Labelless Scene Classification with Semantic Matching

Meng Ye
meng.ye@temple.edu
Computer and Information Sciences
Temple University, Philadelphia, USA
Yuhong Guo
yuhong.guo@carleton.ca
School of Computer Science
Carleton University, Ottawa, Canada

Abstract

Using high-level representation of images, e.g., objects and discriminative patches, for scene classification has recently drawn increasing attention. Compared with low-level image features, the high-level features carry rich semantic information that is useful for improving semantic scene classification. Nevertheless, acquiring scene level annotations remains a bottleneck for automatic scene classification, although plenty of related auxiliary resources such as images with object tags are free available on the Internet. In this paper we propose a simple and novel methodology that exploits the rich auxiliary image and text resources to perform *labelless* automatic scene classification without acquiring training images annotated with scene labels. The key of our methodology is to utilize existing object detectors to represent images in terms of high-level objects and then automatically categorize them based on the semantic relatedness of the object names and scene labels. We further incorporate a label propagation step to refine the automatic scene categorization results. Experiments are conducted on three standard scene classification datasets. The results show that our labelless semantic method can achieve reasonable performance and alleviate considerable amount of scene annotation effort by comparing with supervised scene categorization baselines.

1 Introduction

Scene classification has been a challenging problem in computer vision due to its highly flexible structural layout over high-level visual entities. Traditional visual features [3, 15] only capture low level color or texture information in the image and are not discriminative enough to generate good classification results. The recent development of deep convolutional neural networks (CNN) that learn useful high-level features of images with a deep hierarchical structure has led to state-of-the-art results in both general image classification [8] and scene classification [26]. The CNN training however requires a huge amount of labeled data for a large number of categories. Hence much effort on finding good features for scene classification in the literature has been focused on developing intermediate representations such as patch-based representations [14, 23, 25], semantic-based representations [11, 21, 24], and object-based representations [10]. In particular, the work in [21] uses semantic attributes produced by crowd-sourcing to represent images and exploits the confidence of attributes for scene classification. The method in [24] also uses manually specified semantic visual

attributes as mid-level representations. It has been found that semantic attributes are much more informative than low-level features for object description [5]. Objects, being semantic entities at even higher level, are then expected to be informative for scene descriptions. The *Object Bank* method developed in [10] exploits auxiliary resources, pre-trained object detectors, to produce object-based image representations for scene classification, which demonstrates good performance. These existing works however all require a sufficient amount of labelled data with scene category annotations for training and hence induce significant annotation cost for scene classification.

Recently, the idea of semantic attributes has also been exploited to reduce annotation effort for image classification through zero shot learning [1, 9]. These zero shot learning works use shared semantic attributes to represent the labels of class categories, based on which, intermediate attribute predictors trained on the labeled data can be used to perform classification in new classes that are unseen in the training stage. Although many zero-shot learning works require human defined semantic attributes, recent zero shot learning methods have proposed to explore auxiliary natural language processing (NLP) resources such as WordNet [6, 13] and semantic word embeddings [7, 12, 13, 19] to build inter-class connections for cross-class information adaptation. Nevertheless, zero-shot learning still requires sufficient amount of labeled data in a set of seen classes that are at the same level as the unseen classes.

Motivated by these developments in the literature, in this paper we extend previous effort in reducing annotation effort into a broader vision by exploiting annotation resources at different levels of the semantic output space. We propose a novel labelless learning method for scene classification, which exploits auxiliary image resources such as pre-trained object detectors and textual resources such as semantic word embeddings to build semantic connections between images and scene categories and automatically classify scene images. Different from previous works, the proposed approach does not require annotated data from any scene classes and it can easily handle scene category expansions. Specifically, it first uses auxiliary pre-trained object detectors to produce object-based high-level representations for the images. Then it maps both images and scene labels into the same semantic space by expressing the object names and scene category labels using a common set of word embedding vectors pre-trained from auxiliary large text corpora. Finally, automatic image classification can be conducted by assigning each image into the scene class whose label has the largest matching score with the image in the semantic vector space. We further exploit label propagation to refine the automatic classification results. To our best knowledge, this is the first work that pursues labelless scene categorization. We conducted experiments on three standard scene classification datasets. The experimental results show that the proposed approach can alleviate considerable amount of scene annotation effort from supervised learning.

