Proceedings of Machine Learning Research 95:879-892, 2018 ACML 2018

Relative Attribute Learning with Deep Attentive Cross-image Representation

Zeshang Zhang ZESHANG@ZJU.EDU.CN
Yingming Li * YINGMING@ZJU.EDU.CN
Zhongfei Zhang ZHONGFEI@ZJU.EDU.CN
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

In this paper, we study the relative attribute learning problem, which refers to comparing the strengths of a specific attribute between image pairs, with a new perspective of cross-image representation learning. In particular, we introduce a deep attentive cross-image representation learning (DACRL) model, which first extracts single-image representation with one shared subnetwork, and then learns attentive cross-image representation through considering the channel-wise attention of concatenated single-image feature maps. Taking a pair of images as input, DACRL outputs a posterior probability indicating whether the first image in the pair has a stronger presence of attribute than the second image. The whole network is jointly optimized via a unified end-to-end deep learning scheme. Extensive experiments on several benchmark datasets demonstrate the effectiveness of our approach against the state-of-the-art methods.

1. Introduction

Visual attribute learning has attracted much attention in many real-world applications such as image searching Chen et al. (2013); Wang et al. (2013); Huang et al. (2014); Zhang et al. (2013), face verification Kumar et al. (2009), object recognition Wang and Mori (2010); Branson et al. (2010), video retrieval and recommendation Chen et al. (2014a); Cui et al. (2014), and zero-shot learning Lampert et al. (2014); Li et al. (2014); Han et al. (2014). It aims to learn mid-level semantic properties as the abstraction between the low-level features and the high-level labels. In general, visual attribute learning is considered as binary concept learning which indicates the presence or absence of certain semantic property.

Further, Parikh and Grauman (2011) introduced relative attribute learning which extends the traditional binary attribute learning through comparing the relative strengths of particular attributes. Given a set of manually labeled relative orderings of image pairs, relative attribute learning considers to learn a global ranking function for each attribute so that the strengths of each attribute between two images can be compared. Figure 1 shows some examples of relative attribute learning. Different from the binary attributes, relative attributes bear more semantic information and have been exploited in many applications Kovashka et al. (2012); Biswas and Parikh (2013); Shrivastava et al. (2012); O'Donovan et al. (2014).

* Corresponding author

Figure 1: Relative attribute learning. Given the training image pairs for a specific attribute, the goal is to compare a pair of novel images with respect to the same attribute.

For relative attribute learning, early efforts focus on the learning-to-rank framework such as RankSVM, which learns a ranking function for each attribute so that the ranking scores of the two samples can be compared to determine the relative strengths for specific attributes. Inspired by the success of convolutional neural networks Krizhevsky et al. (2012), recent works propose to apply the CNN architecture to the task of relative attribute learning. In particular, single-image representations are first obtained using deep CNN approaches, and then a ranking measure is utilized to predict the relative strengths of the two images with respect to specific attributes. For example, Souri et al. (2016) introduced a RankNet framework which trained a deep ranking network in an end-to-end fashion. Further, Yang et al. (2016) also proposed to jointly learn visual features and a nonlinear ranking function in a unified framework and presented a deep relative attribute (DRA) algorithm. While RankNet and DRA significantly outperform the prior shallow models in relative attribute prediction due to the incorporation of deep representation learning, they mainly focus on single-image representation learning and ignore the learning of cross-image representation which has the ability of capturing deep relationship between the two images.

In this paper, we investigate the relative attribute learning problem with a new perspective of considering cross-image representation. In particular, we present the deep attentive cross-image representation learning (DACRL) model, an end-to-end convolutional neural network which takes a pair of images as input, and outputs a posterior probability that indicates the relative strengths of a specific attribute, based on cross-image representation learning. DACRL first learns the respective discriminative representations for each image with one shared subnetwork and then employs an attentive cross-image convolution module to adaptively learn the non-linear cross-image representation, which helps capture the correspondence among the semantic properties of the two images. Further, a posterior probability with respect to the specific attribute is predicted based on the learned cross-image representation. The above processes are jointly optimized via a unified end-to-end deep learning scheme.

The proposed framework is evaluated on six real-world datasets. Extensive experiments on these benchmark datasets demonstrate the effectiveness of our approach against the state-of-the-art methods. The main contributions of this paper are as follows:

(1) We deal with the problem of relative attribute learning from a new perspective of considering cross-image representation learning; (2) The proposed framework introduces the attentive cross-image convolutional strategy to further enhance the learning of cross-image representation.

