CS676: IMAGE PROCESSING AND COMPUTER VISION

Scene Recognition
Using Mid-level features from CNN

Ashudeep Singh
10327162
ashudeep@iitk.ac.in

Anant Raj
10327xxx
anantraj@iitk.ac.in

Vishal Kumar Gupta
10xxx
vishalkg@iitk.ac.in

In this project we try to explore how the features extracted from the activation of a deep convolutional neural network trained in a supervised fashion on the ImageNet dataset can be used to classify in nivel generic tasks such as scene recognition. We use the mid-level features from the pretrained CNN hypothesising that they contain semantic information as relevant for the task of scene recognition. We consider a spatial pyramid kind of setting for the image, which represents ImageNet objects of scale upto the size of the image. We present that the results, obtained on MIT-67 Scene Dataset and 15 scene dataset, are quite impressive as compared to the state-of-the-art.

1 INTRODUCTION

A convolutional neural network is a type of neural network where the individual neurons are tiled in such a way that they respond to overlapping regions in the visual field.[1] Convolutional networks were inspired by biological processes and are variations of multilayer perceptrons which are designed to use minimal amounts of preprocessing. The results obtained by classification using convolutional neural networks have been the state-of-the-art on many datasets and hence they are widely used models for image recognition.

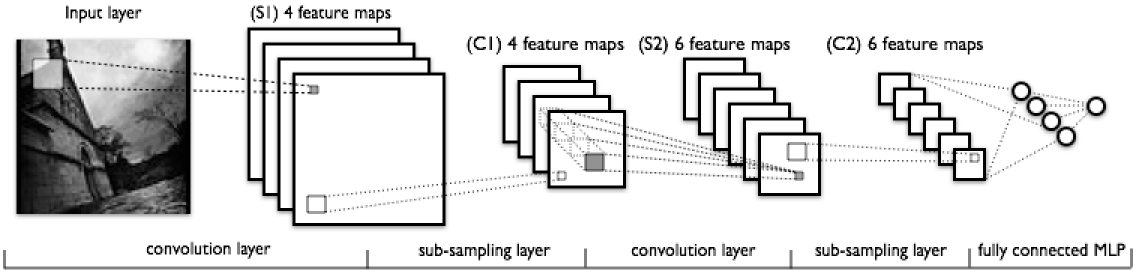


Figure 1.1: A representation of the structure of a Convolutional Neural Network characterized by alternating convolution and spooling layers[2]

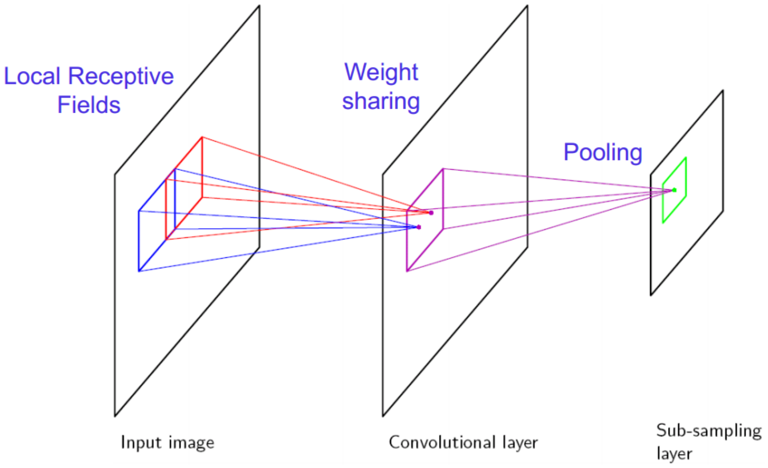
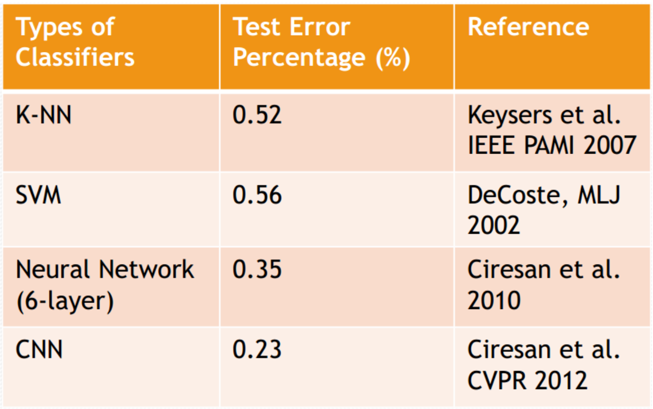