2 Related Work

This section provides a brief review on scene classification works that exploit intermediate representations and image classification works that exploit semantic word embeddings.

2.1 Scene Classification with Intermediate Representations

Due to its highly flexible structural layout, scene images are difficult to classify based on low-level visual features. Much work in the literature has proposed to address scene classifi-

cation by learning intermediate or high-level representations such as semantic attributes [11, 21, 24], discriminative patches [14, 23] and objects [9, 10, 25]. The work in [24] uses manually specified semantic visual attributes as a mid-level representation. It trains a classifier for each individual attribute and then uses the outputs of these classifiers as image descriptors for classification. The work in [21] uses semantic attributes produced by crowd-sourcing to represent images and exploits the confidence of the attributes for scene classification. The work in [11] proposes a hierarchical model to automatically learn latent semantic representations of the images. In [23], the authors use a set of discriminative patches discovered in an unsupervised way as mid-level visual representations, while the work in [14] proposes to increase the discriminative power of image patches. The *Object Bank* method in [10] uses pre-trained generic object detectors to produce object-based image representations for scene classification. The work in [9] also uses objects as intermediate semantic representations. It relates objects to scenes by prior contextual information computed from the frequencies of objects in each scene in the labeled training data. More recently, some researchers make use of the generic CNN features to harvest discriminative visual objects and parts, called Meta Objects, for scene classification [25]. While these previous works successfully construct intermediate representations of scene images, they still need to have a sufficient amount of labeled data with scene annotations to train their classification models. By contrast, our proposed work builds connections between the intermediate representations and the target scene labels by using auxiliary textual resources such as semantic word embeddings, and it does not require any labeled data with scene annotations.

2.2 Semantic Word Embeddings

Learning semantic word embeddings for linguistic words and phrases from large text corpus has been a recent advance in Natural Language Processing (NLP). Notable models for this advance include the Skip-gram model [18] and the Continuous Bag of Words (CBoW) model [16]. Both models use neural networks to learn real valued vector representations for the linguistic words (or phrases). But CBoW performs learning by predicting a word given its context while Skip-gram conducts learning by predicting the context given a word. The word embedding vectors induced by these models can successfully capture the underlying semantic meanings of the words from the contextual information of the text corpus.

Many previous works have exploited semantic word embeddings to build inter-class semantic connections and address image classification in the context of zero shot learning [2, 7, 12, 13, 19]. The work in [2] evaluates both human specified supervised semantic attributes and unsupervised word embedding vectors for fine-grained classification, and it shows the unsupervised semantic embeddings achieve compelling or even superior results than the supervised attributes. In [7], the authors proposed a deep visual-semantic model to incorporate both word embeddings and deep image features to improve image classification and zero-shot prediction performance. In [12], the authors used word embeddings to build inter-class relationship matrix to perform max-margin zero-shot learning. In [13], the authors proposed a hierarchical semantic embedding model that exploits the WordNet hierarchy to improve label embedding and image embedding for zero-shot image tagging. The work in [19] proposes to train classifiers on labeled classes, and then maps a test image into the semantic word embedding space by taking a weighted convex combination of the seen classes' embedding vectors with the classifier outputs. It assigns the test image to the novel class that has the largest similarity value with the image in the embedding space. Different from all these works, our proposed work in this paper exploits word embeddings to build semantic

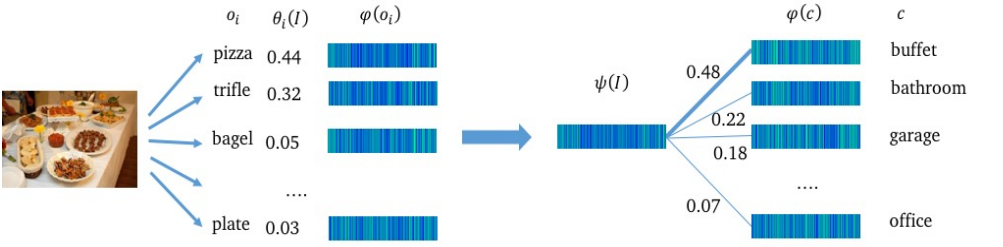

Figure 1: An illustration example of the semantic matching process for the proposed method. First, a set of objects $\{o_i\}$ are detected from image I, each with a corresponding normalized detection probability $\theta_i(I)$. Next, the semantic embedding vector $\varphi(o_i)$ for each object o_i and the semantic embedding vector $\varphi(c)$ for each scene category c are computed from the auxiliary NLP word embeddings. The semantic embedding vector $\psi(I)$ for image I is then computed as a probability weighted sum of the object semantic embedding vectors. Finally, the similarity-based matching scores between $\psi(I)$ and each $\varphi(c)$ are computed and the image can be assigned into the scene category ('buffet' in this example) that has the largest matching score (0.48 in this example).