2. Related Works

Relative attributes. Since the introduction of relative attributes Parikh and Grauman (2011), relative attribute learning has attracted the attention of many researchers for its variety of applications, such as image retrieval Kovashka et al. (2012), zero-shot learning Parikh and Grauman (2011); Yang et al. (2016); Chen et al. (2014b); Biswas and Parikh (2013), and font selection O'Donovan et al. (2014).

In Parikh and Grauman (2011), relative attributes are first proposed and the original approach adopted the learning-to-rank framework. A linear ranking function for each attribute is trained based on the hand-crafted features (GIST and Color). Extended from this, Li et al. (2012) trained non-linear functions for attribute prediction. To capture the correlations among multiple attributes, multi-task learning (MTL) is introduced in Chen et al. (2014b). More recently, Wang et al. (2016b) fused pointwise and pairwise labels to capture the relations between class labels, tags, and attributes. Yu and Grauman (2014) proposed a learning-to-rank framework for fine-grained visual comparisons. In another work Yu and Grauman (2015), Yu and Grauman developed a Bayesian local learning strategy to infer when images are hardly distinguishable for a given attribute. All the above methods are based on the hand-crafted features. The success of deep learning has motivated end-to-end frameworks for learning features and attribute ranking simultaneously Souri et al. (2016); Yang et al. (2016); Singh and Lee (2016); He et al. (2016b). To overcome the sparsity of supervision for visual comparisons, Yu and Grauman (2017) proposed to augment real training image pairs with synthetic images. However, all the methods simply focus on the single-image representation learning and ignore the correlation between the cross-image representation and relative attributes.

Cross-image representation Cross-image representation learning has been explored in some pairwise comparison tasks, such as person re-identification. Ahmed et al. (2015) proposed to capture the cross-image representation for person re-identification by computing the local difference between the features of the two input images. Chen et al. (2016) simply stitched a pair of input images to learn the joint feature representation through CNN. Extended from these efforts, Wang et al. (2016a) jointly learned single-image representation and cross-image representation based on CNN for person re-identification. Mao et al. (2018) employed multi-rate atrous convolution layers to match the cross-image semantic components. As far as we know, no existing works capture the cross-image representation for relative attribute learning. Besides, channel-wise attention is incorporated to enhance the learning of cross-image representation.

3. Our Approach

As shown in Figure 2, our proposed model consists of three parts, the single-image feature learning part, the attentive cross-image representation block, and the prediction part. In

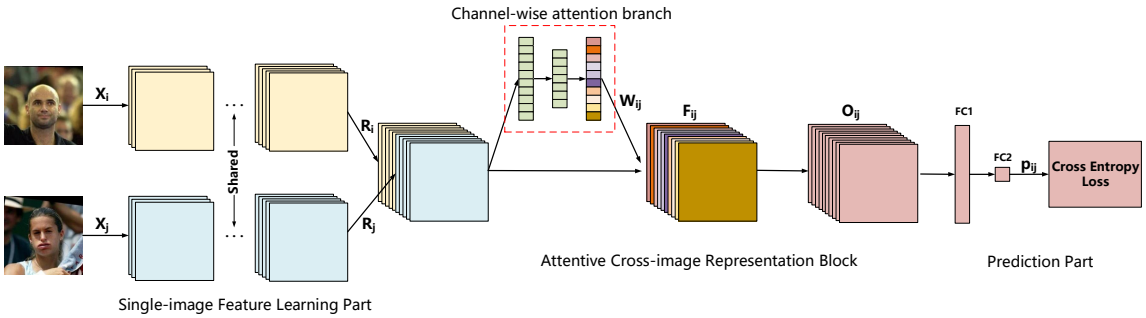

Figure 2: The overall architecture of the proposed deep attentive cross-image representation learning model. The network consists of three parts, the single-image feature learning part, the attentive cross-image representation block, and the prediction part. Pairs of images are fed into the network with their corresponding target relative strengths for a specific attribute, which is transformed into probability scores. Cross-entropy loss is computed and back-propagated through the network to update the weights.

this section, we first introduce the problem description of the relative attribute learning, and then present the details of the model.

