JMLR: Workshop and Conference Proceedings 80:1–13, 2017 ACML 2017

IMCNN: An Identity-aware Multi-task Deep Convolutional Neural Nerwork for Face Attribute Classification

Author Name1 ABC@SAMPLE.COM
Address 1

Author Name2 XYZ@SAMPLE.COM
Address 2

Editors: Editor Name

Abstract

In this paper, we study the face attribute learning problem with a new perspective of considering identity information and attribute relationships simultaneously. In particular, we first introduce an Identity-aware Multi-task deep Convolutional Neural Network (IMCNN), with low-level layers shared with all attributes and high-level layers shared within attribute groups. Meanwhile, the selected feature maps are merged for an auxiliary face recognition task. IMCNN is able to learn spatial attribute relationships via grouping and inner-person attribute consistence through face recognition. Consequently, better global attribute relationships are modeled. The experimental results on CelebA and LFWA demonstrate the promise of the proposed method.

Keywords: Face Attribute Learning, Multi-task Learning, Deep Learning

1. Introduction

Face attribute learning has attracted a great attention in many real-world applications such as face identification, verification and retrieval. It aims at learning mid-level representations as the bridge between the low-level features and the high-level labels. Face attributes usually include *gender, race, age, hair, nose, eyes*, etc. Since such attributes provide much more informative descriptions for objects and people, they have been applied in many tasks which require detailed descriptions. For example, given low quality imagery in surveillance video, the common strategy for identity verification is to describe suspects in terms of attributes to speed up the search process. However, large-scale face attribute learning problem is still very challenging as the faces captured in the real-world are usually influenced by some factors such as illumination, pose and expression.

Motivated by the success of convolutional neural network (CNN) in image classification Krizhevsky et al. (2012); Szegedy et al. (2015); Ren et al. (2015); Girshick (2015), the discriminative CNN representations have been widely used for face attribute learning. L-Nets Liu et al. (2015) are pre-trained on CelebFace Sun et al. (2014) dataset to get a better face location for attribute prediction. In Zhong et al. (2016), a face recognition network is trained and facial features are extracted from this network to train SVMs for attribute classification. However, these identity-based methods benefit from the deep representations by CNNs, but completely ignore relationships among all the attributes and consider all the attributes independent. This is not appropriate as attribute relationships provide additional

information to attribute learning and help improve the accuracy of attribute classification. For example, in gender recognition a person wearing lipstick and earrings has a high probability of being classified as a woman. Further, identity information is usually available in public dataset and real application scenarios. For the same person, some attributes are always the same such as the face shape, the mouth size and gender. An example is shown in Fig 3. We can see that the attribute Attractive is consistent for this person.

On the other hand, recent work, for example Hand and Chellappa (2016), considers exploiting the attribute correlations to boost the performance of attribute classification. It introduces a multi-task deep CNN (MCNN) by sharing the lower layers of the deep architectures for all the attributes and sharing the higher layers for similar attributes. MCNN assumes that many attributes are strongly related and splits all the 40 attributes into several attribute groups, where high-level layers are shared within a group and irrelevant to the rest. Although MCNN sets up the state-of-the-art in face attribute learning, it neglects attribute relationships related to the identity information, since attributes are highly correlated within the same person but less correlated among different identities.

In this paper we investigate the face attribute learning problem with a new perspective of considering identity information and attribute relationships simultaneously. We hypothesize that combining identity information and attribute relationship modeling together enables us to develop more accurate attribute learning algorithms. Our hypothesis is based on the following insights: (1) most attributes show consistency for the same identity while attributes from different persons are independent; (2) by performing face recognition on attribute features, these inner-person attribute relationships can be modeled, since identity information forces attribute features to be similar for the same person.

We present an Identity-aware Multi-task deep Convolutional Neural network (IMCNN), a novel deep learning framework, which performs identity recognition and attribute classification in a single framework. IMCNN exploits all the attributes and identity information by sharing lower layers of IMCNN for all attributes, sharing higher layers of IMCNN for strongly correlated attributes, and performing identity recognition in the merged higher layer.

