A Latent Model for Visual Disambiguation of Keyword-based Image Search

Kong-Wah WAN[1]
kongwah@i2r.a-star.edu.sg
Ah-Hwee TAN[2]
asahtan@ntu.edu.sg
Joo-Hwee LIM[1]
joohwee@i2r.a-star.edu.sg
Liang-Tien CHIA[2]
asltchia@ntu.edu.sg
Sujoy ROY[1]
sujoy@i2r.a-star.edu.sg

[1] Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613

[2] School of Computer Engineering
Nanyang Technological University
Singapore

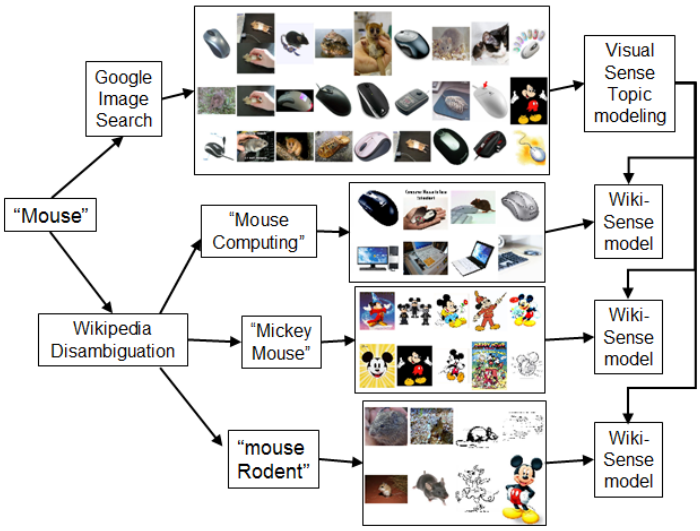

Figure 1: An example of 3 Wiki-word senses for mouse. A visual sense topic model is learnt on the mouse images and incorporated into the wiki-word-sense classifier model.

The problem of polysemy in keyword-based image search arises mainly from the inherent ambiguity in user queries. We propose a latent model based approach that resolves user search ambiguity by allowing sense specific diversity in search results. Given a query keyword and the images retrieved by issuing the query to an image search engine, we first learn a latent visual sense model of these polysemous images. Next, we use Wikipedia to disambiguate the word sense of the original query, and issue these Wiki-senses as new queries to retrieve sense specific images. A sense-specific image classifier is then learnt by combining information from the latent visual sense model.

Figure 1 shows an illustration of our proposed framework. Given a (ambiguous) search keyword, we retrieve a set of images from the web using Google Image Search. Using a probabilistic topic model [2], we attempt to uncover the latent visual sense topics in these polysemous images. We treat each of the visual sense topics as a visual sense underlying the images. Hence, we uncover a latent visual representation of the various senses for each image.

At the same time, we exploit the collaborative structures in Wikipedia Disambiguation to suggest possible word-senses of the keyword [1, 3]. Each Wiki-sense is then used as a new query to pull sense-specific web images using the same search engine. We now construct a probabilistic classifier for each Wiki-sense by incorporating the latent visual topics from the polysemous images.

Given a query keyword P, we treat the images retrieved by each Wiki-sense-specific image queries as the (primary) senses S_i of P, $i = 1,2,\ldots,N_P$, N_P is the number of Wiki-senses of P. For example, in figure 1, the "mouse" keyword has 3 wiki-word-senses: "mouse computing", "Mickey mouse" and "mouse rodent". We define the likelihood of the i^{th} sense S_i given the topic $z = z_j$ as:

$$P(S_i|z=z_j) = \frac{1}{|S_i|}\sum_{a\in S_i} P(a|z=z_j) = \frac{1}{|S_i|}\sum_{a\in S_i} KL(W_a, Z_j) \qquad (1)$$

where W_a is the visual word distribution of image a, Z_j is the visual word distribution of topic z_j, and $KL(\cdot)$ is the Kullback Leibler divergence between the two. For an image d, the model computes the probability of d belonging to the i^{th} sense S_i as:

$$P(S_i|d) = \sum_{j=1}^{K} P(S_i|z=z_j)P(z=z_j|d) \qquad (2)$$

Equation 2 assigns visual sense probabilities to an image according to how similar it is to the sense-specific images. $P(S_i|d)$ provides a way to re-rank the images in the original polysemous order. Images belonging to some sibling sense are given lower probabilities and pushed to the back of the rank list. We call this method VSD-LDA. VSD-LDA extends the method in [4]. The main difference is that [4] is a text-based method, where latent topic discovery is performed on a collection of web text crawled using the polysemous keyword. In contrast, we propose here a VSD-based visual domain topic modelling framework.

Given that we are also retrieving sense-specific images using Wiki-sense queries, an obvious approach is to bootstrap sense-specific classifiers from these images. We shall call this method Sense-Specific SVM. While we expect that returned images from Wiki-sense queries will be more homogeneous, polysemy will still be a problem in learning the sense-specific SVM. In contrast, our VSD-LDA method alleviates these issues by incorporating a latent model of the visual sense of the original keyword. The key idea is that in these images, there is a rich source of information about the various senses (visual or word) of the word, of which Wikipedia merely provides a subset list denoting the primary senses that are more commonly used. These visual senses capture the salient visual characteristics of images associated with the keyword, and offer a more robust visual model than learning on just the Wiki-sense-specific images. Each Wiki-word sense is then represented in the latent space of hidden visual topics for the polysemous keyword.

We evaluate our VSD-LDA approach over the Sense-Specific SVM using a dataset of 17K images collected from a set of 10 polysemous keywords issued to Google Image Search. The keywords are "bank", "bar", "bass", "mouse", "plant", "speaker", "temple", "tiger", "watch" and "window". We evaluate re-ranking retrieval using AUC. Table 1 shows the overall results. For details, please refer to the main paper.

Keyword	Google rank	SVM	VSD-LDA
Total-AUC	28.0172	45.5596	48.1998

Table 1: Area Under Curve (AUC) of all senses of each keyword

[1] Wikipedia:disambiguation. http://en.wikipedia.org/wiki/Dab_page.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] R. Mihalcea. Using wikipedia for automatic word sense disambiguation. In *Proc. NAACL*, 2007.

[4] K. Saenko and T. Darrell. Unsupervised learning of visual sense models for polysemous words. In *Proc. NIPS*, 2008.