Pattern Recognition xxx (xxxx) xxx

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Exploiting textual queries for dynamically visual disambiguation

Zeren Sun[a], Yazhou Yao[a,*], Jimin Xiao[b], Lei Zhang[c], Jian Zhang[d], Zhenmin Tang[a,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[b] Department of Electrical and Electronic Engineering, Xian Jiaotong-Liverpool University, Suzhou, China
[c] School of Computer Science, Northwestern Polytechnical University, Xian, China
[d] School of Computing and Communication, University of Technology Sydney, Sydney, Australia

ARTICLE INFO

Article history:
Received 17 January 2020
Revised 15 August 2020
Accepted 29 August 2020
Available online xxx

Keywords:
Visual disambiguation
Image search
Text queries
Web images

ABSTRACT

Due to the high cost of manual annotation, learning directly from the web has attracted broad attention. One issue that limits the performance of current webly supervised models is the problem of visual polysemy. In this work, we present a novel framework that resolves visual polysemy by dynamically matching candidate text queries with retrieved images. Specifically, our proposed framework includes three major steps: we first discover and then dynamically select the text queries according to the keyword-based image search results, we employ the proposed saliency-guided deep multi-instance learning (MIL) network to remove outliers and learn classification models for visual disambiguation. Compared to existing methods, our proposed approach can figure out the right visual senses, adapt to dynamic changes in the search results, remove outliers, and jointly learn the classification models. Extensive experiments and ablation studies on CMU-Poly-30 and MIT-ISD datasets demonstrate the effectiveness of our proposed approach.

1. Introduction

Over the past several years, labeled image datasets have played a critical role in high-level image understanding [1,2]. However, the process of constructing labeled datasets is both time-consuming and labor-intensive [3,4]. To reduce the time and labor cost of manual annotation, several works have focused on active learning. For example, Collins et al. [5] proposed to label some seed images to train the initial classifiers. Then, they leveraged these classifiers to carry out classification on other unlabeled images and find low confidence images for manual labeling. The process was iterated until sufficient classification accuracy was achieved. In [6], a system for online learning of object detectors was proposed. This system refined its models by actively requesting annotations on images. However, active learning methods require pre-existing annotation, which often is one of the most significant limitations when it comes to scalability.

To further reduce the cost of manual annotation, learning directly from web images has attracted more and more attention [7]. 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 [8,9] like Google[1], Bing[2], or Baidu[3]. Unfortunately, the precision of images returned from these search engines is still unsatisfactory. For example, Schroff et al. [8] reported that the average precision of the top 1000 images for 18 categories from the Google Image Search Engine is only 32%.

One of the main 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., “bus”, “handbag”, sports “instructor”, or “company”). This is commonly referred to as word-sense disambiguation in Natural Language Processing.

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 [10] or Wikipedia [11]. However, visual disambiguation is a *dynamic data-driven problem* that is specific to images 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 apple before the “Apple” company was founded.

* Corresponding authors.
E-mail addresses: yazhou.yao@njust.edu.cn (Y. Yao), tzm.cs@njust.edu.cn (Z. Tang).

[1] https://www.google.com/imghp?
[2] https://www.bing.com/images/discover?
[3] https://image.baidu.com/

https://doi.org/10.1016/j.patcog.2020.107620
0031-3203/

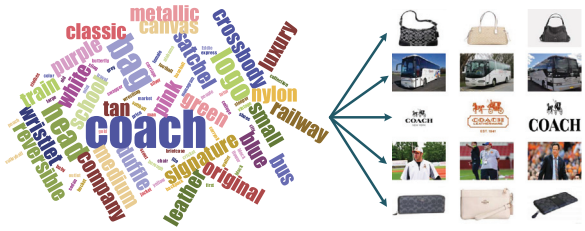

Fig. 1. Visual polysemy. The keyword "coach" can refer to multiple text semantics, resulting in images with various visual senses being included in the returned results by an image search engine.

The traditional way of handling visual polysemy falls back to expert knowledge, WordNet or Wikipedia. However, this human-developed knowledge suffers from the problem of missing visual information and still requires manual annotation to bridge the text semantics and visual senses [12,13]. Some existing works attempt to reduce the influence of visual polysemy by filtering out irrelevant images [14,15]. For example, Li et al. [15] utilized the few top-ranked images returned from an image search engine to learn the initial classifier. Then, the classifier refined its model through an incremental learning strategy. With the increase in the number of positive images accepted by the classifier, the learned model would reach a robust level. Hua et al. [16] leveraged a clustering-based strategy to remove "group" noisy images and a propagation-based mechanism to filter out individual noisy images. These existing methods have the advantage of eliminating manual intervention. However, none of them can directly address the problem of visual polysemy.

Since the text semantics and visual senses of a given keyword are highly related, recent works have also concentrated on combining text and image features [17]. Most of these methods assume that there exists a one-to-one mapping between semantic and visual senses for a given keyword. However, this assumption is not always true in practice. For example, while there are two predominant text semantics for the word "apple", there exist multiple visual senses due to appearance variation (green vs. red apples). To deal with the multiple visual senses, Chen et al. [12] adopted 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 effectiveness of this approach is greatly reduced if we fail to collect webpages that contain enough text semantics and visual senses [29].

Instead of relying on human-developed resources, we focus on automatically solving visual disambiguation in an unsupervised way. The motivation behind this work comes from the fact that keyword-based image search may yield multiple visual senses, and those returned results change dynamically. Therefore, the proposed approach should have a better time adaptability. Unlike the common unsupervised paradigm, which jointly clusters text and image features to solve visual disambiguation, we present a novel framework that resolves it by dynamically matching candidate text queries with images retrieved for 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 three major steps: we first discover and then dynamically select the text queries according to the keyword-based image search results, 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 and demonstrate its superiority, we conduct extensive experiments on visual polysemy datasets CMU-Poly-30 [12] and MIT-ISD [18]. The main contributions of this work can be summarized as follows:

1) Compared to existing methods, our proposed framework can adapt to the dynamic changes in search results and carry out visual disambiguation accordingly. Therefore, it has a better time adaptability.

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 achieves state-of-the-art performance.

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

The rest of the paper is organized as follows: Section 2 elaborates the related works of visual disambiguation. We propose our framework and associated algorithms in Section 3. The experimental evaluations and ablation studies are presented in Sections 4 and 5, respectively. Section 6 concludes this paper.

2. Related work

Dynamically discovering and distinguishing multiple visual senses for polysemous words is a difficult task [19]. Several authors have proposed to clean the web images and learn visual classification models, although none have specifically addressed the problem of visual polysemy [14,15,20–22]. Fergus et al. [21] proposed the use of visual classifiers learned from the Google Image Search Engine to re-rank the images based on their visual consistency. Subsequent methods [15,20] have employed similar removing mechanisms to automatically construct clean image datasets for training visual classifiers. Berg et al. [14] discovered topics using LDA in the text domain, and then used them to cluster the images. However, these works are category-independent and do not learn which words are predictive of a specific sense.

