Deep Relative Attributes

Yaser Souri[1], Erfan Noury[2], Ehsan Adeli[3]

[1]Sobhe [2]Sharif University of Technology
[3]University of North Carolina at Chapel Hill

Abstract. Visual attributes are great means of describing images or scenes, in a way both humans and computers understand. In order to establish a correspondence between images and to be able to compare the strength of each property between images, relative attributes were introduced. However, since their introduction, hand-crafted and engineered features were used to learn increasingly complex models for the problem of relative attributes. This limits the applicability of those methods for more realistic cases. We introduce a deep neural network architecture for the task of relative attribute prediction. A convolutional neural network (ConvNet) is adopted to learn the features by including an additional layer (ranking layer) that learns to rank the images based on these features. We adopt an appropriate ranking loss to train the whole network in an end-to-end fashion. Our proposed method outperforms the baseline and state-of-the-art methods in relative attribute prediction on various coarse and fine-grained datasets. Our qualitative results along with the visualization of the saliency maps show that the network is able to learn effective features for each specific attribute. Source code of the proposed method is available at https://github.com/yassersouri/ghiaseddin.

1 Introduction

Visual attributes are linguistic terms that bear semantic properties of (visual) entities, often shared among categories. They are both human understandable and machine detectable, which makes them appropriate for better human machine communications. Visual attributes have been successfully used for many applications, such as image search [1], interactive fine-grained recognition, [2,3] and zero-shot learning [4,5].

Traditionally, visual attributes were treated as binary concepts [6,7], as if they are present or not, in an image. Parikh and Grauman [5] introduced a more natural view on visual attributes, in which pairs of visual entities can be compared, with respect to their relative strength of any specific attribute. With a set of human assessed relative orderings of image pairs, they learn a global ranking function for each attribute that can be used to compare a pair of two novel images respective to the same attribute (Figure 1). While binary visual attributes relate properties to entities (*e.g.*, a dog being furry), relative attributes make it possible to relate entities to each other in terms of their properties (*e.g.*, a bunny being furrier than a dog).

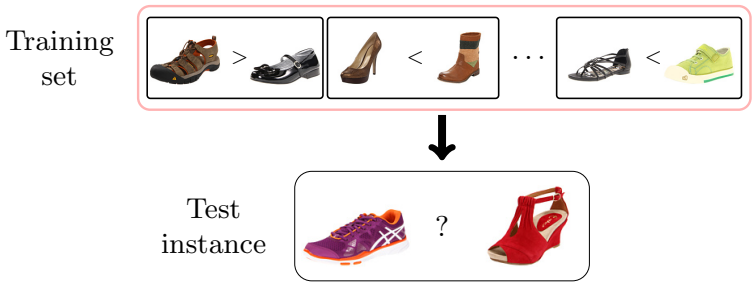


Fig. 1: Visual Relative Attributes. This figure shows samples of training pairs of images from the UT-Zap50K dataset, comparing shoes in terms of the *comfort* attribute (top). The goal is to compare a pair of two novel images of shoes, respective to the same attribute (bottom).

Many have tried to build on the seminal work of Parikh and Grauman [5] with more complex and task-specific models for ranking, while still using hand-crafted visual features, such as GIST [8] and HOG [9]. Recently, Convolutional Neural Networks (ConvNets) have proved to be successful in various visual recognition tasks, such as image classification [10], object detection [11] and image segmentation [12]. Many ascribe the success of ConvNets to their ability to learn multiple layers of visual features from the data.

In this work, we propose to use a ConvNet-based architecture comprising of a feature learning and extraction and ranking portions. This network is used to learn the ranking of images, using relatively annotated pairs of images with similar and/or different strengths of some particular attribute. The network learns a series of visual features, which are known to perform better than the engineered visual features for various tasks [13]. These layers could simply be learned through gradient descent. As a result, it would be possible to learn (or fine-tune) the features through back-propagation, while learning the ranking layer. Interweaving the two processes leads to a set of learned features that appropriately characterizes each single attribute. Our qualitative investigation of the learned feature space further confirms this assumption. This escalates the overall performance and is the main advantage of our proposed method over previous methods. Furthermore, our proposed model can effectively utilize pairs of images with equal annotated attribute strength. The equality relation can happen quite frequently when humans are qualitatively deciding about the relations of attributes in images. In previous works, this is often overlooked and mainly inequality relations are exploited. Our proposed method incorporates an easy and elegant way to deal with equality relations (*i.e.*, an attribute is similarly strong in two images). In addition, it is noteworthy to pinpoint that by exploiting the saliency maps of the learned features for each attribute, similar to [14], we can discover the pixels which contribute the most towards an attribute in the image. This can be used to coarsely localize the specific attribute.

Our approach achieves very competitive results and improves the state-of-the-art (with a large margin in some datasets) on major publicly available datasets for relative attribute prediction, both coarse and fine-grained, while many of the previous works targeted only one of the two sets of problems (coarse or fine-grained), and designed a method accordingly.

The rest of the paper is organized as follows: Section 2 discusses the related works. Section 3 illustrates our proposed method. Then, Section 4 exhibits the experimental setup and results, and finally, Section 5 concludes the paper.

