Deep Multiple Instance Learning for Image Classification and Auto-Annotation

Jiajun Wu[1], Yinan Yu[2,3], Chang Huang[2], Kai Yu[2]
[1]Massachusetts Institute of Technology [2]Institute of Deep Learning, Baidu [3] Tsinghua University

Abstract

The recent development in learning deep representations has demonstrated its wide applications in traditional vision tasks like classification and detection. However, there has been little investigation on how we could build up a deep learning framework in a weakly supervised setting. In this paper, we attempt to model deep learning in a weakly supervised learning (multiple instance learning) framework. In our setting, each image follows a dual multi-instance assumption, where its object proposals and possible text annotations can be regarded as two instance sets. We thus design effective systems to exploit the MIL property with deep learning strategies from the two ends; we also try to jointly learn the relationship between object and annotation proposals. We conduct extensive experiments and prove that our weakly supervised deep learning framework not only achieves convincing performance in vision tasks including classification and image annotation, but also extracts reasonable region-keyword pairs with little supervision, on both widely used benchmarks like PASCAL VOC and MIT Indoor Scene 67, and also a dataset for image- and patch-level annotations.

1. Introduction

Deep learning, as a recent breakthrough in artificial intelligence, has been successfully applied to multiple fields including speech recognition [12] and visual recognition [16, 19, 15, 18], mostly with full supervision. A typical deep learning architecture for visual recognition builds upon convolutional neural network (CNN) [17, 19, 16, 38]. Given large-scale training data and the power of high-performance computational infrastructure, deep learning has achieved tremendous improvement in visual recognition with thousands of categories [6].

While deep learning shows superior performance on fully supervised learning tasks, research on learning deep representations with weak supervision is still in its early stage; *i.e.*, human labels still play a key role in these popular frameworks [12, 16]. This is in a sense anathema to the very nature of large-scale web or real-world data — namely, big data is largely data with no labels or noisy labels. The emergence of image search engines like Google, social network sites like Facebook, and photo and video sharing sites like Flickr provides vision researchers with abundant visual data; unfortunately, strong labels for these images are in much shorter supply. Therefore, unsupervised or weakly-supervised methods are particularly favored as they can better utilize the available large-scale web resources.

Weakly supervised learning can in general be viewed as mechanisms to learn from sparse or noisy labels. As web data usually comes with high diversity but much noise, these weakly supervised methods have been successfully applied to learn effective visual representations for classification [27], detection [27, 11], and segmentation [11], all using weak labels alone.

In terms of visual recognition, people have recently proposed a number of techniques to generate object proposals for higher level tasks, apart from traditional exhaustive searches. These approaches either adopt saliency information [13], train generic object models to harvest "objectness" [4], or turn to more adaptive segmentation systems [33, 41], all of which can be viewed as effective ways of reducing the search space.

These proposal generating algorithms usually have very high recalls but adequate precisions, which indicates that although proposals may be noisy, there is almost always an object of interest within a number of most likely proposals [4]. We observe that this property actually corresponds to the assumption in multiple instance learning, which states that there must be at least one positive instance within each positive bag. Therefore, we attempt to incorporate multiple instance learning into a deep learning framework and apply the learned visual knowledge to assist the task of image classification.

We also notice that the multiple instance assumption widely exists in other domains, *e.g.*, image annotation (tagging), a task which both vision and natural language processing communities are interested in. Modern search engines like Google, Bing, and Baidu can already perform image keywording in a fully unsupervised way, though the

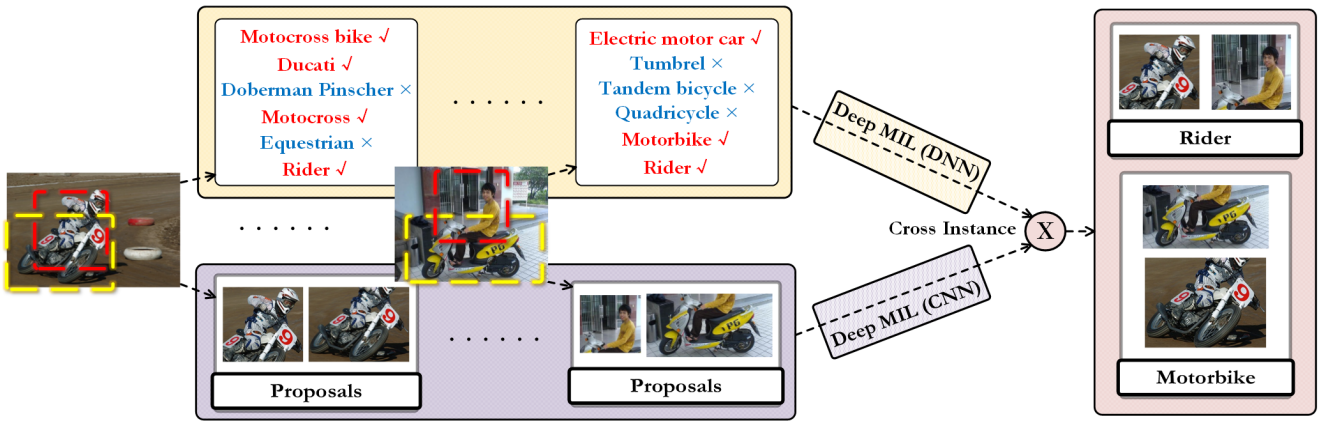

Figure 1. Overview of our framework and the dual multiple instance assumptions

returned keywords might not be accurate enough. This is another type of "noisy input" and can be naturally modeled as a multiple instance learning problem if we consider each tag as an instance and the tags for a certain image as a bag. Here we also develop a deep multiple instance learning framework to identify the relevant keywords for images.

