FACE ATTRIBUTE PREDICTION WITH CLASSIFICATION CNN

Yang Zhong **Josephine Sullivan** **Haibo Li**
Computer Science and Communication
KTH Royal Institute of Technology
100 44 Stockholm, Sweden
{yzhong, sullivan, haiboli}@kth.se

Abstract

Predicting facial attributes from faces in the wild is very challenging due to pose and lighting variations in the real world. The key to this problem is to build proper feature representations to cope with these unfavourable conditions. Given the success of convolutional neural network (CNN) in image classification, the high-level CNN feature as an intuitive and reasonable choice has been widely utilized for this problem. In this paper, however, we consider the mid-level CNN features as an alternative to the high-level ones for attribute prediction. This is based on the observation that face attributes are different: some of them are locally oriented while others are globally defined. Our investigations reveal that the mid-level deep representations outperform the prediction accuracy achieved by the high-level abstractions. We demonstrate that the mid-level representations achieve state-of-the-art prediction performance on CelebA and LFWA datasets. Our investigations also show that by utilizing the mid-level representations one can employ a single deep network to achieve both face recognition and attribute prediction.

1 Introduction

Telling attributes from face images in the wild is known to be very useful in large scale face search, image understanding and face recognition. However, the problem indeed is very challenging since the faces captured in the real world affected by many unfavourable influences, such as illumination, pose and expression. To build usable attribute prediction, it is important to preserve the essential traits and at the same time, make the representations less sensitive to interference.

The representations describing faces in prior literature generally formed two groups: the hand-crafted representations and learned representations. Exemplified as in [7, 2], local low-level features were constructed from detected face regions. The use of local features were mostly based on the consideration that they are less likely to be influenced than the holistic features by pose changes and facial expressions. In recent work [8], hierarchical Gabor features were used as a holistic face representation, which is then converted by learned hashing process for efficient face retrieval and attribute prediction.

Driven by the great improvements brought by the CNN in image classification [6, 14, 12] and face recognition [13, 9, 15, 11] , features extracted from deep architectures became a natural and reasonable choice to represent faces for attribute prediction. In [17], local semantic image patches were first detected and fed into deep networks to construct concatenated, pose normalized representation. In [18], Liu et al. trained two concatenating CNNs to locate and predict attributes from arbitrary size of faces in the wild. The first CNN was trained by image categories and face attribute tags to locate face from complex background; the second stage was trained by identities and face attributes to achieve an effective fusion of the discrimination of inter-person representations and the variability

of non-identity related factors in face the representations. As a common character, the high-level hidden (fully connected) layer was used for feature representation in these work.

The best representation has been shown to be the first fully connected (FC) layer for retrieval tasks [1, 10]. Naturally, it is commonly used for attribute prediction. Considering that different levels of ConvNet encode different levels of abstraction, one can expect that such representations may not be optimal to describe the physical facial characteristics, especially the local attributes, such as "open mouth" and "wearing glasses". It is therefore rational to consider earlier layers of representation in the CNN for attribute prediction due to their discriminating power and the embedded rich spatial information.

In this work, our focus is on the face representations for face attribute prediction. Our intuition is that the representations from the earlier layers in CNNs are likely to better describe the facial appearance than high level features from the last layer. To validate this, we employ publicly available data, architecture and a deep learning framework to train a classification CNN and extract hierarchical face representations for further study. Through intensive investigations, we empirically show the effectiveness of the hierarchical deep representations and demonstrate the advantages of the mid-level representations for tackling the face attribute prediction problem. The major contributions of this work are:

- Our investigations reveal the fact of the diverse utilities of deep hierarchical representations for face attribute prediction; intermediate representations are shown very effective for predicting describable face attributes. (Section 2.2)
- By jointly leveraging the mid-level hierarchical representations, we further improve the state-of-the-art on two large scale datasets. (Section 2.3)
- One can construct a single deep network for both face recognition and attribute prediction.

