Extracting Visual Knowledge from the Web with Multimodal Learning

Dihong Gong, Daisy Zhe Wang
Department of Computer and Information Science and Engineering
University of Florida
{gongd, daisyw}@ufl.edu

Abstract

We consider the problem of automatically extracting visual objects from web images. Despite the extraordinary advancement in deep learning, visual object detection remains a challenging task. To overcome the deficiency of pure visual techniques, we propose to make use of meta text surrounding images on the Web for enhanced detection accuracy. In this paper we present a multimodal learning algorithm to integrate text information into visual knowledge extraction. To demonstrate the effectiveness of our approach, we developed a system that takes raw webpages and a small set of training images from ImageNet as inputs, and automatically extracts visual knowledge (e.g. object bounding boxes) from tens of millions of images crawled from the Web. Experimental results based on 46 object categories show that the extraction precision is improved significantly from 73% (with state-of-the-art deep learning programs) to 81%, which is equivalent to a 31% reduction in error rates.

1 Introduction

Recent progresses on computer vision research community such as large scale object detection [Girshick, 2015][Simonyan and Zisserman, 2014], age invariant face recognition [Gong *et al.*, 2013; 2015], and region-to-phrase correspondences [Plummer *et al.*, 2015], largely benefit from ever increasing amount of visual knowledge as training data. Collecting visual knowledge in a crowdsourcing manner, such as the ImageNet database [Deng *et al.*, 2009] and Visipedia [Perona, 2010], has major limitation in the lack of both diversity and scalability. In addition, manually annotating large collection of images is usually an expensive and time consuming process. For example, based on the most recent statistics, there are only 7.28% images annotated in the ImageNet database. Given these limitations, in this paper we explore an alternative approach to build large visual knowledge base by extracting visual knowledge from Web data.

Mining useful visual knowledge automatically from the Web can be challenging. Even with the recent state-of-the-art visual object detection algorithms, the precision of content based image retrieval is still unacceptable. For example, GoogLeNet [Szegedy *et al.*, 2015] which finished at the 1^{st} place in the ImageNet Large-Scale Visual Recognition Challenge 2014, only achieves mean average precision of 43.9%.

Given the limitation of visual knowledge extraction methods purely relying on visual content, this paper explores an alternative method to integrate multimodal information for better extraction accuracy. The Web is mostly a vast collection of unstructured information of various modalities (e.g. text, image, and video), where multimodal information is usually correlated. For instance, images of a news article usually illustrate the corresponding text content and meta text like *alt* or *src* of an image describes the content of that image. These observations suggest that a potentially greater visual knowledge extraction accuracy can be achieved by making use of information from alternative modalities.

2 Related Work and Our Contributions

Traditionally, visual knowledge bases such as ImageNet [Deng *et al.*, 2009] and Visipedia [Perona, 2010] are constructed by manual annotations with motivated teams of people or power of crowds. These approaches however are quite limited because annotations become expensive, prone to errors and do not scale. To overcome these disadvantages, recent studies have been focused on leveraging machine learning technologies to reduce human intervention. For example, Vijayanarasimhan et al. [Vijayanarasimhan and Grauman, 2014] proposed an active learning algorithm based on crowdsourcing using Amazon Mechanical Turk to train object detectors with crawled data. Later, Chen et al. [Chen *et al.*, 2013b] presented a completely autonomous system called Never Ending Image Learner (NEIL) to automatically mine visual knowledge from the Web. While these systems are designed to reduce the human intervention for visual knowledge mining, their retrieval precision is still quite low (e.g. NEIL has mean average precision of 51% for 15 object categories) due to limited source of information.

