suitable network.

In this paper, we propose a collaborative learning network to simultaneously predict 40 face attributes via a single model. Intuitively, modeling each attribute independently would be the best way to gain the superior accuracy and maximize the extraction of knowledge in the data whereas we consider it a heavy computational engineering more than a suitable model. Hence, we put forward a CNN structure equipped with well-designed both width and depth to jointly predict all the facial attributes with a shared feature extraction hierarchy and independent classification layers. Noticeably, our end-to-end model outperforms existing methods, gaining significant improvement in terms of the predicting accuracy evaluated on the CelebA [19]. On this basis, we further optimize our model inspired by [20], which described the hierarchical organization of face processing in the ventral stream. Similar to the visual cortical area organization which transforms the same low-level input into different representations in different visual cortical areas, we cluster the facial attributes into groups and the CNN only shares features within each group in later feature extraction stages to jointly predicts the attributes in each group respectively like the different visual cortical areas. This novel structure helps to further improve the predicting performance that significantly advances state-of-the-art attribute recognition on the CelebA. In summary, the contributions of our paper include:

- A well-designed convolutional neural network for predicting all the attributes both end to end and simultaneously.

- A simple yet useful algorithm to explore a pair-wise relationship between attributes and cluster them into different groups so that a new CNN structure can be built refer to above clusters.

- A novel framework that attributes share low-level feature extraction in the earlier stages of CNN, while focus on the relative tasks within a group in later stages that we called collaborative learning framework;

- Experiments which demonstrate that the collaborative learning framework significantly advances state-of-the-art attribute predicting on the CelebA dataset and the second best on the LFW dataset.

2 Relative work

Attribute recognition methods are generally categorized into two groups: hand-crafted features and deep learning methods. Extracting hand-crafted features at pre-defined land-marks is mentioned in [1], [8], [4] and [21]. [1] extracted hand-crafted features from different regions as "low-level" features, and then compute visual traits for attribute classification and face verification. [8] adopted color, texture, visual words, and edges as a base feature, shape, part and material as semantic attributes , embedding auxiliary discriminative attributes for describing objects. [21] and [4] improve the discriminativeness of hand-crafted features via specializing a particular domain and set of parts or stronger classifier such as three-level SVM system. Deep learning method including [19], [22], [23], [24], [25], [26] and [27] has achieved great success in attribute predicting and any other vision tasks. Among these significant previous work, [19] and [22] are the most relevant work with ours. The method in [19] cascades two CNNs, LNet and ANet, where LNet locates the entire face region and ANet extracts high-level face representation from the located region. The FC layer of Anet, namely, high-level representation is then adopted to train forty SVM classifiers for each attribute prediction. During the training process, LNet is pre-trained with one thousand object categories of ImageNet [28], while ANet is pre-trained by distinguishing massive face identities. In other word, both CNNs need large extra datasets in order to achieve good initialization. Besides, the overall method can be called Features+Classifier that motivated by the AdaBoost algorithm taken by Kumar et al. [29] which seems like optimizing each attributes model independently rather than end-to-end training simultaneously. Instead of relying on manually annotated images, Wang et al. [22] proposed a feature learning method relies on processing extra identity-unlabeled data and embeddings from a few supervised tasks. Actually, they still need extra datasets.

To sum up, all above significant prior work provided a variety of useful methods, especially [19] and [22] which equipped with the competitive results and we will present their results as our strong baseline. The evaluation of above methods will be conducted on the benchmarks CelebA [30] and LFW datasets [31]. The attributes labels of these two datasets is provided by [19].

3 Our Approach

Framework Overview Fig.2 illustrates our pipeline where a CNN model predicts all the facial attributes end-to-end as show in (a), while attributes clustering and the improved model as show in (b) and (c).

In the first step, we design a convolutional neural network learns multiple attribute labels simultaneously and predicts every attributes end to end.

In the second step, we exploit some latent information among the attributes. If the dataset consists of n people equipped with m attributes, it can be regard as n samples to explore the relevance to each pair of m attributes. If two attributes always appear to be the same output, namely presence or absence, we can consider them as positive correlation whereas one attribute is presence/absence, but the other one always appears to be the opposite state, we call them negative correlation. Mining from these large samples, we can draw a similarity matrix of these m attributes. And then, all the attributes can be clustered into k groups base on above matrix which is of great importance for the following step.

