Visual Contextual Advertising: Bringing Textual Advertisements to Images

Yuqiang Chen[†]   **Ou Jin**[†]   **Gui-Rong Xue**[†]   **Jia Chen**[†]   **Qiang Yang**[‡]

[†]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[‡]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong

[†]{yuqiangchen, kingohm, grxue, chenjia}@apex.sjtu.edu.cn   [‡]qyang@cs.ust.hk

Abstract

Advertising in the case of textual Web pages has been studied extensively by many researchers. However, with the increasing amount of multimedia data such as image, audio and video on the Web, the need for recommending advertisement for the multimedia data is becoming a reality. In this paper, we address the novel problem of *visual contextual advertising*, which is to directly advertise when users are viewing images which do not have any surrounding text. A key challenging issue of visual contextual advertising is that images and advertisements are usually represented in image space and word space respectively, which are quite different with each other inherently. As a result, existing methods for Web page advertising are inapplicable since they represent both Web pages and advertisement in the same word space. In order to solve the problem, we propose to exploit the social Web to link these two feature spaces together. In particular, we present a unified generative model to integrate advertisements, words and images. Specifically, our solution combines two parts in a principled approach: First, we transform images from a image feature space to a word space utilizing the knowledge from images with annotations from social Web. Then, a language model based approach is applied to estimate the relevance between transformed images and advertisements. Moreover, in this model, the probability of recommending an advertisement can be inferred efficiently given an image, which enables potential applications to online advertising.

1   Introduction

Online advertising is one of the booming sectors of the Web based business world. Huge revenue has been made each year in this area. Traditional online advertising researches, such as contextual advertising (Ribeiro-Neto et al. 2005), focus on delivering advertisements for textual Web pages so that they are matched as closely as possible. Researchers on contextual advertising area have proposed various ways to deal with the problem, e.g. (Ribeiro-Neto et al. 2005; Lacerda et al. 2006; Broder et al. 2007). However in recent years, the rapid increase in multimedia data, such as image, audio and video on the Web, presents new opportunities for

context based advertisement. Therefore, adverting for multimedia data is increasingly in need.

Considering the problem of recommending advertisements for Web images, traditional methods largely rely on the textual contexts of images, such as surrounding text and social tags, to extract keywords and then obtain relevant advertisements through textual information retrieval. However, there are a large amount of Web images with little or no text contexts. Furthermore, text can be noisy and ambiguous, which could reduce the accuracy for the recommended advertisements.

In this paper, we present a novel approach to bring textual advertisements (ADs) to images, so that images and ADs are closely matched to the true meaning of images rather than the textual context of images. Specifically, we focus on the *visual contextual adverting* problem, which aims to recommending textual advertisements for Web images without the help of any textual context, such as surrounding text for the images. Unlike previous methods for contextual advertising, in our approach, advertisements are recommended entirely based on the visual contextual of an image rather than its textual context in our setting. As a result, our approach can recommend advertisements even for images with little or no textual context.

Visual contextual advertising can be applied to various Web applications in reality. For example, many images in online albums lack annotations, and it will be helpful to advertise the images based on the content of them. Another example is that when visiting entertainment websites, it will be beneficial if we can advertise the cloths or watch worn by a fashion star when browsing the star's image, in which case no corresponding text is available. Yet another example is when one browse his own images on a mobile phone, in which case appropriate Ads are required to be chosen for the user with high accuracy.

Ideally, in order to perform visual contextual advertising, the algorithm must first understand the images and then make appropriate recommendations based on the understanding. However, images are often in a image feature space, such as color histogram, or scale invariant feature transform (Lowe 2004) descriptors, while the advertisements are usually represented in a word space, which usually consists of textual features from bid phrases, landing pages, etc. As a result, traditional contextual text advertising meth-
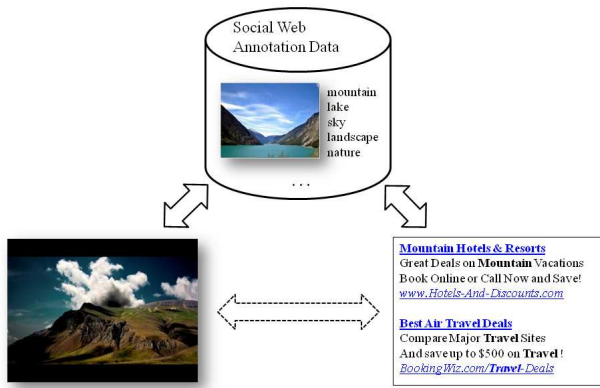
Figure 1: An illustrative example of our solution for *visual contextual adverting* problem. Originally images are represented in different feature space from advertisements, which poses difficulties for advertising. With the help of annotation data from social Web, we can link the images and advertisements, and thus advertise directly based on the context of images.

ods are unable to handle this problem.

