Dynamically Visual Disambiguation of Keyword-based Image Search

Yazhou Yao[1,2] , **Zeren Sun**[1] , **Fumin Shen**[3*] **Li Liu**[2] , **Limin Wang**[4] , **Fan Zhu**[2] , **Lizhong Ding**[2] , **Gangshan Wu**[4] and **Ling Shao**[2]
[1]Nanjing University of Science and Technology, Nanjing, China
[2]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[3]University of Electronic Science and Technology of China, Chengdu, China
[4]Nanjing University, Nanjing, China
{yaoyazhou, zerensun, fumin.shen, liuli1213, lmwang.nju, fanzhu1987, lizhongding}@gmail.com,
gswu@nju.edu.cn, ling.shao@ieee.org

Abstract

Due to the high cost of manual annotation, learning directly from the web has attracted broad attention. One issue that limits their performance is the problem of visual polysemy. To address this issue, we present an adaptive multi-model framework that resolves polysemy by visual disambiguation. Compared to existing methods, the primary advantage of our approach lies in that our approach can adapt to the dynamic changes in the search results. Our proposed framework consists of two major steps: we first discover and dynamically select the text queries according to the image search results, then we employ the proposed saliency-guided deep multi-instance learning network to remove outliers and learn classification models for visual disambiguation. Extensive experiments demonstrate the superiority of our proposed approach.

1 Introduction

In the past few years, labeled image datasets have played a critical role in high-level image understanding [Min, 2016; Zhang, 2017; Xie, 2019; Shu, 2018; Hu, 2017; Hua, 2017]. However, the process of constructing manually labeled datasets is both time-consuming and labor-intensive [Deng, 2009]. To reduce the time and labor cost of manual annotation, learning directly from the web images has attracted more and more attention [Chen, 2013; Yao, 2018; Shen, 2019; Zhang, 2016; Yao, 2019; Liu, 2019; Tang, 2018; Hua, 2017; Shu, 2015]. Compared to manually-labeled image datasets, web images are a rich and free resource. For arbitrary categories, potential training data can be easily obtained from an image search engine [Zhang, 2016]. Unfortunately, the precision of returned images from an image search engine is still unsatisfactory.

One of the most important reasons for the noisy results is the problem of visual polysemy. As shown in Fig. 1, visual polysemy means that a word has multiple semantic senses that are visually distinct. For example, the keyword "coach" can refer to multiple text semantics and visual senses (*e.g.,* the "bus", the "handbag", the sports "instructor", or the "company"). This is commonly referred as word-sense disambiguation in Natural Language Processing [Wan, 2009].

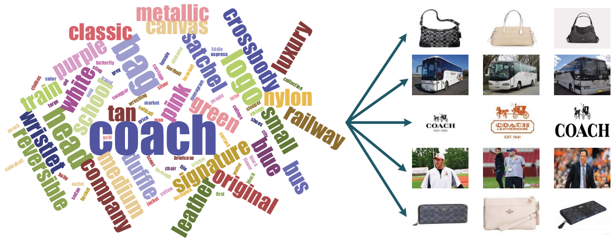


Figure 1: Visual polysemy. The keyword "coach" can refer to multiple text semantics, resulting in images with various visual senses in the image search results.

Word-sense disambiguation is a top-down process arising from ambiguities in natural language. The text semantics of a word are robust and relatively static, and we can easily look them up from a dictionary resource such as WordNet or Wikipedia. However, visual-sense disambiguation is a *data-driven dynamic problem* which is specific to image collection. For the same keyword, the visual senses of images returned from the image search engine may be different at different time periods. For example, the keyword "apple" might have mainly referred to the fruit before the company was founded.

Since the text semantics and visual senses of a given keyword are highly related, recent works also concentrated on combining text and image features [Chen, 2015; Wan, 2009; Loeff, 2006; Yao, 2018]. Most of these methods assume that there exists a one-to-one mapping between semantic and visual sense for a given keyword. However, this assumption is not always true in practice. For example, while there are two predominant text semantics of the word "apple", there exist multiple visual senses due to appearance variation (green vs. red apples). To deal with the multiple visual senses, the method in [Chen, 2015] adopts a one-to-many mapping between text semantics and visual senses. This approach can help us discover multiple visual senses from the web but overly depends on the collected webpages. The effect of this approach will be greatly reduced if we can't collect webpages

[*] Corresponding Author

that contain enough text semantics and visual senses [Shen, 2019; Yao, 2018].

