Learning Visual Representations using Images with Captions

Ariadna Quattoni   Michael Collins   Trevor Darrell
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge MA 02139
ariadna, mcollins, trevor @csail.mit.edu

January 22, 2007

1 Overview

Current methods for learning visual categories work well when a large amount of labeled data is available, but can run into severe difficulties when the number of labeled examples is small. When labeled data is scarce it may be beneficial to use unlabeled data to learn an image representation that is low-dimensional, but nevertheless captures the information required to discriminate between image categories. We describe a method for learning representations from large quantities of unlabeled images which have associated captions; the aim is to learn a representation that aids learning in image classification problems. Experiments show that the method significantly outperforms a fully-supervised baseline model as well as a model that ignores the captions and learns a visual representation by performing PCA on the unlabeled images alone. Our current work concentrates on captions as the source of meta-data, but more generally other types of meta-data could be used (e.g., video sequences with accompanying speech).

2 Background

When few labeled examples are available most current supervised learning methods [9, 3, 4, 7, 5] for image classification may work poorly–for example when a user defines a new category and provides only a few labeled examples. To reach human performance, it is clear that knowledge beyond the supervised training data needs to be leveraged.

There is a large literature on semi-supervised learning approaches, where unlabeled data is used in addition to labeled data. Our work is related to work in multi-task learning, where training data in related tasks is used to aid learning in the problem of interest. Multi-task learning has a relatively long history in machine learning [8, 2, 6, 1], but has only recently been addressed in machine vision. We build on the structure learning approach of Ando and Zhang [1], who describe an algorithm for transfer learning, and suggest the use of auxiliary problems on unlabeled data as a method for constructing related tasks. In some cases unlabeled data may contain useful meta-data that can be used to learn a low-dimensional representation that reflects the semantic content of an image. As one example, large quantities of images with associated natural language captions can be found on the web.

3 Approach

We propose to use the meta-data to induce a representation that reflects an underlying part structure in an existing, high-dimensional visual representation. The new representation groups together synonymous visual features—features that consistently play a similar role across different image classification tasks. Our approach exploits learning from *auxiliary problems* which can be created from images with associated captions. Each auxiliary problem involves taking an image as input, and predicting whether or not a particular content word (e.g, *man*, *official*, or *celebrates*) is in the caption associated with that image. In structural learning, a separate linear classifier is trained for each of the auxiliary problems; manifold learning (e.g., SVD) is then applied to the resulting set of parameter vectors, in essence finding a low-dimensional space which is a good approximation to the space of possible parameter vectors. If features in the high-dimensional space correspond to the same semantic part, their associated classifier parameters (weights) across different auxiliary problems may be correlated in such a way that the basis functions learned by the SVD step collapse these features to a single feature in a new, low-dimensional feature-vector representation.

Topic: visual processing and pattern recognition.

Preference: oral / poster. (Ariadna Quattoni)

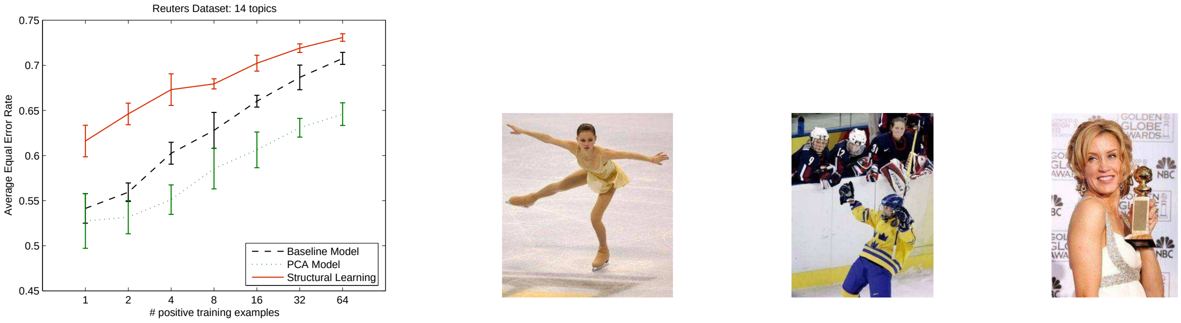

Figure 1: Equal error rates averaged across topics with standard deviation calculated for ten runs for each topic (left). Example images from the Figure Skating, Ice Hockey, and Golden Globes (right).

4 Experiments

In a first set of experiments, we use synthetic data examples to illustrate how the method can uncover latent part structures. A second set of experiments involves classification of news images into different topics. Images on the Reuters website are partitioned into stories which correspond to different topics in the news; each image has a topic label as well as associated caption meta-data. For both experiments we compare a baseline model that uses a bag-of-words SIFT representation of image data, to our method, which replaces the SIFT representation with a new representation that is learned from images with associated captions. In addition, we compare our method to a baseline model that ignores the meta-data and learns a new visual representation by performing PCA on the unlabeled images. Note that our goal is to build classifiers that work on images alone (i.e., images which *do not* have captions), and our experimental set-up reflects this, in that training and test examples for the topic classification tasks include image data only. The experiments show that our method significantly outperforms both baseline models. See http://people.csail.mit.edu/ariadna/TransferLearning for further details on the method and the experiments.

5 Summary

We have described a method for learning visual representations from large quantities of unlabeled images which have associated captions. The method makes use of auxiliary training sets corresponding to different words in the captions, and structural learning, which learns a manifold in parameter space. The induced representations significantly speed up learning of image classifiers applied to topic classification. Our results show that when meta-data labels are suitably related to a target (core) task, the structure learning method can discover feature groupings that speed learning of the target task. Future work includes exploration of automatic determination of relevance between target and auxiliary tasks, and experimental evaluation of the effectiveness of structure learning from more weakly related auxiliary domains.

References

[1] A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] J. Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28:7–39, 1997.

[3] K. Grauman and T. Darrell. The pyramid match kernel:discriminative classification with sets of image features. In *Proceedings fo the International Conference on Computer Vision (ICCV)*, 2005.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of CVPR-2006*, 2006.

[5] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *Proceedings of CVPR-2006*, 2006.

[6] R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine learning*, pages 713–720, 2006.

[7] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 2005.

[8] S. Thrun. Is learning the n-th thing any easier than learning the first? In *In Advances in Neural Information Processing Systems*, 1996.

[9] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Proceedings of CVPR-2006*, 2006.