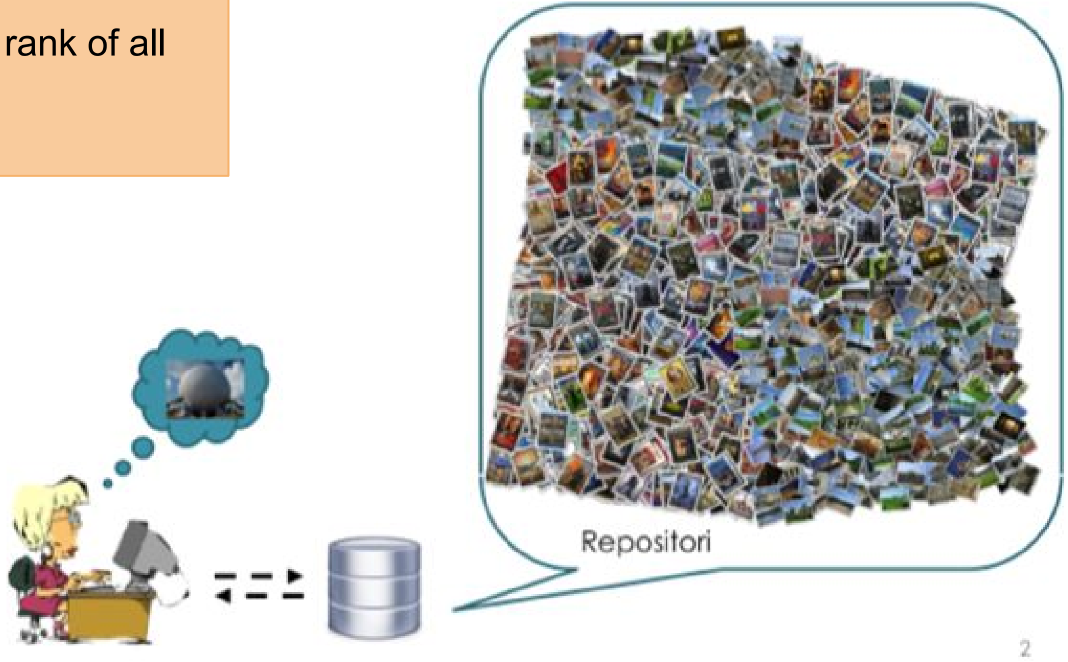
Day 3 Lecture 2

Rankings

Content Based Image Retrieval

Given an image query, generate a rank of all similar images.



Classification

Query: This chair

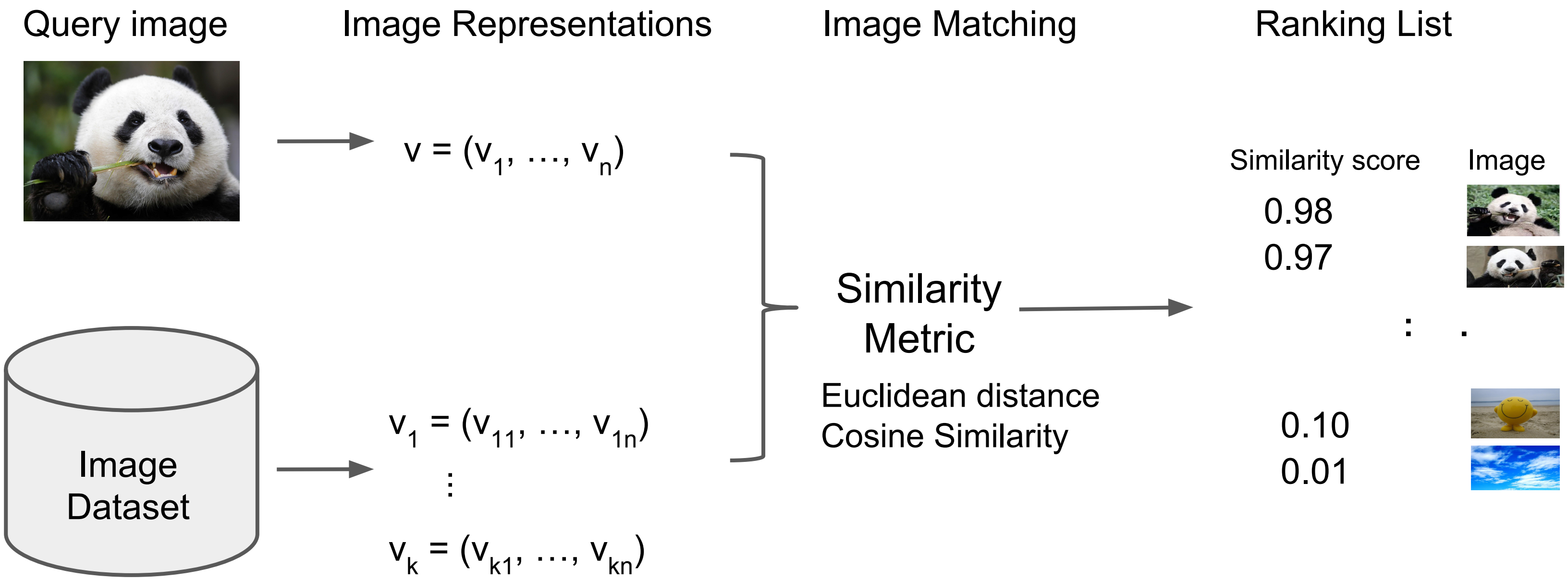
Results from dataset classified as “chair”

