Labelless Scene Classification

Meng Ye
Computer and Information Sciences
Temple University
Philadelphia, USA
meng.ye@temple.edu

Yuhong Guo
School of Computer Science
Carleton University
Ottawa, Canada
yuhong.guo@carleton.ca

Abstract

In this paper we propose a simple and novel methodology that exploits the rich auxiliary image and text resources to perform *labelless* automatic scene categorization 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 label phrases induced from textual resources. 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. Much effort on finding good features for scene classification in the literature has been focused on developing intermediate representations such as patch-based representations [21, 13, 23], semantic-based representations [22, 19, 10], and object-based representations [9]. 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 [9] 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, 8]. Recent zero-shot learning methods have proposed to explore auxiliary natural language processing (NLP) resources such as WordNet [6, 12] and semantic word embeddings [7, 11, 12, 17] 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 label 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 conveniently handle scene category expansion. We conducted experiments on three scene datasets, and the results show the proposed approach can alleviate considerable amount of scene annotation effort from supervised learning.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Related Work

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 classification by learning intermediate representations. The works in [19, 22] use semantic visual attributes, while the authors of [21, 13] use a set of discriminative patches as mid-level visual representations. The *Object Bank* method in [9] uses pre-trained generic object detectors to produce object-based image representations for scene classification. The work in [4] also uses objects as intermediate semantic representations. More recently, some researchers make use of the generic CNN features to harvest discriminative visual objects and parts, called Meta Objects, for scene classification [23]. Nevertheless, these previous works still need 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, and it does not require any labeled data with scene annotations.

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 [16] and the Continuous Bag of Words (CBoW) model [14]. 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, 11, 12, 17]. Different from all these works, our proposed work exploits word embeddings to build semantic 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 scene 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 first represent the images in terms of high-level visual concepts, objects, and then exploit the semantic relatedness between the scene concepts and objects 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.

High-level Object-based Image Representation Natural scenes are abstractive high-level semantic concepts, and can be expressed as a collections of high-level visual objects in variable layouts. 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, 18]. 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 $\boldsymbol{\theta}(\cdot)$ 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 $\boldsymbol{\theta}(\mathcal{I}) = [\theta_1(\mathcal{I}), \ldots, \theta_i(\mathcal{I}), \cdots, \theta_m(\mathcal{I})]$, where each value $\theta_i(\mathcal{I})$ indicates how likely the image $\mathcal{I}$ contains the object o_i. This vector $\boldsymbol{\theta}(\mathcal{I})$ hence forms the object-based high-level representation of image $\mathcal{I}$. To reduce the impact of noisy object detections, we only consider the top T detected objects for each image (T=10 in experiments), while setting the remaining $\{\theta_i(\mathcal{I})\}$ to zeros. We further normalize the vector $\boldsymbol{\theta}(\mathcal{I})$ to sum to 1, to represent each image at the same quantity level.

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. The availability of the semantic word embedding vectors from NLP field provides a natural way of

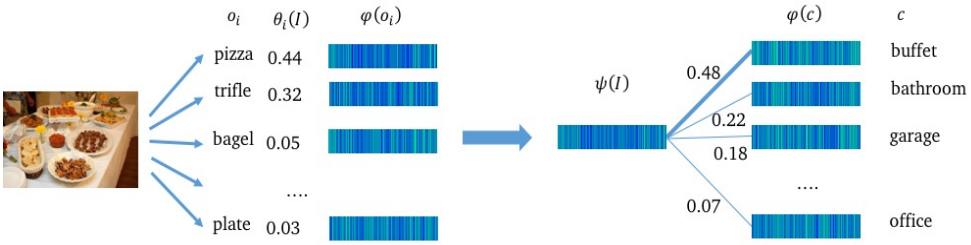

Figure 1: An illustration example of the semantic matching process for the proposed method. First, a set of objects $\{o_i\}$ are detected from image I, each with a corresponding normalized detection probability $\theta_i(I)$. Next, the semantic embedding vector $\varphi(o_i)$ for each object o_i and 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).

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: $\phi : \mathcal{W}_p \mapsto \mathbb{R}^d$. 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 $\varphi(\text{'dining table'}) = \phi(\text{'dining'}) + \phi(\text{'table'})$.

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.

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. We hence compute the semantic embedding vector of an image $\mathcal{I}$ by taking a weighted sum of the embedding vectors for all the objects it contains: $\psi(\mathcal{I}) = \sum_{i=1}^{m} \theta_i(\mathcal{I})\varphi(o_i)$. 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 $s(\mathcal{I}, c)$ between an image $\mathcal{I}$ and a scene category c as the cosine similarity score between their semantic embedding vectors, $\psi(\mathcal{I})$ and $\varphi(c)$. The matching scores of the i-th image $\mathcal{I}_i$ with all the scene categories then form a row vector: $Y(i,:) = [s(\mathcal{I}_i, c_1), s(\mathcal{I}_i, c_2), \cdots, s(\mathcal{I}_i, c_N)]$.

