#DLUPC

Day 1 Lecture 4

Content-based Image Retrieval

Eva Mohedano
eva.mohedano@insight-centre.org

Postdoctoral Researcher
Insight-centre for Data Analytics
Dublin City University

DCU

Insight
Centre for Data Analytics

http://bit.ly/dlcv2018

Overview

- What is content-based image retrieval?
- The classic SIFT retrieval pipeline
- Using off the shelf CNN features for retrieval
- Learning representations for retrieval

Overview

- **What is content-based image retrieval?**
- The classic SIFT retrieval pipeline
- Using off the shelf CNN features for retrieval
- Learning representations for retrieval

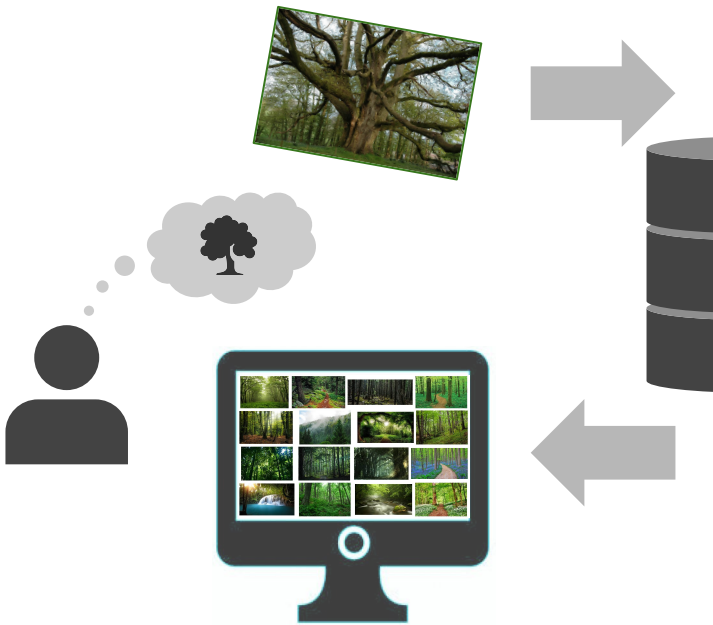
The problem: query by example

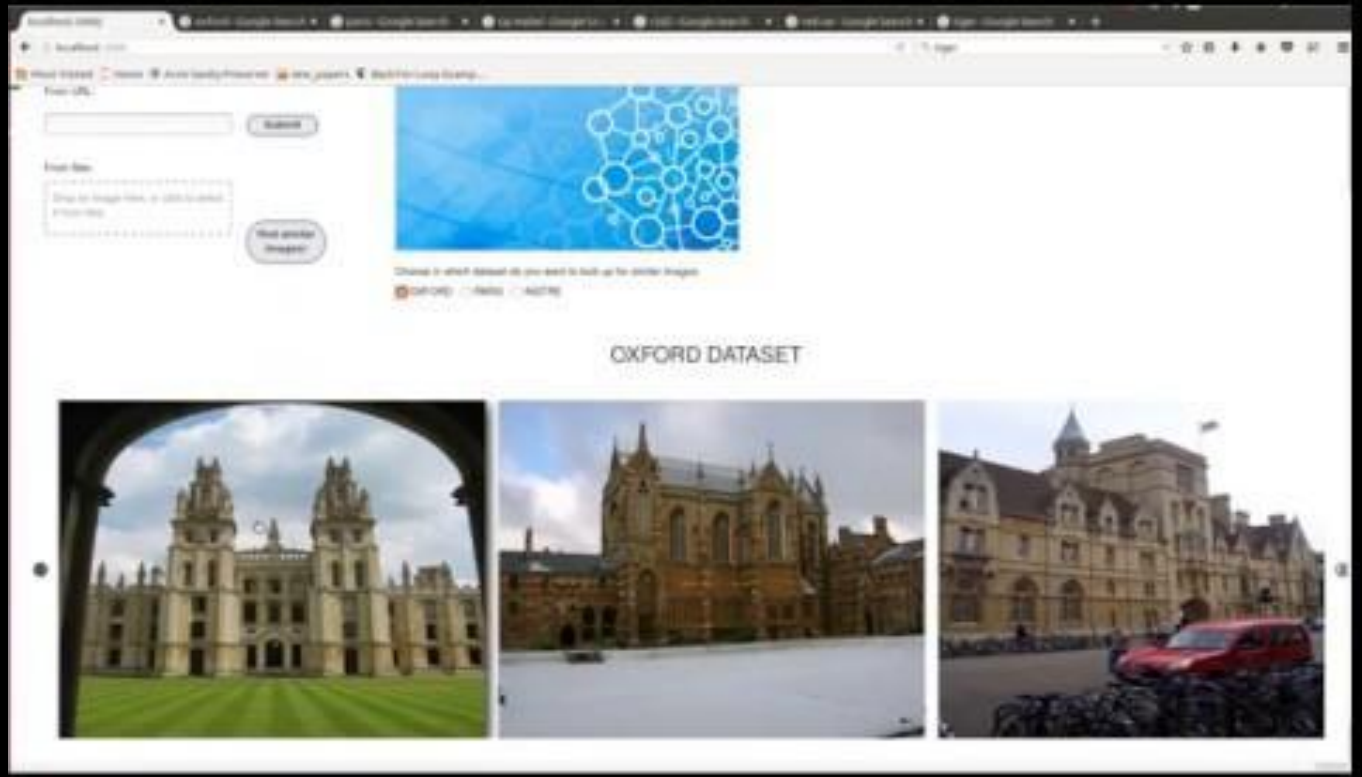
Given:

- An example query image that illustrates the user's information need
- A very large dataset of images

Task:

- Rank all images in the dataset according to how likely they are to fulfil the user's information need



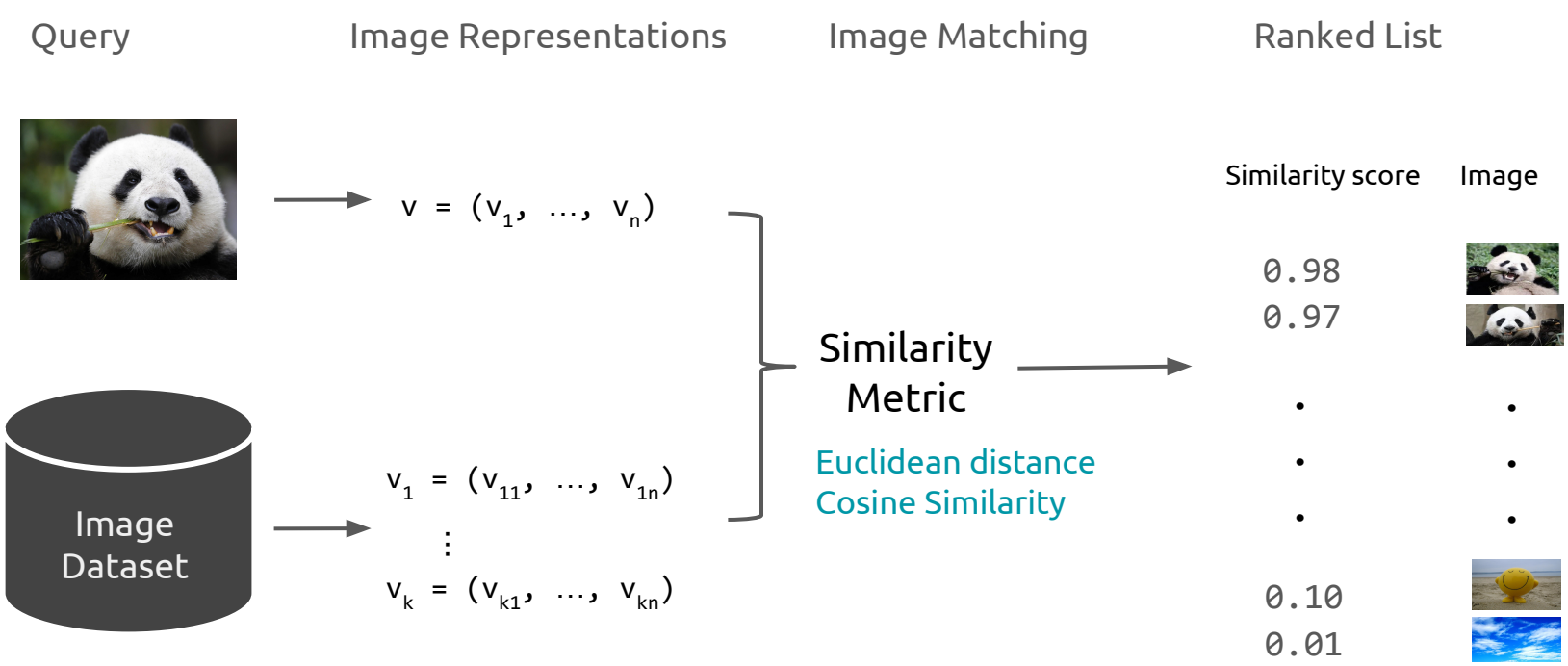


Demo by Paula Gomez Duran
Dockerized visualization tool

Overview

- What is content-based image retrieval?
- **The classic SIFT retrieval pipeline**
- Using off the shelf CNN features for retrieval
- Learning representations for retrieval

The retrieval pipeline



The classic SIFT retrieval pipeline

Initial Search

variable number of
feature vectors per image

$v_1 = (v_{11}, \ldots, v_{1n})$

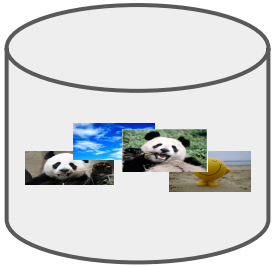
$\vdots$

$v_k = (v_{k1}, \ldots, v_{kn})$

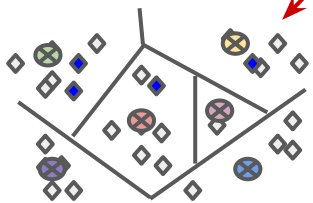
Bag of Visual Words

INVERTED FILE

word	Image ID
1	1, 12,
2	1, 30, 102
3	10, 12
4	2,3
6	10



N-Dimensional
feature space



M visual words
(M clusters)

Large vocabularies (50k-1M)
Very fast!
Typically used with SIFT features

The classic SIFT retrieval pipeline

Spatial re-ranking

Re-ranking the **top-ranked** results using **spatial constraints**

RAndom SAmple Consensus (RANSAC)