Retrieval

Query: This chair

Similar images

Retrieval Pipeline



Retrieval Pipeline

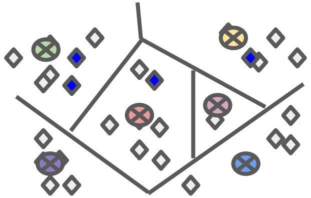
k feature vectors per image

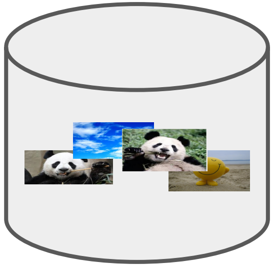
$$v_1 = (v_{11}, \ldots, v_{1n})$$

$$\vdots$$

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

N-Dimensional feature space





M visual words (M clusters)

Bag of Visual Words

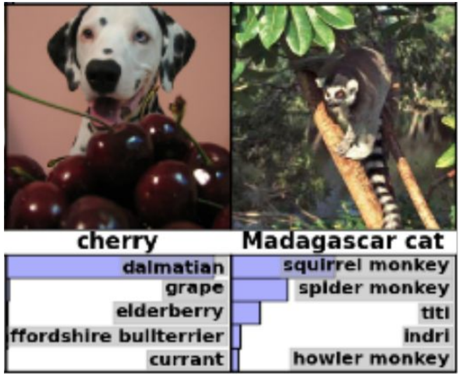
INVERTED FILE

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

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

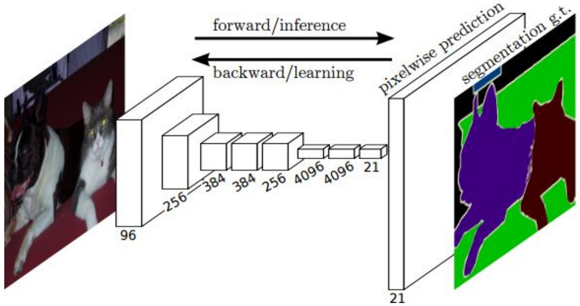
CNN for retrieval

Classification



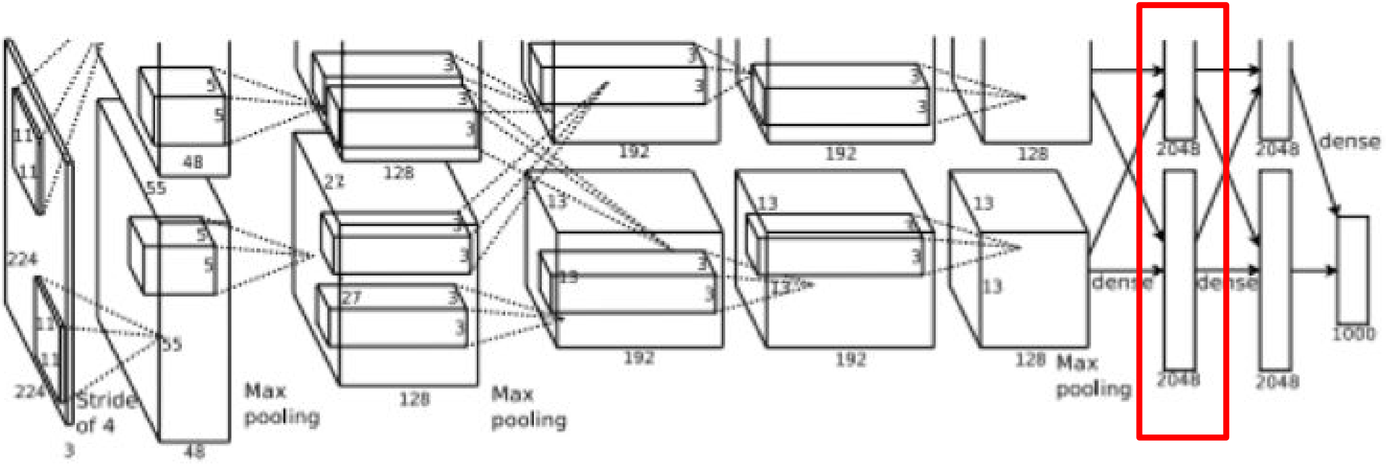
Object Detection

Segmentation



Off-the-shelf CNN representations

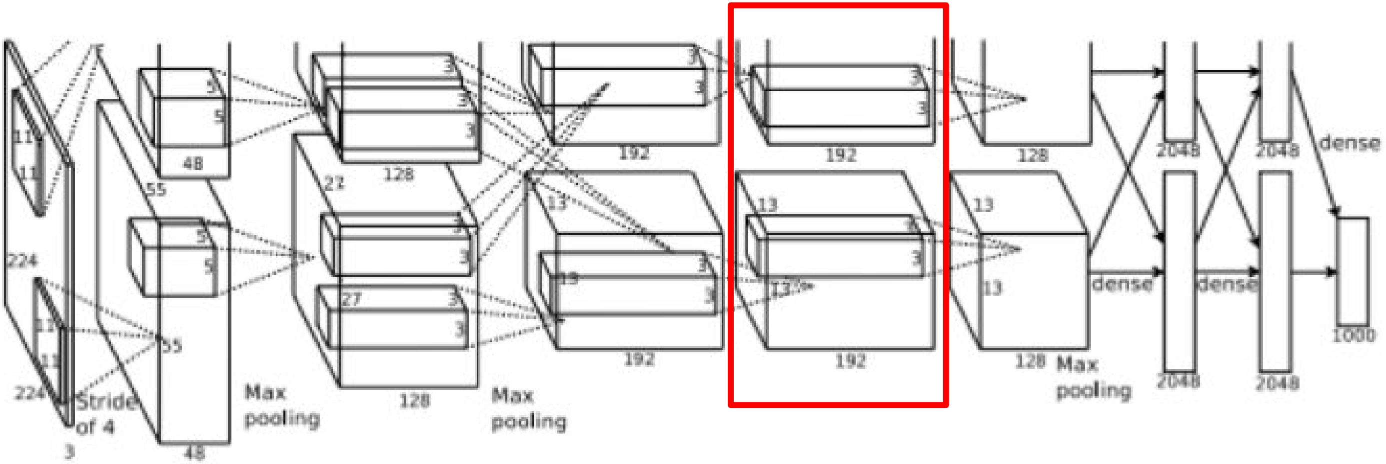
FC layers as global feature representation



Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. In *ECCV*
Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *CVPRW*