connections between the high-level image descriptors, objects, and the scene class labels. We do not transfer information from any labeled scene classes to novel classes, but address the overall scene classification problem in an unsupervised manner at the scene level.

3 Proposed Approach

In this section we present a novel labelless scene classification method that exploits auxiliary image and textual resources to automatically build semantic connections between the images and the scene categories without acquiring images annotated with *scene labels*.

3.1 Labelless Semantic Scene Categorization

Given a set of D images $\{\mathcal{I}_1, \cdots, \mathcal{I}_D\}$ and a set of N scene category labels $\{c_1, \cdots, c_N\}$, we aim to automatically classify each image $\mathcal{I}_i$ into one of the N scene categories. The main idea is to first represent the images in terms of high-level visual concepts, objects, and then exploit the semantic relatedness between the abstractive scene concepts and the objects that visually express scene concepts to automatically categorize the images into scene classes. The process of our methodology is illustrated in Figure 1. Below we will present it in detail.

3.1.1 High-level Object-based Image Representation

Natural scenes are abstractive high-level semantic concepts. Most scenes are expressed as a collections of high-level visual entities, such as objects, in variable layouts. It has been shown in the literature that high-level semantic representations such as objects are very useful for scene classification [10]. Meanwhile there are rich image resources with object labels such as ImageNet available on the Internet to be used for training generic object detectors [20, 22].

We hence propose to exploit the generic object detectors pre-trained on the auxiliary image resources to produce high-level object-based representations for the target unlabeled images.

Let $\theta(\cdot)$ denote the object detection function produced by the object detectors on a set of m objects with object labels $\{o_1, o_2, ..., o_m | o_k \in \mathcal{W}, k = 1, 2, ..., m\}$. Given an image $\mathcal{I}$, the output of the object detection will be a probabilistic vector over the m objects such that $\theta(\mathcal{I}) = [\theta_1(\mathcal{I}), \ldots, \theta_i(\mathcal{I}), \cdots, \theta_m(\mathcal{I})]$, where each value $\theta_i(\mathcal{I})$ indicates how likely the image $\mathcal{I}$ contains the object o_i. This vector $\theta(\mathcal{I})$ hence forms the object-based high-level representation of image $\mathcal{I}$. To reduce the impact of noisy object detections, we choose to only consider the top T detected objects for each image (we used T=10 in experiments), while setting the remaining smaller detection probabilities, $\{\theta_i(\mathcal{I})\}$, as zeros. We further normalize the vector $\theta(\mathcal{I})$ to sum to 1, in order to represent each image at the same quantity level. For example, in Figure 1, the $\theta(\mathcal{I})$ function returns a vector of probability values [*0.44, 0.32, 0.05, ···, 0.03*] over a set of objects {*'pizza', 'trifle', 'bagle', ···, 'plate'*}.

3.1.2 Semantic Embeddings of High-level Visual Concepts

Computer vision tasks have natural connections with natural language processing (NLP) since each high-level visual concept, such as an object name or a scene label, is described using the key linguistic elements of NLP, words or phrases. Recent advances in NLP techniques have allowed the semantic meanings of linguistic words and phrases to be learned in the form of distributed embedding vectors from the contextual information of large text corpora without human supervision [18].

