Multiple Instance Visual-Semantic Embedding

Zhou Ren[1]
zhou.ren@snap.com
Hailin Jin[2]
hljin@adobe.com
Zhe Lin[2]
zlin@adobe.com
Chen Fang[2]
cfang@adobe.com
Alan Yuille[3]
alan.yuille@jhu.edu

[1] Snap Inc.
64 Market Street,
Venice, CA, USA

[2] Adobe Inc.
345 Park Avenue,
San Jose, CA, USA

[3] Department of Cognitive Science and Computer Science
Johns Hopkins University
Baltimore, MD, USA

Abstract

Visual-semantic embedding models have been recently proposed and shown to be effective for image classification and zero-shot learning. The key idea is that by directly learning a mapping from images into a semantic label space, the algorithm can generalize to a large number of unseen labels. However, existing approaches are limited to single-label embedding, handling images with multiple labels still remains an open problem, mainly due to the complex underlying correspondence between an image and its labels. In this work, we present a novel Multiple Instance Visual-Semantic Embedding (MIVSE) model for multi-label images. Instead of embedding a whole image into the semantic space, our model characterizes the subregion-to-label correspondence, which discovers and maps semantically meaningful image subregions to the corresponding labels. Experiments on two challenging tasks, multi-label image annotation and zero-shot learning, show that the proposed MIVSE model outperforms state-of-the-art methods on both tasks and possesses the ability of generalizing to unseen labels.

1 Introduction

Image classification is a fundamental problem in computer vision. Most existing work focuses on single-label image classification [3, 17, 34], where each image is assumed to have only one class label. Classic approaches try to categorize images into a fixed set of classes by training a multi-class classifier. However, due to the complex relationships among classes (*e.g.*, hierarchical, disjoint, *etc.*), it is hard to define a perfect classifier encoding all those semantic relations [4]. Also, since the set of classes is predefined, such systems need to be re-trained whenever a new visual entity emerges.

To address these shortcomings, visual-semantic embedding models [7, 15, 24, 29] were recently proposed, which leverage semantic information contained in unannotated text data

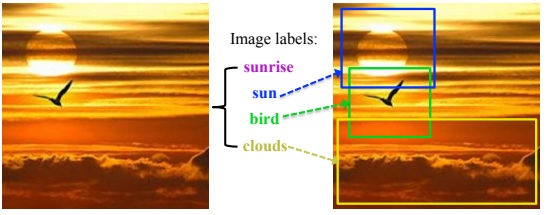

Figure 1: (a) An example of image with multiple labels; (b) We observe that different labels may correspond to various image subregions, but not necessarily the whole image, such as the labels *clouds*, *sun*, *bird*, which are associated with the subregions in the bounding boxes.

to learn semantic relationships among labels, and explicitly map images into a rich semantic space (in such space, certain semantic relationships are encoded, *e.g.*, related labels like *sun* and *sunrise* locate at close positions. And images from these two classes may share some common visual appearance.). By resorting to classification in the semantic space with respect to a set of label embedding vectors, visual-semantic models have shown comparable performance to state-of-the-art visual object classifiers and demonstrated *zero-shot learning* capability, *i.e.*, the ability to predict unseen image categories without training with them.

Although visual-semantic embedding models have shown impressive results for images with single labels, no attempts have been made on optimizing it for multi-label annotation. It is important to develop such a model due to the following reasons. Firstly, real-world images are often associated with multiple description labels. Multi-label annotation is a practical and challenging task, since the image labels can be diverse, which may describe the image foreground, image background, or the whole image. Secondly, it is nontrivial to extend a single-label visual-semantic embedding model to a multi-label one. The implicit assumption that each label corresponds to the whole image does not hold for multi-label cases. For a typical multi-label image, some labels may correspond to different image subregions, instead of the whole image, as shown in Figure 1.