Off-the-shelf CNN representations

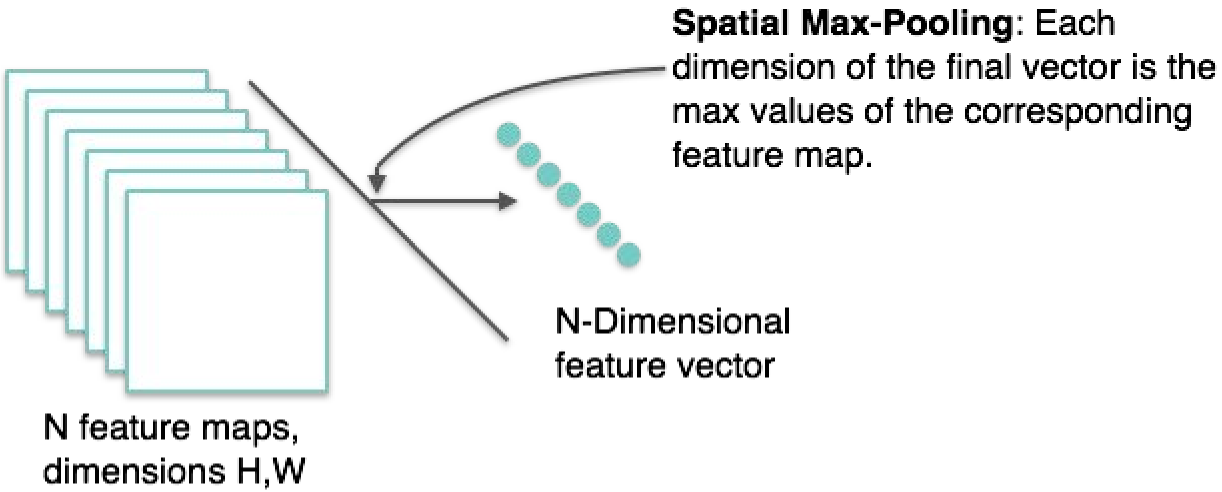
sum/max pool conv features across filters



Babenko, A., & Lempitsky, V. (2015). Aggregating local deep features for image retrieval. ICCV
Tolias, G., Sicre, R., & Jégou, H. (2015). Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879*.
Kalantidis, Y., Mellina, C., & Osindero, S. (2015). Cross-dimensional Weighting for Aggregated Deep Convolutional Features. *arXiv preprint arXiv:1512.04065*.