In order the overcome the above-mentioned problems, we exploit the annotated image data from social Web sites such as Flickr[1] to link the visual feature space and the word space. An illustrative example of our proposed method for visual contextual advertising is given in Figure 1. To be specific, we present a unified generative model, `ViCAD`, to deal with the visual contextual advertising. `ViCAD` runs in several steps. First, we model the visual contextual advertising problem with a Markov chain which utilizes annotated images to transform images from the image feature space to the word space. With the representations of images in word space, a language model for information retrieval is then applied to find the most relevant advertisements. Moreover, we show that the inference of the model can be performed efficiently by constraining the word space for representing image to be a smaller subspace, which allows potential applications of using our model in online image advertising.

2 Related Work

2.1 Contextual Advertising

The online advertising problem is getting increasingly important with the development of the Web business. A popular approach in online advertising is contextual advertising, which is to place advertisements in a target web page based on the similarity between the content of target page and advertisement description. Various researchers have addressed the problem of contextual adverting (Ribeiro-Neto et al. 2005; Lacerda et al. 2006; Broder et al. 2007). In (Ribeiro-Neto et al. 2005), ten strategies of contextual advertising were proposed and compared. Lacerda et al. proposed a framework using learning method based on genetic programming . Broder et al. combined the semantic and

syntactic features to calculate the relevance score between a web page and advertisements (Broder et al. 2007).

All the mentioned studies intend to advertise based on textual web page context, rather than visual textual. Mei, Hua, and Li presented a novel contextual advertising algorithm using both surrounding textual information and local visual relevance (Mei, Hua, and Li 2008). An image advertisement would be seamlessly inserted into the target image relying on a saliency map. However in (Mei, Hua, and Li 2008), image information was mainly used as a complement for textual annotation when annotations were insufficient or the quality was low. In this paper, we consider a more challenging scenario where no surrounding text is given and advertisements are recommended entirely based on the visual context.

2.2 Image Annotation

Another closely related area is image annotation. Duygulu et al. regarded the image annotation as a machine translating process (Duygulu et al. 2002). Some other researchers model the joint probability of images regions and annotations. Barnard et al. (Barnard et al. 2003) investigated image annotation under probabilistic framework and put forward a number of models for the joint distribution of image blobs and words . Blei and Jordan (Blei and Jordan 2003) developed *correspondence latent Dirichlet allocation* to model the joint distribution. In (Jeon et al. 2003), *continuous-space relevance model* was proposed to better handle continues features and be free from the influence of image blobs clustering. In (Carneiro et al. 2007), image annotation is posed as classification problems where each class is defined by images sharing a common semantic label. While visual contextual advertising presented in this paper has some similarity with image annotation, some key differences exist. A major difference is that the advertisements correspond to groups of fixed keywords rather than collections of independent keywords as in the case of image annotation. As such, there is a need to tradeoff advertisement selection with the accuracy of individual words. Advertisement selection also relates to diversity of selected advertisements as a whole, and other important factors such profit.

3 `ViCAD` Algorithm for Visual Contextual Advertising

3.1 Problem Formulation

First we define the problem of visual contextual advertising formally. Let $\mathcal{W} = \{w_1, w_2, \ldots, w_m\}$ be the vocabulary space, where w_i is a word and m is the size of vocabulary. Let $\mathcal{T}$ be the advertisement space. In this space, each advertisement $t_i \in \mathcal{T}$ is represented by a feature vector $(t_i^1, t_i^2, \ldots, t_i^m)$ on the word space $\mathcal{W}$. We denote $\mathcal{V}$ as the image space, in which each image $v_i \in \mathcal{V}$ is represented by feature vector $(v_i^1, v_i^2, \ldots, v_i^n)$ on the image feature space $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$. To link two different feature spaces, the word space and the image feature space, we make use of auxiliary annotated image data $\mathcal{D} = \{(v_i, w_i)\}_{i=1}^{|\mathcal{D}|}$, where $v_i \in \mathcal{V}$ and $w_i \in \mathcal{W}$ are a image-word pair which is semantically related.
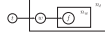
lar advertisement. Therefore, Eq. (2) can be reformed as

$$\Pr(t|v) \propto \sum_w \left(\underbrace{\Pr(w|t)}_{\text{language model}} \prod_{f \in v} \underbrace{\Pr(f|w)}_{\text{feature translator}} \right). \quad (3)$$