Instead of relying on human-developed resources, we focus on automatically solving the visual disambiguation in an unsupervised way. Unlike the common unsupervised paradigm which jointly clusters text features and image features to solve the visual disambiguation, we present a novel framework that resolves the visual disambiguation by dynamically matching candidate text queries with retrieved images of the given keyword. Compared to human-developed and clustering-based methods, our approach can adapt to the dynamic changes in the search results. Our proposed framework includes two major steps: we first discover and dynamically select the text queries according to the keyword-based image search results, then we employ the proposed saliency-guided deep multi-instance learning (MIL) network to remove outliers and learn classification models for visual disambiguation. To verify the effectiveness of our proposed approach, we conduct extensive experiments on visual polysemy datasets CMU-Poly-30 and MIT-ISD to demonstrate the superiority of our approach. The main contributions are:

1) Our proposed framework can adapt to the dynamic changes in search results and do visual disambiguation accordingly. Our approach has a better time adaptation ability.

2) We propose a saliency-guided deep MIL network to remove outliers and jointly learn the classification models for visual disambiguation. Compared to existing approaches, our proposed network has achieved the-state-of-the-art performance.

3) Our work can be used as a pre-step before directly learning from the web, which helps to choose appropriate visual senses for sense-specific images collection, thereby improving the efficiency of learning from the web.

2 The Proposed Approach

As shown in Fig. 2 and Fig. 3, our proposed approach consists of two major steps. The following subsections describe the details of our proposed approach.

2.1 Discovering and Selecting Text Queries

Inspired by recent works [Yao, 2018; Divvala, 2014], untagged corpora Google Books [Lin, 2012] can be used to discover candidate text queries for modifying given keyword. Following the work in [Lin, 2012] (see section 4.3), we discover the candidate text queries by using n-gram dependencies whose modifiers are tagged as NOUN.

The image search results are dynamically changing, not all the candidate text queries have enough images in the search results representing their visual senses. Therefore, we can dynamically purify the candidate text queries by matching them with the retrieved images. Suppose the given keyword is kw, then we discover $E(kw)$ candidate text queries through Google Books. We collect the top K images for given keyword kw. We perform a clean-up step for broken links and set the rest images $I(kw)$ as the selected images for given keyword kw (*e.g.,* "apple"). In addition, we retrieve the top $I(tq) = 5$ images for each candidate text query tq (*e.g.,* "Apple laptop"). A text query $tq \in E(kw)$ is expected to frequently appear in $I(kw)$. To well obtain the visual senses of the images, some subset images which all have tq are required to contain visual similar content. To this end, $E(kw)$ can be selected in the following way.

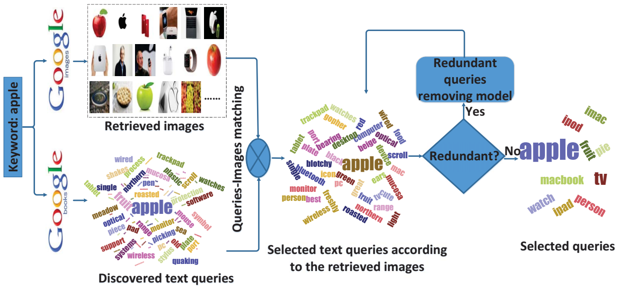


Figure 2: Framework of discovering and dynamically selecting text queries. The input is a keyword. We first discover a list of candidate text queries and retrieve the top images for the given keyword. Then we dynamically purify the candidate text queries according to the retrieved images. We remove the redundant and set the rest as selected text queries.

For each image $x_i \in I(tq)$, all images in $I(kw)$ are matched with x_i on the basis of the visual similarity. In our work, the visual features and similarity measure methods in [Wang, 2014] are leveraged. We set $\vartheta_i(I)$ to be the number of images in $I(kw)$ which can match with x_i. The overall number of a candidate text query tq matching with the search results is its accumulated numbers over all the $I(tq)$ images:

$$\vartheta(tq) = \sum_{i=1}^{I(tq)} \vartheta_i(I). \tag{1}$$

A large $\vartheta(tq)$ indicates that tq matches in a good number of images in $I(kw)$. When tq only presents in a few images or images involving tq are visually different, $\vartheta(tq)$ will be set to be zero. Accordingly, when tq contains a big accumulated value $\vartheta(tq)$, we can notice that lots of images within $I(kw)$ contain tq and the images involving tq have similar visual senses. These N text queries with the highest numbers are chosen as the selected candidate text queries $E(kw)$ for the given keyword kw.

Among the list of selected candidate text queries, some of them share visual similar distributions (*e.g.,* "Apple MacBook" and "apple laptop"). To lessen the computing costs, these text queries which increase the discriminative power of the semantic space are kept and others are removed. To calculate the visual similarity between two text queries, half data in both text queries are used to learn a binary SVM classifier to do classification on the other half data. We believe that the two text queries are not similar if we can easily separate the testing data. Assume we obtain N candidate text queries from the above step. We split the retrieved images of text query m into two groups, I_m^t and I_m^v. To calculate the distinctness $D(m,n)$ between m and n, we train a binary SVM by using I_m^t and I_n^t. We then obtain the probability of image in I_m^v belonging to the class m with the learned SVM classifier. Suppose the average score over I_m^v is $\bar{\rho_m}$. Similarly, we can also obtain the average score $\bar{\rho_n}$ over I_n^v. Then $D(m,n)$ can be calculated by:

$$D(m,n) = \chi((\bar{\rho_m} + \bar{\rho_n})/2) \tag{2}$$

where χ is a monotonically increasing function. In this work, we define

$$\chi(\bar{\rho}) = 1 - e^{-\beta(\bar{\rho}-\alpha)} \tag{3}$$

in which the parameters α and β are two constants. When the value of $(\bar{\rho_m} + \bar{\rho_n})/2$ goes below the threshold α, $\chi(\bar{\rho})$ decreases with a fast speed to penalize pair-wisely similar text queries. In our work, the value of α and β are set to be 0.6 and 30 respectively.