There have been an increasing research interest for multimodal learning in the recent years. The general goal of multimodal learning is to utilize information across multiple modalities (e.g. text, image, video, and audio) for either enabling cross-modality query or improving retrieval accuracy for a wide variety of machine learning tasks. For instances, Zhu et al. [Zhu *et al.*, 2015] proposed a scalable algorithm to build multimodal knowledge base for an-

swering visual queries. Their system takes annotated multimodal data (e.g. images with text descriptions and attributes) as input, and establishes relations between entities using MRF models. The system however doesn't completely solve the automatic visual knowledge mining problem because it relies on annotated multimodal data, the acquisition of which is another challenging problem yet to be addressed. Other representative approaches such as [Kiros *et al.*, 2014; Norouzi *et al.*, 2013; Lazaridou *et al.*, 2015; Frome *et al.*, 2013] primarily focus on learning semantic embeddings using deep neural networks for multimodal object representations. The basic idea of these approaches is to map objects in different modalities into a common vector space so that correspondence between multimodal objects can be established. For example, Kiros et al. [Kiros *et al.*, 2014] proposed a multimodal skip-gram model to learn word embeddings closely related to the corresponding vision concepts (e.g. embedding of word **kitten** is close to embedding of visual objects in **cat** category). Similarly, Frome et al. [Frome *et al.*, 2013] developed a system called "DeViSE" that learns to transform embeddings from visual modality to textual modality to allow prediction of unseen visual categories based on text labels (so called *Zero-Shot Learning* [Norouzi *et al.*, 2013]). However, like the system by Zhu et al. [Zhu *et al.*, 2015], all of these methods rely on image data manually annotated (e.g. annotation is noise-free) with text descriptions by human workers (e.g. Flickr 8K [Hodosh *et al.*, 2013] and ImageNet [Deng *et al.*, 2009]). Annotating images with text descriptions is another challenging problem yet to be solved. Consequently, these approaches may not be suitable for open-domain visual knowledge extraction, where noise-free image descriptions are usually unavailable.

Our work is closely related to the aforementioned approaches, with major contributions summarized as follows:

- The major novelty of this paper is to present a multimodal learning approach to integrate meta text surrounding web images for large scale open-domain visual knowledge extraction. Compared to existing approaches, our approach is proved to be effective for real-world web data that is noisy, incomplete or redundant.
- We develop a sophisticated end-to-end system for large-scale visual knowledge extraction from real web data. Unlike existing multimodal approaches which test on standard datasets, in this paper we take hundreds of millions of raw webpages as input, and automatically extract visual object bounding boxes as outputs for testing.
- Finally, we demonstrate a significant improvement in extraction precision over the state-of-the-art visual object detection algorithms. Experimental results based on 46 object categories show that by making use of textual and visual information jointly, the extraction precision is improved significantly from 73% to 81%, which is equivalent to a 31% reduction in error rates.

3 Multimodal Embeddings

In this section, we describe an algorithm to learn embeddings for textual words and visual concepts. The learned representative embeddings in continuous space preserve relative distance between multimodal objects, such that if two objects have similar meaning (e.g. "car" and "truck"), their embeddings are also close to each other.

3.1 Review of Skip-Gram Model

Our algorithm is closely related to the skip-gram model [Guthrie *et al.*, 2006] which is a language modeling algorithm used to learn embeddings of text words. Given text corpus, the training objective of the Skip-Gram model is to find word representations that are useful for predicting the surrounding words in a sentence. Mathematically, it maximizes the objective function

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t), \tag{1}$$

where $w_1, w_2, \dots w_T$ is a sequence of training words in the corpus, and c is the size of the window around target w_t. In the basic Skip-Gram model, the conditional probability $p(w_{t+j}|w_t)$ is defined using the softmax function as

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^T v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^T v_{w_I}\right)}, \tag{2}$$

where v_w and v'_w are the "input" and "output" vector representations of w, and W is the size of the vocabulary. Due to the normalization term in denominator, the Equation (2) requires $O(W)$ time complexity, which makes it computionally impractical since the size of vocabulary is usually very large (e.g. $10^5 - 10^7$). To speedup the computation, hierarchical softmax [Mikolov *et al.*, 2013] is proposed to approximate the standard softmax in Equation (2) with a binary tree, which reduces the time complexity from $O(W)$ to $O(\log(W))$ on average. In this paper, we shall apply the hierarchical softmax whose implementation is based on Google word2vec[1].