2 Investigation

Our motivation in this work is to study the effectiveness of hierarchical CNN representations for attribute prediction of faces captured in the wild. Our procedure is to first extract hierarchical representations from aligned faces using our trained CNN and then we construct and evaluate attribute classifiers to identify the most effective representation of each attribute. In the following sections, we first describe the experimental details and then present our investigations and results on the CelebA dataset and annotated LFW dataset.

2.1 Implementation details

CNN: The face representations studied in this work were extracted from a face classification CNN. The architecture of the CNN with configuration details is shown in Table 1. We adopted the structure of the convolutional filters of "FaceNet NN1 [11]" , with slight modifications, into our CNN and concatenated two $512d$ FC layers and trained it in a N-way classification manner. The network was initialized from random and started with a learning rate of 0.015, which was then decreased two times when the classification accuracy stopped increasing on the validation set. PReLU [3] rectification was used after all the convolutional and FC layers. Dropout layers were inserted between FC layers with dropout rate of 0.5 to prevent overfitting. We used around 10000 identities with 350000 face images from the publicly accessible dataset WebFace [16] for training. The training instances were augmented with random flipping and slight rotation. The trained CNN had a verification accuracy of 97.5% on LFW dataset [4], which is totally comparable to that of the DeepID [13] structure used by [18].

Feature Representations: In our experiments faces were aligned based on the detected feature points [5] (or provided feature points). The aligned image, covering from the top of head to neck, had a size of 120×120. The center patch, with a size of 112×112, and its horizontal flipped patch, were fed in to the CNN. We extracted the average representations at $Conv2$, $Conv3$, $Conv4$, $Conv5$, $Conv6$, $FC1$ and $FC2$ layers and explored the utility of these representations for attribute prediction. To reduce the redundancy, we additionally applied multi-scale spatial pooling to the output to improve the invariance of CNN activations without degrading their discriminative power. The first average pooling halved the dimension and the second overlapping max. pooling selected

Table 1: CNN architecture used in our experiments. The perception size is 112 × 112 (except experiments in Section 2.4). The dimensions of the investigated deep representations are shown in the last column.

Layer	Kernel	Output	Representation Dimension
Conv 1	conv3-64		
Pool 1	max 2,2	56*56*64	
RNorm 1	5		
Conv 2a	conv1-64		
Conv 2	conv3-192	28*28*192	3* 3* 192
RNorm 2	5		
Pool 2	max 2,2		
Conv 3a	conv1-192		
Conv 3	conv3-384	14*14*384	3* 3* 384
Pool 3	max 2,2		
Conv 4a	conv1-384		
Conv 4	conv3-256	14*14*384	3* 3* 384
Conv 5a	conv1-256		
Conv 5	conv3-256	14*14*384	3* 3* 384
Conv 6a	conv1-256		
Conv 6	conv3-256	7*7*256	3* 3* 256
Pool 6			
FC 1	512	1*1*512	
FC 2	512	1*1*512	

the most locally activated neuron [1]. That is, all the resulting representations from convolutional layers had the same feature map size of 3 × 3 as shown in the last column of Table 1. The resulting features are denoted by $C2$, ... , $C6$, $F1$ and $F2$ in the following.

Dataset. In this work, we used two datasets, CelebA and LFWA [2], to conduct benchmark evaluations for attribute prediction accuracy. The CelebA contains around $200,000$ images of about $10,000$ identities and LFWA has $13,233$ images of $5,749$ identities. Images on both datasets are labeled with 40 binary codes to represent the presence of facial attributes. These attributes ranges from demographic information to physical characteristics such as nose size, eyebrow shape and expressions.

2.2 Exploration

The attribute prediction power of the hierarchical representations was first investigated on the training set of the CelebA dataset. We used the training set defined in [18] on both datasets to construct linear SVM attribute classifiers for each representation. The prediction accuracy on all the 40 attributes for each type of representation is demonstrated in Figure 1. It can be observed that features the intermediate conv. layers (Conv3 to Conv6) demonstrate an obvious advantage over the final FC layer on average, especially for attributes describing motions of the mouth area where the gap is almost 20%. Feature $C6$ had the highest prediction accuracy on average, which was more than 2% higher than the last FC layer. One can also observe that the mid-level conv. representations also featured slightly different prediction power on different attributes, e.g. $C2$ outperformed on "Rosy Cheeks" but not on hair related attributes. Through the exploration, therefore, we were able to identify the best representation for each attribute.