3.1. Problem Formulation

The goal of the relative attribute learning is to learn a function which predicts the relative strengths of paired images in any attribute. Existing methods are based on the learning-to-rank framework to learn a ranking function for a specific attribute. Given a set of ordered image pairs $O_m = \{(x_i, x_j)\}$ and a set of un-ordered image pairs $U_m = \{(x_i, x_j)\}$ for any attribute a_m, image x_i has a stronger presence of attribute a_m than image x_j if image pair $(x_i, x_j) \in O_m$, and image x_i and image x_j have similar presence of attribute a_m if $(x_i, x_j) \in U_m$. Let $f_m(x)$ denote the ranking function corresponding to a specific attribute a_m. With these notations, the existing relative attribute learning can be formulated as learning $f_m(x)$ that satisfies the following constraints:

$$(x_i, x_j) \in O_m \quad f_m(x_i) > f_m(x_j), \quad (x_i, x_j) \in U_m \quad f_m(x_i) = f_m(x_j) \tag{1}$$

Different from the existing methods, our approach explores to make full use of cross-image representation to perform relative attribute learning. Taking an image pair as input, we first learn single-image features for each image with one shared subnetwork. Consequently, we perform cross-image representation learning based on the learned features of the two images. Further, prediction functions are learned with the cross-representations to output the probability that indicates the relative relationship between the two images. Let $f_m(x_i, x_j)$ denote the prediction function of a specific attribute a_m. The relative attribute

learning is formulated as learning $f_m(x_i, x_j)$ which satisfies the following constraints:

$$\begin{aligned} (x_i, x_j) \in O_m \quad & f_m(x_i, x_j) = 1.0, \ f_m(x_j, x_i) = 0.0 \\ (x_i, x_j) \in U_m \quad & f_m(x_i, x_j) = 0.5, \ f_m(x_j, x_i) = 0.5 \end{aligned} \tag{2}$$

3.2. Singe-image Feature Learning

Inspired by the success of deep learning, we explore a deep convolutional neural network to learn and extract single-image representation. In particular, VGG-16 Simonyan and Zisserman (2014) architecture with all fully connected layers removed is employed to capture the semantic feature maps for each image. As the architecture shown in figure 2, our approach takes a pair of images as input which is processed with the same deep network and performs the feature learning part separately. Let $f_{CNN}(X, \theta_1)$ denote the single-image feature learning part, which takes X as input and θ_1 as parameters. We formulate the learning process as follows:

$$\{R_i, R_j\} = \{f_{CNN}(X_i, \theta_1), f_{CNN}(X_j, \theta_1)\} \tag{3}$$

where R_i and R_j denote the learned representations of images X_i and X_j respectively, and θ_1 denotes the shared parameters. Consequently, the pair of feature maps (R_i, R_j) would be fed into the attentive cross-image representation block.

3.3. Attentive Cross-image Representation Learning

With the learned singe-image representation, we introduce the particular cross-image representation block for relative attribute learning. While traditional cross-image representation learning has been exploited in other vision tasks, the existing efforts consider the single-image feature maps from the two images equally and directly perform joint convolution on them, which ignores some intrinsic characteristics of those feature maps. For example, different feature maps may contribute differently to a specific attribute. The features which attend to the face may have less relevance with the attribute 'Dark-Hair' than the features that focus on the hair intuitively.

Based on this assumption, we propose an attentive cross-image representation block which performs channel-wise attention and assigns different weights to different feature maps accordingly. As shown in Figure 2, the attention W_{ij} is generated by the channel-wise attention branch, which consists of a global average layer and two fully connected layers. Afterwards, the channel-wise multiplication is performed to get attentive feature maps F_{ij}.

With these features F_{ij}, the attentive cross-image representation block would perform cross-image convolution to fuse the features and learn the cross-image representation. Given a pair of generated single-image features (R_i, R_j), the block concatenates the feature maps along the channel axis, generates the channel-wise attention W_{ij}, and performs channel-wise multiplication between $[R_i, R_j]$ and W_{ij}. Afterwards, cross-image representation would be further learned with convolutional operations. The attentive cross-image representation is formulated as:

$$O_{ij} = f_M(F_{ij}) = f_M([R_i, R_j] \otimes W_{ij}) = f_M([R_i, R_j] \otimes f_W([R_i, R_j])) \tag{4}$$

where O_{ij} denotes the output of the block, $f_M(\cdot)$ is the operation that captures the cross-image representation implemented as a convolutional network, $f_W(\cdot)$ denotes the branch that generates channel-wise attention, and $\otimes$ is the channel-wise multiplication.

