Deep Relative Attributes

Anonymous CVPR submission

Paper ID ****

Abstract

Relative Attributes are a very natural way of thinking in terms of attributes and communicating with machines. The idea was introduced in the award winning ICCV 2011 paper by D. Parikh and K. Grauman [6]. In this project we want to improve their system by using a Deep Neural Network instead of a RankSVM to do the ranking. This way we can also use Convolutional Layers to learn the features end-to-end or fine-tune the features.

1. Introduction

After the "Relative Attributes" paper [6], there was a stream of papers that aimed to solve the same or similar task ([4, 8, 7, 3]) using a different and often more complex model instead of the original RankSVM model. I think as progress in the "Visual Recognition" field has shown, for solving the problem you actually need to change the features not the model. So in this project I actually want to experiment with how learning the features end-to-end (or fine-tuning the features) can improve Relative Attributes accuracy and power.

2. Related works

3. End-to-end deep relative attributes

3.1. Fine-tuned representation of relative attributes

3.2. Ranking layer

4. Experiments

4.1. Datasets

To assess the performance of the proposed method, we have evaluated our method on five datasets. **PubFig** [1] (faces) and **OSR** [5] (outdoor scenes) datasets are used to compare the results of the proposed method with previous works. The PubFig dataset consists of 800 facial images of 8 random subjects. 11 attributes are defined in this dataset and attribute ordering of images is annotated in category level, *i.e.*all images of a category may be ranked higher, equal, or lower than all images of another category, with respect to an attribute. The OSR dataset contains 2688 images in 8 categories, for which 6 relative attributes are defined. Like the PubFig dataset, relative ranking of attributes for this dataset have been annotated in category level. Also **UT-Zap50K** [8] (shoes) and **LFW-10** [7] (faces) datasets, which are more challenging, have been used to assess the quality of the proposed method. The UT-Zap50K dataset consists of two collections, namely UT-Zap50K-1 in which *coarse* relative attributes are compared for image pairs, and UT-Zap50K-2 in which *fine-grained* relative attributes are compared for image pairs. The LFW-10 dataset consists of 2000 images and 10 attributes and for each attribute a random subset of 500 pairs of images have been annotated for each train and test set. Large number of categories in the UT-Zap50K and LFW-10 datasets makes them more challenging than the PubFig and OSR datasets. In addition to these datasets, to further analyze the properties of this end-to-end model and the feature hierarchy obtained, we have also experiemted with the MNIST [2] dataset. We have used class labels for images as the relative attribute and used the value of class label to rank images.

4.2. Experimental setup

4.3. Baseline and compared methods

4.4. Results

4.5. Discussions

5. Conclusion

References

[1] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009. 1

[2] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, 1998. 1

[3] Y. J. Lee, A. A. Efros, and M. Hebert . Style-aware mid-level representation for discovering visual connections in space and time. In *ICCV*, 2013. 1

[4] S. Li, S. Shan, and X. Chen. Relative forest for attribute prediction. In *ACCV*, volume 7724, pages 316–327. 2013. 1

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 1

[6] D. Parikh and K. Grauman. Relative attributes. *CVPR*, pages 503–510, 2011. 1

[7] R. N. Sandeep, Y. Verma, and C. V. Jawahar. Relative parts: Distinctive parts for learning relative attributes. 2014. 1

[8] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *CVPR*, 2014. 1