2 Related Works

We usually describe visual concepts with their attributes. Attributes are, therefore, mid-level representations for describing objects and scenes. In an early work on attributes, Farhadi *et al.* [7] proposed to describe objects using mid-level attributes. In another work [15], the authors described images based on a semantic triple "object, action, scene". In the recent years, attributes have shown great performance in object recognition [7,16], action recognition [17,18] and event detection [19]. Lampert *et al.* [4] predicted unseen objects using a zero-shot learning framework, incorporating the binary attribute representation of the objects.

Although detection and recognition based on the presence of attributes appeared to be quite interesting, comparing attributes enables us to easily and reliably search through high-level data derived from *e.g.*, documents or images. For instance, Kovashka *et al.* [20] proposed a relevance feedback strategy for image search using attributes and their comparisons. In order to establish the capacity for comparing attributes, we need to move from binary attributes towards describing attributes relatively. In the recent years, relative attributes have attracted the attention of many researchers. For instance, a linear relative comparison function is learned in [5], based on RankSVM [21] and a non-linear strategy in [22]. In another work, Datta *et al.* [23] used trained rankers for each facial image feature and formed a global ranking function for attributes.

For the process of learning the attributes, different types of low-level image features are often incorporated. For instance, Parikh and Grauman [5] used 512-dimensional GIST [8] descriptors as image features, while Jayaraman *et al.* [24] used histograms of image features, and reduced their dimensionality using PCA. Other works tried learning attributes through *e.g.*, local learning [25] or fine-grained comparisons [26]. Yu and Grauman [26] proposed a local learning-to-rank framework for fine-grained visual comparisons, in which the ranking model is learned using only analogous training comparisons. In another work [27], they proposed a local Bayesian model to rank images, which are hardly distinguishable for a given attribute. However, none of these methods leverage the effectiveness of feature learning methods and only use engineered and hand-crafted features for predicting relative attributes.

As could be inferred from the literature, it is very hard to decide what low-level image features to use for identifying and comparing visual attributes. Recent studies show that features learned through the convolutional neural net-

works (CNNs) [28] (also known as deep features) could achieve great performance for image classification [10] and object detection [29]. Zhang *et al.* [30] utilized CNNs for classifying binary attributes. In other works, Escorcia *et al.* [31] proposed CCNs with attribute centric nodes within the network for establishing the relationships between visual attributes. Shankar *et al.* [32] proposed a weakly supervised setting on convolutional neural networks, applied for attribute detection. Khan *et al.* [33] used deep features for describing human attributes and thereafter for action recognition, and Huang *et al.* [34] used deep features for cross-domain image retrieval based on binary attributes.

Neural networks have also been extended for learning-to-rank applications. One of the earliest networks for ranking was proposed by Burges*et al.* [35], known as RankNet. The underlying model in RankNet maps an input feature vector to a Real number. The model is trained by presenting the network pairs of input training feature vectors with differing labels. Then, based on how they should be ranked, the underlying model parameters are updated. This model is used in different fields for ranking and retrieval applications, *e.g.*, for personalized search [36] or content-based image retrieval [37]. In another work, Yao *et al.* [38] proposed a ranking framework for videos for first-person video summarization, through recognizing video highlights. They incorporated both spatial and temporal streams through 2D and 3D CNNs and detect the video highlights.

3 Proposed Method

We propose to use a ConvNet-based deep neural network that is trained to optimize an appropriate ranking loss for the task of predicting relative attribute strength. The network architecture consists of two parts, the *feature learning and extraction* part and the *ranking* part.

The feature learning and extraction part takes a fixed size image, I_i, as input and outputs the learned feature representation for that image $\psi_i \in \mathbb{R}^d$. Over the past few years, different network architectures for computer vision problems have been developed. These deep architectures can be used for extracting and learning features for different applications. For the current work, outputs of an intermediate layer, like the last layer before the probability layer, from a ConvNet architecture (*e.g.*, AlexNet [10], VGGNet [39] or GoogLeNet [40]) can be incorporated. In our experiments we use the VGG-16 architecture [39] with the last fully connected layer (the class probabilities) removed. This architecture takes as input a 224x224 RGB image and consists of 13, 3x3 convolutional layers with max pooling layers in between. In addition, it has 2 fully connected layers on top of the convolutional layers. For details on the architecture see [39].

One of the most widely used models for relative attributes in the literature is RankSVM [21]. However, in our case, we seek a neural network-based ranking procedure, to which relatively ordered pairs of feature vectors are provided during training. This procedure should learn to map each feature vector to an absolute ranking, for testing purpose. Burges *et al.* [35] introduced such a neural network based ranking procedure that exquisitely fits our needs. We adopt a

similar strategy and thus, the ranking part of our proposed network architecture is analogous to [35] (referred to as RankNet).

During training for a minibatch of image pairs and their target orderings, the output of the feature learning and extraction part of the network is fed into the ranking part and a ranking loss is computed. The loss is then back-propagated through the network, which enables us to simultaneously learn the weights of both feature learning and extraction (ConvNet) and ranking (RankNet) parts of the network. Further with back-propagation we can calculate the derivative of the estimated ordering with respect to the pixel values. In this way, we can generate saliency maps for each attribute (see section 4.6). These saliency maps exhibit interesting properties, as they can be used to localize the regions in the image that are informative about the attribute.