Our work is related to the text-based word sense discovering methods [23,24]. Pantel et al. [23] presented a clustering algorithm, called Clustering By Committee (CBC), that automatically discovers word senses from the text. It first discovers a set of tight clusters, named committees, that are well-scattered in the similarity space. Then proceeds by assigning words to their most similar clusters. This allows CBC to discover the less frequent senses of a word and to avoid discovering duplicate senses. Each cluster that a word belongs to represents one of its senses. A subsequent method in [24] also employed a similar Clustering by Committee algorithm to congregate similar words.

Our work is also related to the manually labeled expert knowledge works [11,25]. The method in [25] proposed to disambiguate word-senses using statistical models. This method overcomes the knowledge acquisition bottleneck faced by word-specific sense discriminators. By entirely circumventing the issue of polysemy resolution in training material acquisition, the system has acquired an extensive set of sense discriminators from unrestricted monolingual texts without human intervention. In addition, class models also offer the additional advantages of smaller model storage requirements and increased implementation efficiency due to reduced dimensionality. Mihalcea et al. [11] proposed an approach for using Wikipedia as a source of sense annotations for word sense disambiguation. Nevertheless, all methods in [11,25] lack visual information and require human annotation to bridge the text semantics and visual senses.

Our work is related to methods that leverage images for visual disambiguation [18,26–28]. Barnard et al. [28] proposed a method that can be alone, or in conjunction with traditional text-based methods for visual disambiguation. This approach is based on a method developed for automatically annotating images using a statistical model for the joint probability of image regions and words. The method in [27] proposed an unsupervised algorithm based on

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

Lesk, which performs visual sense disambiguation using textual, visual, or multi-modal embeddings. Saenko et al. [18] proposed to use a dictionary to learn models of visual word senses, from a large collection of unlabeled web data. Due to the text semantics and visual senses are highly related, the performance when leveraging only images is not satisfying.

Our work is more related to the methods that combine text and images [29–32]. Wan et al. [30] proposed a method that combines a dictionary and the visual content of web images to disambiguate keyword-based image search. The motivation is that these images contain a rich source of information about the various senses (visual and word) of a word. Both methods in [30,31] assumed that there exists a one-to-one mapping between the semantic and visual sense. However, this assumption is not always true in practice. Chen et al. [12] proposed a co-clustering based approach for sense discovery. Specifically, they relaxed the assumption and allowed a one-to-many mapping, which is useful when the granularity of clustering in two domains is different. However, clustering-based methods have a scalability problem. This is because the images are directly retrieved from the web, and thus have no bounding boxes. Every image creates millions of data points, the majority of which are outliers. Our work is inspired by Zhang et al. [13], which has two major drawbacks. Our work solves these two problems very well. Firstly, our proposed approach can adapt to the dynamic changes in the keyword-based image search results. Secondly, our proposed saliency-guided deep MIL network can remove outliers and capture the key regions of the web images to learn classification models.

3. Framework and methods

As shown in Fig. 2, our proposed approach consists of three major steps. The following subsections describe our proposed approach in detail.

For ease of presentation, we denote kw as a keyword and $E(kw)$ as the number of discovered candidate text queries for kw. $I(kw)$ and $I(tq)$ are the selected images for keyword kw and text query tq, respectively. We denote each image as x_i and $\vartheta_i(I)$ to be the number of images in $I(kw)$, which can match with x_i. $D(m, n)$ represents the distinctness between text query m and n.

3.1. Discovering candidate text queries

Manually developed dictionaries (e.g., WordNet [10] or ConceptNet [33]) usually serve as a source for word senses. However, these dictionaries tend to include many rare senses, while missing corpus/domain-specific senses. In addition, the process of constructing manually compiled dictionaries is time-consuming and labor-intensive. To ease the limitations of missing information, as well as to reduce the dependency on manually labeled data, Pantel et al. [23] and Chatterjee et al. [24] proposed to discover semantic senses from text via clustering. The disadvantage is that these methods overly depend on the quality of the collected text. The performance of these methods is greatly reduced when we fail to collect enough useful text. Inspired by recent works [13,34], untagged corpora Google Books [35] can be used to discover candidate text queries for modifying a given keyword. Following the work in [35] (see Section 4.3), we discover the candidate text queries by using n-gram dependencies whose modifiers are tagged as NOUN.

3.2. Dynamically selecting text queries

Image search results are dynamically changing, and not all 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 kw. Then we perform a clean-up step for broken links and set the rest images $I(kw)$ as the selected images for 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 $I(kw)$. In addition, to well obtain the visual senses of the images, some subset images which all have tq must contain visually similar content. To this end, $E(kw)$ can be dynamically selected in the following way.

For each image $x_i \in I(tq)$, all images in $I(kw)$ are matched with x_i on the basis of their visual similarity. In our work, the visual features and similarity measure methods from Wang et al. [36] are leveraged. We set $\vartheta_i(I)$ to be the number of images in $I(kw)$ that can match with x_i. The overall number of a candidate text query tq matching with the search results is its accumulated number 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 with a good number of images in $I(kw)$. When tq is only present in a few images or images involving tq are visually different, $\vartheta(tq)$ will be set to zero. Accordingly, when tq contains a large accumulated value $\vartheta(tq)$, this indicates that many images within $I(kw)$ contain tq and the images involving tq have similar visual senses. These N text queries with the

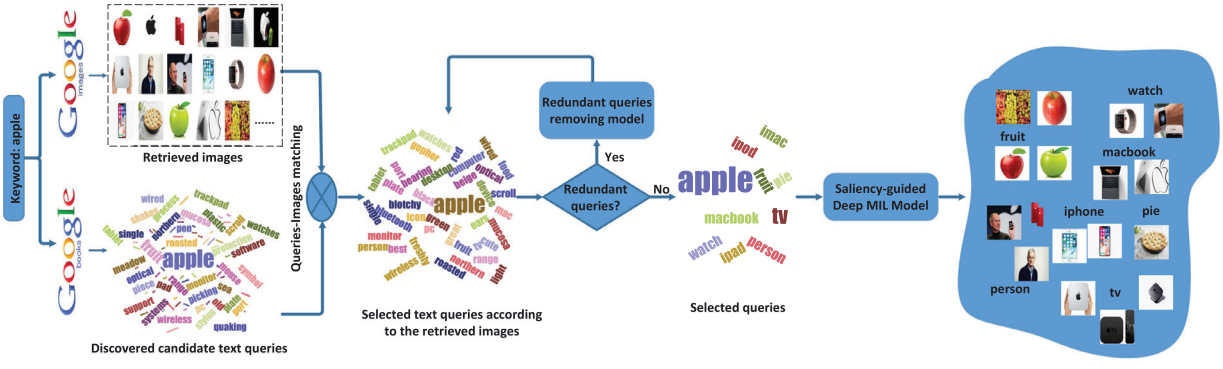

Fig. 2. The proposed dynamically visual disambiguation framework. The input is a keyword. We first discover a list of candidate text queries and retrieve the top images for the given keyword. Then, we dynamically purify the candidate text queries according to the retrieved images. We remove the redundant queries and set the rest as the final selected text queries. We retrieve the top images for each selected text query and leverage a saliency-guided deep MIL model to remove outliers and perform visual disambiguation.