The availability of the semantic word embedding vectors from NLP field provides a natural way of expressing the high-level visual concepts, such as object names and scene category labels, in the *same* semantic embedding space. Let $\phi(\cdot)$ denote the word/phrase embedding function produced by NLP techniques, which maps a word/phrase into a d-dimensional embedding vector space: $\phi : \mathcal{W}_p \mapsto \mathbb{R}^d$. However, the domain $\mathcal{W}_p$ usually contains all single words but only a subset of phrases which may not cover all the phrases in our object names or scene labels. Hence below we define a general word/phrase embedding function $\varphi : \mathcal{W} \mapsto \mathbb{R}^d$ based on $\phi(\cdot)$, which maps any input word/phrase, e.g., an object name o_i, into a d-dimensional embedding vector space:

$$\varphi(o_i) = \begin{cases} \phi(o_i) & \text{if} \quad o_i \in \mathcal{W}_p \\ \sum_{w \in o_i} \phi(w) & \text{otherwise} \end{cases} \tag{1}$$

where w denotes a single word. For example, for an object name *'dining table'* $\notin \mathcal{W}_p$, we will have $\varphi(\text{'dining table'}) = \phi(\text{'dining'}) + \phi(\text{'table'})$; in Figure 1, the embedding vectors of the object names, {*'pizza', 'trifle', ···, 'plate'*}, are demonstrated in the column under "$\varphi(o_i)$".

Sometimes we have more than one phrases to describe an object; e.g., *'trash can'*, *'garbage can'* and *'dustbin'* all refer to the same object. In this case, we propose to use the average of embedding vectors of the phrases to represent the object:

$$\varphi(o_i) = \frac{1}{K_i} \sum_{k=1}^{K_i} \varphi(o_i^{(k)}) \tag{2}$$

where $\{o_i^{(k)}, k = 1, 2, ..., K_i\}$ refer to all the K_i phrases that describe the i-th object.

Similarly we can compute the embedding vectors of the scene category labels, $\{c_1, ..., c_N\}$, in the same semantic embedding space using the same embedding function $\varphi(\cdot)$ defined above. In Figure 1, the embedding vectors of the scene labels, {*'buffet', 'bathroom', 'garage', ···, 'office'*}, are demonstrated in the column under "$\varphi(c)$".

3.1.3 Scene Classification with Semantic Matching

The visual relationships of high-level visual concepts are typically consistent with their semantic relationships since images and text descriptions can be two parallel ways of recording the same observations of life and nature. For example, in an image of a *'bedroom'* scene, it is most likely to see objects such as *'bed', 'lamp'* and *'shade curtain'*, while in an article talking about bedroom it is also very likely to see these object phrases. In addition, the word/phrase embedding vectors produced by NLP tools can carry semantic meanings induced from natural text resources. They often demonstrate interesting properties in expressing various semantic relationships. For example, it has been shown that $\varphi(\text{'Paris'})$- $\varphi(\text{'France'})$ + $\varphi(\text{'Italy'})$ result in a similar embedding vector to $\varphi(\text{'Rome'})$ [17]. Given the consistent visual and semantic relationships and the availability of semantic word/phrase embedding vectors, the high-level object-based representation of images creates a great opportunity for automatically building connections between the images and scene labels in the semantic embedding vector space. We are hence enlightened to develop a labelless image classification approach that performs scene classification by conducting automatic semantic matching between the images and the scene labels in the semantic embedding space.

We view the semantic representation of an image as the combination of the semantic embedding vectors for all the objects it contains. Since our object detection function outputs probabilities for the appearance of the considered objects in a given image, we compute the semantic embedding vector of an image $\mathcal{I}$ by taking a weighted sum of the embedding vectors for all the objects:

$$\psi(\mathcal{I}) = \sum_{i=1}^{m} \theta_i(\mathcal{I})\varphi(o_i) \tag{3}$$

This $\psi(\cdot)$ function maps an image into the same semantic embedding space as the object and scene category labels. We can then compute the matching score between an image $\mathcal{I}$ and a scene category c as the cosine similarity score between their semantic embedding vectors:

$$s(\mathcal{I}, c) = \frac{\psi(\mathcal{I})^\top \varphi(c)}{\|\psi(\mathcal{I})\| \, \|\varphi(c)\|} \tag{4}$$

where $\|\cdot\|$ denotes the Euclidean norm. Note the usage of cosine similarity function can eliminate normalization issues over the image and scene category representation vectors. The matching scores of the i-th image $\mathcal{I}_i$ with all the scene categories then form a row vector:

$$Y(i,:) = [s(\mathcal{I}_i, c_1), s(\mathcal{I}_i, c_2), \cdots, s(\mathcal{I}_i, c_N)] \tag{5}$$

According to our assumption that visual and semantic relationships are consistent, we expect each image will have the largest matching score with its underlying scene category. Hence we can automatically classify the image $\mathcal{I}_i$ by assigning it into the scene category c_{y_i} that has the highest matching score among all the N scene categories; that is, $y_i = \arg\max_j Y(i,j)$. For the example demonstrated in Figure 1, the image is successfully classified into the scene category with label '*buffet*' that has the highest matching score 0.48.

