2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Transferring from Face Recognition to Face Attribute Prediction through Adaptive Selection of Off-the-shelf CNN Representations

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

***Abstract*—This paper addresses the problem of transferring CNNs pre-trained for face recognition to a face attribute prediction task. To transfer an off-the-shelf CNN to a novel task, a typical solution is to fine-tune the network towards the novel task. As demonstrated in the state-of-the-art face attribute prediction approach, fine-tuning the high-level CNN hidden layer by using labeled attribute data leads to significant performance improvements. In this paper, however, we tackle the same problem but through a different approach. Rather than using an end-to-end network, we select face descriptors from off-the-shelf hierarchical CNN representations for recognizing different attributes. Through such an adaptive representation selection, even without any fine-tuning, our results still outperform the state-of-the-art face attribute prediction approach on the latest large-scale dataset for an error rate reduction of more than 20%. Moreover, by using intensive empirical probes, we have identified several key factors that are significant for achieving promising face attribute prediction performance. These results attempt to gain and update our understandings of the nature of CNN features and how they can be better applied to the transferred novel tasks.**

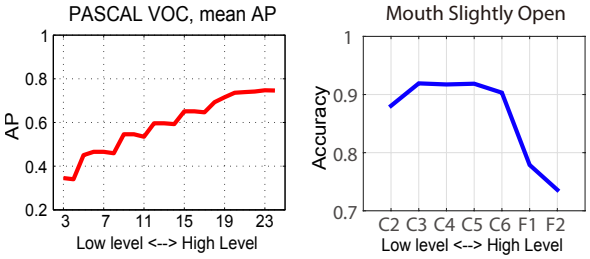

Fig. 1. The utility of off-the-shelf CNN representations depends on the classification tasks. As shown in the left figure [12], it is well-known that in many classification tasks off-the-shelf deep representations get incrementally more discriminative as going deeper in the network. Therefore, features from the high-level Fully Connected (FC) layers have been widely used for detection, classification, and recognition. But off-the-shelf representations might exhibit different behaviors. As exemplified in the right figure, to identify faces with slightly open mouth, mid-level representations from a pre-trained face recognition CNN outperform the high-level FC features. Left figure was adapted with authors' permission.

I. Introduction

Although people can tell face attributes effortlessly, predicting these human describable properties has been a very challenging task for computers [8], [9]. Recently, given that Convolutional Neural Networks (CNNs) have achieved great advancements in computer vision and pattern recognition, they have become a natural tool for predicting face attributes [3], [18], [20]. Training a CNN to predict face attributes, one intuitive way is to explicitly learn the visual-semantic correspondence in an end-to-end manner as in [6], [15] where CNNs were trained to describe image contents. But end-to-end training is hardly feasible when training instances are scarce. An alternative, but more practical, solution is to transfer a pre-trained CNN to the face attribute prediction task.

A common practice when transferring a CNN to a novel task is to fine-tune the off-the-shelf network towards the target problem with a moderate size of training data, and this procedure has been shown beneficial [1], [2]. The state-of-the-art face attribute prediction approach [20] employed cascaded CNNs for face detection and learning the best target representation consecutively, and fine-tuning on these cascaded components brought about significantly better performance. However, it seems that fine-tuning must be applied to both the face detection and the feature learning networks; fine-tuning only on the feature learning network did not significantly improve the accuracy compared to the previous work [18]. This reminds us of reconsidering the way of utilizing CNNs for attribute prediction and we have to think about the selection of the most suitable face representations to facilitate face attribute prediction.

To construct representations for face attribute prediction, the feature from the high-level hidden layer of CNNs has been an intuitive choice as witnessed in the recent attribute prediction solutions [10], [18], [20]. This might be because the discrimination power increases along a classification CNN and the high-level hidden layer features are the most discriminative [12]. But recall that face attributes are mostly explanatory, it is therefore natural to consider employing the intermediate CNN representations, where useful spatial information are better preserved, to form effective face representations for face attribute prediction.

