Scene Classification with Deep Convolutional Neural Networks

Yangzihao Wang
University of California, Davis
yzhwang@ucdavis.edu

Yuduo Wu
University of California, Davis
yudwu@ucdavis.edu

Abstract

The use of massive datasets like ImageNet and the revival of Convolutional Neural Networks (CNNs) for learning deep features has significantly improved the performance of object recognition. However, performance at scene classification has not achieved the same level of success since there is still semantic gap between the deep features and the high-level context. In this project we proposed a novel scene classification method which combines CNN and Spatial Pyramid to generate high-level context-aware features for one-vs-all linear SVMs. Our method achieves the state-of-the-art result: 68.04% accuracy rate on MIT indoor67 dataset using only the deep features trained from ImageNet.

1. Related Work

Scene classification means to provide information about the semantic category or the function of a given image. Among different kidn of scene classification tasks, the indoor scene classification is considered to be one of the most difficult since the lack of discriminative features and contexts at the high level [6]. Spatial pyramid representation[5] is a popular method used for scene classification tasks. It is a simple and computationally efficient extension of an orderless bag-of-features image representation. However, without a proper high-level feature representation, such schemes often fail to offer sufficient semantic information of a scene. Object bank[3] is among the first to propose a high-level image representation for scene classification. It uses a large number of pre-trained generic object detectors to create response maps for high level visual recognition tasks. The combination of off-the-shelf object detectors and a simple linear prediction model with a sparse-coding scheme achieves superior predictive power over similar linear prediction models trained on conventional representations. However, this method also limits the performance of their system to the performance of the object detectors they choose. Recently, Convolutional Neural Networks (CNNs) with flexible capacity makes training from large-scale dataset such as ImageNet [1] possible. In the work of A. Krizhevsky et al.[4], they trained one of the largest CNNs on the subsets of ImageNet and achieved better results than any other state-of-the-art methods in 2012. While their CNN system focuses on object detection, the features generated can be used for other applications such as scene classification. Two types of improvements has been done on top of their CNN works. The first type of improvement tries to address the problem of generating possible object locations in an image. Selective search method [7] combines the strength of both an exhaustive search and segmentation and results in a small set of data-driven, class-independent, high quality locations. Girshick et al. propose the Regions with CNN features (R-CNN) method [2] as a more effective feature generation method. Alternatively, Zhou et al. try to increase the performance of scene classification using CNN by creating a new scene-centric database [8].

1.1. Technical Approach

Describe in detail the feature representation(s) and algorithm(s) you employed. The description should be self-contained (i.e., the reader should not have to rely on outside sources for your points to be clear), and should provide enough detail so that the reader could re-implement the approach. Clearly state the method's input and output, and any assumptions or design choices;

1.2. Experiments

Describe the experiments you conducted to evaluate the approach. For each experiment, describe what you did, what was the main purpose of the experiment, and what you learned from the results. Provide figures, tables, and qualitative examples, as appropriate.

1.3. Conclusions

briefly summarize the main idea and results, and possible future work.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In

Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255, June 2009. 1

[2] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. 1

[3] L. jia Li, H. Su, L. Fei-fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1378–1386. Curran Associates, Inc., 2010. 1

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178, 2006. 1

[6] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420, June 2009. 1

[7] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 1

[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014. 1