Finally, we select a set of text queries from the N candidates. The selected text queries are most relevant to the given keyword kw. We define the relevance in (1). Meanwhile, to characterize the visual distributions of the given keyword, the selected text queries are required to dissimilar with each other from the visual relevance perspective. The distinctiveness can be calculated through matrix D in (2). We can solve the following optimization problem to satisfy the two criteria.

γ is used to indicate text query n is selected or removed. Specifically, we set $\gamma_n = 1$ when selected and $\gamma_n = 0$ when removed. We can estimate the value of γ by solving:

$$\arg \max_{\gamma \in \{0,1\}^N} \{\lambda \phi_\gamma + \gamma^N D_\gamma\} \tag{4}$$

Let tq_n be the text query of keyword kw. $\phi = (\vartheta(tq_1), \vartheta(tq_2), ..., \vartheta(tq_N))$, where $\vartheta(tq_n)$ is defined in (1). λ is the scaling factor. Due to the integer quadratic programming is NP hard, γ is relaxed to be in $\mathbb{R}^T$ and we choose text query n whose $\gamma_n \geqslant 0.5$ as the final selected text query.

2.2 Saliency-guided Deep MIL Model

Due to the error indexing of image search engine, even we retrieve the top few sense-specific images, some noise may still be included [Shen, 2019; Yao, 2018]. As shown in Fig. 3, our model consists of two stream networks SGN and DMIL. SGN is to localize object for generating instance. DMIL is to encode the discriminative features for learning the deep classification models to remove outliers and perform visual disambiguation.

Different from existing methods which attempt to follow a multi-instance assumption, where its object proposals can be regarded as one "instance" sets and each image can be treated as one "bag", our approach treats each selected text query as a "bag" and each image therein as one "instance". The main reason for this is that our images come from the web and may contain noise. If we treat each web image as a "bag", the generated proposals ("instances") by existing methods RPN can't always satisfy such a condition: object lies in at least one of the proposals. However, when we treat each image returned from the image search engine as one "instance", and each selected text query as one "bag", then it becomes natural to formulate outliers removal as a multi-instance learning problem.

The selected text queries are leveraged to collect sense-specific images from the image search engine. To reduce the interference of noisy background objects in web images, we propose to use a saliency extraction network (SGN) to localize the discriminative regions and generate the instance for the web image. Specifically, we follow the work in [Zhou, 2016] to model this process by leveraging global average pooling (GAP) to produce the saliency map. The feature maps of the last convolutional layer with weights were summed to generate the saliency map for each image. Finally, we conduct binarization operation on the saliency map with a adaptive threshold, which is obtained through OTSU algorithm. We leverage the bounding box that covers the largest connected area as the discriminative region of object. For a given image *I*, the value of spatial location (*x*, *y*) in saliency map for category *c* is defined as follows:

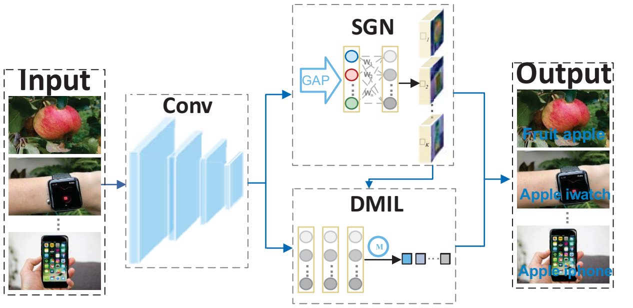


Figure 3: The framework of saliency-guided deep MIL model. Our framework includes two stream networks SGN and DMIL. SGN is to localize object and generate an instance for the web image. DMIL is to encode the discriminative features for learning the deep MIL classification models to remove outliers and perform visual disambiguation.

$$M_c(x, y) = \sum_u w_u^c f_u(x, y) \tag{5}$$

where $M_c(x, y)$ directly indicates the importance of activation at spatial location (*x*, *y*) leading to the classification of an image to category *c*. $f_u(x, y)$ denotes the activation of neuron *u* in the last convolutional layer at spatial location (*x*, *y*), and w_u^c denotes the weight that corresponding to category *c* for neuron *u*. Instead of treating the whole image as one instance, we use the generated bounding box result as the instance.

