ConceptLearner: Discovering Visual Concepts from Weakly Labeled Image Collections

Bolei Zhou[†], Vignesh Jagadeesh[‡], Robinson Piramuthu[‡]
[†]MIT [‡]eBay Research Labs
bolei@mit.edu, [vjagadeesh, rpiramuthu]@ebay.com

Abstract

Discovering visual knowledge from weakly labeled data is crucial to scale up computer vision recognition systems, since it is expensive to obtain fully labeled data for a large number of concept categories. In this paper, we propose ConceptLearner, which is a scalable approach to discover visual concepts from weakly labeled image collections. Thousands of visual concept detectors are learned automatically, without human in the loop for additional annotation. We show that these learned detectors could be applied to recognize concepts at image-level and to detect concepts at image region-level accurately. Under domain-specific supervision, we further evaluate the learned concepts for scene recognition on SUN database and for object detection on Pascal VOC 2007. ConceptLearner shows promising performance compared to fully supervised and weakly supervised methods.

1. Introduction

Recent advances in mobile devices, cloud storage and social network have increased the amount of visual data along with other auxiliary data such as text. Such big data is accumulating at an exponential rate and is typically diverse with a long tail. Detecting new concepts and trends automatically is vital to exploit the full potential of this data deluge. Scaling up visual recognition for such large data is an important topic in computer vision. One of the challenges in scaling up visual recognition is to obtain fully labeled images for a large number of categories. The majority of data is not fully annotated. Often, they are mislabeled or labels are missing or annotations are not as precise as name-value pairs. It is almost impossible to annotate all the data with human in the loop. In computer vision research, there has been great effort to build large-scale fully labeled datasets by crowd sourcing, such as ImageNet [7], Pascal Visual Object Classes [9], Places Database [38] from which the state-of-the-art object/scene recognition and detection systems are trained [18, 12]. However, it is cumbersome and expensive to obtain such fully labeled datasets. Recently, there has been growing interest to harvest visual concepts from internet search engines [2, 8]. These approaches re-rank the search results and then learn concept detectors. The learned detectors largely depend on the quality of image search results, while image search engines themselves have sophisticated supervised training procedures. Alternatively, this paper explores another scalable direction to discover visual concepts from weakly labeled images.

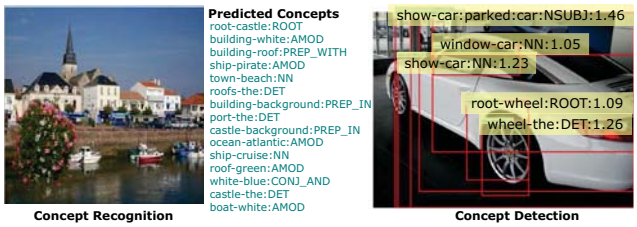

Figure 1. **ConceptLearner:** Thousands of visual concepts are learned automatically from weakly labeled image collections. Weak labels can be in the form of keywords or short description. ConceptLearner can be used to *recognize* concepts at image level, as well as *detect* concepts within an image. Here we show two examples done by the learned detectors.

Weakly labeled images could be collected cheaply and massively. Images uploaded to photo sharing websites like Facebook, Flickr, Instagram typically include tags or sentence descriptions. These tags or descriptions, which might be relevant to the image contents, can be treated as weak labels for these images. Despite the noise in these weak labels, there is still a lot of useful information to describe the scene and objects in the image. Thus, discovering visual concepts from weakly labeled images is crucial and has wide applications such as large scale visual recognition, image retrieval, and scene understanding. Figure 1 shows our concept recognition and detection results by detectors discovered by the ConceptLearner from weakly labeled image collections[1].

[1]Project page is at http://conceptlearner.csail.mit.edu/

The contributions of this paper are as follows:

- scalable max-margin algorithm to discover and learn visual concepts from weakly labeled image collections.
- domain-selective supervision for application of weakly-learned concept classifiers on novel datasets.
- application of learned visual concepts to the tasks of concept recognition and detection, with quantitative evaluation on scene recognition and object detection under the domain-selected supervision.

The rest of the paper is organized as follows. Section 2 gives an overview of related work. Description of the model for weakly labeled image collections is in Section 3. This is followed by max-margin visual concept discovery from weakly labeled image collections using hard instance learning in Section 4. Section 5 shows how we can use the discovered concepts on a novel dataset using domain-selected supervision. We show 3 applications of concept discovery in Section 6. We conclude with Section 7 that gives a summary and a list of possible extensions.

2. Related Work

Discovering visual knowledge without human annotation is a fascinating idea. This idea dates back to the early 90's where a system PICTION [32] identifies human faces in newspaper photographs using associated captions. A real-time system ALIPR was proposed in [20] to recognize hundreds of semantic concepts using example pictures from each concept. Recently there have been a line of work on learning visual concepts and knowledge from image search engines. For example, NEIL [2] uses a semi-supervised learning algorithm to jointly discover common sense relationships and labels instances of the given visual categories; LEVAN [8] harvests keywords from Google Ngram and uses them as structured queries to retrieve all the relevant diverse instances about one concept; [21] proposes a multiple instance learning algorithm to learn mid-level visual concepts from image query results.