3.2 Classification Refinement with Label Propagation

The labelless scene classification method proposed above can output unreliable matching scores for images on which the object detection function has poor detection results. We hence

propose an additional label propagation step on a k-NN graph built over all the D images based on the extracted CNN features to refine the semantic matching results Y. Moreover, since we only want to propagate the most confident predictions through the graph, we further prepare Y for propagation by only keeping the top-δ fraction of scores in each class (the columns of Y) and setting other values to zeros. The label propagation is expected to exploit the intrinsic manifold structure of the images and propagate confident predictions to improve the ultimate scene classification performance.

To build a k-NN graph, we first compute the squared Euclidean distance between each pair of images, such that $d(\mathbf{x}_i,\mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|^2$, where $\mathbf{x}_i$ (or $\mathbf{x}_j$) denotes the CNN feature vector of image $\mathcal{I}_i$ (or $\mathcal{I}_j$). Then we construct the k-NN graph by computing the RBF kernel based affinity matrix W in the following way:

$$W_{ij} = \begin{cases} \exp\left(\frac{-d(\mathbf{x}_i,\mathbf{x}_j)}{2\sigma^2}\right), & \text{if } i \in \text{KNN}(j) \text{ or } j \in \text{KNN}(i) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where KNN(i) denotes the k-nearest neighbors of the i-th image. After computing the normalized Laplacian matrix L from W such as $L = Q^{-1/2}WQ^{-1/2}$ where Q is a diagonal matrix with $Q_{ii} = \sum_j W_{ij}$, we can take the standard label propagation technique in [27] to perform label propagation, which provides the following refined prediction score matrix:

$$Y^* = (I - \alpha L)^{-1} \times Y \tag{7}$$

where I is an identity matrix of size D and $\alpha \in [0,1]$ is a regularization trade-off parameter. Same as before, in each row of Y^*, the class with the largest score is selected to be the predicted class.

4 Experiments

We conducted experiments on three standard scene classification datasets, *MIT-Indoor*, *15-Scene* and *SUN* datasets. The MIT-Indoor database contains 67 indoor scene categories and the number of images varies across categories, with at least 100 images per category. We used a subset of 30 classes with some common scene labels such as 'game room', 'kitchen', 'bathroom', 'office', etc.. The 15-Scene dataset contains images in 15 natural scene categories, including 'bedroom', 'highway', 'kitchen' and 'mountain'. Each of its classes contains more than 200 grayscale images. The SUN dataset contains a large number of categories and images. We chose a subset of 50 categories to use by dropping categories that are semantically close; for example, 'bookstore' and 'library' are semantically close to each other and we picked only one of them. Each of these 50 categories contains at least 100 images and in total we have 15,812 images.

4.1 Free Auxiliary Resources

Two types of free auxiliary resources exploited in this work are object detectors and NLP tools that produce word embedding vectors. We used an existing state-of-the-art object prediction/detection technique, *OverFeat* [27], to produce object detectors. The *OverFeat* tool is pre-trained on ImageNet with 1000 classes. We dropped the classes that can potentially overlap with our scene categories and used the remaining classes that focused on objects.

Table 1: Performance of the proposed approach in terms of the three evaluation metrics.

		mcAccu	avgAccu	mAUC
15-Scene	LSM	34.18	91.22	78.30
	LSM+LP	53.36	93.78	89.87
MIT-Indoor	LSM	34.62	95.64	77.58
	LSM+LP	42.05	96.14	89.28
SUN	LSM	30.41	97.22	77.85
	LSM+LP	34.55	97.38	84.69

For producing the semantic embeddings used in our approach, we used one most popular NLP model for learning word embeddings, the Skip-gram model [18]. It learns word representation vectors by predicting the context words (or phrases) in a sentence or a document. The model is trained on a large document collection, Google News dataset, and its published results contain 300-dim embedding vectors for 3 million words and phrases.