3.2 Definition

Suppose we have a set of N images from which visual knowledge is to be extracted, denoted as

$$\mathcal{I} = \{(I_n, T_n)|n = 1 \dots N\},$$

where I_n represents an image, and T_n denotes a set of text phrases describing image I_n. Then we apply visual object detection program on each of the image I_n, and arrive at a set of detected visual objects denoted as

$$\mathcal{I}_{\mathcal{D}} = \{(D(I_n), T_n)|I_n \in \mathcal{I} \wedge \mathbf{card}(D(I_n)) > 0\}, \tag{3}$$

where $D(\cdot)$ is a detection operator gives a set of detected visual categories, and $\mathbf{card}(\cdot)$ is the cardinality of a set. We only retain images that have at least one detected object from predefined set of visual categories (denoted as $\mathcal{C}$).

We have two vocabularies, one for text and one for image. The text vocabulary (denoted as $\mathcal{V}$) is derived from text phrases associated with images in $\mathcal{I}_{\mathcal{D}}$, and visual vocabulary (denoted as $\mathcal{C}$) is derived from visual categories.

[1] https://code.google.com/archive/p/word2vec

3.3 Formulation

Our goal is to learn embeddings for text phrase $x \in \mathcal{V}$ and visual category $y \in \mathcal{C}$, such that:

- The embeddings $\vec{v}(x)$ and $\vec{v}(y)$ are close to each other (e.g. in Euclidean *n-space* with cosine distance) if text phrase x is an appropriate description of image category y. For example, $\vec{v}$(**kitten**) should be close to $\vec{v}$(**cat**), where **kitten** $\in \mathcal{V}$ and **cat** $\in \mathcal{C}$, since **kitten** is an appropriate description of **cat**.
- The embeddings $\vec{v}(y_1)$ and $\vec{v}(y_2)$ are close to each other if objects in image category y_1 *co-occur* with objects in image category y_2 with high frequency. Two objects *co-occur* if they occur in the same image. For example, $\vec{v}$(**car wheel**) is close to $\vec{v}$(**car**) because objects of **car wheel** and **car** *co-occur* with high frequency.

These two items are essential. The first item identifies a set of important text phrases that are useful for image retrieval. For example, if we have text **kitten** as tag of an image, then the confidence about that image contains objects in **cat** category becomes higher. The second item identifies important visual categories that are useful for image retrieval of other visual categories. For example, if we have detected an object of **car**, then the confidence of that image contains **car wheel** objects becomes higher. This kind of visual co-occurrence regularity has been shown to be useful in improving detection precision in NEIL [Chen *et al.*, 2013a].

Our model jointly encodes the intuition of multimodal information and visual co-occurrence regularity, by maximizing the following objective function:

$$\frac{1}{\textbf{card}(\mathcal{I}_\mathcal{D})} \sum_{(D,T)\in\mathcal{I}_\mathcal{D}} \sum_{\substack{u\neq v \\ u,v\in D\cup T}} \log p(v|u) \qquad (4)$$

For each image, we have a set of visual categories D obtained from image detection (e.g. if we detect **car** and **car wheel** objects in an image, then $D = \{\textbf{car}, \textbf{car wheel}\}$), and a set of text phrases T obtained from extracting text information surrounding the image (e.g. if title of an image contains text phrases **autmobile** and **dealer** then $T = \{\textbf{autmobile}, \textbf{dealer}\}$). For an image we calculate the sum of conditional probabilities predicting elements between each other, and the overall objective function maximizes average of these sums. It can be verified that this objective function satisfies the two items we have previously declared.