3.4. Prediction Learning

With the learned cross-image representation, we perform relative attribute learning with a prediction block. The prediction part consists of two fully connected layers and outputs p_{ij} which indicates the probability of that image X_i exhibits more of the attribute than image X_j. p_{ij} is expected to be larger than 0.5, if image X_i exhibits more of the attribute than image X_j. Similarly, if image X_i exhibits less of the attribute than image X_j, p_{ij} is expected to be smaller than 0.5, and if it is desired that the two images have the same strength, p_{ij} is expected to be 0.5.

To learn the parameters of our network, we optimize it by minimizing the standard cross-entropy loss:

$$L_{ij} = -t_{ij}\log(p_{ij}) - (1 - t_{ij})\log(1 - p_{ij}) \tag{5}$$

where t_{ij} is the target probability of the image pair (X_i, X_j). Because of the nature of the datasets, t_{ij} is chosen from set $\{0, 0.5, 1\}$, according to the available labels in the dataset.

4. Experiments

To validate our method, we quantitatively compare it with several state-of-the-art methods on all publicly available benchmark datasets for relative attributes to our knowledge. At the same time, we analyze the capability and effectiveness of our method through the qualitative results.

4.1. Datasets

UT-Zap50K Yu and Grauman (2014) dataset is a collection of 50025 images with annotations for relative comparison of 4 attributes collected from Zappos.com. This dataset contains two collections: **Zappos50K-1**, in which relative attributes are annotated for coarse pairs, where the comparisons are relatively easy to interpret, and **Zappos50K-2**, where the relative attributes are annotated for fine-grained pairs, for which making the distinction between them is hard. Zappos50K-1 contains approximately 1500-1800 training image pairs for each attribute. They are divided into 10 train/test splits which are provided alongside the dataset and are used in this work. Zappos50K-2 contains a test set of approximately 4300 pairs, while the training set is the combination of the training and testing sets of Zappos50K-1.

Zappos50K-lexi Yu and Grauman (2017) is an augmented collection of images with the crowd-mined lexicon for 10 additional attributes based on the UT-Zap50K. It contains approximately 1300-2100 image pairs for each attribute.

LFW-10 Sandeep et al. (2014) dataset is a subset of the *Labeled faces in the wild* (LFW) and contains 2000 images of faces: 1000 for training and 1000 for testing. Annotations for 10 attributes are available. For each attribute, a subset of 500 image pairs have been annotated for each training and testing set. We use the same train-test splits provided in Sandeep et al. (2014).

PubFig Parikh and Grauman (2011) dataset is a subset of the *Public Figure Face Dataset* (PubFig) and contains 800 facial images (GIST+Color features) from 8 random identities. 11 attributes are available. The same train-test splits are used as in Parikh and Grauman (2011); Singh and Lee (2016).

OSR Parikh and Grauman (2011) dataset contains 2688 images (GIST features) of outdoor scenes in 8 categories, for which 6 relative attributes are annotated. We use the same train-test splits as in Parikh and Grauman (2011); Singh and Lee (2016). The ordering of samples in both PubFig and OSR datasets are annotated in a category level; all images in a specific category are annotated higher, more equal, and lower than all images in another category, with respect to an attribute.

4.2. Implementation Details

We train a separate network for each attribute. We implement our proposed model based on the Lasagne Dieleman et al. (2015) deep learning framework. In all our experiments, we use the VGG-16 model of Simonyan and Zisserman (2014) and trim out all the fully connected layers (all the convolutional layers are used) as the single-image feature learning part. The weights of the model are initialized with a pre-trained model on ILSVRC 2014 Russakovsky et al. (2015) dataset for the task of image classification. While the network learns to predict the relative attributes, these weights are fine-tuned. The remaining layers are newly introduced. The dimensions of the fully connected layers in the prediction part are set to 1024 and 1. All the weights of the attentive cross-image representation block and prediction part are initialized using the weights sampled from the Gaussian distribution with a standard deviation of 0.01, and the biases are initialized to 0.01. To prevent the loss from diverging, we clip the probability p_{ij} to be in the range $[10^{-7}, 1-10^{-7}]$.

For training, we use stochastic gradient descent with RMSProp updates with a mini-batch size of 48 (24 pairs of image). The learning rate of the single-image feature learning part is set to 5e-5 and the learning rate of all the other layers is set to 5e-4 initially. For the sparsity of the supervision, we adopt the learning rate decay and weight decay to prevent overfitting. In particular, to determine the step of the learning rate decay, we randomly sample 1/5 of the training set from the Zappos50K-2 as the validation set. A fixed 5e-4 multiplier is used for the weight decay.