There are alternative approaches of discovering visual patterns from weakly labeled data that do not depend strongly on results from search engine. For example, [1] uses multiple instance learning and boosting to discover attributes from images and associated textual description collected from the Internet. [24] learns object detectors from weakly annotated videos. [36, 31] use weakly supervised learning for object and attribute localization, where image-level labels are given and the goal is to localize these tags on image regions. [29] learns discriminative patches as mid-level image descriptors without any text label associated with the learned patch patterns. In our work, we take on a more challenging task where both image and image-level labels are noisy in the weakly labeled image collections.

Figure 2. **NUS-WIDE Dataset [3]:** Images have multiple tags/keywords. There are 1000 candidate tags in this dataset. Here are three examples, with original true tags shown in black, original noisy tags in green, and possible missing tags in red.

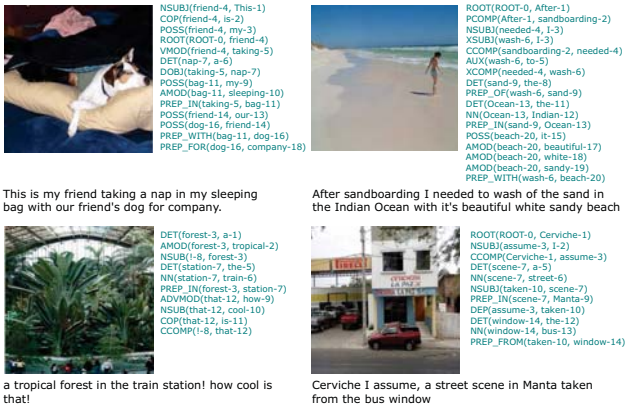

Figure 3. **SBU Dataset [23]:** Each image has a short description. Typically, this is a sentence, as shown below each image. We extract phrases from each sentence, as shown on the side of each image. Each phrase represents a relationship between two ordered words. The relationship is shown in capital letters. For example, AMOD dependency is like attribute+object, PREP are preposition phrases. Details of dependency types can be found in [6].

Other related work include [23, 19, 13, 17], which generate sentence description for images. They either generate sentences by image retrieval [23], or learn conditional random field among concepts [19], or utilize image-sentence embedding [13] and image-fragment embedding [17] to generate sentences. Our work focuses more on learning general concept detectors from weakly labeled data. Note that the predicted labels obtained from our method could also be used to generate sentence description, but it is beyond the scope of this paper.

3. Modeling Weakly Labeled Images

Generally speaking, there are two categories of weakly labeled image collections: (i) multiple tags for each image as in NUS-WIDE dataset [3] and (ii) sentence description for each image as in SBU dataset [23]. Here we analyze the representative weakly labeled image collections NUS-WIDE and SBU dataset respectively.

Figures 2 and 3 illustrate samples from NUS-WIDE dataset [3] and SBU dataset [23]. Note that tags in Figure 2 can be incorrect or missing. In Figure 3 sentences associated with images in [23] are also noisy, as they were written by the image owners when the images were uploaded. Im-