In this paper, we approach the face attribute prediction problem from a different perspective. We aim to identify the off-the-shelf CNN representations best for attribute prediction. Rather than fine-tuning a pre-train network, we study how to use the off-the-shelf features from multiple levels of a

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

pre-trained CNN to predict face attributes. We obtained the following findings through our study:

- Off-the-shelf deep hierarchical representations provide diverse utilities for a novel face attribute prediction task; the widely used high-level feature is mostly not the best choice (see Figure 1). Through adaptive selection of hierarchical representations, we build a very efficient and accurate solution that outperforms the state-of-the-art with a big margin (Section III).
- The common practice that fine-tuning the high level hidden layers for novel related tasks may not be the only means to improve the performance; off-the-shelf representations from mid-level CNNs could be equally effective or better.
- We further analyze the driving power that enables the performance advantage by untangling the feature magnitude and spatial activation pattern. Our experiments show that spatial information is more important than magnitude of filter response (Section IV).
- We demonstrate that an accurate face classifier can be built merely on attributes, which indicates that one can construct a single deep network for both face recognition and efficient search (Section V).

We hope our work could provide useful knowledge for using off-the-shelf CNNs in face related tasks and also could attract more attentions to the intermediate CNN representations, which would be very useful in computer vision tasks as recently proposed by [17].

II. Constructing Off-the-shelf CNN Representations

Architecture of pre-trained CNNs. Two off-the-shelf CNN models were used in this work. In both models, convolutional filter stacks were followed by two Fully Connected (FC) layers (denoted by $F1$ and $F2$). The $F1$ layer was intended for studying the change of attribute prediction power between the convolutional layer representations and the high-level discriminatively trained feature $F2$. In the first model, we chose the filter architecture (with non-linearity and pooling) used by Google's FaceNet NN.1 [13] (shorten as "FaceNet" in below). The length of FC layers was set to 512 to prevent overfitting. The other model we used was the publicly available pre-trained VGG-Face model[1] [11] (no more training was applied to VGG-Face).

Dataset and training of the FaceNet. Around 350000 image instances of 10000 identities from WebFace dataset [16] was used for training the FaceNet. Faces were segmented and normalized to a size of 140×140 and random crops of 128×128 were fed in. Training instances were augmented with random mirroring, slight rotation and jittering.

The FaceNet was trained from scratch and Softmax loss function was used during training. We used dropout regularization between FCs and the dropout rate was set to 0.5 for all the FC layers. PReLU [4] rectification was attached to each convolution and FC layer.

The initial learning rate was set to 0.012. It is decreased by a factor of 10 when the validation set accuracy stopped increasing. In total, FaceNet was trained by 3 decreasing learning rates.

TABLE I
Dimensions of constructed CNN representations.

	Convolutional Feature						FC
	Map Size	Depth (Low → High)					Layers
FaceNet	4*4	192	384	384	384	256	512
VGG-Face	3*3	64	128	512	512	512	4096

Constructing off-the-shelf CNN representations. Hierarchical representations along the CNNs were constructed to study their effectiveness for attribute prediction. First, faces were aligned based on detected landmarks [7], and the aligned center crops with the corresponding horizontally flipped counterparts were fed to the trained CNNs. Then, spatial poolings were applied to the mean responses of the earlier convolutional layers to reduce the dimensionality and improve the invariance of intermediate representations (features from FC layers were not processed).

Specifically, for the FaceNet, we first applied average spatial poolings to the mean output from "Conv2" layer to "Conv6" layer (with window sizes of $8, 4, 2, 2, 2$ correspondingly). We then applied overlapping 3×3 spatial pooling (stride $= 2$) to generate our intermediate face representations. This ensured the same feature map size of the constructed convolutional representation to be 4×4. The yielded features from the pre-trained FaceNet were denoted by $C2$, $C3$, $C4$, $C5$, $C6$, $F1$ and $F2$, where C is short for $Conv$, F is short for FC, representations are named after the layer number specified by the architecture.

Similarly, for VGG-Face, we applied average pooling with size of 32, 16, 8, 4 to the mean convolutional filter responses (the filter output right before 2×2 max. pool in VGG-Face model, e.g., layer "conv1_2") . The resulting features are then max. pooled in the same way as in the FaceNet. Features named $C1$, $C2$, $C3$, $C4$, $C5$, $F1$ and $F2$ were finally constructed from VGG-Face model. Detailed dimensions of the CNN representations are summarized in Table I.