We further attempt to jointly learn keywords and object instances of interests among candidates, and propose a system to automatically extract insightful keyword-proposal pairs with little supervision. As existing datasets mostly provide image-level annotations [6, 21] or assign pixel/region to a single label [39], we construct a new dataset where both images and patches are associated with multiple relevant tags, which can be used for image- and patch-level annotations or region-keyword pairing.

Our contributions are three-fold: first, we observe the generic existence of the multiple instance assumption in object and keyword candidates; second, we incorporate deep learning into a weakly supervised learning framework in a principled manner; third, we demonstrate that our deep multiple instance learning system achieves convincing performance in both image classification and image annotation.

2. Related Work

Deep architectures consist of feature detector units arranged in layers. Lower layers detect simple and local features and then feed them into higher layers, which in turn capture more complex features [12, 16, 19, 15]. Recently, convolutional neural network (CNN) has been successfully applied in many vision tasks including object recognition [16], image classification [17], and video classification [15].

In the machine learning literature, Dietterich et al. [7] introduced multiple instance learning (MIL) for drug activity prediction. Since then, researchers have proposed a large number of algorithms for the MIL tasks. For example, Andrews *et al*. [1] developed mi-SVM and MI-SVM for instance-level classification and bag-level classification, respectively. There have been some explorations [31, 48, 5] on solving multiple instance learning using traditional neural networks. Their approaches are inspiring; however, neither did they consider learning deep representations, nor did they study computer vision tasks. Numerous computer vision applications can naturally fit into the MIL framework. Examples include object and face detection [43], visual categorization [42], segmentation [45] and image retrieval [24].

Most deep learning approaches are in fully supervised settings. Recently, researchers started to study weakly supervised learning using features learned with deep representations [47, 35]. Specifically, Xu *et al*. [47] proposed to use deep learning to compute features for multi-instance learning in medical imaging; Song *et al*. [35] also used CNN features for weakly supervised object localization. Different from these methods, we propose an integrated framework to learn deep representations with MIL assumptions for the tasks of image classification and annotation.

Russell *et al*. [33] first proposed to use multiple segmentations for higher level vision tasks. These years, researchers have developed a number of methods for finding salient regions or detecting generic objects, *e.g*., spectral methods [13] and those adopting machine learning techniques [25]. Recently, selective search [41], a novel segmentation technique for object recognition, has been shown to be effective in object detection. Note that Zhu *et al*. [49] also studied multiple instance learning with salient windows, but with a focus on unsupervised object discovery and without learning deep representations.

In terms of image classification, PASCAL Visual Object Classes [10] has long been a popular benchmark. Deep learning systems from different research groups all reported impressive performance on the classification task [3, 9, 36]. In this paper, we combine deep learning with multi-instance

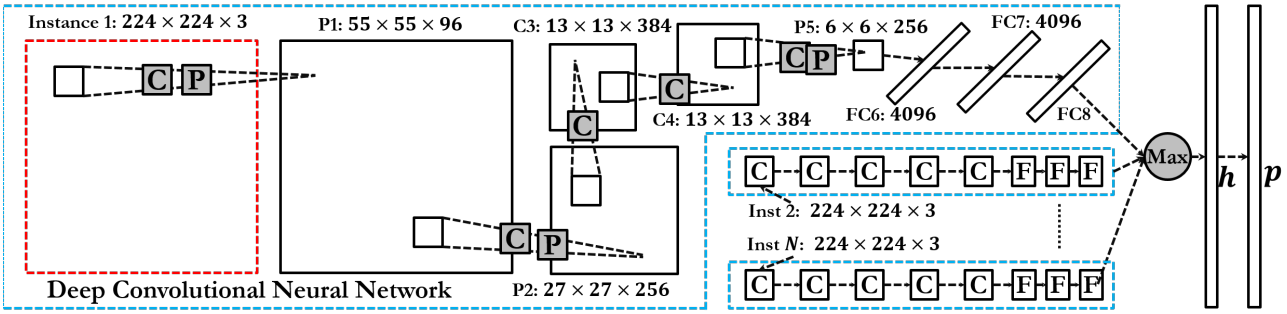

Figure 2. Illustration of our framework for learning deep visual representations within a multiple instance learning setting. Here P stands for a pooling layer, C for a convolution layer, and FC for a fully connected layer.

learning and utilize object-level information to assist the task of classification.

