Good Practice on Deep Scene Classification: from Local Supervision to Knowledge Guided Disambiguation

Yu Qiao, Limin Wang, Sheng Guo, Zhe Wang, Weilin Huang, and Yali Wang
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{yu.qiao, sheng.guo, zhe.wang, wl.huang,yl.wang}@siat.ac.cn, {07wanglimin,buptwangzhe2012}@gmail.com

Abstract

Recent years witness remarkable progresses of scene classification task, mainly with deep convolutional neural network trained on large scale datasets like Place. This short paper summarize our recent works toward improving the performance of large scale scene classification with deep networks [3, 4, 1]. These works includes: 1) encoding locally-supervised convolutional features for scene representation, 2) weakly training patch-level CNNs to extract local discriminative descriptors, and 3) exploiting knowledge from extra networks to release label ambiguity problem of scene categories. We will describe the key insights, approaches and results of our works with analysis on the connection and difference among them. Our methods achieves the second place at the Places2 challenge in ILSVRC 2015, and the first place at the LSUN challenge in CVPR 2016. One key question we try to answer with this paper is how scene classification is different from object classification. We hope these investigation and analysis will inspire novel ideas toward the challenging problem of complex scene understanding.

1. Introduction

Scene recognition is a fundamental yet challenging problem in computer vision, with wide applications in image retrieval, object detection, robot vision, image editing etc. Cognitive researches indicate that human has remarkable ability to ability scenes in a very short time. With a glance, we remember the meaning and layout (scene information) of an image but forget object details [2]. After years of research, there still exists a clear gap between computer algorithms and human in understanding scenes. Along with the popularity and success of using deep learning methods in computer vision tasks, deep Convolutional Neural Networks (CNNs) have been widely exploited for scene classification tasks and significantly improve scene recognition accuracy. These progresses can be mainly ascribed to two facts, 1) the design of powerful deep architectures like AlexNet, Inception, VGGNet, ResNet and 2) the construction of large scale scene datasets like Place [6].

Our recent works address the challenging large scale classification problem from different aspects, ranging from exploiting locally-supervised middle-level convolutional features [3], training PatchNet in a supervised way [4], and utilization knowledge from extra well-trained networks [1]. These approaches investigate novel deep architectures for large scale scene classification from different aspects, which further indicate facts on how scene recognition is related and different with object recognition. Our methods achieve state-of-the-art classification performance on three large scene datasets, including MitIndoor (), SUN [5] and Places [6]. The reminder of this paper will summarize and analyze our recent approaches. We hope these analysis can inspire new ideas toward complex scene understanding.

2. Three works

2.1. Locally-Supervised Deep Hybrid Model [3]

Local object information provide rich cues to describe and discriminate scenes. In classical CNN, the classification is based on fully connected layer (FC) features, which may ignore the local details due to the compression operation in the top CNN layers. On the other hand, convolutional features of CNN in middle layer can be seen as detectors of local objects and semantic parts. This fact inspires us to propose a Deep Hybrid Model to integrate FC features and convolutional features for scene classification. To enhance the discriminative ability of convolutional features, we propose a Local Convolutional Supervision (LCS) layer, which directly propagates the label information to the low/midlevel convolutional layers. LCS releases the problem that important scene cues may be undermined by transforming them through the highly-compressed FC layers. We introduce Fisher Convolutional Vector (FCV) that effectively encodes meaningful local detailed information by pooling the convolutional features into a fixed length representation. The FCV contains rich middle-level semantics which is discrim-

inate the ambiguous scenes and robust to local image distortions. FCV with LCS enhancement is strongly complementary to the high-level FC-features, leading to significant performance improvements. Experimental results show that combing FCV with FC-feature improve the accuracy with 7.72% on MIT-Indoor67 and 5.61% on SUN397 using VGGNet.