- Estimates an homography between the query and a dataset image
- Re-rank based on number of inlier local features
- Improves quality of the initial search

$$x_i' = Hx_i$$

Expensive to compute

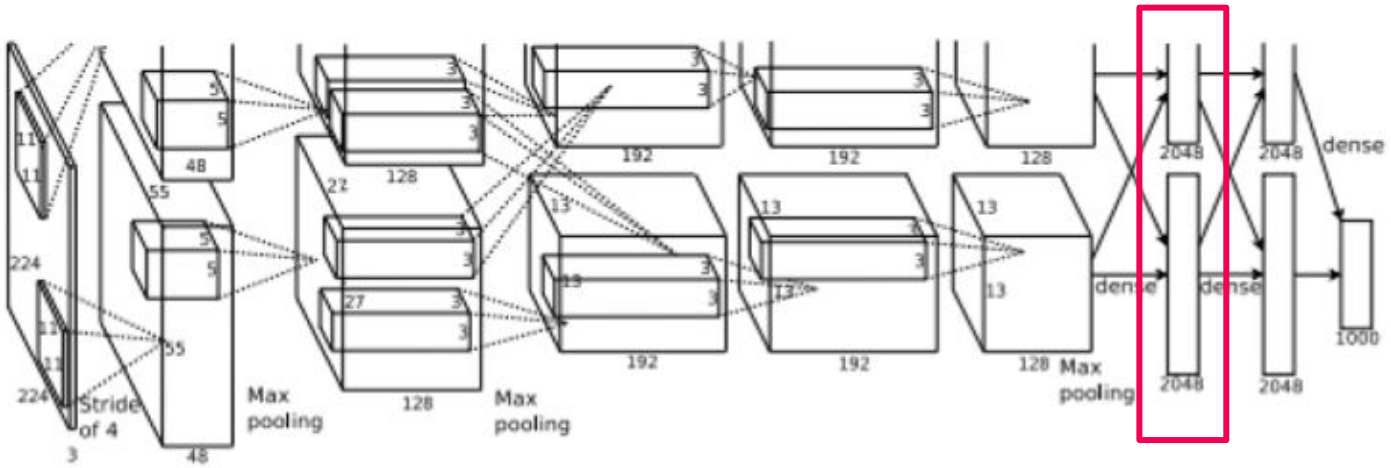
Philbin, J., Chum, O., Isard, M., Sivic, J. and Zisserman, Object retrieval with large vocabularies and fast spatial matching. *CVPR 2007*
Kalantidis, Y., Tolias, G., Spyrou, E., Mylonas, P. and Avrithis, Visual image retrieval and localization. *7th International Workshop on Content-Based Multimedia Indexing, 2009.*

Overview

- What is content-based image retrieval?
- The classic SIFT retrieval pipeline
- **Using off the shelf CNN features for retrieval**
- Learning representations for retrieval

Off-the-shelf CNN representations

FC layers as global feature representation



Razavian et al. CNN features off-the-shelf: an astounding baseline for recognition. CVPR 2014
Babenko et al. Neural codes for image retrieval. ECCV 2014

Off-the-shelf CNN representations

Neural codes for retrieval [1]

- FC7 layer (4096D)
- L^2 norm + PCA whitening + L^2 norm
- Euclidean distance
- Only better than traditional SIFT approach after fine tuning on similar domain image dataset.

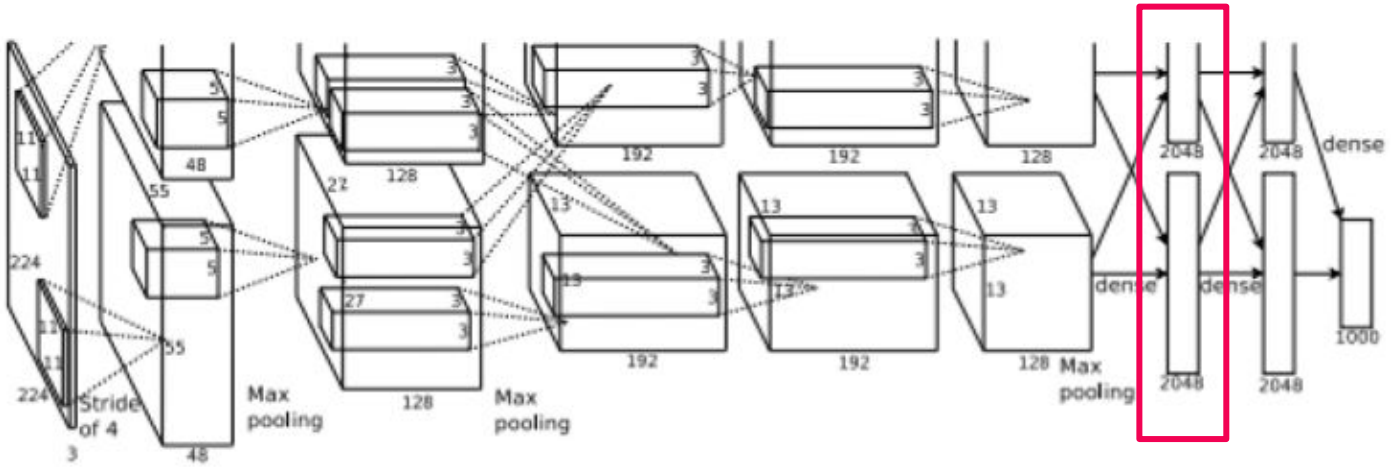
CNN features off-the-shelf: an astounding baseline for recognition [2]

- Extending Babenko's approach with spatial search
- Several features extracted by image (sliding window approach)
- Really good results but too computationally expensive for practical situations

[1] Babenko et al, Neural codes for image retrieval CVPR 2014
[2] Razavian et al, CNN features off-the-shelf: an astounding baseline for recognition, CVPR 2014

Off-the-shelf CNN representations

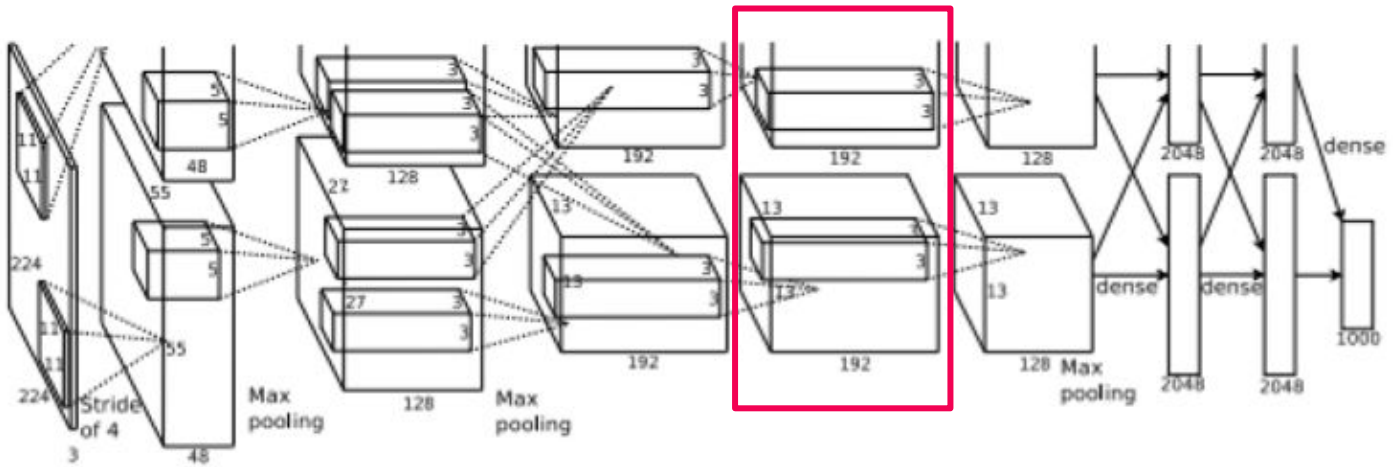
FC layers as global feature representation



Is there any other way to pool the local information from a convolutional layer?

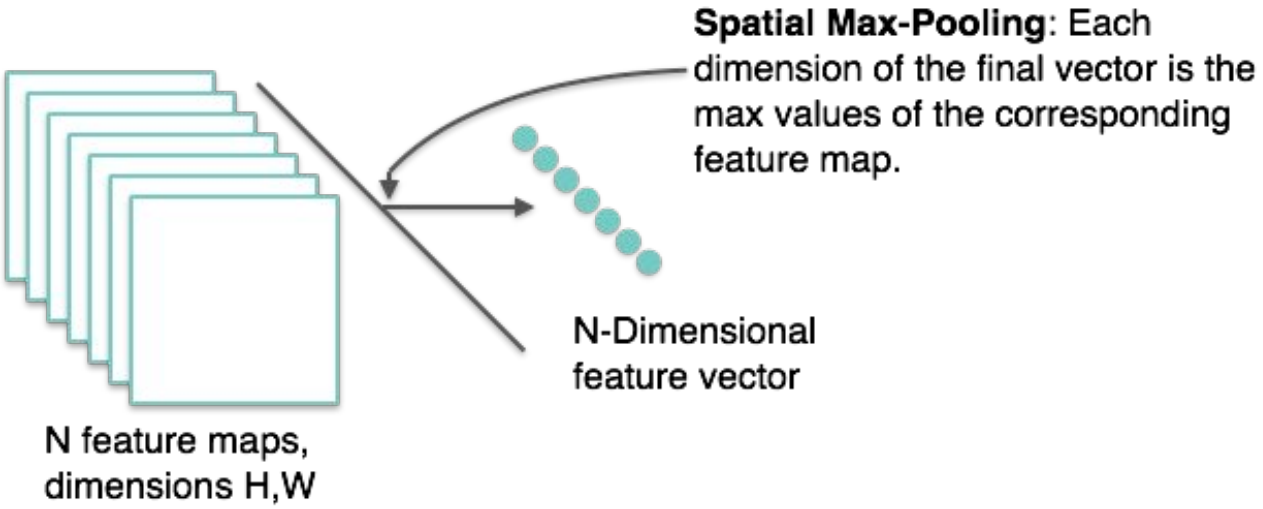
Pooling method #1: Spatial sum/max pooling