Image annotation aims at producing rich image descriptions at different levels [26]. During the past decades, there have been numerous inspiring works in this area [2, 20, 44]. To name a few, Barnard *et al*. [2] presented some correspondence models on matching segmented images with associated text; Li *et al*. [20] proposed a real-time annotation framework for Internet images; Wang and Forsyth [44] made progress on jointly learning attributes and object classes via multi-instance learning. Different from these methods which did not learn deep representations, our deep multiple instance learning framework can achieve high accuracy on both image classification and annotation.

3. Deep Multiple Instance Learning (DMIL)

In this section, we present our method for learning deep representations in a weakly supervised manner. Based on the existence of the multiple instance property in both object and keyword proposals, we attempt to unify the learned deep features within the MIL framework.

3.1. The Setting of DMIL

As described earlier, recent approaches are able to generate a number of object proposals with very high recalls. In this sense, we notice that it is reasonable to assume that the object lies in at least one of the proposals. In other words, it becomes natural to treat the object proposals of each image as a positive bag in multiple instance learning.

From a different angle, there have been many techniques for collecting keywords from the web for a given image. These keywords alone are often too noisy for tasks like image classification. However, it is justifiable to assume that there must be at least one relevant keyword within a number of most confident keywords. This again corresponds to the multiple instance assumption. These findings encourage us to design a multi-instance learning scheme to jointly learn about visual objects and verbal keywords.

As we know, different from traditional supervised learning in which training instances are given as pairs $\{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{-1, 1\}$ is the corresponding label. In multiple instance learning, data are organized as bags $\{\mathbf{X}_i\}$, and within each bag there are a number of instances $\{\mathbf{x}_{ij}\}$. Labels $\{Y_i\}$ are only available at the bag level, while labels of instances $\{y_{ij}\}$ are unknown. Given the MIL assumption lies generally in object and keyword proposals, we therefore propose to exploit this property by incorporating multiple instance learning into a deep learning framework.

3.2. Our Formulation

Considering the recent advances achieved by deep learning, it is a natural choice to employ deep representations instead of a shallow model. We use deep convolutional neural network as our architecture for learning visual representation with multiple instance learning. The structure is inspired by [17]. As shown in Figure 2, it contains five convolutional layers, followed by a pooling layer and three fully connected layers.

We redesign the last hidden layer for multiple instance learning. Given one training sample x, the network extracts layer-wise representations from the first convolutional layer to the output of the last fully connected layer $fc_8 \in \mathbb{R}^m$, which can be viewed as high level features of the input image. Followed by a softmax layer, fc_8 is transformed into a probability distribution $\mathbf{p} \in \mathbb{R}^m$ for objects of m categories, and cross entropy is used to measure the prediction loss of the network. Specifically, we have

$$p_i = \frac{\exp(h_i)}{\sum_i \exp(h_i)}, i = 1, \ldots, m, \text{ and } L = -\sum_i t_i \log(p_i), \tag{1}$$

where L is the loss of cross entropy. The gradients of the deep convolutional neural network is calculated via back-propagation

$$\frac{\partial L}{\partial h_i} = p_i - t_i, \tag{2}$$

Figure 3. Difference between human labels and automatically crawled keywords

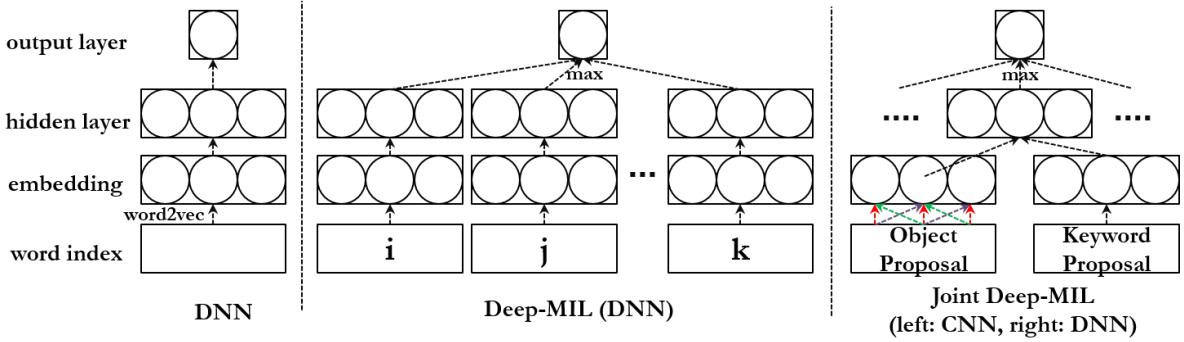

Figure 4. Comparison of our joint deep multiple instance learning framework for learning correspondences between keywords and image regions, with the DMIL framework and traditional DNN for image keywording

where $t = \{t_i | t_i \in \{0,1\}, i = 1,\ldots,m, \sum_{i=1}^{m} t_i = 1\}$ denotes the true label of the sample x.