From Eq. (3), it is clear that our model can be decomposed into two parts. The first part $\Pr(w|t)$ can be estimated using a language model. The second part is the product of probabilities of image features give a word, $\Pr(f|w)$, which requires a feature translator in order link the image feature and word space. We will describe how to estimate these two parts in the following subsections.

Parameter estimation of $\Pr(w|t)$ via language model Statistical language model is originally used in speech recognition, machine translation problems. With the develop of information retrieval, researchers begin to introduce the language models into the information retrieval area (Ponte and Croft 1998).

In this paper, we estimate the probability $\Pr(w|t)$ via a language model. The unsmoothed language model of $\Pr(w|t)$ is given by the maximum likelihood estimation

$$\Pr_{ml}(w|t) = \frac{n(w,t)}{\sum_{w' \in t} n(w', t)}. \quad (4)$$

where $n(w,t)$ is the number of occurrences a word w in advertisement t.

However, Eq. 4 underestimates the unseen words in the advertisements. Many researchers have proposed smoothing methods to overcome the problem (Jelinek and Mercer 1980; Zhai and Lafferty 2004). We utilize the Jelinek-Mercer's smoothing method (Jelinek and Mercer 1980; Hiemstra 2002) in the model estimation. Jelinek-Mercer's smoothing involves a linear interpolation between the maximum likelihood model of an advertisement and the model of all advertisements, with a coefficient λ controlling the influence of those two models. Formally,

$$\Pr(w|t) = (1-\lambda) \cdot \Pr_{ml}(w|t) + \lambda \cdot \Pr(w|\mathcal{T}). \quad (5)$$

where $\mathcal{T}$ is the advertisement space containing all the advertisements. In practise, we can effectively calculate $\Pr(w|t)$ without enumerating all the advertisements by using a index. More detailed descriptions would be given in Section 3.3 .

Parameter estimation of $\Pr(f|w)$ via social annotation data Since $\Pr(f|w)$ involves two feature spaces, the image feature and word space, we can not estimate the $\Pr(f|w)$ directly. Intuitively we want to find a kind of data that act as a bridge for transferring knowledge between different kinds of knowledge sources. In this work, we choose to use socially annotated images from the Web, such as Flickr, as a bridge to link the word and image feature spaces. We assume a Markov chain $w \rightarrow v \rightarrow f$ underlying the image-word co-occurrence data and image features. Then $\Pr(f|w)$ can be decomposed as

$$\Pr(f|w) = \sum_v \underbrace{\Pr(f|v)}_{\text{feature extractor}} \cdot \underbrace{\Pr(v|w)}_{\text{co-occurrence data}} . \quad (6)$$

Figure 2: Graphical model representation of ViCAD model. In this figure t represents an advertisement, w represents a word and f represents an image feature. n_t is the number of words in t and n_w is the number of image features corresponding to w.

The image set to be advertised is denoted as $\mathcal{I}$. For each image $v \in \mathcal{I}$, the objective is to find a function $h_v(t) : \mathcal{T} \mapsto \mathbb{R}$ that estimates the relevance between a given image v and an advertisement t as accurately as possible.

3.2 Visual Contextual Advertising Model

This section gives a detailed description of our model for visual contextual advertising. Our assumption is that given the word space $\mathcal{W}$, the advertisements $\mathcal{T}$ and the image features $\mathcal{F}$ are conditionally independent. Under the independence assumption of words in advertisements, the joint distribution of t, w, and v is given by

$$\begin{aligned} \Pr(t,w,v) &= \Pr(t)\Pr(w|t)\Pr(v|w) \\ &= \Pr(t)\Pr(w|t)\prod_{f \in v}\Pr(f|w). \quad (1) \end{aligned}$$

The graphical model representation of our proposed model ViCAD is shown in Figure 2.

