Multi-task Deep Neural Network for Joint Face Recognition and Facial Attribute Prediction

Zhanxiong Wang[1†], Keke He[1†],Yanwei Fu[2*]
Rui Feng[1], Yu-Gang Jiang[1], and Xiangyang Xue[12]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, China
[2]School of Data Science, Fudan University, China
{15210240046,kkhe15,yanweifu,fengrui,ygj,xyxue}@fudan.edu.cn

ABSTRACT

Deep neural networks have significantly improved the performance of face recognition and facial attribute prediction, which however are still very challenging on the million scale dataset, i.e. MegaFace. In this paper, we for the first time, advocate a multi-task deep neural network for jointly learning face recognition and facial attribute prediction tasks. Extensive experimental evaluation clearly demonstrates the effectiveness of our architecture. Remarkably, on the largest face recognition benchmark – MegaFace dataset, our networks can achieve the Rank-1 identication accuracy of 77.74% and face verication accuracy 79.24% TAR at 10^{-6} FAR, which are the best performance on the small protocol among all the publicly released methods.

KEYWORDS

face identification, face verification, deep learning, multi-task learning and facial attribute prediction

ACM Reference format:
Zhanxiong Wang[1†], Keke He[1†],Yanwei Fu[2] and Rui Feng[1], Yu-Gang Jiang[1], and Xiangyang Xue[12] . 2017. Multi-task Deep Neural Network for Joint Face Recognition and Facial Attribute Prediction. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania,* , 9 pages.
DOI: http://dx.doi.org/10.1145/3078971.3078973

1 INTRODUCTION

Face recognition as a typical biometric-based technique, has received a great deal of research attention over the last few decades. Instead of authenticating people by passwords, PINs, tokens and so on, face recognition enables the identification or verification by biological facial traits. This thus facilitates a wide range of applications such as financial monitoring, building access control, station ticket check system and so forth.

Recently, with the renaissance of convolution neural networks, encouraging breakthroughs have been achieved on the leading Labeled Faces in the Wild (LFW) benchmark [15]. Particularly, some methods [32, 36] have surpassed human performance and achieved near perfect results. To further study face recognition in a dataset of billions of faces, MegaFace dataset [18] has been collected to simulate a more realistic application scenario and evaluate the face recognition algorithms in a million scale of distractors. Critically, recognizing faces is still very challenging "in the wild", since the human face image often changes dramatically caused by the variances of head pose, camera viewpoints, occlusion and illumination conditions.

The standard practices looked at solving this problem by increasing either the depth of neural network, or the number of training images, or both. This nevertheless requires prohibitive cost of computational resources. In contrast, humans recognize one's identity mostly by identifying the facial traits or facial attributes [31]. For example, the police are interested in identifying a suspect, in most scenarios, by collecting the facial traits of the suspect from the eyewitnesses. Interestingly, previous work [24, 35] revealed that identity-related attributes such as gender are implicitly encoded in the nodes of deep neural network for identity discrimination. Thus orthogonal to the standard practices and inspired by humans' ability, we advocate utilizing facial attributes to help face recognition and vice versa.

In this paper, we propose a novel multi-task deep learning architecture to enable the jointly learning facial attribute prediction and face recognition tasks. Specifically, since the existing face recognition dataset do not have pre-labelled facial attributes, the attribute labels are generated by majority-voting manner of prediction of attribute classifiers pre-trained on the auxiliary dataset. With the labelled attributes, we propose our multi-task face recognition network of both Full-Model and Fast-Model. Our architecture learns for both face recognition and facial attribute prediction tasks. The experiments on the challenging MegaFace dataset clearly show that the state-of-the-art performance is obtained. To further reveal the insights of our network, further ablation study on both MegaFace and LFW dataset is conducted and validates the efficacy of our framework.

To sum up, the contributions of our work are three points. (1) To the best of our knowledge, we for the first time, propose a multi-task deep neural network structure of jointly learning the two highly corrected tasks – facial attribute prediction and face recognition. These two tasks can greatly help each other and thus

*Yanwei Fu is the corresponding authour.
†The first two authors contribute equally to this paper.

ICMR '17, June 6–9, 2017, Bucharest, Romania
 ACM ISBN 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3078973

improve the performance over the state-of-the-art algorithms. Our experimental results achieve the best performance on the MegaFace challenge with the comparable evaluation protocol.(2) Due to the highly cost of our Full-Model, we propose a Fast-Model for learning the multiple tasks. The Fast-Model is a reasonable good and fast version of the Full-Model in term of running cost and performance. (3) We also systematically explore the alternative choices of different structures and configures in our network in both our intuitive explanations and experimental validation. The validation reveals the insights of the details how we design the network.

2 RELATED WORK

2.1 Face Recognition Analysis

Face detection and face alignment. Viola and Jones [40, 41] made the first work to enable face detection applicable in real-world scenarios by building cascaded classifiers upon Harr-like features. On top of HOG features, the generic object detector – deformable part model (DPM) [9] can be used for face detection. The detected faces will be aligned by face alignment algorithms [4, 14, 30, 39, 46, 47, 49, 50, 54]. Recently, with the CNNs based face detection algorithms [6, 29, 55] and face alignment algorithms [4, 49, 50], extremely remarkably high and reliable performance on face detection and face alignment can be obtained on the LFW benchmark.