Hence, in this paper, we present a novel Multiple Instance Visual-Semantic Embedding (MIVSE) to model images and their corresponding multiple textual labels. Our method characterizes an image-subregion-to-label correspondence by learning an embedding function that maps semantically meaningful image subregions to their corresponding labels in the semantic space. In order to discover those semantically meaningful subregions, we construct a bag of subregion instances using state-of-the-art region-proposal method [16]. And we propose an objective function that models such correspondence with a weighting scheme encoded to optimize the label prediction. The proposed model has shown superior performance in multi-label image annotation, which outperforms state-of-the-art method [10] on NUS-WIDE dataset, besides, it possesses the generalizing ability to predict unseen labels, which outperforms existing methods [7, 24] on MIT Places205 dataset.

2 Related Work

Multi-label image annotation. Modeling images and their corresponding textual labels have attracted increasing interest recently. Early work in this area focused on learning statistical models based on hand-crafted features [11, 21, 35, 38]. As the learned image represen-

tation using deep convolutional neural network (CNN) has shown superior performance in various vision tasks [8, 17, 19, 28, 30, 32, 34, 36], recently Gong *et al*. [10] applied the CNN architecture to multi-label image annotation problem and achieved nice performance. With the help of additional metadata, methods [12, 14] further boost the annotation performance. However, since the metadata used in [12, 14] were collected separately, for fair comparison, we do not utilize metadata in this paper. Without metadata, [10] is state-of-the-art method, which is classes-predefined and thus lacks the ability of generalizing to unseen labels.

It is also important to note the difference between multi-label image annotation and attribute prediction [1]. Attributes [6, 26] are commonly used to encode the visual properties of objects. However, labels are different from attributes, since attributes are on object-level while labels are on image-level. For example, in Figure 1, *circular* can be an attribute of the object *sun*, but it is not a suitable label of this image.

Zero-shot learning. Zero-shot learning is commonly used to evaluate the generalizing ability of a system, whose goal is to classify images from untrained classes. Early work [22, 25, 31, 33] attempted to solve this problem relying on curated source of semantic information of the labels, such as a knowledge base of labels, the WordNet hierarchy, *etc*. Recently, visual-semantic embedding models [7, 24, 29] were proposed to leverage semantic representation learned directly from unannotated text data online. There are various attribute-based zero-shot learning methods in the literature [18, 20]. However, because of the difference between attributes and labels as explained before, in this paper, we only handle multi-label zero-shot learning problem in the context of label prediction.

3 Overview of Visual-Semantic Embedding

Given a multi-label image dataset $\mathcal{D} \equiv \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, each image is represented by a d-dimensional feature vector, $\mathbf{x}_i \in \mathcal{X} \overset{def}{=} \mathbb{R}^d$, and each image is associated with multiple labels, $\mathbf{y}_i = (y_i^1, ..., y_i^{T_i})$, where the number of labels T_i can be varied across different images. There are total n distinct labels in dataset $\mathcal{D}$, *i.e.*, $y_i^t \in \mathcal{Y} \equiv \{1, 2, ..., n\}, \forall i, t$. Previous methods [10, 11] formulate multi-label image annotation as a label assignment problem, which predefine a fixed set of class labels $\mathcal{Y}$, and learn to predict the labels given image input, *i.e.*, $\mathcal{X} \rightarrow \mathcal{Y}$. However, such classes-predefined approaches lack the ability of generalizing to unseen labels, and need to be retrained when a new label emerges. Given a system trained with labels from $\mathcal{Y}$, it is infeasible to apply it to images with unseen labels from $\mathcal{Y}'$, if $\mathcal{Y} \cap \mathcal{Y}' = \emptyset$.

Fortunately, visual-semantic embedding (VSE) models [7, 24] have been proposed to address this issue for single-label image classification. Instead of learning a mapping from images to the labels ($\mathcal{X} \rightarrow \mathcal{Y}$), it aims to construct a continuous semantic space $\mathcal{S} \equiv \mathbb{R}^e$ which captures the semantic relationship among all labels in $\mathcal{Y} \cup \mathcal{Y}'$, and explicitly learn the embedding function from images to such space, $f : \mathcal{X} \rightarrow \mathcal{S}$. The semantic space $\mathcal{S}$ is constructed such that two labels y and y' are semantically similar if and only if their semantic embeddings $\mathbf{s}(y)$ and $\mathbf{s}(y')$ are close, where $\mathbf{s}(y)$ is the semantic embedding vector of label y in $\mathcal{S}$. Thus the training and unseen test labels become related via the semantic space $\mathcal{S}$. Once $f(\cdot)$ is learned, it can be applied to a test image $\mathbf{x}'$ to obtain $f(\mathbf{x}')$, and this image embedding vector of $\mathbf{x}'$ is then compared with the unseen label embedding vectors, $\{\mathbf{s}(y'); y' \in \mathcal{Y}'\}$. This allows VSE models to generalize to unseen labels.