4.2 Labelless Scene Classification Results

We first investigated the classification performance of the proposed methodology. We compared two variants of the proposed methodology, *LSM* and *LSM+LP*, where *LSM+LP* denotes the proposed full approach with label propagation refinement and *LSM* denotes the variant with only the semantic matching procedure. We conducted the comparison to investigate the effectiveness of the simple semantic matching procedure and the label propagation refinement procedure separately. For the label propagation step, we used $k = 30$ for the k-NN graph construction, and set the radius of the RBF kernel, σ^2, as the average of the squared distances of the edges in the k-NN graph. We used $\delta = 0.2$ to keep the top 20% of the confident matching scores in each class as the initial matrix for propagation, and used $\alpha = 0.5$ to give the initial matrix sufficient weights while allowing modifications from the propagation.

Our methodology is entirely unsupervised at the scene label level. Both *LSM* and *LSM+LP* automatically predict the scene labels for all the images in each dataset. We used three evaluation metrics, *multi-class accuracy* (mcAccu), *average of per-class accuracy* (avgAccu), and *mean of per-class AUC (Area Under ROC-curve)* (mAUC), to evaluate the classification performance. The results are reported in Table 1. We can see the results are reasonably good. The average per-class accuracy of both variants are over 95% on *MIT-Indoor* and *SUN* and over 90% on *15-Scene*. *LSM* achieved above 77% of mAUC on all three datasets and the full approach *LSM+LP* further increases the performance to over 89% on *15-Scene* and *MIT-Indoor* and to over 84% on SUN. In terms of multi-class accuracy, *LSM* achieved a performance of over 30%, while *LSM+LP* boosted the performance to 53.36%, 42.05% and 34.55% on *15-Scene, MIT-Indoor* and *SUN* respectively. Considering there are 15, 30 and 50 classes in these three datasets respectively, these multi-class accuracy values are reasonably good; without scene annotations, the expected naive random guess results in terms of multi-class accuracy will be around 6.7% on *15-Scene*, 3.3% on *MIT-Indoor* and 2.0% on *SUN*. By comparing the results of *LSM* and *LSM+LP*, we can see that the label propagation refinement step is very helpful, and it induces notable large performance increases in terms of multi-class accuracy and mAUC on the three datasets.

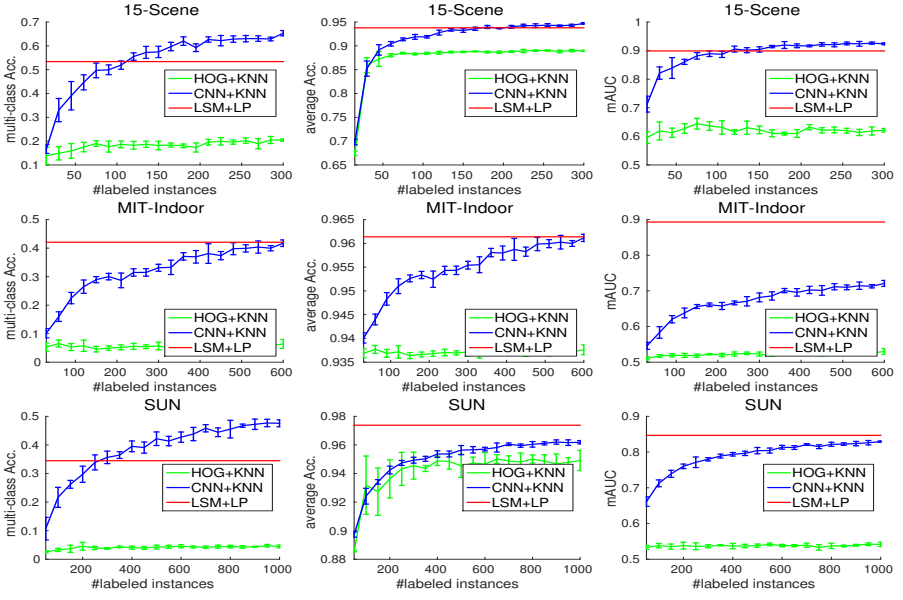

Figure 2: Classification performance v.s. annotation effort in terms of the number of labeled training instances. The proposed LSM+LP does not use the scene annotation information.