Evaluation and comparison. The face attribute prediction performance was evaluated on the recently released CelebA, LFWA and LFWA-extended datasets[2]. CelebA contains around more than 200000 images of around 10000 identities and LFWA has 13233 images of 5749 identities. Each image in CelebA and LFWA was annotated with 40 binary face attribute labels. Face images in LFWA-extended, the extended version of LFWA, were annotated by 73 binary attribute labels. Identical procedures for training and building attribute classifier as in [20] were used in this work, i.e., binary linear SVM classifiers were trained directly for each type of representation to classify

[1]Available at http://www.robots.ox.ac.uk/~vgg/software/vgg_face/.

[2]Project page: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, available from Oct 2015.

face attributes. The same training and test instances defined by both datasets were used in performance evaluations.

For fair comparisons, we selected the equivalent approach, denoted by "[17]+ ANet" in [20], as the **baseline method**. This is because the baseline method also utilized conventional face alignment as in our approach, but the employed CNN was built by fine-tuning a pre-trained face classification CNN with attribute labels. That is, the discrepancy between our approach and the baseline mostly lies in the way of constructing face representations for attribute prediction. The current state-of-the-art in [20], which was a fine-tuned end-to-end attribute prediction CNN, is denoted by "**LNet+ANet**" in this paper. By comparing our approach with the baseline and the state-of-the-art, one is able to quantitatively identify the power of our adaptive representation selection approach over the commonly adapted fine-tuning approach.

III. Leveraging Hierarchical CNN Representations for Attribute Prediction

A. Diverse Utility of Deep Representations

To investigate the prediction utility of the deep features, we initially train attribute classifiers for each type of the FaceNet representations on the training set of CelebA. The prediction accuracy on all the 40 attributes for each representation type is plotted in Figure 2.

It can be observed that features from the intermediate convolutional layers ($C3$ to $C6$) demonstrate an obvious advantage over the final FC features on average, especially for attributes describing motions of the mouth area where the gap is almost 20%. Feature $C6$ has the highest prediction accuracy on average, which is more than 2% higher than the last FC layer. One can also observe that the mid-level convolutional representations also featured slightly different prediction power on different attributes, e.g. $C2$ outperformed on "Rosy Cheeks" but not on hair related attributes. Thus, it is intuitive to consider adaptive selections of face representation to best facilitate accurate classification for each face attribute.

B. Adaptive Representation Selection Outperforms

Next, we used the identified best FaceNet representations for each attribute and compared their effectiveness with the baseline and the state-of-the-art. The best VGG-Face representations were also evaluated on the CelebA to validate that our finding was not specific to a certain network realization. The comparative results are shown in Figure 3. It is clear that the identified best representations from both the FaceNet and the VGG-Face model form an envelop above the baseline and state-of-the-art approach. They outperform on almost all the attributes. Compared to the baseline approach, using the best representations can reduce the error rate for nearly 40%. This validates the general effectiveness of using off-the-shelf attribute-specific representation; flexible representation selection brings more important performance improvements than fine-tuning.

TABLE II
Comprehensive performance comparisons.

	Baseline	LNet+ANet	Ours
CelebA	83%	87%	90%
LFWA	76%	84%	86%
LFWA-ext.	N.A.	83%	88%

In addition, we evaluated the effectiveness of our approach with the FaceNet representations on the LFWA and LFWA-extended datasets. The prediction accuracy on the 19 extended attributes (of LFWA-extended dataset), with the corresponding best representations, are compared to the state-of-the-art (reported in Table 2 in [20]) in Figure 4. It is easy to see that most of the attributes can be more effectively predicted by mid-level convolutional representations; only the identity-related (e.g., ethnic, gender) attributes are best represented by features from the high-level hidden layers. The same trend was observed on CelebA, which highlights the significance of feature representation selection when using an off-the-shelf CNN on a related novel task.

To summarize, we conducted extensive evaluations of the proposed adaptive representation selection approach on CelebA, LFWA and LFWA-extended datasets. The results are comprehensively compared in Table II. Apart from these results, we also noticed that, comparing with [19], the slight increment of the feature map size in this work brought about marginal performance improvement. But, larger feature map made training SVM classifiers costlier, which infers that a proper selection of feature map size is potentially important. Exploring the effectiveness of smaller feature map might be interesting for face attribute prediction.

IV. Identifying Key Factors

To experimentally identify the key factors that enabled the performance advantage, we analyze the significance of representation magnitude versus the spatial information in this section. Since the spatial information in CNN representations has been highlighted previously in [1], [14], we are more interested in examining the impact of magnitude.