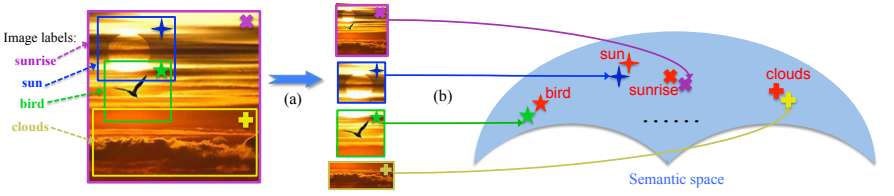

Figure 2: Illustration of our Multiple Instance Visual-Semantic Embedding model, which is composed of two key components: (a) construct image subregion set; (b) establish the subregion-to-label correspondence by embeding semantically meaningful subregions close to their corresponding labels in the semantic space (the red symbols illustrate the embedding of text labels, and symbols of other colors indicate that of different image subregions.). Note that the bounding boxes are for visualization only; they are not provided in training.

3.1 Constructing the semantic label space

Constructing $\mathcal{S}$ is the first step. Distributed representations of texts [23, 27] have shown the capacity to provide semantically meaningful embedding features for semantically related text terms. In this paper, we utilize the GloVe model [27] to construct a 300-dim semantic label space $\mathcal{S}$ which embodies the semantic relationship of labels.

3.2 Naive multi-label VSE baseline

Unfortunately, existing VSE models only handle single label images. We can extend a VSE model to multi-label scenerio by developing the following loss function. Such straightforward extension serves as our multi-label baseline:

$$L_{\text{rank}}(\mathbf{x}_i, \mathbf{y}_i) = \sum_{y_p \in \mathbf{y}_i^+} \sum_{y_q \in \mathbf{y}_i^-} \max\left(0, m + ||f(\mathbf{x}_i) - \mathbf{s}(y_p)||_2^2 - ||f(\mathbf{x}_i) - \mathbf{s}(y_q)||_2^2\right), \tag{1}$$

where m is the margin that we cross-validate, $f(\mathbf{x})$ is the embedding vector of image $\mathbf{x}$ in $\mathcal{S}$, $\mathbf{s}(y)$ is the embedding vector of label y in $\mathcal{S}$. $\mathbf{y}_i^+$ denotes the ground truth label set of $\mathbf{x}_i$, *i.e.*, $\mathbf{y}_i^+ = \mathbf{y}_i$, and $\mathbf{y}_i^-$ denotes the negative label set excluding the labels in $\mathbf{y}_i$, *i.e.*, $\mathbf{y}_i^- = \mathcal{Y} \setminus \mathbf{y}_i$.

4 Multiple Instance Visual-Semantic Embedding

However, the naive extension of VSE models from single-label to multi-label cases is non-trivial. The baseline model in Equation 1 has a critical problem: each image $\mathbf{x}_i$ may correspond to multiple labels in $\mathbf{y}_i$. One or more of those labels could be located far away from others in the semantic space $\mathcal{S}$, such as *bird* and *sun* in Figure 2. Trying to push the embedding of a whole image, $f(\mathbf{x}_i)$, to be close to multiple distant points in $\mathcal{S}$ will confuse the embedding function. In the worst case, the image could be mapped to a near average position of those label embedding vectors, which might correspond to a totally different concept.

The key observation for overcoming this problem is that different image labels often correspond to different subregions in the image. For example, in Figure 2, the image on the left has four labels. Among them only *sunrise* corresponds to the whole image and other labels correspond to image subregions in the bounding boxes. This motivates us to derive a new idea for multi-label embedding, in which one generate multiple subregion proposals from an image (including the whole image) and use the resulting subregion set to match

the labels in the semantic space. This requires a subregion-to-label correspondence to be constructed on-the-fly during the learning process. Now we introduce our solution in stages.

