A Deep Face Identification Network Enhanced by Facial Attributes Prediction

Fariborz Taherkhani, Nasser M. Nasrabadi, Jeremy Dawson
West Virginia University
ft0009@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu, Jeremy.Dawson@mail.wvu.edu

Abstract

In this paper, we propose a new deep framework which predicts facial attributes and leverage it as a soft modality to improve face identification performance. Our model is an end to end framework which consists of a convolutional neural network (CNN) whose output is fanned out into two separate branches; the first branch predicts facial attributes while the second branch identifies face images. Contrary to the existing multi-task methods which only use a shared CNN feature space to train these two tasks jointly, we fuse the predicted attributes with the features from the face modality in order to improve the face identification performance. Experimental results show that our model brings benefits to both face identification as well as facial attribute prediction performance, especially in the case of identity facial attributes such as gender prediction. We tested our model on two standard datasets annotated by identities and face attributes. Experimental results indicate that the proposed model outperforms most of the current existing face identification and attribute prediction methods.

1. Introduction

Deep neural networks, particularly deep Convolutional Neural Networks (CNNs), have provided significant improvement in visual tasks such as face recognition, attribute prediction and image classification [16, 26, 28, 12, 19, 22]. Despite this advancement, designing a deep model to learn different tasks jointly while improving their performance by sharing learned parameters remains a challenging problem.

Providing auxiliary information to a CNN-based face recognition model can improve its recognition performance; however, in some cases such information is available only during training and may not be available during the testing phase. Despite the potential advantages of using auxiliary data, these problems have diminished the popularity and flexibility of using both soft and hard modalities for biometric applications [30].

We propose a model which jointly predicts facial attributes and identifies faces while simultaneously leverages the predicted facial attributes as an auxiliary modality to improve face identification performance. We also show that when our model is trained jointly to recognize face images and predict facial attributes, the model performance on facial attribute prediction increases as well. In other words, in our model the two modalities improve each other's performance once they are trained jointly. We show that some soft biometric information, such as age and gender which on their own are not distinctive enough for face identification, but, nevertheless provide complementary information along with other primary information, such as the face images.

Despite significant improvements in face recognition performance, it is still an ongoing problem in computer vision [3, 11, 24, 25, 27, 29]. There are a number of approaches in the literature that use facial attributes for biometrics applications such as face recognition. For example, Wang et al [33] propose an attribute-constrained face recognition model for joint facial attributes prediction and face recognition. In this model, the parameters of the network are first updated for attributes prediction and then same network is fine-tuned for face recognition. While Ranjan et al [23] add other face related tasks to improve overall performance. Their model is a single multi-task CNN network for simultaneous face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation and face recognition.

Facial attributes as semantic features can be predicted from face images directly, or from other facial attributes indirectly [32]. Attribute prediction methods are generally classified into local or global approaches. Local methods consist of three steps; first they detect different parts of the object and then extract features from each part. Finally, these features are concatenated to train a classifier [18, 4, 7, 2, 20, 37]. For example, Kumar et al's method [18] is based on extracting hand-crafted features from ten facial parts. Zhang et al [37] extract poselets aligning face parts to predict facial attributes. This method works improperly if object localization and alignment are not perfect. Global approaches, however, extract features from entire image disregarding object parts and then train a classifier on extracted features; these methods perform improperly if

2160-7516/18 $31.00 © 2018 IEEE
DOI 10.1109/CVPRW.2018.00097

large face variations such as occlusion, pose and lighting are present in the image [19, 13, 20].

Attribute prediction has been improved in recent years. Bourdev et al [5] propose a part-based attribute prediction method which deploys semantic segmentation in order to transfer localization information from the auxiliary task of semantic face parsing to the facial attribute prediction task. Liu et al [19] use two cascaded CNNs; the first of which, LNet, is used for face localization, while the second, ANet, is used for attribute description. Zhong et al [38] first localize face images and then use an off-the-shelf architecture designed for face recognition to describe face attributes at different levels of a CNN. He et al [36] propose a multi-task framework for relative attribute prediction. The method uses a CNN to learn local context and global style information from the intermediate convolution and fully connected layers, respectively.