3.1 RankNet: Learning to Rank Using Gradient Descent

This section briefly overviews the RankNet procedure in our context. Given a set (of size n) of pairs of sample feature vectors $\left\{(\psi_1^{(k)}, \psi_2^{(k)}) | k \in \{1, \ldots, n\}\right\} \in \mathbb{R}^{d \times d}$, and target probabilities $\left\{t_{12}^{(k)} | k \in \{1, \ldots, n\}\right\}$, which indicate the probability of sample $\psi_1^{(k)}$ being ranked higher than sample $\psi_2^{(k)}$. We would like to learn a ranking function $f : \mathbb{R}^d \mapsto \mathbb{R}$, such that f specifies the ranking order of a set of features. Here, $f(\psi_i) > f(\psi_j)$ indicates that the feature vector ψ_i is ranked higher than ψ_j, denoted by $\psi_i \triangleright \psi_j$. The RankNet model [35] provides an elegant procedure based on neural networks to learn the function f from a set of pairs of samples and target probabilities.

Denoting $r_i \equiv f(\psi_i)$, RankNet models the mapping from rank estimates to posterior probabilities $p_{ij} = P(\psi_i \triangleright \psi_j)$ using a logistic function

$$p_{ij} := \frac{1}{1 + e^{-(r_i - r_j)}}. \tag{1}$$

The loss for the sample pair of feature vectors (ψ_i, ψ_j) along with target probability t_{ij} is defined as

$$C_{ij} := -t_{ij} \log(p_{ij}) - (1 - t_{ij}) \log(1 - p_{ij}), \tag{2}$$

which is the binary cross entropy loss. Figure 2 (left) plots the loss value C_{ij} as a function of $r_i - r_j$ for three values of target probability $t_{ij} \in \{0, 0.5, 1\}$. This function is quite suitable for ranking purposes, as it acts differently compared to regression functions. Specifically, we are not interested in regression instead of ranking for two reasons: First, we cannot regress the absolute rank of images, since the annotations are only available in pairwise ordering for each attribute, in relative attribute datasets (see section 4.1). Second, regressing the difference $r_i - r_j$ to t_{ij} is inappropriate. To understand this, let's consider the squared loss

$$R_{ij} = \left[(r_i - r_j) - t_{ij}\right]^2, \tag{3}$$

which is typically used for regression, illustrated in Figure 2 (right). We observe that the regression loss forces the difference of rank estimates to be a specific value and disallows over-estimation. Furthermore, its quadratic natures makes it sensitive to noise. This sheds light into why regression objective is the wrong objective to optimize when the goal is ranking.

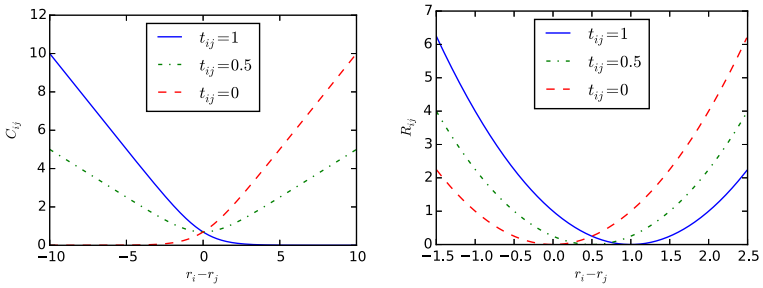


Fig. 2: The ranking loss value for three values of the target probability (left). The squared loss value for three values of the target probability, typically used for regression (right).

Note that when $t_{ij} = 0.5$, and no information is available about the relative rank of the two samples, the ranking cost becomes symmetric. This can be used as a way to train on patterns that are desired to have similar ranks. This is somewhat not much studied in the previous works on relative attributes. Furthermore, this model asymptotically converges to a linear function which makes it more appropriate for problems with noisy labels.

Training this model is possible using stochastic gradient descent or its variants like RMSProp. While testing, we only need to estimate the value of $f(\psi_i)$, which resembles the absolute rank of the testing sample. Using $f(\psi_i)$s, we can easily infer both absolute or relative ordering of the testing pairs.

3.2 Deep Relative Attributes

Our proposed model is depicted in figure 3. The model is trained separately, for each attribute. During training, pairs of images (I_i, I_j) are presented to the network, together with the target probability t_{ij}. If for the attribute of interest $I_i \triangleright I_j$ (image i exhibits more of the attribute than image j), then t_{ij} is expected to be larger than 0.5 depending on our confidence on the relative ordering of I_i and I_j. Similarly, if $I_i \triangleleft I_j$, then t_{ij} is expected to be smaller than 0.5, and if it is desired that the two images have the same rank, t_{ij} is expected to be 0.5. Because of the nature of the datasets, we chose t_{ij} from the set $\{0, 0.5, 1\}$, according to the available annotations in the dataset.