Face verification and face identification. They are two primary tasks of face recognition. Face verification is a one-to-one matching by given a pair of query images and judge whether this pair is the same person. Starting from [7], the idea of selecting a pair of face images to a distance begins. Given positive pairs, siamese networks [20, 27, 36] restrict the similarity metric to be small, and vice versa. Face identification is a one-to-many matching which identifies the face image from many existing facial images. It has been widely studied recently [3, 11, 32–35, 37]. Wen *et al.* [43] proposed a novel discriminative feature learning approach for deep face recognition by joint supervision of the center loss and soft-max loss.

Different from all previous works, we for the first consider jointly learning the facial attributes and face identity in a single framework with the shared deep architecture. We argue that facial attribute can help face recognition significantly and vice versa. For example, in criminal investigation when police identify the face images of a suspect from their own face dataset, they can identify the suspect by personal facial traits such as facial hair type, eye types, nose and so on. Our experiments validate that the remarkable high results have been achieved on the benchmark dataset.

2.2 Facial Attribute Analysis

Facial attribute analysis has also been studied recently and achieved a series of breakthroughs in recent years. In general, the works of facial attribute analysis can be categorized as three classes according to the different visual features and varying network structures.

Firstly, traditional manually designed features such as SIFT [25] and LBP [28] are employed to either holistically describe the entire face, or selected key facial regions such as check and mouth. The attribute classifiers [19, 21] are thus built upon extracted features.

Table 1: Selected attributes of the CelebA Dataset.

Group	Attributes
eye	Narrow Eyes
nose	Big Nose, Pointy Nose
global	Chubby, Double Chin, High Cheekbones, Male

Secondly, deep features based facial attribute analysis has been studied recently [24, 42, 52]. Zhong et al. [52] adopted the off-the-shelf CNN face recognition features for Support Vector Machine (SVM) classifiers to predict facial attributes.

Thirdly, facial attribute prediction is one sub-task of multi-task framework in [1, 31]. Rudd *et al.* [31] introduced a mixed objective optimization network which took account of different distribution of attribute labels. Facial attributes have been directly used to build the face identification classifier in [19]. Comparing with [1, 19, 31], we for the first time, equally and jointly model the facial attribute and face identity.

2.3 Multi-Task Learning

Multi-task Learning (MTL) is one type of transfer learning and addresses the problem of learning to share information and exploiting the similarity among different tasks on the same dataset. It has been applied to many computer vision tasks such as pose estimation and action recognition [53] and semantic classification and information retrieval [23]. Recently MTL has also been employed in face-related tasks, for example, facial landmark detection [29, 51] and [1, 8, 31]. Comparing with these works and to the best of our knowledge, this is the first published work of jointly learning facial attributes and face recognition in the same single deep architectures.

3 THE ATTRIBUTE-CONSTRAINED DEEP FACE RECOGNITION ARCHITECTURE

We advocate the face attribute-constrained deep architecture for face recognition. In contrast, the methods on face recognition [43, 45] did not directly use the information of face attributes. Actually, the face attributes are very useful cues for face related tasks.

3.1 Architecture Overview

Figure 1 illustrates our full model for attribute-constrained face recognition architecture. The whole architecture has two steps: facial attribute prediction step (Sec. 3.2) and face recognition network step (Sec. 3.3).

As most existing face recognition datasets do not have facial attribute labelled, the first step is to generate attribute labels by an off-the-shelf facial attribute prediction algorithm in Sec.3.2. In the second step, we take as input of our network, the generated attribute labels, together with cropped training images and the corresponding face identity labels in Sec. 3.3. We also introduce a fast version of our network in Sec. 3.4.

3.2 Facial Attribute Prediction

In this work, we propose employing the facial attributes as the constraints for face recognition. Such facial attributes thus should be invariant to the visual appearance of one person in difference situation. In other word, the attributes related to biometrics (e.g.