In the third step, differ from the general CNN structure, we present a novel *collaborative network*: on the bottom five stages, the net shares weights globally that works on extracting the rough features from the entire image; in the later feature extraction stage, the net is split to k branches based on the second step with each branch sharing weights within their own group that focuses on the the correlated tasks. At last, the net break into m mini-branches corresponding to m attributes classifiers. Benefit from this architecture, the bottom stage can generate better features because of more supervised labels and the top stage would promote each other when they have correlated tasks, that we called collaborative network.

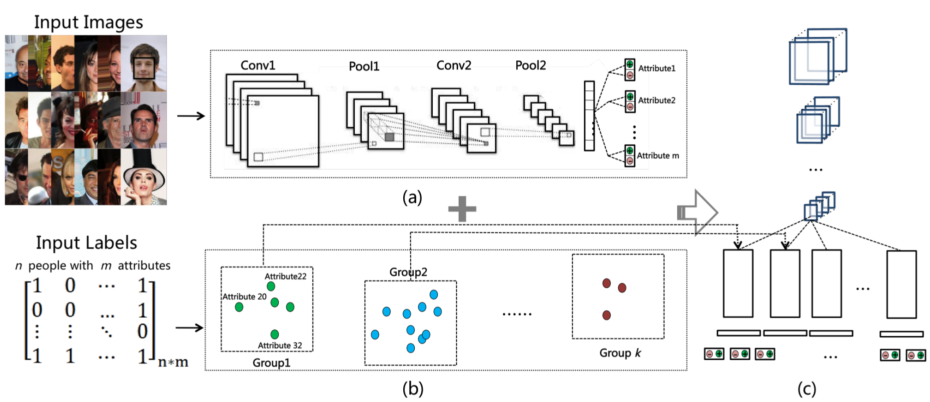


Fig. 2. The proposed pipeline of attribute prediction.

3.1 A well-designed convolutional neural network

To predict the m attributes of a given image end-to-end, we adopt CNN as our feature extractor and classifier while the network in [19] is learned to extract features and SVM classifiers are used to predict attributes. We design a CNN structure as follows: the whole net can be divided into six stages with five max-pooling layers and each stage consists of four general convolution layers, batch normalization layers, non-linear(ReLU) layers whose depth is designed such that neurons in the highest layer have the effective receptive fields of approximately 1.5 ? 2 times of the input sizes. And the number of filters in convolution layer preceding each pooling layer is always doubly expanded which used to be an empirical trick. After the last stage, all the feature maps are flattened into a vector, and connected to a FC layer whose size is 256 that is suitable for acting as a high-level feature. The classification layer is made up of m mini-classifiers corresponding to m attributes. Although the ANet in [19] was trained by classifying massive face identities in the pre-train stage, and there are some tricks during training in order to learn discriminative features with a large number of identities, our network is only fed by given dataset from scratch without pre-training. Our model achieves an impressive performance in CelebA and LFW datasets with the training process straight and concise. The detail of our network can be seen in Fig.3.

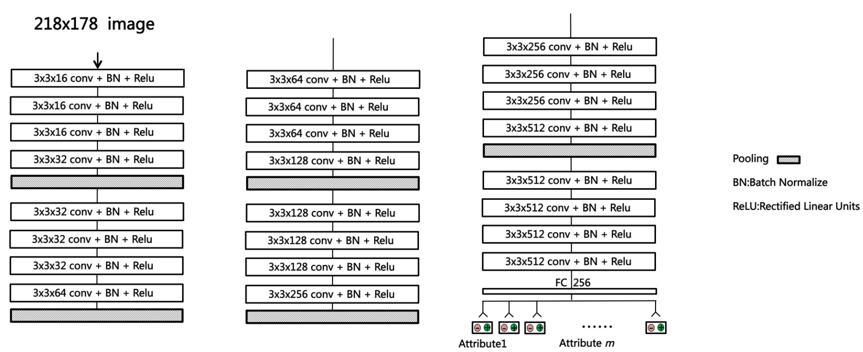


Fig. 3. The structure of our well-designed convolutional neural network.