2.3 Benchmark evaluations

We then identified the best representations of each attribute on CelebA and LFWA datasets respectively and compared our results with a baseline approach and the current state-of-the-art selected from [18] in Figure 2. The baseline approach (denoted by [17]+ ANetin in [18]) used CNN trained

[1] E.g. , for Conv2 layer, the first stage is average pooling (window 4 × 4, stride 4) and the second one is max. pooling (3 × 3, 2).

[2] Available at: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html. Released in Oct 2015.

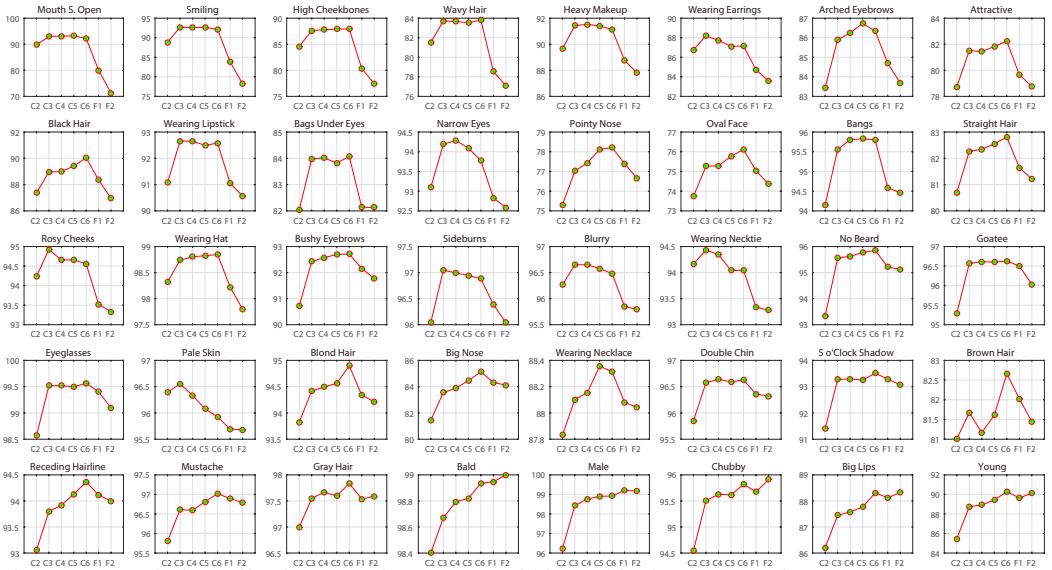

Figure 1: Prediction effectiveness discrepancy of hierarchical representations over 40 attributes on the training set of CelebA. In each grid, y-scale stands for the prediction accuracy in %. The mean prediction accuracy values of each representation (from $C2$ - $F2$) are 89.8% , 91.3%, $91,3\%$, 91.4%, 91.5%, 90.0% and 89.2%.

by identities to extract features from aligned faces; the process is the same as ours. The current best, denoted by "LNet+ANet" in [18], employed concatenated CNNs to locate faces and extract features in order to construct attribute classifiers. One can observe that: by jointly leveraging the best representations, superior performance has been achieved over the equivalent baseline approach and even the current best. For some attributes with relatively lower (baseline) accuracy, e.g. , "Blurry" "Oval Face", "Wearing Necktie" and "Rosy Cheeks", the advantage of our approach is especially significant. Among the evaluated 40 attributes, the mid-level conv. features dominated the best representations (only 4 best representations came from FC on CelebA and 1 on LFWA).