sum/max pool conv features across filters



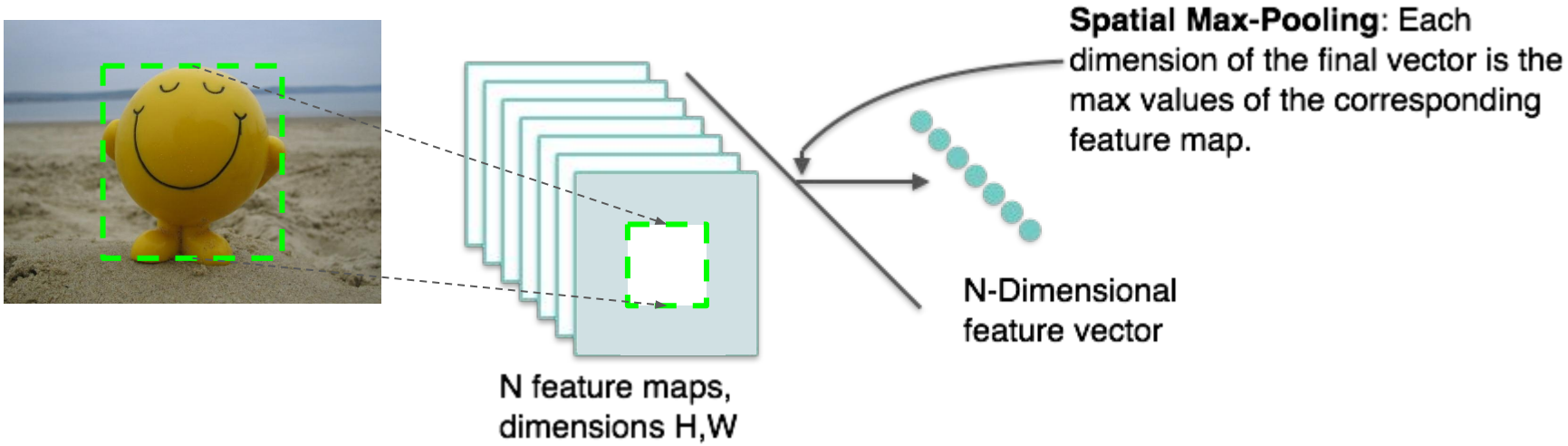
Babenko and Lempitsky, Aggregating local deep features for image retrieval. ICCV 2015
Tolias et al. Particular object retrieval with integral max-pooling of CNN activations. *arXiv 2015.*
Kalantidis et al. Cross-dimensional Weighting for Aggregated Deep Convolutional Features. ECCV 2016 Workshops.

Pooling method #1: Spatial sum/max pooling



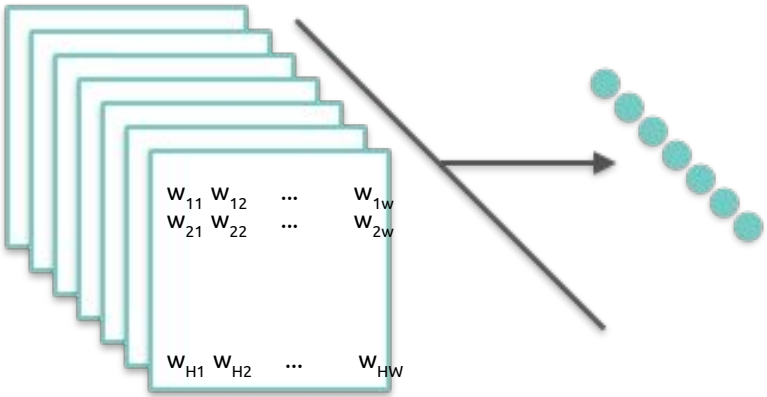
Pooling method #1: Spatial sum/max pooling

Sum/max pooling over specific regions



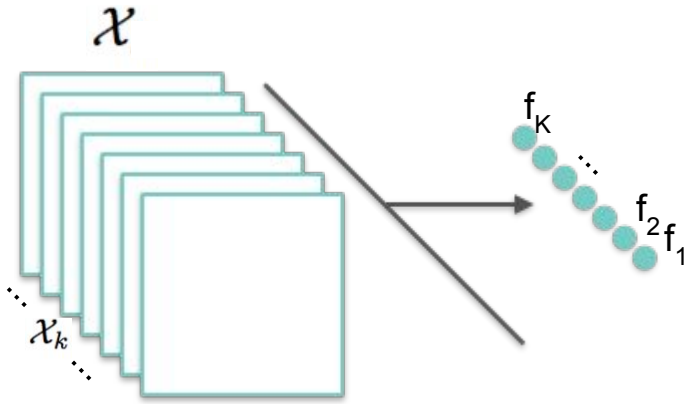
Pooling method #1: Spatial sum/max pooling

Spatial weighting on convolutional activations



Babenko and Lempitsky, Aggregating local deep features for image retrieval. ICCV 2015
Kalantidis et al. Cross-dimensional Weighting for Aggregated Deep Convolutional Features. ECCV 2016 Workshops.
Mohedano et al. Saliency Weighted Convolutional Features for Instance Search, arXiv 2017

Pooling method #2: Generalized-mean pooling



Max-pooling (MAC)

$$\mathbf{f}^{(m)} = [\mathrm{f}_1^{(m)} \dots \mathrm{f}_k^{(m)} \dots \mathrm{f}_K^{(m)}]^\top, \qquad \mathrm{f}_k^{(m)} = \max_{x \in \mathcal{X}_k} x,$$

Average pooling (SPoC)

$$\mathbf{f}^{(a)} = [\mathrm{f}_1^{(a)} \dots \mathrm{f}_k^{(a)} \dots \mathrm{f}_K^{(a)}]^\top, \qquad \mathrm{f}_k^{(a)} = \frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x$$

Generalized-mean pooling (GeM)

$$\mathbf{f}^{(g)} = [\mathrm{f}_1^{(g)} \dots \mathrm{f}_k^{(g)} \dots \mathrm{f}_K^{(g)}]^\top, \quad \mathrm{f}_k^{(g)} = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^{p_k} \right)^{\frac{1}{p_k}}$$

→ p_k can be manually set or learnt

Radenović, F., Tolias, G. and Chum, O, Fine-tuning CNN image retrieval with no human annotation. TPAMI 2018

Pooling method #2: Generalized-mean pooling

$p = 1$ $p = 3$ $p = 10$

Pooling	Initial p	Learned p	Oxford5k	Oxford105k	Paris6k	Paris106k	Holidays	Hol101k
MAC	inf	–	62.2	52.8	68.9	54.7	78.4	66.0
SPoC	1	–	61.2	54.9	70.8	58.0	79.9	70.6
GeM	3	–	**67.9**	60.2	74.8	61.7	83.2	73.3
	[2, 5]	–	66.8	59.7	74.1	60.8	**84.0**	73.6
	[2, 10]	–	65.6	57.8	72.2	58.9	81.9	71.9
	3	2.32	67.7	**60.6**	**75.5**	**62.6**	83.7	**73.7**
	3	[1.0, 6.5]	66.3	57.8	74.0	60.5	83.2	72.7
	[2, 10]	[1.6, 9.9]	65.3	56.4	71.4	58.6	81.4	70.8

Radenović, F., Tolias, G. and Chum, O, Fine-tuning CNN image retrieval with no human annotation. TPAMI 2018

Pooling method #3: Region pooling

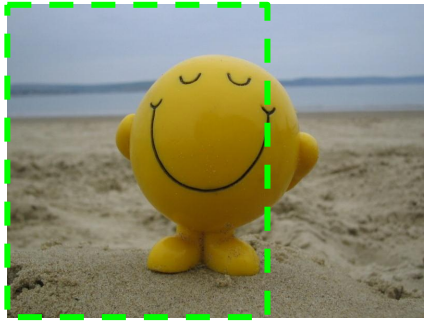
Regional Maximum Activation of Convolution (R-MAC)

Settings

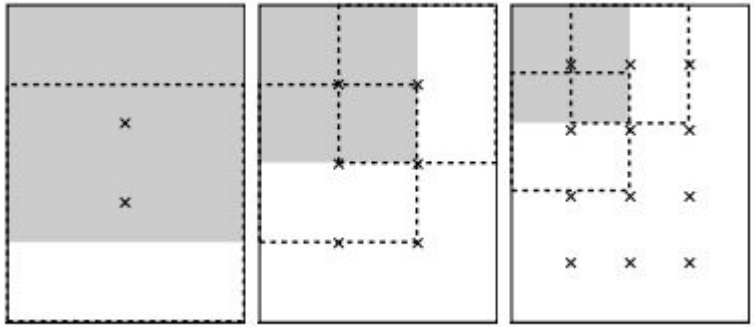
- Fully convolutional off-the-shelf VGG16
- Pool5
- Spatial Max pooling
- High Resolution images
- Global descriptor based on aggregating region vectors
- Sliding window approach

Tolias et al. Particular object retrieval with integral max-pooling of CNN activations. *arXiv 2015.*