Table 1. Summary of notations used in this paper

Variable[1]	Meaning
$\mathcal{D}$	Collection of weakly labeled images with associated tags (which are used as weak labels) (*i.e.*) $\mathcal{D} = \{(I_i, \mathcal{T}_i) \vert I_i \in \mathcal{I}, \mathcal{T}_i \in \mathcal{T}\}_{i=1}^{N}$
$\mathcal{I}$	Set of all images in $\mathcal{D}$
$\mathcal{T}$	Set of unique tags in $\mathcal{D}$ (*i.e.*) $\bigcup_{i=1}^{N} \mathcal{T}_i$
N	Number of images in $\mathcal{I}$ (*i.e.*) $\vert\mathcal{I}\vert$
T	Number of tags in $\mathcal{T}$ (*i.e.*) $\vert\mathcal{T}\vert$
I_i	An image in $\mathcal{I}$
$\mathcal{T}_i$	The set of tags associated with I_i. For collections with sentence description for each image (as opposed to set of tags/keywords), the extracted phrases using [6] are the weak labels
τ_t	A tag in $\mathcal{T}$ (*i.e.*) $\mathcal{T} = \{\tau_t\}_{t=1}^{T}$
$\mathcal{P}_t$	Set of images associated with tag τ_t (*i.e.*) $\mathcal{P}_t = \{I_i \vert \tau_t \in \mathcal{T}_i\}_{i=1}^{N}$
$\mathcal{N}_t$	Set of images not associated with tag τ_t (*i.e.*) $\mathcal{N}_t = \{I_i \vert \tau_t \notin \mathcal{T}_i\}_{i=1}^{N}$
V	Dimensionality of visual feature vector of an image
$\mathbf{V}$	Stacked visual features for $\mathcal{D}$. Row i is a visual feature vector for image I_i. $\mathbf{V} \in \Re^{N \times V}$
$\mathbf{T}$	Stacked indicator vectors for $\mathcal{D}$. Row t is an indicator vector for image I_i. Entry (i,t) of $\mathbf{T}$ is 1 when tag τ_t is associated with image I_i. It is 0 otherwise. $\mathbf{T} \in [0,1]^{N \times T}$
$\mathbf{w}_c$	SVM weight vector, including the bias term for classifying concept c
$f^{hard}_{\mathbf{w}_c,\eta}(\cdot)$	Operator that takes a set of images and maps to *hard* subset, based on SVM concept classifier $\mathbf{w}_c$ such that $y\mathbf{w}_c \cdot \mathbf{x} < \eta$, where $\mathbf{x}$ is the visual feature vector and $y \in \{-1, 1\}$ label for concept c
$f^{easy}_{\mathbf{w}_c,\eta}(\cdot)$	Operator, similar to $f^{hard}_{\mathbf{w}_c,\eta}(\cdot)$, that takes a set of images and maps to *easy* subset, such that $y\mathbf{w}_c \cdot \mathbf{x} > \eta$
$Rand_k(\cdot)$	Operator that takes a set of images and randomly pick k images without replacement. (*i.e.*) $Rand_k(\mathcal{I}_s) = \{I_{r(j)} \vert I_{r(j)} \in \mathcal{I}_s, \mathcal{I}_s \subset \mathcal{I}\}_{j=1}^{k}$, where $r(j)$ picks a unique random integer from $\{i \vert I_i \in \mathcal{I}_s\}$

[1] Sets are denoted by scripts, matrices by bold upper case, vectors by bold lower case, scalars by normal faced lower or upper case.

age owners usually selectively describe the image content with personal feelings, beyond the image content itself.

There is another category of image collection with sentence description such as Pascal Sentence dataset [25] and Pascal30K dataset [14]. These sentence descriptions are generated by the paid Amazon Mechanical Turk workers rather than the image owners, and are more objective and accurate to the image contents. However the labeling is expensive and not scalable to millions of images. Our approach could work on all of the three categories of weakly labeled image collections, but we focus on the first two

Algorithm 1: ConceptLearner

Data: See Table 1 for notations.
(i) $\mathbf{V}$, matrix of visual feature vectors
(ii) $\mathbf{T}$, matrix of tag indicator vectors

Parameters:

(i) α, ratio of cardinalities of negative and positive instance sets

(ii) M_t, number of image clusters for tag τ_t

(iii) η, threshold to determine hard and easy instances

(iv) K, the top number of tags based on tf-idf for each concept cluster.

Result: (i) Matrix $\mathbf{W}$ of SVM weight vectors, where c^{th} row is concept detector $\mathbf{w}_c^T$ (ii) name set for each concept c

for *label* $t = 1 : T$ **do**
  $c = 0$; */* Initialize concept count */*
  Construct $\mathcal{P}_t, \mathcal{N}_t$;
  Use $\mathbf{V}, \mathbf{T}$ to cluster images $\mathcal{P}_t$ into M_t clusters.
  Each such cluster is a concept;
  for *cluster* $m = 1 : M_t$ **do**
    $c = c + 1$;
    Construct the positive training set
    $\mathcal{P}_t^{train} := \{I_i \vert I_i \in \mathcal{P}_t, I_i \in \text{cluster } m\}$;
    $N_p := \left\vert \mathcal{P}_t^{train} \right\vert$, size of positive training set;
    $N_n := \lceil \alpha N_p \rceil$, size of negative training set;
    Initialize the negative training set
    $\mathcal{N}_t^{train} \leftarrow Rand_{N_n}(\mathcal{N}_t)$;
    / Fix $\mathcal{P}_t^{train}$ and mine hard negative instances */*
    while $\mathcal{N}_t^{train}$ *is updated* **do**
      Train SVM on $\mathcal{P}_t^{train}$ and $\mathcal{N}_t^{train}$ to get weight vector $\mathbf{w}_c$;
      Easy positives $\mathcal{P}_t^{easy} := f^{easy}_{\mathbf{w}_c,\eta}(\mathcal{P}_t^{train})$;
      Hard negatives $\mathcal{N}_t^{hard} := f^{hard}_{\mathbf{w}_c,\eta}(\mathcal{N}_t^{train})$;
      Easy negatives $\mathcal{N}_t^{easy} := f^{easy}_{\mathbf{w}_c,\eta}(\mathcal{N}_t^{train})$;
      Update $\mathcal{N}_t^{train} \leftarrow$
      $\mathcal{N}_t^{hard} \bigcup Rand_{N_n - \vert\mathcal{N}_t^{easy}\vert}(\mathcal{N}_t \setminus \mathcal{N}_t^{easy})$;
    / Cache tag frequency for the positive set */*
    Calculate tag frequency vector $\mathbf{f}_m \in \mathbb{Z}_{\geq 0}^{T}$ based on images in $\mathcal{P}_t^{easy}$;
  / Name each concept using tf-idf across the label frequencies, w.r.t. M_t clusters */*
  Compute tf-idf based on $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_{M_t}\}$;
  Create a name set for each concept $m \in [1, M_t]$, by taking the top K labels based on tf-idf;