The objective of visual contextual advertising $h_v(t)$ is set to be $\Pr(t|v)$ in this paper, which can be computed using

$$\begin{aligned} h_v(t) &\triangleq \Pr(t|v) = \sum_w \frac{\Pr(t,w,v)}{\Pr(v)} \\ &= \sum_w \Pr(t)\Pr(w|t)\prod_{f \in v}\frac{\Pr(f|w)}{\Pr(v)}. \quad (2) \end{aligned}$$

Since we keep an image v fixed to find related advertisements, $\Pr(v)$ in Eq. (2) can be omitted. Furthermore, we assume the prior over words $\Pr(t)$ to be uniform since we do not have any preference or prior knowledge on a particu-

Algorithm 1 Visual Contextual Adverting (ViCAD) Algorithm.

Input: Image set $\mathcal{I}$ which is to be advertised, auxiliary image word concurrence data $\mathcal{D}$, target advertisement set $\mathcal{T}$.
Output: $h_v(t)$ for each $v \in \mathcal{I}$ and $t \in \mathcal{T}$.

1: **for all** $w \in \mathcal{W}$ **do**
2:   **for all** $f \in \mathcal{F}$ **do**
3:     Calculate $\Pr(f|w)$ based on $\mathcal{D}$ and Eq. (6), (7), (8).
4:   **end for**
5: **end for**
6: **for all** $v \in \mathcal{I}$ **do**
7:   **for all** $t \in \mathcal{T}$ **do**
8:     **for all** $w \in \mathcal{W}$ **do**
9:       Calculate $\Pr(w|t)$ based on Eq. (5).
10:     **end for**
11:     Calculate $\Pr(t|v)$ based on $\Pr(w|t)$, $\Pr(f|w)$ and Eq. (3)
12:   **end for**
13: **end for**
14: **return**: $\Pr(t|v)$ as hypothesis $h_v(t)$.

The probability $\Pr(v|w)$ in Eq. 6 can be estimated with image-word co-occurrence data $\mathcal{D} = \{(v_i, w_i)\}_{i=1}^{|\mathcal{D}|}$ using:

$$\Pr(v|w) = \frac{\#\{(v,w)|(v,w) \in \mathcal{D}\}}{\sum_{v'} \#\{(v',w)|(v',w) \in \mathcal{D}\}}. \quad (7)$$

Another probability in Eq. 6, $\Pr(f|v)$, can be estimated by first applying a feature extractor, such as SIFT (scale invariant feature transform) (Lowe 2004), on v, which is used for converting v into a vector on a image feature space $\mathcal{F}$, and then be estimated using

$$\Pr(f|v) = \frac{n(f,v)}{\sum_{f' \in \mathcal{F}} n(f',v)}. \quad (8)$$

where $n(f,v)$ is the number of occurrences of image feature f appearing in image v.

3.3 Algorithm and Complexity Analysis

A formal description of our algorithm of visual contextual advertising, named ViCAD, is given in Algorithm 1. Step 1 to Step 5 estimate the model parameters required by the algorithm. Step 6 to Step 13 generate a hypothesis that measures the relevance of an image to advertisements for every image $v \in \mathcal{I}$ and advertisement $t \in \mathcal{T}$. $\Pr(t|v)$ is returned as the final hypothesis.

The time cost for computing $\Pr(f|w)$ is $O(|\mathcal{F}| \cdot |\mathcal{D}|)$, where $\mathcal{F}$ is image feature space and $\mathcal{D}$ is the auxiliary word image co-occurrence data. Next we investigate the time cost for computing $\Pr(w|t)$ for all $t \in \mathcal{T}$ and $w \in \mathcal{W}$. Recall our computation of $\Pr(w|t)$ via language model in Eq. 5, $\Pr(w|t)$ equals $\lambda \cdot \Pr(w|\mathcal{T})$ when word w does not occur in t. Since $\lambda \cdot \Pr(w|\mathcal{T})$ is a constant with respect to all advertisements t, when calculating $\Pr(w|t)$ we only need to consider the advertisements which contain word w. With a inverted index on advertisements, this can be achieved efficiently. Therefore the time cost of computing $\Pr(w|t)$ is

$O(nz_{ad})$ where nz_{ad} is the number of nonzero items in the advertisement-word co-occurrence matrix, which is proportional to the input of advertisements.