In traditional supervised learning paradigm, training samples are given as pairs $\{(x_i, y_i)\}$, where $x_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{-1, 1\}$ is the label. However, in MIL, data are organized as bags $\{\mathbf{X}_i\}$. Each bag contains a number of instances $\{x_{i,j}\}$. Labels $\{\mathbf{Y}_i\}$ are only available for the bag. The labels of instances $\{y_{i,j}\}$ are unknown. Considering the recent advances achieved by deep learning, in this work, we propose to exploit deep CNN as our architecture for learning visual representation with multi-instance learning. Our structure is based on VGG-16 and we redesign the last hidden layer for MIL. For a given training image *x*, we set the output of the last fully connected layer $fc_{15} \in \mathbb{R}^m$ as high-level features of the input image. Followed by a softmax layer, fc_{15} is transformed into a probability distribution $\rho \in \mathbb{R}^m$ for objects belonging to the m text queries. Cross-entropy is taken to measure the prediction loss of the network. Specifically, we have

$$L = -\sum_i t_i \log(\rho_i) \quad \text{where} \quad \rho_i = \frac{\exp(h_i)}{\sum_i \exp(h_i)}, i = 1.., m. \tag{6}$$

We can calculate the gradients of the deep CNN through back-propagation

$$\frac{\partial L}{\partial h_i} = \rho_i - t_i, \tag{7}$$

where

$$t = \{t_i | \sum_{i=1}^{m} t_i = 1, \;\; t_i \in \{0, 1\}, i = 1, ..., m\} \tag{8}$$

represents the true label of the sample x. To learn multiple instances as a bag of samples, we incorporate deep representation with MIL and name it as DMIL. Assume a bag $\{x_j | j = 1, ..., n\}$ contains n instances and the label of the bag is $t = \{t_i | t_i \in \{0, 1\}, i = 1, ...m\}$; DMIL extracts representations of the bag: $h = \{h_{ij}\} \in R^{m \times n}$, in which each column is the representation of an instance. The aggregated representation of the bag for MIL is:

$$\tilde{h_i} = f(h_{i1}, ..., h_{in}) \tag{9}$$

where function f can be $\max_j(h_{ij})$, $\text{avg}_j(h_{ij})$, or $\log[1 + \sum_j \exp(h_{ij})]$. For this work, we use the $\max(\cdot)$ layer. In the ablation studies, we show experiments with these possible choices. Then we can represent the visual distribution of the bag and the loss L as:

$$\rho_i = \frac{\exp(\tilde{h_i})}{\sum_i \exp(\tilde{h_i})}, i = 1, ..., m. \tag{10}$$

and

$$L = -\sum_i t_i \log(\rho_i) \tag{11}$$

respectively. To minimize the loss function of DMIL, we employ stochastic gradient descent (SGD) for optimization. The gradient can be calculated via back propagation [Rumelhart, 1986]:

$$\frac{\partial L}{\partial \tilde{h_i}} = \rho_i - t_i \text{ and } \frac{\partial \tilde{h_i}}{\partial h_{ij}} = \begin{cases} 1, & h_{ij} = \tilde{h_i} \\ 0, & \text{else} \end{cases}. \tag{12}$$

For the task of disambiguating the keyword-based image search results, we first employ SGN to generate the saliency map for localizing the discriminative region and generating the "instance" of the image. Then the proposed DMIL is to encode the discriminative features for learning deep models to remove outliers and perform visual disambiguation.

3 Visual Disambiguation Experiments

3.1 Datasets and Evaluation Metric

Two widely used polysemy datasets CMU-Polysemy-30 [Chen, 2015] and MIT-ISD [Saenko, 2009] are employed to validate the proposed framework. Specifically, we set the images corresponding to various keywords in CMU-Poly-30 and MIT-ISD as the results of keyword-based image search. We follow the setting in baselines [Chen, 2015; Yao, 2018] and exploit web images as the training set, human-labeled images in CMU-Polysemy-30 and MIT-ISD as the testing set. Average Classification Accuracy (**ACA**) is adopted as the evaluation metric. If there is no special statement, the image features are 4096-dimensional deep features based on VGG-16 model.

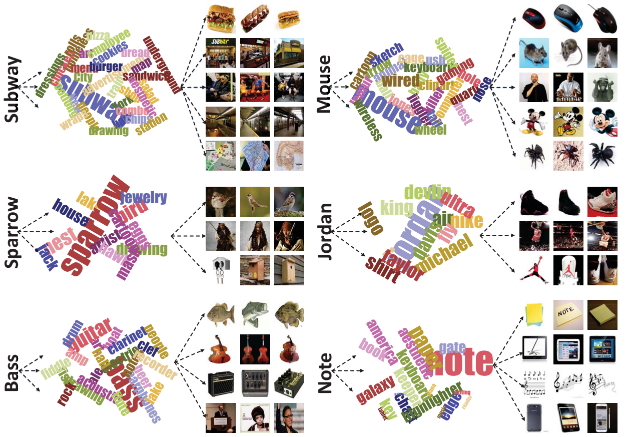


Figure 4: A snapshot of multiple text queries discovered from Google Books and visual senses disambiguated from the CMU-Poly-30 dataset by our proposed framework. For example, our proposed method automatically discovers and disambiguates five senses for "$Subway$": subway sandwich, subway store, subway people, subway station and subway map. For "$Mouse$", it discovers multiple visual senses of the computer mouse, mouse animal, mouse man, and cartoon mouse *etc.*

