Discovering and Distinguishing Multiple Visual Senses for Polysemous Words

Yazhou Yao[†,‡], Jian Zhang[†], Fumin Shen[§,◇*], Wankou Yang[♮], Pu Huang[♯], Zhenmin Tang[‡]
[†]University of Technology Sydney, Australia, [‡]Nanjing University of Science and Technology, China
[§]University of Electronic Science and Technology of China, [♮]Southeast University, China
[♯]Nanjing University of Posts and Telecommunications, China, [◇]Tencent AI Lab, China

Abstract

To reduce the dependence on labeled data, there have been increasing research efforts on learning visual classifiers by exploiting web images. One issue that limits their performance is the problem of polysemy. To solve this problem, in this work, we present a novel framework that solves the problem of polysemy by allowing sense-specific diversity in search results. Specifically, we first discover a list of possible semantic senses to retrieve sense-specific images. Then we merge visual similar semantic senses and prune noises by using the retrieved images. Finally, we train a visual classifier for each selected semantic sense and use the learned sense-specific classifiers to distinguish multiple visual senses. Extensive experiments on classifying images into sense-specific categories and re-ranking search results demonstrate the superiority of our proposed approach.

Introduction

In the past few years, labeled images have played a critical role in high-level image understanding (Guo et al. 2017; Shen et al. 2017). For example, ImageNet (Deng et al. 2009) has acted as one of the most important factors in the recent advance of developing and deploying visual representation learning models (e.g., deep CNN). However, the process of constructing ImageNet is both time-consuming and labour-intensive. To reduce the cost of manual annotation, learning directly from the web images has attracted broad attention (Schroff et al. 2011; Yao et al. 2016; Yao et al. 2017; Guo et al. 2017). Compared to manual-labelled image datasets, web images are a rich and free resource.

For arbitrary categories, the potential training data can be easily obtained from the image search engines like Google or Bing. Unfortunately, due to the error index of image search engine, the precision of the returned images from image search engine is still unsatisfactory. For example, method (Schroff et al. 2011) reports the average precision of the top 1000 images from Google Image Search engine is only 32%. One of the most important reasons for the noisy results is the inherent ambiguity in the user query. As shown in Fig. 1, when we submit the query "mouse" into the Google Image Search engine, the returned results can refer to the animal "mouse", or the electronic product "mouse".

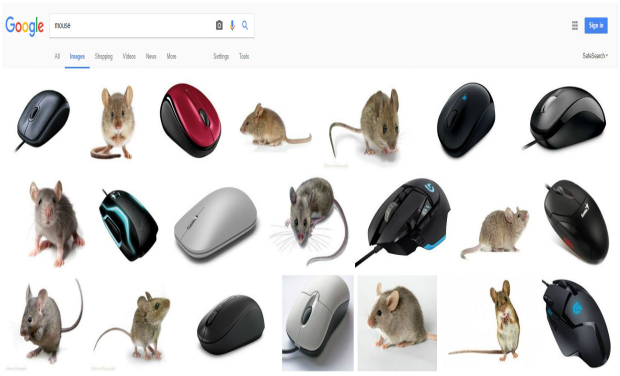

Figure 1: Visual polysemy. For example, the query "mouse" returns multiple visual senses on the first page of results. The retrieved web images suffer from the low precision of any particular visual sense.

Visual polysemy means that a word has several semantic senses that are visually distinct. The traditional way to handle polysemy is to leverage the expert knowledge WordNet (Miller et al. 1995) or Wikipedia (Mihalcea et al. 2007). These human-developed knowledge suffer from the problem of missing information (Prakash et al. 2007). For example, WordNet has a large coverage of NOUN category, it contains very few entities (e.g., organizations, locations). Wikipedia can help to bridge this gap, but a great deal of information is still missing (Chen et al. 2015). What's more, both of the WordNet and Wikipedia have no information about the visual senses and still need manual annotation to bridge the semantic and visual senses.

Since the semantic and visual senses of a given query are highly related, recent works also concentrated on jointly clustering text and images (Loeff et al. 2006; Wan et al. 2009; Saenko et al. 2009). Most of these methods assume that there exists a one-to-one mapping between semantic and visual sense towards to the given query. This assumption is not always true in the practice. To deal with the multiple visual senses, Chen *et al.*(Chen et al. 2015) adopt a one-to-many mapping between semantic and visual spaces. This

*Corresponding author: Fumin Shen

approach can help us to find multiple visual senses from the web but overly depends on the collected web pages. If we can not collect webpages that contain multiple semantic and visual senses for the given query, the effect of this method will be greatly reduced.

Inspired by the situation described above, we seek to automate the process of discovering and distinguishing multiple visual senses for polysemous words through web data. We propose a weakly supervised method that resolves visual polysemy by allowing sense-specific diversity in search results. We take a three-step approach. Firstly, we discover a list of possible semantic senses through Google Books Ngram Corpus (Michel et al. 2011), to retrieve sense-specific images. Secondly, we merge visual similar semantic senses and prune noises by using the retrieved sense-specific images. Thirdly, we learn a visual classifier for each selected semantic sense and use the learned sense-specific classifiers to group and re-rank the polysemous images into its specific senses. To verify the effectiveness, we conducted experiments on datasets CMU-Poly-30 and MIT-ISD to demonstrate the superiority of our proposed approach. The main contributions of this work are summarized as follows:

- We propose a novel approach for discovering and distinguishing multiple visual senses for polysemous words without explicit supervision.
- Our work shows substantial improvement over existing weakly supervised state-of-the-art methods.
- Our work can be used as a pre-step before directly learning from the web, which help to choose appropriate visual senses for sense-specific images collection, thereby improving the efficiency of learning from the web.