We observe that the multimodal objective function in Equation (4) can be transformed into objective function of the Skip-Gram model in Equation (1) by:

1. Mapping each object $u_n \in D(I_n) \cup T_n$ to a word w_t;
2. Mapping a set of multimodal objects $D(I_n) \cup T_n$ to words $w_{t-c} \dots w_{t+c}$ around the target word w_t;

Thus, with these transformations we can solve the optimization problem in Equation (4) using the standard Skip-Gram model. In this paper, we have applied the hierarchical softmax version of the Skip-Gram model. After embeddings are learned, we normalize the embedding vectors by dividing their L_2 norms such that all vectors are of unit magnititude.

3.4 Image Tagging

The image tagging program automatically assigns each image a set of noun phrases (tags) that best describe the image [Chen *et al.*, 2013a]. We extract tags of each image based on both image meta and web page context information, following these steps:

- Retrieve top-k noun phrases (denoted as $TopNP$) from a web page that containing the target image to be tagged. The importance of noun phrases are measured by tfidf score
$$tfidf(t,d) = 0.5 + 0.5\frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \times log\frac{N}{n_t},$$
where d represents the webpage document, $f_{t,d}$ is the frequency of term t in document d, N is the total number of webpages, and n_t is the total number of webpages containing the term t. We found $k = 30$ performs well in our system.
- For each noun phrase in $TopNP$, if either *src* or *alt* attribute of $\langle$img$\rangle$ tag contains the noun phrase, then assign it as a tag of the image.
- Retain top-k tags for each image with the highest tfidf scores. We set $k = 3$ in our system.

The Table 1 illustrates some example images annotated with tags. We can see the extracted tags can be inaccurate or incomplete, which makes it challenging to apply the multimodal learning to predict visual objects. In the next section, we shall describe an algorithm using structure learning approach to efficiently select an optimal set of tags that are most useful for prediction.

4 Structure Learning and Prediction

In this section, we present a prediction model based on multimodal embeddings for visual knowledge extraction.

4.1 Multimodal Vocabulary

The multimodal vocabulary, denoted as $\mathcal{W}$, is a union of text vocabulary $\mathcal{V}$ and visual vocabulary $\mathcal{C}$.

$$\mathcal{W} = \mathcal{V} \cup \mathcal{C}$$

In our system, we require a word to have minimum frequency of 5 to be considered as a member of $\mathcal{W}$. Then, an image in Equation (3) can be represented using the $\mathcal{W}$ as

$$\mathcal{I}_\mathcal{D} = \{W_n | W_n = (D(I_n) \cup T_n) \cap \mathcal{W}\},$$

where W_n is a set of words from vocabulary $\mathcal{W}$ corresponding to image I_n. For a concrete example, let's suppose we have an image I_n, where

$$D(I_n) = \{\textbf{car, car wheel}\}$$

indicates that the visual detector can detect objects in categories of **car, car wheel** from I_n, and

$$T_n = \{\text{automobile, dealer}\}$$

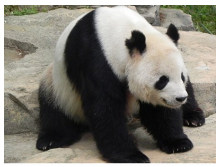
means I_n has two text tags: "automobile" and "dealer". Then, according to the definition,

$$W_n = \{\textbf{car, car wheel}, \text{automobile, dealer}\}. \qquad (5)$$

In the next section, we shall describe an effective learning algorithm to predict if objects of a category (e.g. **car**) present in I_n using W_n.


bear


tours


national zoo, zoo


bears, glacier bay, lewis


bears

Table 1: Example web images with tags automatically annotated

4.2 Structure Learning

We predict the visual objects in an image based on the multimodal words W_n describing the image. Mathematically, we model the probability that an image I_n contains objects of category c with a logistic regression model

$$p_{\boldsymbol{\theta}}(c|W_n) = \frac{e^{\theta_0 + \sum_{w_k \in W_n, w_k \neq c} \theta_k \vec{v}(w_k)^T \vec{v}(c)}}{1 + e^{\theta_0 + \sum_{w_k \in W_n, w_k \neq c} \theta_k \vec{v}(w_k)^T \vec{v}(c)}}, \quad (6)$$