According to our model, (R_i, R_j) and (R_j, R_i) are different image pairs. In order to balance the dataset, we invert the order of images in a pair to get a new pair with an opposite label. For example, we can get a new pair (R_j, R_i) labeled $1 - t_{ij}$ from the pair (R_i, R_j) with label t_{ij}. After this operation, the composition of the dataset is balanced with the number of the training image pairs doubled. For Zappos50K and LFW-10 datasets, we train with 16 epochs and divide the learning rate by 10 at epoch 5 and epoch 12. Due to the large number of the sample pairs, we train with 2 epochs for PubFig and OSR datasets.

4.3. Baseline

We use the RankNet model proposed in Souri et al. (2016) as our baseline, which is based on single-image representation learning with the VGG-16 network pre-trained on ILSVRC 2014 as well. With this baseline, we can evaluate the effectiveness of the attentive cross-image representation for relative attribute learning. In addition, we also include other

Table 1: Results for the Zaps50K-lexi dataset

Method	Comfort	Casual	Simple	Sporty	Colorful	Durable	Supportive	Bold	Sleek	Open	Mean
RankSVM	84.03	86.11	86.89	87.27	83.84	85.15	87.75	83.71	86.06	84.41	85.52
DeepSTN	84.95	87.04	89.46	88.79	94.30	83.29	85.75	87.42	85.82	84.68	87.15
Ours-without-attention	**90.26**	91.05	**90.17**	**93.01**	**95.23**	**90.70**	91.76	91.12	**88.51**	**87.90**	**90.97**
DACRL(ours)	89.56	**91.98**	89.23	92.71	94.29	**90.70**	**91.99**	**91.32**	**88.51**	87.37	90.77
RankNet (224)	90.48	90.43	**90.40**	93.31	95.43	90.47	91.98	91.53	86.31	82.53	90.29
Ours-without-attention(224)	**91.88**	**94.44**	89.93	93.01	**97.33**	**92.65**	**92.65**	91.12	89.24	87.90	92.02
DACRL(ours)(224)	**91.88**	91.36	90.16	**94.22**	95.81	92.33	**92.65**	**92.56**	**90.71**	**88.98**	**92.07**

Table 2: Results for the PubFig dataset

Method	Male	White	Young	Smiling	Chubby	Forehead	Eyebrow	Eye	Nose	Lip	Face	Mean
FG-LP	91.77	87.43	91.87	87.00	87.37	94.00	89.83	91.40	89.07	90.43	86.70	89.72
RankNet	95.50	94.60	94.33	95.36	92.32	97.28	94.53	93.19	94.24	93.62	94.76	94.42
DRA	90.82	87.12	91.49	92.68	89.30	94.39	90.19	90.60	91.03	90.35	91.99	90.91
Local Global	92.39	90.75	91.10	90.24	93.00	93.00	91.78	87.62	88.38	92.84	93.22	91.30
Ours-without-attention	**97.70**	**97.82**	97.10	97.03	97.05	**98.30**	97.36	**97.99**	97.26	94.36	**98.04**	97.27
DACRL(ours)	96.49	97.80	**97.96**	**97.42**	**97.22**	98.05	**97.48**	96.91	**97.74**	**96.83**	96.27	**97.29**

representative results from recent efforts, such as DeepSTN Singh and Lee (2016), FG-LP Yu and Grauman (2014), DRA Yang et al. (2016), Local Global He et al. (2016b), Spatial Extent Xiao and Jae Lee (2015), and RankSVM Parikh and Grauman (2011).

4.4. Quantitative Results

Following Parikh and Grauman (2011); Yu and Grauman (2014); Sandeep et al. (2014); Souri et al. (2016); Yang et al. (2016); He et al. (2016b); Singh and Lee (2016); Yu and Grauman (2017), we report the accuracy in terms of the percentage of correctly ordered pairs. In addition, we have added the results of the proposed model without channel-wise attention (referred as Ours-without-attention), which only relies on the cross-image convolution, to further verify the effectiveness of cross-image representation for relative attribute learning.