Related Work

Automatically discovering and distinguishing visual senses for polysemous words from the web is an extremely difficult problem. Several authors proposed to clean the retrieved images and learn visual classifiers, although none have specifically addressed the problem of polysemy. Method (Fergus et al. 2004) proposed the use of visual classifiers learned from Google Image Search engine to re-rank the images based on the visual consistency. Subsequent method (Yao et al. 2016; Shen et al. 2017; Yao et al. 2017) has employed similar removing mechanisms to automatically construct clean image datasets for training visual classifiers. However, these methods are category-independent and do not learn which words are predictive of a specific sense.

To discover multiple semantic and visual senses for polysemous words, previous works have also concentrated on clustering both of the text and image sources on the web (Loeff et al. 2006; Saenko et al. 2009; Wan et al. 2009; Chen et al. 2015). Method ISD (Loeff et al. 2006) involves two major steps: (1) extracting and weighting text features from the web pages, visual features from the retrieved images, (2) running spectral clustering on both of the text features and visual features to derive the multiple semantic senses. Method VSD (Wan et al. 2009) and ULVSM (Saenko et al. 2009) proposed a latent model to learn multiple visual senses from a large collection of unlabeled web data, but rely on Wikipedia and WordNet's sense inventory respectively. Method SDCIT (Chen et al. 2015) proposed a one-to-many mapping between the text-based feature space and image-based visual space to discover multiple semantic and visual senses of a Noun Phrase. However, clustering presents a scalability issue for this problem. Since our images are sourced directly from the web and have no bounding boxes, every image creates millions of data points, the majority of which are outliers.

The Proposed Approach

The inspiration for our work comes from the fact that web images indexed by a polysemous word can be in a rich diversity. Our main idea of solving the problem of polysemy by allowing sense specific diversity in search results. Specifically, our proposed framework consists of three major steps: 1) discovering a list of possible semantic senses, to retrieve sense-specific images, 2) merging visual similar semantic senses and pruning noises, 3) training a visual classifier for each selected semantic sense and distinguishing the multiple visual senses for the given polysemous word.

Discovering Possible Semantic Senses

Inspired by recent work (Michel et al. 2011), we can use Google Books Ngram English 2012 Corpus to discover the possible semantic senses for modifying the given polysemous word. Our motivation is to find all the possible semantic senses the human race has ever written down in books. Following (Lin et al. 2012) (see section 4.3), we specifically use the dependency gram data with parts-of-speech (POS) for possible semantic senses discovering. For example, given a word (e.g., "mouse") and its corresponding POS tag (e.g., 'mighty, ADJ'), we find all its occurrences annotated with POS tag within the dependency gram data. Of all the ngram dependencies retrieved for the given word, we choose those whose modifiers are tagged as NOUN, VERB, ADJECTIVE and ADVERB as the possible semantic senses. We use these possible semantic senses to retrieve sense-specific web images from the image search engine.

Merging and Pruning Semantic Senses

Among the list of possible semantic senses, some of them are sharing the visually similar distributions (e.g., "jerry mouse", "mininie mouse" and "cartoon mouse"). To avoid training separate models for visually similar semantic senses, and to pool valuable training data across them, we need to merge and sample these visually similar semantic senses. In addition, not all the discovered semantic senses are useful, some noise may also be included (e.g., "figure mouse" and "flying mouse"). To better distinguish multiple visual senses, we need to prune these noises.

Merging visual similar semantic senses For each possible semantic sense, we use the top N images from Google Image Search Engine to represent its visual distribution. We denote the visual similarity space of all discovered semantic senses by a graph $G = \{V, W\}$, where each node represents

a semantic sense and each edge represents the visual similarity between two nodes.

Each node has a score S_i which corresponds to the quality of its classifier. Specifically, we assume the top N images are positive instances (based on the fact that the top few images returned from image search engine tend to be positive), then randomly split these images into a training set and validation set $I_i = \{I_i^t, I_i^v\}$. We gather a random pool of negative images and split them into a training set and validation set $\overline{I} = \{\overline{I}^t, \overline{I}^v\}$. We train a linear support vector machine (SVM) classifier f_i with I_i^t and $\overline{I}^t$ using the 4096 dimensional deep features (based on AlexNet (Krizhevsky et al. 2012)). We then use $\{I_i^v, \overline{I}^v\}$ as validation images to calculate the classification results. We set the score S_i equal to the classification results on its own validation set $\{I_i^v, \overline{I}^v\}$. The edge weights $W_{i,j}$ correspond to the visual similarity between two nodes, and is measured by the score of the *i*th node classifier f_i on the *j*th node validation set $\{I_j^v, \overline{I}^v\}$.