3.2 Implementation Details and Parameters

For each keyword, we first discover the candidate text queries by searching in the Google Books. We set the corresponding images in CMU-Polysemy-30 and MIT-ISD as the results of keyword-based image search. Then we retrieve the top $I(tq)$ images for each candidate text query. The value of $I(tq)$ is selected from $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. We dynamically purify the candidate text queries by matching them with the results of keyword-based image search. Specifically, we select the top N text queries with the highest numbers. N is selected from $\{10, 20, 30, 40, 50, 60\}$. For removing redundancy and selecting representative text queries, we retrieve the top 100 images for the selected candidate text queries and assume the retrieved images are positive instances (in spite of the fact that noisy images might be included). The collected 100 images for each selected text query were randomly split into a training set and testing set (*e.g.,* $I_m = \{I_m^t = 50, I_m^v = 50\}$ and $I_n = \{I_n^t = 50, I_n^v = 50\}$). We train a linear SVM classifier with I_m^t and I_n^t for classifying I_m^v and I_n^v to obtain the value of $\bar{\rho_m}$ and $\bar{\rho_n}$. We then get the distinctness $D(m, n)$ by calculating (2) and remove redundant queries by solving (4). α is selected from $\{0.2, 0.4, 0.5, 0.6, 0.8\}$ and β is selected from $\{10, 20, 30, 40, 50\}$ in (2). γ_n is set $\gamma_n \geqslant 0.5$ in (4).

The structure of SGN is based on VGG-16. To obtain a higher spatial resolution, we remove the layers after conv5_3 and get a mapping resolution of 14 × 14. Then we add a convolutional layer of size 3 × 3, stride 1, pad 1 with 1024 neurons, followed by a global average pooling (GAP) layer and a softmax layer. SGN is pre-trained on the 1.3 million images of ImageNet dataset [Deng, 2009] and then fine-tuned on the collected web images. The number of neurons in the softmax layer is set as the number of selected text queries. The structure of DMIL is also based on VGG-16. We remove the last hidden layer and use the $\max(\cdot)$ layer instead. The initial parameters of the modified version of the model are

	Method	**Dataset**	
		CMU-Poly-30	MIT-ISD
§	VSD [Wan, 2009]	0.728	0.786
	ULVSM [Saenko, 2009]	0.772	0.803
	WSDP [Barnard, 2005]	0.791	0.743
	NEIL [Chen, 2013]	0.741	0.705
	ConceptMap [Golge, 2014]	0.726	0.758
	VSCN [Qiu, 2013]	0.802	0.783
$	ISD [Loeff, 2006]	0.554	0.634
	IWSD [Lucchi, 2012]	0.643	0.725
	SDCIT [Chen, 2015]	0.839	0.853
	VSDE [Gella, 2016]	0.747	0.763
	LEAN [Divvala, 2014]	0.827	0.814
	DRID [Zhang, 2016]	0.846	0.805
	DDPW [Shen, 2019]	0.884	0.897
¶	SG-DMIL (Ours)	**0.925**	**0.938**

$: combination of text and image based methods

§ : image-based methods ¶ : our proposed approach

Table 1: Visual disambiguation results (ACA) on two evaluated datasets CMU-Poly-30 and MIT-ISD.

inherited from the pre-trained VGG-16 model. During training, we leveraged "instances" generated by SGN and set the selected text queries as "bags" to fine-tune the model. DMIL is trained for 100 epochs with an initial learning rate selected from [0.0001, 0.002] (which is robust). In order to generate test "bags", we solely sampled images from the CMU-Polysemy-30 and MIT-ISD datasets.

3.3 Baselines

The method in [Shen, 2019] reproduced nearly all leading methods on the CMU-Polysemy-30 and MIT-ISD dataset. Specifically, we compare our approach with two groups of baselines: image-based methods and the combination of text and image based methods.

3.4 Results and Analysis

Fig. 4 shows a snapshot of multiple text queries discovered from Google Books and visual senses disambiguated from the CMU-Poly-30 dataset by our proposed framework. It should be noted that for some keywords, CMU-Poly-30 dataset only annotates one or two visual senses. However, our proposed approach successfully discovers and distinguishes more visual senses. For example, for keyword "bass" in CMU-Poly-30 dataset, only "bass fish" and "bass guitar" are annotated. Our approach additionally discovered two other visual senses "bass amp" and "Mr./Miss Bass". This is mainly due to our approach can dynamically select text queries based on image search results.