To learn multiple instances as a bag of samples, we incorporate deep representation with multiple instance learning. Denote $\{\mathbf{x_j}|j = 1,2,\ldots,n\}$ as a bag of n instances and $t = \{t_i|t_i \in \{0,1\}, i = 1,\ldots,m\}$ as the label of the bag; a multiple instance convolutional neural network extracts representations of the bag: $h = \{h_{ij}\} \in R^{m\times n}$, in which each column is the representation of an instance. The aggregated representation of the bag for MIL is:

$$\hat{h_i} = f(h_{i1}, h_{i2}, \ldots, h_{in}), \tag{3}$$

where function f can be $\max_j(h_{ij})$, $\text{avg}_j(h_{ij})$, or $\log\left[1 + \sum_j \exp(h_{ij})\right]$, among others. Here we continue our reasoning with the $\max(\cdot)$ layer, but formulations with other choices are straightforward. Also in Section 4.5, we show experiments with these possible choices.

The distribution of visual categories of the bag and the loss L are therefore

$$p_i = \frac{\exp(\hat{h_i})}{\sum_i \exp(\hat{h_i})} \quad \text{and} \quad L = -\sum_i t_i \log(p_i). \tag{4}$$

In order to minimize the loss function of the DMIL, we employ stochastic gradient descent for optimization. The gradient is calculated via back propagation [32],

$$\frac{\partial L}{\partial \hat{h_i}} = p_i - t_i \quad \text{and} \quad \frac{\partial \hat{h_i}}{\partial h_{ij}} = \begin{cases} 1, & h_{ij} = \hat{h_i} \\ 0, & \text{else} \end{cases}. \tag{5}$$

For the task of image classification, we first employ existing methods to generate object proposals within each image; we then apply the deep multiple instance learning framework to perform image classification.

3.3. Automatic Image Annotation

We now explain our method for automatic image annotation. Again, besides purely using deep learning for keyword extraction and image annotation, we integrate deep features and weakly supervised learning to truly find out discriminative and relevant keywords for each image.

Keywords Extraction from Web Data: Human-labeled datasets, *e.g.*, PASCAL VOC [10] and MIT Indoor Scene 67 [30], usually come with a few entry-level labels. In contrast, images on the web are connected with rich documents, consisting of titles, captions, alternate texts, and articles in webpages, all of which may describe images in more detail. There have been abundant classical techniques to extract keywords from texts [40]. In this paper, we use a much simpler strategy to extract keywords: given an arbitrary image I, we firstly use the image similarity search engine from Baidu to find a set of most similar images $\{I'\}$ from the

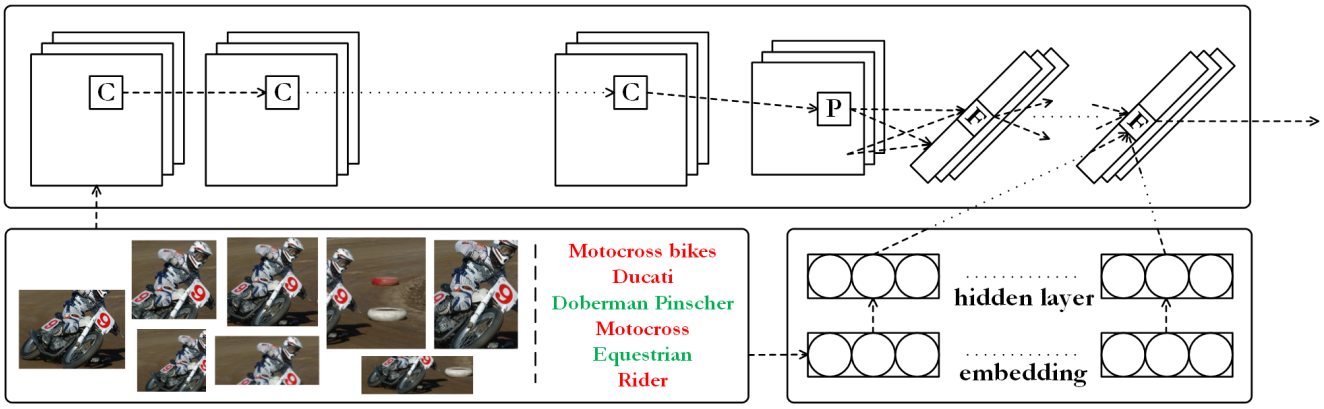

Figure 5. Illustration of our framework for jointly learning image regions and keywords. Here P stands for a pooling layer, C for a convolution layer, and F for a fully connected layer.

web. We then crawl the surrounding documents of each retrieved image I'. The nouns which appear in the surrounding documents are considered as the keywords of the image. Figure 3 provides some examples. The keywords extracted from webpages are highly noisy. However, although many of the keywords are irrelevant to the given image, some words actually provide more detailed and more informative descriptions than the category label does (for instance, F-22 Raptor vs aeroplane). This offers us an opportunity to obtain more specific image annotations than human-labeled tags.