Then the problem of merging visually similar semantic senses can be formulated as sampling a representative subset of space $v \subseteq V$ which maximizes the quality of the subset:

$$\begin{aligned} \max_{v} \quad & \sum_{i \in V} S_i \cdot \phi(i, v) \\ \text{s.t.} \quad & |v| \leqslant k \end{aligned} \tag{1}$$

where k is the number of semantic senses for the given word and ϕ is a soft coverage function that implicitly ensure the diversity of representative subset:

$$\phi(i, v) = \begin{cases} 1 & i \in v \\ 1 - \prod_{j \in v}(1 - W_{i,j}) & i \notin v \end{cases} \tag{2}$$

Similar to recent work (Batra et al. 2012), our formulation is to find a subset of representive space v which can cover the space of variance within the space V. Since our objective function is sub-modular, we can get a constant approximation of the optimal solution. We use an iterative mechanism for discovering the most representative subset. Particularly, we add one semantic sense i at each iteration by maximizing the current space:

$$\arg\max_{i} S(v \cup i) - S(v). \tag{3}$$

By setting the cost of adding semantic sense in v to a large value, each new semantic sense can be merged to its closest member in v.

Pruning noisy semantic senses After we merge the visually similar semantic senses, we get a relatively few discrete sense. Among these discrete senses, there are still some noisy semantic senses need to be removed to distinguish multiple effective visual senses for polysemous words. Our basic idea is noisy semantic senses have no specific visual patterns (e.g., "figure mouse"). Thus, we can prune these noises from the perspective of visual consistency.

We represent each discrete semantic sense as a "bag" and the retrieved images therein as "instances". In particular, we represent each semantic sense G_I with the compound feature $\delta_{f,k}$ of its first k positive images:

$$\delta_{f,k}(G_I) = \frac{1}{k} \sum_{x_i \in \Phi^*_{f,k}(G_I)} x_i \tag{4}$$

with

$$\Phi^*_{f,k}(G_I) = \underset{\Phi \subseteq G_I, |\Phi| = k}{\arg\max} \sum_{x_i \in \Phi} f(x_i). \tag{5}$$

The images in $\Phi^*_{f,k}(G_I)$ are referred to the top k positive instances of G_I according to the SVM classifier f_i. Since the closer of images in G_I from the bag center, the higher probability of these images to be relevant to the bag. The assignment of relatively heavier weights to these images would increase the accuracy of classifying bag G_I to be positive or negative, then increase the efficiency of pruning noisy semantic senses. Following (Carneiro et al. 2007), we assume the form of weighting function is

$$\rho_i = [1 + \exp(\alpha \log d(x_i) + \beta)]^{-1}. \tag{6}$$

$d(x_i)$ represents the Euclidean distance of image x_i from the bag center, $\alpha \in \mathbb{R}_{++}$ and β are scaling and offset parameters which can be determined by cross-validation. Then the representation of (4) for semantic sense G_I can be generalized to a weighted compound feature:

$$\delta_{f,k}(G_I) = \delta(X, h^*) = \frac{Xh^*}{\rho^\top h^*} \tag{7}$$

with

$$h^* = \underset{h \in \mathrm{H}}{\arg\max} \; f(\frac{Xh}{\rho^\top h}) \;\; \text{s.t.} \sum_i h_i = k. \tag{8}$$

$X = [x_1, x_2, x_3.., x_i] \in \mathbb{R}^{D \times i}$ is a matrix whose columns are the instances of bag G_I, $\rho = [\rho_1, \rho_2, \rho_3...\rho_i]^\top \in \mathbb{R}^i_{++}$ are the vectors of weights, and $h^* \in \mathrm{H} = \{0,1\}^i \setminus \{0\}$ ($\sum_i h_i = k$) is an indicator function for the first k positive instances of bag G_I. Then the classifying rule of semantic sense G_I to be selected or pruned is:

$$\begin{aligned} f_{\mathrm{w}}(X) = & \max_{h \in \mathrm{H}} \mathrm{w}^\top \delta(X, h) \\ & \sum_i h_i = k \end{aligned} \tag{9}$$

where $\mathrm{w} \in \mathbb{R}^D$ is the vector of classifying coefficients, $\delta(X, h) \in \mathbb{R}^D$ is the feature vector of (7), h is a vector of latent variables and H is the hypothesis space $\{0,1\}^i \setminus \{0\}$. In order to solve the classifying rule of (9), we need to solve the below following problem:

$$\max_{h \in \mathrm{H}} \frac{\mathrm{w}^\top Xh}{\rho^T h} \;\; \text{s.t.} \sum_i h_i = k. \tag{10}$$

This is an integer linear-fractional programming problem. Since $\rho \in \mathbb{R}^i_{++}$, (10) is identical to the relaxed problem:

$$\max_{h \in \lambda^i} \frac{\mathrm{w}^\top Xh}{\rho^\top h} \;\; \text{s.t.} \sum_i h_i = k. \tag{11}$$

Algorithm 1 Concave-Convex Procedure for solving (13)

1: Initialize w with SVM by setting $h = \mathbf{1} \in \mathbb{R}^i$;
2: Compute a convex upper bound using the current model for the second term of (13);
3: Minimize this upper bound by solving a structural SVM problem via the proximal bundle method (Kiwiel et al. 1990);
4: Repeat step 2 and step 3 until convergence.

where $\lambda^i = [0,1]^i$ is a unit box in $\mathbb{R}^i$. (11) is a linear-fractional programming problem and can be reduced to a linear programming problem of $i+1$ variables and $i+2$ constraints (Boyd et al. 2004).

