Deep Relative Attributes

Anonymous CVPR submission

Paper ID 725

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 [24]. 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 [24], there was a stream of papers that aimed to solve the same or similar task ([19, 30, 25, 18]) 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

Attributes are mid-level representations for describing objects, scenes and literally everything. We usually describe visual concepts with their attributes, and how they look. Therefore, attributes are good means of describing visual concepts, in a way that both computers and humans understand. In an early work on attributes, Farhadi *et al.* [4] proposed to describe objects using mid-level attributes. In another work [5], they described images based on their attributes, basically a semantic triple ¡object, action, scene¿. Later, Han *et al.* [7] proposed to describe images at different semantic levels. In the recent years, attributes have shown great performance in object recognition [4, 28], action recognition [11, 20] and event detection [21]. Lampert *et al.* [15] predicted unseen objects using a zero-shot learning framework, in which binary attribute representation of the objects were incorporated.

On the other hand, 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.* [12] 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, in which a global function is learned for each single attribute. For instance, a linear relative comparison function is learned in [24], based on RankSVM [9] and a non-linear strategy in [19]. In another work, Datta *et al.* [2] used trained rankers for each facial image feature and formed a global ranking function for attributes.

Through the process of learning the attributes, different types of low-level image features are incorporated. For instance, Parikh *et al.* [24] used 512-dimensional GIST [22] descriptors as image features, while Jayaraman *et al.*[8] used histograms of image features, and reduced their dimensionality using PCA. Other works tried learning attributes through *e.g.*, local and learning [32] or fine-grained comparisons [30]. Yu and Grauman [30] 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 [31], they proposed a local bayesian model to rank images, which are indistinguishable for a given attribute.

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 networks (CNNs) [16] (also known as deep features) could achieve great performance for image-based recognition [13] and object detection [6]. Zhang *et al.* [33] utilized CNNs for classifying binary attributes. In other works, Escorcia *et al.* [3] proposed CCNs with attribute centric nodes within the net for establishing the relationships between visual attributes, Shankar *et al.* [26] proposed a weakly supervised setting on convolutional neural networks, applied for attribute detection, and Khan *et al*. [10] used deep features for describing human attributes and thereafter for action recognition.

CNNs and neural networks have also been extended for learning-to-rank applications. One of the earliest networks for ranking was proposed by Burges*et al*. [1], 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 with pairs of input training feature vectors with differing labels. Then based on how they should have been ranked, the underlying model parameters are updated. This model has been used in different fields for ranking and retrieval applications, *e.g*., for personalized search [27] or content-based image retrieval [29].

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** [14] (faces) and **OSR** [23] (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** [30] (shoes) and **LFW-10** [25] (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 [17] 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] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005. 2

[2] A. Datta, R. Feris, and D. Vaquero. Hierarchical ranking of facial attributes. In *FG*, pages 36–42, 2011. 1

[3] V. Escorcia, J. Carlos Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. 2015. 1

[4] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1

[5] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29. 2010. 1

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 1

[7] Y. Han, Y. Yang, Z. Ma, H. Shen, N. Sebe, and X. Zhou. Image attribute adaptation. *IEEE TMM*, 16(4):1115–1126, 2014. 1

[8] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1629–1636, 2014. 1

[9] T. Joachims. Optimizing search engines using clickthrough data. In *ACM KDD*, pages 133–142, 2002. 1

[10] F. Khan, R. Anwer, J. van de Weijer, M. Felsberg, and J. Laaksonen. Deep semantic pyramids for human attributes and action recognition. In *SCIA*, pages 341–353. 2015. 2

[11] F. Khan, J. van de Weijer, R. Anwer, M. Felsberg, and C. Gatta. Semantic pyramids for gender and action recognition. *IEEE TIP*, 23(8):3633–3645, 2014. 1

[12] A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *ICCV*, pages 297–304, 2013. 1

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105. 2012. 1

[14] 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. 2

[15] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2014. 1

[16] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2*, pages 396–404. 1990. 1

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

[18] 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

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

[20] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, pages 3337–3344, 2011. 1

[21] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney. Video event recognition using concept attributes. In *WACV*, pages 339–346, 2013. 1

[22] 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

[23] 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. 2

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

[25] R. N. Sandeep, Y. Verma, and C. V. Jawahar. Relative parts: Distinctive parts for learning relative attributes. 2014. 1, 2

[26] S. Shankar, V. K. Garg, and R. Cipolla. Deep-carving: Discovering visual attributes by carving deep neural nets. 2015. 1

[27] Y. Song, H. Wang, and X. He. Adapting deep ranknet for personalized search. In *WSDM*, 2014. 2

[28] R. Tao, A. W. Smeulders, and S.-F. Chang. Attributes and categories for generic instance search from one example. In *CVPR*, pages 177–186, 2015. 1

[29] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 157–166, 2014. 2

[30] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 1, 2

[31] A. Yu and K. Grauman. Just noticeable differences in visual attributes. In *ICCV*, 2015. 1

[32] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, volume 2, pages 2126–2136, 2006. 1

[33] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose aligned networks for deep attribute modeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1637–1644, 2014. 1