In particular, all the attributes are split into several attribute groups according to spatial information and the selected feature maps of the attribute groups are merged to perform face recognition. The grouping scheme encourages IMCNN modeling the spatial attribute relationships, while the loss function of face recognition can be considered as a regularization term to model inner-person attribute consistency, since the merged features of the same person are forced to be close in the feature space, which leads to similar attribute predictions for the same person. Consequently, combining spatial attribute relationships and inner-person attribute consistency, IMCNN is capable of modeling better global attribute relationships. Extensive evaluations are conducted to investigate the performances of IMCNN. The experimental results on CelebA Liu et al. (2015) as well as LFWA Huang et al. (2007) validate the superior performance of it.

2. Related Work

Attribute classification often extracts features from landmarks or facial parts Berg and Belhumeur (2013); Bourdev et al. (2011); Kumar et al. (2009); Zhang et al. (2014). Hand-picked

facial regions are used for learning AdaBoost-based features for each attribute in Kumar et al. (2009), which are fed into independent SVMs for a final prediction. In Berg and Belhumeur (2013), HOG features are extracted from facial parts located by landmarks, where a single SVM is trained for each attribute to perform attribute classification. Bourdev et al. (2011) collects body parts with different poses and HOG-like features are learned from these poselets; then, SVMs are trained as classifiers.

Thanks to the great breakthrough of deep learning Krizhevsky et al. (2012); Szegedy et al. (2015); Simonyan and Zisserman (2014); Girshick (2015); Ren et al. (2015); Liu et al. (2015); Zhong et al. (2016); Günther et al. (2016), high-level features can be extracted from the whole image without landmarks or parts information. Zhong et al. (2016) extracts features at different layers of a face recognition CNN. These features are used as training data for linear SVM to generate attribute scores. Liu et al. (2015) localizes face by training two cascade LNets and then extracting the 40 features for the 40 attributes via an ANet. Finally, 40 SVMs are trained for the 40 attributes.

Though most deep-learning methods do not rely on landmarks or parts, they consider attributes to be independent. Using independent SVMs as classifiers, they are not able to model attribute relationships, which are crucial for attribute prediction. In addition, all the previous methods do not make full use of the available identity information, which has been proven closely related with attribute learning.

3. Method

3.1. IMCNN Architecture

Multi-task learning Seltzer and Droppo (2013); Luo et al. (2013); Huang et al. (2014); Maurer et al. (2013); Romera-Paredes et al. (2013) aims at learning multiple tasks simultaneously. It assumes that related tasks should be correlated via a certain structure. This is especially true for face attribute learning with deep architectures as there is a hierarchy of attributes, which should be related through shared low-level features. Meanwhile, the high-level features are still task-specific to retain the differently discriminative information for different attribute tasks.

In this section, we investigate the attribute learning problem with the multi-task learning scenario. In particular, we propose to share low-level features for all tasks and to share high-level features for similar tasks. For CNNa, it is natural to bifurcate them so that all the layers before the bifurcation are shared by all the attributes and the layers after the bifurcation are shared within strongly related attributes. Besides, since the identity information of samples usually exists in the training data, we develop face recognition tasks from the same data as an auxiliary task, forcing the attribute learning and identity recognition to benefit from each other. Consequently, we propose IMCNN to improve the generalization performance of multi-task attribute learning.

Fig.1 illustrates the architecture of IMCNN. Low-level layers from Conv1 to Pool3 are shared by all the 40 attributes. After that, group strategy is used to separate layers. There are 6 groups according to locations after the first split: Nose, Mouth, Eyes, Face, Gender and Rest; then, the Rest group is further separated into 4 smaller groups in the second split: AroundHead, FacialHair, Cheeks and Fat. Further, we merge the 6 attribute groups by concatenating the feature maps, which are the inputs of our face recognition