Given a training set $\{G_I, Y_I\}_{I=1}^N$, the learning problem is to determine the parameter vector w in (9). This is a latent SVM problem:

$$\min_{\mathrm{w}} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{I=1}^{N}\max\left(0, 1 - Y_I f_{\mathrm{w}}\left(X_{G_I}\right)\right). \quad (12)$$

In this work, we take the concave-convex procedure (CCCP) algorithm (Yuille et al. 2003) to solve (12). We rewrite the objective of (12) as two convex functions:

$$\min_{\mathrm{w}} \left[\frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{I\in D_N}\max\left(0, 1 + f_{\mathrm{w}}\left(X_{G_I}\right)\right) + C\sum_{I\in D_P}\max\left(f_{\mathrm{w}}\left(X_{G_I}\right), 1\right)\right] - \left[C\sum_{I\in D_P} f_{\mathrm{w}}\left(X_{G_I}\right)\right] \quad (13)$$

where D_P and D_N are positive and negative training sets respectively. The detailed solutions of the CCCP algorithm for (13) are described in Algorithm 1. Lastly, we obtain the pruning rule as (9) to remove noisy semantic senses which have no specific visual senses.

Distinguishing Visual Senses

After pruning the noisy semantic senses, we set the rest as the final selected semantic senses. Due to the error index of image search engine, even we retrieve the sense-specific images, some instance-level noise may also be included. The last step of our approach is to prune these instance-level noisy images and train visual classifiers for distinguishing multiple visual senses. Particularly, we train a classifier for each selected semantic sense.

By treating each selected semantic sense as a "bag" and the retrieved images therein as "instances", we formulate a multi-instance learning problem by selecting a subset of images from each bag to learn the classifier for the selected semantic senses. Since the precision of images returned from Google Image Search engine tend to have a relatively high accuracy, we define each bag at least has a portion of δ positive instances.

We denote each instance as x_i with its label $y_i \in \{\pm 1\}$, where *i*=1,...,*n*. We also denote the label of each bag as $Y_I \in \{\pm 1\}$. The element-wise product between two matrices $\mathbf{P}$ and $\mathbf{Q}$ is represented by $\mathbf{P} \odot \mathbf{Q}$. Moreover, we define the identity matrix as $\mathbf{I}$ and $\mathbf{0}$, $\mathbf{1} \in \Re^n$ denote the column vectors of all zeros and ones, respectively. The inequality $\mathbf{u} = [u_1, u_2 ... u_n]^\top \geq \mathbf{0}$ means that $u_i \geq 0$ for *i*=1,...,*n*.

The decision function is assumed in the form of $f(x) = \mathrm{w}^\top\varphi(x) + b$ and it will be used to prune instance-level noisy images. We employ the formulation of Lagrangian SVM, in which the square bias penalty b^2 and the square hinge loss for each instance are used in the objective function. Then the decision function can be learned by minimizing the following structural risk functional:

$$\min_{\mathbf{y},\mathrm{w},b,\rho,\varepsilon_i} \quad \frac{1}{2}\left(\|\mathrm{w}\|^2 + b^2 + C\sum_{i=1}^{n}\varepsilon_i^2\right) - \rho \quad (14)$$

$$\text{s.t.} \quad y_i(\mathrm{w}^\top\varphi(x_i) + b) \geq \rho - \varepsilon_i, i = 1, ...n. \quad (15)$$

$$\sum_{i:x_i\in G_I}\frac{y_i+1}{2} \geq \delta\left|G_I\right| \quad for \quad Y_I = 1, \qquad y_i = -1 \quad for \quad Y_I = -1 \quad (16)$$

where φ is a mapping function that maps *x* from the original space into a high dimensional space $\varphi(x)$, $C > 0$ is a regularization parameter and ε_i values are slack variables. The margin separation is defined as $\rho/\|\mathrm{w}\|$. $\mathbf{y} = [y_1 ... y_n]^\top$ means the vector of instance labels, $\lambda = \{\mathbf{y} | y_i \in \{\pm 1\}\}$ and $\mathbf{y}$ satisfies constraint (16).

We employ the cutting-plane algorithm (Kelley et al. 1960) to solve the optimization problem (14). Finally, we can derive the decision function for the selected semantic sense as:

$$f(x) = \sum_{i:\alpha_i\neq 0}\alpha_i\widetilde{y}_i\widetilde{k}(x, x_i) \quad (17)$$

where $\widetilde{y}_i = \sum_{t:\mathbf{y}^t\in\lambda} u_t y_i^t$ and $\widetilde{k}(x, x_i) = k(x, x_i) + 1$. The decision function will be used to prune instance-level noisy images in each selected semantic sense. In addition, it will also be leveraged to distinguish different visual senses.

Experiments

To verify the effectiveness of our proposed approach, in this section, we first conduct experiments on the task of classifying images into sense specific categories. Then we quantitative analyze the search results re-ranking ability of our approach with several baseline methods.

Classifying Sense-specific Images

Experimental setting We follow the setting in baseline methods (Chen et al. 2015; Loeff et al. 2006; Wan et al. 2009; Saenko et al. 2009) and exploit web images as the training set, human-labelled images as the testing set. Instead of using co-clustering on web text and images, we use general corpus information and web images to discover and distinguish multiple visual senses for polysemous words. Particularly, we evaluate the performance on the dataset:

1) CMU-Poly-30 (Chen et al. 2015). The CMU-Poly-30 dataset consists of 30 polysemy categories. Each category contains a varying number of images.