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

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 visually similar distributions (e.g., "Apple MacBook" and "Apple laptop"). To reduce the computing costs, the text queries that increase the discriminative power of the semantic space are kept and the others are removed. To calculate the visual similarity between two text queries, half of the data from each text query is used to learn a binary SVM classifier to carry out classification on the other half of the data. We conclude 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 previous step. We split the retrieved images for text query m into two groups, I_m^t and I_m^v. To calculate the distinctness $D(m, n)$ between text queries m and n, we train a binary SVM using I_m^t and I_n^t. We then obtain the probability of an image in I_m^v belonging to 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 as:

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

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

γ^N is an N-dimensional indicator vector $\gamma \in \{0, 1\}^N$ such that $\gamma_n = 1$ indicates the nth text query is selected and $\gamma_n = 0$ indicates it is 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. $\phi = (\vartheta(tq_1), \vartheta(tq_2), \ldots, \vartheta(tq_N))$, where $\vartheta(tq_n)$ is defined in Eq. (1). λ is the scaling factor. Then Eq. (4) is formulated as an integer quadratic programming problem where the variable $\gamma \in \{0, 1\}$. Although the integer quadratic programming is NP hard, we can use the Label Generating MMC (LG-MMC) algorithm [51] to solve this integer programming problem. To be specific, γ is relaxed to be in $\mathbb{R}^T$ and we choose the text query n whose $\gamma_n \geq 0.5$ as the final selected text query.

3.3. Outliers removal and visual disambiguation

Due to the error indexing of an image search engine, even if we retrieve the top sense-specific images, some noise may still be included. The last step of our approach is to train saliency-guided deep MIL visual models for pruning these instance-level outlier images and distinguishing visual senses of image search results. Our proposed saliency-guided deep MIL visual model consists of two-stream networks, SGN and DMIL. SGN is used to localize an object for generating instances and learning object features. DMIL is used to encode the discriminative features for learning the deep classification models that will remove outliers and perform visual disambiguation.

Different from existing methods which attempt to follow a multi-instance assumption, where the object proposals are regarded as one "instance" sets and each image is 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 proposals ("instances") generated by existing methods like selective search [37] or RPN [38] 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 an instance for the web image. Specifically, we follow the work in [39] 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 are summed to generate the saliency map for each image. Finally, we conduct a binarization operation on the saliency map with an adaptive threshold, which is obtained through the OTSU algorithm [40], and leverage the bounding box that covers the largest connected area as the discriminative region of the object. For a given image I, the value of the spatial location (x, y) in the saliency map for category c is defined as follows:

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

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 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 for the image.

In the 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 a deep CNN as our architecture for learning visual representations with multi-instance learning. Our structure is based on VGG-16 [41] 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 the 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

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

represents the true label of the sample x. To learn multiple instances as a bag of samples, we incorporate a deep representation

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

into MIL and name it DMIL. Assume a bag $\{x_j|j=1,\ldots,n\}$ contains n instances and the label of the bag is $t=\{t_i|t_i\in\{0,1\}, i=1,\ldots m\}$; 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:

$$\tilde{h}_i = f(h_{i1},\ldots,h_{in}), \tag{9}$$

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 for each possible choice. Then, we can represent the visual distribution of the bag and the loss L as:

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

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:

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

To disambiguate the keyword-based 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 used to encode the discriminative features for learning deep models to remove outliers and perform visual disambiguation.

4. 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 compare the search results re-ranking ability of our approach with baseline methods. In addition, we conduct ablation studies on coefficients, domains, hidden layers, deep models, more web images, and the different contributions of each step.

4.1. Classifying sense-specific images

The goal of this experiment is to compare the image sense-specific categorization ability of our proposed approach with four sets of baseline works.

4.1.1. Datasets and evaluation metric

Two widely used polysemy datasets, CMU-Polysemy-30 [12] and MIT-ISD [18], are employed to validate the proposed dynamic visual disambiguation framework. Specifically, CMU-Polysemy-30 and MIT-ISD include 30 and 5 keywords, respectively. We set the images corresponding to various keywords in CMU-Polysemy-30 and MIT-ISD as the results of a keyword-based image search. We follow the settings in baselines [12,13] and exploit web images as the training set, and human-labeled images in CMU-Polysemy-30 and MIT-ISD as the testing set to evaluate the visual disambiguation performance. Average Classification Accuracy (**ACA**) is adopted as the evaluation metric. The image features used in the experiments are 4096-dimensional deep features based on the VGG-16 model.

4.1.2. Implementation details and parameters

For each keyword, we first discover the candidate text queries by searching in Google Books. We set the corresponding images in CMU-Polysemy-30 and MIT-ISD as the results of the 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 the 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 the positive instances (in spite of the fact that noisy images might be included). The 100 images collected for each selected text query are 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 values of $\tilde{\rho_m}$ and $\tilde{\rho_n}$. We then get the distinctness $D(m, n)$ by calculating Eq. (2) and remove redundant queries by solving Eq. (4). The value of 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 Eq. (2). γ_n is set to $\gamma_n \geq 0.5$ in Eq. (4).

The structure of SGN is based on VGG-16 [41]. 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, and 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 the ImageNet dataset [4] 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 [41]. We remove the last hidden layer and use a $\max(\cdot)$ layer instead. The initial parameters of the modified version of the model are inherited from the pre-trained VGG-16 model. During training, we leverage "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 only sample images from the CMU-Polysemy-30 and MIT-ISD datasets.

4.1.3. Baselines

To quantify the performance of our proposed approach, we compare its sense-specific image classification ability with four sets of baselines: knowledge-based methods, text-based methods, image-based methods, and the combination of text and images based methods. The knowledge-based methods consist of Wiki-MD [11], Dict-MD [42], and Copr-MD [25]. The text-based methods include EDWD [43], DWST [23], and TMWSD [44]. The image-based methods contain VSD [30], ULVSM [18], VSCN [45], NEIL [46], ConceptMap [47], and WSDP [28]. The combination of text and images based methods are ISD [31], SDCIT [12], DRID [9], DDPW [29], LEAN [34], VSDE [48], and IWSD [49]. The method [29] reproduced nearly all leading methods on the CMU-Polysemy-30 and MIT-ISD datasets. We directly use the results of these methods from [29].

