Induction of Composite Features via Grouping and Composition For Image Classification

Omid Madani
madani@ai.sri.com

Brian Burns
burns@ai.sri.com

There remains a wide gap between the low-level features that are commonly currently extracted from rich multimedia data such as images, and the many semantic classes that humans can recognize. Techniques such as deep belief nets aim to lower this gap [2, 3]. We explore the unsupervised induction of hierarchical composite features in the setting of image classification. In this work, we begin with a vocabulary of 1000 discretized SIFT features [4], and apply grouping (a version of "clustering") and composing ("concatenating") operations as well as statistical filters in order to create composite higher-level features. Grouping is akin to synonymy discovery (features that tend to have similar meaning), and a feature may belong to zero or more (multiple) groups. The discovery of groups is based on a novel use of the analysis of confusions while predicting the features. Our composition operation is currently based on creating new features out of two spatially consecutive features in the image (vertical or horizontal). Both operations apply filters that only keep statistically significant composite features.

Our experiments are performed on a random 100-class subset of ImageNet [1]. We find that the addition of grouping features improves binary-class accuracies (over the use of plain features) by an average of almost 1% (in Max F1), over a simple bag of features representation of images.[1] Composing to create vertical and horizontal bigrams improves the accuracy by another 1.5%. Both improvments are significant at $p \leq 0.01$ level in pairwise sign tests.

The composite features as constructed remain are still close to the level of the raw features. We expect further improvements by repeating this grouping and composition process. We explain the techniques used and the challenges ahead.

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *CVPR*, 2009.

[1] In all the experiments, we use regularized linear SVMs, where the best regularization parameter $C \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1, 5\}$, is picked using a validation set separate from training and test sets.

[2] G. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006.

[3] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition. In *International Conference on Computer Vision*, 2009.

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2004.