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 with the image among all the N scene categories.

3.2 Label Propagation Refinement

The labelless scene classification method proposed above can output unreliable matching scores when the object detection function has poor detection results on the images. We hence propose an additional label propagation step to refine the semantic matching results Y. Moreover, in order to propagate the most confident predictions through the graph, we clean Y by only keeping the top-δ fraction of scores in each class (the columns of Y) and setting other values to zeros. We first build a k-NN graph over all the D images based on the squared Euclidean distances between each pair of images. Then we construct the k-NN graph by computing the RBF kernel based affinity matrix W such that $W_{ij} = \exp(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$ if j-th image is within the k-nearest neighbors of the i-th image, or vice

Table 1: Performance of the proposed labelless 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

versa. 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 [24] to perform label propagation, which provides the following refined prediction score matrix:

$$Y^* = (I - \alpha L)^{-1} \times Y \tag{2}$$

where I is an identity matrix of size D and $\alpha \in [0, 1]$ is a regularization trade-off parameter. 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.

4 Experiments

We conducted experiments on three scene classification datasets, *MIT-Indoor*, *15-Scene* and *SUN* datasets, over 30, 15 and 50 scene classes respectively. 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 technique, *OverFeat* [20], to produce object detectors, and used the Skip-gram model [16] for learning semantic word embeddings on Google News.

4.1 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. For the label propagation step, we used $k = 30$, $\delta = 0.2$ and$\alpha = 0.5$. Our methodology is entirely unsupervised at the scene label level. We used three evaluation metrics, *multi-class accuracy* (mcAccu), *average of per-class accuracy* (avgAccu), and *mean of per-class AUC* (mAUC), to evaluate the classification performance. The results are reported in Table 1. We can see the results are reasonably good. Considering there are 15, 30 and 50 classes in these three datasets respectively, 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.

4.2 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 ($K = 5$) to perform supervised scene classification on HOG features [3] CNN features [20] respectively. 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. 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. These results show the proposed labelless methodology can alleviate considerable amount of annotation effort at the scene level.

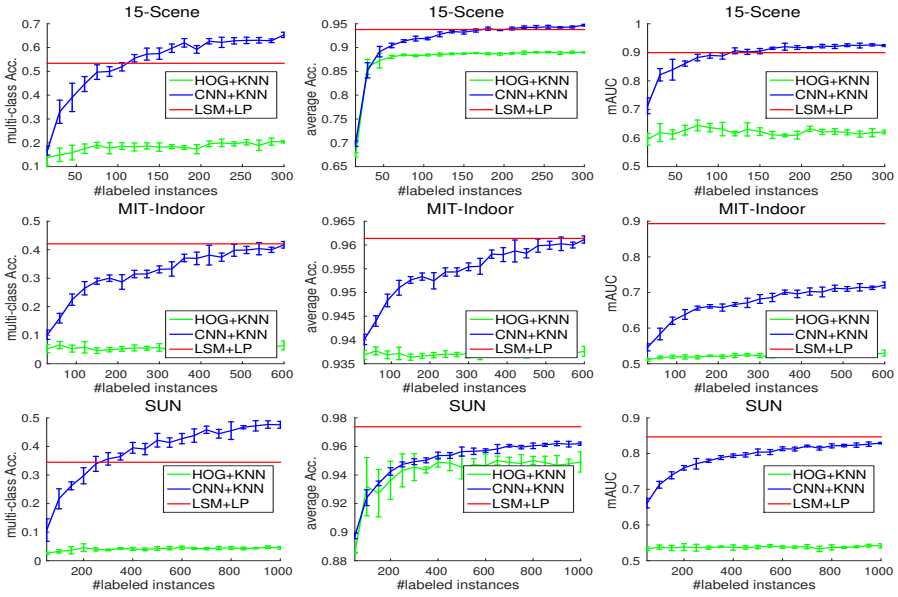

Figure 2: Classification performance v.s. annotation effort in terms of the number of labeled training instances. The proposed LSM+LP does not use the scene annotation information.

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 build connections between images and scene labels based on the object names detected from the images and the scene label phrases. Automatic scene classification is conducted by semantic matching, and further improved using label propagation. We conducted experiments on three scene classification datasets. The results show that the proposed method can achieve reasonable performance and alleviate considerable scene annotation effort.

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

[9] 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.

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

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

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

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

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

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

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

[17] 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.

[18] 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.

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

[20] 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.

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

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

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

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