Scene Classification with Deep Convolutional Neural Networks

Yangzihao Wang and Yuduo Wu
University of California, Davis
{yzhwang,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 information 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 pre-trained on such dataset with the hope to reduce the chance of over-fitting 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]. To better observe the impact of a good feature representation, we choose a very difficult dataset: MIT-Indoor67 dataset, which includes 15,620 images of over 67 indoor scenes. Object Bank achieves only 37.6% recognition rate on this dataset. We expect that using deep features extracted from CNNs can significantly improve the results on this dataset.

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.

Within this general framework, several methods can be explored to improve the feature representation. Instead of using the entire image for deep feature extraction, we can first select a set of region proposals (usually around 2000 for a high-resolution image) which are most informative about the image, then extract 4096-dimensional feature vectors for each region proposal. This improvement puts more weights on more informative regions. However, using the concadenated feature vector would result in an extremely high dimension (2000 $\times$ 4096), it is necessary to use some sparse coding scheme to represent the global feature pattern. In this paper, we adopt spatial pyramid because of its simplicity and effectiveness. To increase the generalization ability of our features and to reduce the impact of overfitting, we further apply $l2$ normalization to the achieved feature vectors before we feed them to the one-vs-all SVMs. Figure 1 shows the pipeline of our system. We will describe the details in the following sections.

2.1. Generating Region Proposals

A variety of recent research offers methods for generating category-independent region proposals for possible object locations. Selective search is one of the most widely used methods for generating possible object locations for use in object recognition[9]. We argue that same strategy can be adopted on the indoor scene classification task, because most indoor scenes can be well characterized by objects they contain. 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 a built in interface for selective search. By changing the setting of "CROP_MODES" to "selective_search", we can load the selective search method to generate roughly 2000 region proposals for an image to feed to the CNN instead of using an entire image.

2.2. Feature Extraction

As we mentioned, we use Caffe, an open source convolutional architecture for fast feature embedding which contains pre-trained models. Specifically, we use pre-trained BVLC Reference CaffeNet to extract 4096-dimensional FC 7 feature vectors from each region proposal. In the Python interface of Caffe, there is 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 FC 7 feature vectors. The reason to choose FC7 is because it is the last hidden layer of the CNN, which is supposed to contain the most informative features. After this step, for one input image, we obtain a 4096-dimension feature vector for each proposed region.

2.3. Spatial Pyramid Feature Representation

The feature vectors created in the previous step has too many dimensions for both training and prediction. We adopt spatial pyramid matching to generate a more compact feature vector for an image while still preserve most visual information in the extracted deep features. Our spatial pyramid has three levels, each is generated by equally dividing each rectangular spatial bin of the previous level into four sub-bins. We generate a single 4096-dimension feature vector for each spatial bin by max pooling over all the deep features of the proposed regions that fall in the spatial bin. For three pyramid levels, there would be $1 + 4 + 16$ spatial bins in total, where the first 4096-dimension feature vector represents the overall visual information of the image, the following 4$\times$4096-dimension feature vector represents the mid-level visual information of the image, and the final 16$\times$4096-dimension feature vector the low-level. We will show the improvement brought by this feature representation over a simple entire image 4096-dimension feature in Section 3. A $l2$ normalization is followed for better convergence and less over-fitting.

2.4. Model Training and Prediction

Support Vector Machine (SVM) is 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 use libSVM's linear classifier with confidence value option to train 67 one-vs-all models each for one category in our MIT-indoor67 dataset. During the scene classification phase, for every testing image, we run the prediction against all 67 categories and classify the image to the category which has the highest confidence score.

3. Experiments

In this section, we evaluate our method on the MIT-indoor67 dataset. 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 jpeg format.

Multi-class classification is done with a support vector machine (SVM) trained using one-versus-all rule, that is,

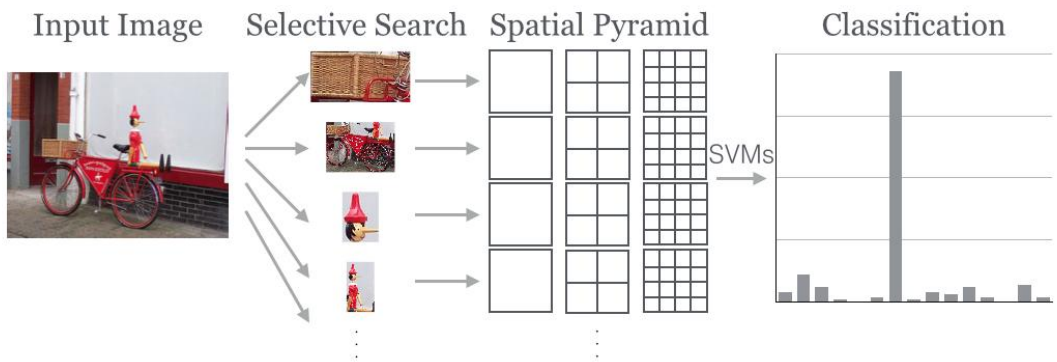

Figure 1. The overview of our system. For an input image, a selective search algorithm is applied first to get roughly 2000 regions of interest. We then apply a pre-trained Convolutional Neural Network (CNN) on each region of interest to get a deep feature vector of length 4096. A three-level spatial 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 4096 $\times$ (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.