4.1.4. Experimental results

Fig. 3 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, the 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 the keyword "bass" in the CMU-Poly-30 dataset, only "bass fish" and "bass guitar" are annotated. Our approach additionally discovers two other visual senses "bass amp" and "Mr./Miss

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

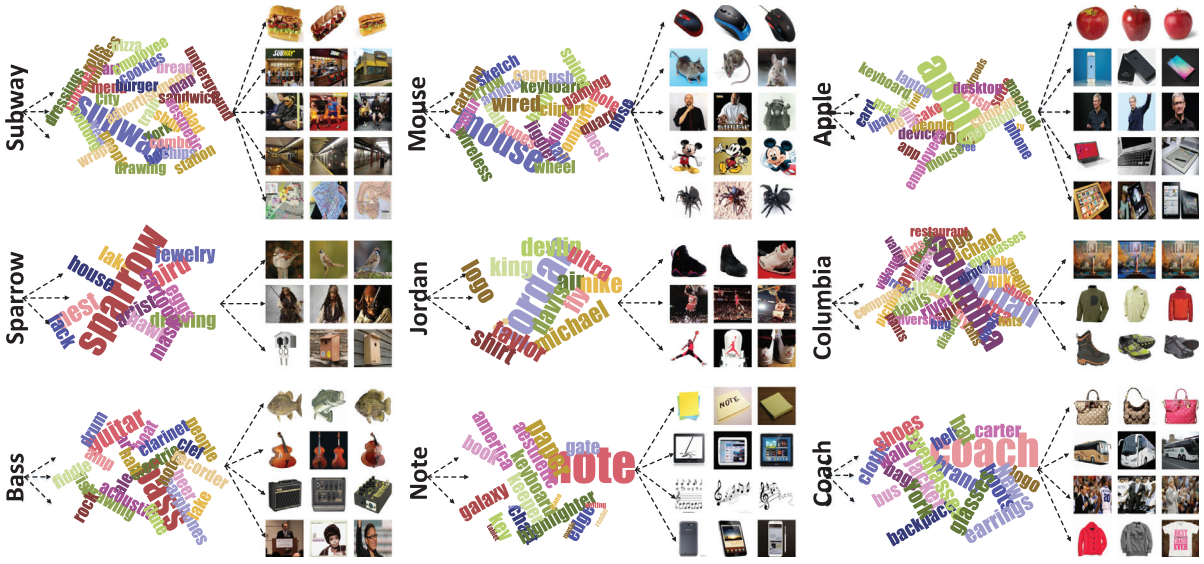

Fig. 3. A snapshot of multiple text queries discovered from Google Books and visual senses disambiguated from the CMU-Poly-30 dataset by our proposed framework. For example, our proposed method automatically discovers and disambiguates five senses for "*Subway*": subway sandwich, subway store, subway 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.

Bass". This is mainly due to the fact that our approach can dynamically select text queries based on image search results. In other words, our proposed framework can adapt to the dynamic changes in search results and carry out visual disambiguation accordingly. Therefore, our approach has a better time adaptability.

To leverage the ground truth labels in CMU-Poly-30 and fairly compare with other baseline methods, we remove the text query discovering and selecting procedures 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 datasets.

From Table 1, we can observe that image-based and the combination of text and images based methods are generally better than the knowledge-based and text-based methods. The reason is that knowledge-based and text-based methods directly leverage web images for training. Due to the error indexing of an image search engine, the web images tend to contain outliers, and we need to remove these to train better models.

By observing Table 1, our proposed approach achieves state-of-the-art ACA performance on both CMU-Poly-30 and MIT-ISD, producing significant improvements over knowledge-based methods, text-based methods, image-based methods, and combination of text and images 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.

Table 1
Visual disambiguation results (ACA) on two evaluated datasets CMU-Poly-30 and MIT-ISD. The best result is marked in **bold**.

	Method	Dataset	
		CMU-Poly-30	MIT-ISD
[a]	Wiki-MD [11]	0.498	0.487
	Dict-MD [42]	0.529	0.522
	Copr-MD [25]	0.549	0.593
[b]	EDWD [43]	0.469	0.483
	DWST [23]	0.563	0.627
	TMWSD [44]	0.593	0.646
[c]	VSD [30]	0.728	0.786
	ULVSM [18]	0.772	0.803
	WSDP [28]	0.791	0.743
	NEIL [46]	0.741	0.705
	ConceptMap [47]	0.726	0.758
	VSCN [45]	0.802	0.783
[d]	ISD [31]	0.554	0.634
	IWSD [49]	0.643	0.725
	SDCIT [12]	0.839	0.853
	VSDE [48]	0.747	0.763
	LEAN [34]	0.827	0.814
	DRID [9]	0.846	0.805
	DDPW [29]	0.884	0.897
[e]	Ours	**0.925**	**0.938**

[a] Knowledge-based methods.
[b] Text-based methods.
[c] Image-based methods.
[d] Combination of text and images based methods.
[e] Our proposed approach.

4.2. Re-ranking search results

The goal of this experiment is to compare the image search results re-ranking ability of our approach with three sets of baseline works.

4.2.1. Datasets and evaluation metric

We leverage the "Bass" and "Mouse" datasets introduced in [29] to evaluate the re-ranking search results ability. Detailed information on these datasets is summarized in Table 2. Following [29,30], we evaluate the re-ranking performance by computing the Area Under Curve (AUC) of all senses for "bass" and "mouse".

4.2.2. Implementation details and parameters

For each query, the sense-specific classifiers are trained with sense-specific web images. Specifically, we leverage the previously trained sense-specific classifiers for the classifying sense-specific images experiment. 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. $P(S_i|d)$ 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.

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

Table 2
"Bass" and "Mouse" polysemy datasets introduced in [29]. For each term, the number of annotated images, semantic senses, visual senses and their distributions are provided, with core semantic senses marked in **bold**.

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

Table 3
Area Under Curve (AUC) of all senses for "bass" and "mouse". The best results are marked in **bold**.

	Method	Semantic Senses & Visual Senses						Average
		bass fish	bass guitar	M. Bass	computer mouse	little mouse	cartoon mouse	
[a]	Wiki-MD [11]	0.364	0.429	0.132	0.536	0.623	0.114	0.366
	Dict-MD [42]	0.443	0.635	0.205	0.464	0.573	0.186	0.418
	Copr-MD [25]	0.504	0.486	0.305	0.624	0.675	0.263	0.476
[b]	VSD [30]	0.547	0.538	0.239	0.684	0.652	0.226	0.481
	ULVSM [18]	0.526	0.615	0.326	0.732	0.735	0.314	0.541
[c]	ISD [31]	0.453	0.526	0.243	0.614	0.536	0.218	0.432
	LEAN [34]	0.623	0.658	0.413	0.753	0.785	0.336	0.595
	SDCIT [12]	0.658	0.773	0.386	0.815	0.845	0.337	0.636
	DDPW [29]	0.713	0.736	0.572	0.834	0.873	0.434	0.694
[d]	Ours	**0.746**	**0.782**	**0.623**	**0.876**	**0.915**	**0.478**	**0.737**

[a] Knowledge-based methods.
[b] Image-based methods.
[c] Combination of text and images based methods.
[d] Our proposed approach.

4.2.3. Baselines

We compare the search results re-ranking ability of our method with three sets of baselines, which include knowledge-based methods, image-based methods, and combination of text and images based methods. The knowledge-based methods consist of Wiki-MD [11], Dict-MD [42] and Copr-MD [25]. The image-based methods contain VSD [30] and ULVSM [18]. The combination of text and images based methods include ISD [31], LEAN [34], SDCIT [12], and DDPW [29]. The method [29] reproduced nearly all leading methods. We directly leverage the results of these methods from [29].

