The 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI 2015)
October 28 ∼ 30, 2015 / KINTEX, Goyang city, Korea

Relative Attributes with Deep Convolutional Neural Network

Dong-Jin Kim, Donggeun Yoo, Sunghoon Im, Namil Kim,
Tharatch Sirinukulwattana, and In So Kweon

Department of Electrical Engineering, KAIST, Daejeon, Korea
(Tel : +82-42-350-5465; E-mail: {djkim, dgyoo, shim, nikim, tharatch }@rcv.kaist.ac.kr , iskweon77@kaist.ac.kr)

Abstract - Our work is based on the idea of relative attributes, aiming to provide more descriptive information to the images. We propose the model that integrates relative-attribute framework with deep Convolutional Neural Networks (CNN) to increase the accuracy of attribute comparison. In addition, we analyzed the role of each network layer in the process. Our model uses features extracted from CNN and is learned by Rank SVM method with these feature vectors. As a result, our model outperforms the original relative attribute model in terms of significant improvement in accuracy.

Keywords - Deep learning, relative attributes, convolutional neural networks

1. Introduction

Visual attributes have come to the fore in today's society visual recognition [1]. Rather than recognizing object-level classes, attributes can provide richer image understanding and further be used for mid-level image representation. However, if the attributes are binary, the settings would be restrictive [2]. To avoid this problem, the model "relative attributes" has been introduced [2]. It compares important attributes within an image, therefore helps understand more details.

Krizhevsky et al. [3] show the state-of-the-art image classification accuracy on ImageNet Large Scale Visual Recognition Chellenge (ILSVRC 2012) by using deep Convolutional Neural Networks (CNN). Since then, activations from a pre-trained CNN have been used as a generic image representation for wide visual recognition tasks. We, therefore, have attempted to improve the relative attribute model by adopting the deep image representation.

There are two major contributions in our work. The first is to employ the deep image representation instead of GIST features. The overview of the concept is depicted in Figure 1. The second contribution is the layer-level analysis of CNN through which we can distinguish the most suitable layer in the relative attributes. We show through the result how our proposed system has improved the accuracy compared to the baseline method.

2. Proposed Method

We train an 8-layered Alex network [3] with ILSVRC'12 dataset. Given a target image, we extract an activation vector at a specific layer, to be fed to Rank SVM. We utilize Caffe toolkit [1] to train a network and extract activations.

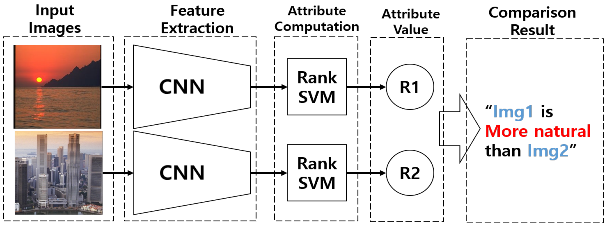


Fig. 1 Our model represents images by the activations from a pre-trained CNN. The image vectors are fed to the Rank SVM, and result in the attribute score. A pair of scores are compared to identify the stronger one.

2.1 Training with Rank SVM

Our task is to estimate relativeness of attributes. We solve the task by using the modified version of Rank SVM from the base-line reference [2]. If an input image goes through 5 convolution layers and 2 fully-connected layers, we are collect up to 4096 dimensional feature vector x_i from the image. The ranking function $R(.)$ of attribute m for image j is:

$$R_m(x_j) = w_m^\top x_j \tag{1}$$

For a pair of images, we can compute the gap of the Rank SVM scores.

The goal of this training step is to find the optimal model parameter of $w_m^\top$ that maximizes the gap for the ordered pairs and minimize that of un-ordered pairs as follows.

$$\hat{w}_m = arg\min_{w_m} \frac{1}{2}|w_m^\top|_2^2 + \lambda \cdot (\sum_{(i,j)} \alpha_{i,j}^2 + \sum_{(i,j)} \beta_{i,j}^2) \tag{2}$$

$$\forall(i,j) O_m : w_m^\top x_m \geq 1 - \alpha_{i,j},\ \alpha_{i,j} \geq 0 \tag{3}$$

$$\forall(i,j) U_m : |w_m^\top x_m| \leq \beta_{i,j},\ \beta_{i,j} \geq 0 \tag{4}$$

where ordered pairs of image O_m and un-ordered pairs of image U_m satisfy the following constraints:

$$\forall(i,j) O_m : w_m^\top x_i > w_m^\top x_j \tag{5}$$

$$\forall(i,j) U_m : w_m^\top x_i = w_m^\top x_j \tag{6}$$

Because the objective is composed of max-margin formulation, our model has advantages in terms of accuracy comparing to other methods such as multivariant regression [2].

2.2 Test

Testing process is similar to the training step. After going through the CNN, image vectors will be multiplied by the trained Rank SVM. We compare the function output in order to distinguish the difference of a certain attribute between two images.