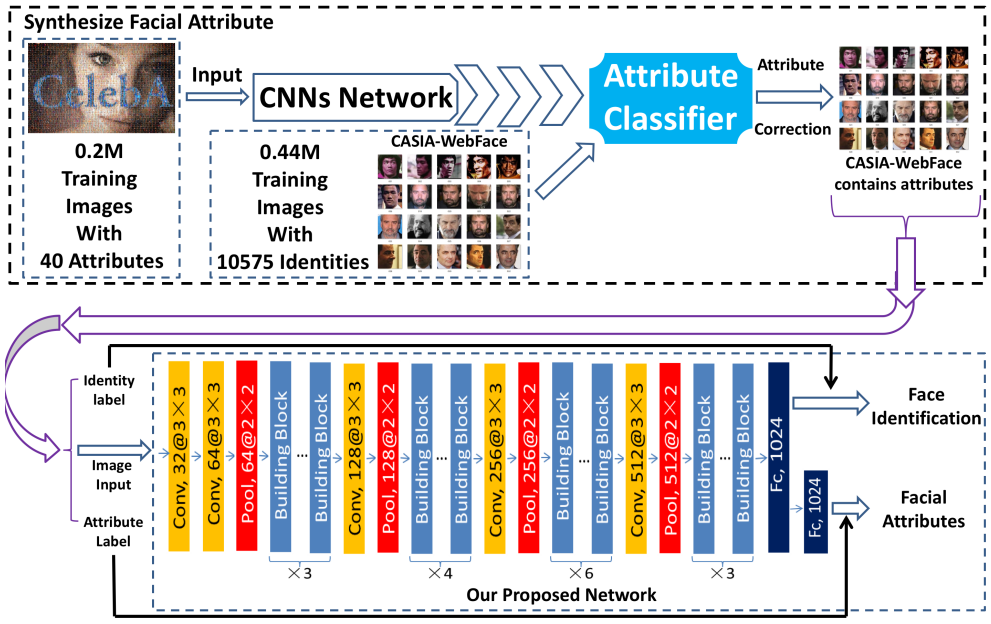

Figure 1: The overview of deep face recognition architecture. Note that: (1) the ×3,×4,×6 and ×3 indicates the corresponding number of building blocks. (2) number@size indicates the number and the size of convolutional filters.

Big Nose) are more desirable than the other appearing facial attributes (e.g. wearing glasses). The full list of attributes used in our framework are shown in Tab. 1.

We predict these facial attributes of each identity in a transferrable manner. Specifically, we use CelebA dataset [24] (totally 0.2 M facial images) to train the facial attribute classifier (MOON model [31]). The trained model can thus be utilized to predict the facial attributes on the CAISIA-WebFace [48] dataset, which is the training data of our proposed network. For each identity, we assume that the selected facial (in Tab. 1) should be consistent. Thus majority voting can be further employed to adjust the results of facial attribute prediction. For example, the dataset includes 15 face images of Bruce Lee: if 14 of images are predicted to male and only one is predicted to female, then the majority voting will constrain all images of Bruce Lee with male attribute. Such a majority voting manner intrinsically helps prune noise and outlying prediction of classifiers for biometrics-related attributes in Tab. 1.

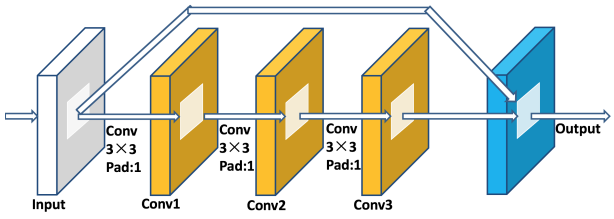

Figure 2: Building Block Structure(Full-Model). Each convolutional layer has padding 1 and stride 1.

3.3 Face Recognition Network Architecture

As illustrated in Fig. 1, this face recognition step intrinsically is organized as the network directly for face identification task, *i.e.*, a classification task for each identity. As for face verification task, we extract the features of each image by using the output of network as the representation; and given a pair of images, we compute their cosine distance and set the threshold to judge whether this pair is the same person or not. The network thus has four types of layers, namely,

Convolutional layers, aims at analyzing the input data stream with the size of 3×3 of receptive field, the stride 1 and padding 0. The four convolutional layers (yellow blocks in Fig. 1) have 32, 64, 64, and 64 filters respectively. From the first convolutional layer to the second convolutional layer, the number of filters is doubled to preserve enough feature information for the following layers.

Max-pooling layers, have the filters of size 2×2 and stride 2. All four layers maintain the same number filters of previous layers and halves the size of feature map to save the computational cost.

Building blocks layers, have four groups with the number of building blocks of 3, 4, 6 and 3 respectively. For each group, we use the same structure as shown in Fig. 2, i.e., it has 3 convolutional filters as main road and one bypass [12] which is element-wise summed as the output. The number of convolutional filters of each group is 64, 128, 256 and 512 individually.

Fully connected layers, extract the 1024−dimensional vector for face recognition tasks. To better model the correlations among the facial attributes, we use one additional fully connected layer with 1024−dimensional output for facial attribute prediction only. The output vectors will model the visual features of face images. The batch normalization [16] is used after each convolutional layer and

fully connected layer, then we use PReLU activation [13] after the batch normalization.

To better facilitate face recognition task, we employ facial attributes generated in Sec. 3.2 to help face recognition tasks. We argue that if our network wants to successfully recognize one identity from face image, it should also be able to trust-ably predict its corresponding biometric traits on faces, i.e. the facial attributes in Tab. 1; and vice versa. Thus, we build a deep network to jointly learn the facial attribute prediction and face identification tasks.

The network in Fig. 1 is denoted as $f(\mathbf{I};\Theta)$, where Θ is the parameter set of the deep architecture and we use $\mathbf{I}$ to denote the training images. Suppose we have M facial attributes and P face identities. We model the minimization of the expected loss as follows,

$$\{\Theta, W_a, W_p\} = \text{argmin}\ \mathcal{L}\left(\mathbf{I};\Theta, W_a, W_p\right) \tag{1}$$

where $\mathcal{L}\left(\mathbf{I};\Theta, W_a, W_p\right)$ is the loss function of the task and defined as

$$\mathcal{L}\left(\mathbf{I};\Theta, W_a, W_p\right) = \lambda_1 \mathcal{L}_a\left(W_a \cdot g\left(f\left(\mathbf{I};\Theta\right)\right)\right) + \lambda_2 \mathcal{L}_p\left(W_p \cdot f\left(\mathbf{I};\Theta\right)\right) \tag{2}$$

where $W_a \subseteq \mathbb{R}^{1024\times M}$ and $W_p \subseteq \mathbb{R}^{1024\times P}$ are the learned weights for facial attribute and face identification tasks; we use mean square error loss for facial attribute $\mathcal{L}_a(\cdot)$; and we use soft-max loss for face identification task $\mathcal{L}_p(\cdot)$. The shared parameters Θ are jointly optimized by these two tasks. $g(\cdot)$ indicates the second fully connected layer for facial attribute only in Fig. 1.