4.1 Modeling subregion-to-label correspondence

Inspired by Multiple Instance Learning [5], we first construct a bag of image subregion instances. In order to interpret each ground truth label, there should be at least one subregion that maps close to it. The subregion with the closest distance to a certain label is more likely to represent that label. Thus, we define a preliminary loss function of MIVSE as follows:

$$L'_{\text{MIVSE}}(\mathbf{x}_i,\mathbf{y}_i) = \sum_{y_p\in\mathbf{y}_i^+}\sum_{y_q\in\mathbf{y}_i^-} \max(0, m + \min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-\mathbf{s}(y_p)||_2^2 - \min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-\mathbf{s}(y_q)||_2^2), \tag{2}$$

where $\mathcal{C}_i$ is the set of all subregions of image $\mathbf{x}_i$ (we will introduce how to obtain $\mathcal{C}_i$ later in Section 4.3.2), $\mathbf{x}_i^c$ indicates one subregion of image $\mathbf{x}_i$, $\mathbf{y}_i^+$ denotes the ground truth label set.

Given numerous subregion candidates in $\mathcal{C}_i$, we want to map meaningful subregions close to the labels it interprets, thus the rank of predicted labels is essential. One limitation of the preliminary loss function in Equation 2 is that it does not explicitly optimize the label prediction. Thus, we propose a weighting scheme to optimize such ranking.

4.2 Optimizing the label prediction

The rank r of a label y is defined as:

$$r = \sum_{y_t\neq y,\ y_t\in\mathcal{Y}} \mathbb{1}\left(\min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-\mathbf{s}(y_t)||_2^2 \leq \min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-\mathbf{s}(y)||_2^2\right), \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function. As we see, given a label y, we rank it according to its minimal distance to all image subregions, *i.e.*, by $\min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-\mathbf{s}(y)||_2^2$. By optimizing the ranking of labels, we are supposed to encourage ground truth (positive) labels to have smaller matching-distance than the negative labels. Thus, we give larger penalties to false predictions of ranking positive labels. The final loss function of MIVSE is defined as follows:

$$L_{\text{MIVSE}}(\mathbf{x}_i,\mathbf{y}_i) = \sum_{y_p\in\mathbf{y}_i^+}\sum_{y_q\in\mathbf{y}_i^-} w(r_p)\cdot\max(0, m + \min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-\mathbf{s}(y_p)||_2^2 - \min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-\mathbf{s}(y_q)||_2^2), \tag{4}$$

where r_p is a positive label y_p's rank, $w(\cdot)$ is a weighting function empirically defined as:

$$w(r) = \begin{cases} 1 & \text{if } r < |\mathbf{y}_i^+|, \\ r & \text{otherwise.} \end{cases} \tag{5}$$

Note that we use exact ranks, which is a major difference with the approximated-weighting methods. If a ground truth label is ranked within top-$|\mathbf{y}_i^+|$, we give small and constant penalty weights to its loss. Otherwise, we assign larger weight. The intuition is to push the positive labels to top ranks, thus mapping meaningful subregions closer to the true labels.

4.3 Learning MIVSE model

4.3.1 MIVSE network architecture.

Figure 3 illustrates the overall network architecture of our MIVSE model. As discussed in previous section, we utilize the GloVe model [27] to extract 300-dim label embedding features $\mathbf{s}(y)$. And we adopt GoogleNet [34] to extract the 1024-dim image feature $\mathbf{x}$. To

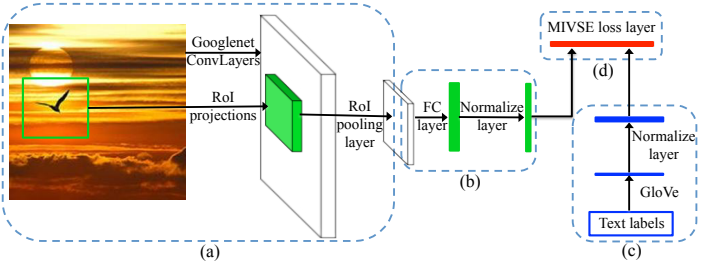

Figure 3: Network architecture of MIVSE, composed of 4 components: (a) subregion image features extraction; (b) image features embedding; (c) text features embedding; (d) embedding learning guided by the MIVSE loss layer.

learn the embedding from image to semantic space, $f : \mathcal{X} \to \mathcal{S}$, a fully connected layer is used following the image feature output. And two L_2 normalize layers are added on both image and text embedding vectors, to make the embeddings of different modalities comparable. Finally, we add the MIVSE loss layer on the top to guide the training.

