Pyramid Pooling of Convolutional Feature Maps for Image Retrieval

Lehrstuhl und Institut für Nachrichtentechnik
RWTH AACHEN UNIVERSITY

Outline

- Motivation
- Neural network model
- Spatial bins
- Pyramid pooling
- Experiments and results
- Conclusions

Motivation

- With advent of Convolutional Neural Networks (CNNs), neural network based feature extraction is used in image retrieval

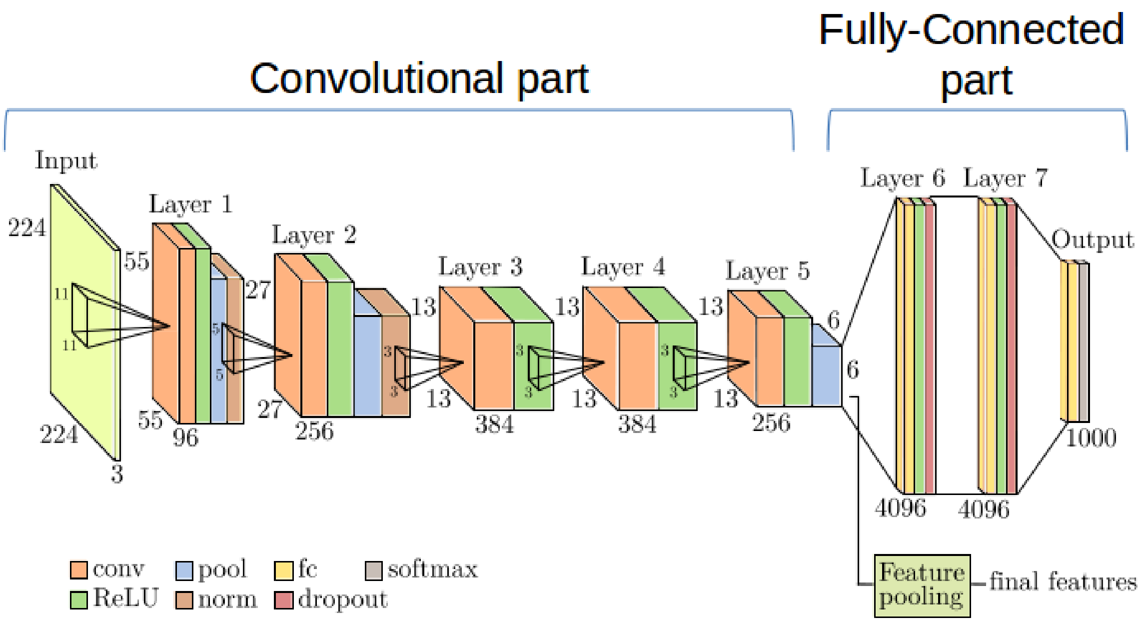
- We address 2 main issues in this work:

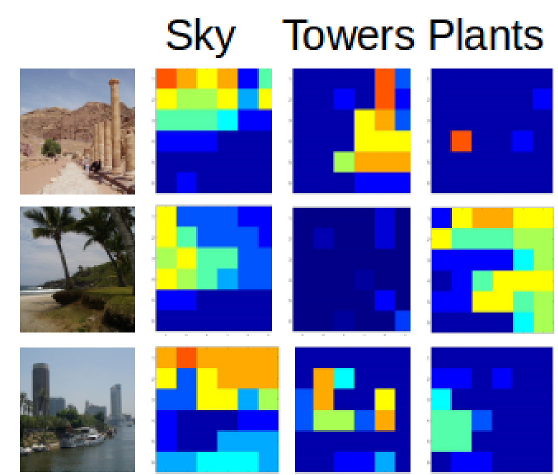
  - How to compress the high dimensional feature vectors without loosing the discriminating capability ?

  - How to incorporate the spatial signature of images into the feature vectors ?

Basic neural network model

- Alexnet model
- Fully connected layer - 4096 dimensional
- Final convolutional layer - 256 different filter responses at a resolution of 6*6
- Each image gives a unique response to the learned filters
- Filter responses carries spatial information about the images





Problems with existing approaches

- Neural codes [1] uses the feature vectors from fully connected layers -
  - **Problem** - Mainly features are high dimensional and lose spatial information
- Hybrid pooling [2] pools the feature activations from convolutional layer (Average and Max pooling)
  - **Problem** - Max pooling ignores all other local maxima in neighborhood
    - Average pooling does not include any spatial cues