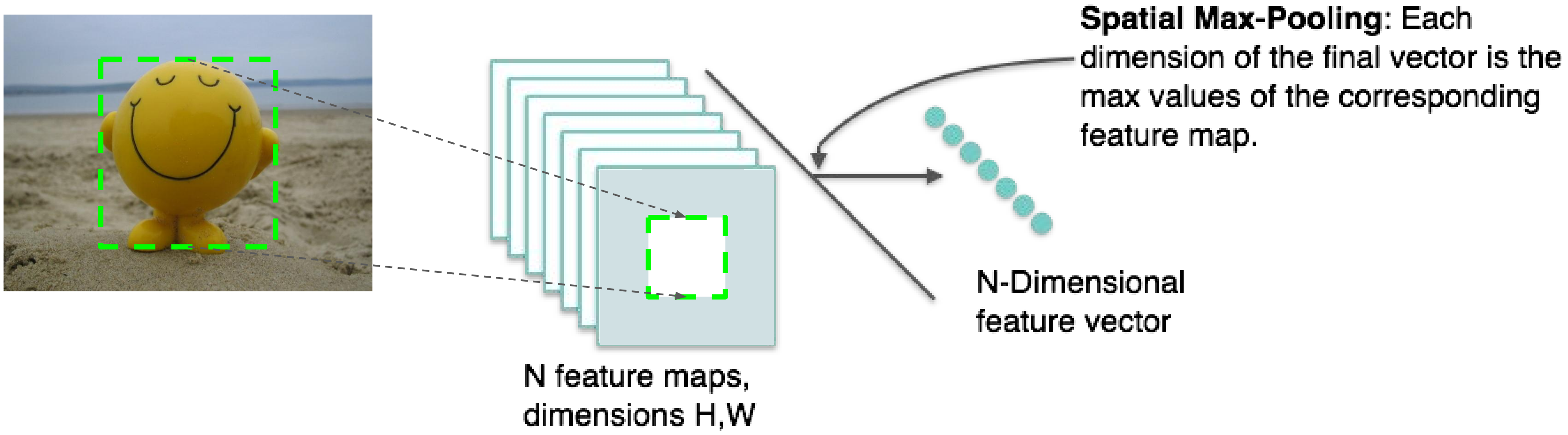
Off-the-shelf CNN representations

Descriptors from convolutional layers



Off-the-shelf CNN representations

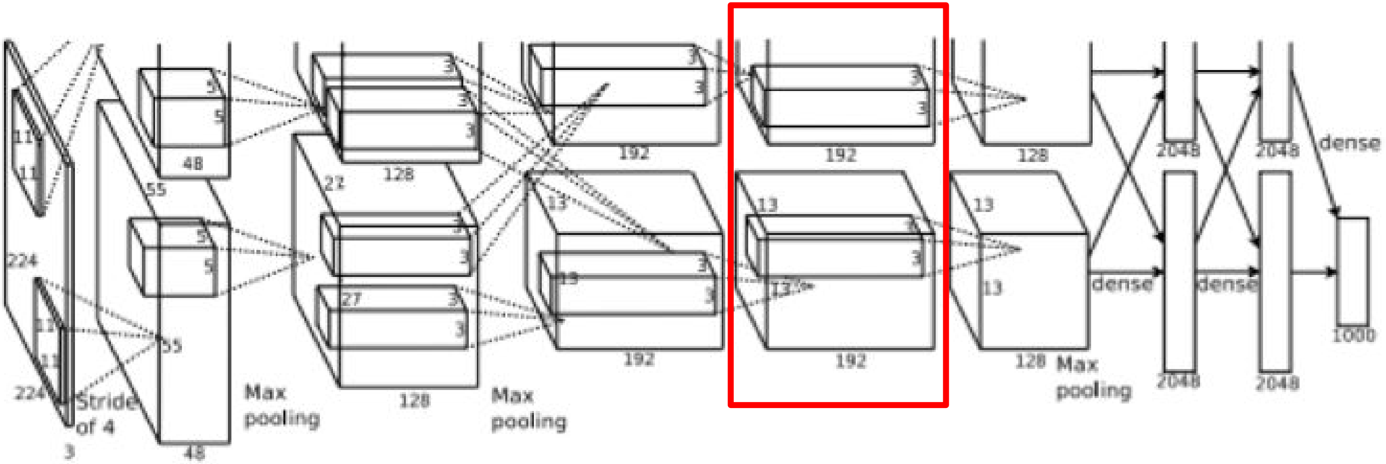
R-MAC: Regional Maximum Activation of Convolutions



Tolias, G., Sicre, R., & Jégou, H. (2015). [Particular object retrieval with integral max-pooling of CNN activations](). *arXiv preprint arXiv: 1511.05879*.

Off-the-shelf CNN representations

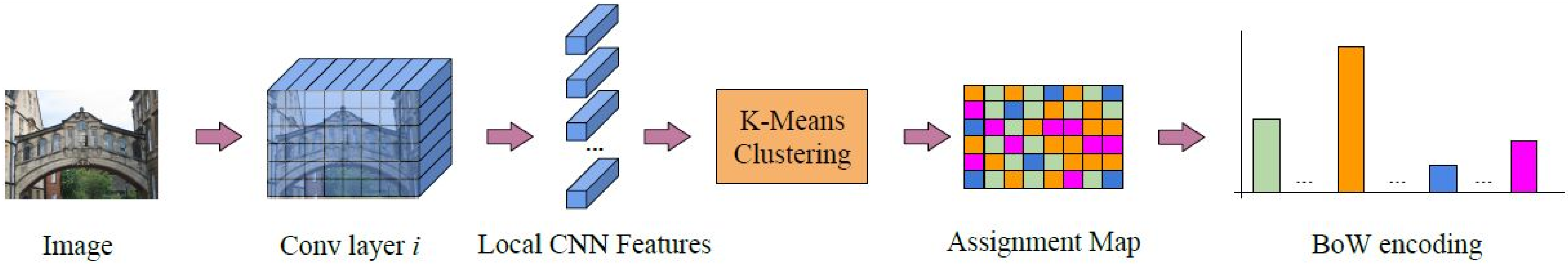
BoW, VLAD encoding of conv features



Ng, J., Yang, F., & Davis, L. (2015). [Exploiting local features from deep networks for image retrieval](). In *CVPRW*
Mohedano, E., Salvador A., McGuinnes K, Marques F, O'Connor N, Giro-i-Nieto X (2016). [Bags of Local Convolutional Features for Scalable Instance Search](). In ICMR

Off-the-shelf CNN representations

Descriptors from convolutional layers



(336x256) Resolution

conv5_1 from VGG16 (42x32)