Figure 1.2: Local Receptive Fields and Weight sharing

2 Motivation

There is a good enough motivation to use Convolutional Neural Network based models for various computer vision tasks because of the state-of-the-art results obtained by such models in recognition tasks. Considering MNIST dataset, CNN error rates from Ciresan et al.[3] are state-of-the-art at 0.23%.



Types of Classifiers	Test Error Percentage (%)	Reference
K-NN	0.52	Keysers et al. IEEE PAMI 2007
SVM	0.56	DeCoste, MLJ 2002
Neural Network (6-layer)	0.35	Ciresan et al. 2010
CNN	0.23	Ciresan et al. CVPR 2012

Figure 2.1: A few results on MNIST datasets

Besides this comparing the results of various techniques at the Large scale Visual Recognition challenge 2012, we may see that the models based on CNN are the best performaing ones. Also we must note here that minimal pre-processing is required while using CNNs to classify images.

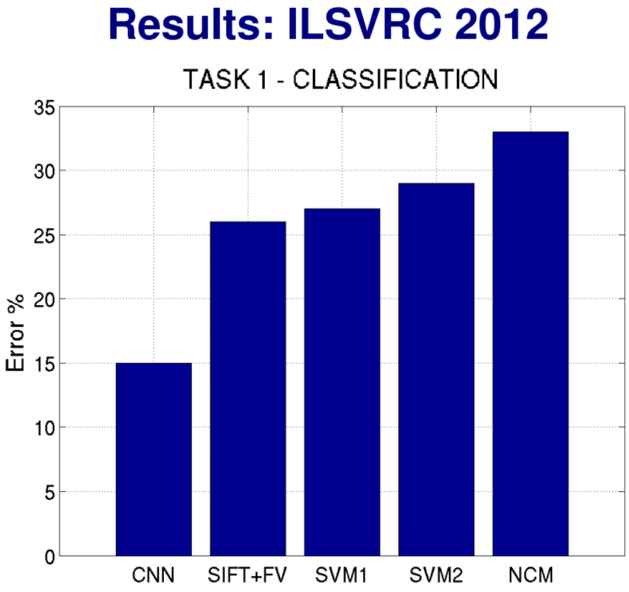


Figure 2.2: Results for various techniques in Large scale Visual Recognition challenge 2012[4]

2.1 Motivation for using Mid-level Features

In human also at very first layer we perceive edgesi.e. low level features. At the end we perceive the whole object as semantics. It is lesser known that what happens in middle layers of the human neural network. Mid-level features represent a mixture of low-level and high-level

features which could be important to distinguish intra-class variations. Hence, using the middle level features from a trained Convolutional Neural Network, which is a model of human vision system, has become a trend in the recent times.
Consider the following visualization of the features at various depths inside a Convolutional neural network trained on different objects. We can observe that:

- The shallow level features represent edges.
- As we go deeper into the network, the filters tend to reflect semantic information also.