We first report the results on the Zappos50K-lexi Yu and Grauman (2017) dataset in Table 1. This is an newly collected dataset with crowd-mined lexicon for 10 attributes. For an fair comparison, our methods are trained and tested on the same 64×64 images as the methods in Yu and Grauman (2017). As shown in Table 1, our methods achieve state-of-the-art results on all the attributes. To fully illustrate the capacity of our model, we train our model on 224×224 images and compare them with that in Souri et al. (2016). Our models surpass it by a considerable margin on most of the attributes. Finally we report the results on the Zappos50K-1 and Zappos50K-2 datasets (Table 4). Our methods show an excellent capacity on most attributes as well.

Table 3: Results for the OSR dataset

Method	Natural	Open	Perspective	Large Size	Diag	ClsDepth	Mean
FG-LP	95.70	94.10	90.43	91.10	92.43	90.47	92.37
RankNet	99.40	97.44	96.88	96.79	98.43	97.65	97.77
DRA	99.47	97.81	97.19	96.88	98.46	97.24	97.84
DeepSTN	98.89	97.20	96.31	95.98	97.64	96.10	97.02
Local Global	98.91	96.32	94.20	94.93	97.01	92.29	95.61
Ours-without-attention	99.43	97.95	**97.73**	97.07	**98.61**	98.13	98.15
DACRL(ours)	**99.77**	**98.54**	97.56	**97.56**	98.48	**98.62**	**98.42**

Table 4: Results for the Zappos dataset

Dataset	Zappos50K1					Zappos50K2				
Method	Open	Pointy	Sporty	Comfort	Mean	Open	Pointy	Sporty	Comfort	Mean
RankSVM	87.77	89.37	91.20	89.93	89.57	60.18	59.56	62.70	64.04	61.62
Spatial Extent	95.03	94.80	96.37	95.60	95.45	-	-	-	-	-
FG-LP	90.67	90.83	92.67	92.37	91.64	74.91	63.74	64.54	62.51	66.43
RankNet	95.37	94.43	97.30	95.57	95.67	73.45	68.20	73.07	70.31	71.26
DeepSTN	94.87	94.93	97.47	95.87	95.79	-	-	-	-	-
Local Global	95.50	**95.98**	97.56	96.00	**96.26**	74.10	69.99	71.92	**71.34**	71.84
Ours-without-attention	**96.63**	95.07	**97.70**	**96.37**	**96.44**	75.45	69.80	73.78	68.54	71.89
DACRL(ours)	96.07	95.03	**97.70**	96.23	**96.26**	**75.66**	**70.65**	**73.87**	69.56	**72.44**

Table 5: Results for the LFW-10 dataset

Method	Bald	DkHair	Eyes	GdLook	Mascu	Mouth	Smile	Teeth	FrHead	Young	Mean
Spatial Extent	83.21	88.13	82.71	72.76	93.68	88.26	86.16	86.46	90.23	75.05	84.67
FG-LP	67.90	73.60	49.60	64.70	70.10	53.40	59.70	53.50	65.60	66.20	62.43
RankNet	81.14	88.92	74.44	70.28	98.08	85.46	82.49	82.77	81.90	76.33	82.18
DeepSTN	83.94	**92.58**	90.23	71.21	96.55	**91.28**	84.75	**89.85**	87.89	**80.81**	**86.91**
Local Global	83.09	90.01	**93.14**	**75.70**	97.93	89.12	**89.50**	85.89	86.11	75.58	86.61
Ours-without-attention	83.21	91.99	87.97	69.97	97.70	89.93	85.03	88.00	89.45	74.84	85.81
DACRL(ours)	**85.04**	**92.58**	90.23	70.28	**98.28**	**91.28**	85.03	89.23	**90.63**	76.55	**86.91**

Table 2 and Table 3 show our results on the PubFig and OSR datasets respectively. Our proposed framework outperforms the state-of-the-art methods on all the attributes of the dataset by a considerable margin. Our methods show outstanding results both on the low level attributes (Nose, White,) and generic, high level attributes (Young, Perspective,). Taking the capacity of the model into account, our models surpass not only the shallower model Yang et al. (2016) which is based on the AlexNet Krizhevsky et al. (2012) but also the deeper model He et al. (2016b) which is based on ResNet-34 He et al. (2016a).

Table 5 shows our results on the LFW-10 Sandeep et al. (2014) dataset. Our models perform competitively with respect to the state-of-the-art methods. Different from the other datasets, the images in this dataset typically contain a large range of background while the face area typically only occupies a small portion of an image. The background part is considered as noise for most of the attributes in this dataset that are highly local, such as eyes and smile. Therefore, the methods that outperform us for certain attributes pay much attention to the local region of the images and incorporate local context and global information.

