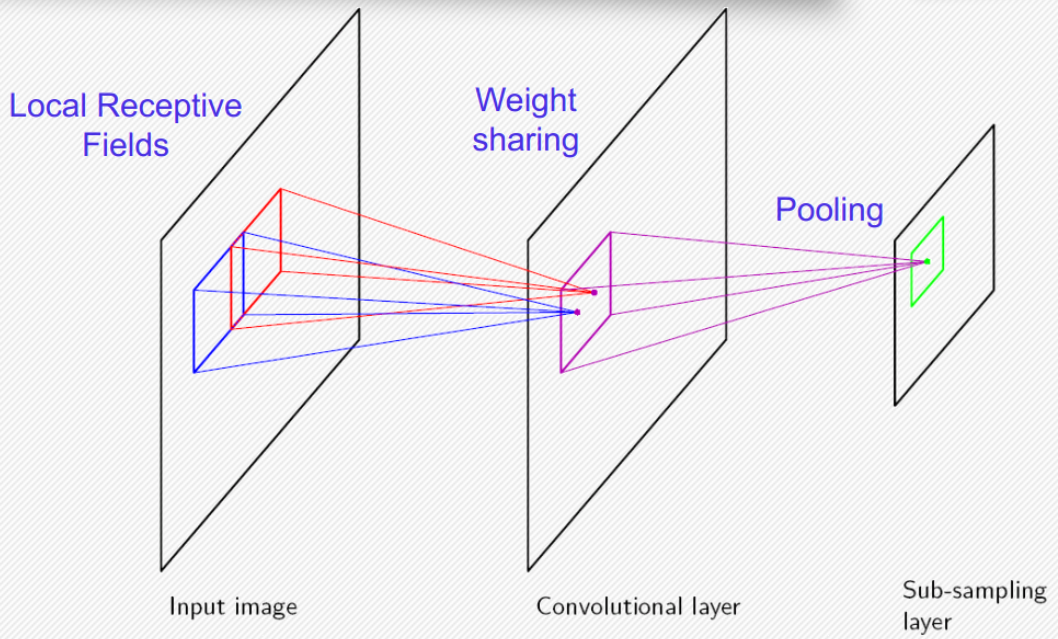
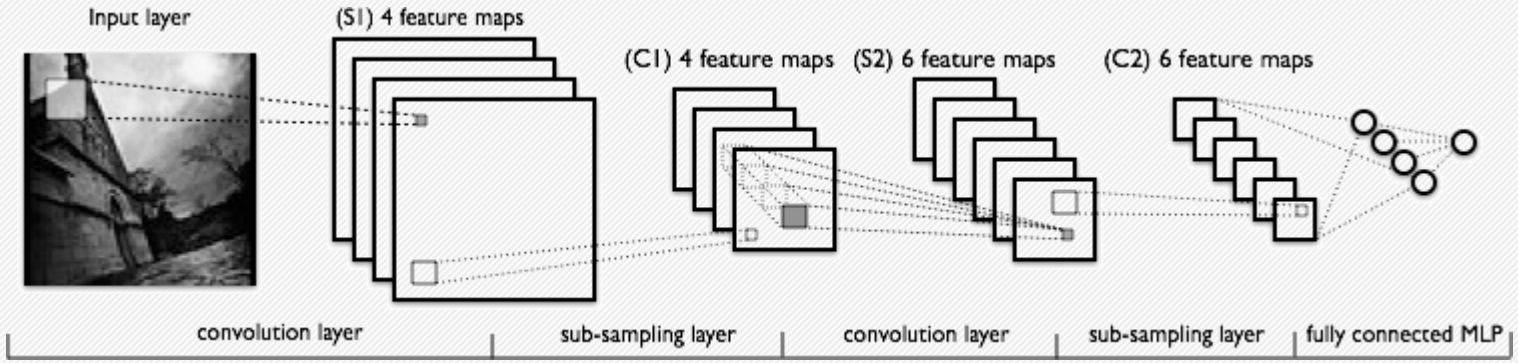
Scene Recognition
using Mid-level features from CNN

Ashudeep Singh
Anant Raj
Vishal Kumar Gupta

Convolutional Neural Networks

- A convolutional neural network is a type of feed-forward neural network where the individual neurons are tiled in such a way that they respond to overlapping regions in the visual field.
- Convolutional networks were inspired by biological processes and are variations of multilayer perceptrons which are designed to use minimal amounts of preprocessing.
- Widely used models for image recognition.



