RGB-D Scene Recognition based on Object-Scene Relation(Student Abstract)

Yuhui Guo, Xun Liang

Renmin University of China, No.59 Zhongguancun Road, Beijing, China 100872
{yhguo,xliang}@ruc.edu.cn

Abstract

We develop a RGB-D scene recognition model based on object-scene relation(RSBR). First learning a Semantic Network in the semantic domain that classifies the label of a scene on the basis of the labels of all object types. Then, we design an Appearance Network in the appearance domain that recognizes the scene according to local captions. We enforce the Semantic Network to guide the Appearance Network in the learning procedure. Based on the proposed RSBR model, we obtain the state-of-the-art results of RGB-D scene recognition on SUN RGB-D and NYUD2 datasets.

Introduction

The goal of scene recognition is to annotate images with scene categories. Humans have an innate talent to recognize those abstract scenes without hard training, whereas the same task is typically challenging for computers. For humans, some scenes (e.g., rivers, mountains, and forest) can be directly distinguished simply at a glance, while some other scenes (e.g., bed room, living room, dining room, and classroom) may require looking from some local points of view such as objects and their relations. For computers, previous scenes may be recognized by directly training the deep learning models with massive data (e.g., Places (Zhou et al. 2018)) under the supervision of global scene labels. By contrast, recognizing latter scenes generally requires a complete understanding of image contents (from the local point of view). The main components of scenes are objects, such as table, vase, tree, bed, and so on (Herranz, Jiang, and Li 2016). Extracting objects based on representations is an intuitive way for scene recognition. However, the diversity of spatial layouts (of objects) and the object co-occurrences (between scenes) may lead to the intra-class difference and inter-class similarity of scenes, which reduce the accuracy of scene recognition. In order to solve the problem, we propose a RGB-D scene recognition model based on object-scene relation.

Methodology

This paper proposes to detect object and their relations on RGB-D data for scene recognition. Particularly, two self-attention modules are introduced to guide object-to-scene
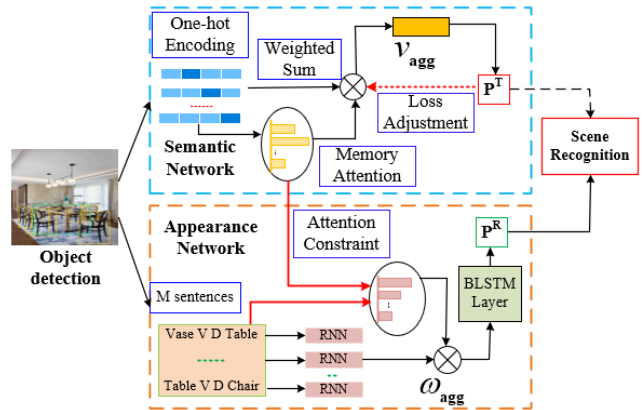
Figure 1: Our proposed RSBR model contains two subnetworks. We first learn a Semantic Network in the semantic domain that classifies the label of a scene on the basis of the labels of all object types. Then, we design an Appearance Network in the appearance domain that recognizes the scene according to local captions. We enforce the Semantic Network to guide the Appearance Network in the learning procedure, i.e., attention constraint, loss adjustment.

relation on the basis of representations, and a RGB-D scene recognition model based on object-scene relation (RSBR) is proposed for scene recognition. Figure 1 illustrates the framework. First, we learn a high-performance model with typical attention mechanism (i.e., Semantic Network) to map the importance of object presence to the object-to-scene relation in the scene. As the inputs of our Semantic Network are generated from the detected single-object labels, object labels are embedded into a latent space and weighted by attention mechanism. These weighted results are added to obtain a scene representation and guide the attention weight of another model (i.e., Appearance Network). Second, we develop an Appearance Network, which predicts the scene label from the appearance information between objects. Richer appearance information is used to describe images, such as their relations between objects and scene. Object-to-scene relations are sequentially represented in our proposed Appearance Network, denoted as SOSR. SOSRs are generated by a template which is typically de-