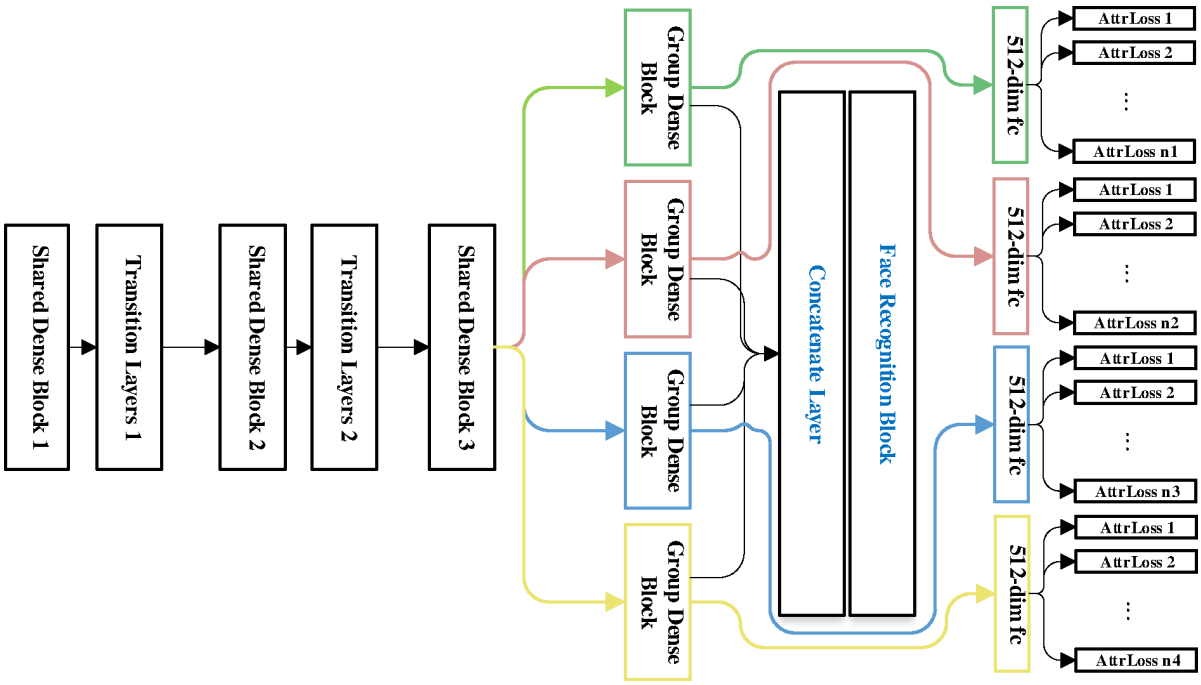

Figure 1: Architecture of IMCNN.

task. This split-and-merge structure efficiently combines the two tasks by using rich spatial information from attribute classification for face recognition. In addition, attribute relationships are explicitly learned by separating attributes based on their similarities into groups. Grouping configuration follows Hand and Chellappa (2016), as shown bellow and we list hyper-parameters of IMCNN in Table 2.

Upper Group: *Arched Eyebrows, Bags Under Eyes, Bushy Eyebrows, Narrow Eyes, Eyeglasses, Black Hair, Blond Hair, Brown Hair, Gray Hair, Balding, Receding Hairline, Bangs, Wearing Hat.*

Middle Group: *Big Nose, Pointy Nose, Wearing Earrings, Sideburns, High Cheekbones, Rosy Cheeks.*

Lower Group: *Big Lips, Lipsticks, Mouth Slightly Open, Wearing Necklace, Wearing Necktie, Mustache, No Beard, Goatee, Double Chin.*

Whole Image Group: *Male, Smiling, Attractive, Blurry, Oval Face, Pale Skin, Young, Heavy Makeup, Straight Hair, Wavy Hair, 5 o'Clock Shadow, Chubby.*

3.2. Joint Optimization of Attributes and Identities

Multiple loss functions are employed when there are multiple tasks to be learned in a neutal network Redmon and Farhadi (2016); Liu et al. (2016); Long et al. (2015). Therefore, we jointly optimize attribute classification and face recognition via combining two loss

Layers	Output Size	IMCNN
Stem	112×112	7×7 conv, s=2
	56×56	3×3 max pool, s=2
Shared Dense Block(1)	56×56	$\begin{matrix}1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv}\end{matrix} \times 4$
Transition Layers(1)	56×56	1×1 conv
	28×28	2×2 ave pool, s=2
Shared Dense Block(2)	28×28	$\begin{matrix}1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv}\end{matrix} \times 8$
Transition Layers(2)	28×28	1×1 conv
	14×14	2×2 ave pool, s=2
Shared Dense Block(3)	14×14	$\begin{matrix}1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv}\end{matrix} \times 16$
Group Dense Block	14×14	$\begin{matrix}1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv}\end{matrix} \times 4$
Face Recognition Block	14×14	maxout
	14×14	$\begin{matrix}1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv}\end{matrix} \times 4$
	512-dim	fc
	1-dim	softmax loss

Table 1: Aechitecture of IMCNN. Batch normalization Ioffe and Szegedy (2015) and ReLU Nair and Hinton (2010) are adopted after each convolutional layer.