Image Recognition by Keywords: we subsequently aim to predict image category from keywords. As discussed before, keywords gathered from the web, as a type of "noisy input", fit the multi-instance assumption well. Here we use another deep neural network formulation with multi-instance learning. As shown in Figure 4, the deep network contains one input layer, one hidden layer, and one output layer with softmax. Instead of using original word indices as input, a 128-dimensional word-to-vector feature is used to relieve the computational burden. For this task, we keep the same learning strategy as that for object proposals.

3.4. Joint Learning of Image Regions and Keywords

We now consider a novel framework for learning correspondences between image regions and keywords, which serves as the basis for patch-level annotations. Object proposals and keywords are two sets of instances satisfying the multiple instance assumption; a cross combination of the regions and the words leads to the possibility that we can label regions with proper words. We build a joint deep multi-instance learning architecture to learn the object proposals and keywords simultaneously.

Specifically, we combine the outputs of image and text understanding systems in the final fully connected layer, as illustrated in Figure 5. This can be viewed as a straightforward generalization of the aggregate equation Eq. 3. Now we have

$$\hat{h_i} = f\begin{pmatrix} h_{i11} & h_{i12} & \dots & h_{i1n} \\ h_{i21} & h_{i22} & \dots & h_{i2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{im1} & h_{im2} & \dots & h_{imn} \end{pmatrix}, \quad (6)$$

where m is the number of keywords and n is the number of patches. Because m is not large, this intuitive formulation is also computationally affordable.

4. Experiments

In this section, we conduct experiments of our weakly supervised deep learning framework on both image classification and image auto-annotation. We test our method on two widely used datasets for object and scene classification, PASCAL VOC 2007 [10] and MIT Indoor Scene 67 [30]. For image annotation, we apply our framework on PASCAL VOC as well as a new dataset for both image-level and region-based annotations.

4.1. Setup

4.1.1 Datasets

PASCAL 07: The PASCAL Visual Object Classes 2007 database [10] contains 9963 images of 20 categories including people, animals, and various objects. This dataset is considered more challenging than datasets like ILSVRC [6] as the objects are not centered and their appearances are more diverse.

MIT Indoor: The MIT Indoor Scene 67 dataset [30] contains 15620 images of 67 categories of scenes. The dataset consists of various types of indoor scenes including public spaces, stores, leisure places, working places, and residential rooms. Many of these indoor scenes are highly similar to each other, which makes the dataset especially difficult compared to traditional outdoor scene datasets.

	aero	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	hrs	mbk	per	plant	shp	sofa	train	tv	mAP
GHM [3]	76.7	74.7	53.8	72.1	40.4	71.7	83.6	66.5	52.5	57.5	62.8	51.1	81.4	71.5	86.5	36.4	55.3	60.6	80.6	57.8	64.7
AGS [9]	82.2	83.0	58.4	76.1	56.4	77.5	88.8	69.1	62.2	61.8	64.2	51.3	85.4	80.2	91.1	48.1	61.7	67.7	86.3	70.9	71.1
NUS [36]	82.5	79.6	64.8	73.4	54.2	75.0	77.5	79.2	46.2	62.7	41.4	74.6	85.0	76.8	91.1	53.9	61.0	67.5	83.6	70.6	70.5
CNN-SVM [34]	91.2	81.4	82.1	81.1	51.6	81.6	84.4	83.9	54.5	61.0	53.8	72.3	74.9	75.6	83.7	47.4	71.7	60.0	88.3	79.4	73.0
DMIL (region)	92.9	81.6	86.0	82.5	53.9	81.8	86.8	83.4	53.7	66.8	51.8	72.3	79.4	77.3	86.1	50.1	74.6	61.7	90.3	80.1	74.7
DMIL (keyword)	81.4	70.3	76.1	71.3	34.7	66.7	71.8	68.1	50.7	49.7	37.0	55.0	57.9	63.9	71.1	46.5	59.0	43.4	87.2	73.9	61.8
DMIL (joint)	93.5	83.4	86.9	83.6	54.2	81.6	86.6	85.2	54.5	68.9	53.8	73.2	78.8	79.0	86.6	51.2	74.4	63.7	91.5	80.4	75.5

Table 1. Image classification results on Pascal VOC 2007, compared to other methods which also use outside training data. The CNN representation is not tuned for the Pascal VOC dataset. Note that in comparison, GHM [3] jointly learns from bag-of-words and contextual information on VOC. By clustering the VOC data, AGS [9] learns layered representation. NUS [36] learns a codebook for descriptors from VOC. CNN-SVM [34] is the popular OverFeat representation.

	aero	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	hrs	mbk	per	plant	shp	sofa	train	tv	avg
DNN	65.2	50.6	73.0	34.9	9.9	60.9	61.3	77.0	40.8	59.8	20.5	69.9	20.1	68.5	48.2	18.8	66.0	54.7	39.4	40.6	49.0
DMIL (ours)	80.9	51.9	69.5	72.1	22.6	53.4	73.1	74.2	43.9	73.2	46.9	85.6	48.9	67.1	34.7	14.7	80.4	57.8	25.1	52.8	56.5