signed for scene recognition. Then, we design a unified framework to utilize the attention knowledge in the Semantic Network to guide the Appearance Network, and SOSRs are encoded with a sequential model, i.e., recurrent neural network (RNN). These encoded features are weighted by attention mechanism to feed the scene classifier (implemented as Binary Long Short-Term Memory, BiLSTM). To obtain an accurate recognition result, a RGB-D proposal fusion method is proposed for RGB-D scene recognition. We evaluate our approach on the RGB-D datasets such as SUN RGB-D and NYUD2, where the experimental results show that the RSBR model outperforms the state-of-the-art works for scene recognition.

Experiment and Evaluation

Different types of features are combined to compare with other state-of-the-art works. Table 1 and 2 illustrate the comparison results on SUN RGB-D and NYUD2 datasets, respectively.

Models	Accuracy (%)		
	RGB	Depth	RGB-D
State-of-the-art works			
Places CNN+R-CNN (Wang et al. 2016)	36.5	40.4	48.1
MSMM (Herranz et al. 2017)	41.5	40.1	52.3
Global+Local+SOOR (Song et al. 2020)	50.5	44.1	55.5
Proposed			
RSBR (our model)	**57.39**	**50.90**	**62.39**

Table 1: Comparison results on SUN RGB-D dataset

Models	Accuracy (%)		
	RGB	Depth	RGB-D
State-of-the-art works			
Places CNN+R-CNN (Wang et al. 2016)	-	-	63.9
MSMM (Herranz et al. 2017)	-	56.4	65.8
Global+Local+SOOR (Song et al. 2020)	64.2	62.3	67.4
Proposed			
RSBR (our model)	**66.67**	**64.90**	**69.90**

Table 2: Comparison results on NYUD2 dataset

The work was (Wang et al. 2016) proposed to extract both global and local features for scene recognition. Although the work (Herranz, Song, and Jiang 2017) did not explicitly extract local features, the depth model (D-CNN) of that work is trained on the basis of local patches with weak supervision. Thus, local information is implicitly included in that work. Those types of intermediate representations are either extracted from local patches in dense grid or from global images. However, without precisely locating the object regions (e.g., with object detection), the extracted features may not be sufficiently reliable for scene recognition. Based on the object detection results, the information is used to generate SOSR in our work and combine SOSR features with attention mechanism, which contain other types of relations. Although the work (Song et al. 2020) also generated other types of representations for object-object relations, our method uses attention mechanism to weight each SOSR by calculating the correlation between each object and the scene. Moreover, our method RSBR contains two subnetworks. The first is the Semantic Network used to guide the Appearance Network absorbing attention knowledge. The other is Appearance Network,which absorbs attention knowledge of Semantic Network to weight each SOSR,thereby obtaining scene labels by training RGB-D scene images. Thus, our method better addresses the intra-class difference and inter-class similarity of scenes, which then improves the accuracy of scene recognition. Comparing the above works, our proposed RSBR model achieves a better performance.

References

Herranz, L.; Jiang, S.; and Li, X. 2016. Scene Recognition with CNNs: Objects, Scales and Dataset Bias. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 571–579.

Herranz, L.; Song, X.; and Jiang, S. 2017. Depth CNNs for RGB-D Scene Recognition: Learning from Scratch Better than Transferring from RGB-CNNs. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4271–4277.

Song, X.; Jiang, S.; Wang, B.; Chen, C.; and Chen, G. 2020. Image Representations With Spatial Object-to-Object Relations for RGB-D Scene Recognition. *IEEE Trans. Image Process.* 29: 525–537.

Wang, A.; Cai, J.; Lu, J.; and Cham, T. 2016. Modality and Component Aware Feature Fusion for RGB-D Scene Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 5995–6004.

Zhou, B.; Lapedriza, À.; Khosla, A.; Oliva, A.; and Torralba, A. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(6): 1452–1464.