functions, namely, $Loss_{ID}$ and $Loss_{ATTR}$. $Loss_{ATTR}$ is the sum of the losses of all the attributes. For the i-th attribute, we formulate its loss as follows:

$$loss_i = \frac{1}{N}\sum_{n=1}^{N}\{[p_n log[\sigma(f_{g(i)})] + (1-p_n) log[1-\sigma(f_{g(i)})]]\},$$

where N indicates batch size, p_n is a binary label, which is assigned as 1 when a face has the corresponding attribute or 0 otherwise, σ is sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, and $f(j)$ represents the output of the FC2Attr layer of the j-th group. The $g(i)$ function returns the group index for each attribute index. For example, since the attribute Male belongs to the group Gender, we have $g(1) = 2$ when the attribute index of Male is 1 and the group index of Gender is 2. Note that the value range of $g(i)$ varies from 1 to 9, since we have 9 different FC2Attr layers, which significantly reduces the model complexity while learning the task-specific parameters for all the attributes. Then, we compute $Loss_{ATTR}$ by summing over $loss_i$ as follows:

$$Loss_{ATTR} = \sum_{i=1}^{T} loss_i,$$

where T is the number of the attributes.

Layer	Kernel Size	Output Size
Conv1	$3 \times 3 \times 64$	$112 \times 112 \times 64$
Pool1	2×2	$56 \times 56 \times 64$
Conv2a	$3 \times 3 \times 64$	$56 \times 56 \times 64$
Conv2	$3 \times 3 \times 192$	$56 \times 56 \times 192$
Pool2	2×2	$28 \times 28 \times 192$
Conv3a	$3 \times 3 \times 192$	$28 \times 28 \times 192$
Conv3	$3 \times 3 \times 384$	$28 \times 28 \times 384$
Pool3	2×2	$14 \times 14 \times 384$
ConvAttr	$3 \times 3 \times 512$	$7 \times 7 \times 512$
FC1Attr	1024	$1 \times 1 \times 1024$
FC2Attr	1024	$1 \times 1 \times 1024$

Table 2: Aechitecture of IMCNN. Batch normalization Ioffe and Szegedy (2015) and ReLU Nair and Hinton (2010) are adopted after each convolutional layer.

By merging features from different attribute groups, we can compute $Loss_{ID}$ on the merged features:

$$Loss_{ID} = -log(\hat{p}_k), \hat{p}_k = \frac{h(F)_k}{\sum_{n=1}^{N} h(F)_n}, k \in [1, N],$$

where F is the merged features. In IMCNN, 6 ConvAttr output feature maps are concatenated to get F; $h(\cdot)$ represents a fully connected mapping between F and an N-dimensional vector. $\hat{p_k}$ is the predicted probability of the k-th identity and is computed from $h(F)$ in a softmax function. Finally, $LOSS$ of our IMCNN network is the weighted sum of $Loss_{ID}$ and $Loss_{ATTR}$ as follows:

$$LOSS = Loss_{ATTR} + \lambda \times Loss_{ID}, \tag{1}$$

where $Loss_{ID}$ acts like a regularization term, which improves the generalization power of IMCNN and makes it more robust. As shown in Sec. 3, the choice of λ is essential for the performance of IMCNN.

Under the regularization of $Loss_{ID}$, the strength of the joint optimization in Eq.(1) lies on two aspects. First, the identity information and attributes are complementary in the process of learning face representations. Attributes usually respond to local facial parts, while the identity information provides a global constraint via $Loss_{ID}$. Therefore, as shown in Fig 2 IMCNN is able to learn fine-grained feature maps. Second, regularized by $Loss_{ID}$, IMCNN can model better attribute relationships by exploiting inner-person attribute correlations. In particular, attributes are consistent for the same person. For example, as illustrated in Fig. 2, we make attribute correlation analysis on a collection of 35 images within a person and among different identities respectively. It indicates that most images of the same person have similar attribute labels, while attributes for different

identities are relatively independent. $Loss_{ATTR}$ fails to model this inner-person attribute correlations as it treats every image independently, while $Loss_{ID}$ forces the merged feature maps F to be as similar as possible for the same person via performing face recognition on F. Since F consists of feature maps from all the attribute groups, predictions for all the attributes are likely to be similar for the same person.