3.2 Modeling the relationship among attributes

Inspired by [20], we expect to build a model works as the visual cortical areas that receiving the low-level input, selective attention is paid to a particular attribute. Consequently, a pair-wise relationship among attributes should be established based on the relationship. With the face dataset consisted of n images, each labeled by m attributes, we will provide a simple yet useful algorithm to model the relationship among attributes. In CelebA, n denotes 162770 training samples while m indicating 40 attributes. As each attribute is represented as a binary code, 1 denoting presence while 0 denoting absence of a certain attribute, we can generate a $n * m$ matrix to present these data:

$$\begin{array}{cc} & \begin{matrix} A_1 & A_2 & \cdots & A_j & \cdots & m \end{matrix} \\ \begin{matrix} P_1 \\ P_2 \\ \vdots \\ P_i \\ \vdots \\ P_n \end{matrix} & \begin{pmatrix} I(P_1,A_1) & I(P_1,A_2) & \cdots & I(P_1,A_j) & \cdots & I(P_1,A_m) \\ I(P_2,A_1) & I(P_2,A_2) & \cdots & I(P_2,A_j) & \cdots & I(P_2,A_m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ I(P_i,A_1) & I(P_i,A_2) & \cdots & I(P_i,A_j) & \cdots & I(P_i,A_m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ I(P_n,A_1) & I(P_n,A_2) & \cdots & I(P_n,A_j) & \cdots & I(P_n,A_m) \end{pmatrix} \end{array} \tag{1}$$

Where P_i presents the i-th person and A_j donotes the j-th attribute. For a given sample P_i, let $I(P_i, A_j) : P_i, A_j \rightarrow \{0,1\}$ be a function yielding the binary ground truth label for P_i and A_j, where $i \in \{1, \cdots, n\}$ is the people index and $j \in \{1, \cdots, m\}$is the attribute index.

After this, all of the samples have been organized as above matrix(Eg. (1)) and each column vector belongs to a specified attribute. In order to explore a pair-wise relationship among attributes, we can calculate the co-occurrence of each two attributes. If they always appear to be the same output, we can consider them as positive correlation. Hence ,we define the following equation as their similarity:

$$s(A_j, A_{j'}) = p(A_j = 1, A_{j'} = 1) + p(A_j = 0, A_{j'} = 0) \tag{2}$$

$$p(A_j = 1, A_{j'} = 1) = \frac{N(I(P_i, A_j) = I(P_i, A_{j'}) = 1)}{n} \tag{3}$$

$$p(A_j = 0, A_{j'} = 0) = \frac{N(I(P_i, A_j) = I(P_i, A_{j'}) = 0)}{n} \tag{4}$$

A_j and $A_{j'}$ are two attributes, and $N(I(P_i, A_j) = I(P_i, A_{j'}) = 1/0)$ corresponds to the num of both A_j and $A_{j'}$ are equal to 1/0.

After computing the similarity between two attributes nodes through above measures, we may model the relationship as a similar matrix:

$$\begin{array}{cc} & \begin{matrix} A_1 & A_2 & \cdots & A_m \end{matrix} \\ \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} & \begin{pmatrix} s(A_1,A_1) & s(A_1,A_2) & \cdots & s(A_1,A_m) \\ s(A_2,A_1) & s(A_2,A_2) & \cdots & s(A_2,A_m) \\ \vdots & \vdots & \vdots & \vdots \\ s(A_m,A_1) & s(A_n,A_2) & \cdots & s(A_m,A_m) \end{pmatrix} \end{array} \tag{5}$$

We cluster these m attributes according to the similar matrix in a way motivated by K-means.

step 1: We choose k attributes as initial cluster centers;

step 2: Each attribute is assigned to its closest cluster center base on the similar matrix;

step 3: Each cluster center is updated according to the above new groups;

step 4: Back to step2 until no further change in cluster centers.

In the later experiments analysis, we choose $k = 6$. In fact, attribute prediction accuracy is improved consistently for different numbers of clusters (K), i.e., K=4,5,6,7,8. Besides, we find that the attribute clusters have carried semantic concepts.

3.3 Collaborative learning network

After exploring the relationship between attributes, a new CNN architecture is presented in this section. We still believe that end-to-end is a better way exploiting all correlation instead of separating of feature extraction and attribute classification. Therefore, a collaborative learning network is presented in Figure.4.