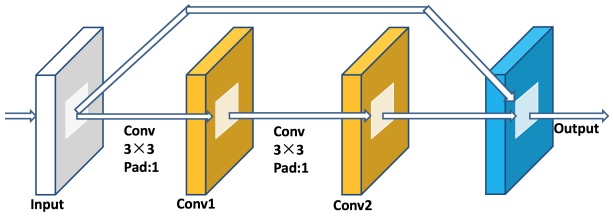

Figure 3: Building Block Structure(Fast-Model).

3.4 The building blocks of Fast-Model

Our network requires significant amount of computational cost on the building blocks in Fig. 2. Thus, in order to be able to balance the computational efficiency and accuracy, we simplify the building block structure in Fig. 3 and propose a Fast-Model. Particularly, the building block in our Fast-Model contains only 2 convolutional layers (Fig.3) comparing with the 3 convolutional layers in Full-Model. This will greatly reduce the computing cost and thus improve training speed, while our experiments show that the performance of Fast-Model is still very good.

4 EXPERIMENTS

4.1 Datasets and settings

Our models are trained on CASIA-WebFace [48] and CelebA [24], and we test our results on the widely used face datasets, MegaFace [18]. We further employ the LFW [15] dataset to evaluate the key components of our networks. Specifically,

Figure 4: Some face images in LFW dataset. The face image pairs in red bounding box are the same person pairs and the yellow images are the negative pairs.

CASIA-WebFace is currently the largest publicly released dataset for face verification and identification tasks. It contains 10,575 celebrities and 494,414 face images which are crawled from the web. Each person has 46.8 images on average.

CelebA is a facial attribute dataset which contains approximately 200k images. Each image is annotated with 5 landmarks (two eyes, the nose tips, the mouth corners) and binary labels of 40 attributes.

MegaFace is a very challenging dataset to evaluate the performance of face recognition algorithms. It includes gallery set and probe set. The gallery set consists of 690K different individuals with more than 1 million images. This images are collected from a subset of Flickr photos [38]. The sample images are shown in Fig. 5. The probe set used in the challenge are two existing databases: *Facescrub* [26] and *FGNet* [10]. *Facescrub* dataset is publicly available dataset, containing over 100K face images of 530 identities. *FGNet* dataset is a face aging dataset, consists of 1002 images of 82 individuals. Each identity has average 12 images ranging from 0 to 69. The gallery set are used as the distractors[1] to evaluates the performance of algorithms with the increasing size of "distractors" (going from 10 to 1M). It has been showed that training image size have a significant impact on the performance of the algorithm. To evaluate to algorithm more fairly, MegaFace challenge has two protocols (large or small training set). The training set is defined as large if it contains more than 0.5M images and 20K subjects. Otherwise, it is defined as small. Our training set has 0.44M images and 10k subjects which belong to the small training set protocol, in contrast to the previous work of 200M images [32] and 1.3M images [22]. We conduct our experiments on Facescrub.

LFW contains 13,233 web-collected images of 5749 different identities. These images has large variations in illuminations, pose and expression. LFW includes 6000 face pairs in 10 splits; each split with half positive pairs (the same persons) and half negative pairs (different persons). The sample images can be viewed in Fig 4.

[1]it means who are not in the testing set, i.e. faces of unknown persons.

Figure 5: Some face images in MegaFace dataset.

4.2 Competitors

Our model is compared against the exiting methods on face recognition tasks reported from the official websites of MegaFace[2]. We mainly compare those publicly released methods including Google FaceNet [32], Center Loss [43], Lightened CNN [44], and LBP [2], and Joint Bayes model [5].

As an open challenge, there are lots of methods from the commercial companies such as FaceAll, NTechLAB, SIAT_MMLAB, BareBonesFR, 3DiVi Company. Since the details of these methods are not still kept unknown to the community, it is hard to judge whether these methods are comparable with us. However, we still list these methods as a general reference.

For facial attribute prediction, we compare with the state-of-the-art method – MOON [31].

4.3 Evaluation metrics

For MegaFace dataset [18], there are two recognition tasks, i.e., *identification* and *verification*. We evaluate our models on both settings with the provided toolkit. The face attribute prediction is also evaluated here.

Face Identification. Given a probe photo (Facescrub) and a gallery containing at least one photo of the same identity, we need to compute the similarity between each of the gallery images and the probe face image given, as well as the rank orders of all the gallery images. Results are presented with Cumulative Match Characteristics (CMC) curves. It indicates the probability that a correct gallery image can be found in the first thousan images.