However when nz_{ad} is very large, the computation can be potentially expensive. In order to furthermore accelerate our algorithm, instead computing $\Pr(t|v)$ under the word space $\mathcal{W}$, we calculate under a smaller subspace of $\mathcal{W}$. Formally we use the formula $\Pr(t|v) = \sum_{w \in \mathcal{W}'} \Pr(w|t) \prod_{f \in v} \Pr(f|w)$ instead of Eq. 3, where $\mathcal{W}'$ is the set of word with large values of $\prod_{f \in v} \Pr(f|w)$. This approximation would not affects the result of our objective of ad recommendation significantly since those words with small $\prod_{f \in v} \Pr(f|w)$ do not tend to affects the final ranking of $h_v(t) \triangleq \Pr(t|v)$ much. Therefore the complexity of computing $\Pr(w|t)$ turns into $O(nz'_{ad})$ where nz'_{ad} is the number of nonzero items in the advertisement-subword space $\mathcal{W}'$ co-occurrence matrix.

The time cost for computing $\Pr(t|v)$ based on $\Pr(t|w)$ and $\Pr(f|w)$ is $O(|\mathcal{W}| \cdot |\mathcal{F}|)$.

To sum up, the total time cost of ViCAD is $O(|\mathcal{I}| \cdot (nz'_{ad} + |\mathcal{W}| \cdot |\mathcal{F}|) + |\mathcal{F}| \cdot |\mathcal{D}|)$. In practise, $|\mathcal{W}| \cdot |\mathcal{F}|$ is usually much smaller than nz'_{ad}. In addition, $|\mathcal{F}|$ can be taken as a constant usually no larger than 1000. So the overall complexity of algorithm ViCAD is $O(|\mathcal{I}| \cdot nz'_{ad} + |\mathcal{D}|)$, which shows our algorithm is quite scalable and fitful for online applications.

4 Experiments

In this section, we describe the data used in our experiment and empirically analyze the effective of our algorithm ViCAD on two different image corpora.
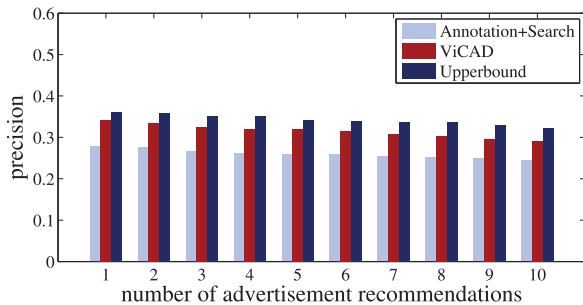
4.1 Data Sets

In order to evaluate our algorithm ViCAD, we conduct experiments on two image corpora, Flickr Images[2] and Fifteen scene categories (Lazebnik, Schmid, and Ponce 2006). Flickr is a social network website where users can freely upload images and add annotations to images. We crawled 59,759 image from Flickr during May to August 2009. Semantic topics of images in Flickr varied widely from artifacts for nature, from human to animal. The Fifteen scene categories is a publicly available image data set of 4,485 images containing fifteen different scenes such as store and forest. From these two corpora, we randomly select 200 images from each data set as the evaluation data sets.

Our advertisement data were crawled from the MSN search engine[3] in 2008. First, a list of popular queries were selected from AOL query log, and then with this query list we searched using MSN search engine to obtain the advertisements appearing in the search results. Overall, 312,759 advertisements with both title and content were crawled as our advertisement set.
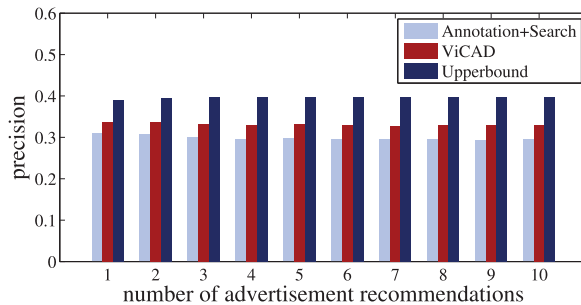
As described in the Section 3.3 , in order to link the spaces of image feature and word, we used image-word co-occurrence data in Flickr as the auxiliary data. Since the

---

[2] http://www.flickr.com
[3] http://search.msn.com

(a) Flickr image set

(b) Fifteen scene data set

Figure 3: Figure representation of $P@n$ results of all compared methods with respect to Flickr and Fifteen scene data sets.

original Flickr annotations are quite noisy, we manually selected 1,055 nouns of widely seen general objects, such as apple, building or mountain, as the word space $\mathcal{W}$. Word selection is independent of advertisements since we would like to have a general purposed word space. Based the selected words and images we crawled, we obtained 673,547 image-word co-occurrence data from Flickr, which form our auxiliary data $\mathcal{D}$. Note that in our experiments, images for evaluating ad-recommendation are different from those images used as auxiliary data.