Our network is inspired by multi-task network but we fuse the output of the attribute predictor into the face recognition layers which makes it different from other existing multi-task methods such as Wang et al's [33] approach. Our deep CNN model is constructed from two cascaded networks in which the final one consists of two branches, each of which are used for facial attribute prediction and face identification, respectively. Both these two branches communicate information together by sharing parameters of the first network in the model as well as fusing attribute branch with the last pooling layer of the face identification branch. In our model, all the parameters (i.e. the parameters of the two cascaded networks) are updated simultaneously in each training step.

The Contributions of our work are summarized as follows:

1) We design a new end to end CNN architecture that learns to predict facial attributes while simultaneously being trained with the objective of face identification. Our model shares learned parameters to train both tasks and also fuses attribute information and the face modality to improve face identification performance.

2) Contrary to the existing multi-task methods that only use a shared CNN feature space to train these two tasks jointly, our model uses a feature level fusion approach to leverage facial attributes for improving face identification performance. Furthermore, we observe that our jointly trained network is a more capable face attribute predictor than one trained on facial attributes alone.

The rest of this paper is organized as follows: The CNN architecture is described in section 2, fusion of attribute and face modalities is described in section 3, model training parameters are described in section 4, and finally, results and concluding remarks are provided in sections 5 and 6, respectively.

2. Deep Joint Facial Attributes Prediction and Face Identification Model

The proposed architecture predicts facial attributes and uses them as an auxiliary modality to recognize face images. The model is constructed from two successive cascaded networks as shown in Fig.1. The first network (net@1) uses the VGG 19 structure [26] with identical filter size, convolutional layers, and pooling operation. The first network applies filters with 3×3 receptive field. The convolution stride is set to 1 pixel. To preserve spatial resolution after convolution, spatial padding of the convolutional layer is fixed to 1 pixel for all 3×3 convolutional layers. Spatial pooling is performed by four max-pooling layers placed after the second, fourth, eighth, and twelfth convolutional layers and one global average pooling (GAP) layer which is placed after the sixteenth convolutional layer. Max-pooling is carried out on a 2×2 pixel window with a stride of 2. Each hidden layer is followed by a Rectified Linear Units (ReLU) [16] activation function. A GAP layer is a substantial process in our model because by disregarding the GAP layer and replacing it by a max-pooling layer, the output of the fusion layer will have a very high dimension when we fuse face and attribute modalities together. The GAP layer simply takes average of each feature map obtained from last convolutional layer. Since no parameter is optimized at the GAP layer, overfitting is prevented at this layer.

The second network (net@2) is divided into two separate branches trained simultaneously while communicating information together through the training process. Both of these branches consist of two fully connected (FC) layers operating on the output of the first network. The first FC layer of each branch (Fc1 and Fc $'1$ in Fig.1) consists of 4096 units. The next layers of (Fc1) and (Fc $'1$) are fully connected layers on which the soft-max operation is conducted. The first branch performs the attribute prediction task, and the output of the last FC layer in this branch before performing soft-max operation is fused with the GAP layer of net@1 by using Kronecker product [10]. Finally this fused layer is employed to train the second branch - the face identification task. As shown in Fig.1, attributes are predicted by net@1 and first branch of net@2 parameters while face images are identified by net@1 and all parameters in net@2; the overall proposed architecture is shown in Fig.1.

3. Fusion Layer on Facial Attributes and Face Modalities

Previously, feature concatenation has been used as an approach for multimodal fusion. In this work, we use the Kronecker product to fuse facial attributes features with face features. Since the Kronecker product of two vectors (i.e. attributes and face features) is mathematically formed by a