R-MAC

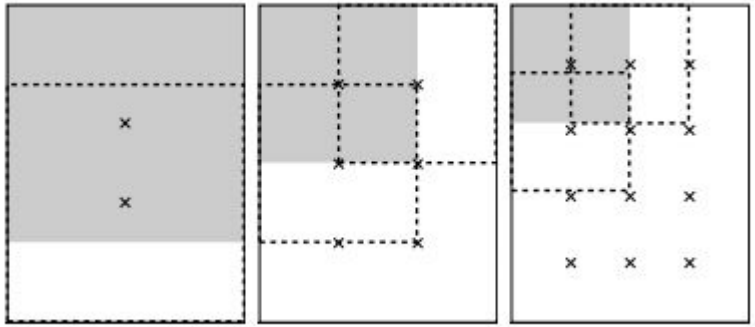


Region1

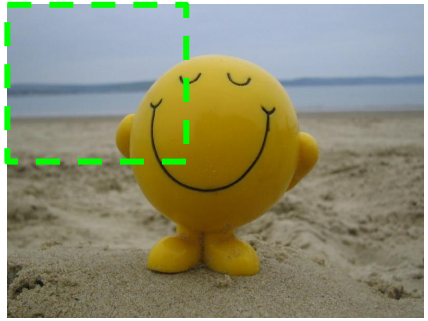


R-MAC

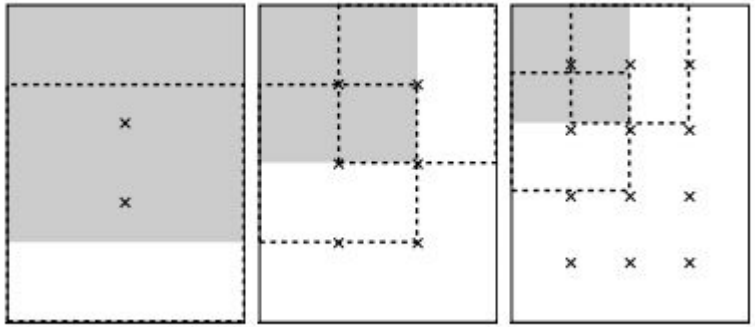
Region1
Region2



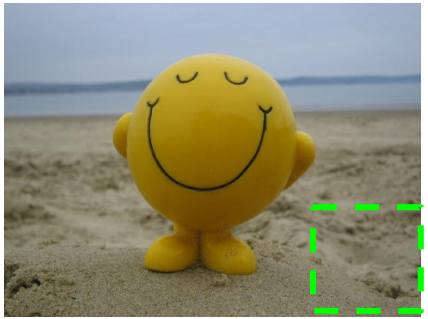
R-MAC



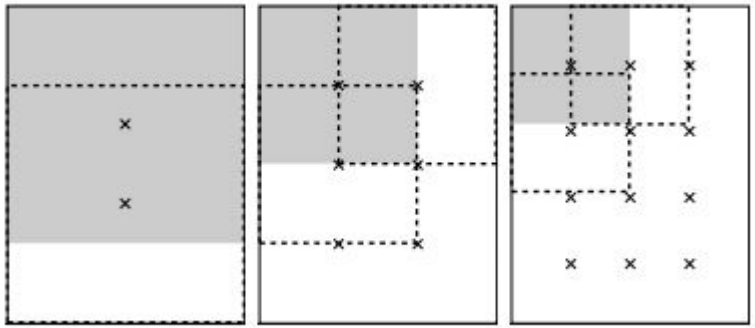
Region1
Region2
Region3



R-MAC



Region1
Region2
Region3
Region4
...
RegionN



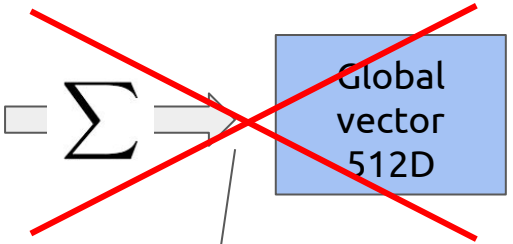
R-MAC

Region vectors 512D

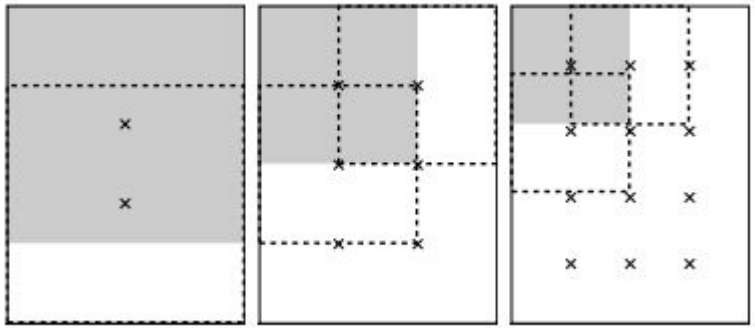
Region1
Region2
Region3
Region4
...
RegionN

l2-PCAw-l2

Region1
Region2
Region3
Region4
...
RegionN

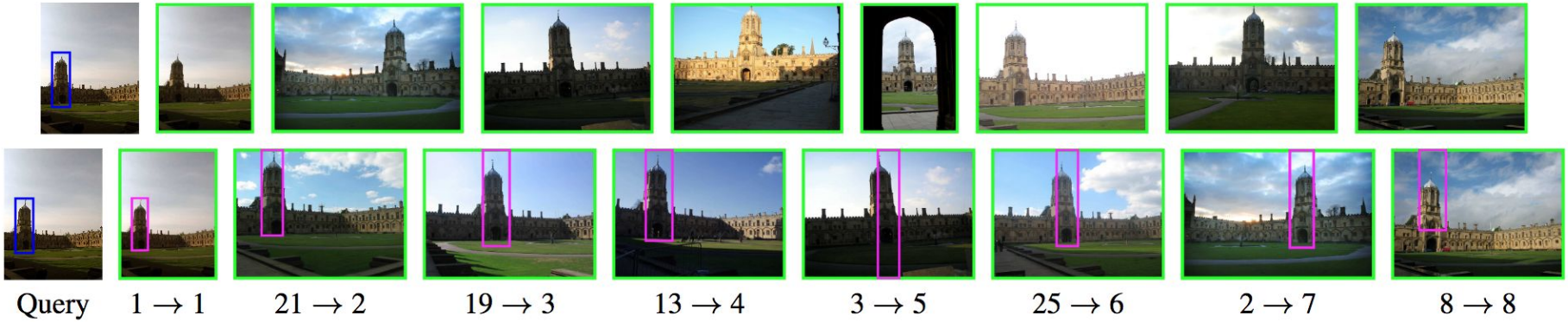


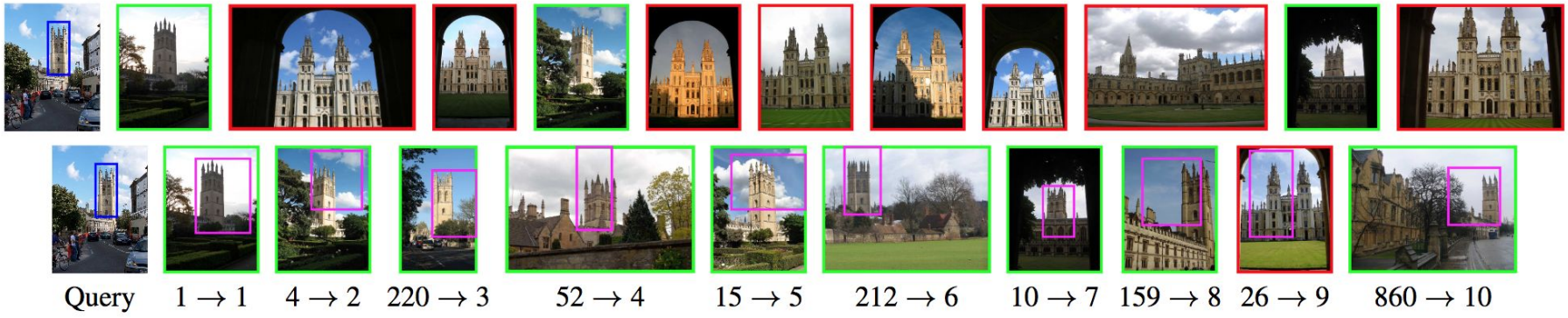
Spatial re-ranking



- Only on top 100 ranked results, compare query with dataset regions
- Re-rank images based on best region score per image

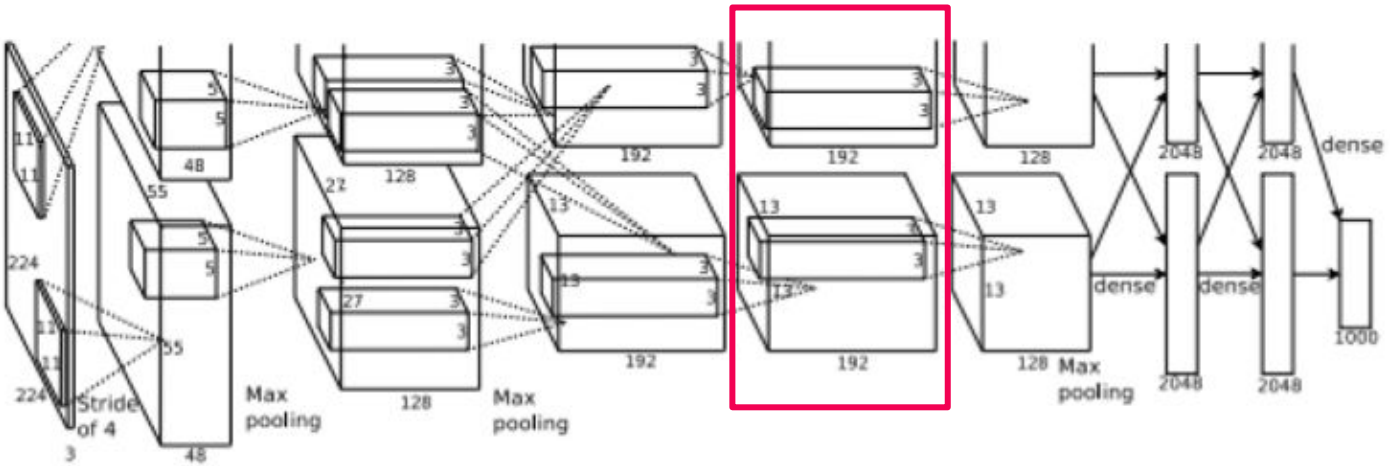
R-MAC


Query
1 → 1
21 → 2
19 → 3
13 → 4
3 → 5
25 → 6
2 → 7
8 → 8


Query
1 → 1
4 → 2
220 → 3
52 → 4
15 → 5
212 → 6
10 → 7
159 → 8
26 → 9
860 → 10

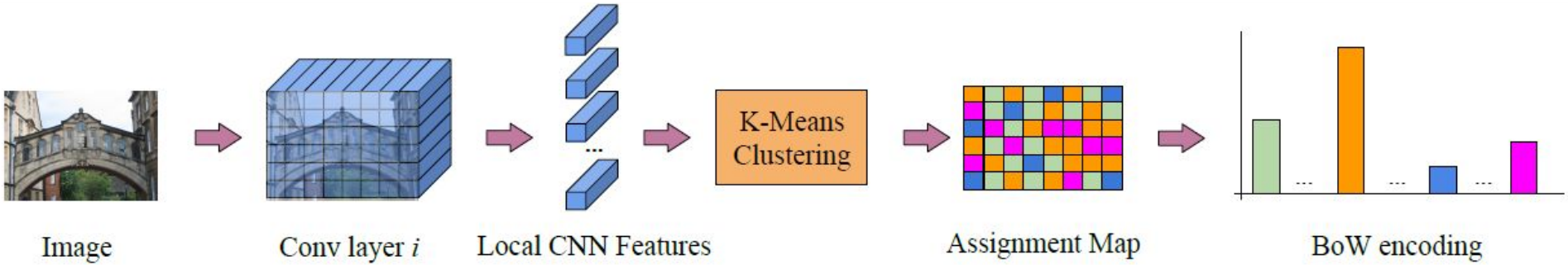
Pooling method #4: Traditional pooling techniques

BoW, VLAD encoding of conv features



Ng et al. Exploiting local features from deep networks for image retrieval. CVPR Workshops 2015
Mohedano et al. Bags of Local Convolutional Features for Scalable Instance Search. ICMR 2016