more challenging categories.

For image collections with multiple tags, we just take the sparse tag count vector as the weak label feature of each image. For image collections with sentence description, we extract phrases, which are semantic fragments of sentence, as weak label feature for each image. A sentence contains not only several entities such as multiple weak tags for the image, but also contains relationships between the entities. These relationships between entities, composed as phrases, could be easily interpreted and effectively used by human. The phrase representation is more descriptive than a single keyword to describe the image content. Figure 3 shows some examples of extracted phrases from sentences. For simplicity, we adopt the Stanford typed dependencies system [6] as the standard for sentence parsing. All sentences are parsed into short phrases and only those that occur more than 50 times are kept. Note that in [27], 17 visual phrases are manually defined and labeled, corresponding to chunks of meaning bigger than objects and smaller than scenes as intermediate descriptor of the image. In contrast, our approach is data-driven and extracts thousands of phrases from image sentence descriptions automatically. We use these extracted phrases as weak labels for images and learn visual concepts automatically at scale. Notations used in this paper are summarized in Table 1.

4. Max Margin Visual Concept Discovery

Learning visual patterns from weakly labeled image collection is challenging because the labels for training images are noisy. Existing learning methods for this task include semi-supervised learning as in [2] and multiple instance learning as in [1, 21]. In this paper, we formulate this problem as max-margin hard instance learning of visual concepts using SVM.

Since the labels for every image are noisy and there are a lot of missing labels, there is no clear separation of positive set and negative set. If images with a specific label are considered as positive images for that label and images without that label as negative images, there would be a lot of false positives (image with some concept label but has no noticeable image content related to that concept) in the positive set and false negatives (image with some visible concept inside but without that concept labeled) in the negative set. Inspired by the idea of hard instance mining used in face detection and object detection [4, 11], we consider false positives and false negatives as hard instances in the learning of visual concepts. The algorithm will iteratively seek the max-margin decision boundary that separates hard instances.

The detailed steps of our algorithm for concept discovery are listed in Algorithm 1. Our algorithm starts with an initial cache of instances, where the positive set includes all the examples with label t and the negative set is a random sample of images without that label t. In each iteration, we remove easy instances from the cache and add additional randomly selected negative images. The SVM is then retrained on the new cache of positive and negative sets. Here we keep the positive set fixed and only do hard negative instance sampling.

α is the ratio of the number of negatives over the number of positives. Since the number of hard negative instance might be high, we keep a relatively large ratio $\alpha = 5 \sim 10$. On the other hand, as there are various views or subcategories related to the same concept, it is better to learn several sub-category detectors for the same concept than to learn a single detector using all the positive set. Hence we cluster images in each positive set, based on visual features, before learning concept detectors. The cluster number M_t for the t^{th} tag controls the diversity of the learned detectors. Tf-idf [22], short for term frequency inverse document frequency, is used to find the important contextual labels in the label frequency for each sub-categories so that we could better name each learned sub-category detectors.

5. Selecting Domain-Specific Detectors

After the concept detectors are learned, we could directly apply all of them for concept recognition at image-level. But in some applications, we need to apply one concept detector or subset of detectors from the pool of detectors learned from source dataset (say, SBU) to some specific tasks on target dataset (say, Pascal VOC 2007). Here we simply use a winner-take-all selection protocol for the detector selection. We define a selection set, which contains some labeled instances from the target dataset. Then the relevant concept detector with the highest accuracy/precision on the target dataset is selected. Note that the selection set should be separated from the test set of the target dataset. In the following experiments on scene recognition and object detection, we follow this selection protocol to automatically select the most relevant detectors for evaluation on test set. We call this as *domain selected supervision*. This is related to the topic of domain adaptation [28, 33], but we do not use the instances in the target domain to fine-tune the learned detectors. Instead, we only use a small subset of the target domain to select the most relevant concept detectors from a large pool of pre-trained concept detectors. It is also related to the issue of dataset bias [34] existing in current recognition datasets. Domain-selected supervision provides a nice way to generalize the learned detectors to novel datasets.