2.2. Weakly Supervised PatchNets [4]

The previous study [] already verifies the importance using middle-level representation from local regions. So there are questions whether we can train deep networks directly for local regions and how to encode the deep representation in a more effective way? Our work [4] address this question with two contributions. Firstly, we develop PatchNets to extract effective deep representation from local patches. PatchNets are trained a weakly-supervised way with image level annotations. Specially, we train two types of PatchNets with object and scene categories. Secondly, we exploit both output probabilities and FC features from PatchNet to construct a hybrid visual representation, namely vector of semantically aggregating descriptor (VSAD). For an input image, we can extract a set of patch-level FC descriptors by applying Scene-PatchNet on a number of random cropped patches. VSAD encode these patch descriptors with the output probabilities from Object-PatchNet where each type of object is treated as a word in the dictionary for encoding. Experiments show that these methods can further improve the scene recognition performance.

2.3. Knowledge Guided Disambiguation [1]

Large scale image dataset with rich annotation is key to train effective deep models. Scene categories are usually broad and abstractive. Photos within the same category can include different objects and have various layouts, while examples from different scene categories may contain similar objects and structures. Moreover, most large scene datasets include noisy samples. These facts make it challenging to train deep CNNs with scene-category supervision, compared with object classification where categories are well-defined. To address these problems, we propose to use the knowledge from extra domain or dataset to guide the training process of a deep scene CNN (main network). Extra-knowledge deep networks trained on a relatively smaller and well-labeled dataset (e.g. ImageNet) are exploited to provide additional supervision. Main network receives two types of supervision, one is the traditional classification loss and the other is the prediction loss of extra-knowledge deep networks,

$$\ell(D) = -(\sum_{\mathbf{I}_i \in D} \sum_{k} \mathbb{I}(y_i = k) \log p_{i,k} + \lambda \sum_{\mathbf{I}_i \in D} \sum_{m} q_{i,m} \log f_{i,m}) \quad (1)$$

Table 1. Performance on MIT-Indoor 67 and SUN397.

Model	MIT Indoor67	SUN397
Places205-GoogLeNet [3]	74.0%	58.8%
LS-DHM [3]	83.8%	67.6%
PatchNet-VSAD [4]	84.9%	71.7%
MS-KD [1]	**86.7%**	**72.0%**

where $\mathbf{I}_i$ is the i^{th} image in training dataset D, y_i is the ground-truth scene label (hard label), and p_i is corresponding predicted scene label. f_i is produced by extra knowledge network, and q_i is the predicted soft code of image $\mathbf{I}_i$. λ is a parameter balancing these two terms.

The performance of the three works on MIT-Indoor67 and SUN397 are compared in Table 1. The last method with multi-resolution extension [1] achieved the second place at the Places2 challenge in ILSVRC 2015, and the first place at the LSUN challenge in CVPR 2016.

3. Conclusions

Scene classification is the first task toward understanding complex scenes. We have several conclusions from the above three works. Firstly, scene classification is high related with local objects and semantic stuffs. Constructing middle-level representation with convolutional features or PatchNets can yield rich complementary cues to the FC feature extracted with classical CNNs, which further lead to performance gain. Secondly, rich and clean supervision is vital for training deep networks with high performance. This is partly important for scene classification where the categories can be abstractive and ambiguous. One can enrich the supervision with patch-level object information or with the knowledge from extra networks trained in different domains.

References

[1] W. H. Y. X. L. Wang, S. Guo and Y. Qiao. Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns. *IEEE T-IP*, 26(4):2055C–2068, 2017.

[2] A. Oliva. Gist of the scene. *Neurobiology of attention*, 696(64):251–258, 2005.

[3] L. W. S. Guo, W. Huang and Y. Qiao. Locally supervised deep hybrid model for scene recognition. *IEEE T-IP*, 26(2):808C–820, 2017.

[4] Y. W. B. Z. W. Wang, L. Wang and Y. Qiao. Weakly supervised patchnets: Describing and aggregating local patches for scene recognition. *IEEE T-IP*, 26(4):2028C–2041, 2017.

[5] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[6] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.