Specifically, to remove the impact of magnitude, it is intuitive to binarize the representations by thresholding, i.e., set x=1 if x>0, otherwise x=0 (x stands for the filter response). To study the importance of spatial information, we naturally focus on $C6$ representation, because it can be seen as the best representative that maintain both high discriminative power and spatial information. To remove the impact of spatial information while retain the magnitude, we applied spatial 4×4 max. pooling which yielded 1×1 feature map (denoted by Down Sampled (DS) representations). The prediction performance of binarized, down-sampled and the original representations is shown in Figure 5 and compared in Table III.

From Figure 5, it is easy to observe a trend that binarizing the deep representations has gradually less impact on attribute prediction accuracy as going to higher level; the impact of binarizing the $F2$ features is negligible. It is very interesting

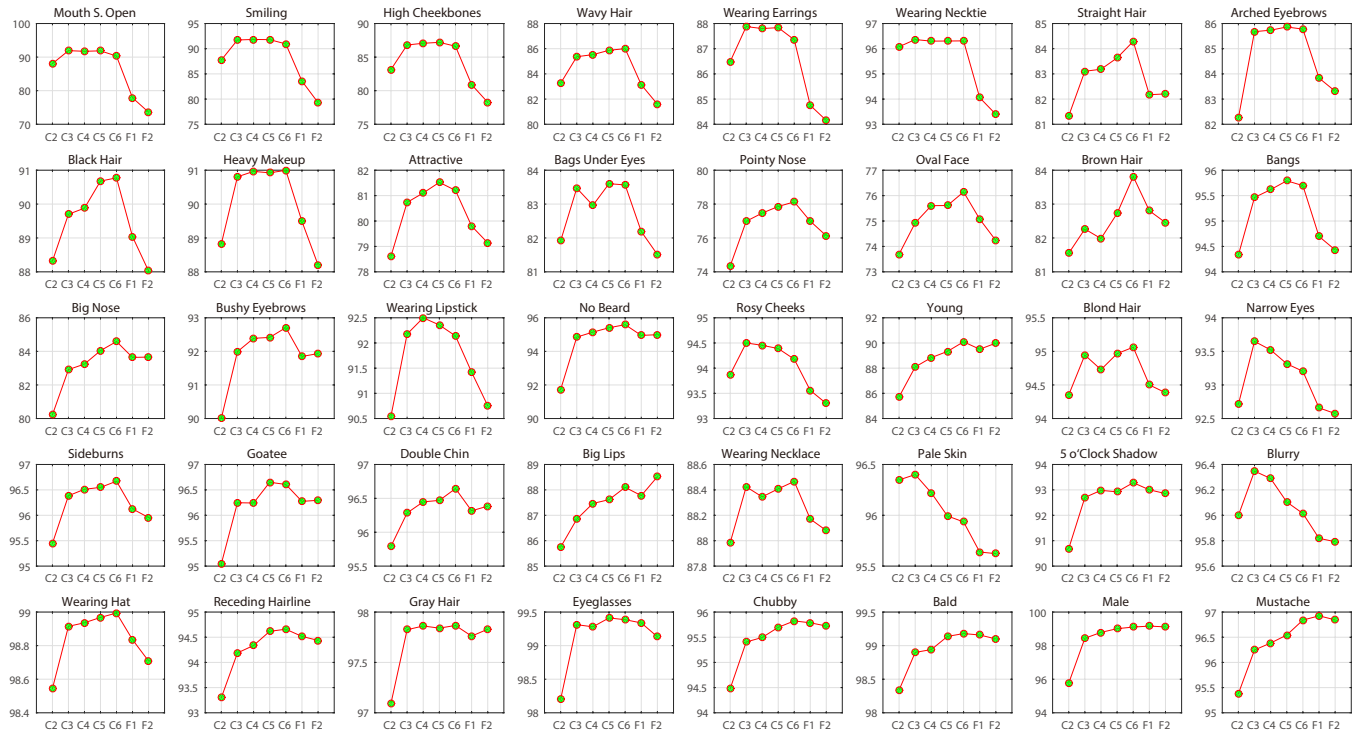

Fig. 2. Exploring the prediction accuracy of off-the-shelf CNN representations for 40 attributes on CelebA. In each grid, y-axis stands for the prediction accuracy in %. These deep representations demonstrate diverse prediction power on different attributes. It is therefore reasonable to perform adaptive representation selection for each attribute.