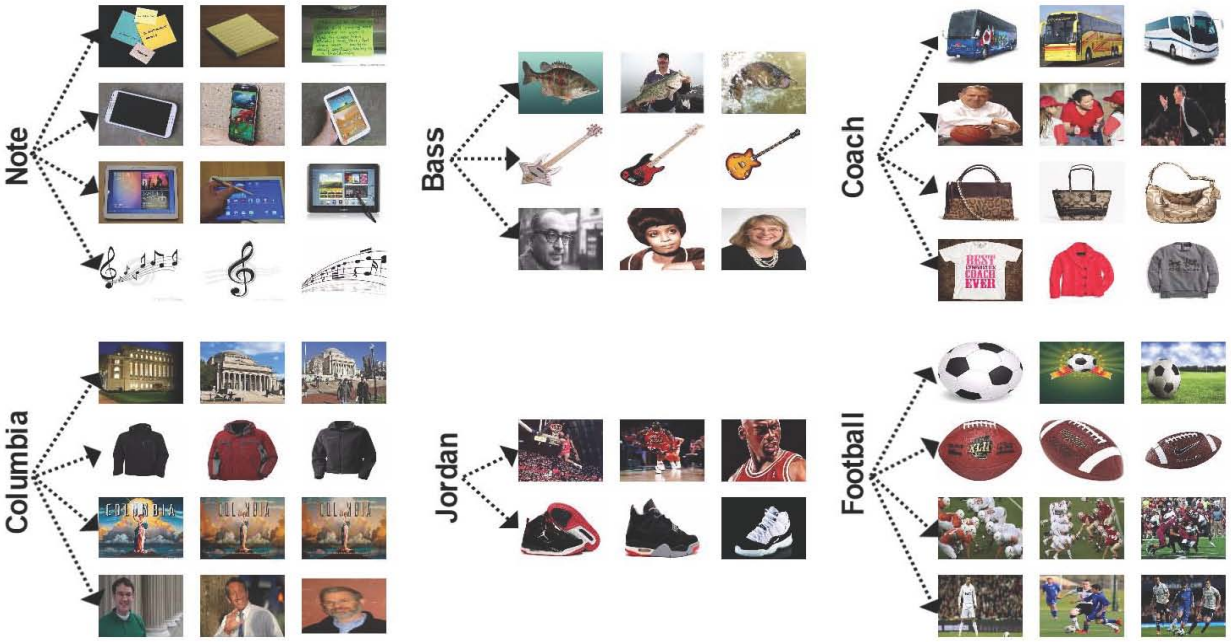

Figure 2: Examples of multiple visual senses discovered by our proposed approach. For example, our approach automatically discovers and distinguishes four senses for "$Note$": notes, galaxy note, note tablet and music note. For "$Bass$", it discovers multiple visual senses of: bass fish, bass guitar and Mr./Mrs. Bass etc.

2) MIT-ISD (Saenko et al. 2009). The MIT-ISD dataset contains 5 categories. Each of which has three sizes. We are concerned with the "keyword" based size as it has the ground truth.

For each category, we first discover the possible semantic senses by searching in the Google Books Ngram Corpus. Then we retrieve the top $N = 100$ images from the Google Image Search engine for each discovered semantic sense. We assume the retrieved images as the positive instances (in spite of the fact that noisy images might be included). We randomly split the retrieved 100 images for each semantic sense into a training set and validation set $I_i = \{I_i^t = 50, I_i^v = 50\}$. We gather a random pool of negative images and split them into a training set and validation set $\overline{I} = \{\overline{I}^t = 50, \overline{I}^v = 50\}$. We train the SVM classifier f_i and calculate the score S_i using the validation set. The edge weights $W_{i,j}$ are obtained by calculating the score of the *i*th node classifier f_i on the *j*th node validation set $\{I_j^v, \overline{I}^v\}$. We merge the visually similar semantic senses and sample the representative subset of space by setting the cost to be 0.3 (obtained by cross validation).

In order to prune noisy semantic senses, we retrieve the top 500 images for each selected semantic sense. We then use the previously trained classifier f_i to select the most positive $k = 100$ images from the rest 450 images (the training data and testing data have no duplicates). We represent the selected semantic sense G_I with the compound feature $\delta_{f,k}$ of the most positive 100 images. There are multiple methods for learning the weighting function (e.g., logistic regression or cross-validation), here we follow (Carneiro et al. 2007) and use cross-validation to learn the weighting function. To this end, we label $M = 500$ positive bags and 500 negative bags. Labeling work only needs to be carried out once to learn the weighting function and the weighted bag classification rule (9). The learned weighted bag classification rule (9) will also be used to prune noisy bags (corresponding to noisy semantic senses) which have no specific visual senses.

After pruning the noisy semantic senses, we set the rest as the final selected semantic senses. For each selected semantic sense, we collect the training data (500 images) from the image search engine. We take the MIL based method to handle instance-level noisy images and select the positive training data, to train the visual classifier. The negative training data is drawn from a "background" category, which in our case is the union of all other categories that we are asked to classify. The visual feature in our experiment is 4096 dimensional deep features (based on AlexNet (Krizhevsky et al. 2012)).

Baselines In order to quantify the performance of our proposed approach, we compare the sense-specific image classification ability of our approach with four weakly supervised baseline methods including ISD (Loeff et al. 2006), VSD (Wan et al. 2009), ULVSM (Saenko et al. 2009) and SDCIT (Chen et al. 2015). For all of the baseline methods, we adopt the same parameter configuration as described in their original works.