Table 2. Image annotation results on Pascal VOC 2007

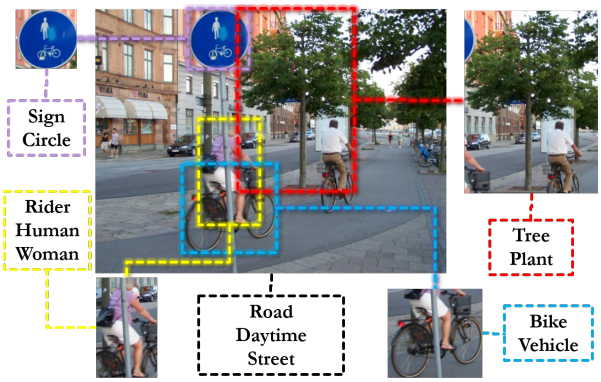

Figure 6. Sample data from the new dataset. Note that human labeling helps to remove noisy tags and construct a clean dataset with both image-level (the black box) and patch-level (the colored boxes) annotations.

Dataset for annotation: We evaluate on a new dataset specifically designed for both image- and patch-level annotation (tagging). The dataset contains images of 50 categories, 25 of which are the most popular object categories in ImageNet [6] and the others are the most popular scene categories in SUN database [46]. For each category, we collect 50 images from these existing databases and the web.

For each of the 2500 images, we manually label bounding boxes for several salient objects. We then apply the keyword extraction technique described in Section 3.3 with Baidu search engine to collect keywords while restricting them to be from a dictionary of 981 nouns for both the image and these boxes. The 981 nouns are chosen from a set of most frequently searched keywords on Baidu. To remove noise, five external experts are invited to decide whether each tag is correct or not, and we retain those tags that are endorsed by at least four of the five experts. Figure 6 provides a snapshot of our dataset.

4.1.2 Metrics and Measures

For image classification on PASCAL VOC 2007 dataset, we adopt the traditional mean average precision (mAP) as our evaluation metric. Following [8] and [14], We use mean accuracy as the evaluation metric for image classfication on MIT Indoor Scene 67 dataset and for image annotation.

4.1.3 Implementation Details

Following previous works [3, 9, 36], we first conduct pre-training of CNN on the ILSVRC dataset [6]. Based on the parameters obtained from pre-training, we then train the DMIL framework on the PASCAL 07 training set. We use BING [4] as the proposal generating system. For each image, windows with confidence scores larger than -0.97 are retained for further use.

4.2. Image Classification on PASCAL VOC 2007

Table 1 shows the results of our deep multiple instance learning (DMIL) system on image classification. As mentioned earlier, the performance is measured in mean average precision (mAP). As our system uses training data outside the standard Pascal VOC 2007 dataset, we also compare the results only with those methods which have used outside training data. We can see that DMIL outperforms previous efforts by a significant margin in mean average precision. Specifically, compared to the popular OverFeat

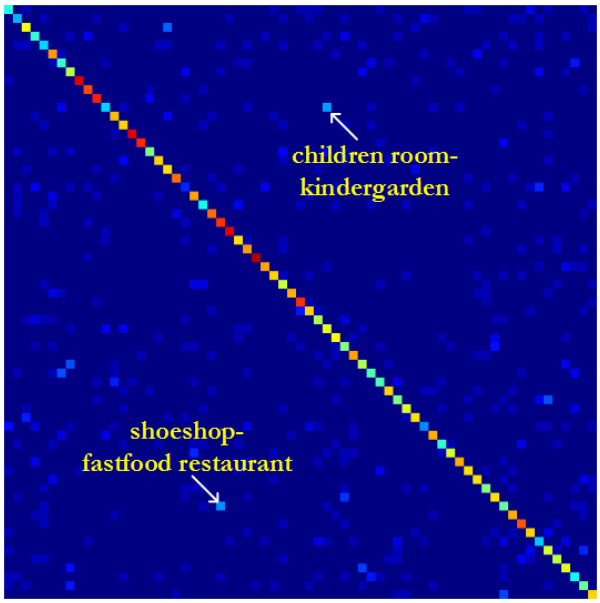

Figure 7. Confusion matrix for MIT Indoor 67. Some of the mistakes (children room-kindergarden) are very hard even for humans.

(CNN-SVM) architecture [34], it has superior average precision on 14 out of 20 classes. Also, both region proposals and keyword proposals contribute to the overall performance, although regions proposals play a more central role. Note that all baselines in Table 1 require to learn dataset-specific knowledge (codebook, contextual information) from VOC, while our representation is not fine-tuned for the VOC data.