Data preprocessing is applied to the raw data. We use the "bag-of-words" model (Li and Perona 2005) to represent each image. First interesting point were detected and described by SIFT descriptor (Lowe 2004). Then we cluster a randomly sample subset of all interesting points to obtain a codebook, which is our image feature space $\mathcal{F}$. Similar to (Sivic et al. 2005), we set the number of clusters to be 1,000. Using this codebook, each image is converted into a vector for further ad-recommendation uses.

As mentioned in the Section 3.3 , we use the inverted index to accelerate the computation of $\Pr(w|t)$. For each advertisement, its title and content are concatenated, and then be indexed using Lucene[4]. To further accelerate computation, a smaller word space $\mathcal{W}'$ is exploited (refer to Section 3.3). The size of $\mathcal{W}'$ is set to ten in all our experiments. The smoothing parameter λ for the language model in Eq. 5 is set to be 0.15 as indicated in (Jelinek and Mercer 1980).

4.2 Evaluation Methods

Since few existing work investigate the problem of visual contextual advertising to our best knowledge, we compare our algorithm with two baselines. One baseline is to first annotate images (Makadia, Pavlovic, and Kumar 2008)[5] and then search for related advertisements based on the annotations. We refer to this baseline as Annotation+Search. The other baseline is the *upper bound* performance of recommending advertisements. To estimate upper bound, we peek the true labels of the images used for evaluation and search for the related advertisements in $\mathcal{T}$, to which we referred to as Upperbound method. Specifically in Flickr

---

[4]http://lucene.apache.org/

[5]We adapted color features and SIFT feature as basic features and combined them using *joint equal contribution* (JEC).

	FLICKER DATA			FIFTEEN SCENE DATA		
n	A+S	ViCAD	UB	A+S	ViCAD	UB
1	0.280	0.342	0.362	0.310	0.338	0.39
2	0.276	0.334	0.359	0.307	0.336	0.395
3	0.266	0.324	0.352	0.302	0.332	0.397
4	0.262	0.321	0.351	0.297	0.329	0.398
5	0.261	0.319	0.342	0.298	0.331	0.398
6	0.259	0.316	0.338	0.297	0.329	0.398
7	0.255	0.309	0.337	0.295	0.328	0.398
8	0.253	0.303	0.338	0.295	0.329	0.398
9	0.249	0.297	0.329	0.294	0.329	0.398
10	0.246	0.291	0.322	0.296	0.331	0.398

Table 1: The $P@n$ results of all compared methods with respect to Flickr and Fifteen scene data sets. Due to space limitation, Search+Annotation is abbreviated as A+S, and Upperbound is abbreviated as UB.

data set, the true labels are the tags annotated by users, while in Fifteen scene data set, the true labels are the scene names of images. Notice that the Upperbound method is unavailable in our visual contextual advertising problem setting since no textual knowledge about an image is given. In our experiments, the upper bound is used for reference.

As for evaluate metrics, we choose the *precision at n* (or $P@n$), which is the portion of related advertisements on the topmost n recommendations to evaluate the experimental results. To obtain ground truth results, four participants manually judged whether target images and advertisements are relevant or not. The final result is the average precision of all participants.

4.3 Empirical Analysis

The $P@n$ results with respect to Flickr and Fifteen scene data sets are given in Figure 3. Detailed results are given in Table 1. On one hand, ViCAD shows comparable performance to the Upperbound on both data corpora. On the other hand, we can see that ViCAD performs better than Annotation+Search regardless the number of advertisements recommended. This indicates that ViCAD, which integrates the visual contextual advertising into a unified model, performs better than adverting with two separate process.