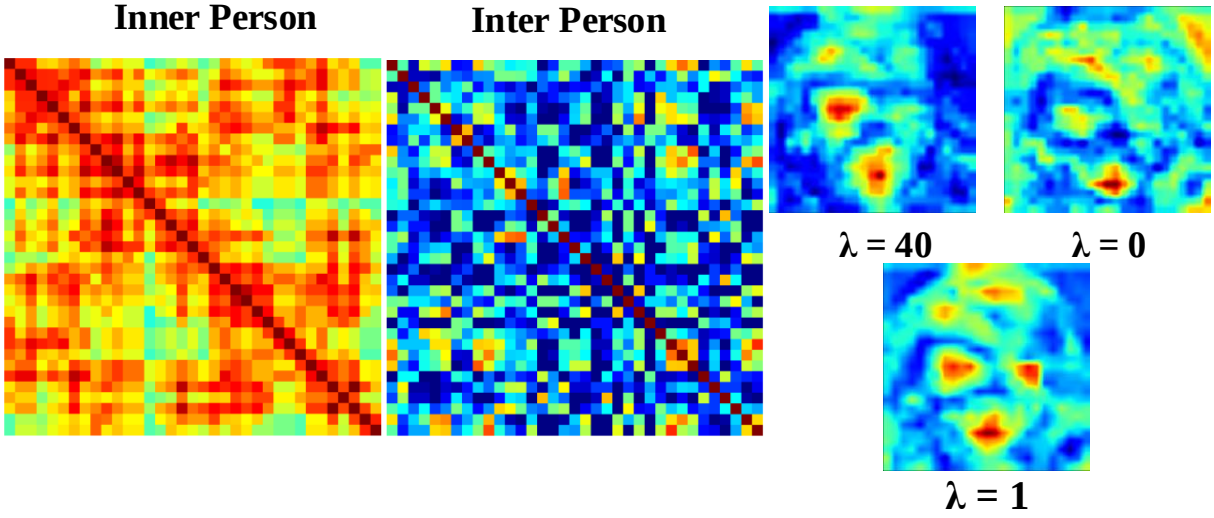

Figure 2: Correlation Analysis and feature maps of IMCNN with different λ values.

4. Experiment

4.1. Experimental Configurations

A large face dataset with identity and attribute labels is needed for training IMCNN. One option is LFW, consisting of 13,323 images from 5,749 identities. However, since many identities only have 1 image in LFW, it is not suitable for jointly training with attribute classification and face recognition tasks. While CelebA Liu et al. (2015) has 202,599 images of 10,000 persons, each person has 20 images on average. Consequently, we consider CelebA as our training set for IMCNN and its identity information is public available.

We split the CelebA into two parts, one for training and another for test. The training set contains 180,000 images of 8,000 identities while the test set contains 20,000 images from the rest identities. Note that the training set doesn't share identities with the test set. Therefore, IMCNN needs to model a general relationship between identity information and attributes, so that it can perform well on new identities.

We train all the models in the Caffe Jia et al. (2014) framework on single GTX1080ti graphic card. For our IMCNN, we use a batch size of 64 and a initial learning rate of 0.001 with the SGD strategy. We further decrease the learning rate two times after every 10 epoches. All the filters of convolutional layers and fully connected layers are initialized with the xavier method. During the test process, we resize the images to 112×112 and feed them into the models. Different from the training process, which requires identity information,

the test process only requires a face image. We report accuracies on both the test set of CelebA and the whole LFWA dataset.

4.2. Baselines

To verify the effectiveness of IMCNN, we compare them with the state-of-the-art methods: LNets+ANet Liu et al. (2015), face attribute prediction with classification CNN (referred as ID Net) Zhong et al. (2016) and multi-task convolutional neural network (referred as MCNN) Hand and Chellappa (2016). Further, we also study different configurations of IMCNN. Specifically, we train an ATTR Net, which is optimized with $Loss_{ATTR}$ only. IMCNNs with different λ values are also trained to explore the influence of the identity information. For methods with public available results including LNets+ANet, ID Net and MCNN, we report the public results. For ATTR Net, we train it with the same configurations as the IMCNN for fair comparisons.