Experimental results Fig. 2 presents the examples of multiple visual senses discovered by our proposed approach on CMU-Poly-30 dataset. Fig. 3 and Fig. 4 demonstrate the

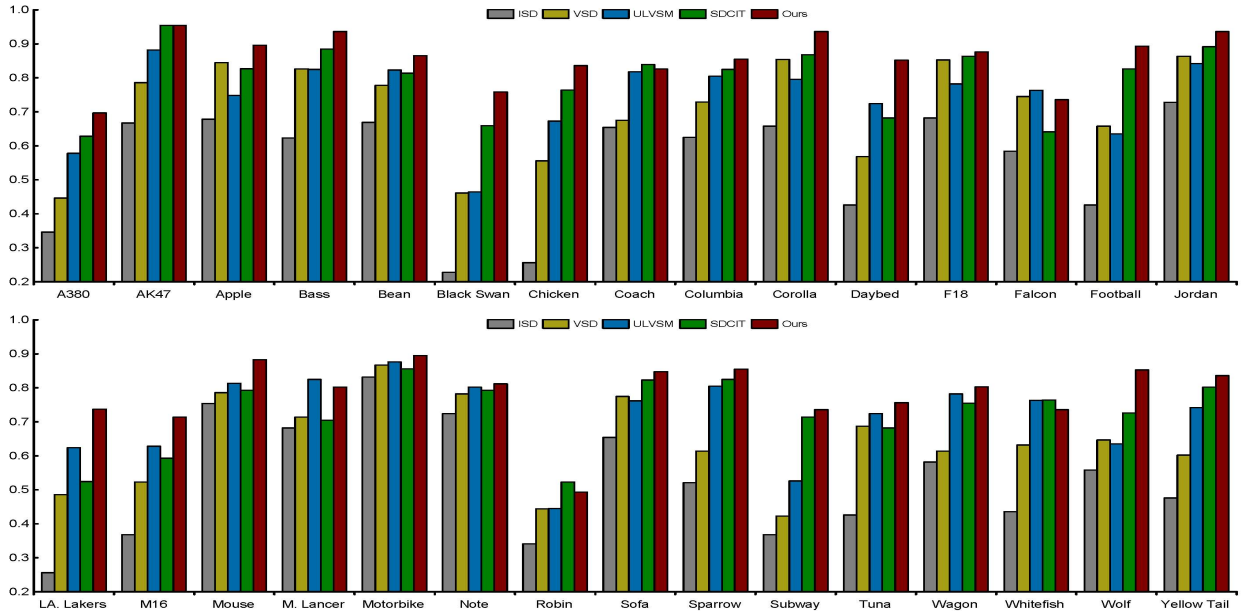

Figure 3: The detailed performance comparison of classification accuracy over 30 categories on the CMU-Poly-30 dataset.

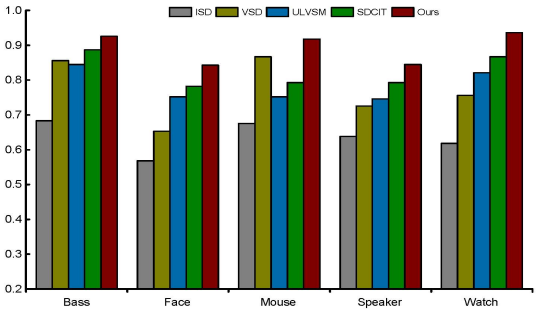

Figure 4: The detailed performance comparison of classification accuracy over 5 categories on the MIT-ISD dataset.

detailed performance comparison of classification accuracy on the CMU-Poly-30 and MIT-ISD dataset respectively. Table 1 shows the average performance comparison of classification accuracy on the CMU-Poly-30 and MIT-ISD dataset.

From Fig. 3 and Fig. 4, we achieved the best results in 26 categories on the CMU-Poly-30 dataset. In the 5 categories of dataset MIT-ISD, we obtained the best results in all 5 categories. By observing Table 1, the best average performance is achieved by our approach, which produces significant improvements over method ISD, VSD, ULVSM and SDCIT. One possible explanation is that the automatically generated sense-specific terms by our approach could return relatively high-precision web images. Meanwhile, the MIL based method can handle the few noise in the training data and train a robust classifier.

It is interesting to note in Fig. 2, our proposed approach not only discovers and distinguishes the sense of "notes" for "$Note$", but also "galaxy note", "note tablet" and "music note". For "$Bass$", in addition to "bass fish" and "bass guitar", our approach also discovers and distinguishes the sense of "Mr./Mrs. Bass". Compared to method VSD and ULVSM which discovers possible semantic senses through Wikipedia or WordNet, our proposed approach that adopts Google Books Ngram Corpus to discover possible semantic senses is much more exhaustive and general. Method ISD and SDCIT which uses webpages can discover multiple semantic senses, but overly depends on the collected webpages. For example, method ISD fails to collect webpages that contain enough semantic senses and visual senses for the given query, it can be seen that in Table 1, the performance of this method is greatly reduced.

Table 1: The average performance comparison of classification accuracy on the CMU-Poly-30 and MIT-ISD dataset.

