Advanced Computer Vision – 1/30 Reviews

Michael Cogswell

1 Using Multiple Segmentations to Discover Objects and their Extents in Image Collections

The authors continue previous work which brought already popular models in text mining to Computer Vision (CV). Instead of just discovering categories from large(ish) sets of unlabeled data they extend previous work to the harder challenge of discovering categories and their segmentations by treating superpixels as documents in Latent Dirichlet Allocation (LDA). In four steps they establish potential segmentations and find the top segments for each topic.

First, for each image, they generate multiple segmentations and extract segments hoping at least a few will correspond to similar segments in other images, thus forming a topic. These multiple segmentations are generated by varying parameters in Shi and Malik's Normalized Graph Cuts algorithm.

Second, 2000 visual words are extracted for the data set and each segment is described as a histogram of those words.

Third, topics are discovered. One segment can be about many topics, each of which is described by a distribution over visual words. By adding these topic variables LDA describes segments as more than a bag of words. Unfortunately, training (maximizing the likelihood of words given knowledge of the English language and beliefs about how topics occur in segments) can only be approximated using Gibbs sampling.

Finally, segments are sorted in a particular topic by comparing the distribution of words within them to the distribution of words given that topic using Kullback-Leibler divergence.

Some thoughts:

- Describing the problem as discovering volumes in a stack of images was particularly intuitive and appealing to me.
- I like seeing research that jumps out of a community's bubble, even if it is only a short hop over to Natural Language Processing (NLP). Doing so presents a writing challenge as the ideas should be as accessible as possible to the wider CV community while focusing on novel contributions. Instead of mapping LDA to the CV world then describing it in NLP terms they should have provided the mapping then described LDA in CV terms.
- Their presentation of results was well done. Lots of useful qualitative visual results were included as well as a quantitative break down of how individual parts contributed to the results. Nowadays, it seems that a larger data set would be more appropriate for unsupervised mining.
- Extensions include using an algorithm other than (in addition to), normalized cuts to generate segmentations; perhaps Dhruv's diverse solutions would work. Mining topics could help the reranker; there may be signal in segments which are sorted higher in certain topic lists. Finally, a more sophisticated version of visual words might give an improvement: finding a good dictionary size (as in the first paper); using features other than (in addition to) SIFT.

2 Foreground Focus: Finding Meaningful Features in Unlabeled Images

Lee and Grauman present a method which refines a set of weights on semi-local features to better discriminate between common visual appearances in a chosen data set. Interestingly, they treat locations of heavily weighted features in a given cluster as foreground, resulting in some strange segmentation(ish) things.

In one iteration feature weights are adjusted to favor features which occur together in the same cluster – to get better features for the current class proposals. In the next iteration they update cluster assignments – to get better class assignments from the more discriminative features.

This seems like some sort of Expectation Maximization, but it is not mentioned. Perhaps someone will formalize that intuition if possible. Also, as my mind does (should not) blindly apply deep learning to everything, replacing their semi-local features with DeCAF may be a good idea.