In our model, we need to extract many subregion features $\{\mathbf{x}^c\}_{c\in\mathcal{C}}$ for each image. For efficiency, we follow the Fast RCNN [8] scheme: Given an image $\mathbf{x}$ and the regions of interests (RoI) $\mathcal{C}$, we pass the image through the fully convolutional network once, and all subregions $c \in \mathcal{C}$ are pooled into a fixed-size feature map to obtain $\{\mathbf{x}^c\}$.

4.3.2 Subregion set construction.

Note that we do not have bounding box annotations in training, thus a key problem of MIVSE is how to construct the image subregion set $\mathcal{C}$. Inspired by the recent region-proposal-based methods [8, 9] in object detection, we construct this set using Geodesic Object Proposals [16] followed by post-processing. Semantically meaningful subregions of an image do not necessarily contain objects. We adopt Geodesic Object Proposals since we empirically find that it covers both foreground and background regions well. After constructing the subregion set above, there are around 100 subregions per image on average. But among them only very few correspond to the ground truth labels. Thus, in order to save computational time, we post-process the generated region proposals to discard regions of too small sizes or extreme aspect ratios (in experiments we constraint the subregions' side length to be at least 0.3 of the image and the extreme aspect ratio to be 1:4 or 4:1). Finally, we keep the top 35 refined-subregions sorted by region-proposal score, to construct $\mathcal{C}$. The whole image is included.

4.4 Inference with MIVSE model

Given a trained MIVSE model and a new test image $\mathbf{x}'$, firstly, the subregion set $\mathcal{C}'$ is constructed as above. And we pass $\mathbf{x}'$ and $\mathcal{C}'$ through our MIVSE network in Figure 3 (a) (b) to obtain the subregion embedding vectors, $\{f(\mathbf{x}'^c)\}_{c\in\mathcal{C}}$. Then, for all testing labels $y' \in \mathcal{Y}'$, the distances between $\mathbf{x}'$ and y' are computed by $\min_{c\in\mathcal{C}'}||f(\mathbf{x}'^c) - \mathbf{s}(y')||_2^2$. Thus, for image $\mathbf{x}'$ there is a ranking list of label prediction according to such distances. In addition, given a predicted label y^*, we can locate the corresponding semantically meaningful image subregion c^* via, $c^* = \text{argmin}_{c\in\mathcal{C}}||f(\mathbf{x}'^c) - \mathbf{s}(y^*)||_2^2$.

	Approach	Rec_L	$Prec_L$	Rec_A	$Prec_A$	N_+
1. $k=3$	CNN + Ranking [10]	26.83	31.93	58.00	46.59	95.06
	CNN + WARP [10]	35.60	31.65	60.49	48.59	96.29
	upgraded [10]	37.81	35.23	62.02	50.91	98.77
	upgraded kNN-voting [21]	28.53	32.76	58.92	46.41	95.06
	multi-label VSE baseline	31.59	34.75	60.26	49.17	98.77
	Our MIVSE model	**40.15**	**37.74**	**65.03**	**52.23**	**100.00**
2. $k=5$	CNN + Ranking [10]	42.48	22.74	72.78	35.08	97.53
	CNN + WARP [10]	52.03	22.31	75.00	36.16	**100.00**
	Upgraded [10]	55.27	25.93	78.01	38.04	**100.00**
	multi-label VSE baseline	50.25	26.08	75.62	36.94	98.77
	Our MIVSE model	**59.81**	**28.26**	**80.94**	**39.00**	**100.00**

Table 1: Performance of the proposed MIVSE model in image annotation on NUS-WIDE, shown in %, with $k=3$ and $k=5$ annotated labels per image, respectively.