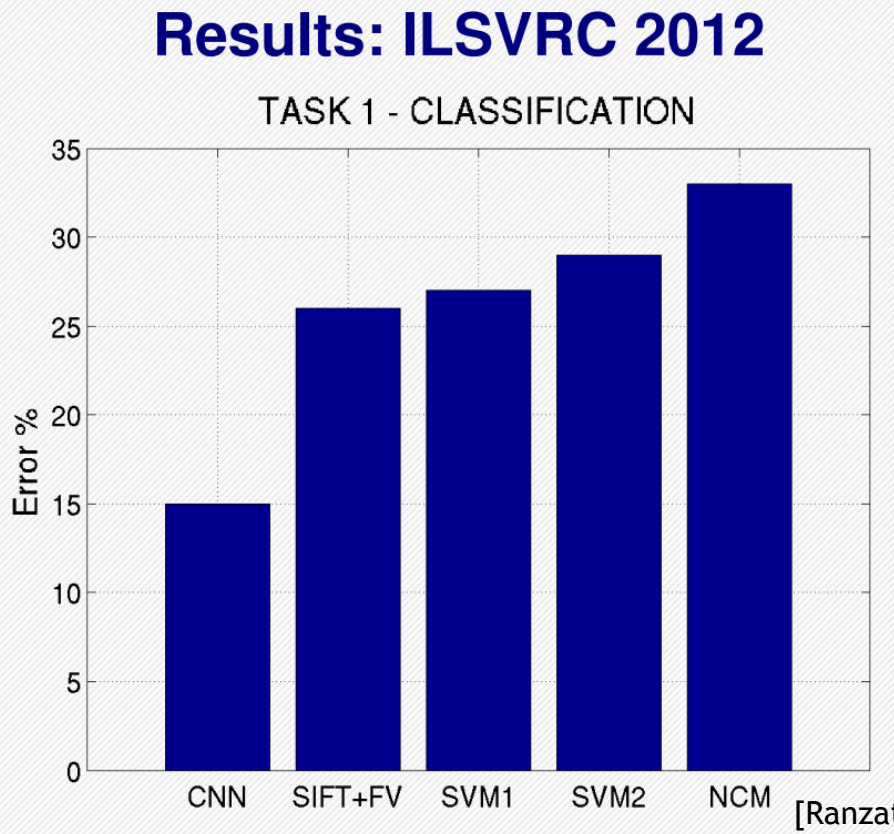


Why CNN?

- **Results on MNIST**

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

- **Results on ImageNet**



[Ranzato et al '13]

Scene Recognition

MIT Indoor Scene Recognition Dataset

- The database contains 67 Indoor categories, and a total of 15620 images.
- The number of images varies across categories, but there are at least 100 images per category.

Scene Recognition

15 Scene Dataset

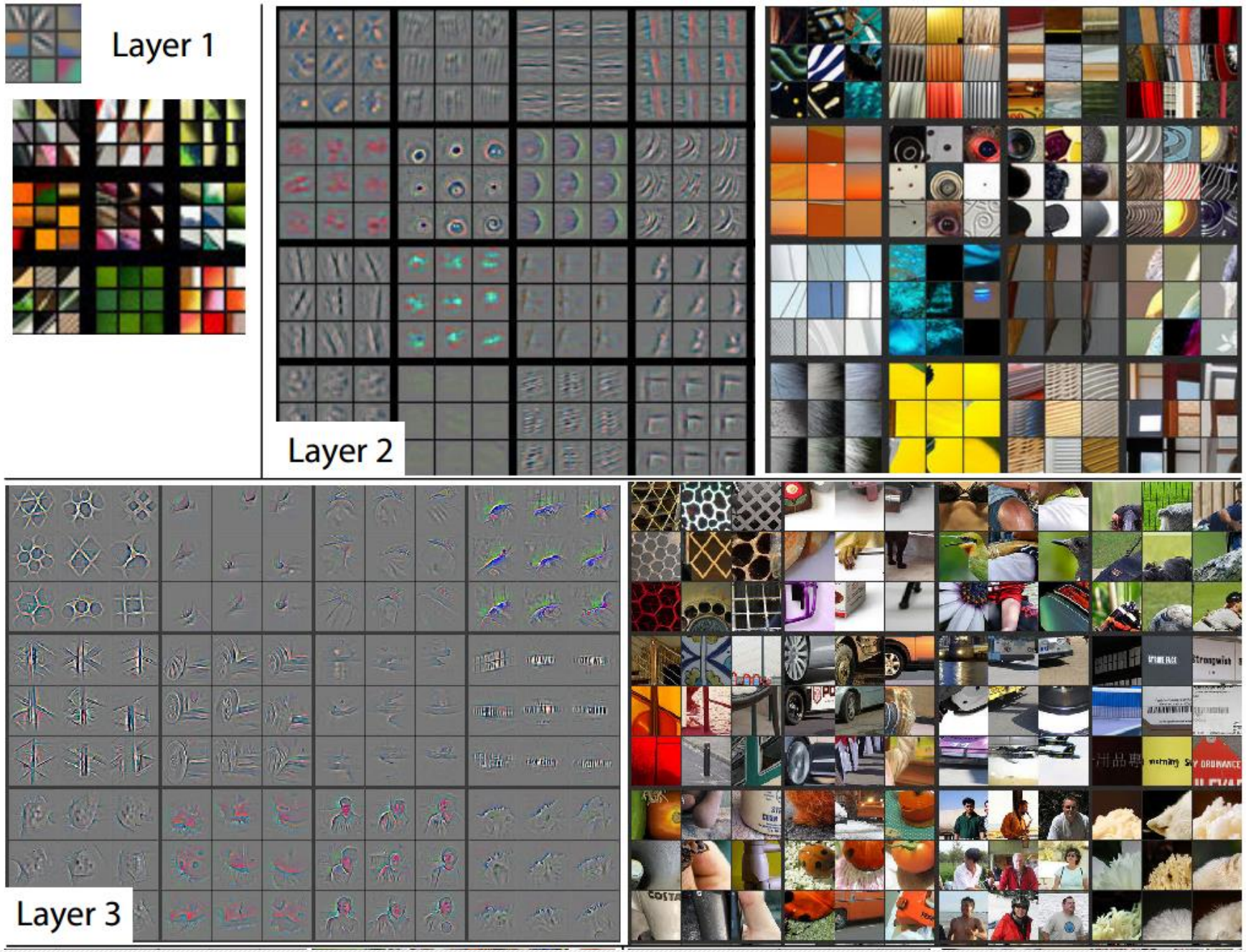
- ~4500 Grayscale images
- Outdoor scene classes