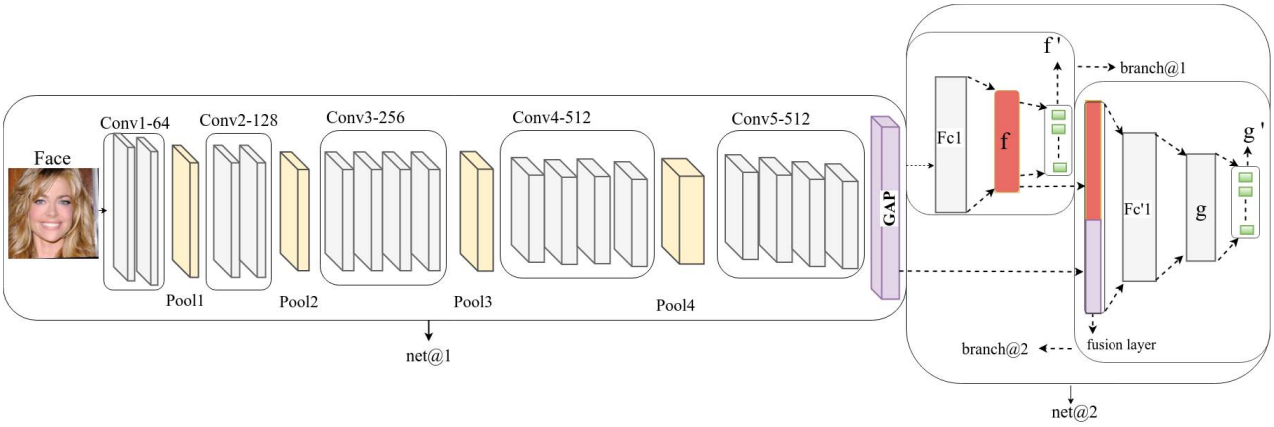

Figure 1. Proposed CNN architecture, face identification and attribute prediction are trained jointly.

matrix direct product, there are no learnable parameters at this layer and, consequently, chances of overfitting are low at this layer. Furthermore, we argue that, due to existing correlation between facial attributes features and face features, the output neurons of the fusion layer are simple to interpret and are semantically meaningful. (i.e., the manifold that they will lie on is not complex, however, it is just high dimensional). Therefore, it is simple for the following layers of the network to decode such meaningful information. Assume that **v** and **u** are the feature vectors of attributes and face, respectively. The Kronecker product of these two vectors is defined as follows:

$$\mathbf{u} \otimes \mathbf{v} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \otimes \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} = \begin{bmatrix} u_1 v_1 \\ u_1 v_2 \\ \vdots \\ u_1 v_m \\ u_2 v_1 \\ \vdots \\ u_n v_m \end{bmatrix} \tag{1}$$

4. Training our CNN architecture

In this section, we describe how we train our model. Thousands of images are needed to train such a deep model. For this reason, we initialize net@1 parameters by a CNN pre-trained on the ImageNet dataset and then we fine tune it as a classifier by using the CASIA-Web Face dataset. CASIA-Web Face contains 10,575 subjects and 494,414 images. As far as we know, this is the largest publicly available face image dataset, second only to the private Facebook dataset.

The proposed deep network is described as a succession of two cascaded networks. net@1 is constructed from 16 layers of convolutional operations on the inputs, intertwined with ReLU non-linear operation and five pooling operations. Weights in each convolutional layer form a sequence of *4-d* tensors; $W \in \mathbb{R}^{l \times c \times p \times q}$ where l, c, p and q are dimensions of the weights along the axes of filter, channel, and spatial width and height, respectively. For notational simplicity, we denote all the weights in net@1 with W_1 and the weights in net@2 with W_2. W_2 is separated into two groups of $W_{2,1}$ and $W_{2,2}$ representing all weights in the first and second branches, respectively.

$\mathcal{L}_1$ and $\mathcal{L}_2$ described in (2) and (3) are the loss functions designed to perform attribute prediction and face identification tasks, respectively. We use the cross entropy as our network loss functions. T, C and $\mathbf{X} = \{x_i\}_{i=1}^{N}$ indicate the number of facial attributes used in the model, number of classes and the training samples, respectively. L'_i and L_{ji} represent face label and facial attribute label for attribute j and the training sample i, respectively. f and g functions are outputs of the network for attribute prediction and face identification tasks, respectively. f' and g' are soft-max functions performed on the f and g outputs, respectively. The loss functions represented in (2) and (3) show how two branches of net@2 communicate information and update their learning parameters with each other. As shown in (2) and (3), the f function (attribute prediction output) takes W_1 and $W_{2,1}$ as input. The g function (face identification output) takes W_1, $W_{2,2}$ and f as input. Therefore, both attribute prediction and face identification use W_1 as shared parameters. Furthermore, attribute prediction parameters and $W_{2,2}$ are used for face identification.

We use an Adam optimizer [15] to minimize our network's loss functions. The Adam optimizer is a robust and well-adapted optimizer that can be applied to a variety of non-convex optimization problems in the field of deep neural networks. All parameter values used in Adam optimizer are initialized using the authors' suggestion; we set learning rate to 0.001 to minimize our network's loss functions.