Pooling method #4: Bags of Local Convolutional Features



(336x256) Resolution

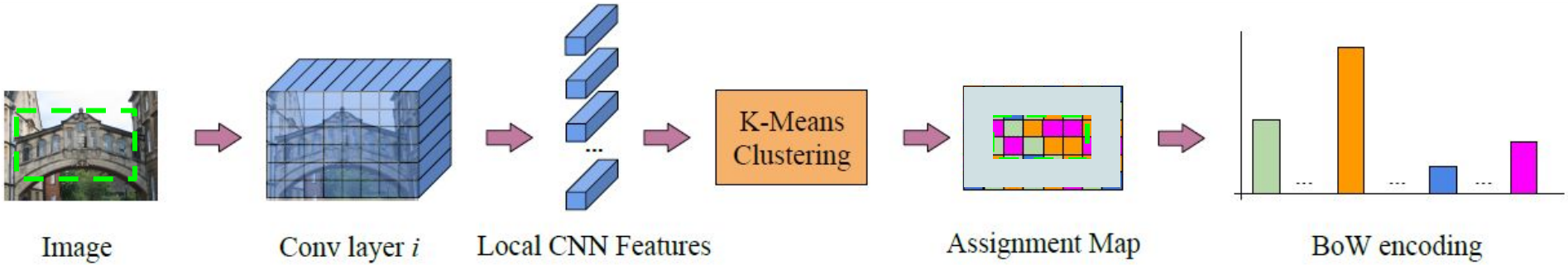
conv5_1 from VGG16 (42x32)

25K centroids

25K-D vector

Mohedano et al. Bags of Local Convolutional Features for Scalable Instance Search. ICMR 2016

Pooling method #4: Bags of Local Convolutional Features



(336x256) Resolution

conv5_1 from VGG16 (42x32)

25K centroids

25K-D vector

Pooling method #4: Bags of Local Convolutional Features

Paris Buildings 6k

Oxford Buildings 5k

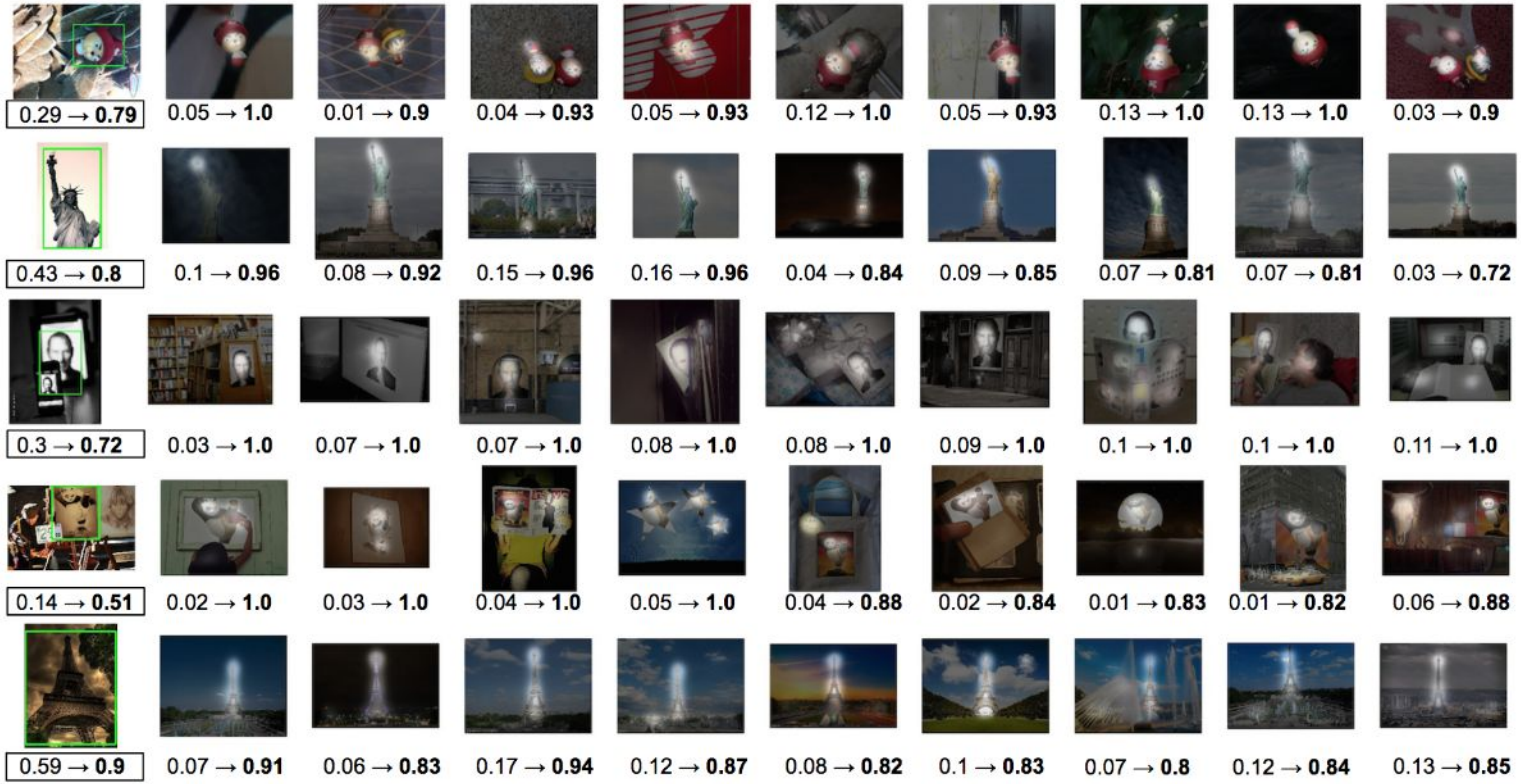
TRECVID Instance Search 2013
(subset of 23k frames)

		Oxford 5k	Paris 6k	INS 23k
BoW	GS	0.650	0.698	**0.323**
	LS	**0.739**	**0.819**	0.295
Sum pooling (as ours)	GS	0.606	0.712	0.156
	LS	0.583	0.742	0.097
Sum pooling (as in [7])	GS	0.672	0.774	0.139
	LS	0.683	0.763	0.120

[7] Kalantidis et al. Cross-dimensional Weighting for Aggregated Deep Convolutional Features. ECCV 2016 Workshops.

Mohedano et al. Bags of Local Convolutional Features for Scalable Instance Search. ICMR 2016

Pooling method #4: Bags of Local Convolutional Features



Mohedano et al. Saliency Weighted Convolutional Features for Instance Search, arXiv 2017

Overview

- What is content-based image retrieval?
- The classic SIFT retrieval pipeline
- Using off the shelf CNN features for retrieval
- **Learning representations for retrieval**

Classification

Query: This chair

Results from dataset classified as “chair”

Retrieval

Query: This chair

Results from dataset ranked by similarity to the query

What loss function should we use?

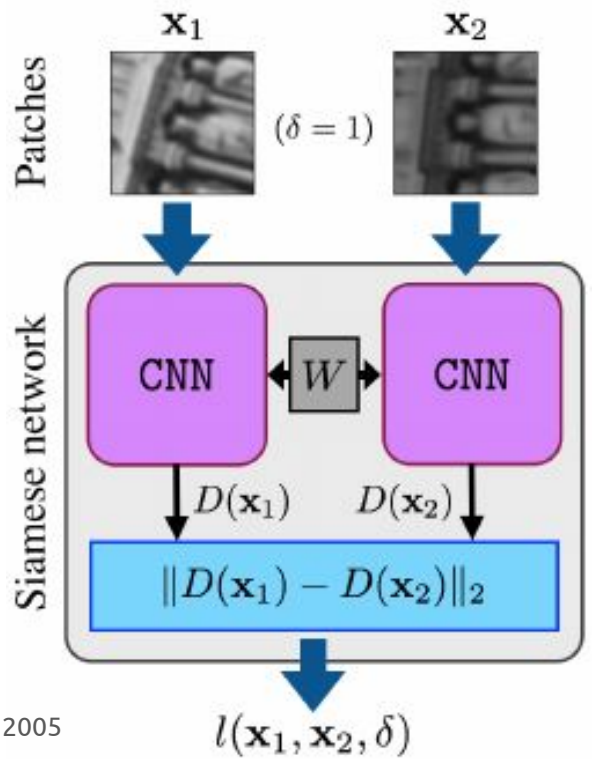
Learning representations for retrieval

Siamese network: network to learn a function that maps input patterns into a target space such that L^2 norm in the target space approximates the semantic distance in the input space.

$$l(\mathbf{x}_1, \mathbf{x}_2, \delta) = \delta \cdot l_P(d_D(\mathbf{x}_1, \mathbf{x}_2)) + (1-\delta) \cdot l_N(d_D(\mathbf{x}_1, \mathbf{x}_2))$$

$$l_P(d_D(\mathbf{x}_1, \mathbf{x}_2)) = d_D(\mathbf{x}_1, \mathbf{x}_2)$$

$$l_N(d_D(\mathbf{x}_1, \mathbf{x}_2)) = \max(0, m - d_D(\mathbf{x}_1, \mathbf{x}_2))$$

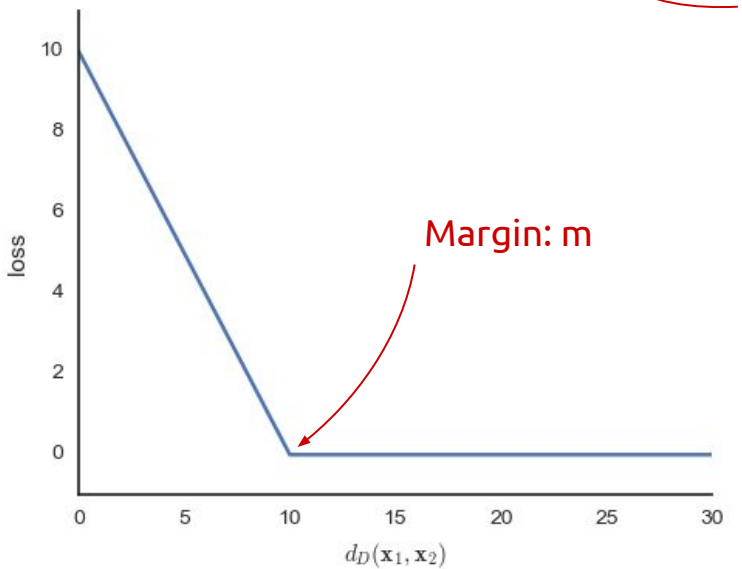


Song et al.: Deep metric learning via lifted structured feature embedding. CVPR 2015
Chopra et al. Learning a similarity metric discriminatively, with application to face verification . CVPR 2005
Simo-Serra et al. Discriminative Learning of Deep Convolutional Feature Point Descriptor . ICCV 2015

$$l_N(d_D(\mathbf{x}_1, \mathbf{x}_2)) = \max(0, m - d_D(\mathbf{x}_1, \mathbf{x}_2))$$

