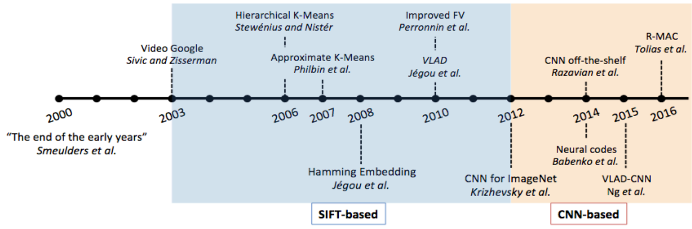
Vision par Ordinateur :
Indexation et recherche d'images

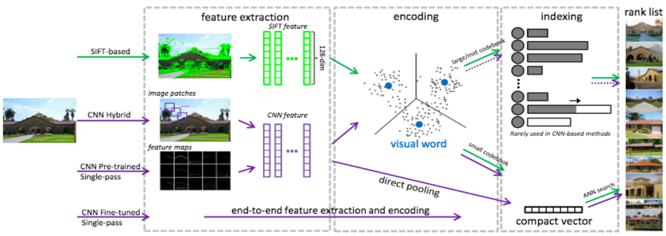
CNN-based descriptors

Milestones of instance retrieval

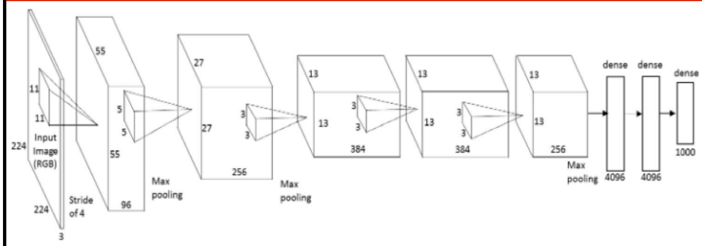


« SIFT Meets CNN: A Decade Survey of Instance Retrieval » Liang Zheng, Yi Yang, and Qi Tian, 2015

General pipeline of SIFT- and CNN-based retrieval models



AlexNet



- 3-channel RGB input, 224 x 224
- 8 layers : 5 conv + 3 fc
- ReLU follows each convolutional and fully connected layer
- Data augmentation, dropout
- Stochastic gradient descent with momentum

Pre-trained CNNs

- AlexNet pre-trained on ImageNet for classification
- last fully connected layer (fc6): global descriptor of dimension k = 4096
- nearest neighbors in ImageNet according to Euclidean distance

Krizhevsky, Sutskever, Hinton. NIPS 2012. Imagenet Classification with Deep Convolutional Neural Networks.

Neural codes for image retrieval

- fine-tuning by softmax on 672 classes of 200k landmark photos
- outperforms VLAD and Fisher vectors on standard retrieval benchmarks, but still inferior to SIFT local descriptors

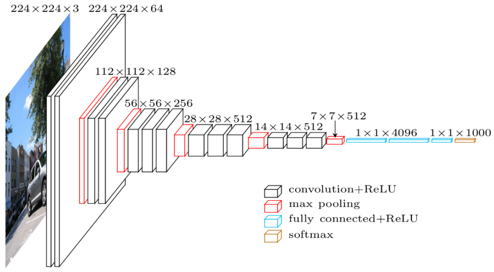
Babenko, Slesarev, Chigorin, Lempitsky. ECCV 2014. Neural codes for Image Retrieval.

CNN features Off-the-shelf

- For each image, extract multiple sub-patches of different sizes at different locations
- For each extracted sub-patch, its CNN representation is the L2 normalized output of the first fully connected layer (dim=4096)
- PCA dimensionality reduction → whitening →L2 renormalization (500-D)

A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in CVPR Workshops, 2014

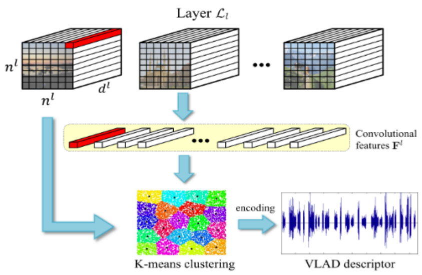
VGG-16



- Depth increased up to 19 layers,
- Kernel sizes reduced to 3, strides to 1

Simonyan and Zisserman 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition.