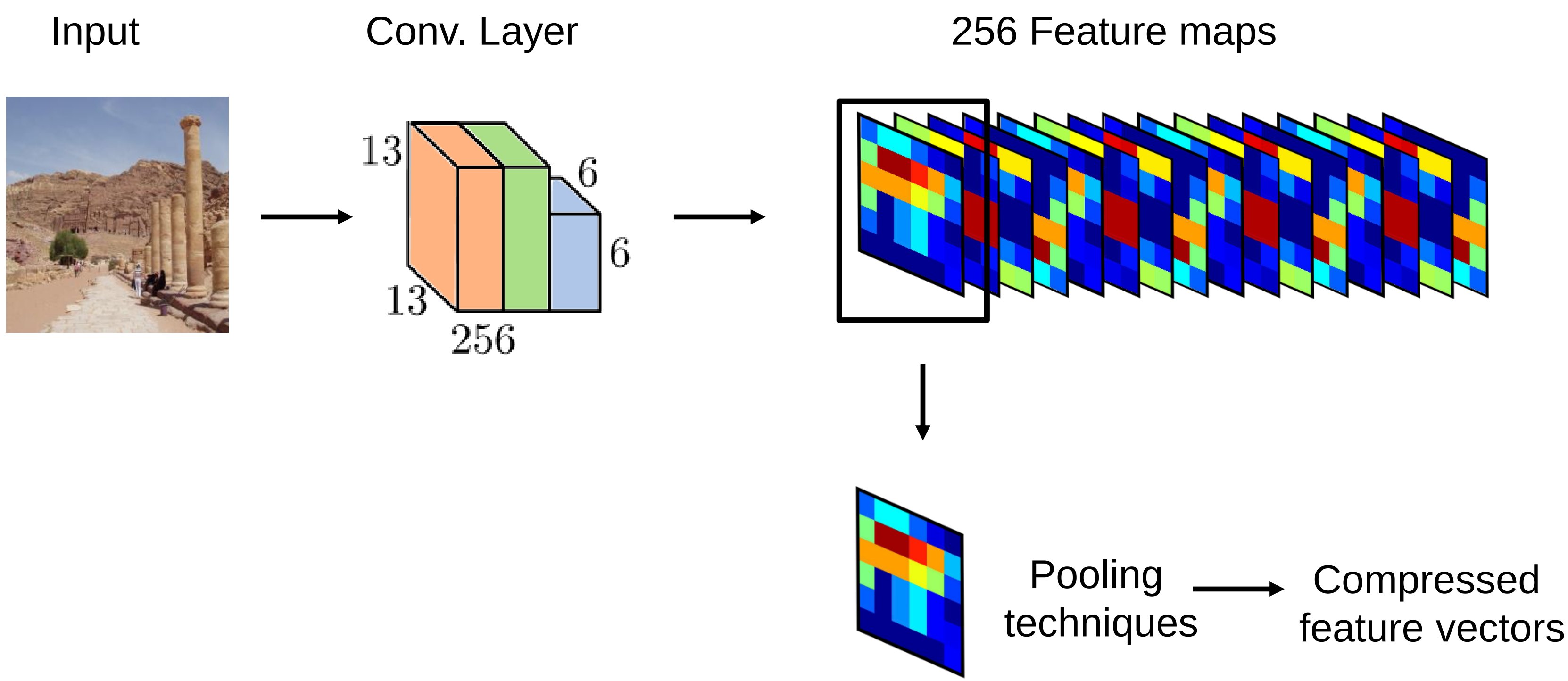
Lehrstuhl und Institut für Nachrichtentechnik
RWTH AACHEN UNIVERSITY

Our approach

- **How to compress high dimensional feature vectors without loosing discriminating capability?**
  - Solved by Max-pooling
  - Robust to scale changes as maximum response of a feature activation will not change with scale
- **How to incorporate the spatial information into feature vectors ?**
  - Addressed by using spatial pyramid pooling method - explained later

Feature extraction and pooling

- Uses Alexnet model
- Extracts the feature activations from final convolutional layer
- We extract the feature vectors from the final 6*6 convolutional feature maps



Feature map and need for spatial pooling

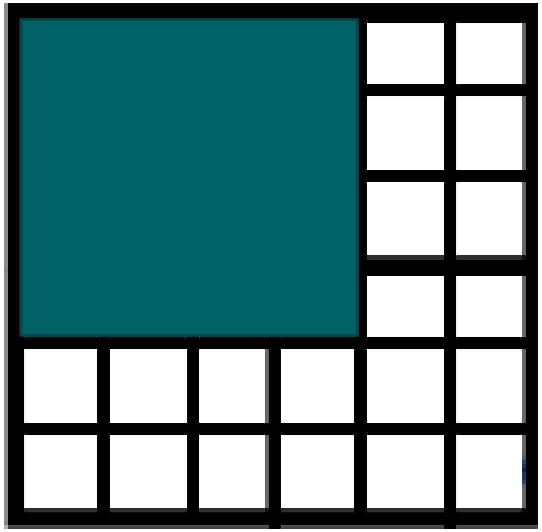
- Max pooling - information about immediate maxima in adjacent bins lost
- Taking a single maximum activation from filter will not form - good descriptor
- Activation map carries spatial signature
- **Solution -** Apply sliding window based pooling
- The remaining 255 filters will have different responses which correspond to different regions such as water, sky, trees and ship