The pair of images then go though the feature learning and extraction part of the network (ConvNet). This procedure maps the images onto feature vectors ψ_i

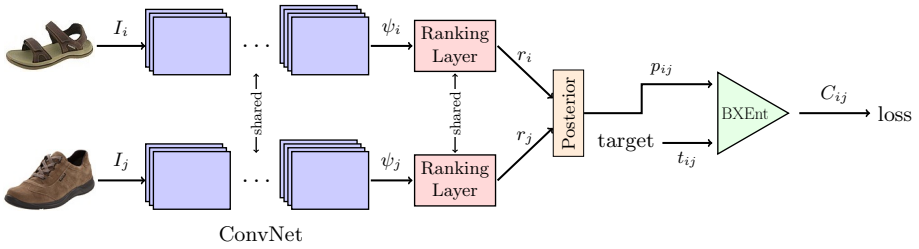


Fig. 3: The overall schematic view of the proposed method during training. The network consists of two parts, the *feature learning and extraction* part (labeled ConvNet in the figure), and the *ranking* part (the Ranking Layer). Pairs of images are presented to the network with their corresponding target probabilities. This is used to calculate the loss, which is then back-propagated through the network to update the weights.

and ψ_j, respectively. Afterwards, these feature vectors go through the ranking layer, as described in section 3.1. We choose the ranking layer to be a fully connected neural network layer with linear activation function, a single output neuron and weights w and b. It maps the feature vector ψ_i to the estimated absolute rank of that feature vector, $r_i \in \mathbb{R}$, where

$$r_i := w^T \psi_i + b. \tag{4}$$

The two estimated ranks r_i and r_j, for the two images I_i and I_j in comparison, are then combined (using Equation (1)) to output the estimated posterior probability $p_{ij} = P(I_i \triangleright I_j)$. This estimated posterior probability is used along with the target probability t_{ij} to calculate the loss, as in Equation (2). This loss is then back-propagated through the network and is used to update the weights of the whole network, including both the weights of the feature learning and extraction sub-network and the ranking layer.

During testing (Figure 4), we need to calculate the estimated absolute rank r_k for each testing image I_k. Using these estimated absolute ranks, we can then easily infer both the relative or absolute attribute ordering, for all testing pairs.

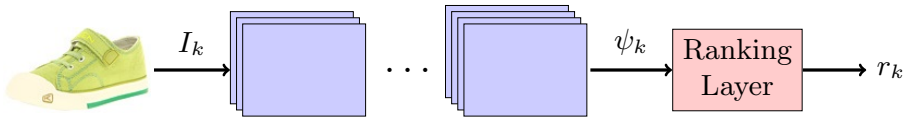


Fig. 4: During testing, we only need to evaluate r_k for each testing image. Using this value, we can infer the relative or absolute ordering of testing images, for the attribute of interest.

4 Experiments

To evaluate our proposed method, we quantitatively compare it with the state-of-the-art methods, as well as an informative baseline on all publicly available benchmarks for relative attributes to our knowledge. Furthermore, we perform multiple qualitative experiments to demonstrate the capability and superiority of our method.

4.1 Datasets

To assess the performance of the proposed method, we have evaluated it on all publicly available datasets to our knowledge: **Zappos50K** [26] (both coarse and fine-grained versions), **LFW-10** [41] and for the sake of completeness and comparison with previous works, on **PubFig** and **OSR** datasets of [5].

UT-Zap50K [26] dataset is a collection of images with annotations for relative comparison of 4 attributes. 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 relative attributes are annotated for fine-grained pairs, for which making the distinction between them is hard according to human annotators. Training set for Zappos50K-1 contains approximately 1500 to 1800 annotated pairs of images for each attribute. These are divided into 10 train/test splits which are provided alongside the dataset and used in this work. Meanwhile, Zappos50K-2 only contains a test set of approximately 4300 pairs, while its training set is the combination of training and testing sets of Zappos50K-1.

We have also conducted experiments on the **LFW-10** [41] dataset. This dataset has 2000 images of faces of people and annotations for 10 attributes. For each attribute, a random subset of 500 pairs of images have been annotated for each training and testing set.

PubFig [5] dataset (a set of public figure faces), consists of 800 facial images of 8 random subjects, with 11 attributes. **OSR** [5] dataset contains 2688 images of outdoor scenes in 8 categories, for which 6 relative attributes are defined. The ordering of samples in both PubFig and OSR datasets are annotated in a category level, *i.e.*, all images in a specific category may be ranked higher, equal, or lower than all images in another category, with respect to an attribute. This sometimes causes annotation inconsistencies [41]. In our experiments, we have used the provided training/testing split of PubFig and OSR datasets.

4.2 Experimental setup

We train our proposed model (described in Section 3) for each attribute, separately. In our proposed model, it is possible to train multiple attributes at the same time, however, this is not done due to the structure of the datasets, in which for each training pair of images only a certain attribute is annotated.

We have used the Lasagne [42] deep learning framework to implement our model. In all our experiments, for the feature learning and extraction part of the

network, we use the VGG-16 model of [39] and trim out the probability layer (all layers up to fc7 are used, only the probability layer is not included). We initialize the weights of the model using a pretrained model on ILSVRC 2014 dataset [43] for the task of image classification. These weights are fine-tuned as the network learns to predict the relative attributes (see section 4.5). The weights w of the ranking layer are initialized using the Xavier method [44], and the bias is initialized to 0.