4.5. Ablation Study

We study the contribution that cross-image representation versus the channel-wise attention has to the relative attribute prediction performance. For this, we compare three models: (1) the baseline model (**RankNet Souri et al. (2016)**) which is based on the VGG-16 architecture with the last fully connected layer replaced with one unit linear layer; (2) the proposed model (shown in Figure 2) with the channel-wise attention branch (the part in the red dotted box) removed (**Ours-without-attention**); and (3) the proposed model which combines cross-image representation and channel-wise attention (**DACRL(ours)**).

According to the quantitative results of the three models on the datasets, our models (with and without attention) can outperform the baseline model by a large margin over most of the attributes. It demonstrates that cross-image representation contributes to the relative

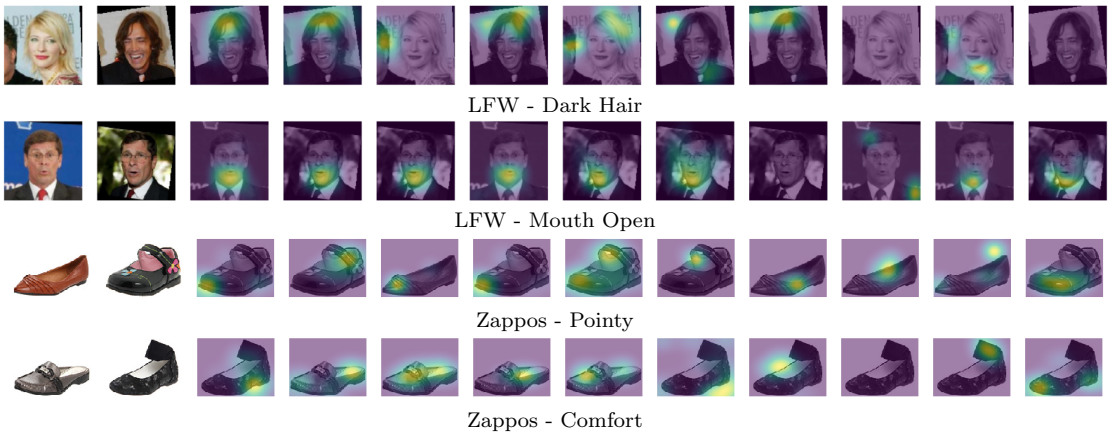

Figure 3: Visualized channel-wise attention results for LFW-10 and Zappos dataset. The first two columns are the original images. The third to seventh columns are the feature maps assigned largest weights and eighth to twelfth columns are the ones with smallest weights.

attribute learning. Furthermore, our combined model (**DACRL(ours)**) outperforms the model without attention (**Ours-without-attention**) by a considerable margin on most attributes, especially on dataset LFW-10. To reveal the effectiveness of the channel-wise attention, we visualize some results in Figure 3. For attribute "Mouth open", the feature maps which focus on the mouth region get more attention, and features that activate more at the tiptoe are assigned more weights for attribute "Pointy". We find that the channel-wise attention attends to the features relevant to the specific attribute adaptively. Therefore, the combined model produces the best accuracy for most of the attributes, which shows the channel-wise attention and cross-image representation well cooperate complementarily in relative attribute learning.

4.6. Saliency Maps

According to Simonyan et al. (2013), we compute the derivative of the final output with respect to the input images. These saliency maps visualize the pixels in the input images which contribute most to the prediction result. We visualize the saliency maps of 4 datasets obtained from our model in Figure 4. For most attributes, our method correctly discovers the relevant image regions: hair for dark hair, plants and road for natural, regions around eyes for eyebrow, and toe end for pointy. We find that our model can localize easily not only local attributes such as "Dark Hair" but also abstract attributes such as "Natural".

5. Conclusion

In this paper, we present the deep attentive cross-image representation learning (DACRL) model for relative attribute learning, which first extracts single-image representation with one shared subnetwork, and then learns attentive cross-image representation through

Figure 4: Saliency maps generated by our model. For each dataset, 4 test images and their overlaid saliency maps are shown (the warmer the color of the overlay image, the more salient that pixel is). As shown, our model pays attention to the relevant regions: hair for dark hair, plants and road for natural, regions around eyes for eyebrow, and toe end for pointy.

considering the channel-wise attention of concatenated single-image feature maps. Taking a pair of images as input, DACRL outputs a posterior probability indicating the relative strengths of a specific attribute. Extensive experiments on several benchmark datasets demonstrate the effectiveness of our approach against the state-of-the-art methods.