Motivation for using Mid Level Features

- In human also at very first layer we perceive *edges* i.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.

Visualizing the features

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



Pre-Trained ImageNet Features

- DeCAF - **A Deep Convolutional Activation Feature for Generic Visual Recognition**
  - **We use DeCAF to 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.

Our approach

Use the following features out of the pretrained CNN:

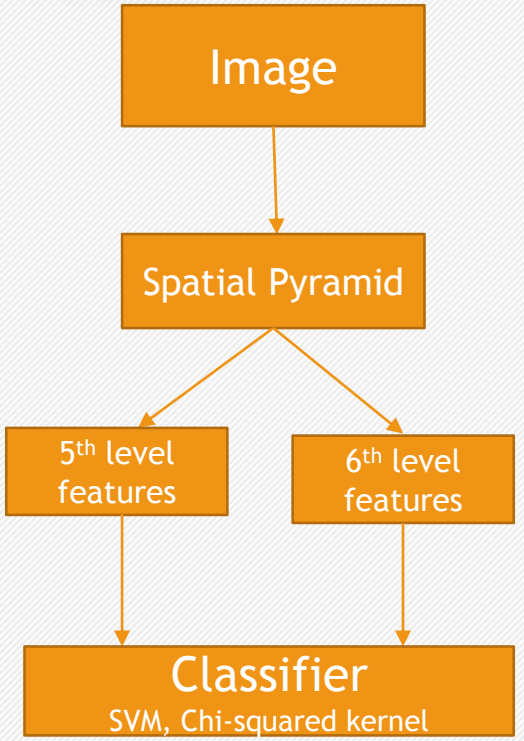
- *5th layer features*

  the last convolutional layer output, of size 6x6x256.

- 6th layer features:

  the 4096 dimensional feature after the first fully connected layer.

  - *Divide the image into 4 quadrants to get the above features for the whole image as well as the 4 parts.*
  - *Bag of words of 5th level image features.*
  - *Spatial histogram*



Results

Features used	Method	Accuracy
5th layer	Normalized features SVM, Chi-squared kernel	68.23%
6th layer	Normalized features SVM, Chi-squared kernel	60%
5th layer, 3 layer spatial histogram		
	Bag-of-words, vocab-size=16	34.6%

Results on 37 classes out of 67 in CVPR MIT-67 Dataset

Further Work

- Explore the use of features at the 5^{th} layer by modifying:
  - The normalization of features
  - Using a Bag of Words histogram approach.
- Use a reconfigurable part model: A model that is based on the assumption that the different parts of the scene can be reconfigured in certain ways.

Reference:

- Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *arXiv preprint arXiv:1311.2524* (2013).
- Donahue, Jeff, et al. "Decaf: A deep convolutional activation feature for generic visual recognition." *arXiv preprint arXiv:1310.1531* (2013).
- Zeiler, Matthew D., and Rob Fergus. "Visualizing and Understanding Convolutional Neural Networks." *arXiv preprint arXiv:1311.2901* (2013).
- Razavian, Ali Sharif, et al. "CNN Features off-the-shelf: an Astounding Baseline for Recognition." *arXiv preprint arXiv:1403.6382* (2014).