where W_n is a set of multimodal words describing image I_n, and $\theta_0, \theta_1, \ldots$ (collectively denoted as $\boldsymbol{\theta}$) are the bias and combination coefficients. When summarizing over W_n, we exclude the word c because $\vec{v}(w_k)^T \vec{v}(c) = 1$ is constant when $w_k = c$ (all vectors are L_2 normalized). Note that each multimodal word is corresponding to one θ (excluding $w_k = c$), so the total number of parameters is $\textbf{card}(\mathcal{W})$, including bias θ_0. The operator $\vec{v}(\cdot)$ converts a word into its corresponding distributed vector representation as before. In this manner, the probability is determined by weighted combination of W_n. Note that we train one prediction model per category, thus different categories have different $\boldsymbol{\theta}$ parameters.

To learn the model parameters $\boldsymbol{\theta}$ w.r.t. category c, we maximize the following regularized objective function

$$L(\boldsymbol{\theta}, c) = \sum_{n \in \mathcal{P}} \frac{\ln p_{\boldsymbol{\theta}}(c|W_n)}{\textbf{card}(\mathcal{P})} + \sum_{n \in \mathcal{N}} \frac{\ln\left(1 - p_{\boldsymbol{\theta}}(c|W_n)\right)}{\textbf{card}(\mathcal{N})} - \lambda |\boldsymbol{\theta}|_1, \quad (7)$$

where $\mathcal{P}$ denotes a set of indices for image samples in $\mathcal{I}_{\mathcal{D}}$ that are relevant to category c, and $\mathcal{N}$ denotes the irrelevant samples. An image is relevant to category c if visual detector detects objects of category c (e.g. $c \in W_n$). Equation (7) is a balanced and regularized version of the log likelihood. Usually the number of negative training samples is much more than the positive samples, so we balance uneven number of training samples by dividing the cardinalities. The L_1 regularization term is used to encourage sparse solutions while at the same time keep the optimization problem convex. We need sparse solution for $\boldsymbol{\theta}$ because the size of $\mathcal{W}$ is usually very large (e.g. $10^5 - 10^7$) and only a very small fraction of words are useful for specific category c. With proper sparsity level, words that are not useful will have zero θ value, which makes these useless words inactive at prediction stage. Intuitively speaking, in the example of Equation (5), we expect word "image" to have zero θ value when c is **car** because "image" is not closely related to category **car**. On the contrary, if θ value of "image" is nonzero, then this word can easily lead to false positive prediction if θ is positive and false negative detection if θ is negative.

The prediction model in Equation (6) predicts visual objects based on both visual co-detection and text information. For example, assuming that target category c is **car** and W_n is given by Equation (5), then Equation (6) predicts that image I_n contains **car** objects based on weighted combination of: **car wheel**, automobile, image. Intuitively, the confidence about I_n containing **car** objects becomes higher if we know that **car wheel** objects can be detected from I_n, which we called visual co-detection information. Similarly, the same confidence increases as we learn that text tags of I_n contains "automobile", which we called text information.

4.3 Unifying Predictions

The probabilistic model in Equation (6) exploits both visual co-detection and text information. However, this probabilistic function doesn't take the prediction given by object detectors into consideration. To include the confidence predicted by object detectors, denoted as $q(c|I_n)$, we unify the two predictions to give the final scoring function as

$$score(I_n, c) = p_{\boldsymbol{\theta}}(c|W_n) \cdot q(c|I_n), \quad (8)$$

where W_n is a set of multimodal words corresponding to image I_n as before. As a result, for each visual category c we rank all candidate images in $\mathcal{I}_{\mathcal{D}}$ by this confidence score and then retrieve the top candidates as output.

5 Experiments and Results

In this section, we evaluate the proposed multimodal algorithm in the context of Web visual knowledge mining.