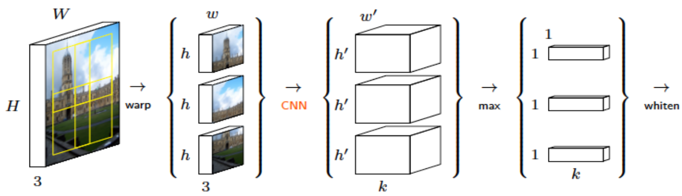
VLAD-CNN



- Consider different layers of VGG-16
- For each layer, VLAD encoding (k=100)
- L2-normalization, PCS-whitening (128-D)

J. Ng, F. Yang, and L. Davis, "Exploiting local features from deep networks for image retrieval," CVPR Workshops, 2015.

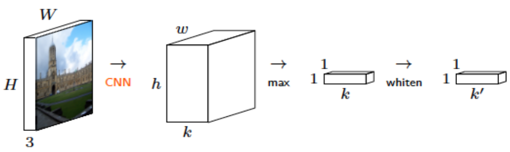
Regional CNN features



- 3-channel RGB input, largest square region extracted
- fixed multiscale overlapping regions, warped into w x h = 227 x 227
- each region yields a w' x h' x k = 36 x 36 x 256 dimensional feature at the last convolutional layer of AlexNet
- global spatial max-pooling
- L2-normalization, PCA-whitening of each descriptor

Razavian, Sullivan, Maki and Carlsson 2015. Visual Instance Retrieval with Deep Convolutional Networks.

Global max-pooling (MAC)



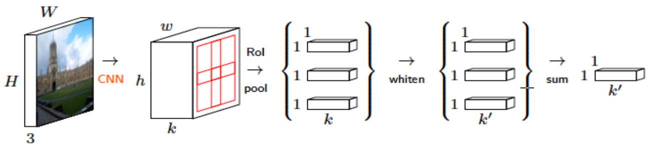
- VGG-16 last convolutional layer, k = 512
- global spatial max-pooling
- L2-normalization, PCA-whitening, L2-normalization
- MAC: maximum activation of convolutions

Tolias, Sicre and Jegou. ICLR 2016. Particular Object Retrieval with Integral Max-Pooling of CNN Activations.

Global max-pooling (MAC)

- receptive fields of 5 components of MAC vectors that contribute most to image similarity

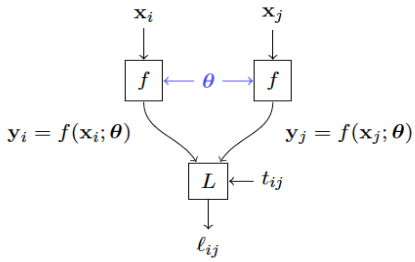
Regional max-pooling (R-MAC)



- VGG-16 last convolutional layer, k = 512
- fixed multiscale overlapping regions, spatial max-pooling
- L2-normalization, PCA-whitening, L2-normalization
- sum-pooling over all descriptors, L2-normalization

Tolias, Sicre and Jegou. ICLR 2016. Particular Object Retrieval with Integral Max-Pooling of CNN Activations.

Siamese architecture

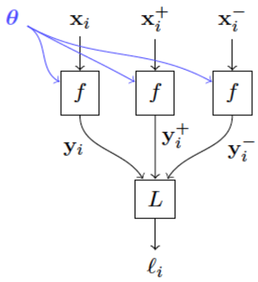


- an input sample is a pair (x_i, x_j)
- both x_i, x_j go through the same function f with shared parameters θ
- Contrastive loss l_{ij} is measured on output pair (y_i, y_j) and target t_{ij}

Chopra, Hadsell, Lecun, CVPR 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification.

Triplet architecture

- an input sample is a triplet (x_i, x^+_i, x^-_i)
- x_i, x^+_i, x^-_i go through the same function f with shared parameters
- loss l_i measured on output triplet (y_i, y^+_i, y^-_i)



Wang, Song, Leung, Rosenberg, Wang, Philbin, Chen, Wu. CVPR 2014. Learning Fine-Grained Image Similarity with Deep Ranking.