The optimization algorithm mainly consists of two steps, the first of which calculates the gradient of the loss functions with respect to the model parameters, and then, for the

second step, updates the biased first moment estimate and the model parameters, successively.

$$\mathcal{L}_1(W_1, W_{2,1}, X) = -\sum_{j=1}^{T}\sum_{i=1}^{N} L_{ji} log(f'(f(L_{ji}|x_i, W_1, W_{2,1}))) + (1 - L_{ji}) log(f'(f(1 - L_{ji}|x_i, W_1, W_{2,1}))) \quad (2)$$

$$\mathcal{L}_2(W_1, W_{2,1}, W_{2,2}, X) = -\sum_{i=1}^{N}\sum_{k=1}^{C} L'_{ik}\ log(g'(g(L'_{ik}|x_i, W_1, W_{2,2}, f(x_i, W_1, W_{2,1})))) \quad (3)$$

We iterate this algorithm through several epochs for the complete training batches until training error convergence is achieved.

5. Experiment

We conducted experiments for two different cases to examine if our model improves overall performance in identification and prediction tasks. In the first case, we train and test the model to perform two tasks separately in isolation, while in the second case we employ our model to train both tasks jointly. In the second case, however, we predict facial attributes assuming that such information is not available during the testing phase, and then outputs of the attribute prediction branch before performing the soft-max operation is fused with the last pooling layer of net@1 by using the Kronecker product. We fuse the face modality with those facial attributes such as gender and face shape which remain the same in all images of a person. Experimental results show that our model increases overall performance in face identification as well as attribute prediction in comparison to the first case. We performed our experiments on two GeForce GTX TITAN X 12GB GPU. We ran our model through 100 epochs using batch normalization (i.e. shifting inputs to zero-mean and unit variance) after each convolutional and fully connected layer before performing nolinearity. Batch normalization potentially helps to achieve faster learning as well as higher overall accuracy. Furthermore, batch normalization allows us to use a higher learning rate, which potentially provides another boost in speed. We used TensorFlow to implement our network. The batch size in all experiments is fixed to 128.

5.1. Datasets

We conducted our experiments on the CelebA dataset [19] for facial attribute prediction, as well as MegaFace [14] which is a widely used and well-known face datasets for face identification.

Figure 2. : First and second rows are image samples in CelebA dataset; third and forth rows are samples of aligned face images in MegaFace dataset.

CelebA is a large-scale, richly annotated face attribute dataset containing more than 200K celebrity images, each of which is notated with 40 facial attributes. CelebA has about ten thousand identities with twenty images per identity on average. This dataset is also annotated by five landmarks. The dataset can be used as the training and testing sets for facial attribute prediction, face detection, and landmark (or facial part) localization. To compare our method fairly with the other methods, we use the same setup that they have used. We use images of 8000 identities for training and remaining 1000 identities for testing. Train and test sets are available here.[1]

MegaFace is a publicly available and very challenging dataset which is used for evaluating the performance of face recognition algorithms with up to a million distractors (i.e., up to a million people who are not in the test set). MegaFace contains 1M images from 690K individuals with unconstrained pose, expression, lighting, and exposure. MegaFace captures many different subjects rather than many images of a small number of subjects. The gallery set of MegaFace is collected from a subset of Flickr [31]. The probe set of MegaFace used in the challenge consists of two databases; Facescrub [21] and FGNet [9]. FG-NET contains 975 images of 82 individuals, each with several images spanning ages from 0 to 69. Facescrub dataset contains

[1] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

more than 100K face images of 530 people. The MegaFace challenge evaluates performance of face recognition algorithms by increasing the numbers of distractors (going from 10 to 1M) in the gallery set. Training size is important, since it has been shown that face recognition algorithms that were trained on larger sets tend to perform better at scale. In order to evaluate the face recognition algorithms fairly, MegaFace challenge has two protocols including large or small training sets. If a training set has more than 0.5M images and 20K subjects, it is considered as large. Otherwise, it is considered as small. We use a small training set which has 0.44M images and 10k subjects. The prob set in our experiments is Facescrub.