Sensitivity of λ: To verify the effectiveness of the auxiliary face recognition task, we train IMCNN with different λ values. Note that we refer IMCNN with a zero λ to ATTR Net. As illustrated in Fig 4, accuracy under different λ values varies and achieves the best when λ equals 1. Interestingly, accuracy drops dramatically when λ increases or decreases. Further, a properly chosen λ value encourages IMCNN to reach significantly lower training and validation losses. As we can see in Fig 4, the validation loss drops from 15 to 14 when λ increases from 0 to 1. And it drops from 16 to 14 when λ decreases from 40 to 1. The change of accuracy and loss can be explained when $Loss_{ID}$ is interpreted as a regularization term to the attribute classification task. Since $Loss_{ID}$ forces feature maps of the same person to be as close as possible, a proper λ helps IMCNN to force features of the same person to be close. As a result, a higher performance is obtained as the inner-person consistency is modeled. However, when λ is too large, IMCNN is not able to model attribute diversity for the same person, so that the performance is harmed. On the other hand, when λ is too small, the regularization influence of the identity information is limited.

Comparison with other methods: According to our experimental results in Table 3, IMCNN achieves an average accuracy of 92.16%, surpassing all the other methods by a large margin. On the one hand, IMCNN outperforms identity-based methods including LNets+ANet and ID Net, especially on FacialHair and AroundHead groups. We attribute this to their pre-training on face recognition, which forces them putting more weights to identity-related facial parts while decreasing its attention on facial hair as well as around head areas. On the contrary, face recognition is an auxiliary task in IMCNN. By choosing proper λ, $Loss_{ID}$ does not harm identity-free attributes but helps model inner-person attribute consistency. On the other hand, IMCNN beats MCNN and ATTR Net on every single attribute, which implies that identity is informative for all the 40 attributes. We attribute this to the supervision of the identity information. In Fig 3, IMCNN takes advantage of the identity information while MCNN fails to make the right predictions. To better understand this phenomenon, we compare feature maps from Conv3 of IMCNN with ATTR Net in Fig 2. We find that ATTR Net puts more weights on facial hair and around head areas, since nearly half of the 40 attributes lie in these areas. But these areas are large and easily polluted by noise and occlusion. Nevertheless, IMCNN learns more fine-grained feature maps putting more weights on facial parts while leaving small regions of