Hinge loss

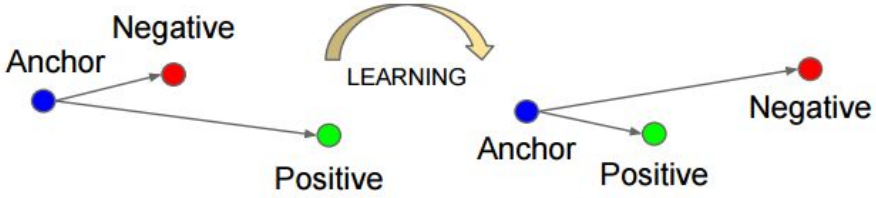
Negative pairs: if nearer than the margin, pay a linear penalty



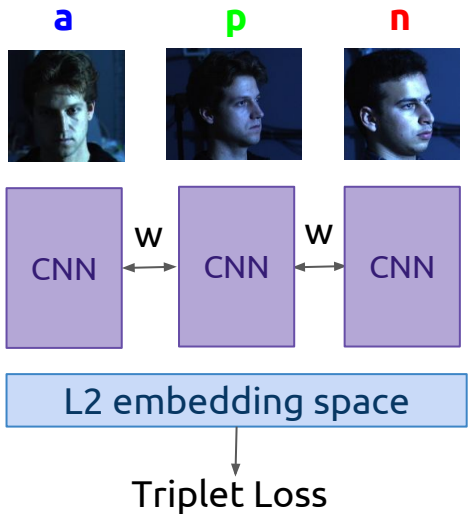
The margin is a **hyperparameter**

Learning representations for retrieval

Siamese network with triplet loss: loss function minimizes distance between query and positive and maximizes distance between query and negative



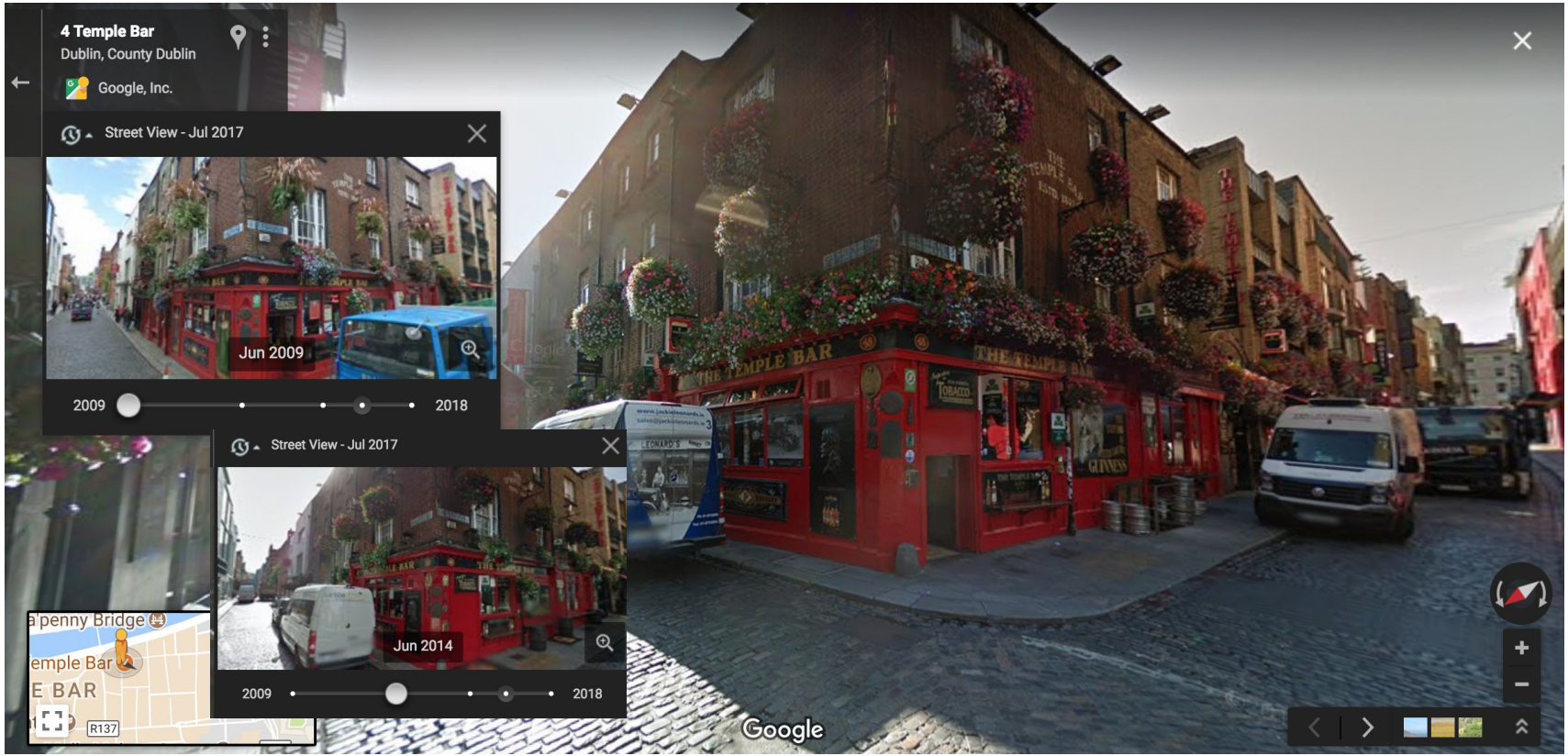
$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$$



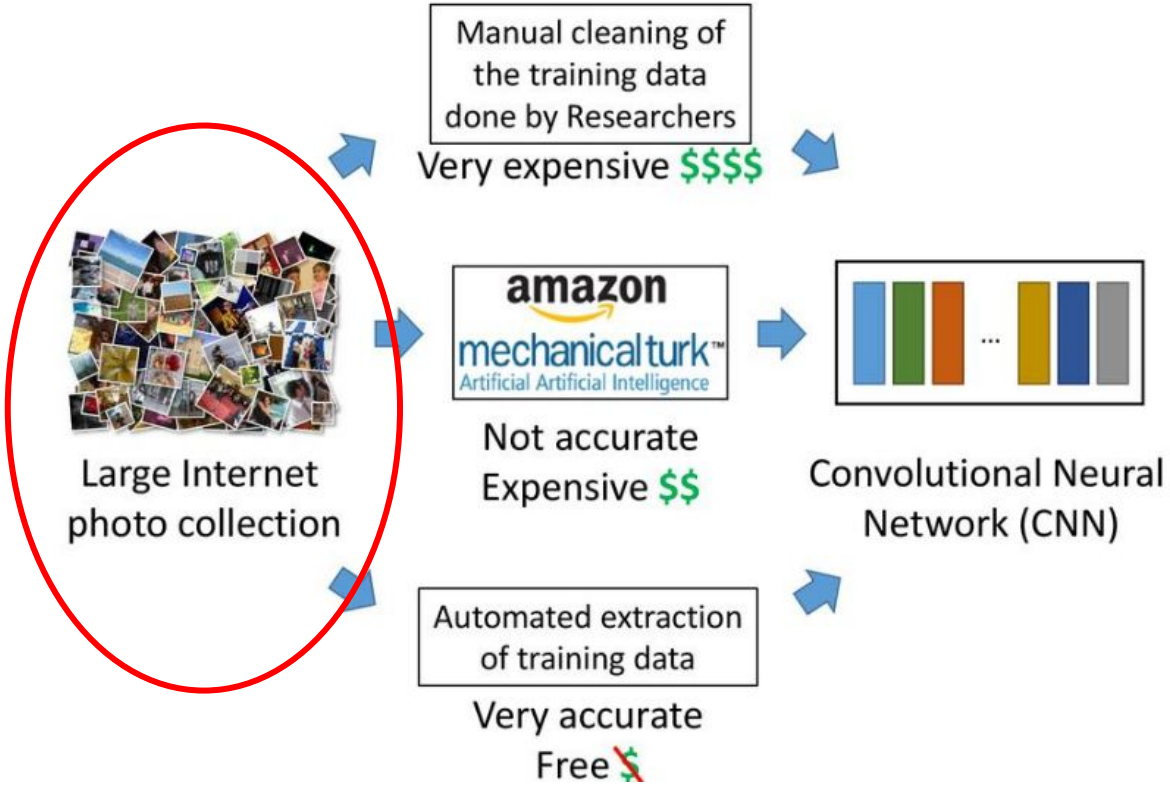
Schroff et al. FaceNet: A Unified Embedding for Face Recognition and Clustering, CVPR 2015

How do we get the training data?

Exploring Image datasets with GPS coordinates (Google Street View Time Machine)



Relja Arandjelović et al, NetVLAD: CNN architecture for weakly supervised place recognition, CVPR 2016



Automatic data cleaning

Strong baseline (SIFT) + geometric verification + 3D camera estimation[1]

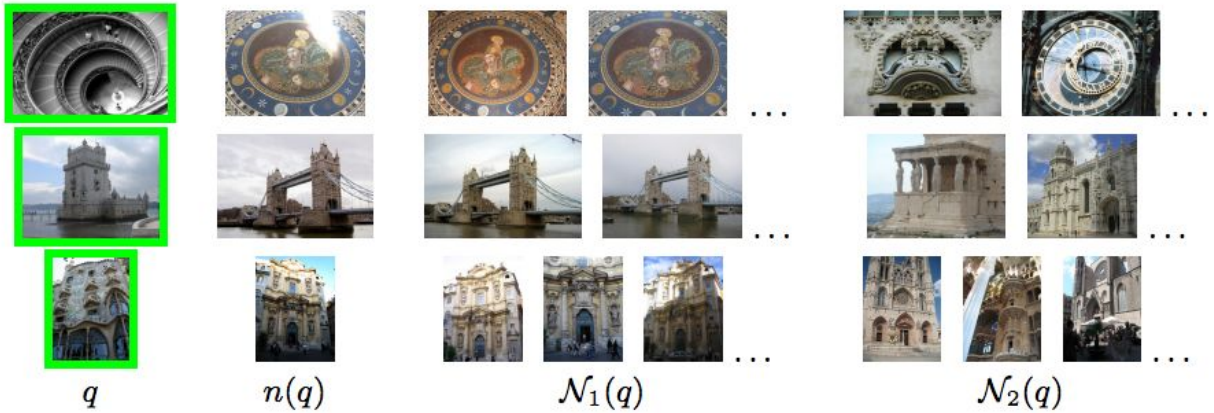
Further manual inspection

Further manual inspection

Radenović el at., CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples, CVPR 2016
Gordo et al., Deep Image Retrieval: Learning global representations for image search, CVPR 2016
[1] Schonberger et al. From single image query to detailed 3D reconstruction, CVPR 2015