Face Verification. Given a pair of photos, we need to determine the person in the two images is the same or not. MegaFace challenges provides 4 billion negative pairs between probe and gallery datasets. The verification result is report with Receiver Operating Characteristic (ROC) curves. It explores the trade off between True Accept Rate (TAR) and False Accept Rate (FAR).

Facial attribute prediction. Given one attribute, we will compute the accuracy of facial attribute prediction. Note that the MegaFace dataset did not provide the ground-truth labels for the facial attributes. Since exhaustively annotating all the images used in MegaFace is expensive, we randomly sample 1000 images from faceScrub dataset and invite three annotators who are unknown to our projects to label the selected facial attributes. The ground-truth facial attributes are majority-voted among these three annotators. This sampled dataset is utilized to evaluate our algorithms.

In the real-world scenario, face recognition should achieve high accuracy with millions of distractors. In the MegaFace challenging setting, we compare Rank-1 identification accuracy with at least 1M distractors on the face identification task. For the face verification, verification accuracy at low false accept rate FAR = 10^{-6} is adopted. These two evaluation settings show the incredible difficulty of MegaFace challenging.

4.4 Parameter settings

To train the CNN model, we randomly select one face image from each identity as the validation set and the remaining images as the training set. We train our model for 70k iterations with $\lambda_1 = 0, \lambda_2 = 1$ in Eq 2. The training images are horizontally flipped for data augmentation. The base learning rate is set as 0.01, and gradually decreased by 1/10 at $40k$, $58k$. The size of input image is $112 \times 96 \times 3$. We use the stochastic gradient descent algorithm with mini-batch size of 40 in Full-Model and mini-batch size of 48 in Fast-Model. Dropout is used for fully connected layers and the ratio is set to 0.5. For fair comparison, all the CNN models share the same base architecture and the details are given in Fig 1.

We also fine-tune our model for 30k iterations with $\lambda_1 = 0.1, \lambda_2 = 1$ in Eq 2. The base learning rate is set as 0.001, and gradually decreased by 1/10 at $10k$, $20k$. The remaining parameter settings are consistent with the above. We implement our method using the open source deep learning framework Caffe [17]. For all the experiments, we train the deep model from the scratch and we only use a single end-to-end model for all the testing.

4.5 Preprocessing

We use the annotations [3] from[**?**] to improve the quality of training set. Then we preprocess the CASIA-WebFace dataset with the following 3 steps, i.e. *face detection*, *face landmark detection* and *face alignment*. (1) For face detection and face landmark detection, we use the the public tools [50]. If the detection step is failed, we discard the image; otherwise, we crop the face image with the provided bounding box. (2) We employ the method of [44] to detect the five facial landmarks (two eyes, nose and mouth corners). (3) We use the codes in [44] for face alignment. The obtained faces are normalized and resized to 112×96 RGB images. Finally, we obtain roughly 0.44M remaining images of 10,575 unique identities from CASIA-WebFace dataset.

4.6 Running cost

Our Full-Model get converged with 100, 000 iteration and it takes 17 hours on CASIA-WebFace with one NVIDIA TITANX GPUs. The batch size is 40, and it takes around 11 GB GPU memory. The Fast-Model gets converged with 100, 000 iteration and it costs 14 hours on one NVIDIA TITANX GPUs. The batch-size is fixed as 48 which uses round 10 GB GPU memory.

4.7 Experiment on MegaFace

We evaluate face identification and face verification tasks on the standard settings of the challenging MegaFace benchmark. For face identification, we plot the CMC curve of different methods with 1M distractors as shown in Fig 6 and list the Rank-1 identification accuracy at Tab. 2. In face verification, we plot the ROC curve of different methods also with 1M distractors listed in Fig. 7 and we compare the verification TAR at 10^{-6} FAR accuracy at Tab. 3.

[2]http://megaface.cs.washington.edu/results/facescrubresults.html

[3]http://zyb.im/research/mm16.html

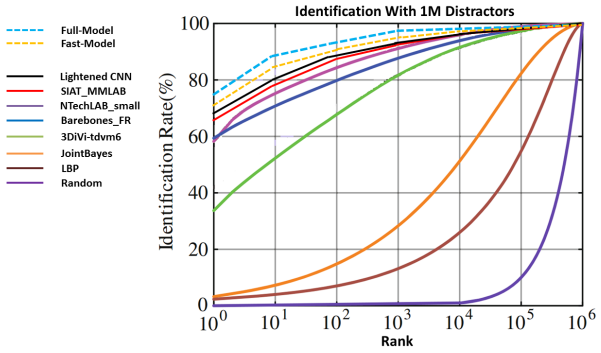

Figure 6: CMC curves of different methods (under the protocol of small training set) with 1M distractors. Note that the results of other other methods are reproduced from the official website of MegaFace dataset.

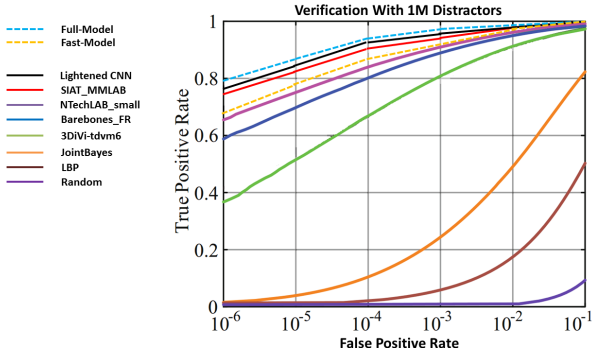

Figure 7: ROC curves of different methods (under the protocol of small training set) with 1M distractors. Note that the results of other other methods are reproduced from the official website of MegaFace dataset.