To leverage the ground truth labels in CMU-Poly-30 and fairly compare with other baseline methods, we remove the text queries discovering and selecting procedure and directly use the annotated labels in the dataset to collect web images. Then we leverage the proposed saliency-guided deep MIL to remove outliers and train classification models for visual disambiguation. Table 1 presents the ACA results on the CMU-Poly-30 and MIT-ISD dataset. By observing Table 1, our proposed approach achieves the-state-of-the-art ACA performance on CMU-Poly-30 and MIT-ISD dataset, which produces significant improvements over image-based methods, and the combination of text and image based methods. One possible explanation is that our proposed saliency-guided deep MIL can effectively remove the outlier images from the image search results and train robust classification models for visual disambiguation.

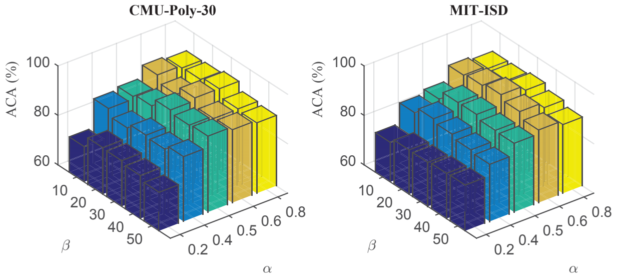


Figure 5: The ACA performance of the interaction between pairs of parameters α and β.

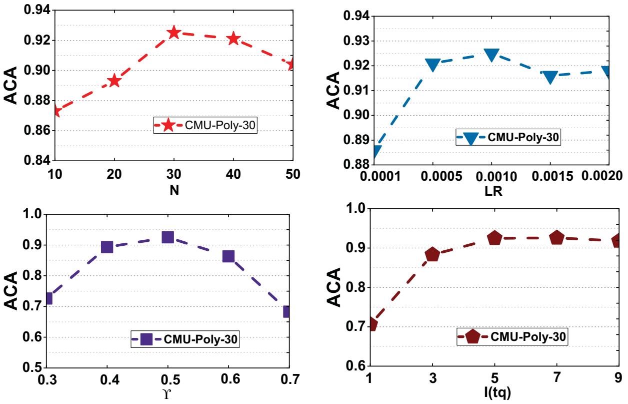


Figure 6: The parameter sensitivities of N, LR, γ, and $I(tq)$ w.r.t. ACA in CMU-Poly-30 dataset.

4 Ablation Studies

4.1 Coefficients in Proposed Framework

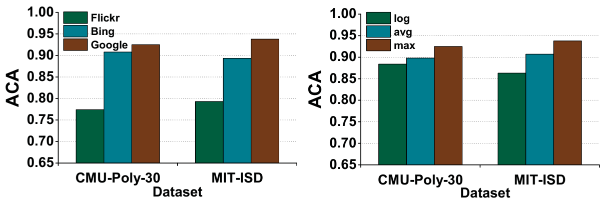
For the coefficients analysis, we mainly concern the parameters α, β, γ, N, and $I(tq)$ in selecting text queries and learning rate (LR) in DMIL. Specifically, we analyze the interaction between pairs of parameters α and β in Eq (3). For other parameters, we analyze the sensitivities by a graphic per parameter. As shown in Fig. 5, the changing tendency of ACA w.r.t. (α, β), overall, is stable and consistent. Fig. 6 presents the parameter sensitivities of N, LR, γ, and $I(tq)$ w.r.t. ACA in CMU-Poly-30 dataset.

4.2 Pre-step before Learning from the Web

Our work can be used as a pre-step before directly learning from the web. To verify this statement, we collected the top 100 web images from the Google Image Search Engine by using the labels in CUB-200-2011 dataset [Wah, 2011]. Then we employed the proposed approach to choose appropriate visual senses and purify the outliers. The outputs are a set of relatively clean web images. We leveraged the relatively clean web images as the training set to perform one of the most popular weakly supervised fine-grained algorithms Bilinear [Lin, 2015] on CUB-200-2011 [Wah, 2011] testing set.

Training data	Algorithm	Accuracy
Original web	Bilinear	0.718
Clean web	Bilinear	0.752
CUB training	Bilinear	0.841
Clean web + CUB training	Bilinear	0.863

Table 2: Fine-grained visual recognition results on CUB-200-2011 testing set.



(a) (b)

Figure 7: (a) Impact of different domains. (b) Impact of different hidden layers.

The results are shown in Table 2. By observing Table 2, we can observe that our proposed approach greatly improves the baseline accuracy.

4.3 Influence of Different Domains

To analyze the influence of using different domain web images for visual disambiguation, we collected web images for selected text queries from the Google Image Search Engine, the Bing Image Search Engine, and Flickr respectively. As shown in Fig 7 (a), the performance of web images coming from Flickr is much lower than from the Google Image Search Engine and the Bing Image Search Engine. The performance of web images coming from the Google Image Search Engine is a little better than from the Bing Image Search Engine.