	LNets+ANet	ID Net	MCNN	ATTR Net	IMCNN	LNets+ANet	ID Net	MCNN	ATTR Net	IMCNN
5 Shadow	91	92.64	94.41	94.87	**95.43**	84	85.32	77.70	77.89	**85.11**
Arch. Eyebrows	79	82.03	83.55	83.33	**84.99**	82	84.58	82.36	82.11	**83.19**
Attractive	81	79.99	82.94	82.74	**83.71**	83	82.16	80.42	80.26	**84.52**
Bags Un. Eyes	79	81.43	84.98	85.63	**86.21**	83	87.31	83.51	84.00	**86.37**
Bald	98	97.16	98.87	98.82	**99.19**	88	86.99	91.99	91.85	**92.17**
Bangs	95	94.64	96.04	96.02	**96.71**	88	88.13	89.99	89.99	**91.03**
Big Lips	68	68.87	71.20	70.98	**72.77**	75	75.68	79.21	78.33	**82.37**
Big Nose	78	82.39	84.50	84.63	**85.66**	81	83.67	84.67	84.99	**86.11**
Black Hair	88	84.14	89.87	89.73	**90.81**	90	88.56	92.35	92.16	**92.52**
Blond Hair	95	94.22	95.97	95.76	**96.55**	97	95.89	97.45	97.10	**98.02**
Blurry	84	95.75	96.08	96.12	**96.86**	74	83.67	85.30	85.38	**86.83**
Brown Hair	80	83.54	88.99	88.85	**90.07**	77	79.47	80.94	80.99	**81.55**
Bushy Eyebrows	90	92.20	92.80	92.63	**93.50**	82	83.68	85.11	85.13	**85.32**
Chubby	91	92.89	95.66	95.68	**96.47**	73	75.84	76.90	76.96	**77.83**
Double Chin	92	94.36	96.41	96.42	**97.22**	78	80.09	81.17	81.17	**96.33**
Eyeglasses	99	99.25	99.63	99.64	**99.80**	95	94.87	91.22	91.20	**96.33**
Goatee	95	95.77	97.30	97.29	**97.74**	78	79.36	82.52	82.54	**83.77**
Gray Hair	97	97.28	98.20	98.16	**98.47**	84	83.69	89.04	89.00	**90.63**
Heavy Makeup	90	88.58	91.34	91.41	**92.47**	95	93.11	95.84	96.00	**96.13**
H. Cheekbones	87	84.21	87.55	87.49	**88.43**	87	84.52	88.25	88.15	**88.38**
Male	98	97.32	98.16	98.18	**98.69**	94	93.66	93.66	93.74	**94.73**
Mouth S. O.	92	92.03	93.74	93.70	**94.79**	82	81.87	83.47	83.44	**84.25**
Mustache	95	95.11	96.93	96.84	**97.45**	92	92.36	93.53	93.48	**94.02**
Narrow Eyes	81	86.04	87.16	87.24	**88.25**	81	81.59	82.37	82.49	**83.17**
No Beard	95	95.69	96.11	96.25	**97.70**	79	80.21	82.13	82.25	**82.68**
Oval Face	66	70.15	75.81	75.12	**76.90**	74	75.39	77.38	76.98	**77.62**
Pale Skin	91	94.91	97.01	96.90	**97.76**	84	87.22	93.41	93.35	**94.52**
Receed. Hairline	89	92.37	93.81	93.76	**94.74**	85	85.39	86.26	86.24	**87.12**
Rosy Cheeks	90	93.28	95.13	95.11	**95.87**	85	85.39	86.26	86.24	**87.12**
Pointy Nose	72	74.62	77.47	77.32	**78.67**	78	83.52	87.52	87.46	**87.15**
Sideburns	96	95.58	97.82	97.84	**97.14**	77	76.33	82.73	82.79	**84.07**
Smiling	92	91.57	92.66	92.92	**93.73**	91	90.17	91.75	91.99	**92.07**
Straight Hair	73	78.38	83.39	83.24	**85.09**	76	77.32	78.72	78.64	**79.35**
Wavy Hair	80	75.88	83.92	83.49	**85.38**	76	75.39	81.96	81.48	**83.01**
Wear. Earings	82	85.73	90.32	90.70	**91.76**	94	93.02	94.71	95.00	**95.08**
Wear. Hat	99	98.87	99.04	99.01	**99.33**	88	89.64	90.20	90.11	**90.80**
Wear. Lipstick	93	90.45	93.95	94.21	**94.76**	95	92.07	94.89	95.12	**95.24**
Wear. Necklace	71	85.01	86.82	87.03	**87.91**	88	88.59	89.66	90.01	**90.11**
Wear. Necktie	93	91.13	96.53	96.55	**97.33**	79	78.82	80.50	80.56	**81.36**
Young	87	84.56	88.48	88.70	**89.65**	86	84.26	85.37	85.67	**86.11**
Average	87	88.75	91.25	91.26	**92.16**	84	84.64	86.27	86.25	**87.38**

Table 3: Performance comparison. Left columns show the performance on CelebA. Right ones give that of LFWA.

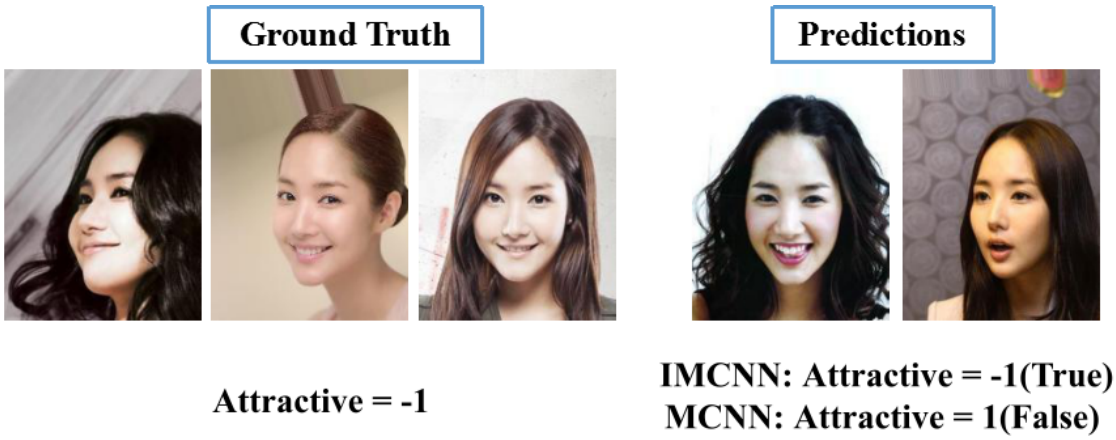

Figure 3: An example illustrates the inner-person attribute consistency. IMCNN takes advantage of the identity information while MCNN fails to make the right predictions.

responses on identity-free areas such as hair and forehead. As a result, IMCNN directly enhances the performance on identity-related groups including Nose Group and Eyes Group. In the meantime, it indirectly boosts the performance of identity-free groups by decreasing the possibility of being influenced by noise, illumination and other factors. Therefore, the accuracies on both identity-related attributes and identity-free attributes are improved significantly.