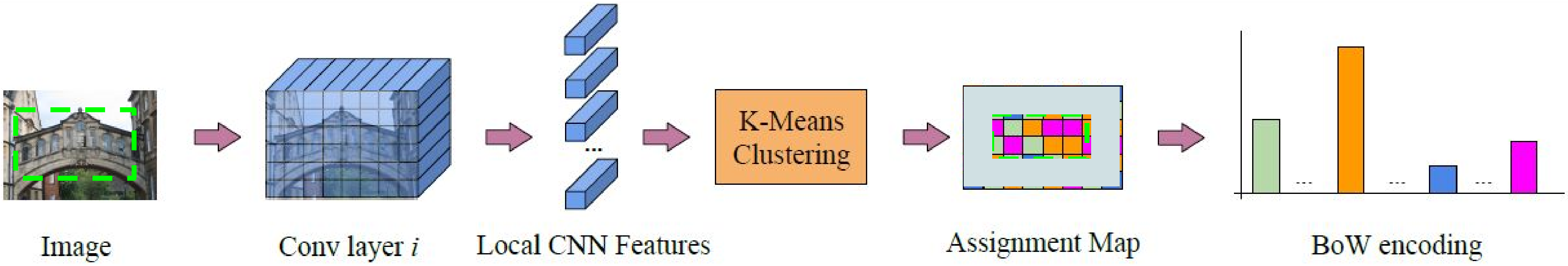
25K centroids

25K-D vector

Mohedano, E., Salvador A., McGuinnes K, Marques F, O'Connor N, Giro-i-Nieto X (2016). Bags of Local Convolutional Features for Scalable Instance Search. In ICMR

Off-the-shelf CNN representations

Descriptors from convolutional layers



(336x256)
Resolution

conv5_1 from
VGG16
(42x32)

25K centroids

25K-D vector

Off-the-shelf CNN representations

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, Y., Mellina, C., & Osindero, S. (2015). Cross-dimensional Weighting for Aggregated Deep Convolutional Features. *arXiv preprint arXiv:1512.04065*.

Mohedano, E., Salvador A., McGuinnes K, Marques F, O'Connor N, Giro-i-Nieto X (2016). Bags of Local Convolutional Features for Scalable Instance Search. In ICMR

Off-the-shelf CNN representations

CNN representations

- l2 Normalization + PCA whitening + l2 Normalization
- Cosine similarity
- Convolutional features better than fully connected features
- Convolutional features keep spatial information → Retrieval+object location
- Convolutional layers allows custom input size.
- If data labels available, fine tuning the network to the image domain improves CNN representations.

Learning representations for retrieval

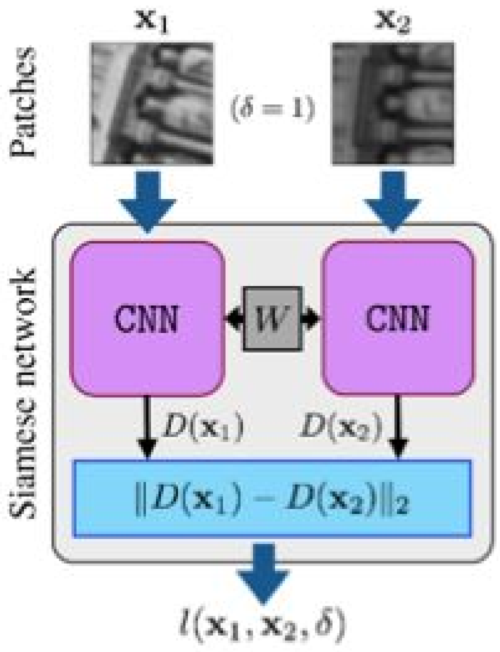
Siamese Network: Network to learn a function that maps input patterns into a target space such that l2-norm in the target space approximates the semantic distance in the input space.

Applied in:

Dimensionality reduction[1]

Face verification[2]

Learning local image representations[3]



[1] Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR.
[2] S. Chopra, R. Hadsell and Y. LeCun, Learning a similarity metric discriminatively, with application to face verification.*(CVPR'05)*[3] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, and F. Moreno-Noguer. Fracking deep convolutional image descriptors. CoRR, abs/1412.6537, 2014

Learning representations for retrieval

Siamese Network: Network to learn a function that maps input patterns into a target space such that l2-norm in the target space approximates the semantic distance in the input space.