4.4 Influence of Different Hidden Layer

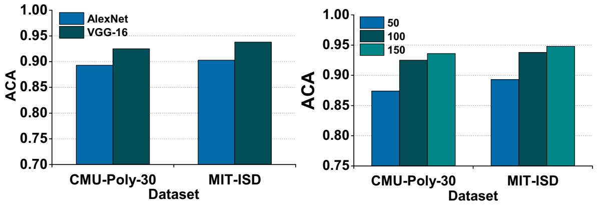
The choice of hidden layer is of critical importance in our proposed saliency-guided deep MIL network. As mentioned in Section 3.2, the $\max(\cdot)$, avg$(\cdot)$, and $\log(\cdot)$ refer to $\max_j(h_{ij})$, avg$_j(h_{ij})$, and $\log[1+\sum_j \exp(h_{ij})]$, respectively. From Fig 7 (b), we can notice that the straightforward $\max(\cdot)$ layer obtains the best ACA performance.

4.5 Are Deeper Models Helpful?

It is well known that the CNN model architecture has a critical impact on object recognition performance. We investigated this issue by replacing VGG-16 with a shallower architecture AlexNet in the saliency-guided deep MIL model and compared the results. As shown in Fig 8 (a), using a deeper model (VGG-16) was better than using shallower models (AlexNet), as expected. In particular, the VGG model was more effective for localizing the objects from the images.

4.6 Are More Web Images Helpful?

Data scale has a large impact on web-supervised learning. We investigated this impact by incrementally increasing or decreasing the number of web images used for each text query. Specifically, we choose $\{50, 100, 150\}$ images from the web for each selected text query. As shown in Fig 8 (b), in general, the performance of ACA improved steadily with the use of more training samples.



(a) (b)

Figure 8: (a) Impact of different CNN architectures. (b) Impact of different training samples.

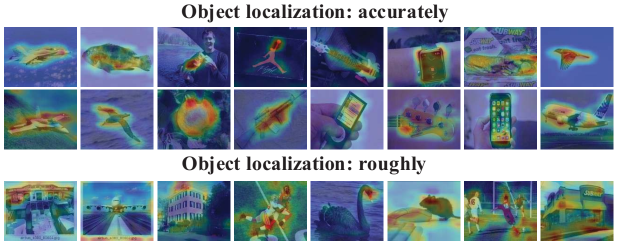


Figure 9: Visualization of object locating via saliency map.

4.7 Visualization

Fig. 9 visualizes the object locating via saliency map. By observing Fig. 9, we can find the SGN can well locate the object for the web image. For some images, although SGN cannot accurately locate the exact location where it is located, the rough region of SGN locating still contains the location of the object.

5 Conclusions

In this work, we focused on one important yet often ignored problem: we argue that the current poor performance of some models learned from the web images is due to the inherent ambiguity in user queries. We solved this problem by visual disambiguation in search results. Our work could be used as a pre-step before directly learning from the web images, which helps to choose appropriate visual senses for images collection and thereby improve the efficiency of learning from the web. Compared to existing methods, our proposed approach can 1) figure out the right visual senses, and 2) adapt to the dynamic changes in the search results. Extensive experiments demonstrated the superiority of our proposed approach.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Project 61502081, Sichuan Science and Technology Program (No.2019YFG0003, 2018GZDZX0032).

References

[Deng, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database. *CVPR*, 248–255, 2009.

[Yao, 2018] Yazhou Yao, Jian Zhang, Fumin Shen, Wankou Yang, Pu Huang and Zhenmin Tang, Discovering and Distinguishing Multiple Visual Senses for Polysemous Words. *AAAI*, 523–530, 2018.

[Shen, 2019] Yazhou Yao, Fumin Shen, Jian Zhang, Li Liu, Zhenmin Tang and Ling Shao, Extracting Multiple Visual Senses for Web Learning. *TMM*, 21(1): 184–196, 2019.

[Chen, 2013] Xinlei Chen, Abhinav Gupta, "Neil: Extracting visual knowledge from web data." *ICCV*, 1409–1416, 2013.

[Chen, 2015] Xinlei Chen, Alan Ritter, Abhinav Gupta, Tom Mitchell, Sense discovery via co-clustering on images and text. *CVPR*, 5298–5306, 2015.

[Shu, 2015] Xiangbo Shu, Guojun Qi, Jinhui Tang, Jingdong Wang, Weakly-Shared Deep Transfer Networks for Heterogeneous-Domain Knowledge Propagation. *ACM MM*, 35–44, 2015.

[Wan, 2009] Kong-Wah Wan, Ah-Hwee Tan, Joo-Hwee Lim, Liangtien Chia, Sujoy Roy, A latent model for visual disambiguation of keyword-based image search. *BMVC*, 163–170, 2009.