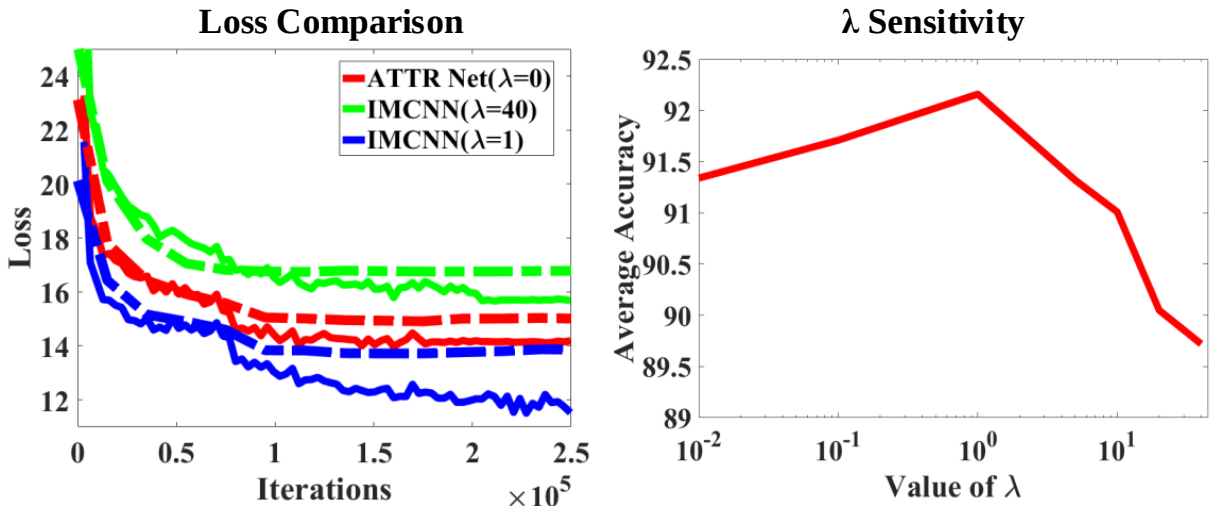

Figure 4: Loss comparison and accuracy under different λ values. The dotted lines represent the validation loss.

5. Conclusion

In this paper, we investigate the face attribute learning problem from a new perspective of considering the identity information and attribute relationships simultaneously. In particular, we introduce an Identity-aware Multi-task deep Convolutional Neural Network (IM-CNN), with low-level layers shared by all attributes and high-level layers shared within attribute groups. Meanwhile, the selected feature maps are merged for an auxiliary face recognition task, to model the inner-person consistency. The experimental results on CelebA and LFWA demonstrate the promise of the proposed methods.

References

Thomas Berg and Peter 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.

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.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

Manuel Günther, Andras Rozsa, and Terrance E. Boult. AFFACT - alignment free facial attribute classification technique. *CoRR*, abs/1611.06158, 2016.

Emily M Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network for attribute classification. *arXiv preprint arXiv:1604.07360*, 2016.

Gary B Huang, Manu Ramesh, Tamara Berg, and Erick Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Technical Report 07-49, University of Massachusetts, Amherst*, 2007.

Wenhao Huang, Guojie Song, Haikun Hong, and Kunqing Xie. Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2191–2201, 2014.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

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.

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.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. pages 3730–3738, 2015.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and Stephen J Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Image Processing*, 22(2):523–536, 2013.

Andreas Maurer, Massi Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 343–351, 2013.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *International Conference on Machine Learning*, pages 1444–1452, 2013.

Michael L Seltzer and Jasha Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6965–6969. IEEE, 2013.

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

Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, pages 1988–1996, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir 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.

Yang Zhong, Josephine Sullivan, and Haibo Li. Face attribute prediction using off-the-shelf cnn features. pages 1–7, 2016.