Method	**Dataset**	
	CMU-Poly-30	MIT-ISD
ISD	0.555	0.634
VSD	0.728	0.786
ULVSM	0.772	0.803
SDCIT	0.839	0.853
Ours	**0.884**	**0.897**

Re-ranking Search Results

Experimental setting We compare the search results re-ranking ability of our approach with four weakly supervised baseline methods ISD (Loeff et al. 2006), VSD (Wan et al. 2009), ULVSM (Saenko et al. 2009), VCL (Divvala et al. 2014), and SDCIT (Chen et al. 2015). We collect the top 500 images from Google Image Search engine for semantically ambiguous words: "bass" and "mouse". We perform

Table 2: Web images for polysemy terms were annotated manually. For each term, the number of annotated images, the semantic senses, the visual senses and their distributions are provided, with core semantic senses marked in boldface.

Query (#Annot. images)	Semantic senses	Visual senses	Numbers of images	Coverage
Bass (349)	1. **bass fish**	fish	159	45.6%
	2. **bass guitar**	musical instrument	154	44.1%
	3. Mr./ Mrs. Bass	people	20	5.7%
	Noise	unrelated	16	4.6%
Mouse (251)	1. **computer mouse**	electronic product	125	49.8%
	2. **little mouse**	animal	81	32.3%
	3. carton mouse	cartoon role	26	10.4%
	Noise	unrelated	19	7.5%

Table 3: Area Under Curve (AUC) of all senses for "bass" and "mouse".

Method	**Semantic senses**						**Average**
	bass fish	bass guitar	M. Bass	Computer mouse	little mouse	carton mouse	
ISD	0.453	0.526	0.243	0.614	0.536	0.218	0.432
VSD	0.547	0.538	0.239	0.684	0.652	0.226	0.481
ULVSM	0.526	0.615	0.326	0.732	0.735	0.314	0.541
VCL	0.623	0.658	0.413	0.753	0.785	0.336	0.595
SDCIT	0.658	**0.773**	0.386	0.815	0.845	0.337	0.636
Ours	**0.713**	0.736	**0.572**	**0.834**	**0.873**	**0.434**	**0.694**

a cleanup step for broken links, webpages, end up with 349 and 251 images for "bass" and "mouse" respectively.

These images were annotated with one of the several semantic senses by one of the authors. The annotator tried to resist name influence, and make judgments based just on the image. For each query, 2 core semantic senses were distinguished from inspecting the data. The detailed information for these retrieved images is summarized in Table 2.

We now evaluate how well the 4 compared methods and our method can re-rank the retrieved images. For each query, the sense-specific classifiers are trained on the sense-specific web images. Particularly, we use the previously trained sense-specific classifiers. Retrieved images are then re-ranked by moving the negatively-classified images down to the last rank. For an image d, we compute the probability $P(S_i|d)$ of image d belonging to the ith sense S_i and rank the corresponding images according to the probability of each sense S.

Experimental results Following (Wan et al. 2009), we evaluate the retrieval performance by computing the Area Under Curve (AUC) of all senses for "bass" and "mouse". The results are shown in Table 3.

From Table 2, we observe that there are only 4.6% and 7.5% true noise in the retrieved images for "bass" and "mouse" respectively. Most of the retrieved images are different forms of visual senses for the given query. This indicates that we should firstly discover the multiple visual senses for the given query. So that we can choose appropriate visual senses as needed to carry out sense-specific images collection. By doing this, we can greatly improve the efficiency of collecting web images, thereby improving the efficiency of learning from the web.

By observing Table 3, we achieve the best average performance which consistent with the results of sense-specific image classification. The reason can be explained by the generated sense-specific terms of our approach. Compared to method ISD, VSD, ULVSM, VCL and SDCIT, using our sense-specific terms to retrieve images can return high precision sense-specific images, thereby can help us to train sense-specific classifiers for re-ranking the search results.

Conclusions

In this work, we focused on one important yet often ignored problem: we argue that the current poor performance of some classification models learned from the web is due to the visual polysemy. We solved the problem of polysemy by allowing sense-specific diversity in search results. Specifically, we presented a new framework for discovering and distinguishing multiple visual senses for polysemous words. Our work could be used as a pre-step before directly learning from the web, which helped to choose appropriate visual senses for sense-specific images collection and thereby improve the efficiency of learning from the web. Compared to existing methods, our proposed method can not only figures out the right sense, but also generates the right mapping between semantic and visual senses. We verified the effectiveness of our approach on the tasks of sense-specific image classification and search results re-ranking. The experimental results demonstrated the superiority of our proposed approach over existing weakly supervised state-of-the-art approaches.

Acknowledgments

This research is supported by the National Science Foundation of China (Grant No. 61473154, 61502081, 61473086, 61503195, 61773210) and the China Postdoctoral Science Foundation (Grant No. 2016M600433).

References

[Deng et al. 2009] Deng, J., Dong, W., Socher, R., Li, L., and Li, Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

[Batra et al. 2012] Batra, D., Yadollahpour, P., Guzman, A., and Shakhnarovich G. 2012. Diverse m-best solutions in markov random fields. In *ECCV*, 1–16.

[Guo et al. 2017] Guo, Y., Ding, G., Han, J., Gao, Y. 2017. Zero-Shot Learning With Transferred Samples. In *IEEE TIP*, 3277-3290.

[Boyd et al. 2004] Boyd, S., and Vandenberghe, L. 2004. Convex Optimization. In *Cambridge University Press*.

[Carneiro et al. 2007] Carneiro, G., Chan, A., Moreno, P., and Vasconcelos, N. 2007. Supervised learning of semantic classes for image annotation and retrieval. In *IEEE TPAMI*, 29(3): 1462–1475.