5 Experiments

In this section, we report experiments in two tasks, multi-label image annotation and zero-shot learning. We use Caffe [13] to implement our model, which is optimized by SGD with momentum of weight 0.9. The ranking loss margin m is set to be 0.3 in our experiments.

5.1 Multi-label image annotation on NUS-WIDE

5.1.1 Dataset.

In the task of multi-label image annotation, Gong *et al.* [10] reported state-of-the-art performance among the methods that do not utilize metadata. We follow [10] to test on one of the largest public multi-label image dataset, NUS-WIDE [2]. This dataset contains 209,347 images from Flickr with 81 ground-truth labels. We follow the train-test split of [10] to use a subset of 150,000 images for training and the rest for testing. It is worth noting that except the data provided in the dataset, we do not use any extra data for training. Thus we do not compare with those methods that utilize rich metadata for image annotation, *e.g.*, [12, 14].

5.1.2 Evaluation metric.

We adopt the 5 metrics used in [10] for fair comparison. Specifically, for each image, we annotate it with the k highest-ranked labels and compare the assigned labels with its ground-truth. Firstly, we compute the recall and precision for each label, and report the per-label recall and per-label precision: $Rec_L = \frac{1}{T}\sum_{i=1}^{T}\frac{N_i^c}{N_i^g}$, $Prec_L = \frac{1}{T}\sum_{i=1}^{T}\frac{N_i^c}{N_i^p}$, where T is the total number of labels, N_i^c is the number of correctly annotated images for label i, N_i^g is the number of ground-truth labeling for label i, and N_i^p is the number of predictions for label i. We also report the overall recall and overall precision: $Rec_A = \frac{\sum_{i=1}^{T}N_i^c}{\sum_{i=1}^{T}N_i^g}$, $Prec_A = \frac{\sum_{i=1}^{T}N_i^c}{\sum_{i=1}^{T}N_i^p}$. The percentage of recalled labels in all labels is evaluated, denoted as N_+. Evaluating these 5 metrics makes the evaluation less biased and more thorough.

5.1.3 Comparing with state-of-the-art methods.

We compare MIVSE with various methods on NUS-WIDE using $k=3$ and $k=5$, respectively. Table 1 shows the results. Note that the results shown here are reported from the

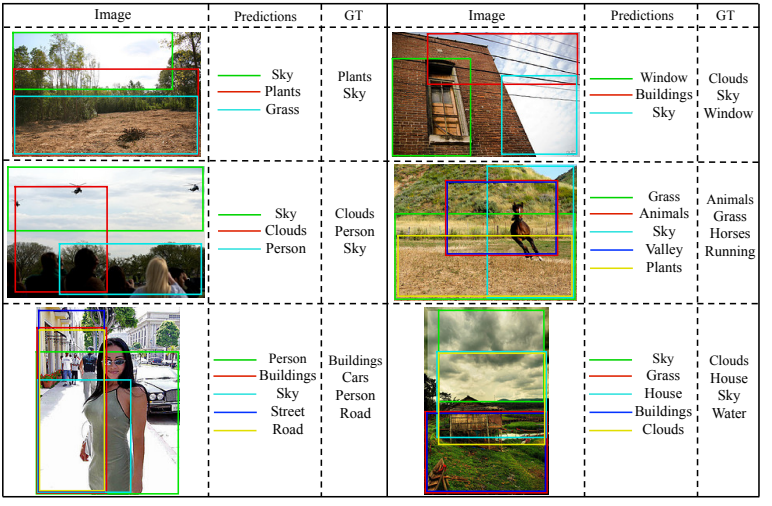

Figure 4: Image annotation results using MIVSE. The predicted labels are listed according to the ranking (we show top-3 predictions if the number of ground truth labels is smaller than or equal to 3, otherwise show top-5 predictions.). The semantically meaningful subregions of predicted labels are shown in bounding boxes of the same color indicated with the labels. The ground truth labels (GT) are listed according to alphabetic order. Better viewed in color.

original papers. The original results reported in [10] were based on AlexNet [17] while our model uses GoogleNet [34]. For fair comparison, we reimplement their model using GoogleNet, named as "upgraded [10]", and "upgraded kNN-voting [21]". Those upgraded methods do not utilize metadata for training. Overall, our model outperforms state-of-the-art results reported in [10] by 4.51% averaged over all metrics for $k = 3$, and 4.50% for $k = 5$. Our model outperforms the upgraded best performer by 2.08% for $k = 3$ and 2.15% for $k = 5$.