5.1 Dataset

We evaluate our approach based on a collection of web pages and images derived from the Common Crawl dataset [Smith *et al.*, 2013] that is publicly available on Amazon S3. The entire Common Crawl dataset comprises billions of raw webpages in warc compressed format, and for our study we take a subset of the data with hundreds of millions of webpages. These webpages are processed following these steps:

1. Parse the HTML webpages, with a C++ open-source program **Gumbo-Parser** by Google[2].
2. Extract all image urls of each web page, along with *alt* and *src* attributes. We only retain images whose dimension (shorest edge) is at least 150 pixels.
3. Clean meta and spam from web pages to obtain plain text, then tokenize and apply part-of-speech tagging. The part of speech tagger is based on **Tree-Tagger** program [Schmid, 1995] for best computational efficiency.

[2]https://github.com/google/gumbo-parser

4. Extract nouns and noun phrases. The nouns are extracted based on part-of-speech tag of a word. The noun phrases are extracted based on the following rules:
   - Common noun phrases (e.g. "computer monitor") consist of a sequence of consecutive common nouns;
   - Proper noun phrases (e.g. "National Aeronautics and Space Administration") are a sequence of proper nouns optionally connected by conjunction or preposition.
5. Assign a set of tags to images by running the image tagging program described in Section 3.4. Then download images from the Web based on image urls in a distributed manner using Amazon S3 and EC2 (40 concurrent spot instances).

This results in a collection of around 10 million tagged images for our study.

5.2 Experimental Settings

In this section, we describe the detailed system parameters, baseline approach and experimental procedures.

System parameters

For each image we retain top 3 tags of the highest **tfidf** scores, and images are resized to fit a 217×217 bounding box. For multimodal embedding, we set the dimension of vector representations as 500 (we found that dimensions between 100 and 1000 give similar results) according to the recommendation from [Frome *et al.*, 2013]. For structure learning, we tune the λ parameter in Equation (7) on training data such that the number of non-zero elements is around 100 for the $\boldsymbol{\theta}$ parameter. We observed that the performance is stable when $\boldsymbol{\theta}$ has number of non-zeros elements between 50 and 200. In this paper, we consider 46 visual categories that are taken from the ImageNet [Deng *et al.*, 2009] database. Each category is initialized with 250 seed images from the ImageNet with annotated object bounding boxes, which are then used to train visual object detector. For visual detection, we use the Fast R-CNN [Girshick, 2015] (a recent state-of-the-art deep learning algorithm) as our object detection program. The reasons we use Fast R-CNN are due to both its computational efficiency and being able to achieve accuracy that is comparable to other state-of-the-art approaches. The CaffeNet models with feature dimension of 4096 were trained on a NVIDIA Tesla K40c GPU. All detection parameters were set as default[3]. The trained detector was then applied to all images, which took around 35 days to complete the detection of 10 million images on a 16-core server machine with NVIDIA Tesla K40c GPU.

Baseline approach

The thesis explored in this paper is that a much greater extraction accuracy for Web visual knowledge extraction can be achieved by exploiting multimodal information. Thus, for each visual category c our comparative baseline ranks all candidate images in $\mathcal{I}_{\mathcal{D}}$ merely based on visual information by