Filter activation which closely represents a tower

Lehrstuhl und Institut für Nachrichtentechnik
RWTH AACHEN UNIVERSITY

Spatial bins proposed for 6*6 activation maps

- Windows are moved in a sliding window manner and max response of the feature map is pooled.
- This captures the strength of feature maps at different spatial positions.
- Final descriptor size is 256*4



Window1 pooling strides

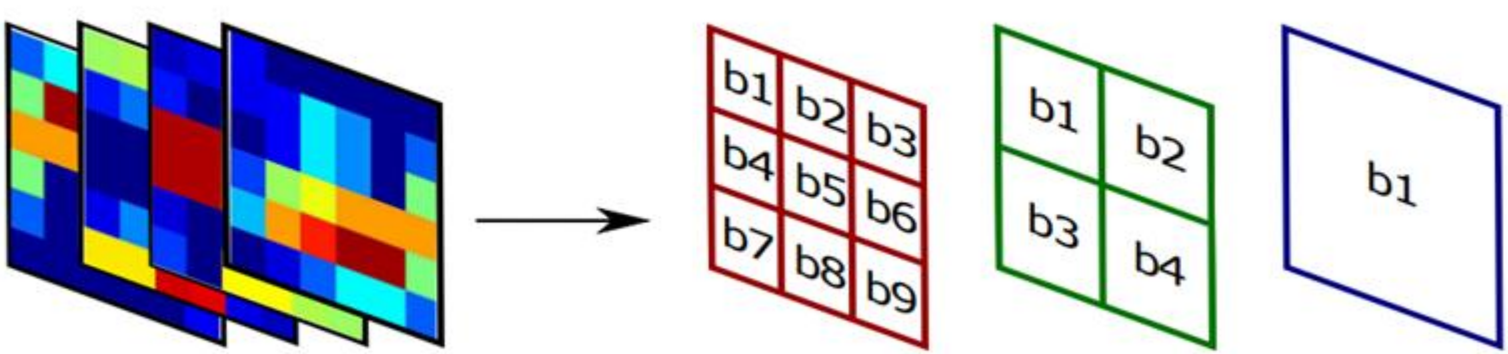
Spatial bins proposed for 6*6 activation maps

- The 6×6 dimensional feature map is divided into different sub regions called bins.
- Table summarizes the different window sizes used for forming bins.
- "Window3" is made of 2 sliding windows calculated independently since the maximum dimension of the feature map is 6.

	HxW size	Stride
Window1	2x2	2
Window2	4x4	2
Window3	3x6 and 6x3	1

Pyramid combinations – Used in our feature extraction pipeline

	Layers
Pyramid 1	Window 1 + Window 3
Pyramid 2	Window 2 + Window 3
Pyramid 3	MAX + Window3
Pyramid 4	MAX + Window 1 + Window 2
Pyramid 5	MAX + Window2 + Window 3



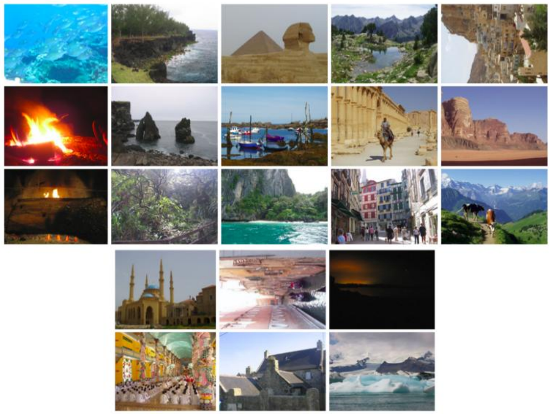
Feature maps

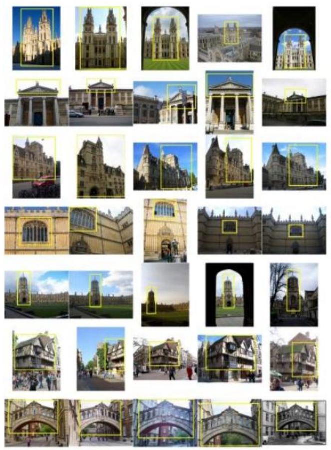
Final descriptor = [b1,...,b9,b1,...b4, b1]

Datasets and experimental settings

- Networks trained on 2 datasets ImageNet and Places used as pre-trained models
- **Oxford5k buildings dataset**
  - Buildings Dataset - 5062 images from Flickr.
  - 11 different landmark images each represented by 5 possible query images
  - Total of 55 different query images