6. Experiments

We evaluate the learning of visual concepts on two weakly labeled image collections: NUS-WIDE [3] and SBU [23] datasets. NUS-WIDE has 226,484 images (the original set has 269,649 URLs but some of them are invalid

now) with 1000 tags (which were used as weak labels) and 81 ground-truth labels. As shown in [3], the average precision and recall of tags with the corresponding ground-truth labels are both about 0.5, which indicates that about half of the tags are incorrect and half of the true labels are missing. We acquired 934,987 images (the original set has 1M URLs but some of them are invalid now) from SBU dataset. Each image has a text description written by the image owner. Examples from these two datasets are shown in Figures 2 and 3.

The 4096 dimensional feature vector from the FC7 layer of Caffe reference network [16] was used as the visual feature for each image, since deep features from pre-trained Convolutional Neural Network on ImageNet [7] has shown state-of-the-art performance on various visual recognition tasks [26]. Each description was converted to phrases using the Stanford English Parser [5]. Phrases with count smaller than 50 were not used. We used 7437 phrases. Figure 3 shows some sample phrases. These phrases contain rich information such as relationships attribute-object, object-scene, and object-object. We use linear SVM from liblinear [10] in the concept discovery algorithm.

Concepts were learned independently from these datasets using Algorithm 1. Once concepts were learned, we consider 3 different applications: (i) concept detection and recognition, (ii) scene recognition and (iii) object detection. For concept detection and recognition, we chose $M_t = 1$ and $M_t = 4$ for learning concepts from SBU and NUS-WIDE datasets respectively. NUS-WIDE dataset has multiple tags and diverse images for each tag. Tag tuples are assigned to each cluster to capture diversity. Concepts in SBU dataset are grouped by compound relationships of multiple words from dependency parser. Hence we use more clusters when learning concepts from NUS-WIDE dataset. For scene recognition and object detection, we varied $M_t = 1 \sim 10$ to learn the selected concepts and then pooled together all possible concept detectors. Note that M_t was determined empirically, a larger M_t might generate near-duplicate or redundant concept detectors, but it might make the concept pool more diverse. Determining M_t automatically for each label t is part of future work. $\eta \in [0, 1]$ is set empirically at 0.5. For small η, the pool of negatives is small and consists of very hard examples. Due to noisy tags, there is a risk that these are actually true positives. For large η, it takes longer to converge and the richness of easy positives is decreased. The illustration of some learned concept detectors along with the top ranked positive images is shown in Figure 4.

For the concepts learned from NUS-WIDE dataset in Figure 4(a), we show the central concept (cat, boat) in each row along with their variations. The titles show 3 tags of which the first one is the central concept. The other two tags are more contextual words ranked from tf-idf scores associated with the central concept name as the sub-category concept name. We can see that there are indeed sub-categories representing different views of the same concepts, the contextual words ranked using tf-idf well describe the diversity of the same concept. For the concepts learned from SBU dataset, we show 8 learned phrase detectors in Figure 4(b). We can see that the visual concepts well match the associated phrases. For example, cat-in-basket and cat-in-tree describe the cat in different scene contexts; sitting-on-beach and riding-horse describe the specific actions; wooden-bridge and rusty-car describe the attributes of objects. Besides, the top ranked hard negatives are also shown below the ranked positive images. We can see that these hard negatives are visually similar to the images in the positive set.

Figure 4. **Discovered Concepts:** Illustration of learned concepts from NUS-WIDE and SBU datasets. Each montage contains the top 15 positive images for each concept, followed by a single row of 5 negative images. 4 sub-category concept detectors for car and boat respectively are illustrated in (a), based on concepts learned from NUS-WIDE. The title shows the name set for each concept from NUS-WIDE. Phrases for SBU dataset are shown in titles as in (b). We use each tag/phrase (see Figures 2 and 3) to represent a concept and group the associated images together. Within each such group (say, cat, boat), we group images based on visual features only, as we want to have visually similar cluster for one concept. Label vectors (say, car-racing-race, car-automobile-truck, car-automobile-vehicle, car-road-light) are further used to name these clusters after hard instance learning using Algorithm 1. This refined collection of groups is then used to learn concept classifiers and detectors. Examples of positive and negative samples for few such concept classifiers are shown in this figure.

To evaluate the learned concept detectors, we use images from the SUN database [37] and Pascal VOC 2007 object detection dataset [9]. These are independent from the NUS-

WIDE and SBU datasets where we discover the concept detectors. We first show some qualitative results of concept recognition and detection done by the learned detectors. Then we perform quantitative experiments to evaluate the learned concept detectors on specific vision tasks through domain-selected supervision (Section 5), for scene recognition and object detection respectively. Compared to the fully supervised methods and weakly supervised methods, our domain-selected detectors show very promising performance[3].