4.3 Alleviation of the Annotation Effort

We have also conducted experiments to compare our proposed full method *LSM+LP* with two supervised baselines, *HOG+KNN* and *CNN+KNN*. Both *HOG+KNN* and *CNN+KNN* use the K-Nearest Neighbor method to perform supervised scene classification, while *HOG+KNN* uses the HOG features [3] to represent each image and *CNN+KNN* uses the advanced CNN features [22]. We used $K = 5$ in the experiments. Since the proposed methodology does not require any labeled data with scene annotations, it is not fair to compare its classification performance with the supervised classifiers which have to be trained on labeled images. Instead we are interested to find out how many labeled training images are required to increase the performance of the supervised baselines to the level of our proposed labelless method, which can be viewed as the amount of annotation effort alleviated by our approach.

On each dataset, we randomly split the data into 80% training/20% test and then ran the supervised baselines with different numbers of labeled images from the training set. We repeated the process five times and collected five sets of results. The *unsupervised* method, *LSM+LP*, is also evaluated on the same set of test images. We tested the following sets of labeled training sizes, $[15, 30, 45, \cdots, 300]$, $[30, 60, 90, \cdots, 600]$, and $[50, 100, 150, \cdots, 1000]$, on *15-Scene, MIT-Indoor* and SUN datasets respectively. The average results are reported in Figure 2. We can see our unsupervised approach consistently outperforms the supervised *HOG+KNN* method, while *CNN+KNN* takes considerable number of labeled instances to reach the performance level of our unsupervised method. In terms of multi-class accuracy, *CNN+KNN* uses about 120, 600 and 270 labeled training images on the three datasets respectively to reach the performance level of *LSM+LP*. In terms of average per-class accuracy, *CNN+KNN* uses 165 and 600 labeled training images on 15-Scene and MIT-Indoor respectively to reach the performance level of *LSM+LP*, and fails to reach the level of *LSM+LP* on SUN with 1000 labeled images. In terms of mAUC, *CNN+KNN* uses 150 labeled training

images on 15-Scene to reach the level of *LSM+LP*, while failing to beat *LSM+LP* with 600 and 1000 labeled images on MIT-Indoor and SUN. These results show the proposed labelless methodology can alleviate considerable amount of annotation effort at the scene level.

5 Conclusion

In this paper we have developed a novel labelless method for scene classification, which does not require labeled data from any scene classes. The proposed approach uses auxiliary object detectors to produce object-based high-level image representations. Then it exploits auxiliary word embeddings to map both images and scene labels into embedding vectors in the same semantic embedding space, based on the object names detected from the images and the scene label phrases. Automatic scene classification is conducted by semantic matching, which assigns each image into the scene class whose label embedding vector has the largest matching score with the embedding vector of the image. We further exploited label propagation to refine the automatic scene classification results. We conducted experiments on three scene classification datasets. The experimental results show that the proposed method can achieve reasonable performance and alleviate considerable scene annotation effort.

Acknowledegements

This research was supported by NSF IIS-1422127 and the Canada Research Chairs program.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

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

[4] P. Espinace, T. Kollar, A. Soto, and N. Roy. Indoor scene recognition through object detection. In *ICRA*, 2010.

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

[6] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *ECCV*, 2010.

[7] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *NIPS*, 2013.

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

[9] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2014.

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

[11] X. Li and Y. Guo. Latent semantic representation learning for scene classification. In *ICML*, 2014.

[12] X. Li and Y. Guo. Max-margin zero-shot learning for multi-class classification. In *AISTATS*, 2015.

[13] X. Li, S. Liao, W. Lan, X. Du, and G. Yang. Zero-shot image tagging by hierarchical semantic embedding. In *ACM SIGIR*, 2015.

[14] D. Lin, C. Lu, R. Liao, and J. Jia. Learning important spatial pooling regions for scene classification. In *CVPR*, 2014.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[19] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[20] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C. Loy, and X. Tang. DeepID-Net: Deformable deep convolutional neural networks for object detection. In *CVPR*, 2015.

[21] G. Patterson and J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[22] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

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

[24] Y. Su and F. Jurie. Improving image classification using semantic attributes. *IJCV*, 100 (1):59–77, 2012.

[25] R. Wu, B. Wang, W. Wang, and Y. Yu. Harvesting discriminative meta objects with deep CNN features for scene classification. In *ICCV*, 2015.

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

[27] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2004.