5.2. Evaluation metrics

We evaluate the face identification performance of our model on the MegaFace dataset; and facial attribute prediction performance on the CelebA dataset. The MegaFace dataset is not annotated by facial attribute. Our model, however, predicts facial attributes and then uses them for face identification. To conduct experiments on the MegaFace dataset, we restore the model parameters trained on the CelebA dataset, which is annotated by facial attributes as well as people identification, and then fine-tune the model parameters on the MegaFace dataset for the objective of face identification on the MegaFace dataset. Our model predicts facial attributes from the first branch of our architecture and employs this auxiliary modality for face identification.

Face Identification: we calculate the similarity between each of the images in the gallery set and given image from the probe set, and then rank these images based on the obtained similarities. In face identification, the gallery set should contain at least one image of the same identity. We evaluate our model by using rank-1 identification accuracy as well as Cumulative Match Characteristics (CMC) curves. CMC is a rank-base metric indicating the probability of the correct gallery image that can be found in the top k similar images from the gallery set.

Facial Attribute Prediction: We leverage identity facial attributes as an auxiliary modality for improving face identification performance. Identity facial attributes are invariant attributes which remain same from different images of a person. For example, gender, nose and lips shapes remain the same in different images of a person; however, attributes such as glasses, mustaches, or beards may or may not exist in different images of a person. We discard such attributes in our model because we look for robust as well as invariant facial attributes. Identity facial attributes in CelebA dataset are listed as follows: narrow eyes, big nose, pointy nose, chubby, double chin, high cheekbones, male, bald, big lips and oval face . We evaluate our attribute predictor by using accuracy metric.

5.3. Methods for Comparisons

Attribute Prediction: We compare our method with several competitive algorithms including FaceTracer, PANDA[37], ANet+LNet [19] and MT-RBM-PCA [8]. FaceTracer [17] extracts handcraft features including color histogram and HOG from some functional face image region and then concatenates these features to train a SVM classifier for predicting attributes. Functional regions are determined by using ground truth landmarks. PANDA mainly was proposed by creating an ensemble of several CNNs for body attributes prediction. Each CNN in this model extracts features from a well-aligned human part using poselet. Next, all of the extracted features are concatenated to train a SVM for body attribute prediction. However, for our case, it is simple to adjust this method for facial attribute prediction such that the face part is aligned using landmark points. In ANet+LNet method, images of the first 8000 identities, which is roughly 162k images, are employed for pre-training and face localization. The images of the next 1000 identities, which is roughly 20k images, are used to train a SVM classifier. We use same testing and training sets to conduct our experiment. We compare our model with the other methods for attribute prediction. Table.1 shows the model improvement on identity facial attribute prediction once the model trains both tasks jointly. The results shows that joint-training has higher contribution for the attributes of gender, bald, narrow eyes, big lip, big nose, oval face, young, high cheekbone and chubby, respectively.

Face Identification: We compare our method with the exiting methods on face identification which are reported from the official websites of MegaFace[2]. We primarily compare with publicly released methods, for which the details are known. These methods are listed as follows: Google FaceNet [24], Center Loss [34], Lightened CNN [35], LBP [1] and Joint Bayes model [6].

There are several other methods from commercial companies such as FaceAll, NTechLAB, SIAT MMLAB, BareBonesFR, 3DiVi companies, the details of which are not known to the community yet. Therefore, we can not compare these methods with ours fairly; however, we report these methods to provide a comprehensive list of references on the Megaface dataset. Fig. 3 represents CMC curves for different methods; it is shown that our model covers larger area under the curve in comparison to the other methods. We also compare our model performance when the model trains facial attributes prediction and face identification jointly and separately. The results show that our face identifier benefits from joint training. We also compare performance of the algorithms by rank-1 identification accuracy; Table.2 compares face identification models on

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

	FaceTracer	PANDA	LNets+ANet	RBM-PCA	Ours-S	Ours-J
Bald	89	96	98	98	96.16	98.93
Big Lips	64	67	68	69	69.25	71.69
Big Nose	74	75	78	81	82.35	84.67
Chubby	86	86	91	95	94.22	95.27
High Cheekbones	84	86	88	83	86.61	87.79
Male	91	97	98	90	95.65	98.61
Narrow Eyes	82	84	81	86	85.45	87.9
Oval Face	64	65	66	73	74.49	75.94
Young	80	84	87	81	87.12	88.54

Table 1. Comparing attribute prediction models on CelebFacesA dataset.