Generating training pairs

Once we have collected a training dataset (with annotations), the way to select the pairs/triplets is crucial



Select the hardest negative

- Select triplet which generates larger loss
- If 3D-geometric verification model, select images based on number of inliers

Radenović et al, CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples, CVPR 2016
Gordo et al., Deep Image Retrieval: Learning global representations for image search, CVPR 2016

End-to-End learning for retrieval #1

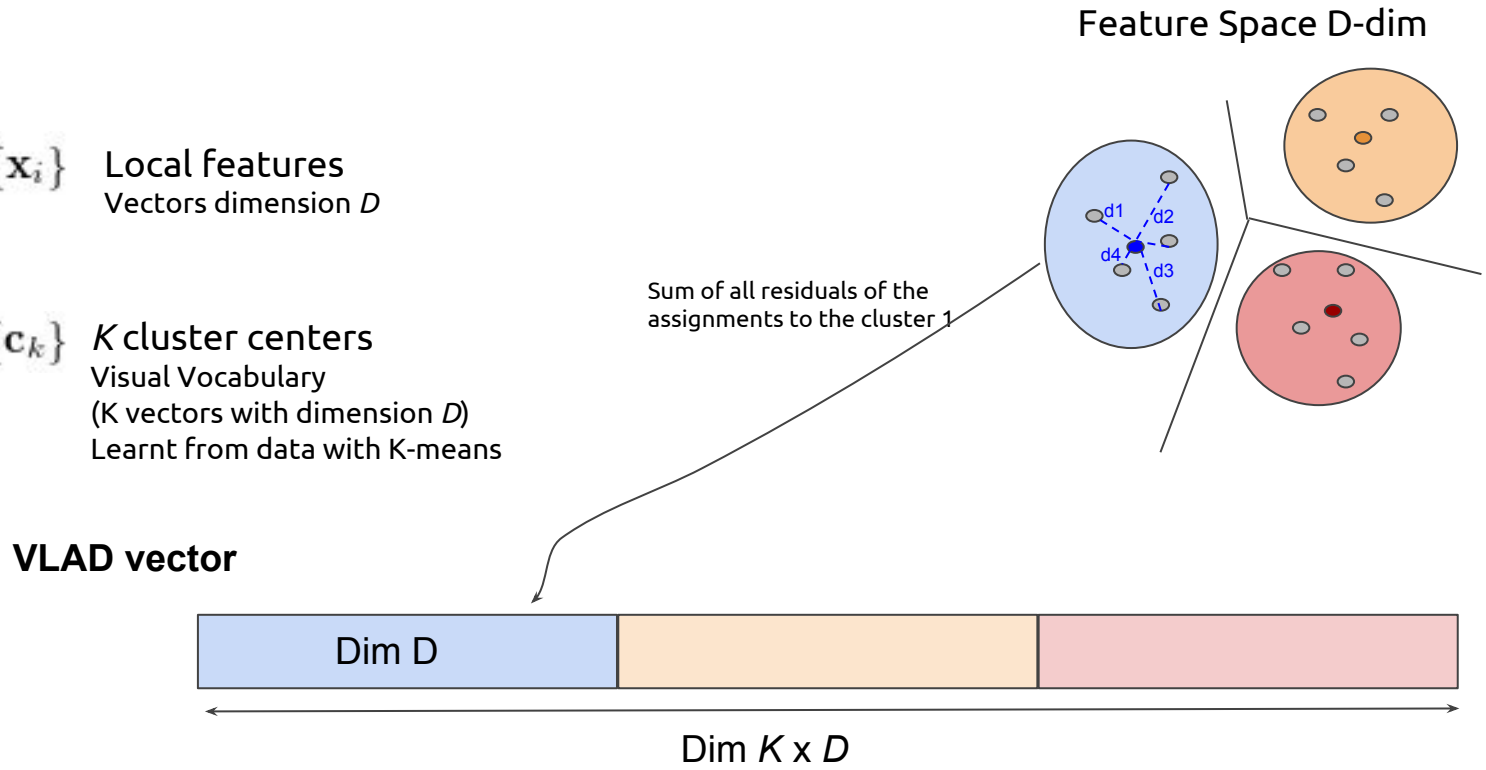
NetVLAD

Relja Arandjelović et al, NetVLAD: CNN architecture for weakly supervised place recognition, CVPR 2016

VLAD

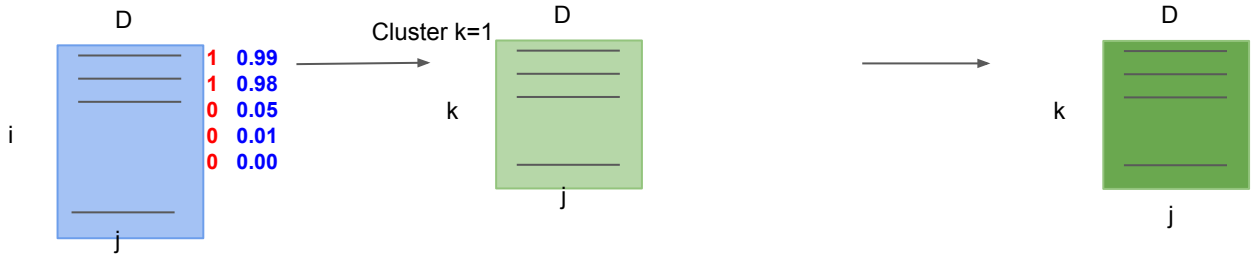
$\{\mathbf{x}_i\}$ Local features
Vectors dimension *D*

$\{\mathbf{c}_k\}$ *K* cluster centers
Visual Vocabulary
(K vectors with dimension *D*)
Learnt from data with K-means



VLAD

N Local features $\{X_i\}$ K Clusters features $\{C_k\}$ V Matrix VLAD (residuals)



$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i)\,(x_i(j) - c_k(j))\,,$$

Hard Assignment
Not differentiable !

Soft Assignment

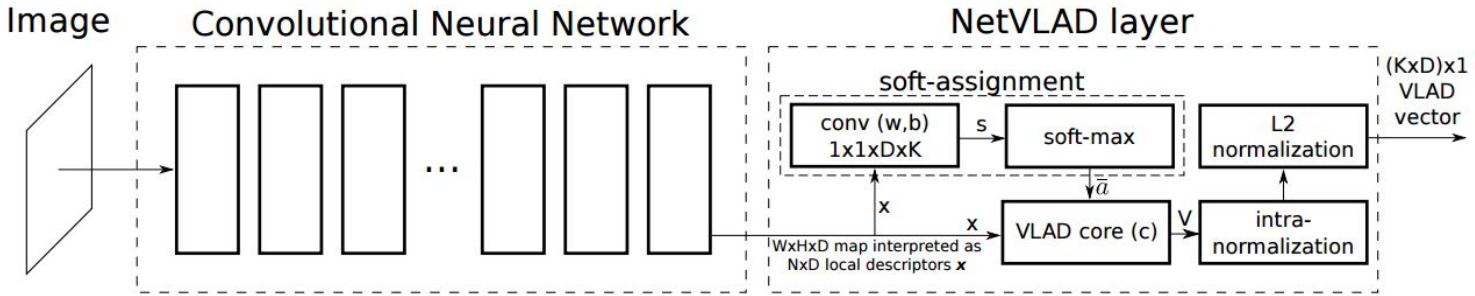
$$\bar{a}_k(\mathbf{x}_i) = \frac{e^{-\alpha\|\mathbf{x}_i - \mathbf{c}_k\|^2}}{\sum_{k'} e^{-\alpha\|\mathbf{x}_i - \mathbf{c}_{k'}\|^2}},$$

Softmax function :)

$$\bar{a}_k(\mathbf{x}_i) = \frac{e^{\mathbf{w}_k^T \mathbf{x}_i + b_k}}{\sum_{k'} e^{\mathbf{w}_{k'}^T \mathbf{x}_i + b_{k'}}},$$

Relja Arandjelović et al, NetVLAD: CNN architecture for weakly supervised place recognition, CVPR 2016

NetVLAD



Training based on GPS annotated data

$$L_\theta = \sum_j l\left(\min_i d_\theta^2(q, p_i^q) + m - d_\theta^2(q, n_j^q)\right)$$

$$l(x) = \max(x, 0),$$

Consider the positive example closer to the query

Consider all negative candidates

With distance inferior to a certain margin

Relja Arandjelović et al, NetVLAD: CNN architecture for weakly supervised place recognition, CVPR 2016

End-to-End learning for retrieval #2

Fine-tuned R-MAC

Gordo et al, Deep Image Retrieval: Learning global representations for image search. ECCV 2016

R-MAC

Region vectors
512D

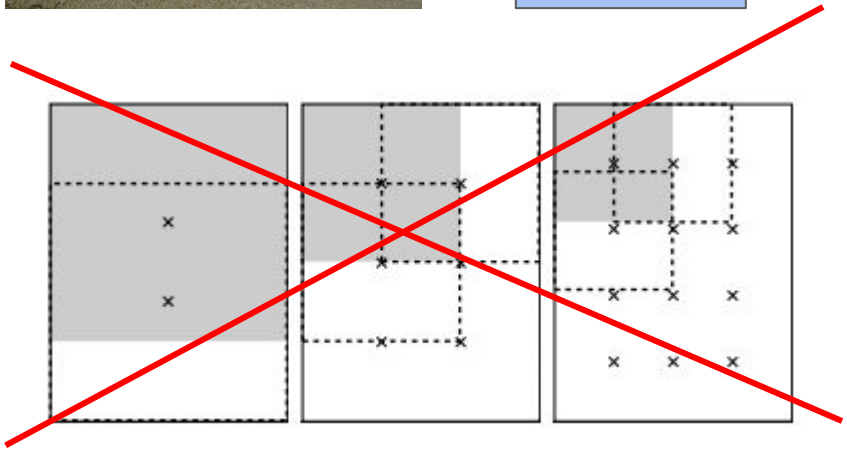
Region1
Region2
Region3
Region4
...
RegionN