For training, we use stochastic gradient descent with RMSProp [45] updates and minibatches of size 32 (16 pair of images). We set the learning rate of the feature learning and extraction layers of the network to 10^{-5} and the ranking layer to 10^{-4} for all experiments initially, then RMSProp changes the learning rates dynamically during training. We have also used weight decay (ℓ_2 norm regularization), with a fixed 10^{-5} multiplier. Furthermore, when calculating the binary cross entropy loss, we clip the estimated posterior p_{ij} to be in the range $[10^{-7}, 1 - 10^{-7}]$. This is used to prevent the loss from diverging.

In each epoch, we randomly shuffle the training pairs. The number of epochs of training were chosen to reflect the training size. For Zappos50K and LFW-10 datasets, we train for 25 and 40 epochs, respectively. For PubFig and OSR datasets, we train for 2 epochs due to the large number of training sample pairs. When performing evaluation on OSR the total number of pairs is too large (around 3 million pairs) we only evaluate on a 5% random subset of them.

4.3 Baseline

As a baseline, we have also included results for the RankSVM method (as in [5]), when the features given to the method were computed from the output of the VGG-16 pretrained network on ILSVRC 2014.

Using this baseline we can evaluate the extent of effectiveness of off-the-shelf ConvNet features [13] for the task of ranking. In a sense, comparing this baseline with our proposed method reveals the effect of features fine-tuning, for the task.

4.4 Quantitative Results

Following [5,26,41], we report the accuracy in terms of the percentage of correctly ordered pairs. For our proposed method, we report the mean accuracy and standard deviation over 3 separate runs.

Table 1 and 2 shows our results on the OSR and PubFig dataset respectively. Our method outperforms the baseline and the state-of-the-art on this dataset by a considerable margin, on most attributes. These are relatively easy datasets but have their own challenges. Specifically the OSR dataset contains attributes like "Perspective" which are very generic, high level and global in the image, which might not correspond easily to local low level image features. We think that our proposed method is specially well suited for such cases.

Table 3 shows our results on the LFW-10 dataset. On this dataset, our method performs competitive with respect to the state-of-the-art, but cannot

Table 1: Results for the OSR dataset

Method	Natural	Open	Perspective	Large Size	Diag	ClsDepth	Mean
Relative Attributes [5]	95.03	90.77	86.73	86.23	86.50	87.53	88.80
Relative Forest [22]	95.24	92.39	87.58	88.34	89.34	89.54	90.41
Fine-grained Comparison [26]	95.70	94.10	90.43	91.10	92.43	90.47	92.37
VGG16-fc7 (baseline)	98.00	94.46	92.92	94.08	94.91	95.02	94.90
RankNet (ours)	99.40 (± 0.10)	97.44 (± 0.16)	96.88 (± 0.13)	96.79 (± 0.32)	98.43 (± 0.23)	97.65 (± 0.16)	**97.77** (± 0.10)

Table 2: Results for the PubFig dataset

Method	Male	White	Young	Smiling	Chubby	Forehead	Eyebrow	Eye	Nose	Lip	Face	Mean
Relative Attributes [5]	81.80	76.97	83.20	79.90	76.27	87.60	79.87	81.67	77.40	79.17	82.33	80.56
Relative Forest [22]	85.33	82.59	84.41	83.36	78.97	88.83	81.84	83.15	80.43	81.87	86.31	83.37
Fine-grained Comparison [26]	91.77	87.43	91.87	87.00	87.37	94.00	89.83	91.40	89.07	90.43	86.70	89.72
VGG16-fc7 (baseline)	85.56	80.59	85.20	84.81	82.56	88.50	83.50	83.11	81.52	85.67	86.23	84.30
RankNet (ours)	95.50 (± 0.36)	94.60 (± 0.55)	94.33 (± 0.36)	95.36 (± 0.56)	92.32 (± 0.36)	97.28 (± 0.49)	94.53 (± 0.64)	93.19 (± 0.51)	94.24 (± 0.24)	93.62 (± 0.20)	94.76 (± 0.24)	**94.52** (± 0.08)

Table 3: Results for the LFW-10 dataset

Method	Bald	DkHair	Eyes	GdLook	Mascu.	Mouth	Smile	Teeth	FrHead	Young	Mean
Fine-grained Comparison [22]	67.9	73.6	49.6	64.7	70.1	53.4	59.7	53.5	65.6	66.2	62.4
Relative Attributes [5]	70.4	75.7	52.6	68.4	71.3	55.0	54.6	56.0	64.5	65.8	63.4
Global + HOG [46]	78.8	72.4	70.7	67.6	84.5	67.8	67.4	71.7	79.3	68.4	72.9
Relative Parts [41]	71.8	80.5	90.5	77.6	67.0	77.6	81.3	76.2	80.2	82.4	78.5
Spatial Extent [47]	83.21	88.13	82.71	72.76	93.68	88.26	88.16	88.46	90.23	75.05	**84.66**
VGG16-fc7 (baseline)	72.26	79.23	55.64	62.85	90.80	62.42	66.38	59.38	64.45	66.31	67.97
RankNet (ours)	81.14 (± 3.39)	88.92 (± 0.75)	74.44 (± 5.97)	70.28 (± 0.54)	98.08 (± 0.33)	85.46 (± 0.70)	82.49 (± 1.41)	82.77 (± 2.15)	81.90 (± 2.00)	76.33 (± 0.43)	**82.18** (± 1.08)

outperform it. We think this might be due to label noise in this dataset and due to the fact that most of the attributes in this dataset are highly local and methods that outperform us on this dataset look locally on regions of the image instead of the whole image.