4.2.4. Experimental results

The experimental results are shown in Table 3. From Table 3, we observe that the combination of text and images based methods SDCIT [12], LEAN [34] and DDPW [29] are generally better than knowledge-based methods Wiki-MD [11], Dict-MD [42], Copr-MD [25] and images-based methods VSD [30], ULVSM [18]. Specifically, SDCIT [12], LEAN [34] and DDPW [29] achieve better results than ISD [31]. This is because it is necessary to remove outlier images from the training set during the process of classifier learning. Learning directly from the web images without outliers removal may affect the performance of the classifier due to the presence of outlier images.

By observing Table 2, we can notice that there is 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 first discover the multiple visual senses for the query. As such we can choose appropriate visual senses as needed to carry out sense-specific image collection. By doing this, we can greatly improve the efficiency of collecting web images, thereby improving the efficiency of learning from the web images.

From Table 3, we achieve the best average performance which is consistent with the results for sense-specific image classification. This can be explained by the generated sense-specific terms and filtered images of our approach. Compared to knowledge-based methods Wiki-MD [11], Dict-MD [42] and Copr-MD [25], our approach does not directly use web images for classifier learning. Instead, we purify the retrieved images to select useful data and then use the selected images to learn classifiers. By doing this, our approach can effectively overcome the impact of outliers on the classifiers due to the error indexing of image search engines. Compared to image-based methods VSD [30], ULVSM [18] and combination of text and images based methods ISD [31], LEAN [34], SDCIT [12], DDPW [29], the sense-specific terms generated by our approach are more accurate and exhaustive, using our sense-specific terms to retrieve images can return high precision web images, and can thereby help us to train sense-specific classifiers to re-rank the search results.

5. Ablation studies

Since our proposed approach incorporates multiple steps to carry out dynamic visual disambiguation, we analyze the contributions of each step in this section. We also analyze the coefficients, domains, hidden layers, models, web images, time and space complexity through ablation studies. In addition, we leverage our approach as a pre-step before directly learning from the web and show its superiority.

5.1. Contributions of each step in proposed framework

Our proposed visual disambiguation framework includes three major steps: candidate text queries discovering, text queries selecting, and outliers removing. To quantify the role of different steps contributing to the final model, we construct two new frameworks.

One is based on candidate text queries discovering and text queries selecting (which we refer to TQDS). The other one is based on text queries discovering and outliers removing (which we refer to TDOR). For framework TQDS, we first obtain the candidate text

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

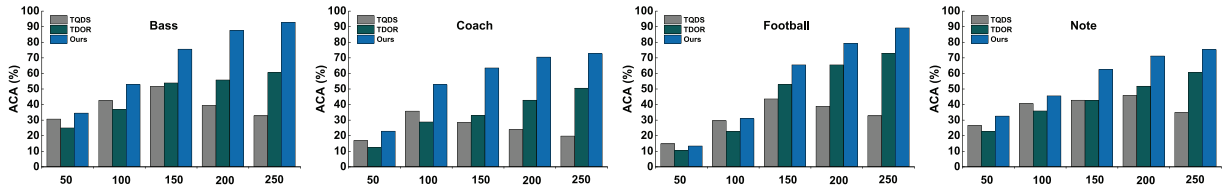

Fig. 4. Sense-specific image classification ability of TQDS, TDOR and ours on CMU-Poly-30 dataset ("bass", "coach", "football" and "note").

queries through searching in Google Books. Then, we employ the text queries selecting procedure to obtain the selected text queries. We directly retrieve the top images from the image search engine for the selected text queries to train image classifiers (without outliers removing). For framework TDOR, we also obtain the candidate text queries by searching in Google Books. Then, we retrieve the top images from the image search engine for all the candidate text queries (without text queries selecting procedure). We employ the saliency-guided deep MIL model to remove outliers and train image classifiers.

We compare the sense-specific image classification ability of these two new frameworks with our complete framework. Following [29], we select "note", "bass", "coach" and "football" as four target categories to evaluate. We sequentially collect [50, 100, 150, 200, 250] images for each selected text query as the positive training samples and use 500 fixed irrelevant negative samples to learn image classifiers. We test the sense-specific image classification ability of these three frameworks on the CMU-Poly-30 dataset. The results are shown in Fig. 4. From Fig. 4, we can observe: Framework TQDS usually performs better than TDOR when the training number for each semantic sense is small (e.g., below 150). The explanation is that the first few returned images tend to have fewer outliers. With an increasing number of images for each text query, the images retrieved from the image search engine contain more and more noise. In this situation, the outlier images caused by the image search engine have a worse effect than those induced by noisy text queries.

Our proposed complete framework outperforms both TQDS and TDOR. The reason is that our complete framework, which combines text queries selecting and outlier images removing, can effectively remove the outliers induced by both the noisy text queries and the error indexing of the image search engine.

5.2. Coefficients in proposed framework

For the coefficients analysis, we are mainly concerned with the parameters α, β, γ, N, and $I(tq)$ when selecting the text queries and learning rate (LR) for the saliency-guided deep MIL. Specifically, we analyze the interaction between pairs of parameters α and β in Eq. (3). For other parameters, we analyze the sensitivities using one 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 on the CMU-Poly-30 dataset.

5.3. Influence of different domains

To analyze the influence of using web images from different domains 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 on the web images from Flickr is much lower than on those from the Google Image Search Engine and the Bing Image Search Engine. One possible explanation is that Flickr's image data comes from people's daily lives, and the background is more complicated, making it difficult to accurately locate the target objects. The performance on web images coming from the Google Image Search Engine is a little better than on those from the Bing Image Search Engine. This may be due to Google's bias toward images with a single centered object and a clean background. This allows us to obtain the bounding boxes of the target objects easily and accurately.

5.4. Influence of different hidden layers

The choice of hidden layer is of critical importance in our proposed saliency-guided deep MIL network. As mentioned in Section 3.3, the $\max(\cdot)$, $\text{avg}(\cdot)$, and $\log(\cdot)$ refer to $\max_j(h_{ij})$, $\text{avg}_j(h_{ij})$, 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.

5.5. Are deeper models helpful?

It is well known that the CNN model architecture has a critical impact on object recognition performance. We investigate this issue by replacing VGG-16 with a new architecture, ResNet-50, in the saliency-guided deep MIL model and compare the results. The experimental results are shown in Fig. 8(a). In particular, the ResNet-50 model is more effective for localizing the objects from the images on CMU-Poly-30 dataset. While on the MIT-ISD dataset, VGG-16 obtains a better performance.

5.6. Are more web images helpful?

Data scale has a large impact on web-supervised learning. We investigate 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 improves steadily with the use of more training samples.