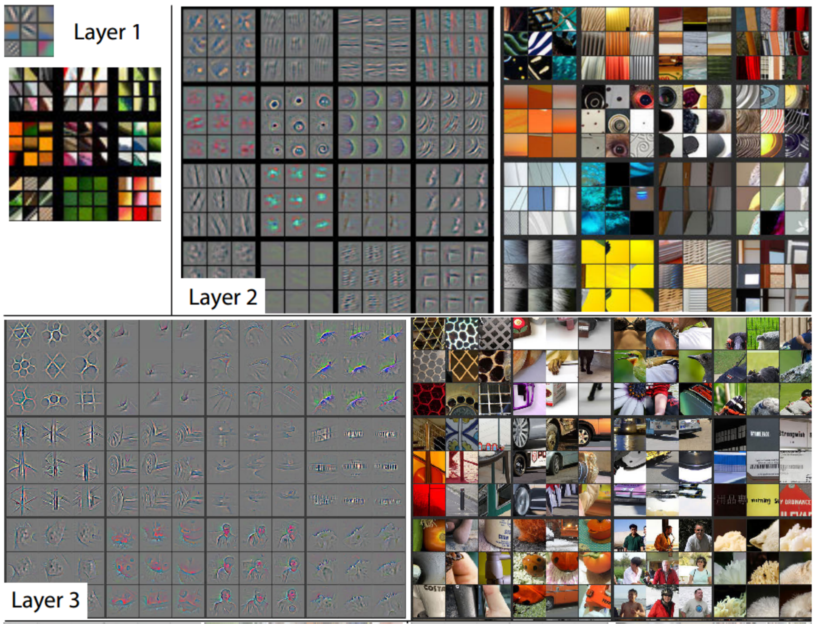


Figure 2.3: Visualization of the features at various depths inside a Convolutional neural network trained on different objects

3 Dataset, Tools and Techniques

3.1 Dataset

1. **MIT Indoor Scene Recognition Dataset**[5]:
The dataset consists of images from 67 indoor scene categories. The total number of the images is 15620. The number of images varies accross categories, but there are at least 100 images per category.

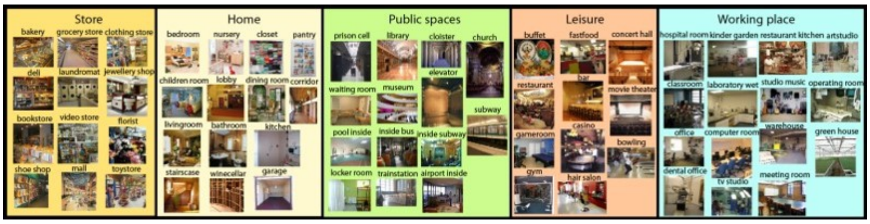


Figure 3.1: MIT 67 Dataset classes

2. **15-scene dataset**[6]:
   The dataset consists of around 4500 images which are from 15 outdoor scenes and the images are grayscale.

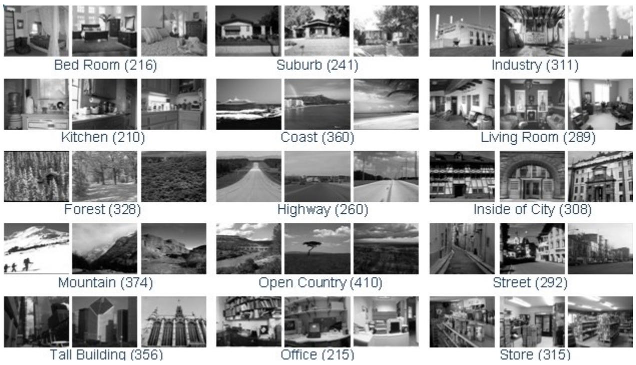


Figure 3.2: 15 scene dataset classes and examples

3.2 DeCaf: A Deep Convolutional Activation Feature for Generic Visual Recognition[8]

We use DeCAFto extract the middle layer pretrained CNN features on ImageNet. We experiment whether the features extracted from the activation of a deep convolutional network trained in a supervised manner on ImageNet Dataset can be re-purposed to novel generic tasks such as scene recognition.

3.3 Other tools and libraries

We use VLFeat library [7] for Chi-squared kernel and training our SVMs. We use our own implementations of feature extraction in a spatial pyramid, then feature map normalization for 5th layer features and then training and testing scripts.