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

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

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

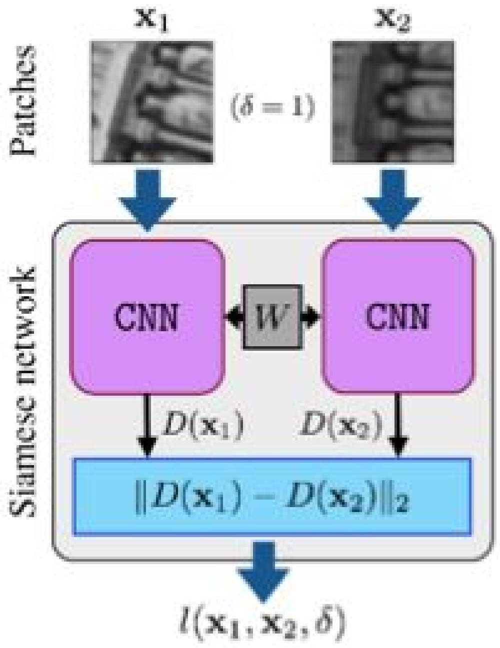
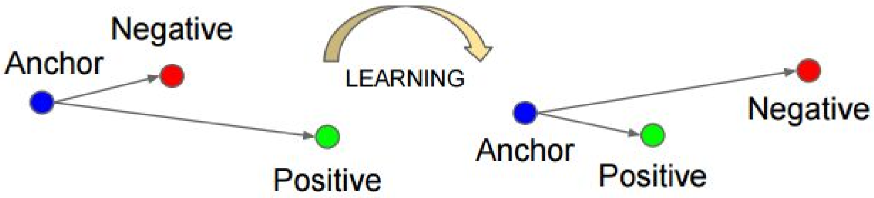
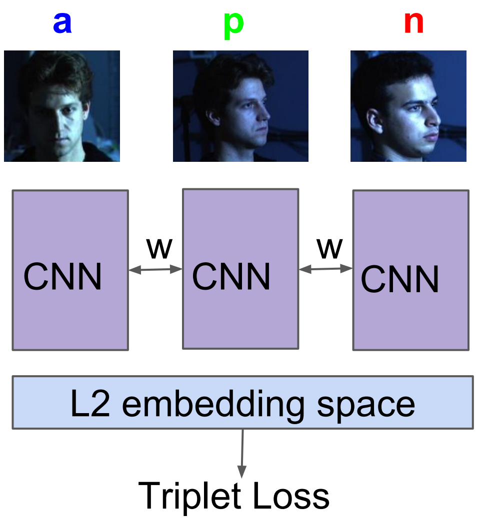


Image from: Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, and F. Moreno-Noguer. Fracking deep convolutional image descriptors. CoRR, abs/1412.6537, 2014

Learning representations for retrieval

Siamese Network with Triplet Loss: Loss function minimizes distance between query and positive and maximizes distance between query and negative





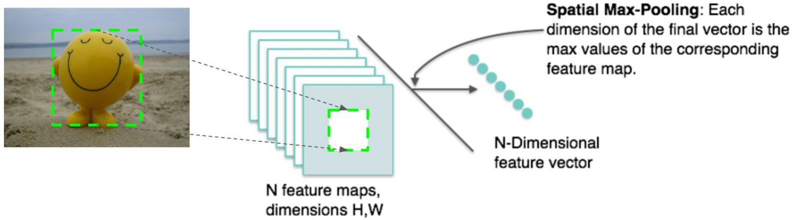
$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$$

https://arxiv.org/pdf/1503.03832v3.pdf

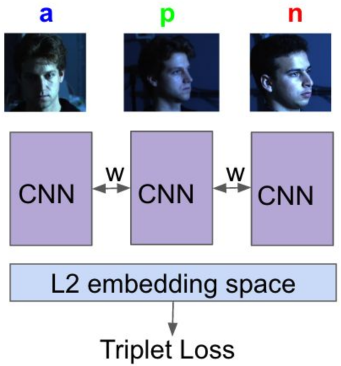
Learning representations for retrieval

[Deep Image Retrieval: Learning global representations for image search](), Gordo A. et al. Xerox Research Centre, 2016

- R-MAC representation
- Learning descriptors for retrieval using three channels siamese loss: Ranking objective