- **INRIA Holidays dataset**
  - Dataset contains 1491 vacation photographs in 500 groups
  - Images taken at same time but with different translation, rotation, and moderate viewpoint changes.
  - First image from each group serves as query





Retrieval results for networks trained on ImageNet dataset

- **Holidays dataset**

  - Pyramid pooling approach - better Mean Average Precision (MAP).
  - MAP has increased in the range of 0.7693 to 0.7732.
  - Dimensions of feature vectors - lower compared to the dimensions of neural codes [1] from layer 5 and from fully connected layers 6 and 7.
  - However, the dimensions of feature vectors from layer 5 is lower for the hybrid pooling approach [2] with slightly lower MAP.

Descriptor	Dimensions	Holidays	Oxford5K
Neural codes layer 5	9216	0.6828	0.3837
Neural codes layer 6	4096	0.7170	0.4004
Neural codes layer 7	4096	0.7162	0.3650
Hybrid pooling	512	0.7634	-
Pyramid 1	3328	**0.7732**	0.4477
Pyramid 2	2048	0.7693	**0.4889**
Pyramid 3	1280	0.7718	0.4471
Pyramid 4	3584	0.7693	0.4422
Pyramid 5	2304	0.7705	0.4461

Lehrstuhl und Institut für Nachrichtentechnik
RWTH AACHEN UNIVERSITY

Retrieval results for networks trained on ImageNet dataset

- **Oxford5K dataset**

  - The MAP values are higher with pyramid pooling approach.
  - The neural codes from layer 5 gives a MAP of 0.3837
  - Pyramid pooling improves the result to an average value of 0.4544

Descriptor	Dimensions	Holidays	Oxford5K
Neural codes layer 5	9216	0.6828	0.3837
Neural codes layer 6	4096	0.7170	0.4004
Neural codes layer 7	4096	0.7162	0.3650
Hybrid pooling	512	0.7634	-
Pyramid 1	3328	**0.7732**	0.4477
Pyramid 2	2048	0.7693	**0.4889**
Pyramid 3	1280	0.7718	0.4471
Pyramid 4	3584	0.7693	0.4422
Pyramid 5	2304	0.7705	0.4461

Retrieval results for networks trained on Places dataset

- **Holidays dataset**
  - Pyramid pooling approach has slightly lower MAP (with an average value = 0.75266)
  - MAP is still better compared to the neural codes from layer 5

Descriptor	Dimensions	Holidays	Oxford5K
Neural codes layer 5	9216	0.6771	0.3717
Neural codes layer 6	4096	0.6914	0.3634
Neural codes layer 7	4096	0.6709	0.3482
Hybrid pooling	512	**0.7924**	-
Pyramid 1	3328	0.7543	0.4228
Pyramid 2	2048	0.7523	**0.4289**
Pyramid 3	1280	0.7514	0.4241
Pyramid 4	3584	0.7539	0.4209
Pyramid 5	2304	0.7514	0.4261

Retrieval results for networks trained on Places dataset

- **Oxford5K dataset**

  - For the Oxford5K dataset, the MAP values are higher than the values obtained using simple pooling layers
  - Retrieval performance here is lower than the values obtained for the network trained with the ImageNet dataset.
  - Reason - Oxford5K dataset is a more object-centric dataset.
  - So the ImageNet pretrained model will give better feature representation.

Descriptor	Dimensions	Holidays	Oxford5K
Neural codes layer 5	9216	0.6771	0.3717
Neural codes layer 6	4096	0.6914	0.3634
Neural codes layer 7	4096	0.6709	0.3482
Hybrid pooling	512	**0.7924**	-
Pyramid 1	3328	0.7543	0.4228
Pyramid 2	2048	0.7523	**0.4289**
Pyramid 3	1280	0.7514	0.4241
Pyramid 4	3584	0.7539	0.4209
Pyramid 5	2304	0.7514	0.4261

Retrieval results for networks trained on Places dataset

- Proposed a novel method for generating the feature vectors from the final convolutional layer by pooling the feature activations from windows of different sizes and strides.
- This spatial pyramid pooling of feature activations helps in capturing the spatial information in the scene.
- This pooling approach reduces the dimension of the feature vectors.
- Our experimental results have shown that this method outperforms state-of-the-art image retrieval methods on 2 standard datasets.

References

[1] Babenko, Artem and Victor, Lempitsky. "Aggregating local deep features for image retrieval." In *Proceedings of the IEEE international conference on computer vision*, pp. 1269-1277, 2015.

[2] Mousavian, Arsalan, and Jana Kosecka. "Deep convolutional features for image based retrieval and scene categorization." *arXiv preprint arXiv:1509.06033* (2015).

Thank you!!