Tables 4 and 5 show the results on Zappos50K-1 and Zappos50K-2 datasets, respectively. Our method, again, achieves the state-of-the-art accuracy on both coarse-grained and fine-grained datasets. Our proposed method learns appropriate features for the task, given the large amount of training data available in this dataset.

4.5 Qualitative Results

Our proposed method uses a deep network with two parts, the feature learning and extraction part and the ranking part. During training, not only the weights for the ranking part are learned, but also the weights for the feature learning and extraction part of the network, which were initialized using a pretrained network, are fine-tuned. By fine-tuning the features, our network learns a set of features that are more appropriate for the images of that particular dataset, along with the attribute of interest. To show the effectiveness of fine-tuning the features of the feature learning and extraction part of the network, we have projected them (features before and after fine-tuning) into 2-D space using the t-SNE [48], as

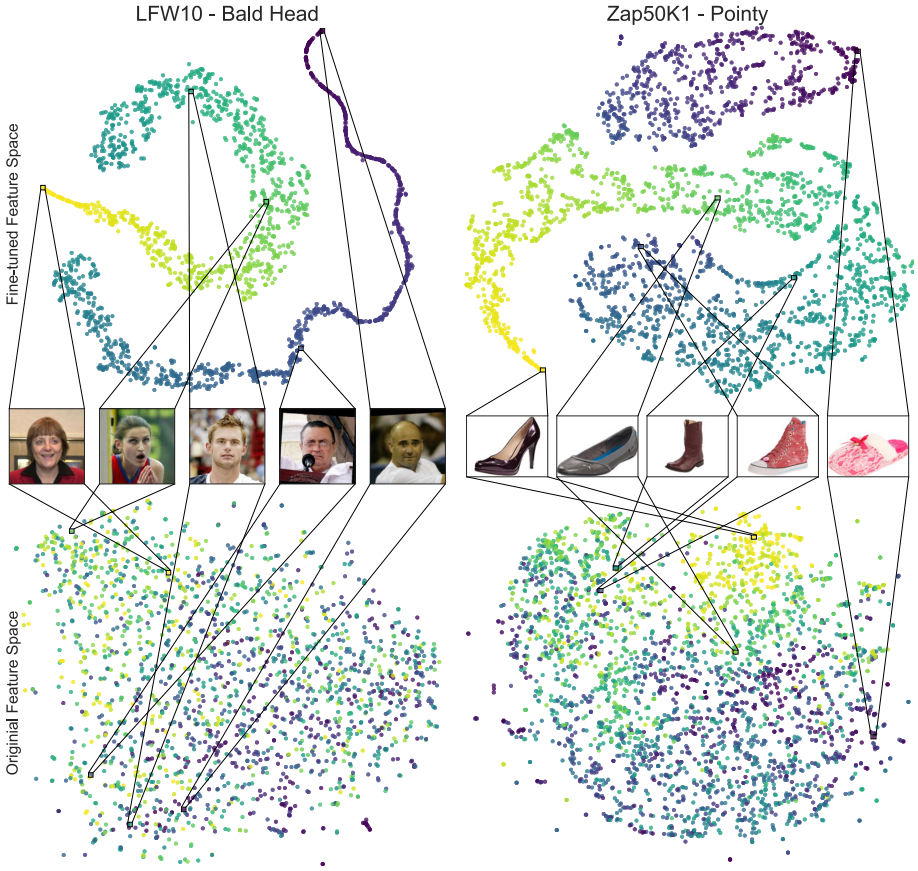


Fig. 5: t-SNE embedding of images in fine-tuned feature space (top) and original feature space (bottom). The set of visualizations on the left are for the *Bald Head* attribute of the LFW-10 dataset, while the visualizations on the right are for the *Pointy* attribute of the Zappos50K-1 dataset. Images in the middle row show a number of samples from the feature space. In the fine-tuned feature space, it is clear that images are ordered according to their value of the attribute. Each point is colored according to its value of the respective attribute, to discriminate images according to their value of the attribute.