Table 2: Face identification Rank-1 results on MegaFace with 1M distractors. 'Release' indicates whether the details of this method are publicly released.

Method	Rels	Protocol	Acc. (%)
Google - FaceNet v8 [32]	✓	Large	70.50
NTechLAB - Large	×	Large	73.30
Faceall Co. - Norm_1600	×	Large	64.80
Faceall Co. - FaceAll_1600	×	Large	63.98
Lightened CNN [44]	✓	Small	67.11
Center Loss [43]	✓	Small	65.23
LBP [2]	✓	Small	3.02
Joint Bayes [5]	✓	Small	2.33
NTechLAB -Small	×	Small	58.22
3DiVi Company	×	Small	33.71
SIAT_MMLAB	×	Small	65.23
Barebones FR	×	Small	59.36
Fast-Model (without Attribute)	✓	Small	70.28
Fast-Model	✓	Small	73.07
Full-Model (without Attribute)	✓	Small	74.25
Full-Model	✓	Small	77.74

Table 3: Face verification on MegaFace with 1M distractors (TAR at 10^{-6} FAR). 'Rels' indicates whether the details of this method are publicly released.

Method	Released	Protocol	Acc. (%)
Lightened CNN [44]	✓	Small	77.46
Center Loss [43]	✓	Small	76.52
LBP [2]	✓	Small	2.17
Joint Bayes [5]	✓	Small	1.46
Barebones FR	×	Small	59.04
NTechLAB - Small	×	Small	66.37
3DiVi Company	×	Small	36.93
SIAT_MMLAB	×	Small	76.72
Fast-Model (without Attribute)	✓	Small	58.20
Fast-Model	✓	Small	68.56
Full-Model (without Attribute)	✓	Small	76.22
Full-Model	✓	Small	**79.24**

Comparing with all competitors, we want to highlight that *our method achieves the best performance on the small protocol among all the publicly released methods on MegaFace dataset*; specifically, we have

(1) our Full-Model beats the other methods on the protocol of small training set on the identification task. For face identification, our methods (both Full-Model and Fast-Model) has larger area under CMC curve than the other competitor methods as shown in Fig 6. This proves the efficacy of our model. Table 2 further shows the results on Rank-1 accuracy. As can be seen from the table, our Full-Model achieves 77.74% Rank-1 accuracy, outperforming all the other methods of the small protocol by clear margins (> 10%). Furthermore, it is also worth noticing that our model even outperform some models trained with very large training set (i.e., large protocol), e.g., Beijing Facecall Co. and Google's FaceNet. For example, Google - FaceNet v8 is trained on 200M face images; in contrast, our training set has only $0.44M$ face images.

(2) our Full-Model beats the other methods on the protocol of small training set on the verification task. The results are compared in Fig. 7 and Tab. 3. As we can see from the figure, our Full-Model can obtain the verification accuracy 79.24%; it shows 1.78% improvement over the second best method – Lightened CNN [44]. Note that, we argue that the verification task is more sensible to the scale of training data, and thus on the large protocol, the other competitors such as Google FaceNet v8 [32] achieving 86.47% enable better performance than ours (79.24%). However, our Full-Model on identification task can still beat Google FaceNet v8 [32] even in large protocol.

(3) our Fast-Model is very efficient and it can beat most of competitors on the small protocol. On face identification task, our Fast-Model achieve 73.07% and outperforms all the other methods on the small protocol. On face verification task, the result of Fast-Model is 68.56% which is higher than almost all the other methods with the only two exceptions – Lightened CNN [44] and Center Loss [43]. It shows that our Fast-Model is a reasonable good and fast version of our Full-Model in terms of running cost and performance.

Table 4: Results of facial attribute prediction. Our model is compared agains MOON. Each row is corresponding to one task of attribute prediction.

	MOON(%)	Ours (%)
Narrow Eyes	83.40	**85.60**
Big_nose	71.00	**75.30**
Pointy_nose	70.50	**73.90**
Chubby	93.30	**93.50**
Double_Chin	94.80	**96.70**
High_cheekbones	74.10	**76.80**
Male	96.90	**97.30**

(4) our Full-Model/Fast-Model with attribute constraints beats the models without learning attribute tasks. We compare another variants of our model, i.e., directly using our network for face recognition without learning facial attribute tasks. The experiments show that the results without attribute constraints are inferior to those of their corresponding models with attribute constraints on both identification and verification tasks. This validates that the jointly learning two tasks can help improve the performance of face recognition.

(5) our facial attribute prediction beat the results of MOON [31]. We also compare the results of jointly learned framework against the model of only learning the facial attribute task, i.e., MOON. Note that without learning the face recognition task, our model can be degraded into MOON. The results are compared in Tab. 4. On all the attributes, our methods achieve higher prediction accuracy than MOON. This validates that our multi-task framework can help with facial attribute prediction.