One reason why the performance of Upperbound on Flickr data set is not as good as on Fifteen scene data set

Flickr Data Set		Fifteen Scene Data Set	
	Dog, cat, fish, bird, reptile supplies Find all reptile supplies online Positron emission tomography, PET scan		Ocean beach flower delivery Ocean drive beach flower delivery Ocean Isle beach flower delivery
	Scenery hill flower delivery Lake hill flower delivery Twin lake hill flower delivery		Send a gift tower or gift baskets. Shop for kitchen wall tiles Shop for wall stencil
	Humour plants excellence Shop for nursery plants Nursery plants		Master building products Pte Ltd Shop for building materials CTC holidays
	You can sail round the world! Shop for pontoon boat Bass boat shoes		Buy or sell cars, cars items on eBay Car loan calculator Buy used cars

Table 2: The visual contextual adverting results with respect to Flickr and Fifteen scene data sets. Texts given in right of each image are the top three advertisements recommended by our algorithm `ViCAD`.

lies in the quality of labels of the images. The Fifteen scene data consists of limited number of scenes and images were delicately selected by human. Therefore the quality of labels is quite high. Nevertheless the Flickr data set contains images with far more scenes and labels were annotated by Web users, which in turn leads to noise in the labels. However, comparing the result of our algorithm `ViCAD` on those two data sets, the performance difference is not large especially for those top three recommended advertisements. This demonstrates that our algorithm is an effective solution when no label or only noisy labels of an image are given.

Table 2 shows the visual contextual adverting result of our algorithm `ViCAD` with respect to Flickr and Fifteen scene data sets. In this figure, images in the first column are the images from Flickr data set and images in the second column are the images from Fifteen scene data set. Top three advertisement recommendations are given on the right of each image. From the table we can see that our algorithm can indeed find related advertisements based on the visual contextual information of an image. Moreover on our dual core 2.4GHz CPU PC, our algorithm without full optimization takes only about 0.2s to advertise an image, which empirically shows our algorithm is quite efficient and applicable for online applications in practice.

5   Conclusion

In this paper, we proposed a new approach for Web based advertising, called image contextual advertising, which is to advertise an image without the help of any surrounding text. A unified generative model was applied to model the advertisements, words, and image features. Based on the model, we proposed an algorithm called `ViCAD`, and showed how to accelerate inference and analyzed the complexity. Finally we performed experiments on two image data sets and demonstrated the effectiveness of our algorithm.

References

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3:1107–1135.

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *SIGIR '03*, 127–134.

Broder, A.; Fontoura, M.; Josifovski, V.; and Riedel, L. 2007. A semantic approach to contextual advertising. In *SIGIR '07*, 559–566.

Carneiro, G.; Chan, A. B.; Moreno, P. J.; and Vasconcelos, N. 2007. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Patt. Anal. Mach. Intell.* 29(3):394–410.

Duygulu, P.; Barnard, K.; Freitas, J. F. G. d.; and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02*, 97–112.

Hiemstra, D. 2002. Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term. In *SIGIR '02*, 35–41.

Jelinek, F., and Mercer, R. 1980. Interpolated estimation of markov source parameters from sparse data. *Pattern Recognition in Practice*.

Jeon, L. M.; Lavrenko, V.; Manmatha, R.; and Jeon, J. 2003. A model for learning the semantics of pictures. In *NIPS '03*. MIT Press.

Lacerda, A.; Cristo, M.; Gonçalves, M. A.; Fan, W.; Ziviani, N.; and Ribeiro-Neto, B. 2006. Learning to advertise. In *SIGIR '06*, 549–556.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR' 06*, 2169–2178.

Li, F.-F., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR' 05*, 524–531.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60:91–110.

Makadia, A.; Pavlovic, V.; and Kumar, S. 2008. A new baseline for image annotation. In *ECCV '08*, 316–329.

Mei, T.; Hua, X.-S.; and Li, S. 2008. Contextual in-image advertising. In *ACM MULTIMEDIA '08*, 439–448.

Ponte, J. M., and Croft, W. B. 1998. A language modeling approach to information retrieval. In *SIGIR '98*, 275–281.

Ribeiro-Neto, B.; Cristo, M.; Golgher, P. B.; and Silva de Moura, E. 2005. Impedance coupling in content-targeted advertising. In *SIGIR '05*, 496–503.

Sivic, J.; Russell, B. C.; Efros, A. A.; Zisserman, A.; and Freeman, W. T. 2005. Discovering object categories in image collections. In *ICCV' 05*.

Zhai, C., and Lafferty, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22(2):179–214.