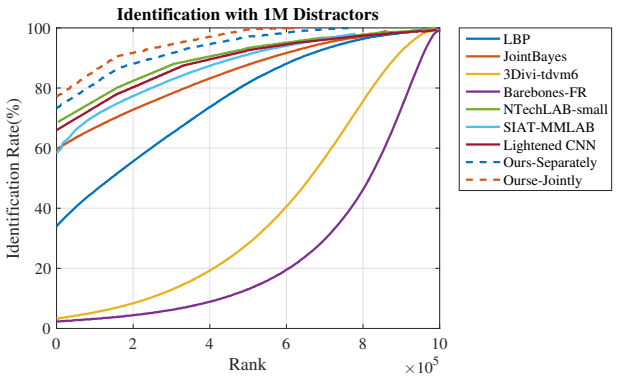

Figure 3. CMC curves of different methods with the protocol of small training set by 1M distractors. Please note that results of the other methods are reported from official website of MegaFace dataset.

Methods	Rels	Protocol	Acc%
Google - FaceNet v8	✓	Large	70.5
NTechLAB - Large	×	Large	73.3
Faceall Co. - Norm-1600	×	Large	64.8
Faceall Co. - FaceAll-1600	×	Large	63.98
Lightened CNN	✓	Small	67.11
Center Loss	✓	Small	65.23
LBP	✓	Small	3.02
Joint Bayes	✓	Small	2.33
NTechLAB -Small	×	Small	58.22
3DiVi Company	×	Small	33.71
SIAT-MMLAB	×	Small	65.23
Barebones FR	×	Small	59.36
Wang et al [33]	✓	Small	77.74
PM-Separately	✓	Small	76.15
PM-Jointly	✓	Small	**78.82**

Table 2. Comparing face identification models on MegaFace dataset using rank-1 identification accuracy metric.

MegaFacedataset using rank-1 identification accuracy metric. The results show the superiority of our model. We also observe that the model performance increases about 2.5% if the model train attributes and face jointly in comparison to the case which the model is trained separately.

5.4. Further Analysis

Experimental results included in Table.2 show that our model improves face recognition performance by leveraging identity facial attributes. To verify this claim, we conducted experiments for two different cases described earlier. In the second case we emphasize predicting facial attributes, because in a real face identification scenario, such information is not available during the testing phase. To use facial attributes as an auxiliary modality in our proposed model for face identification, we fused this modality with the last pooling layer of the model shown in Fig.1. The second case of our model which uses the predicted attributes outperforms the first case which does not use any privilege data.

Experimental results show that training the two tasks jointly increases not only face identification performance, but also facial attribute prediction performance, especially on identity facial attributes such as gender. For example, experiments performed on the CelebA dataset indicate that performance on face attributes including narrow eyes, big nose, pointy nose, chubby, double chin, high cheekbones, male, bald, big lips and oval face is improved around 2% on average if the tasks are trained jointly. Moreover, as shown in Table.2, our proposed model outperforms the accuracy of the state of the art methods for identity facial attributes prediction. One of the intuitive reasons causing this improvement is that, once our deep CNN model is trained to identify face images, it also learns more accurate face attributes in order to perform better face identification. In other words, these two modalities enhance each others' performance once they are trained jointly.

Table.2 also indicates that using facial attributes as privileged data boosts the model performance on face identification task. Our model beats most of the face identification algorithms used in the MegaFace data set challenge.

Inspired by the work in [39] on class activation map, we

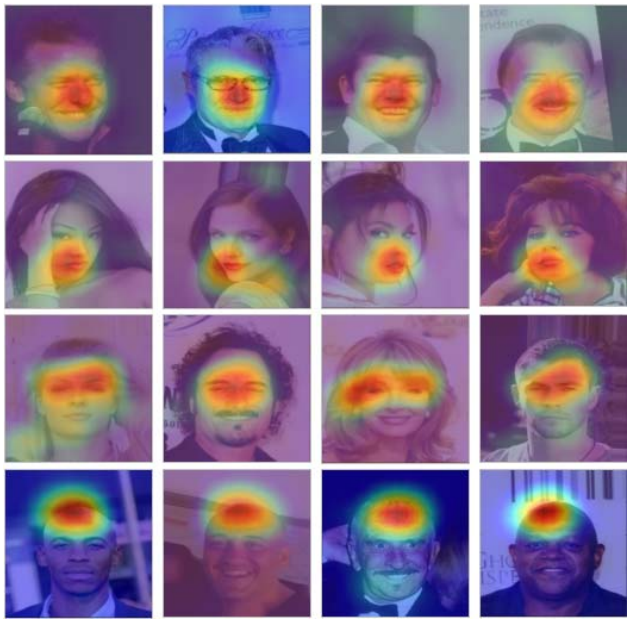

Figure 4. Example of class activation map generated from attribute predictor part of our model. Each row indicates nose attribute , mouth attribute, eyes attribute and head attribute , respectively. We observe that highlighted regions are activated by class activation map algorithm.