4.8 The ablation study on MegaFace and LFW dataset

To better reveal the insights and evaluate the key components of our model, we compare the variants of our model on the LFW and MegaFace dataset. For LFW dataset, we evaluate algorithms with the standard protocol of unrestricted labeled data [15]. We test our algorithm on 6000 face pairs and mean accuracy is reported.

Our building block structure is important to guarantee the good performance on the benchmark. We further investigate different variants of building block structures listed in Fig. 8. Particularly, there are 3 variant structure compared here, i.e. the building blocks of Fast-Model, Full-Model and the variants-block. The variants-Block takes two convolutional layer on the main road, and one convolutional layer on the bypass. We compare the results of Rank-1 identification accuracy on MegaFace and LFW datasets in Tab. 5. For all the three methods, the only difference is the structure of building blocks. We highlight that (1) the building blocks of our Full-Model structure outperforms the other structures on both two datasets. This reveals that arranging the convolutional layers in a deeper structure (*i.e.* 3 layers in the main road of our Full-Model) can indeed extract more discriminative features, and thus improve the recognition accuracy. (2) Comparing the building block of Full-Model block with that of variants-block model, we conjecture that convolutional layer at different lines (*i.e.* both main pass and by-pass in variants-block model) may confuse the residual learning of features, and thus it will harm the performance of corresponding architectures.

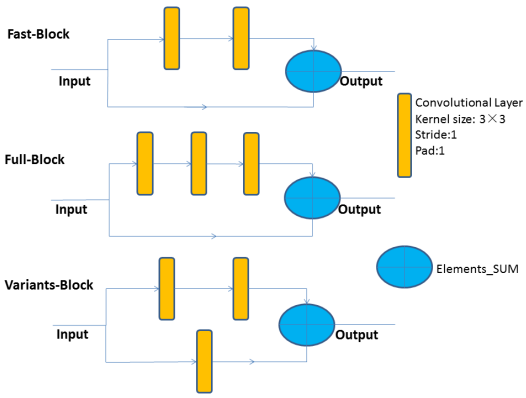

Figure 8: Different building blocks.

Table 5: Comparison of building block structure.

Method Names	MegFace (%)	LFW (%)
Fast-Block	73.07	97.62
Full-Block	**77.74**	**98.15**
Variants-Block	69.63	97.50

Table 6: Comparison of different configurations.

Method Names	MegFace (%)	LFW (%)
Configuration-1	**73.07**	**97.62**
Configuration-2	72.05	97.50
Configuration-3	69.40	97.60

The different configurations of the last layer of our network in Fig. 1. We compare different configurations of the final layer of our network. Specifically, we use the Fast-Model to retrain the network with 3 different structures listed in Fig. 9. For a faire comparison, all the other settings and structures are the same except the last layer. (1) Configuration-1 is the structure is used in our Full-Model and Fast-Model. (2) Configuration-2 is a shared structure of both tasks. (3) Configuration-3 enforces another independent fully connected layer for each individual task. The result on both MegaFace and LFW datasets are shown in Tab. 6. The results of our configuration is better than those of the others. We argue that these attributes have relatively high correlation and such correlation should be modeled separately beyond the face recognition task. Thus another fully connected layer is introduced in our configuration to help model the correlation among attribute. As results, it enables the better performance than the other two configurations.

The benefits of different kinds of facial attributes. It is another interesting question of which type/group of facial attributes in Tab 1 can mostly help the face identification tasks. To answer this question, we design another set of experiments. Particularly, still with the same Fast-Model structure, we compare the methods (1) without attribute, (2) with only eye attribute, (3) with nose attribute,

Figure 9: Different configurations of the last layers.

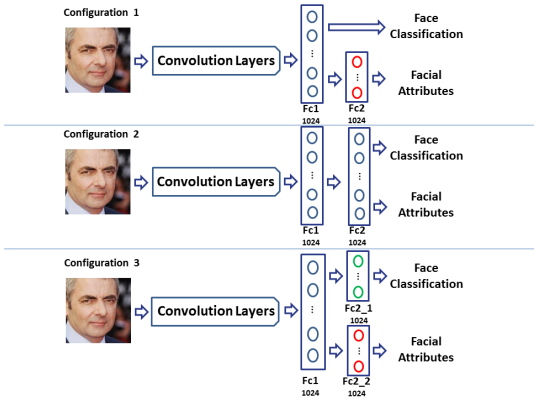

Table 7: Comparison of different groups of face attributes on face identification.

Method Names	MegFace (%)	LFW (%)
No attribute	70.28	97.20
Eye attribute	72.23	97.49
Nose attributes	71.72	97.63
Global attributes	71.92	97.45
All attributes	**73.07**	**97.62**

(4) with global attribute and (5) with all attributes. The results are listed in Tab 7. Comparing the model without attribute, it is obviously that using all facial attributes can improve face recognition performance. Further, we find that the eye attribute is more important than Global attributes in term of the identification accuracy on both dataset.

5 CONCLUSIONS

In this paper, we for the first time propose a novel joint deep architecture for face recognition and facial attribute prediction. In contrast to previous work of only considering one task only, our experimental results clearly show the benefit of the jointly learning structures and learning in such a way can help to capture both global feature and local attribute information simultaneously.