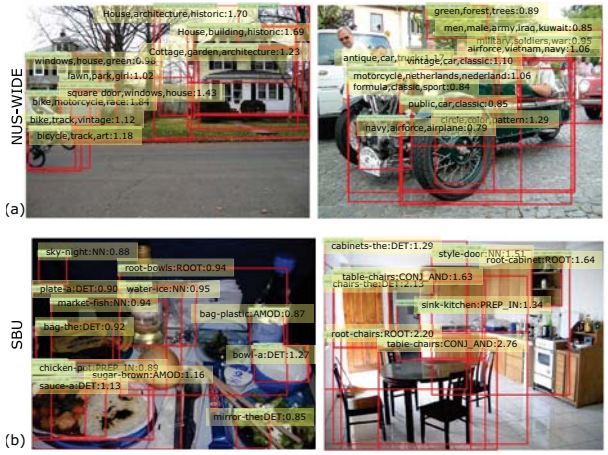

Figure 5. **Concept Detection:** Results of concepts discovered from (a) NUS-WIDE and (b) SBU. Top 20 bounding boxes with high detector responses are shown. Note that for legibility we manually overlaid the text labels with large fonts.

6.1. Concept Recognition and Detection

We apply the learned concept detectors for concept recognition at image level and concept detection at image regions. After the deep feature $\mathbf{x}_q$ for a novel query image I_q is extracted, we multiply the learned detector matrix with the feature vector to get the response vector $\mathbf{r} = \mathbf{W}\mathbf{x}_q$, where each element of the vector is the response value of one concept. Then we pick the most likely concepts of that image by simply sorting the response values on $\mathbf{r}$.

We randomly take the images from SUN database [37] and Pascal VOC 2007 as query images, the recognition results by concept detectors learned from NUS-WIDE and SBU datasets are shown in Figure 6. We can see that the predicted concepts well describe the image contents, from various aspects of description, such as attributes, objects and scenes, and activities in the image.

Furthermore, we could apply the learned concept detectors for concept detection at the level of image regions. Specifically, we mount the learned concept detectors on a detection system similar to the front-end of Region-CNN [12]: Selective search [35] is first used to extract region proposals from the test image. Then CNN features of region proposals are extracted. Finally the deep features of every region proposal are pre-multiplied with the detector matrix and non-maximum suppression is used to merge the responses of the overlapped region proposals. The concept detection results are shown in Figure 5. We can see that this simple detection system mounted with learned concept detectors interprets the images in great detail.

[3]More experimental results are included in supplementary materials.

Table 2. Accuracy and mean average precision (mAP) of baseline, NUS-WIDE concepts and SBU concepts. Mean $\pm$ std is computed from 5 random splits of training and testing (Section 6.2)

Method	Supervision	Accuracy	mAP
Baseline (strong)	Full	69.0$\pm$0.6	59.6$\pm$0.8
NUS concepts (weak)	Selected	55.5$\pm$1.8	47.0$\pm$0.4
SBU concepts (weak)	Selected	60.0$\pm$1.2	50.6$\pm$0.7

6.2. Scene Recognition on SUN database

Here we evaluate the learned concept detectors for scene recognition on the SUN database [37] which has 397 scene categories. We firstly use the scene name to select the relevant concept detectors from the pool of learned concepts *i.e.* the scene name appears in the name of some concept detector. There are 37 matched scene categories among the concept pool of SBU and the concept pool of NUS-WIDE. We take all the images of these 37 scene categories from SUN database and randomly split them into train and test sets. The size of the training set is 50 images per category. We train a linear SVM on the train set as the fully supervised baseline. Note that this baseline is quite strong, since linear SVM plus deep feature is currently the state-of-the-art single feature classifier on the SUN database [38].

To evaluate the learned concepts, we use the domain-selected supervision introduced in Section 5. The train set is used as the selection set. 37 best scene detectors are selected out from the concept pool of SBU and NUS-WIDE based on their top mAP on the selection set, then they are evaluated on the test set. A test image is classified into the scene category which has the highest detector response. Without calibration of detector responses, the classification result is already reasonably good.

The accuracy and mean average precision (mAP) of the fully supervised baseline and our domain-selected supervised methods are listed in Table 2. The AP per category for the three methods are plotted in Figure 7(a). We can see that the SBU concept detectors perform better than the NUS-WIDE concept detectors because of larger amount of data. Both of the learned concept detectors have good performance, compared to the fully supervised baseline with strong labels. SBU concept detectors even outperform the baseline for mountain, castle, marsh, and valley categories shown in Figure 7(a). The concept detectors perform worse on some scene categories like village, hospital, and wave,

Figure 6. **Concept Recognition:** Illustration of concept recognition using concepts discovered from (a) NUS-WIDE and (b) SBU datasets. Top 5 and 15 ranked concepts are shown respectively. These predicted concepts well describe the objects, the scene contexts, and the activities in these images.