[1] http://caffe.berkeleyvision.org/

3. Experimental Results

3.1 Comparison of Attributes by Pair

We tested images from the OSR [4] and PubFig [5] datasets. Pubfig has 11 attributes while OSR has 6 attributes.

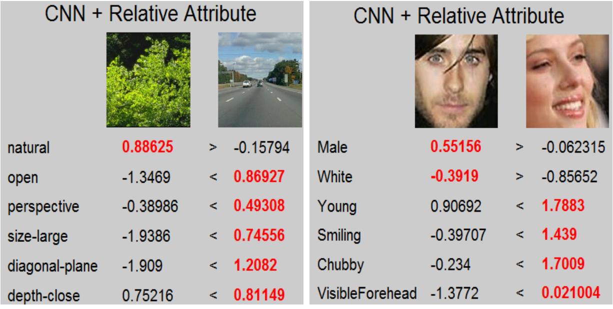
3.2 Comparison of GIST, BOW and CNN

To show the effectiveness, we compare our method against GIST and BOW over OSR datatset. Our deep image representation is noted by CNN. We compute accuracy from all attributes and average them as shown in Table 1. The time for encoding an image was 0.006 seconds. For Rank SVM test, it costs 0.01 seconds.

Table 1 Average accuracy with respect to the feature

Feature	GIST	SIFT+BOW	CNN
Accuracy	69.32%	80.02%	**89.32%**

The attributes from GIST feature have the lowest accuracy, while CNN achieves the highest. Representative examples are shown in Figure 2.



CNN + Relative Attribute

natural	**0.88625**	>	-0.15794
open	-1.3469	<	**0.86927**
perspective	-0.38986	<	**0.49308**
size-large	-1.9386	<	**0.74556**
diagonal-plane	-1.909	<	**1.2082**
depth-close	0.75216	<	**0.81149**

CNN + Relative Attribute

Male	**0.55156**	>	-0.062315
White	**-0.3919**	>	-0.85652
Young	0.90692	<	**1.7883**
Smiling	-0.39707	<	**1.439**
Chubby	-0.234	<	**1.7009**
VisibleForehead	-1.3772	<	**0.021004**

Fig. 2 Sample results with (a) OSR, (b) PubFig dataset

3.3 Analysis of Layer-wise Image Representation

To examine the proper layer for relative attribute when a pre-trained CNN given, we represent images by activations of each layer and obtain each result. The attribute ”male” in PubFig dataset was used for this experiment.

As shown in Figure 3, the accuracy of CNN system analyzed from the first to second convolutional layer is much lower than the accuracy from the third to seventh layer. This is because the first and second convolutional layers extract low-level features, while the layers from the third to seventh extract mid/high-level features, which are more distinctive and informative [6, 7].

Compared to GIST, the activations from the first and second layers are less efficient for recognizing attribute. This implies that low-level representation alone is insufficient. SIFT+BOW representation, which is also well known as a mid-level image representation, shows better performance than GIST but less than CNN. We conclude that the image representation using an pre-trained CNN can also be successfully used for recognizing relative attribute as well as object-level classification.

4. Discussion and Conclusion

By introducing our deep image representation followed by Rank SVM, we succeeded to improve the accuracy of relative attributes. This research in attributes can be further extended to the image caption generators [8, 9]. For the tasks, attributes give more informative description since current image caption generators are rely only on object existence. As the accuracy of relative attributes increases with our method, we anticipate that the description quality will be increased.

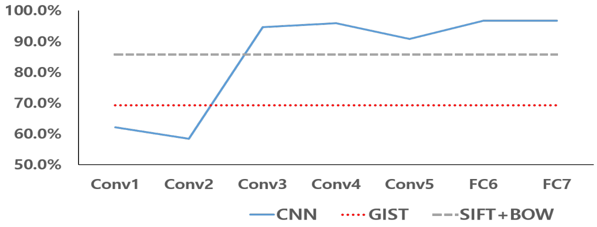


Fig. 3 Accuracy from various feature types

Acknowledgement

This work was supported by the Technology Innovation Program (No. 10048320), funded by the Ministry of Trade, Industry & Energy (MI, Korea).

References

[1] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth, “Describing objects by their attributes,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1778–1785.

[2] Devi Parikh and Kristen Grauman, “Relative attributes,” in *Proc. of Int’l Conf. on Computer Vision (ICCV)*, 2011.

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

[4] Aude Oliva and Antonio Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[5] 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*. IEEE, 2009, pp. 365–372.

[6] Pulkit Agrawal, Ross Girshick, and Jitendra Malik, “Analyzing the performance of multilayer neural networks for object recognition,” in *Computer Vision–ECCV 2014*, pp. 329–344. Springer, 2014.

[7] Matthew D Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, “Show and tell: A neural image caption generator,” *arXiv preprint arXiv:1411.4555*, 2014.

[9] Andrej Karpathy and Li Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *arXiv preprint arXiv:1412.2306*, 2014.