Table 4: Results for the UT-Zap50K-1 (coarse) dataset

Method	**Open**	**Pointy**	**Sporty**	**Comfort**	**Mean**
Relative Attributes [5]	87.77	89.37	91.20	89.93	89.57
Fine-grained Comparison [26]	90.67	90.83	92.67	92.37	91.64
Spatial Extent [47]	95.03	94.80	96.47	95.60	95.47
VGG16-fc7 (baseline)	89.67	90.67	91.67	91.00	90.75
RankNet (ours)	95.37 (± 0.82)	94.43 (± 0.75)	97.30 (± 0.81)	95.57 (± 0.97)	**95.67** (± 0.49)

Table 5: Results for the UT-Zap50K-2 (fine-grained) dataset

Method	**Open**	**Pointy**	**Sporty**	**Comfort**	**Mean**
Relative Attributes [5]	60.18	59.56	62.70	64.04	61.62
Fine-grained Comparison [26]	74.91	63.74	64.54	62.51	66.43
LocalPair + ML + HOG [46]	76.2	65.3	64.8	63.6	67.5
VGG16-fc7 (baseline)	64.82	64.51	67.31	67.01	65.91
RankNet (ours)	73.45 (± 1.23)	68.20 (± 0.18)	73.07 (± 0.75)	70.31 (± 1.50)	**71.26** (± 0.50)

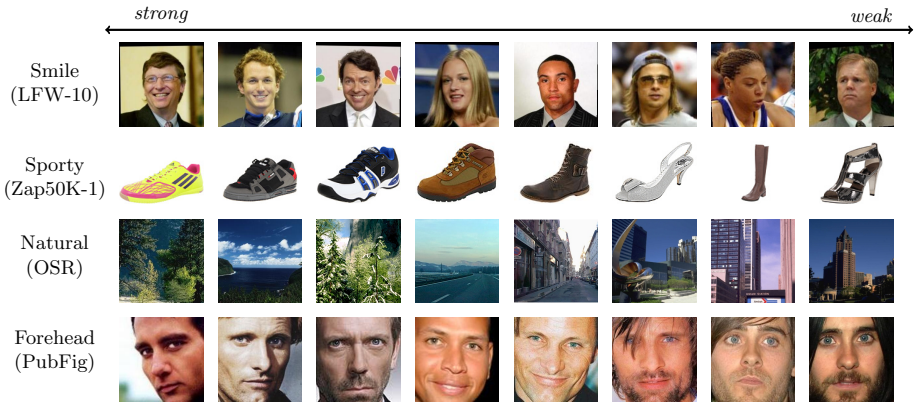


Fig. 6: Sample images from different datasets, ordered according to the predicted value of their respective attribute.

can be seen in Figure 5. The visualizations on the top of each figure show the images projected into 2-D space from the fine-tuned feature space, while the visualizations on the bottom show the images from the original feature space. Each image is displayed as a point and colored according to its attribute strength. It is clear from these visualizations that fine-tuned feature space is better in capturing the ordering of images with respect to the respective attribute. Since t-SNE embedding is a non-linear embedding, relative distances between points in the high-dimensional space and the low-dimensional embedding space are preserved, thus close points in the low-dimensional embedding space are also close to each other in the high-dimensional space. It can, therefore, be seen that

fine-tuning indeed changes the feature space such that images with similar values of the respective attribute get projected into a close vicinity of the feature space. However, in the original feature space, images are projected according to their visual content, regardless of their value of the attribute.

Another property of our network is that it can achieve a total ordering of images, given a set of pairwise orderings. In spite of the fact that training samples are pairs of images annotated according to their relative value of the attribute, the network can generalize the relativity of attribute values to a global ranking of images. Figure 6 shows some images ordered according to their value of the respective attribute.

4.6 Saliency Maps and Localizing the Attributes

We have also used the method of [14] to visualize the saliency of each attribute. Giving two image as inputs to the network, we take the derivative of the estimated posterior with respect to the input images and visualize them. Figure 7 shows some sample visualization for some test pairs. To generate this figure we have applied Gaussian smoothing to the saliency map.

These saliency maps visualize the pixels in the images which contributed most to the ranking predicted by the network. Sometimes these saliency maps are easily interpretable by humans and they can be used to localize attributes using the same network that was trained to rank the attributes in an unsupervised manner, *i.e.*, although we haven't explicitly trained our network to localize the salient and informative regions of the image, it has implicitly learned to find these regions. We see that this technique is able to localize both easy to localize attributes such as "Bald Head" in the LFW10 dataset and abstract attributes such as "Natural" in the OSR dataset.

5 Conclusion

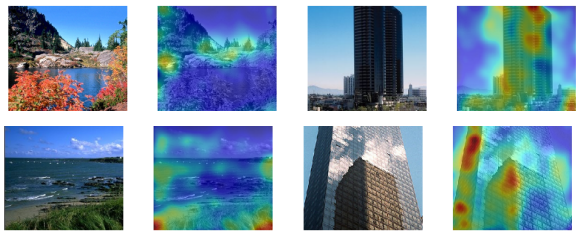
In this paper, we introduced an approach for relative attribute prediction on images, based on convolutional neural networks. Unlike previous methods that use engineered or hand-crafted features, our proposed method learns attribute-specific features, on-the-fly, during the learning procedure of the ranking function. Our results achieve state-of-the-art performance in relative attribute prediction on various datasets both coarse- and fine-grained. We qualitatively show that the feature learning and extraction part, effectively learns appropriate features for each attribute and dataset. Furthermore, we show that one can use a trained model for relative attribute prediction to obtain saliency maps for each attribute in the image.