In addition, we also comprehensively compared the overall performance in Table 2 and layer wise prediction accuracy in Table 3. It can been found that: 1) utilizing the best representations demonstrates **outperforming effectiveness** over the equivalent baseline method and the current art approach on both large scale datasets; 2) the prediction power increases along the conv. layers of the CNN. Even the very early layer $C2$ astonishingly achieved entirely comparable performance as state-of-the-art solution. Indeed, even if we utilize conv. features on the attributes which had FC as the best representations (e.g. "Bald"), the effect on prediction performance is not detectable (this also reflects in Figure 1).

Moreover, it is worth nothing that using features at FC layers is computationally expensive due to the bottle neck between conv. and FC layer. Thus, leveraging the mid-level conv. representations features two inherent advantages: 1) very **efficient** inference with competitive performance; 2) breaking the bounds of the interconnections between conv. and FC layers so that the CNN accepts arbitrary perception size (as demonstrated in Section 2.4).

Recall that the training effort of a face classification CNN is almost minimal compared to the state-of-the-art, two-stage CNN approach and other potential end-to-end learning for attribute prediction. However, the deep hierarchical representations from the trained CNN still allow superior performance. This indicates that these representations from mid-level of CNN have contained rich spatial information which can be utilized more in constructing effective appearances describing, attribute prediction and retrieval applications.

2.4 Discussion

It is easy to expect a number of factors that affect the prediction performance, e.g. image resolution. A small perception size would affect the prediction relating to small scale face traits. We found that