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-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.

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%.

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

[5] 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

[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

[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 Recogni-*

Table 1. Comparison results on MIT-indoor67

Models	Average Precision
$l2$ Norm + Selective Search + Spatial Pyramid	**68.2953%**
Selective Search + Spatial Pyramid	68.0469%
Entire Image CNN Features	59.9507%

tion, 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, 3

[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

Appendix: Confusion Matrix

Confusion matrix is shown in Figure.**??** and some sample training and testing images from the MIT-indoor67 dataset are shown in Figure. **??**

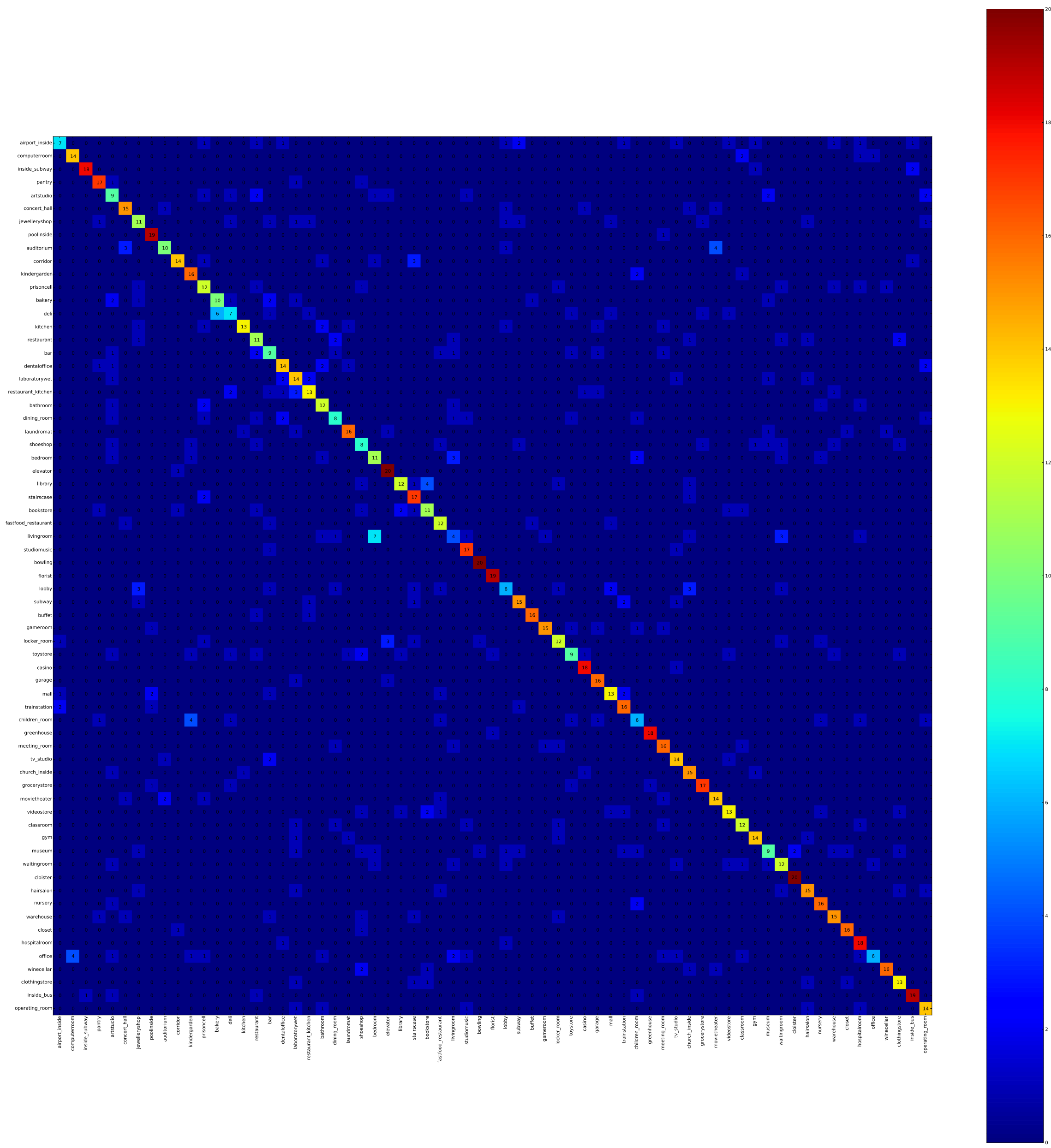

Figure 2. From the confusion matrix of all 67 categories, we notice that several pairs of categories are relatively easier to be confused and misclassified by the SVMs with each other, such as bakery and deli, living room and bedroom as well as bookstore and library.

Figure 3. This figure contains some samples pairs from the MIT-indoor67 dataset. For each line of image, left two are from the same category and right two are from another. The first pair is bakery and deli which have very similar patterns of listing of breads and sandwiches. The second pair is living rooms and bedrooms, living room images may include some beds or beds-like sofas, which is almost identical to the scenes in bedroom. It is extremely hard, even for human beings, to be able distinguish between library and bookstore. the scene in bedrooms.