6 Acknowledgments

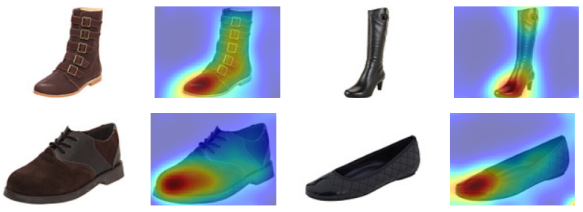
We would like to thank Computer Engineering Department of Sharif University of Technology and HPC center of IPM for their support with computational resources.

LFW10 - Bald Head

LFW10 - Good Looking



OSR - Natural



Zappos50k1 - Pointy

Fig. 7: Saliency maps obtained from the network. First we feed two test images into the network and compute the derivative of the estimated posterior with respect to the pair of input images and use the method of [14] to visualize salient pixels with Gaussian smoothing. In each row, the two input images from the a dataset's test set with their corresponding overlaid saliency maps are shown (the warmer the color of the overlay image, the more salient that pixel is).

References

1. Kovashka, A., Parikh, D., Grauman, K.: Whittlesearch: Image Search with Relative Attribute Feedback. In: CVPR. (2012)
2. Branson, S., Beijbom, O., Belongie, S.: Efficient large-scale structured learning. In: CVPR. (2013)
3. Branson, S., Wah, C., Babenko, B., Schroff, F., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: ECCV. (2010)
4. Lampert, C., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE TPAMI **36** (2014) 453–465
5. Parikh, D., Grauman, K.: Relative attributes. CVPR (2011) 503–510
6. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS. (2007) 433–440
7. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
8. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV **42** (2001) 145–175
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886–893
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440
13. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPRW. (2014) 512–519
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
15. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.A.: Every picture tells a story: Generating sentences from images. In: ECCV. (2010)
16. Tao, R., Smeulders, A.W., Chang, S.F.: Attributes and categories for generic instance search from one example. In: CVPR. (2015) 177–186
17. Khan, F., van de Weijer, J., Anwer, R., Felsberg, M., Gatta, C.: Semantic pyramids for gender and action recognition. IEEE TIP **23** (2014) 3633–3645
18. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR. (2011) 3337–3344
19. Liu, J., Yu, Q., Javed, O., Ali, S., Tamrakar, A., Divakaran, A., Cheng, H., Sawhney, H.: Video event recognition using concept attributes. In: WACV. (2013) 339–346
20. Kovashka, A., Grauman, K.: Attribute pivots for guiding relevance feedback in image search. In: ICCV. (2013) 297–304
21. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM KDD. (2002) 133–142
22. Li, S., Shan, S., Chen, X.: Relative forest for attribute prediction. In: ACCV. (2012)
23. Datta, A., Feris, R., Vaquero, D.: Hierarchical ranking of facial attributes. In: FG. (2011) 36–42

24. Jayaraman, D., Sha, F., Grauman, K.: Decorrelating semantic visual attributes by resisting the urge to share. In: CVPR. (2014) 1629–1636
25. Zhang, H., Berg, A., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: CVPR. Volume 2. (2006) 2126–2136
26. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: CVPR. (2014)
27. Yu, A., Grauman, K.: Just noticeable differences in visual attributes. In: ICCV. (2015)
28. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: NIPS. (1989)
29. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014) 580–587
30. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: PANDA: Pose aligned networks for deep attribute modeling. In: CVPR. (2014) 1637–1644
31. Escorcia, V., Carlos Niebles, J., Ghanem, B.: On the relationship between visual attributes and convolutional networks. In: CVPR. (2015)
32. Shankar, S., Garg, V.K., Cipolla, R.: Deep-carving: Discovering visual attributes by carving deep neural nets. In: CVPR. (2015)
33. Khan, F.S., Anwer, R.M., van de Weijer, J., Felsberg, M., Laaksonen, J.: Deep semantic pyramids for human attributes and action recognition. In: Image Analysis. Springer (2015) 341–353
34. Huang, J., Feris, R.S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: ICCV. (2015)
35. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: ICML. (2005) 89–96
36. Song, Y., Wang, H., He, X.: Adapting deep ranknet for personalized search. In: WSDM. (2014)
37. Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y., Li, J.: Deep learning for content-based image retrieval: A comprehensive study. In: ACM MM. (2014) 157–166
38. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: CVPR. (2016)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
40. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. (2015)
41. Sandeep, R.N., Verma, Y., Jawahar, C.V.: Relative parts: Distinctive parts for learning relative attributes. In: CVPR. (2014)
42. Dieleman, S., Schlter, J., Raffel, C., Olson, E., Snderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J.D., Heilman, M., diogo149, McFee, B., Weideman, H., takacsg84, peterderivaz, Jon, instagibbs, Rasul, D.K., CongLiu, Britefury, Degrave, J.: Lasagne: First release. (2015)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115** (2015) 211–252
44. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. (2010) 249–256

45. Tieleman, T., Hinton, G.: Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning (2012)
46. Verma, Y., Jawahar, C.V.: Exploring locally rigid discriminative patches for learning relative attributes. In: BMVC. (2015)
47. Xiao, F., Jae Lee, Y.: Discovering the spatial extent of relative attributes. In: CVPR. (2015)
48. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. JMLR **9** (2008) 85