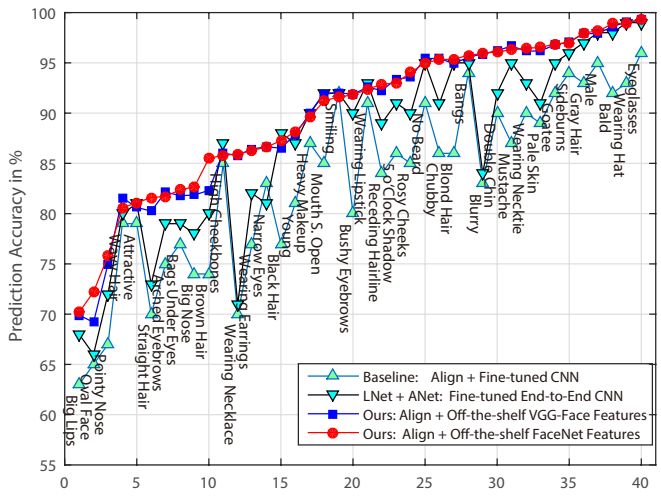

Fig. 3. Comparing adaptive representation selection with the baseline and the state-of-the-art on CelebA. The corresponding average prediction accuracy of the baseline is 83%, state-of-the-art is 87%, and ours are 90% (FaceNet), 90% (VGG-Face).

TABLE III
COMPARING THE EFFECT OF BINARIZATION AND DOWN-SAMPLING ON THE PREDICTION ACCURACY.

	C2	C3	C4	C5	C6	FC1	FC2
Original Rep.	88%	90%	90%	90%	90%	88%	87%
Bin. Rep.	80%	83%	84%	83%	87%	88%	87%
DS Conv6	-	-	-	-	82%	-	-

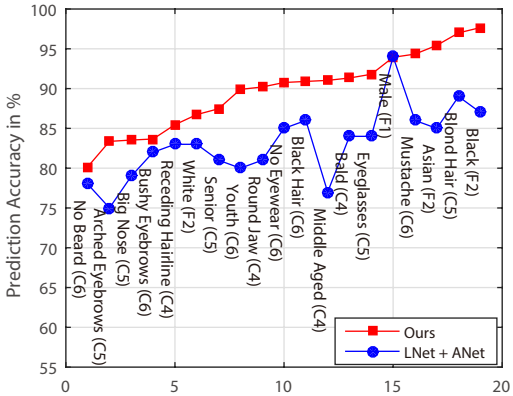

Fig. 4. Comparing the additional 19 attributes on LFWA-extended dataset. The name of the best representations are given in parenthesis after each attribute.

since it means that binarized high-level features can also be utilized for efficient and relatively accurate attribute-based face search. One can also construct face recognition, attribute prediction and image retrieval with a single pre-traind CNN.

To explain the gradually decreased binarization impact on prediction accuracy, we analyzed the representation magnitude and found that it gets gradually weaker from lower to higher level representations. In other words, a trend of incremental sparsity can be clearly observed in the deep representations along the network. Here we demonstrate the average of the filter responses of 10000 random instances along the FaceNet