Acknowledgments

This work was supported by NSFC (No. 61702448, 61672456) and the Fundamental Research Funds for the Central Universities (No. 2017QNA5008, 2017FZA5007). We thank all reviewers for their valuable comments.

References

Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

Arijit Biswas and Devi Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 644–651. IEEE, 2013.

Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision*, pages 438–451. Springer, 2010.

Bor-Chun Chen, Yan-Ying Chen, Yin-Hsi Kuo, and Winston H Hsu. Scalable face image retrieval using attribute-enhanced sparse codewords. *IEEE Trans. Multimedia*, 15(5): 1163–1173, 2013.

Lin Chen, Peng Zhang, and Baoxin Li. Instructive video retrieval based on hybrid ranking and attribute learning: A case study on surgical skill training. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1045–1048. ACM, 2014a.

Lin Chen, Qiang Zhang, and Baoxin Li. Predicting multiple attributes via relative multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1027–1034, 2014b.

Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

Peng Cui, Zhiyu Wang, and Zhou Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 597–606. ACM, 2014.

Sander Dieleman, Jan Schlüter, Colin Raffel, Eben Olson, Søren Kaae Sønderby, Daniel Nouri, Daniel Maturana, Martin Thoma, Eric Battenberg, Jack Kelly, et al. Lasagne: first release. *Zenodo: Geneva, Switzerland*, 3, 2015.

Yahong Han, Yi Yang, Zhigang Ma, Haoquan Shen, Nicu Sebe, and Xiaofang Zhou. Image attribute adaptation. *IEEE Transactions on Multimedia*, 16(4):1115–1126, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Yuhang He, Long Chen, and Jianda Chen. Multi-task relative attribute prediction by incorporating local context and global style information. In *BMVC*, 2016b.

Junshi Huang, Wei Xia, and Shuicheng Yan. Deep search with attribute-aware deep network. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 731–732. ACM, 2014.

Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2973–2980. IEEE, 2012.

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.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

Hanhui Li, Donghui Li, and Xiaonan Luo. Bap: Bimodal attribute prediction for zero-shot image categorization. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1013–1016. ACM, 2014.

Shaoxin Li, Shiguang Shan, and Xilin Chen. Relative forest for attribute prediction. In *Asian Conference on Computer Vision*, pages 316–327. Springer, 2012.

Chaojie Mao, Yingming Li, Zhongfei Zhang, Yaqing Zhang, and Xi Li. Pyramid person matching network for person re-identification. *arXiv preprint arXiv:1803.02547*, 2018.

Peter O'Donovan, Jānis Lībeks, Aseem Agarwala, and Aaron Hertzmann. Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG)*, 33(4): 92, 2014.

Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Ramachandruni N Sandeep, Yashaswi Verma, and CV Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3614–3621. IEEE, 2014.

Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *European Conference on Computer Vision*, pages 369–383. Springer, 2012.

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

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps (2014). *arXiv preprint arXiv:1312.6034*, 2013.

Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pages 753–769. Springer, 2016.

Yaser Souri, Erfan Noury, and Ehsan Adeli. Deep relative attributes. In *Asian Conference on Computer Vision*, pages 118–133. Springer, 2016.

Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1288–1296, 2016a.

Xianwang Wang, Tong Zhang, Daniel R Tretter, and Qian Lin. Personal clothing retrieval on photo collections by color and attributes. *IEEE Transactions on Multimedia*, 15(8): 2035–2045, 2013.

Yang Wang and Greg Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision*, pages 155–168. Springer, 2010.

Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, and Baoxin Li. Ppp: Joint pointwise and pairwise image label prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6005–6013, 2016b.

Fanyi Xiao and Yong Jae Lee. Discovering the spatial extent of relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1458–1466, 2015.

Xiaoshan Yang, Tianzhu Zhang, Changsheng Xu, Shuicheng Yan, M Shamim Hossain, and Ahmed Ghoneim. Deep relative attributes. *IEEE Transactions on Multimedia*, 18(9): 1832–1842, 2016.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

Aron Yu and Kristen Grauman. Just noticeable differences in visual attributes. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2416–2424. IEEE, 2015.

Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 5571–5580. IEEE, 2017.

Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 33–42. ACM, 2013.