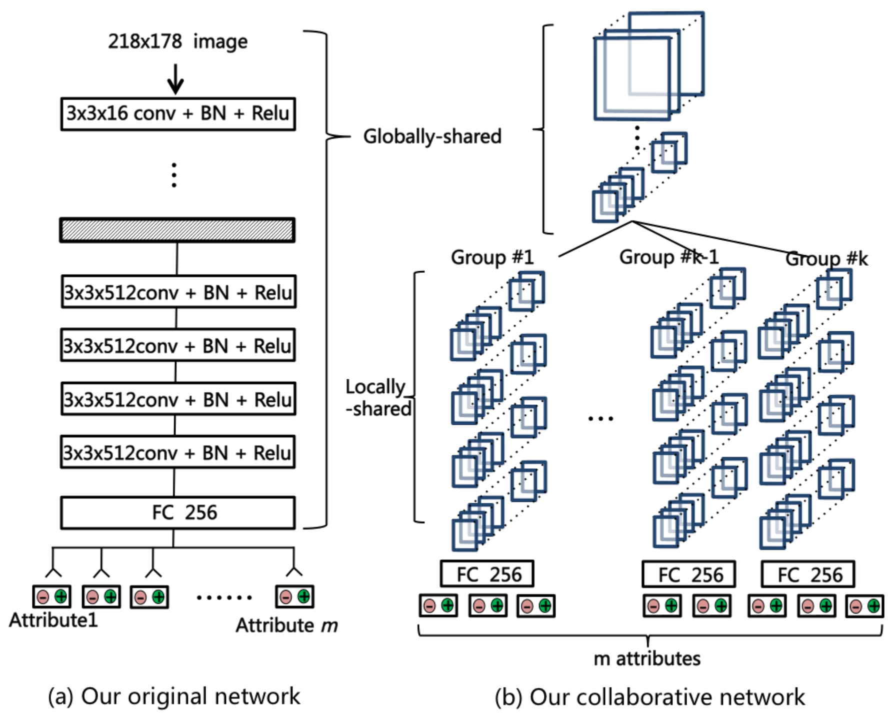


Fig. 4. Collaborative learning network.

During the first five stages, all the filters are globally-shared like our original net which aim at extracting the rough features over the entire image. And then, our net is divided into k sub-nets based on the clustering results provided by section 3.1. In other word, in the last stage, our net generates k branches corresponding to the k clusters that including three standard units (a convolution layer, a batch normalization layer and a non-linear layer) and the filters are shared within their own branch respectively. In this way, each branch focuses on the specified and discriminative features extracting for relative attributes recognition and collaboratively learning by sharing the filters. Benefit from above stages, the features are rich and specialized enough for classification, so that we can apply the m classifiers after that process. Our collaborative network is still fed by the given attribute dataset and initialized with random weights. Contrary to our original net, the final accuracy is improved a lot compared with the strong baseline in [19] and [22] which well demonstrated our assumption.

4 EXPERIMENT

In this section we evaluate the effectiveness of collaborative learning network with quantitative results on two standard facial attribute datasets called CelebA and LFWA datasets. CelebA contains ten thousand identities, each of which has twenty images. For this dataset, there are eight thousand identities with 162770 images for training, another one thousand identities with 19867 images for validation and the remaining 19962 images for testing. LFWA has 13, 233 images of 5, 749 identities. Each image in CelebA and LFWA is annotated with forty face attributes. Our network is implemented based on the framework of caffe [32], and conducted on two GPUs with data parallelism. In all experiments, our models are trained using stochastic gradient descent (SGD) algorithm with a mini-batch of 64. The learning rate is initialized to 0.001 and repeatedly decreased 3 times, until it arrives at 1e-6. We run our net through the whole training data with only 8 epochs and a momentum of 0.9 is used in the entire training process to make SGD stable and fast. In the following sections, we conduct three kinds of experiments. First, we present a comparison between an end-to-end convolutional neural network and an approach separates the feature extraction and classification in [19]; Second, we show the pairwise relationship between attributes in section 4.2 and a new CNN structure based on the above relationship is also presented; Last but not the least, some existing approaches including the sate-of-the-art are reported in section 4.3 that proves our approach has advanced the other methods.

4.1 An end-to-end convolutional neural network for attribute prediction

In this section, we compare the method in [19] and our well-designed convolutional neural network in order to demonstrate a way of end-to-end is able to achieve the higher accuracy. The Although Liu et al. [19] suggested to extract features via network and feed to independent linear SVMs for final attribute classification, we believe that our network has equipped with the ability of classification over 40 attributes labels as supervisory signals. Besides, contract to some suggestion that independently optimize each attributes by different networks, we