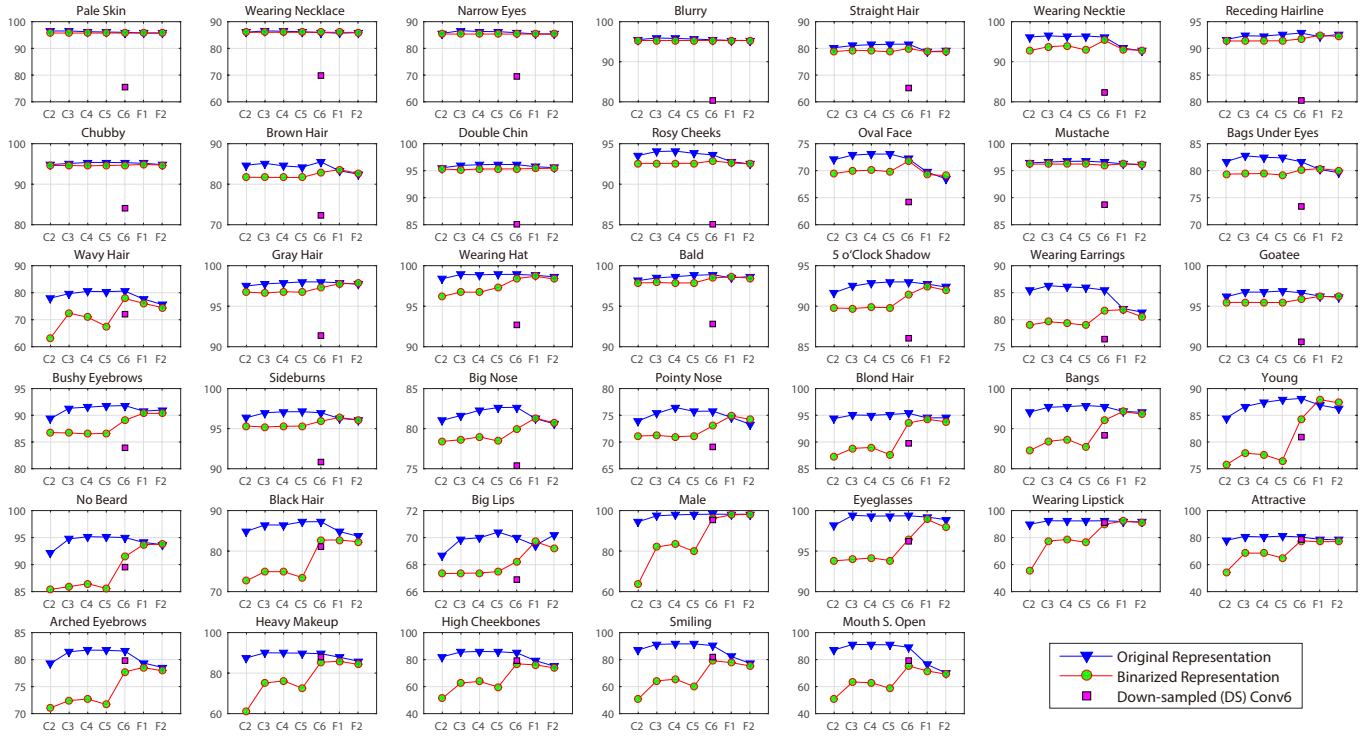

Fig. 5. Comparing the original FaceNet representations with binarized and down-sampled counterparts. The prediction accuracy of these representations is provided in Table III.

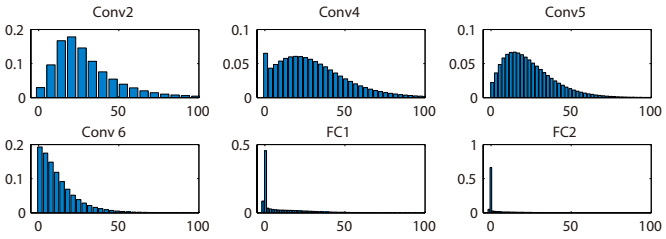

Fig. 6. Magnitude histograms of representation along FaceNet. Incremental sparsity can be observed from lower level representations to higher level ones.

on attribute "Mouth Slightly Open" in Figure 6.

On the other hand, from Figure 6 and Table III, it is hard to qualitatively summarize the relation between the performance variation, the magnitude and spatial information w.r.t the properties of attributes. But down-sampling deteriorate the prediction accuracy for 5.6% more than binarization on average. This implies that, the spatial information (i.e., activation pattern) is more important than the feature magnitude for using off-the-shelf deep representations in attribute prediction.

V. Face Verification Using Attributes

Given that the human describable attributes are much fewer than the dimension of identity preserving feature, using face attributes for efficient screening in large-scale face recognition becomes a natural approach. Driven by the earlier success of face attribute classifier based on low-level features [9], here we investigate how well the attributes predicted by our approach perform in face verification. We evaluated face verification performance on the LFW [5] with the attributes annotated on LFWA.