5.7. Time and space complexity analysis

Our proposed framework primarily contains three major steps: candidate text queries discovering, text queries selecting, and outliers removing. Since the first step of our approach is to search from the corpus, its time complexity is negligible. For the time complexity analysis, we mainly concern the second and third step. We formulate the process of "text queries selecting" as an SVM problem. There are a large number of works have analyzed the time complexity of SVM. According to [50]. The time complexity of SVM is between $O(n^2)$ and $O(n^3)$ with n is the number of training instances. In the third step, we formulate the outliers removing as a saliency-guided deep MIL problem. The time complexity is $O(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l)$ where D is the number of convolution layers, that is, the depth of the network. l represents the lth convolutional layer and C_l is the number of convolution kernels of the lth layer of the neural network. M_l is the side length of lth

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

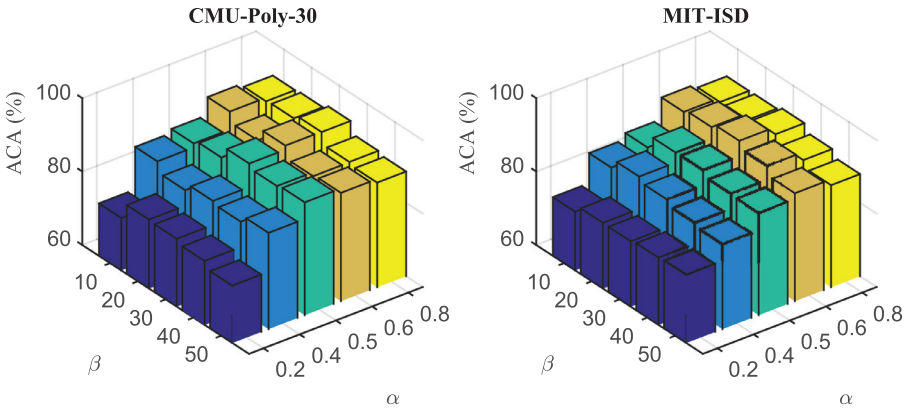

Fig. 5. The ACA performance of the interaction between pairs of parameters α and β.

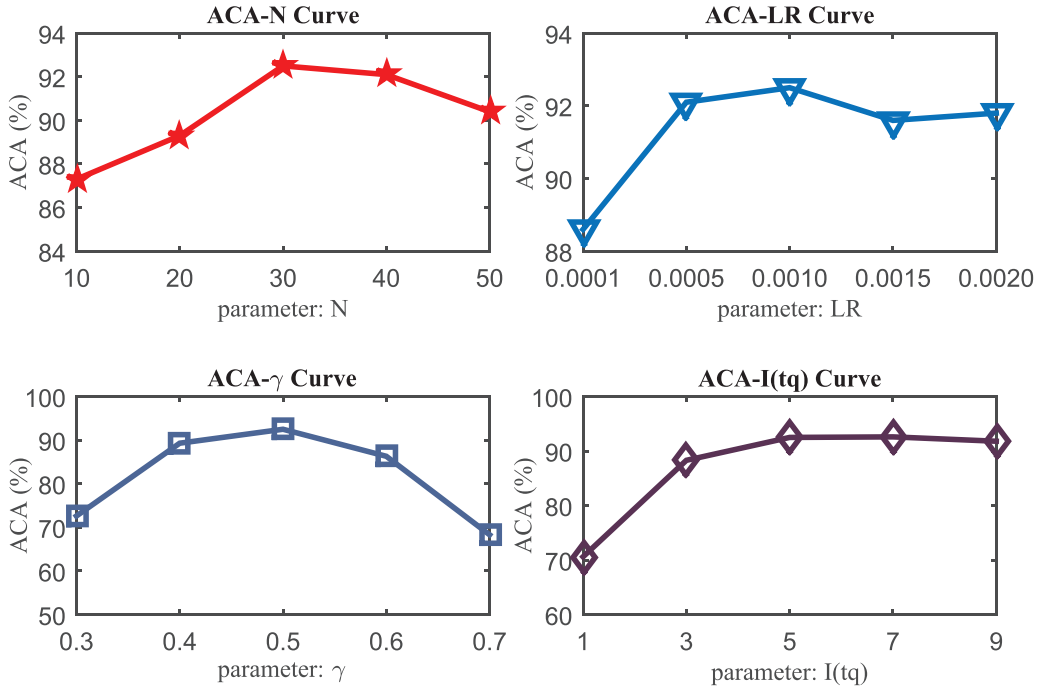

Fig. 6. The parameter sensitivities of *N*, *LR*, γ, and *I*(*tq*) w.r.t. ACA on CMU-Poly-30 dataset.

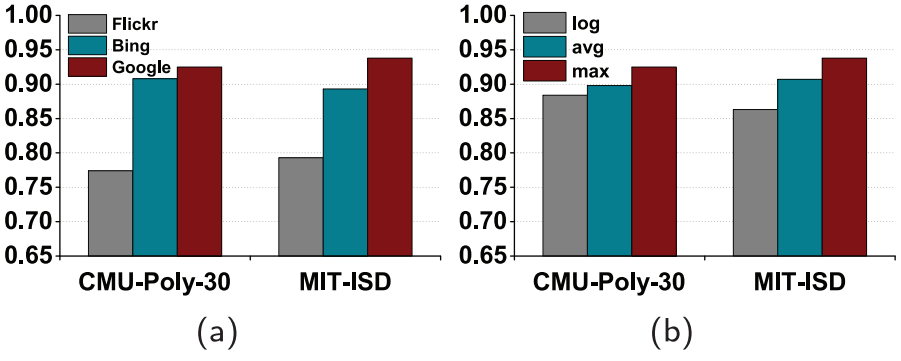

Fig. 7. (a) Demonstration of the impact for different domains. (b) Demonstration of the impact for different hidden layers.

convolution kernel output feature map. K_l is the length of lth convolution kernel.

For the space complexity analysis, we give the hardware configuration of the experiment. All the data processing and experiments were performed on a Dell workstation (Intel Xeon Gold 5120 CPU, 64 GByte RAM and 12 GByte VRAM).

5.8. 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 the CUB-200-2011 dataset [52]. Our overlap removing strat-

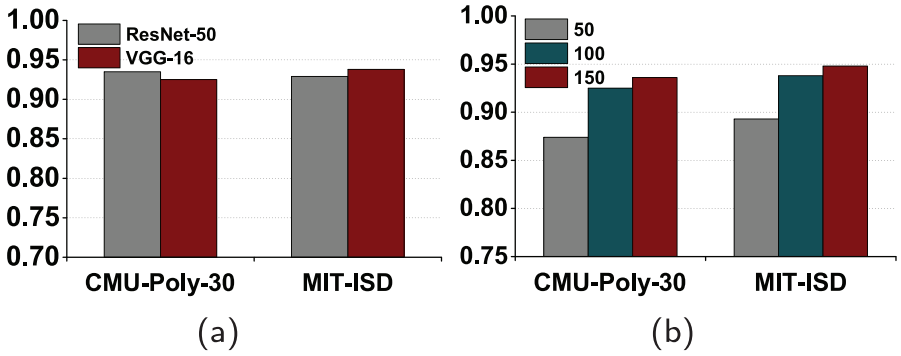

Fig. 8. (a) Demonstration of the impact for different CNN architectures. (b) Demonstration of the impact for different training samples.