5.1.4 Qualitative results.

Our method can discover semantically meaningful image subregion for each label by modeling the subregion-to-label correspondence. Figure 4 shows several sample results on image annotation, as well as the visualization of corresponding subregions for the predicted labels, as indicated by the bounding boxes. As shown in the figure, the predicted labels are associated with subregions of reasonable semantics. For example, *Sky* and *Window* in the first row, *Person* and *Animals* in the second row, *Road* and *Grass* in the third row, *etc.*, are reasonably discovered using our model. There are a few annotation errors or inaccurate bounding boxes. For instance, the right image in the second row is mistakenly annotated with *Plants*. But if we look at the bounding box of *Plants*, the subregion interprets the label concept well. It is important to note that our task is not detection and no bounding box annotation is used in the training. Thus, some objects are not tightly localized by the bounding boxes.

In the following section, we validate the generalizing ability of the proposed MIVSE model on zero-shot learning.

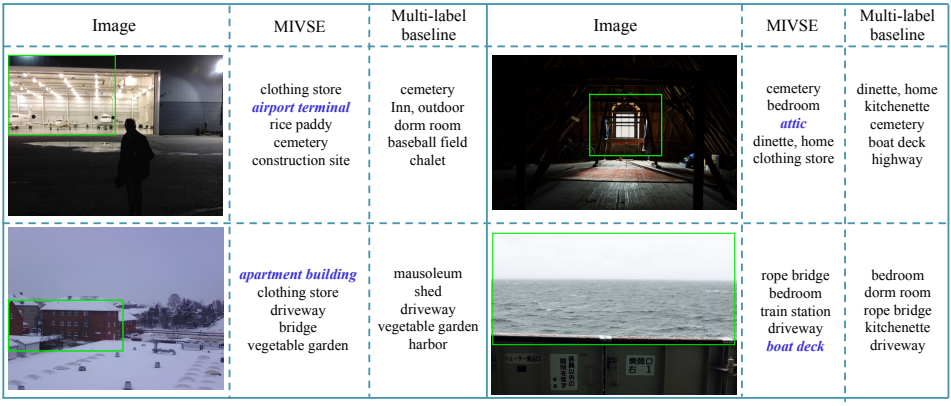

Figure 5: Zero-shot learning results on Places205 dataset using MIVSE model and the multi-label baseline. The correctly predicted labels are shown in blue (note that there is only one ground truth label for each image in Places205). In each image, the semantically meaningful image subregion of MIVSE is shown in green bounding box.

5.2 Zero-shot learning on MIT Places205

5.2.1 Dataset.

As discussed in related work, because of the difference between attributes and labels, we focus on zero-shot learning in the context of label prediction. Unfortunately, most zero-shot learning datasets in the literature are attribute-based, and others are for single-label zero-shot learning. Thus, we have to setup a zero-shot learning dataset for multi-label case. We reconstruct such a dataset based on the NUS-WIDE dataset [7] and the MIT Places205 dataset [37]. The multi-label dataset NUS-WIDE is used for model training, and we test the learned model on the validation set of the single label MIT Places205 dataset. In MIT Places205, there are total 205 classes, and for each class there are 100 validation images. We exclude images from the following 8 classes: *bridge*, *castle*, *harbor*, *mountain*, *ocean*, *sky*, *tower*, and *valley*, since they are included in NUS-WIDE, which results in 197 test classes. It is important to note that in our setup, some test labels may be partially overlapped with training labels like *airport* and *airport terminal*, however, they are considered to be unseen because of the different concepts. Similarly, such setup was widely used by [7, 24].