interpret the prediction decision made by our proposed architecture. Fig.4 shows the class activation map for predicting big nose, big lips, narrow eyes and bald, respectively. We can see that our model is triggered by different semantic regions of the image for different predictions. Fig.4 shows that our model due to using GAP layer also learns to localize the common visual patterns for the same facial attribute. Furthermore, the deep features obtained from our attribute predictor branch can also be used for generic facial attribute localization in any given image without using any extra information such as bounding box.

6. Conclusion

In this paper, we proposed an end to end deep network to predict facial attributes and identify face images simultaneously with better performance. Our model trains these two tasks jointly through shared CNN feature space, and also fuses predicted identity attributes modality with face modality features to improve face identification performance. The model increases both face recognition and face attribute prediction performance in comparison to the case when the model is trained separately. Experimental results show the superiority of the model in comparison to the current face identification models. The model also predicts identity facial attributes better than the state of the art models.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006. 5

[2] T. Berg and P. N. Belhumeur. 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*, pages 955–962, 2013. 1

[3] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, 2014. 1

[4] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1543–1550. IEEE, 2011. 1

[5] L. D. Bourdev. Pose-aligned networks for deep attribute modeling, July 26 2016. US Patent 9,400,925. 2

[6] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision*, pages 566–579. Springer, 2012. 5

[7] J. Chung, D. Lee, Y. Seo, and C. D. Yoo. Deep attribute networks. In *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, volume 3, 2012. 1

[8] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–55, 2016. 5

[9] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao. Interestingness prediction by robust learning to rank. In *European conference on computer vision*, pages 488–503. Springer, 2014. 4

[10] A. Graham. Kronecker products and matrix calculus: With applications (mathematics and its applications) pdf. 1981. 2

[11] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 498–505. IEEE, 2009. 1

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[13] M. M. Kalayeh, B. Gong, and M. Shah. Improving facial attribute prediction using semantic segmentation. *arXiv preprint arXiv:1704.08740*, 2017. 2

[14] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 4

[15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2

[17] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *European conference on computer vision*, pages 340–353. Springer, 2008. 5

[18] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009. 1

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 1, 2, 4, 5

[20] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2864–2871, 2013. 1, 2

[21] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 343–347. IEEE, 2014. 4

[22] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015. 1

[23] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 17–24. IEEE, 2017. 1

[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1, 5

[25] W. R. Schwartz, H. Guo, and L. S. Davis. A robust and scalable approach to face identification. In *European Conference on Computer Vision*, pages 476–489. Springer, 2010. 1

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2

[27] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2746–2754, 2015. 1

[30] V. Talreja, M. C. Valenti, and N. M. Nasrabadi. Multibiometric secure system based on deep learning. *arXiv preprint arXiv:1708.02314*, 2017. 1

[31] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 1(8), 2015. 4

[32] R. Torfason, E. Agustsson, R. Rothe, and R. Timofte. From face images and attributes to attributes. In *Asian Conference on Computer Vision*, pages 313–329. Springer, 2016. 1

[33] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 365–374. ACM, 2017. 1, 2, 6

[34] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 5

[35] X. Wu, R. He, and Z. Sun. A lightened cnn for deep face representation. In *2015 IEEE Conference on IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 4, 2015. 5

[36] L. C. Yuhang He and J. Chen. Multi-task relative attribute prediction by incorporating local context and global style information. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 131.1–131.12. BMVA Press, September 2016. 2

[37] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1644, 2014. 1, 5

[38] Y. Zhong, J. Sullivan, and H. Li. Face attribute prediction using off-the-shelf cnn features. In *Biometrics (ICB), 2016 International Conference on*, pages 1–7. IEEE, 2016. 2

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016. 6