[Min, 2016] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, "Being a Supercook: Joint Food Attributes and Multimodal Content Modeling for Recipe Retrieval and Exploration" *TMM*, 19(5): 1100–1113, 2017.

[Zhang, 2017] Guo-Sen Xie, Xu-Yao Zhang, Shuicheng Yan, Cheng-Lin Liu, "SDE: A novel selective, discriminative and equalizing feature representation for visual recognition" *IJCV*, 124(2): 145–168, 2017.

[Xie, 2019] Guo-Sen Xie, Li Liu, Xiao-Bo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao, "Attentive Region Embedding Network for Zero-shot Learning" *CVPR*, 1245–1253, 2019.

[Shu, 2018] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Wei Liu, Jian Yang, "Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition" *ArXiv:1811.00270*, 2018.

[Hu, 2017] Guosheng Hu, Xiaojiang Peng, Yongxin Yang, Timothy Hospedales, Jakob Verbeek, "Frankenstein: Learning deep face representations using small data" *TIP*, 27(1): 293–303, 2017.

[Hua, 2017] Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, "Attribute-enhanced face recognition with neural tensor fusion networks" *ICCV*, 3744–3753, 2017.

[Liu, 2019] Yazhou Yao, Fumin Shen, Jian Zhang, Li Liu, Zhenmin Tang and Ling Shao, Extracting Privileged Information for Enhancing Classifier Learning. *TIP*, 28(1): 436–450, 2019.

[Tang, 2018] Yazhou Yao, Jian Zhang, Fumin Shen, Wankou Yang, Xiansheng Hua and Zhenmin Tang, Extracting Privileged Information from Untagged Corpora for Classifier Learning. *IJCAI*, 1085–1091, 2018.

[Loeff, 2006] Nicolas Loeff, David Forsyth, Discriminating image senses by clustering with multimodal features. *ACL*, 547–554, 2006.

[Saenko, 2009] Kate Saenko, Trevor Darrell, Unsupervised learning of visual sense models for polysemous words. *NIPS*, 1393–1400, 2009.

[Barnard, 2005] Kobus Barnard, Matthw Johnson, David Forsyth, "Word sense disambiguation with pictures," *AI*, 167(2): 13–30, 2005.

[Hua, 2017] Yazhou Yao, Jian Zhang, Fumin Shen, Xian-Sheng Hua, Jingsong Xu and Zhenmin Tang, Exploiting Web Images for Dataset Construction: A Domain Robust Approach. *TMM*, 19(8): 1771–1784, 2017.

[Lin, 2012] Yuri Lin, Jean Michel, Erez Aiden, Jon Orwant, Will Brockman, Syntactic annotations for the google books ngram corpus. *ACL*, 169–174, 2012.

[Wang, 2014] Xiaogang Wang, Shi Qiu, Ke Liu, Xiaoou Tang, Web image re-ranking usingquery-specific semantic signatures. *TPAMI*, 36(4): 810–823, 2014.

[Zhou, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning deep features for discriminative localization *CVPR*, 2921–2929, 2016.

[Rumelhart, 1986] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, Learning representations by back-propagating errors. *Nature*, 5(3): 1–12, 1986.

[Divvala, 2014] Santosh Kumar Divvala, Ali Farhadi, Carlos Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," *CVPR*, 3270–3277, 2014.

[Zhang, 2016] Yazhou Yao, Xiansheng Hua, Fumin Shen, Jian Zhang and Zhenmin Tang, A domain robust approach for image dataset construction. *ACM MM*, 212–216, 2016.

[Yao, 2019] Yazhou Yao, Jian Zhang, Fumin Shen, Li Liu, Fan Zhu, and Heng-Tao Shen, Towards Automatic Construction of Diverse, High-quality Image Datasets. *TKDE*, 2019.

[Gella, 2016] Spandana Gella, Mirella Lapata, Frank Keller, "Unsupervised Visual Sense Disambiguation for Verbs using Multimodal Embeddings" *ACL*, 1600–1613, 2016.

[Lucchi, 2012] Aurelien Lucchi, Jason Weston, "Joint image and word sense discrimination for image retrieval" *ECCV*, 130–143, 2012.

[Golge, 2014] Eren Golge, Pinar Duygulu, "Concept map: Mining noisy web data for concept learning" *ECCV*, 439–455, 2014.

[Qiu, 2013] Shi Qiu, Xiangang Wang, Xiaoou Tang, "Visual semantic complex network for web images" *ICCV*, 3623–3630, 2013.

[Wah, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie, "The Caltech-UCSD birds-200–2011 dataset", *Tech Report*, 2011.

[Lin, 2015] Tsung-Yu Lin, Aruni Roychowdhury, Subhransu Maji, "Bilinear CNN models for fine-grained visual recognition" *ICCV*, 1449–1457, 2015.