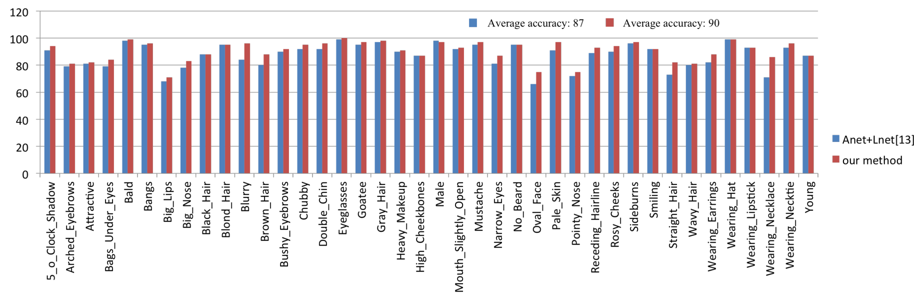


Fig. 5. Attribute prediction results of our method and [19].

tend to gather all the attributes together account for the collaborative learning. Actually, the comparison in figure.5 shows the average accuracy of LNets+ANet [19] is 87% while our net can achieve 90%. Almost every attributes have been improved, and the maximum improvement can obtain 15% (Wearing_Necklace). Consequently, we suggest that predicting face attributes though an end-to-end convolutional neural network would be better and all the attributes can be set as supervised signals in the meanwhile. When [19] prefers pre-taining by massive face identities for attribute prediction, we tend to train from scratch. Recognizing identities and attributes are two intuitively contradictive tasks. The former suppresses face variations unrelated to identities, such as pose, expression, age, wearing hat etc. while the latter reserves such variations. For comparison, we pre-train our model for face recognition on the Celeb face dataset and fine-tune it for attribute recognition. It gets 89.35% accuracy, inferior to 90.41% of the model without pre-training.

4.2 A collaborative learning network based on the pair wise relationship of attributes

We compute the correlation between two attributes as described in Sec.2.2. Note that the relationship between any two sets of attributes can be computed in the same way. The correlation value is in the range of [0, 1], meaning positive correlation, respectively. And then we cluster all the attributes into six groups which presented as:

Cluster #1: High Cheekbones, Mouth Slightly Open, Smiling;

Cluster #2: No Beard, Young;

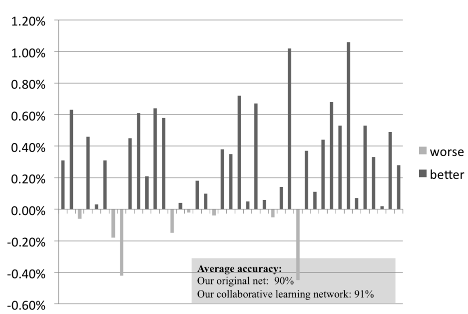
Cluster #3: Arched Eyebrows, Attractive, Heavy Makeup, Pointy Nose, Wavy Hair, Wearing Lipstick;

Cluster #4: Bags Under Eyes, Big Nose, Male;

Cluster #5: Black Hair, Oval Face, Straight Hair;

Cluster #6: 5 o Clock Shadow, Bald, Bangs, Big Lips, Blond Hair, Blurry, Brown Hair, Bushy Eyebrows, Chubby, Double Chin, Eyeglasses, Goatee, Gray Hair, Mustache, Narrow Eyes, Pale Skin, Receding Hairline, Rosy Cheeks, Sideburns, Wearing Earrings, Wearing Hat, Wearing Necklace, Wearing Necktie;

It seems that the clusters have carried some semantics concepts: Mouth Slightly Open has highly positive correlation with Smiling and they are in cluster #1; No Beard is clustered with Young; Attractive, Heavy Makeup and Wearing are in cluster #3. We adopt these group information into our collaborative network by dividing our net into six branches on the last stage of original structure. The average accuracy is further improved from 90% to 91%. All groups get consistent improvement. 0.5%, 0.3%, 0.4%, 0.1%, 0.4%, and 0.2% average improvement is acquired for each of the six groups, respectively. We make a comparison between our original net and collaborative net for each attributes in Fig.6.(a). Why collaborative network helps to further improve the predicting accuracy? We consider that although there are 160 thousand images for training, the labels are not balanced such as Gray Hair which we only have 4% samples of positive labels.When add several relative attributes, the model will be capable of generating better features. In addition, we conduct anther experiment to prove the effectiveness of our framework. We designed three models taking different input sizes and network depth, all of which are based on the design principles in section 3.1. As show in Fig.6.(b), all these three models achieve higher accuracy compared to our original network. Furthermore, we even applied our proposed method to GoogLeNet by branching its later feature extraction stages and using the clustered attributes for supervision. The accuracy is further improved from 90.9% to 91.24%. Therefore our method generally improves the attribute prediction accuracy irrespective of the network architectures that effectively demonstrate our collaborative learning network.