[3]https://github.com/rbgirshick/fast-rcnn

		Precision (%)	
Category	**#Det.**	**Uni.**	**Mul.**
beach	11,457	74	91(+17)
bear	9,751	72	90 (+18)
bed	92,024	95	100 (+5)
bedroom	3,318	96	100 (+4)
bee	4,477	59	70 (+11)
boat	10,215	74	90 (+16)
car	110,898	97	100 (+3)
car mirror	15,756	22	21 (-1)
car tire	10,180	33	48 (+15)
car wheel	75,554	87	100 (+13)
cellular telephone	69,768	93	100 (+7)
chair	81,642	92	100 (+8)
civilian clothing	182,826	69	73 (+4)
computer keyboard	14,959	23	31 (+8)
crane	4,262	31	45 (+14)
female child	402,921	93	100 (+7)
fish	20,694	9	29 (+20)
game fish	3,840	60	84 (+24)
hand-held computer	106,263	95	95 (+0)
helicopter	7,952	96	99 (+3)
insect	29,919	93	99 (+6)
jersey	134,964	100	100 (+0)
kitchen	88,884	82	89 (+7)
laptop	56,676	78	100 (+22)
lifeboat	6,671	63	71 (+8)
locomotive	6,643	83	90 (+7)
man's clothing	300,496	94	99 (+5)
microwave	69,707	83	88 (+5)
musical instrument	67,475	14	10 (-4)
people	468,914	98	100 (2)
pizza	225,795	72	67 (-5)
racer	78,487	94	100 (+6)
riverbed	19,689	78	82 (+4)
salmon	20,652	75	90 (+15)
school bus	6,964	54	49 (-5)
seashore	113,937	68	82 (+14)
skirt	117,309	88	98 (+10)
sky	161,540	95	100 (+5)
suspension bridge	5,841	30	48 (+18)
table	150,542	84	90 (+6)
television	45,690	19	45 (+26)
truck	24,263	92	97 (+5)
vehicle	5,446	88	97 (+9)
wading bird	4,371	91	98 (+7)
warplane	32,506	95	99 (+4)
window	275,872	75	92 (+17)
Average	**81,696**	**72.95**	**81.43 (+8.48)**

Table 2: The comparison of precision between Multimodal (**Mul.**) and Unimodal (**Uni.**) based on top 1,000 detections of each category (totally 46 categories). The **#Det.** is total number of detected objects for a category from 10 million Web images (some images don't contain objects of interest).

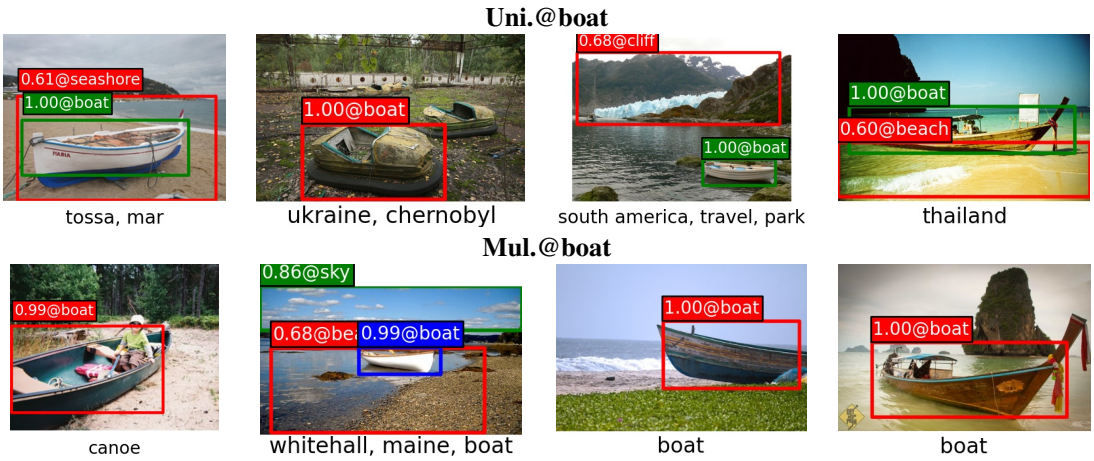


Table 3: The comparison of top extracted objects for the unimodal (**Uni.**) and multimodal (**Mul.**) approaches based on three example visual categories. Each image is visualized with object bounding boxes along with the corresponding confidence scores given by the Fast-RCNN visual object detector. The text descriptions on the bottom of each image are extracted automatically by parsing web pages using our Image Tagging algorithm.

scoring function

$$score(I_n, c) = q(c|I_n), \tag{9}$$

where $q(c|I_n)$ is confidence score predicted by object detector. We used the state-of-the-art visual object detection program (Fast R-CNN) as the baseline object detector.