4.3. Image Classification on MIT Indoor Scene 67

Table 3 shows the results of our deep multiple instance learning (DMIL) system on MIT Indoor Scene 67. We measure our performance in mean accuracy. Figure 7 is the confusion matrix of our DMIL system. As we can see, DMIL achieves encouraging performance and outperforms most of other methods including CNN-SVM, and again both patch and keyword modules are helpful. As illustrated by Figure 7, some of the mistakes made by DMIL are even hard for humans to distinguish. Please note that the dimension of the Improved Fisher Vector (IFV) [14] representation is over $200,000$, while DMIL only employs a feature vector of length $4,096$. Also, the MLrep [8] requires fine tuning on the dataset which takes several weeks.

4.4. Image Annotation

Here we show the application of our framework in image annotation. As illustrated earlier, our method can perform both image-level and region-based annotations. We also provide some exemplar patch-keyword pairs automatically extracted by our system.

Methods	mAcc
ROI+Gist [30]	26.1
DPM [28]	30.4
Object Bank [22]	37.6
RBow [29]	37.9
BoP [14]	46.1
miSVM [23]	46.4
D-Parts [37]	51.4
IFV [14]	60.8
MLrep [8]	64.0
CNN-SVM [34]	57.7
DMIL (region)	60.0
DMIL (keyword)	48.3
DMIL (joint)	61.2

Table 3. Comparison of classification results on MIT Indoor Scene 67. Note that IFV [14] needs a feature representation of length over $200,000$, and MLrep [8] employs the very time-consuming fine tuning on the dataset.

4.4.1 Image-level annotation

We perform image keywording on the test set of PASCAL VOC 2007. We compare our system with a simple deep learning system without the multi-instance learning layer. We evaluate the returned keywords in a class-wise manner. For each image in a specific class, we choose the top one keywords returned by the two systems, and invite external experts to decide whether these keywords are relevant to the object class. For instance, if an image of class "car" is labeled as "hatchback", based on our statistics, the expert would regard the annotation correct; if it is labeled as "bedroom", then it would be considered as an error.

We then compute for each class the accuracy of top one keyword. As shown in Table 2, we find that DMIL comes with very convincing performance. In 13 of 20 classes, DMIL achieves a higher accuracy than a straightforward deep formulation. The average accuracy grows from 49.0% to 56.5%. Specifically, for classes where objects are often localized in images, like "boat", "bottle", "dog", and "table", DMIL provides a significant increase in performance, which indicates that the multiple instance assumption assists in finding objects of interest in image annotation.

4.4.2 Patch-level annotation

Given our framework for learning cross-instance (image regions and keywords) relations, it is intuitive to obtain patch-level annotations. This is also similar to object localization, although the patches we use here are not necessarily for objects. Here we also test our system for learning region-keyword pairs on PASCAL VOC 2007 dataset.

For each test image, we keep the most confident region-

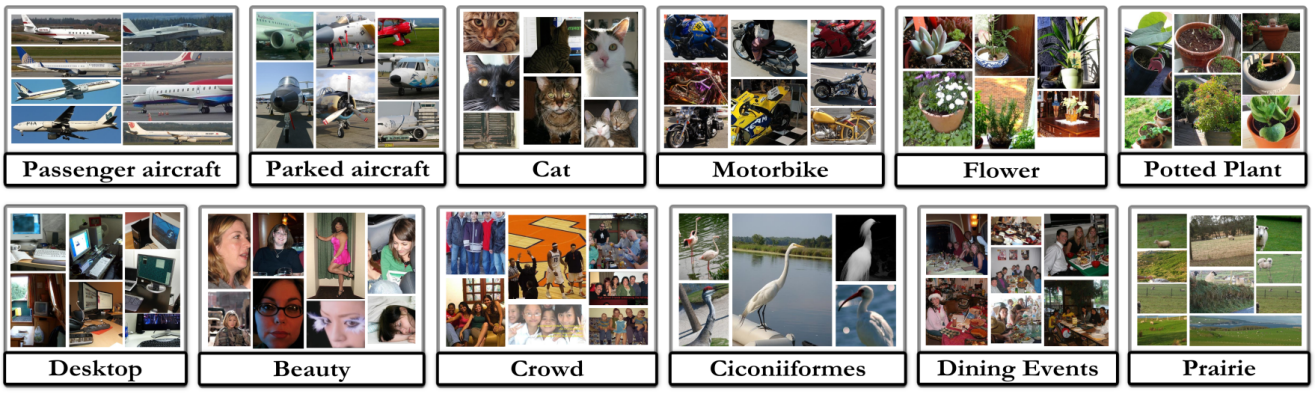

Figure 8. Sample results for patch-level annotations. Note that the image patches are randomly sampled from all patches associated with the keyword in some returned patch-keyword pair.