(a)

	Model Size (original)	Average Accuracy	
		Original Net	Collaborative Net
Model #1	72MB	90.4%	**90.7%**
Model #2	42MB	90.2%	**90.5%**
Model #3	15MB	90.2%	**90.3%**
GoogLeNet	58MB	90.9%	**91.24%**

(b)

Fig. 6. (a) Comparison between our original net and collaborative net for each attributes. (b) Performances of three different models.

4.3 Performance Comparison with Relative Work

In this section, our method is compared with five competitive approaches, i.e. FaceTracer [29], PANDA-w [26], and PANDA-l [26], Lnets+Anet [19] and Walk and Learn [22]. External data are needed during their training process while our net is trained from scratch without extra data. The average accuracy of all above methods in CelebA is 81%, 79%, 85%, 87%, 88% and 91%, respectively(also see Fig.7). We also test these methods on LFWA dataset, and we get the second best. We check labels of LFWA to find some of images are mis-labeled. For example, some images labeled “female’ is shown in Fig.8. In order to further evaluate our model, we cleanse bad labels about “Male” and test our net on that cleaned labels again. The accuracy of the sex predicting can achieve 98% while it is 94% in the previous mis-labeled data that meet our expectations. In particular, our model is even not fine-tuned in LFWA data, and directly tested using the model trained over CelebA.

		5_o_Clock_Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	High Cheekbones	Male
CelebA	FaceTracer[29]	85	76	78	76	89	88	64	74	70	80	81	60	80	86	88	98	93	90	85	84	91
	PANDA-w [26]	82	73	77	71	92	89	61	70	74	81	77	69	76	82	85	94	86	88	84	80	93
	PANDA-l [26]	88	78	81	79	96	92	67	75	85	93	86	77	86	86	88	98	93	94	90	86	97
	Lnets+Anet [19]	91	79	81	79	98	95	68	78	88	95	84	80	90	91	92	99	95	97	90	87	**98**
	Work anf learn [22]	84	**87**	**84**	**87**	92	**96**	**78**	**91**	84	92	**97**	81	**93**	89	93	97	92	95	**96**	**95**	96
	Our method	**94**	82	82	85	**99**	**96**	71	83	**89**	**96**	96	**88**	**93**	**95**	**96**	**100**	**97**	**98**	91	87	**98**
LFWA	FaceTracer[29]	70	67	71	65	77	72	68	73	76	88	73	62	67	67	70	90	69	78	88	77	84
	PANDA-w [26]	64	63	70	63	82	79	64	71	78	87	70	65	63	65	64	84	65	77	86	75	86
	PANDA-l [26]	84	79	81	80	84	84	73	79	87	94	74	74	79	69	75	89	75	81	93	86	92
	Lnets+Anet [19]	84	82	83	83	88	88	75	81	90	97	74	77	82	73	78	95	78	84	95	88	94
	Work anf learn [22]	76	82	82	91	82	93	75	92	93	97	86	83	78	79	81	94	80	91	96	96	93
	Our method	74	80	79	83	91	88	79	84	92	96	85	81	81	78	79	91	81	86	96	90	94

		Mouth_Slightly_Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young		Average
CelebA	FaceTracer[29]	87	91	82	90	64	83	68	76	84	94	89	63	73	73	89	89	68	86	80		81
	PANDA-w [26]	82	83	79	87	62	84	65	82	81	90	89	68	76	72	91	88	67	88	77		79
	PANDA-l [26]	93	93	84	93	65	91	71	85	87	93	92	69	77	78	96	93	67	91	84		85
	Lnets+Anet[19]	92	95	81	95	66	91	72	89	90	96	92	73	80	82	99	93	71	93	87		87
	Work anf learn [22]	**94**	83	79	75	**84**	87	**91**	86	81	77	**97**	76	**89**	**96**	86	**97**	**95**	80	**89**		88
	Our method	**94**	**97**	**87**	**96**	75	**97**	76	**93**	**95**	**98**	92	**82**	81	89	**99**	94	86	**96**	87		**91**
LFWA	FaceTracer[29]	77	83	73	69	66	70	74	63	70	71	78	67	62	88	75	87	81	71	80		74
	PANDA-w [26]	74	77	68	63	64	64	68	61	64	68	77	68	63	85	78	83	79	70	76		71
	PANDA-l [26]	78	87	73	75	72	84	76	84	73	76	89	73	75	92	82	93	86	79	82		81
	Lnets+Anet [19]	82	92	81	79	74	84	80	858	78	77	91	76	76	94	88	95	88	79	86		84
	Work anf learn [22]	98	90	79	90	79	85	77	84	96	92	98	75	85	91	96	92	77	84	86		87
	Our method	78	92	79	81	71	89	83	88	74	81	91	82	80	96	89	96	90	84	85		85