Experimental procedures

We evaluate our approach by comparing the quality of the extracted visual knowledge. For each visual category, we rank all relevant images by scoring functions in Equation (8) and (9) respectively, and then retrieve the top-k images with the highest scores as output. The precision of output images of category c are estimated by

$$Precision(c, k) = \frac{\#\text{relevant}(S_k, c)}{\textbf{card}(S_k)}. \tag{10}$$

The S_k is a set of images sampled randomly from top-k retrieved images, and $\#\text{relevant}(S_k, c)$ denotes the number of images in S_k that contains correct detection of objects in visual category c. In our paper, k is set to be $1,000$ and the size of S_k is set to be 100. Consequently, we compare the quality of the first $1,000$ retrieved images based on estimation of 100 random samples from that $1,000$ images. Retrieved images of different approaches are mixed together and then the correctness of individual retrieval is verified by human workers.

5.3 Results

Quantative evaluation

We first present our quantitive evaluation results by comparing the proposed multimodal algorithm (**Mul.**) against the baseline unimodal approach (**Uni.**). We estimated the precision of each visual category following the procedures described in Section 5.2, and results are shown in Table 2. In the table, the **#Det.** column is the total number of detected visual objects of a category from 10 million image corpus based on Fast R-CNN detector (may contain false detections). Based on the result in Table 2, we see that for most of the visual categories, the proposed multimodal approach outperforms the unimodal baseline by a clear margin, which confirms the effectiveness of the proposed approach. On average, the multimodal approach has improved the precision by **8.48%**, which is equivalent to a **31%** reduction in error rates.

Illustrative examples

To intuitively examine the effectiveness, we visualize extracted examples as shown in Table 3. Due to the limited space, we only illustrate the top four examples of **boat** category. From these examples, we conclude that the baseline **Uni.** approach extracts objects with the highest visual detection score (1st row), while the proposed **Mul.** approach leverages both text and visual information (2nd row). We also observe that the text description for images retrieved with **Mul.** (2nd row) is more consistent with the visual objects in the images. The second image in the first row is a false positive extraction, which also shows the unreliability of algorithms relying on single source of information.

6 Conclusion

We demonstrate an effective multimodal learning algorithm for visual knowledge extraction from the Web. We have developed a mutimodal visual knowledge extraction system which takes raw web pages as input and extract visual objects leveraging both textual and visual information in a fully automatic manner. Evaluational experiments show that when compared with the state-of-the-art algorithms purely relying on visual content, our multimodal algorithm significantly improves the visual knowledge extraction precision.

Acknowledgements

This work is supported by DARPA under FA8750-12-2-0348-2 (DEFT/CUBISM), and NSF IIS Award #1526753.

References

[Chen *et al.*, 2013a] Minmin Chen, Alice Zheng, and Kilian Weinberger. Fast image tagging. In *Proceedings of the 30th international conference on Machine Learning*, pages 1274–1282, 2013.

[Chen *et al.*, 2013b] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[Gong *et al.*, 2013] Dihong Gong, Zhifeng Li, Dahua Lin, Jianzhuang Liu, and Xiaoou Tang. Hidden factor analysis for age invariant face recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[Gong *et al.*, 2015] Dihong Gong, Zhifeng Li, Dacheng Tao, Jianzhuang Liu, and Xuelong Li. A maximum entropy feature descriptor for age invariant face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[Guthrie *et al.*, 2006] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4, 2006.

[Hodosh *et al.*, 2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[Lazaridou *et al.*, 2015] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Norouzi *et al.*, 2013] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[Perona, 2010] Pietro Perona. Vision of a visipedia. *Proceedings of the IEEE*, 98(8):1526–1534, 2010.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015.

[Schmid, 1995] Helmut Schmid. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28, 1995.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Smith *et al.*, 2013] Jason R Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. In *ACL (1)*, pages 1374–1383, 2013.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[Vijayanarasimhan and Grauman, 2014] Sudheendra Vijayanarasimhan and Kristen Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision*, 108(1-2):97–114, 2014.

[Zhu *et al.*, 2015] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base for visual question answering. *arXiv:1507.05670*, 2015.