Table 4
Fine-grained visual recognition results on CUB-200-2011 testing set.

Training data	Algorithm	Accuracy
Original web	Bilinear	0.718
Clean web	Bilinear	0.832
CUB training	Bilinear	0.841
Clean web + CUB training	Bilinear	0.863

egy between the web training and labeled testing set is under the assumption that images with more similar semantic information are more likely to be similar or even identical. To be specific, we first use the VGG-16 model pre-trained on ImageNet to extract the embedding feature vector for each image in both training and testing data. Then, for every single test image per category, we calculate the similarity distance between this testing image and every training image. For each category, we obtain the smallest distance between training and testing data, which is denoted as θ. We set an empirical threshold factor $\eta = 0.05$ to scale the distance and remove the web training images which have a smaller distance than $(1 + \eta) \times \theta$.

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 leverage the relatively clean web images as the training set to perform one of the most popular weakly supervised fine-grained algorithms Bilinear [53], on the CUB-200-2011 [52] testing set. We leverage the retrieved original web images (without outliers removal) as the training set to perform the Bilinear algorithm on CUB-200-2011 and set the result as the baseline performance. In addition, we put the collected relatively clean web data and the training data in the CUB-200-2011 dataset together as the training data to perform the Bilinear algorithm, and then test on the CUB-200-2011 dataset. The results are shown in Table 4. From Table 4, we can observe that our proposed approach greatly improves the baseline accuracy. Our collected relatively clean web data and manually labeled training data in CUB-200 achieve a very close classification accuracy (83.2% VS 84.1%). From the result of "Clean web + CUB training", we can draw a conclusion that the collected clean web data can be used to enhance existing manually labeled datasets and can achieve a more robust classification model.

5.9. Visualization

Our saliency-guided deep MIL model consists of two-stream networks, and SGN is used to localize objects and generate the "instances" for the web images. Whether or not objects are accurately located by SGN network is the basis for extracting deep features and learning the classification models. Fig. 9 visualizes the object

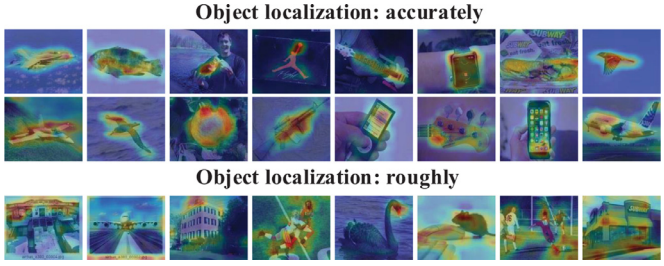

Fig. 9. Visualization of object locating via saliency map.

locating via a 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 of the objects, they can nevertheless be roughly located.

6. Conclusions

In this work, we focused on one important yet often ignored problem: we argue that the current poor performance of models learned from web images is due to the inherent ambiguity in user queries. We solved this problem by visual disambiguation in search results. The contributions mainly contain: 1) our approach can adapt to the dynamic changes in search results and carry out visual disambiguation accordingly; 2) we propose a saliency-guided deep MIL network to remove outliers and jointly learn the classification models for visual disambiguation; 3) our work can be used as a pre-step before directly learning from web images, helping to choose appropriate visual senses for images collection and thereby improving the efficiency of learning from the web. Compared to existing methods, the strengths of our approach are: our proposed approach can figure out the right visual senses, adapt to dynamic changes in the search results, remove outliers, and jointly learn the classification models. Despite the good results we have achieved, the weakness of our approach is that our method still has the problem of domain mismatch. In our future work, how to solve the domain adaptation problem between web images and practical test data is an important research direction. Besides, how to build a large-scale/web-scale noisy data learning system is another important research direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61976116) and Fundamental Research Funds for the Central Universities (No. 30920021135).