Fig. 7. Performance comparison with state of the art methods on 40 binary facial attributes. (Note that all the methods expect ours need external datasets for training.)

5 CONCLUSION

This paper have proposed a facial attribute learning algorithm with deep convolutional neural networks (CNN). First, we provide a well-designed convolutional neural network to predict attributes end to end with noticeable improvement. Second, to take a deep look into all the given attributes, we present a useful

Fig. 8. Some mis-labeled images in LFWA. (labed "female").

algorithm to learn a pairwise relationship between them; Third, based on the above relationship, unlike the previous facial attribute recognition studies which treated all the attributes equally, this paper discovered the grouping properties of the facial attributes and designed the deep network which explored the correlation of the attribute labels. The proposed method consistently improves the attribute prediction accuracy of very competitive baseline networks such as GoogLeNet and our designed deep networks. We have demonstrated the effectiveness of this framework on two challenging face datasets, CelebFaces and LFW datasets. Instead of pre-training on large-scale face recognition datasets like the previous state-of-the-art facial attribute recognition methods, the proposed collaborative learning network is trained from scratch without external data while achieving the best attribute recognition accuracy on the challenging CelebA face dataset and the second best on the LFW dataset.

Acknowledgement. The authors would like to thank the anonymous reviewers for their valuable comments that considerably contributed to improving this paper. This work was supported in part by the National Science Foundation of China (NSFC) under Grant Nos. 91420106, 90820305, and 60775040, and by the National High-Tech R&D Program of China under Grant No. 2012AA041402.

References

1. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 365–372
2. Song, F., Tan, X., Chen, S.: Exploiting relationship between attributes for improved face verification. Computer Vision and Image Understanding **122** (2014) 143–154
3. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011) 1962–1977
4. Berg, T., Belhumeur, P.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 955–962

5. Siddiquie, B., Feris, R.S., Davis, L.S.: Image ranking and retrieval based on multi-attribute queries. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 801–808
6. Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for efficient image retrieval. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 745–752
7. Feris, R., Bobbitt, R., Brown, L., Pankanti, S.: Attribute-based people search: Lessons learnt from a practical surveillance system. In: Proceedings of International Conference on Multimedia Retrieval, ACM (2014) 153
8. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1778–1785
9. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: Computer Vision–ECCV 2010. Springer (2010) 155–168
10. Yu, F., Cao, L., Feris, R., Smith, J., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 771–778
11. Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
12. Xiong, X., Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2013) 532–539
13. Ding, C., Xu, C., Tao, D.: Multi-task pose-invariant face recognition. Image Processing, IEEE Transactions on **24** (2015) 980–993
14. Zhu, S., Li, C., Change Loy, C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 4998–5006
15. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 3659–3667
16. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 1489–1496
17. Zhu, Z., Luo, P., Wang, X., Tang, X.: Deep learning identity-preserving face space. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 113–120
18. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1701–1708
19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3730–3738
20. Ungerleider, L.G., Haxby, J.V.: 'what' and 'where' in the human brain. Current opinion in neurobiology **4** (1994) 157–165
21. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 1543–1550
22. Wang, J., Chen, Y., Wang, Tang, X.: Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. (2016)

23. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2480–2487
24. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: a deep model for learning face identity and view representations. In: Advances in Neural Information Processing Systems. (2014) 217–225
25. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: Computer Vision–ECCV 2014. Springer (2014) 834–849
26. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1637–1644
27. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning social relation traits from face images. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3631–3639
28. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 248–255
29. Kumar, N., Belhumeur, P., Nayar, S.: Facetracer: A search engine for large collections of images with faces. In: Computer Vision–ECCV 2008. Springer (2008) 340–353
30. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems. (2014) 1988–1996
31. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst (2007)
32. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, ACM (2014) 675–678