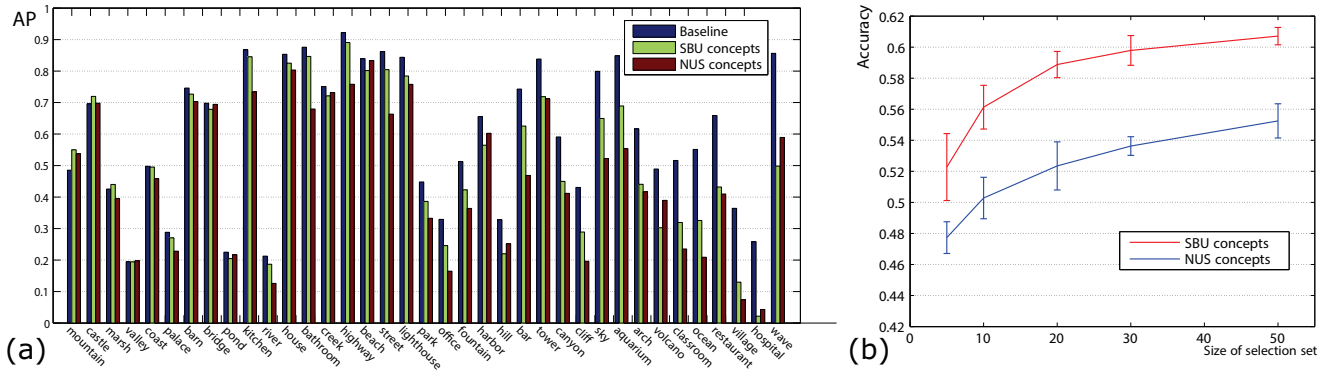

Figure 7. **Scene Recognition on SUN:** (a) AP per category for three methods, ranked by the gap between learned concepts and fully supervised baseline. SBU concept detectors from weak labels outperform the baseline for mountain, castle, marsh, and valley. Concept detectors perform worse for village, hospital, and wave, due to lack of sufficient positive examples in the weakly labeled image collections (b) Recognition accuracy over the size of selection set. Domain-specific detectors work well when there are only a few samples in the selection set.

because there are not so many good positive examples in the weakly labeled image collections.

In Figure 7(b), we further analyze the influence of selection set on the performance of our method. We randomly select the subset of images from the train set as the selection set for our method, we can see that the SBU concepts still achieve 52.5% accuracy when there are only 5 instances per category as the selection set to pick the most relevant concept detectors. It shows that the domain-selected supervision works well even with few samples from the target domain.

6.3. Object Detection on Pascal VOC 2007

We further evaluate the concept detectors on Pascal VOC 2007 object detection dataset. We follow the pipeline of region proposal and deep feature extraction in [12] for the

Table 3. Comparison of methods with various kinds of supervision on Pascal VOC 2007. NUS-WIDE has missing entries since some object classes don't appear in the original tags. See Section 5 for details on "selected" supervision. If we pool the concept detectors learned from SBU dataset and concept detectors learned from NUS-WIDE dataset together as initial concept pool, the final mAP is **23.2**.

Method	Supervision	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbik	pers	plant	sheep	sofa	train	tv	mAP
SBU	Selected	34.5	**39.0**	**18.2**	14.8	8.4	31.0	39.1	20.4	**15.5**	13.1	**14.5**	3.6	20.6	33.9	**9.4**	**17.0**	14.7	**22.6**	27.9	**19.0**	**20.9**
NUS-WIDE	Selected	**34.6**	38.5	16.5	**18.7**	-	27.0	**43.6**	**24.6**	10.9	9.3	-	**20.4**	30.3	**36.6**	3.0	4.7	13.6	-	**36.1**	-	-
CVPR'14 [8]	Webly	14.0	36.2	12.5	10.3	**9.2**	**35.0**	35.9	8.4	10.0	**17.5**	6.5	12.9	**30.6**	27.5	6.0	1.5	**18.8**	10.3	23.5	16.4	17.2
ECCV'12 [24]	Video	17.4	-	9.3	9.2	-	-	35.7	9.4	-	9.7	-	3.3	16.2	27.3	-	-	-	-	15.0	-	-
ICCV'11 [30]	Weakly	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
ICML'14 [31]	Weakly	7.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
CVPR'14 [12]	Full	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7

validation and test sets of Pascal VOC 2007. Under domain-selected supervision (Section 5), we first select the learned concept detectors which have the object name inside their name and compute the AP for each of them on the validation set (thus the validation set of Pascal VOC 2007 is our selection set). Then we evaluate the selected 20 best concept detectors for all 20 objects in VOC 2007 respectively.