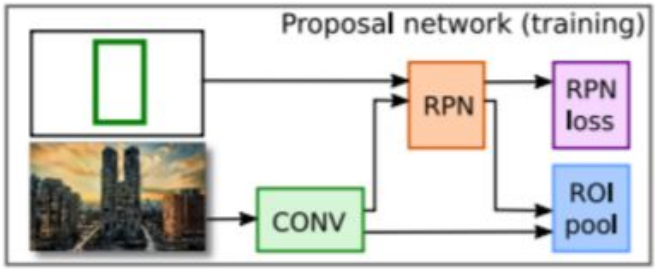
l2-PCAw-l2

Region1
Region2
Region3
Region4
...
RegionN

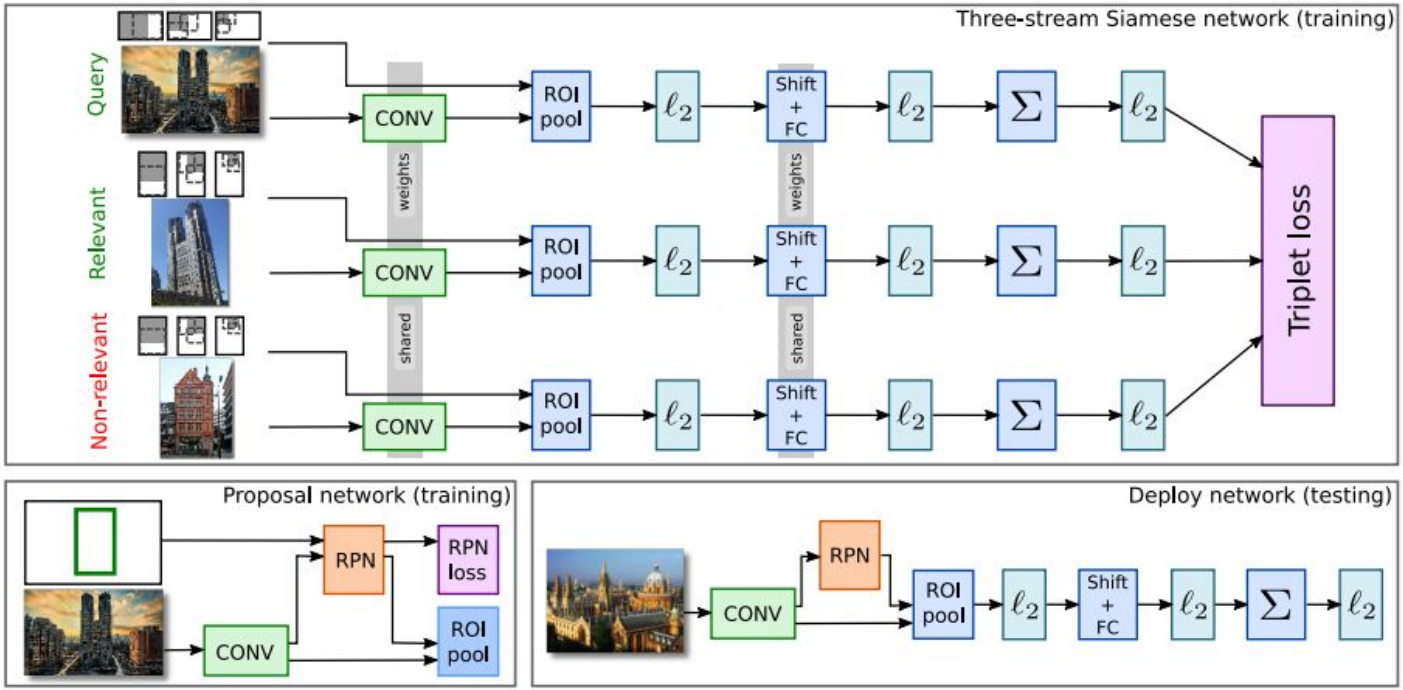
$\sum$

Global
vector
512D





Learning representations for retrieval



Gordo et al, Deep Image Retrieval: Learning global representations for image search. ECCV 2016

Learning representations for retrieval

Comparison between R-MAC from off-the-shelf network and R-MAC retrained for retrieval

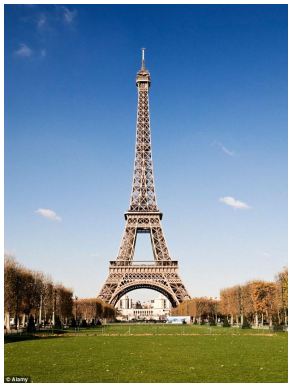
		R-MAC		Learned R-MAC		
Dataset	PCA	[14]	Reimp.	C-Full	C-Clean	R-Clean
Oxford 5k	PCA Paris	66.9	66.1	-	-	-
	PCA Landmarks	-	64.7	75.3	75.9	**78.6**
Paris 6k	PCA Oxford	83.0	82.5	-	-	-
	PCA Landmarks	-	81.6	82.2	83.7	**84.5**

Gordo et al, Deep Image Retrieval: Learning global representations for image search. ECCV 2016

Extra: Ranking refinement

Ranking refinement: Query Expansion

Query Image

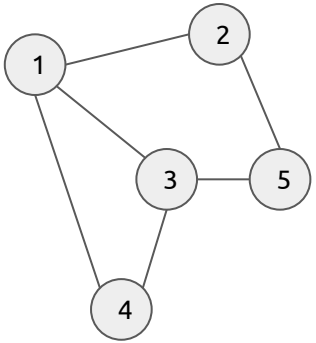


New Query

- Runs at query time
- Boost performance if the initial ranking *is good*
  - Traditionally applied on spatially verified images (RANSAC)
- Only considers the top NN of the image

Ranking refinement: Diffusion

Diffusion mechanism to propagate similarities through a **pairwise affinity matrix**



Affinity Matrix

0	1	1	1	0
1	0	0	0	1
1	0	0	1	1
1	0	1	0	0
0	1	1	1	1

On Random walks and Diffusion on graphs:

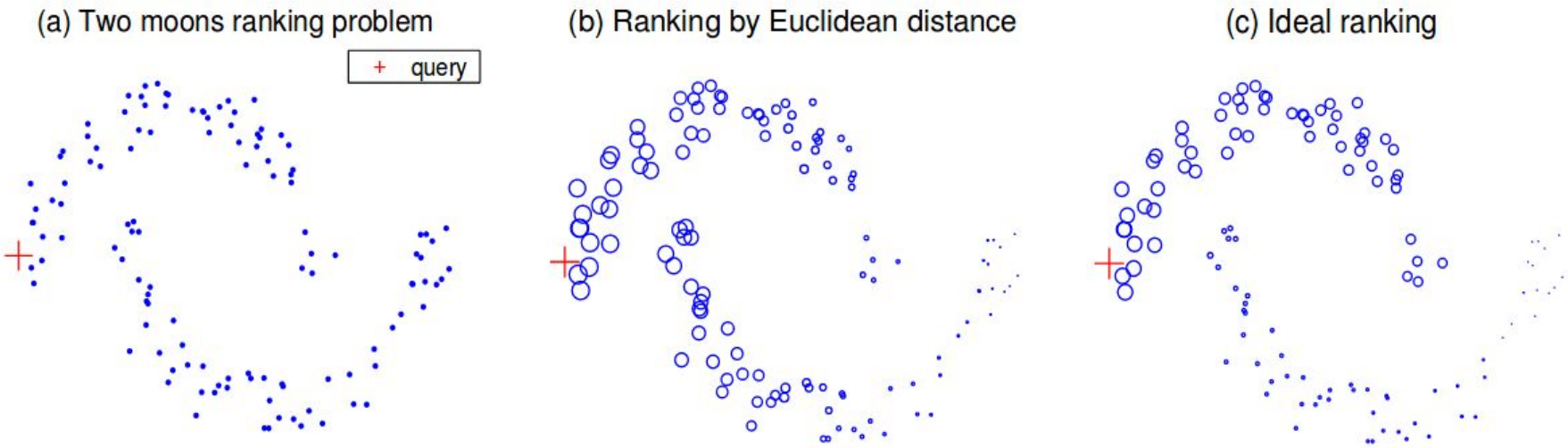
Leonid E. Zhukov, Structural Analysis and Visualization of Networks [slides] [video lecture]

Ranking refinement: Diffusion

Diffusion mechanism to propagate similarities through a **pairwise affinity matrix**

- Most retrieval approaches using pairwise measures (i.e Euclidean distance between query vs dataset images) to generate the final rank.
- This has the limitation of ignoring the structure of the underlying data manifold

Ranking refinement: Diffusion



Zhou et al. Ranking on data manifolds. In *Advances in neural information processing systems* (pp. 169-176), 2004

Method	$m \times d$	INSTRE	Oxf5k	Oxf105k	Par6k	Par106k
Global descriptors - nearest neighbor search						
CroW [30]†	512	-	68.2	63.2	79.8	71.0
R-MAC [43]	512	47.7	77.7	70.1	84.1	76.8
R-MAC [19]	2,048	62.6	83.9	80.8	93.8	89.9
NetVLAD [1]†	4,096	-	71.6	-	79.7	-
Global descriptors - query expansion						
R-MAC [43]+AQE [8]	512	57.3	85.4	79.7	88.4	83.5
R-MAC [43]+SCSM [48]	512	60.1	85.3	80.5	89.4	84.5
R-MAC [43]+HN [42]	512	64.7	79.9	-	92.0	-
Global diffusion	512	70.3	85.7	82.7	94.1	92.5
R-MAC [19]+AQE [8]	2,048	70.5	89.6	88.3	95.3	92.7
R-MAC [19]+SCSM [48]	2,048	71.4	89.1	87.3	95.4	92.5
Global diffusion	2,048	80.5	87.1	87.4	96.5	95.4
Regional descriptors - nearest neighbor search						
R-match [44]	21×512	55.5	81.5	76.5	86.1	79.9
R-match [44]	21×2,048	71.0	88.1	85.7	94.9	91.3
Regional descriptors - query expansion						
HQE [51]	2.4k×128	74.7	89.4†	84.0†	82.8†	-
R-match [44]+AQE [8]	21×512	60.4	83.6	78.6	87.0	81.0
Regional diffusion⋆	5×512	77.5	91.5	84.7	95.6	93.0
Regional diffusion⋆	21×512	80.0	93.2	90.3	96.5	92.6
R-match [44]+AQE [8]	21×2,048	77.1	91.0	89.6	95.5	92.5
Regional diffusion⋆	5×2,048	88.4	95.0	90.0	96.4	**95.8**
Regional diffusion⋆	21×2,048	**89.6**	**95.8**	**94.2**	**96.9**	95.3

Iscen, A et al. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. CVPR 2017

Overview

- What is content-based image retrieval?
- The classic SIFT retrieval pipeline
- Using off the shelf CNN features for retrieval
- Learning representations for retrieval

Questions?