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% average 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 kinds 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 [8]. Spatial pyramid representation[7] 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[5] 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 [2] possible. In the work of A. Krizhevsky et al.[6], 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 [9] 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 [3] 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 [10].

2. Technical Approach

Previous work on Convolutional Neural Networks (CNNs) implies that it may capture the high-level representation of an image using a certain deep layer feature set. Our goal of this project is to answer one single question: *Whether CNNs can help with the feature representation to extract high-level inforamtion of an image scene and thus improve the scene classification precision?* We choose a CNN which is pre-trained on ImageNet dataset (ImageNet-CNN) since its a large-scale general object recognition datasets which consists of over 15 million labeled high-resolution images in over 22,000 categories. We use CNN pretrained on such dataset with the hope to reduce the chance of overfitting to certain scenes. To utilize a pre-trained ImageNet CNN and for the efficiency of the feature extraction process, we use a popular library: Caffe [4].

For the training process, our system takes all images in the training set for each category as the input, use the ImageNet-CNN to perform a prediction for each image. Instead of getting the final 1000 length prediction vector, we extract the FC 7 layer feature set which contains 4096 response values. We then use such features to train one linear SVM model for each scene category. For the testing process, an input image goes through the same ImageNet-CNN and its 4096 length deep feature vector are used to predict its scene classification for each linear SVM model and we assign the one with highest confidence score.

extract around 2000 bottom-up region proposals using selective search[9], then extract 4096-dimensional feature vectors for each region proposal using a large convolutional neural network (CNN) library Caffe[**?**], after we got feature for each region proposal, we apply spatial pyramid and do max-pooling to find the features that contribute most. We then perform a L2 normalization procedure for all feature matrices. The final feature matrices then used for classification using multi-class linear SVMs classifier. Our method achieves a mean average precision (mAP) of 68.2953% on dataset MIT-indoor67[8] without fine-tune on this dataset. For comparison, we implement using only 4096-dimensional feature vectors extracted from Caffe without region proposals, spatial pyramid matching and max-pooling which has the mAP of 59.9507%.

2.1. Selective Search

A variety of recent research offers methods for generating category-independent region proposals for possible object locations. Selective search is widely used for generating possible object locations for use in object recognition[9]. Same strategies can be adopted on indoor scene classification. For the indoor scenes that can be well characterized by objects they contain, the selective search can exploit local discriminative information with greatly reduced number of locations compared to an exhaustive search. We use selective search to generate region proposals. Caffe provides a general Python interface for models and it has built in interface for selective search. We only need to change the setting of CROP_MODES to selective_search, we can operate on around 2000 region proposals instead of the entire image.

2.2. Feature Extraction

Caffe [4] is an open source convolutional architecture for fast feature embedding which contains pre-trained models. In our project, we use pre-trained BVLC Reference CaffeNet to classify images and extract 4096-dimensional layer 7 feature vectors from each region proposal using Caffe [4] implementation of the CNN described by Krizhevsky et al[6]. It provides an option to output the features in certain layer rather than only the final classification results. We set the *blobs* option to *fc7* in order to obtained the Layer 7 feature vectors for each of the region proposals. Therefore, after this step, for one input image, we obtain around 2000 feature vectors and each have the dimension of 4096.

2.3. Max Pooling

2.4. Spatial Pyramid Matching

For each image, a three-level spatial pyramid representation is used, resulting $numImages * numWindows * (1 + 4 + 16)$ length feature vectors.

2.5. L2 Normalization

We then perform the L2 normalizations for each of the feature matrix. L2 normalization is computed by the square root of the sum of each feature's square and for each feature in the feature vector. By dividing each feature vector the L2 normalization value, we have each feature vector's L2 normalization = 1. This L2 normalizations have only a modest effect on the overall performance perhaps simply because the original feature vector data already in the similar scale.

2.6. Training

SVMs (Support Vector Machines) are a useful technique for data classification. LibSVM[1] is an integrated software for support vector classification that is widely used in variety of classification tasks. It supports multi-class classification which is used in this project. We trained 67 binary one-vs-all SVM classifiers each for one category in MIT-indoor67 dataset. In order to fit the required LibSVM training file format, for each category training file, we add label +1 to each feature vector that belongs to the category and label -1 to all feature vectors that belong to rest of categories. We also move all instances that belonging to the current positive category (+1 labeled feature vectors) to be at the top of the feature matrix, this would guarantee the correctness even if LibSVM might internally map the label of the first training instance to be +1 regardless of its actual label value.

3. Experiments

Describe the experiments you conducted to evaluate the approach. For each

In this section, we evaluate our method on the indoor scene dataset: MIT-indoor67 [8], which includes 15,620 images over 67 indoor scenes. Suggested training and testing list of images are used to do the training (80 images per class) and validation (20 images per class). There are at least 100 images per scene and all in .jpg format.

Multi-class classification is done with a support vector machine (SVM) trained using one-versus-all rule, that is, each classifier is learned to separate each class from the rest of classes. Test image is assigned the label of the classifier with the highest response.Scene classification performance is evaluated by average multi-class classification accuracy over all scene classes.

For comparison purposes, we implement with the same procedure but only use the extracted layer 7 4096-

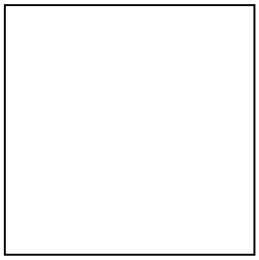

Figure 1. The overview of our system. For an input image, a selective search algorithm is applied first to get roughly 2000 regions of interest. We then apply a pre-trained Convolutional Neural Network (CNN) on each region of interest to get a deep feature vector of length 4096. A three-level spatial pyramid representation of the image with deep feature is used to create the final feature representation. At each level, for each spatial bin, we use max pooling to get the largest feature value of all the feature values of the regions of interest which fall into that spatial bin, resulting in the final feature of length No.Deep Features × (1+4+16) as a high-level representation of the input image. Then multiple one-vs-all linear SVMs are used to do the scene classification.

dimensional feature vectors from Caffe. After we get one feature vectors for each entire image, instead of perform spatial pyramid and L2 normalization, we simply add labels and send them into the multi-class SVMs. Validation image feature vectors are also generated in the same way.

Table 1. Comparison results on MIT-indoor67

Models	Average Precision
Our Method	**68.2953%**
Without Normalization	68.0469%
Only CNN Feature	59.9507%

We compare our scene classification tasks with the performance without perform L2 normalization and the performance of using only the features extracted from CNN, summarized in Table 1. Around 14% improvement of using region proposals, max-pooling and spatial pyramid are shown. In majority of categories, we perform much better on the average precision. Some examples are shoes shop, bedroom, bowling, grocery store, hospital room and operating room. This might due to the region proposals and spatial pyramid technique allow us to better characterize the particular objects belong to the category. However, there are also some drops of average accuracy using our methods and mainly for these three categories: prison-cell, library and living room. These three categories are all relatively easier to be characterized by global spatial properties.

4. Conclusions

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

References

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2

[2] 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

[3] 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

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 1, 2

[5] L. jia Li, H. Su, L. Fei-fei, and E. P. Xing. Object bank: A high-level image representation for scene classification &amp; 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

[6] 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, 2

[7] 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

[8] 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, 2

[9] 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, 2

[10] 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