[Shen et al. 2017] Shen, F., Mu, Y., Yang, Y., Liu, W., Liu, L., Song, J., Shen, H. 2017. Classification by Retrieval: Binarizing Data and Classifier. In *SIGIR*, 595-604.

[Chen et al. 2015] Chen, X., Ritter, A., Gupta, A., and Mitchell, T. 2015. Sense discovery via co-clustering on images and text. In *IEEE CVPR*, 5298–5306.

[Dalal et al. 2005] Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 886–893.

[Yao et al. 2017] Yao, Y., Zhang, J., Shen, F., Hua, X., Xu, J., and Tang, Z. 2017. Exploiting Web Images for Dataset Construction: A Domain Robust Approach. In *IEEE TMM*, 19(8): 1771–1784.

[Kiwiel et al. 1990] Kiwiel, K. C. 1990. Proximity control in bundle methods for convex non differentiable minimization. In *Mathematical Programming*, 46(1): 105–122.

[Lindeberg et al. 2012] Lindeberg, T. 2012. Scale invariant feature transform. In *Scholarpedia*, 7(5): 104-112.

[Fergus et al. 2004] Fergus, R., Perona, P., and Zisserman, A. 2004. A visual category filter for google images. In *ECCV*, 242–256.

[Berg et al. 2006] Berg, T., and Forsyth, D. 2006. Animals on the web. In *IEEE CVPR*, 1463–1470.

[Li et al. 2009] Li, Y., Tsang, I., Kwok, J., and Zhou, Z.H. 2009. Tighter and convex maximum margin clustering. In *AISTATS*, 344–351.

[Mansur et al. 2008] Mansur, A., and Kuno, Y. 2008. Improving recognition through object sub-categorization. In *ISVC*, 851–859.

[Michel et al. 2011] Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., and Orwant, J. 2011. Quantitative analysis of culture using millions of digitized books. In *Science*, 331(6014):176–182.

[Lin et al. 2012] Lin, Y., Michel, J.-B., Aiden, E.-L., Orwant, J., Brockman, W., and Petrov, S. 2012. Syntactic annotations for the google books ngram corpus. In *ACL*, 169–174.

[Loeff et al. 2006] Loeff, N., Alm, C. O., and Forsyth, D. A. 2006. Discriminating image senses by clustering with multimodal features. In *ACL*, 547–554.

[Yao et al. 2016] Yao, Y., Hua, X., Shen, F., Zhang, J., and Tang, Z. 2016. A Domain Robust Approach for Image Dataset Construction. In *ACM MM*, 212–216.

[Miller et al. 1995] Miller, G. A. 1995. Wordnet: a lexical database for english. In *Communications of the ACM*, 38(11): 39–41.

[Prakash et al. 2007] Prakash, R. S. S., and Ng, A. Y. 2007. Learning to merge word senses. In *EMNLP-CoNLL*, 1005–1015.

[Kelley et al. 1960] Kelley, J. E. 1960. The cutting-plane method for solving convex programs. In *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712.

[Saenko et al. 2009] Saenko, K., and Darrell, T. 2009. Unsupervised learning of visual sense models for polysemous words. In *NIPS*, 1393–1400.

[Schroff et al. 2011] Schroff, F., Criminisi, A., and Zisserman, A. 2011. Harvesting image databases from the web. In *IEEE TPAMI*, 33(4): 754–766.

[Fan et al. 2005] Fan, R.-E., Chen, P.-H., and Lin, C.-J. 2005. Working set selection using second order information for training support vector machines. In *JMLR*, 6(11): 1889–1918.

[Hoffman et al. 2012] Hoffman, J., Kulis, B., Darrell, T., and Saenko, K. 2012. Discovering latent domains for multisource domain adaptation. In *ECCV*, 702–715.

[Krizhevsky et al. 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. 2012. "Imagenet classification with deep convolutional neural networks," In *NIPS*, 1097–1105.

[Wan et al. 2009] Wan, K.-W., Tan, A.-H., Lim, J.-H., Chia, L.-T., and Roy, S. 2009. A latent model for visual disambiguation of keyword-based image search. In *BMVC*, 2–7.

[Mihalcea et al. 2007] Mihalcea, R. 2007. Using wikipedia for automatic word sense disambiguation. In *HLT-NAACL*, 196–203.

[Yuille et al. 2003] Yuille, A., and Rangarajan, A. 2003. The concave-convex procedure. In *Neural Computation*, 15(4): 915–936.

[Yao et al. 2016] Yao, Y., Zhang, J., Shen, F., Hua, X., Xu, J., and Tang, Z. 2016. Automatic image dataset construction with multiple textual metadata. In *IEEE ICME*, 1–6.

[Shen et al. 2017] Shen, F., Gao, X., Liu, L., Yang, Y., Shen, H. 2017. Deep Asymmetric Pairwise Hashing. In *ACM MM*.

[Guo et al. 2017] Guo, Y., Ding, G., Liu, L., Han, J., Shao, L. 2017. Learning to Hash With Optimized Anchor Embedding for Scalable Retrieval. In *IEEE TIP*, 1344-1354.

[Divvala et al. 2014] Divvala, S.K., Farhadi, A., Guestrin, C. 2014. Learning Everything about Anything: Webly-Supervised Visual Concept Learning. In *CVPR*, 3270–3277.