4 OUR APPROACH

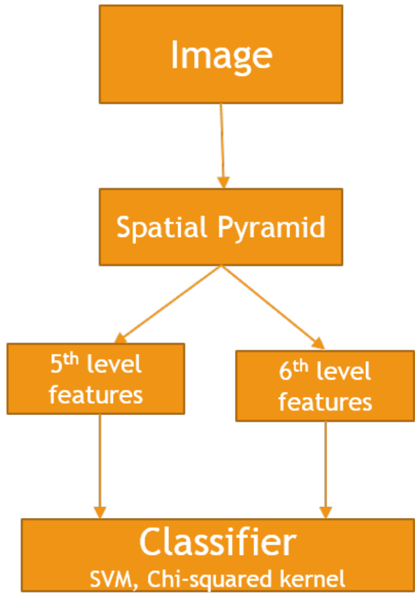


Figure 4.1: The basic flow of out approach

We divide the image into 4 quadrants to get the following two kinds of features using DeCaf for the whole image as well as the 4 parts:

1. 5^{th} layer features: These features come from the pretrained imagenet CNN from the last convolutional layer output. The features are of size $6 \times 6 \times 256$ i.e. 256 feature maps of sizes 6×6. We then normalize these features over depth using L1 norm i.e. dividing each pixel intensity in particular position for all feature maps by the sum of their absolute intensity values. We then concatenate the features for all the levels of the spatial pyramid.

2. 6^{th} layer features: These features come from the first fully connected layer in the pretrained CNN. It is a single vector of dimension 4096×1, all of which from each pyramid level are concatenated to form a feature descriptor.

We use a SVM classifier to train a model to classify for the different classes in a one-vs-all fashion. We use the chi-squared kernel map to map the feature space.

We have used only 37 out of the 67 classes in the MIT-67 dataset classes because of memory and time constraints.

5 Results

We obtain results which are quite comparable to the state-of-the-art. But the training and testing has been done only on 37 classes of the MIT-67 dataset. We expect the results to improve the state-of-the-art in the future work.

Features Used	Method	Accuracy
5^{th} layer	Normalized Features SVM, Chi-squared kernel	68.23%
6^{th} layer	Normalized Features SVM, Chi-squared kernel	60%
5^{th} layer, 2 level spatial pyramid	Normalized Features SVM, Chi-squared kernel	74.5%
5^{th} layer, 3 level spatial pyramid	Bag-of-words, vocab_size=16, SVM classifier	34.6%

We can observe that using a bag of words model to classify is worse than using the 6^{th} layer features, which makes us think that the max-spooling between the 5th and 6th layers is a much better combining technique rather than using a bag-of-words approach.
We obtain our best results using 5^{th} layer features for 2 levels in a spatial pyramid. We wish to improve upon the results by trying this out on another level of the spatial pyramid.

6 Future Work

1. Explore the use of features at the 5th layer by modifying:
   - The normalization of features
   - Using a Bag of Words histogram approach.
2. **Use a reconfigurable part model[10]:** A model that is based on the assumption that the different parts of the scene can be reconfigured in certain ways.
   This problem can be formulated into a Latent Structural SVM problem with latent parameters as the permutation of the various regions in the grid space of the image. The LSSVM is expected to learn the correct dependence of the features on their relative/absolute positions.

References

[1] LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361.

[2] Deep Learning Tutorial. http://deeplearning.net/tutorial/lenet.html

[3] Ciresan, D., Meier, U., & Schmidhuber, J. (2012, June). Multi-column deep neural networks for image classification. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 3642-3649). IEEE.

[4] Ranzato, M., Huang, F. J., Boureau, Y. L., & Lecun, Y. (2007, June). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on (pp. 1-8). IEEE.

[5] A. Quattoni, and A.Torralba. Recognizing Indoor Scenes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[6] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[7] Vedaldi, A., & Fulkerson, B. (2010, October). VLFeat: An open and portable library of computer vision algorithms. In Proceedings of the international conference on Multimedia (pp. 1469-1472). ACM.

[8] Donahue, Jeff, et al. "Decaf: A deep convolutional activation feature for generic visual recognition."arXiv preprint arXiv:1310.1531(2013).

[9] Zeiler, Matthew D., and Rob Fergus. "Visualizing and Understanding Convolutional Neural Networks."arXiv preprint arXiv:1311.2901(2013)

[10] Parizi, Sobhan Naderi, John G. Oberlin, and Pedro F. Felzenszwalb. "Reconfigurable models for scene recognition." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.