5.2.2 Evaluation metric and comparing methods.

To quantify the performance, we use mAP@k as the evaluation metric, which measures the mean average precision of annotating the ground truth label within the top-k prediction. There is seldom zero-shot learning approach for multi-label images. Classes-predefined multi-label image annotation methods do not possess the ability of generalizing to unseen labels. Previous visual-semantic embedding models are developed for single-label images only. We compare our model with the multi-label baseline shown in Equation 1, DeViSE [7], and ConSE [24]. DeViSE and ConSE are state-of-the-art for single-label zero-shot learning. In order to adjust them to fit the multi-label case, we duplicate each multi-label image as multiple single-label images, and train them following the original single-label setting.

Approach	mAP@1	mAP@2	mAP@5	mAP@10
DeViSE	1.31	2.25	3.62	5.40
ConSE	1.82	2.77	4.59	6.48
multi-label baseline	6.53	10.18	18.92	28.17
Our model	**7.14**	**11.29**	**20.50**	**30.27**

Table 2: Zero-shot learning results on the MIT Places205 dataset, shown in %.

5.2.3 Quantitative results.

As shown in Table 2, DeViSE and ConSE perform inferior results because they are inherently designed for single-label zero-shot learning. Separately training it with duplicated images will confuse the embedding learning. MIVSE outperforms the multi-label VSE baseline by 1.35%, which further validates the benefit of modeling subregion-to-label correspondence.

5.2.4 Qualitative results.

Figure 5 shows several sample results of zero-shot learning. The two columns on the right of each image show the top-5 label predictions of our MIVSE model as well as the naive multi-label baseline. As shown in Figure 5, our MIVSE model correctly predicts all ground truth labels in top-5, shown in blue. In addition, using our model, the semantically meaningful subregion associated with each predicted labels is located, as shown in green bounding boxes. The localized image subregions interpret the label concepts reasonably well. For instance, in the upper right image, test label *attic* is semantically close to the training label *window*. Thus, based on the concept of *window* already learned in training and a well-located *window*-like subregion, our MIVSE model can utilize the relationship in the semantic space to transfer a learned concept of *window* to assist the prediction of an unseen label *attic*. This suggests that the subregion-to-label correspondence of our method, which helps identify semantically meaningful subregions in image, can benefit the prediction of unseen labels.

6 Conclusion

In this paper, we proposed a novel multiple instance visual-semantic embedding model for multi-label image representation. Instead of embedding a whole image into the semantic space, our model learns an embedding function that characterizes the subregion-to-label correspondence, which discovers and maps semantically meaningful image subregions to the corresponding labels. Experimental results on two challenging tasks, *i.e.*, multi-label image annotation and zero-shot learning, have demonstrated that the proposed method achieves superior performance over state-of-the-art methods on both tasks and possesses the generalizing ability to make correct predictions for unseen labels.

Acknowledgement

This work was partially supported by Army Research Office ARO 62250-CS and a gift grant from Adobe Research. We also acknowledge the support of Nvidia Corporation with the donation of GPUs.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[2] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, 2009.

[3] Jia Deng, W. Dong, R. Socher, Lijia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierachical image database. In *CVPR*, 2009.

[4] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014.

[5] Thomas G. Dietterich, Richard H. Lathrop, and Tomas Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[7] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[8] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[10] Yunchao Gong, Yangqing Jia, Thomas K. Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. In *ICLR*, 2014.

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[12] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *CVPR*, 2016.

[13] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[14] Justin Johnson, Lamberto Ballan, and Li Fei-Fei. Love thy neighbors: Image annotation by exploiting image metadata. In *ICCV*, 2015.

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[16] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *ECCV*, 2014.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[19] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. In *NIPS*, 1990.

[20] Xin Li, Yuhong Guo, and Dale Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *ICCV*, 2015.

[21] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[22] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*, 2012.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[24] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[25] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[26] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[28] Zhou Ren, Chaohui Wang, and Alan Yuille. Scene-domain active part models for object representation. In *ICCV*, 2015.

[29] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Joint image-text representation by gaussian visual-semantic embedding. In *ACM Multimedia*, 2016.

[30] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*, 2017.

[31] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[33] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[35] Jason Weston, Samy Benjio, and Nicolas Usunier. Wsabie: scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[36] Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

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

[38] Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176:2291–2320, 2012.