Table 3 displays the results obtained using our concept discovery algorithm on NUS-WIDE and SBU datasets and compares the state-of-the-art baselines with various kinds of supervision. CVPR'14 [12] is the R-CNN detection framework, a fully supervised state-of-the-art method on Pascal VOC 2007. It uses the training set and validation set with bounding boxes to train the object detectors with deep features, then generates region proposal and deep feature for testing (we use the scores without fine-tuning). ICML'14 [31] is the state-of-the-art method method for weakly supervised approaches on Pascal VOC 2007. It assumes that there are just image level labeling on the train set and validation set without bounding boxes to train the object detectors. It uses R-CNN framework to compute features on image windows to train the detectors and to generate region proposals and deep features for testing. ICCV'11 [30] is another weakly supervised method using DPM. Since all these three methods only use the training set and validation set of Pascal VOC 2007 to train the detector, they are relevant to our method as "upper bound" baselines.

Another two most relevant comparison methods are the webly supervised method [8] and video supervised method [30]. Webly supervised method uses items in Google N-grams as queries to collect images from image search engine for training the detectors. So their training set of detector could be considered as the unlimited number of images from search engines. Video supervised method [30] trains detectors on manually selected videos without bounding boxes and shows results on 10 classes of Pascal VOC 2007. Since these two methods train detectors on other data source then test on Pascal VOC 2007, which is similar to our scenario, we consider them as direct comparison baselines. Our method outperforms these two methods with better AP on majority of the classes. Besides, if we pool the concepts learned from SBU dataset and concepts learned from NUS-WIDE dataset together as the initial concept pool, the final mAP reaches **23.2**, which even outperforms the weakly supervised method in ICML'14 [31].

7. Conclusion and Future Work

In this paper, we presented ConceptLearner, a max-margin hard instance learning approach to discover visual concepts from weakly labeled image collection. With more than 10,000 concept detectors learned from NUS-WIDE and SBU datasets, we apply the discovered concepts to concept recognition and detection. Based on domain-selected supervision, we further quantitatively evaluate the learned concepts on benchmarks for scene recognition and object detection, with promising results compared to other fully and weakly supervised methods.

There are several possible extensions and applications for the discovered concepts. Firstly, since there are thousands of the concepts discovered, some concept detectors have overlaps. For example, as the predicted labels in the second example in Figure 6(b), there are 'market-fruit:NN','market-local:AMOD','market-a:DET','market-vegetable:NN','market-farmers:NN', and 'fruit-market:PREP IN', which are redundant to describe the same image. Thus some bottom-up or top-down clustering methods could be used to merge the similar concept detectors or to merge the predicted labels for a query image. Besides, some measures could be introduced to characterize the properties of learned concepts, such as the visualness [15] and localizability [1]. Then the subset of concept detectors could be grouped and used in a specific image interpretation task. Meanwhile, in concept recognition and concept detection, since every concept is detected independently, some spatial or co-occurrence constraints could be defined and used to filter out some outlier concepts detected in the same image, in the context of all the other detected concepts. Besides, with the grammatical structure integrated, the predicted phrases and tags could be further used to generate a full sentence description for the image.

References

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *Proc.*

ECCV. 2010.

[2] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proc. ICCV*, 2013.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[5] M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, 2006.

[6] M.-C. De Marneffe and C. D. Manning. Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*, 2008.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[8] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proc. CVPR*, 2014.

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge–a retrospective. *Int'l Journal of Computer Vision*, 2014.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 2008.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2010.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[13] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *Proc. ECCV*. 2014.

[14] P. Y. A. L. M. Hodosh and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. 2013.

[15] J.-W. Jeong, X.-J. Wang, and D.-H. Lee. Towards measuring the visualness of a concept. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012.

[16] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding, 2013.

[17] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. Technical report, 2014.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[19] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proc. CVPR*, 2011.

[20] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):985–1002, June 2008.

[21] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *Proc. CVPR*, 2013.

[22] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[23] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[24] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *Proc. CVPR*, 2012.

[25] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[27] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Proc. CVPR*, 2011.

[28] A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, and S. L. DuVall. Active supervised domain adaptation. In *Machine Learning and Knowledge Discovery in Databases*. 2011.

[29] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proc. ECCV*. 2012.

[30] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *Proc. ICCV*, 2011.

[31] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, T. Darrell, et al. On learning to localize objects with minimal supervision. In *International Conference on Machine Learning*, 2014.

[32] R. K. Srihari. Piction: A system that uses captions to label human faces in newspaper photographs. In T. L. Dean and K. McKeown, editors, *AAAI*, pages 80–85. AAAI Press / The MIT Press, 1991.

[33] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012.

[34] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proc. CVPR*, 2011.

[35] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *Int'l Journal of Computer Vision*, 2013.

[36] S. Wang, J. Joo, Y. Wang, and S.-C. Zhu. Weakly supervised learning for attribute localization in outdoor scenes. In *Proc. CVPR*, 2013.

[37] J. Xiao, J. Hays, K. A. Ehinger, A. Torralba, and A. Oliva. Sun database: Exploring a large collection of scene categories. 2014.

[38] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.