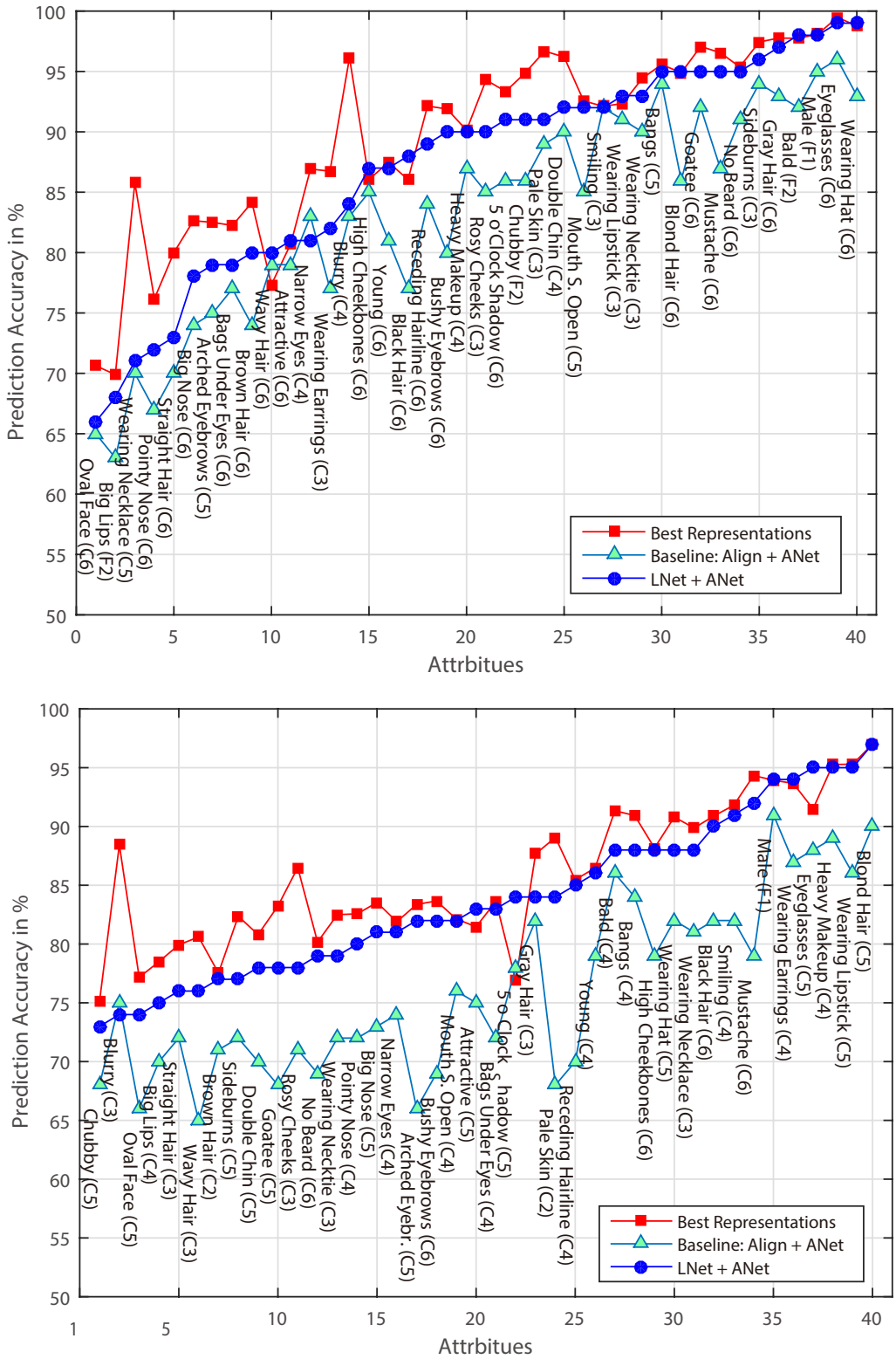

Figure 2: Comparative prediction accuracy on CelebA (up) and LFWA (down) with best performing representation given in parenthesis after each attribute (refer numeric accuracy values to Table 1s in the supplementary document).

Table 2: Overall comparisons of the baseline approach, current state-of-the-art (LNet+ANet) and our approach.

	Baseline	LNet+ANet	Ours
CelebA	83%	87%	**89.8%**
LFWA	76%	84%	**85.9%**

Table 3: Layer wise average prediction accuracy over 40 attributes on the test set of CelebA and LFWA of our approach.

	C2	C3	C4	C5	C6	F1	F2
CelebA	89%	90%	90%	90%	90%	87%	88%
LFWA	83%	86%	86%	86%	85%	82%	81%

by doubling the perception size of our CNN the prediction accuracy for "Bag under eyes" increased from 82.2% to 83.5% with $C3$ as the best representation.

Similarly, attributes that are related to the face components with deterministic distribution, such as "arched eyebrow" always appears in the upper part of face images and "wearing necktie" appears in the lower part, can have condensed representations. We studied the condensed $C6$ feature ($3 \times 2 \times 256$) as the representation for locally related attributes (see complimentary material for details) and found that the average prediction accuracy remained almost the same. This means that we can further reduce the memory footprint of the representations and the linear classifiers for attributes defined by deterministic facial components.

3 Conclusions

In this paper, we have proposed to leverage the mid-level representations from deep convolutional neural network to tackle the attribute prediction problem for faces in the wild. We used an off-the-shelf architecture and a publicly available dataset to train a plain classification network and conduct investigations on the utility of the deep representations from various levels of the network. Although the trained network is not optimized either towards attribute prediction or recognition, it still allows accurate attribute prediction surpassing the state-of-the-art with a noticeable margin.

Our investigations indicate that CNNs trained for face classification have implicitly learned many semantic concepts and human describable attributes have been embedded in the deep representations, which can be separated by simple classifiers. They also reveal the potential utility of the intermediate deep representations for other face related tasks. These findings promise to increase the face attribute prediction performance and to achieve multiple intelligent functions, such as face recognition, retrieval and attribute prediction, in one single deep architecture.

References

[1] Hossein Azizpour, Ali Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–45, 2015.

[2] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1543–1550. IEEE, 2011.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[4] Gary B. Huang, Marwan Mattar, Tamara Berg, and Erik Learned-miller. E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2007.

[5] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[8] Yan Li, Ruiping Wang, Haomiao Liu, Huajie Jiang, Shiguang Shan, and Xilin Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3819–3827, 2015.

[9] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *Proceedings of the British Machine Vision*, 1(3):6, 2015.

[10] Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.

[11] Florian Schroff, Dmitry Kalenichenko, and James 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.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[13] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2013.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[15] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2013.

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

[17] Ning Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1637–1644, June 2014.

[18] Xiaogang Wang Ziwei Liu, Ping Luo and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.