6 ACKNOWLEDGEMENT

This work is supported by the National Key Research and Development Plan (#2016YFC0801003), Shanghai Municipal Science and Technology Commission (#16JC1420401), Shanghai Sailing Program (#17YF1427500), the Shanghai Municipal R&D Foundation (#16511105402), Application of big data computing platform in Smart Lingang New City based BIM and GIS (#ZN2016020103).

REFERENCES

[1] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. 2015. Multi-task cnn model for attribute prediction. *IEEE TMM* (2015).
[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. 2006. Face description with local binary patterns: Application to face recognition. *IEEE TPAMI* (2006).
[3] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan F Klare, and Anil K Jain. 2014. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security* (2014).
[4] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. 2014. Face alignment by explicit shape regression. *IJCV* 107, 2 (2014), 177–190.
[5] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. 2012. Bayesian face revisited: A joint formulation. In *ECCV*.
[6] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. 2014. Joint cascade face detection and alignment. In *ECCV*.
[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*.
[8] Max Ehrlich, Timothy J Shields, Timur Almaev, and Mohamed R Amer. 2016. Facial attributes classification using multi-task representation learning. In *CVPR Workshops*.
[9] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE TPAMI* (2010).
[10] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Yuan Yao, and Shaogang Gong. 2014. Interestingness Prediction by Robust Learning to Rank. In *ECCV*.
[11] M. Guillaumin, J. Verbeek, and C. Schmid. 2009. Is that you? Metric Learning Approaches for Face Identification. In *CVPR*.
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015).
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *CoRR, vol. abs/1502.01852*.
[14] G. B. Huang, V. Jain, and E. Learned-Miller. 2007. Unsupervised Joint Alignment of Complex Images. In *IEEE International Conference on Computer Vision*. 1–8. DOI : http://dx.doi.org/10.1109/ICCV.2007.4408858
[15] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Technical Report.
[16] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*.
[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Gir-shick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv* (2014).
[18] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. 2016. The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. In *CVPR*.
[19] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. 2011. Describable visual attributes for face verification and image search. *IEEE TPMAI* (2011).
[20] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. 2009. Attribute and Simile Classifiers for Face Verification. In *ICCV*.
[21] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *ICCV*. IEEE, 365–372.
[22] Jinguo Liu, Yafeng Deng, and Chang Huang. 2015. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310* (2015).
[23] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval.
[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*.
[25] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60 (2004).
[26] Hong-Wei Ng and Stefan Winkler. 2014. A data-driven approach to cleaning large face datasets. In *ICIP*.
[27] Hieu V. Nguyen and Li Bai. 2010. Cosine Similarity Metric Learning for Face Verification. In *ACCV*.
[28] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. 2002. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE TPAMI* (2002).
[29] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. 2016. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249* (2016).
[30] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. 2014. Face alignment at 3000 fps via regressing local binary features. In *CVPR*. IEEE.
[31] Ethan M. Rudd, Manuel Gunther, and Terrance E. Boult. 2016. MOON:A Mixed Objective Optimization Network for the Recognition of Facial Attributes. In *ECCV*.
[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.
[33] William Robson Schwartz, Huimin Guo, and Larry S Davis. 2010. A robust and scalable approach to face identification. In *ECCV*.
[34] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. 2015. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873* (2015).
[35] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *CVPR*.

[36] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*. 1701–1708.

[37] Yaniv Taigman, Ming Yang, MarcAurelio Ranzato, and Lior Wolf. 2015. Web-scale training for face identification. In *CVPR*.

[38] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817* (2015).

[39] Georgios Tzimiropoulos and Maja Pantic. 2014. Gauss-newton deformable part models for face alignment in-the-wild. In *CVPR*. IEEE, 1851–1858.

[40] Paul Viola and Michael Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. In *CVPR*.

[41] P. Viola, J. Platt, and C. Zhang. 2005. Multiple instance boosting for object detection. In *NIPS*.

[42] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. 2016. Walk and Learn: Facial Attribute Representation Learning from Egocentric Video and Contextual Data. *arXiv preprint arXiv:1604.06433* (2016).

[43] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*.

[44] Xiang Wu, Ran He, and Zhenan Sun. 2015. A Lightened CNN for Deep Face Representation. *arXiv preprint arXiv:1511.02683* (2015).

[45] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. 2016. A Light CNN for Deep Face Representation with Noisy Labels. *arxiv* (2016).

[46] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *CVPR*.

[47] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, and Peter Robinson. 2015. Face Alignment Assisted by Head Pose Estimation. *arXiv preprint arXiv:1507.03148* (2015).

[48] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).

[49] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. 2014. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*. Springer.

[50] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE SP Letters* (2016).

[51] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *ECCV*.

[52] Yang Zhong, Josephine Sullivan, and Haibo Li. 2016. Face Attribute Prediction Using Off-the-Shelf CNN Features. In *arxiv*.

[53] Qiang Zhou, Gang Wang, Kui Jia, and Qi Zhao. 2013. Learning to share latent tasks for action recognition. In *ICCV*.

[54] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. 2015. Face Alignment by Coarse-to-Fine Shape Searching. In *CVPR*.

[55] Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*.