References

[1] G.S. Xie, X.Y. Zhang, W. Yang, M. Xu, S. Yan, C.L. Liu, LG-CNN: From local parts to global discrimination for fine-grained recognition, Pattern Recognit. 71 (2017) 118–131.
[2] Q. Wu, C. Shen, P. Wang, A. Dick, A. van den Hengel, Image captioning and visual question answering based on attributes and external knowledge, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1367–1381.
[3] W. Deng, J. Hu, N. Zhang, B. Chen, J. Guo, Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership, Pattern Recognit. 66 (2017) 63–73.
[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[5] B. Collins, J. Deng, K. Li, L. Fei-Fei, Towards scalable dataset construction: an active learning approach, in: European Conference on Computer Vision, 2008, pp. 86–98.
[6] S. Vijayanarasimhan, K. Grauman, Large-scale live active learning: training object detectors with crawled data and crowds, Int. J. Comput. Vis. 108 (2) (2014) 97–114.
[7] Y. Yao, J. Zhang, F. Shen, L. Liu, F. Zhu, D. Zhang, H. Shen, Towards automatic construction of diverse, high-quality image dataset, IEEE Trans. Knowl. Data Eng. 32 (6) (2020) 1199–1211.
[8] F. Schroff, A. Criminisi, A. Zisserman, Harvesting image databases from the web, IEEE Trans. Pattern Anal. Mach. Intell. 33 (4) (2011) 754–766.
[9] Y. Yao, X.S. Hua, F. Shen, J. Zhang, Z. Tang, A domain robust approach for image dataset construction, in: ACM International Conference on Multimedia, 2016, pp. 212–216.
[10] G.A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (11) (1995) 39–41.
[11] R. Mihalcea, Using wikipedia for automatic word sense disambiguation, Assoc. Comput. Linguist. (2007) 196–203.
[12] X. Chen, A. Ritter, A. Gupta, T. Mitchell, Sense discovery via co-clustering on images and text, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5298–5306.
[13] J. Zhang, F. Shen, W. Yang, P. Huang, Z. Tang, Discovering and distinguishing multiple visual senses for polysemous words, in: AAAI Conference on Artificial Intelligence, 2018, pp. 523–530.
[14] T. Berg, D. Forsyth, Animals on the web, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1463–1470.
[15] L.-J. Li, L. Fei-Fei, OPTIMOL: Automatic online picture collection via incremental model learning, Int. J. Comput. Vis. 88 (2) (2010) 147–168.
[16] X. Hua, J. Li, Prajna: towards recognizing whatever you want from images without image labeling, in: AAAI International Conference on Artificial Intelligence, 2015, pp. 137–144.
[17] Y. Yao, F. Shen, G. Xie, L. Liu, F. Zhu, J. Zhang, H. Shen, Exploiting web images for multi-output classification: from category to subcategories, IEEE Trans. Neural Netw. Learn. Syst. 31 (7) (2020) 2348–2360.
[18] K. Saenko, T. Darrell, Unsupervised learning of visual sense models for polysemous words, Adv. Neural Inf. Process. Syst. (2009) 1393–1400.
[19] Z. Sun, F. Shen, L. Liu, L. Wang, F. Zhu, L. Ding, G. Wu, L. Shao, Dynamically visual disambiguation of keyword-based image search, in: International Joint Conference on Artificial Intelligence, 2019, pp. 996–1002.
[20] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search, in: IEEE International Conference on Computer Vision, 2005, pp. 1816–1823.
[21] R. Fergus, P. Perona, A. Zisserman, A visual category filter for google images, in: European Conference on Computer Vision, 2004, pp. 242–256.
[22] Y. Yao, F. Shen, J. Zhang, L. Liu, Z. Tang, L. Shao, Extracting privileged information for enhancing classifier learning, IEEE Trans. Image Process. 28 (1) (2019) 436–450.
[23] P. Pantel, D. Lin, Discovering word senses from text, in: ACM SIGKDD International conference on Knowledge discovery and data mining, 2002, pp. 613–619.
[24] N. Chatterjee, S. Mohan, Discovering word senses from text using random indexing, in: Computational Linguistics and Intelligent Text Processing, 2008, pp. 299–310.
[25] D. Yarowsky, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, in: Association for Computational Linguistics, 1992, pp. 454–460.
[26] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, Z. Tang, Exploiting web images for dataset construction: adomain robust approach, IEEE Trans. Multimed. 19 (8) (2017) 1771–1784.
[27] S. Gella, M. Lapata, F. Keller, Unsupervised visual sense disambiguation for verbs using multimodal embeddings, 2016. arXiv:1603.09188.
[28] K. Barnard, M. Johnson, Word sense disambiguation with pictures, Artif. Intell. 167 (2) (2005) 13–30.
[29] Y. Yao, F. Shen, J. Zhang, L. Liu, Z. Tang, L. Shao, Extracting multiple visual senses for web learning, 21, 2019, pp. 184–196.
[30] K.-W. Wan, A.-H. Tan, J.-H. Lim, L.-T. Chia, S. Roy, A latent model for visual disambiguation of keyword-based image search, in: British Machine Vision Conference, 2009, pp. 2–7.
[31] N. Loeff, C.O. Alm, D.A. Forsyth, Discriminating image senses by clustering with multimodal features, in: Association for Computational Linguistics, 2006, pp. 547–554.
[32] A. Lucchi, J. Weston, Joint image and word sense discrimination for image retrieval, in: European Conference on Computer Vision, 2012, pp. 130–143.
[33] R. Speer, C. Havasi, ConceptNet 5: a large semantic network for relational knowledge, in: The People's Web Meets NLP, 2013, pp. 161–176.
[34] S. Divvala, A. Farhadi, C. Guestrin, Learning everything about anything: webly–supervised visual concept learning, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2014, pp. 3270–3277.
[35] Y. Lin, J.-B. Michel, E.-L. Aiden, J. Orwant, W. Brockman, S. Petrov, Syntactic annotations for the google books ngram corpus, in: Association for Computational Linguistics, 2012, pp. 169–174.
[36] X. Wang, S. Qiu, K. Liu, X. Tang, Web image re-ranking using query-specific semantic signatures, IEEE Trans. Pattern Anal. Mach. Intell. 36 (4) (2014) 810–823.
[37] J.R. Uijlings, K. Sande, T. Gevers, A. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.
[38] S. Ren, K. He, R. Girshick, J. Sun, Faster r-CNN: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. (2015) 91–99.
[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[40] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014. arXiv:1409.1556.
[42] J. Veronis, N. Ide, Word sense disambiguation with very large neural networks extracted from machine readable dictionaries, in: Association for Computational Linguistics, 1990, pp. 389–394.
[43] S. Cucerzan, Large-scale named entity disambiguation based on wikipedia data, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 708–716.
[44] J. Boyd-Graber, D. Blei, X. Zhu, A topic model forword sense disambiguation, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 1024–1033.
[45] S. Qiu, X. Wang, X. Tang, Visual semantic complex network for web images, in: IEEE International Conference on Computer Vision, 2013, pp. 3623–3630.
[46] X. Chen, A. Shrivastava, A. Gupta, NEIL: Extracting visual knowledge from web data, in: IEEE International Conference on Computer Vision, 2013, pp. 1409–1416.
[47] E. Golge, P. Duygulu, Concept map: mining noisy web data for concept learning, in: European Conference on Computer Vision, 2014, pp. 439–455.
[48] S. Gella, M. Lapata, F. Keller, Unsupervised visual sense disambiguation for verbs using multimodal embeddings, 2016. arXiv:1603.09188.
[49] A. Lucchi, J. Weston, Joint image and word sense discrimination for image retrieval, in: European Conference on Computer Vision, 2012, pp. 130–143.
[50] L. Bottou, C.J. Lin, Support vector machine solvers, in: Large Scale Kernel Machines, 2007, pp. 301–320.
[51] Y. Li, I. Tsang, J. Kwok, Z. Zhou, Tighter and convex maximum margin clustering, in: International Conference on Artificial Intelligence and Statistics, 2009, pp. 344–351.
[52] C. Wah, S. Branson, E. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, Tech Report, 2011.
[53] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in: IEEE International Conference on Computer Vision, 2015, pp. 1449–1457.

Zeren Sun received the B.S. degree in Computer Science from Nanjing University of Science and Technology and the M.S. degree in Robotics Technology from Carnegie Mellon University. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include computer vision, deep learning, fine-grained classification and learning from noise.

Yazhou Yao is currently a Professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology. He obtained the Ph.D. degree in 2018 from Global Big Data Technologies Center (GBDTC), University of Technology Sydney, Australia. From July 2018 to July 2019, he worked as a Research Scientist at Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include multimedia, computer vision, and machine learning.

Jimin Xiao is currently an Associate Professor at the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University. He received the B.S and M.E degrees in Telecommunication Engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004, 2007, respectively. He received the Ph.D. degree in Electrical Engineering and Electronics from University of Liverpool, UK, in 2013. His research interests includes image/video processing, computer vision, deep learning.

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620

Lei Zhang is currently a research scientist in Inception Institute of Artificial Intelligence (IIAI), UAE. He received his Ph.D. degree in the School of Computer Science and Engineering, Northwestern Polytechnical University (NPU), Xi'an, China, in 2018. He was a Research Staff with the School of Computer Science, The University of Adelaide, Adelaide, SA, Australia. His research interests include image processing, machine learning, and video analysis.

Jian Zhang is currently an Associate Professor at the School of Electrical and Data Engineering, University of Technology Sydney (UTS). He is the lab Leader of Multimedia and Data Analytics in the Global Big Data Technologies Centre (GBDTC) at UTS. He earned the Ph.D. degree from School of Information Technology and Electrical Engineering, UNSW@ADFA, Australian Defence Force Academy, at the University of New South Wales in 1999. His research interests include image processing, computer vision, pattern recognition and data analytics, multimedia and social media signal processing, large scale image and video content analytics, and multimedia information retrieval.

Zhenmin Tang is currently a Professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology. He obtained the Ph.D. degree in 2002 from School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include pattern recognition, intelligent system, image processing and object detection.

Please cite this article as: Z. Sun, Y. Yao and J. Xiao et al., Exploiting textual queries for dynamically visual disambiguation, Pattern Recognition, https://doi.org/10.1016/j.patcog.2020.107620