In this work, we took 46 deep attribute predictors constructed using the training set of LFWA (from Section III-B) and trained a linear classifier to verify each matching face pair[3]. Let a vector of predicted attribute scores of a face image I_j be $\{a_i^{(j)}\}_1^N$ (i.e., $N = 46$), we construct a binary linear SVM classifier C, where

$$id(I_j, I_k) = C(< S(I_j, I_k)), D(I_j, I_k), P(I_j, I_k) >), \quad (1)$$

to verify if face j and k are of the same person, where $id(\bullet) = 1$ if I_j, I_k are of the same identity, otherwise $id(\bullet) = 0$. Pairwise distance was formed using the attribute-wise sum (S),

[3]The 46 selected attributes: 'Male' 'Asian' 'White' 'Black' 'Black Hair' 'Blond Hair' 'Brown Hair' 'Bald' 'Eyeglasses' 'Mustache' 'Chubby' 'Curly Hair' 'Wavy Hair' 'Straight Hair' 'Receding Hairline' 'Bangs' 'Sideburns' 'Fully Visible Forehead' 'Partially Visible Forehead' 'Obstructed Forehead' 'Bushy Eyebrows' 'Arched Eyebrows' 'Narrow Eyes' 'Big Nose' 'Pointy Nose' 'Big Lips' 'No Beard' 'Goatee' 'Round Jaw' 'Double Chin' 'Oval Face' 'Square Face' 'Round Face' 'Attractive Man' 'Attractive Woman' 'Indian' 'Gray Hair' 'Bags Under Eyes' 'Heavy Makeup' 'Rosy Cheeks' 'Shiny Skin' 'Pale Skin' '5 o Clock Shadow' 'Strong Nose-Mouth Lines' 'High Cheekbones' 'Brown Eyes'.

difference (D), and product (P), where

$$S(I_j, I_k) = \{abs(a_i^{(j)} + a_i^{(k)})\}_1^N, \quad (2)$$

$$D(I_j, I_k) = \{abs(a_i^{(j)} - a_i^{(k)})\}_1^N, \quad (3)$$

$$P(I_j, I_k) = \{a_i^{(j)} \cdot a_i^{(k)}\}_1^N. \quad (4)$$

The verification accuracy of our approach is 89.5% on the LFW dataset. Compared to [9], where 65 attributes were used, our approach achieved an error reduction rate of 30% by using only 46 attributes with a much simpler classifier. Such a performance advantage is still attributed to the significantly improved attribute prediction, which was driven by our adaptive representation selection strategy.

Considering the computational bottle neck between the last convolutional layer and the following FC layer, one would naturally utilize the earlier convolutional representations for face verification. We selected $C5$ as an example and found the verification accuracy was 89.2% — a very insignificant performance drop compared to using the best representations across the network.

VI. Conclusions

In this work, we address the face attribute prediction problem by transferring pre-trained face classification CNNs. Typically, transferring off-the-shelf CNNs to novel tasks requires fine-tuning. But through our empirical studies, we reveal the diverse utilities of the off-the-shelf deep representations for the novel face attribute prediction task. Therefore, we adaptively select the best off-the-shelf representations for each attribute to facilitate face attribute prediction. Intensive experiments show that our approach outperforms the state-of-the-art on the latest large-scale datasets with a big margin. By jointly using fine-tuning and adaptive representation selection, one can expect even greater performance improvements.

Moreover, we identify several key factors of the off-the-shelf representations that enabled the performance improvements. We hope our work could provide useful understanding for transferring off-the-shelf CNNs and be inspiring for exciting new ideas.

References

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *Computer Vision–ECCV 2014*, pages 329–344. Springer, 2014. 1, 3

[2] H. Azizpour, A. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Factors of transferability for a generic convnet representation (2015 v.3). 2014. 1

[3] 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

[4] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015. 2

[5] G. B. Huang, M. Mattar, T. Berg, and E. Learned-miller. E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2007. 5

[6] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014. 1

[7] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1867–1874. IEEE, 2014. 2

[8] N. Kumar, P. N. Belhumeur, and S. K. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008. 1

[9] 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, 5, 6

[10] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015. 1

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *Proceedings of the British Machine Vision*, 1(3):6, 2015. 2

[12] A. S. Razavian, H. Azizpour, J. Sullivan, and S. 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. 1

[13] 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. 2

[14] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900, 2015. 3

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014. 1

[16] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2

[17] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*, 2016. 2

[18] N. 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. 1

[19] Y. Zhong, J. Sullivan, and H. Li. Leveraging mid-level deep representations for predicting face attributes in the wild. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3239–3243, Sept 2016. 3

[20] X. W. Ziwei Liu, Ping Luo and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3