	CNN-SVM [34]	DMIL (ours)		
		max	avg	log
PASCAL 07	73.0	75.5	72.6	74.7
MIT Indoor	57.7	61.2	59.6	59.3

Table 4. Classification results of CNN-SVM and DMIL with different layers on PASCAL 07 and MIT Indoor 67

	DNN	DMIL (ours)		
		max	avg	log
PASCAL 07	49.0	56.5	50.2	53.3
Annotation (Image)	55.0	62.6	57.3	62.2
Annotation (Patch)	42.0	51.5	46.7	48.5

Table 5. Annotation results of DNN and DMIL with different layers on PASCAL 07 and the new dataset

keyword pair, and group these pairs by keywords. Some of the results are demonstrated in Figure 8. For each keyword listed, the patches shown in the figure are randomly sampled from all patches that are associated with the keyword in some pair. We can see that our system can describe objects (*e.g.*, cat, motorbike, desktop) and scenes (*e.g.*, prairie), as well as recognize activities (*e.g.*, dining events) and perform fine-grained annotation (*e.g.*, parked aircraft, ciconiiformes). All these information can assist tasks like classification, which indicates directions for future research.

4.4.3 Annotation on new dataset

We then evaluate our framework on the newly proposed Words dataset. Here we provide quantitative results on both image- and patch-level annotations. In either case, we first limit the output space of any system to the dictionary of 981 nouns, mentioned in Section 4.1.1. If the top one keyword returned by the system is in the list of tags of that image or patch, we regard this prediction as correct.

As shown in Table 5, again, DMIL achieves an evident and consistent performance boost over DNN on both image- and patch-level annotations on Words dataset.

4.5. Choice of Hidden Layer

The choice of hidden layer is of critical importance in our formulation. In Table 4 and 5, we present results of our formulation with a variety of hidden layers for both image classification and annotation. As mentioned in 3.2, the $\max(\cdot)$, $\text{avg}(\cdot)$, and $\log(\cdot)$ in Table 4 and 5 refer to $\max_j(h_{ij})$, $\text{avg}_j(h_{ij})$, and $\log\left[1+\sum_j \exp(h_{ij})\right]$ for aggregating instance representations, respectively.

We notice that in almost all cases, the straightforward $\max(\cdot)$ layer obtains the best performance. Considering that the $\max(\cdot)$ layer fits the multiple instance assumption best, these empirical results confirm our observation that exploiting the multiple instance property lying in both visual and verbal levels can assist in these vision tasks.

5. Conclusion

In this paper, we proposed to construct a deep learning framework within a weakly supervised learning setting. We demonstrated that our observation of the universal existence of the multiple instance assumption contributes much in solving computer vision tasks, and the deep multiple instance learning system we developed performs well in both image classification and image auto-annotation. Our system is also able to automatically extract correspondences between object and keyword proposals and return meaningful region-keyword pairs on widely used benchmarks. We hope our findings could arouse further research in the fields of deep learning and weakly supervised learning in the vision community.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002. 2

[2] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003. 3

[3] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan. Hierarchical matching with side information for image classification. In *CVPR*, 2012. 2, 6

[4] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014. 1, 6

[5] A. d'Avila Garcez and G. Zaverucha. Multi-instance learning using recurrent neural networks. In *IJCNN*, 2012. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 5, 6

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71, 1997. 2

[8] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013. 6, 7

[9] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan. Subcategory-aware object classification. In *CVPR*, 2013. 2, 6

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2, 4, 5

[11] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. Weakly supervised object localization with stable segmentations. In *ECCV*. 2008. 1

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012. 1, 2

[13] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 1, 2

[14] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013. 6, 7

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 2

[16] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierarchies for visual recognition. In *NIPS*, 2010. 1, 2

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2, 3

[18] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015. 1

[19] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009. 1, 2

[20] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE TPAMI*, 30(6):985–1002, 2008. 3

[21] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007. 2

[22] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010. 7

[23] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013. 7

[24] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang. Text-based image retrieval using progressive multi-instance learning. In *ICCV*, 2011. 2

[25] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2011. 2

[26] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007. 3

[27] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009. 1

[28] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 7

[29] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb. Reconfigurable models for scene recognition. In *CVPR*, 2012. 7

[30] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 4, 5, 7

[31] J. Ramon and L. De Raedt. Multi instance neural networks. In *ICML workshop on attribute-value and relational learning*, 2000. 2

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. 4

[33] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 1, 2

[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014. 6, 7, 8

[35] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, 2014. 2

[36] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011. 2, 6

[37] J. Sun, J. Ponce, et al. Learning discriminative part detectors for image classification and cosegmentation. In *ICCV*, 2013. 7

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 1

[39] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*. 2010. 2

[40] P. D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, 2000. 4

[41] K. E. van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 1, 2

[42] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In *CVPR*, 2008. 2

[43] P. A. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2006. 2

[44] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 3

[45] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *CVPR*, 2014. 2

[46] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, pages 1–20, 2014. 6

[47] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, E. I. Chang, et al. Deep learning of feature representation with multiple instance learning for medical image analysis. In *ICASSP*, 2014. 2

[48] Z.-H. Zhou and M.-L. Zhang. Neural networks for multi-instance learning. In *ICIIT*, 2002. 2

[49] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu. Unsupervised object class discovery via saliency-guided multiple class learning. In *CVPR*, 2012. 2