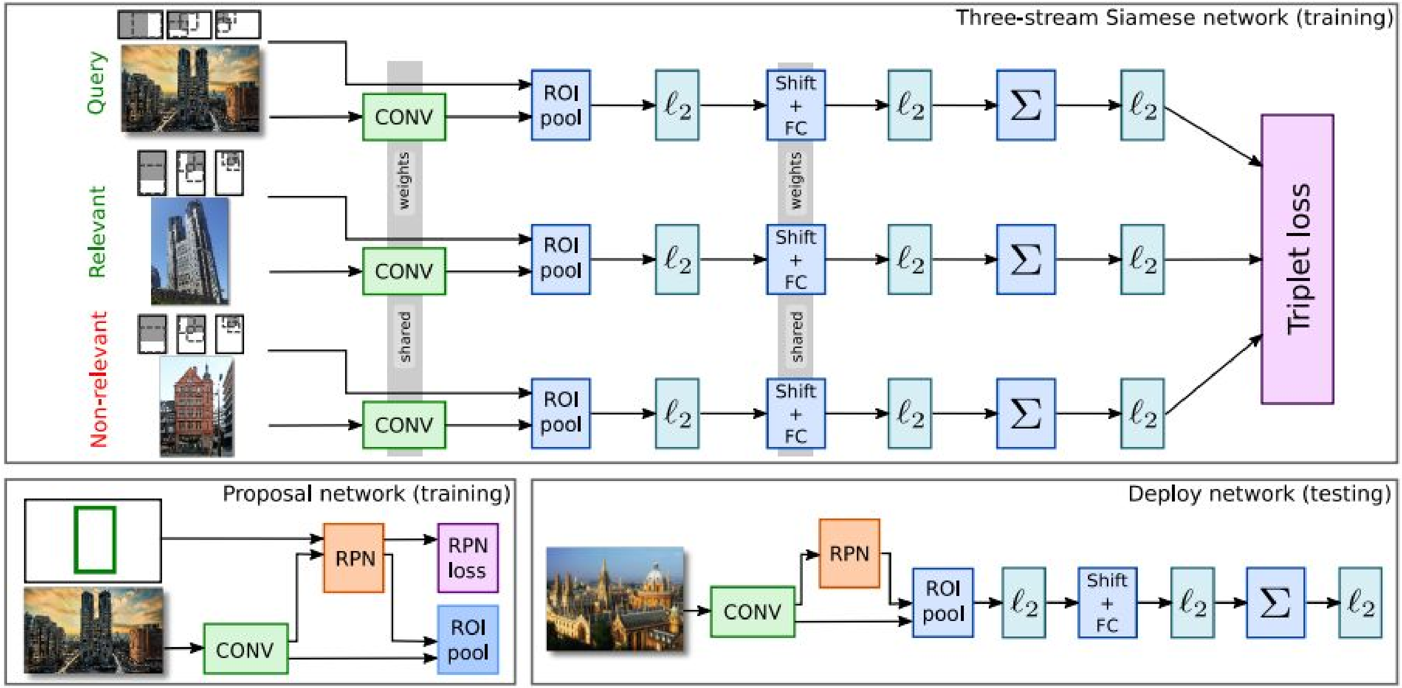


$$L(I_q, I^+, I^-) = \max(0, m + q^T d^- - q^T d^+)$$



- Learning where to pool within an image: predicting object locations
- Local features (from predicted ROI) pooled into a more discriminative space (learned fc)
- Building and cleaning a dataset to generate triplets

Learning representations for retrieval

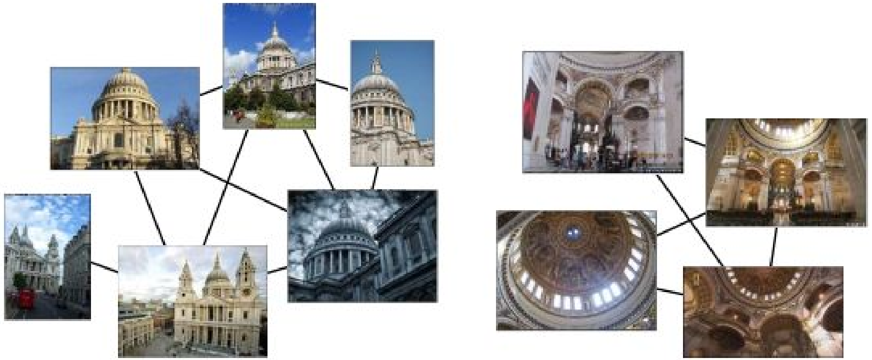


Learning representations for retrieval

[Deep Image Retrieval: Learning global representations for image search](), Gordo A. et al. Xerox Research Centre, 2016

Dataset: Landmarks dataset:

- 214K images of 672 famous landmark site.
- Dataset processing based on a matching baseline: SIFT + Hessian-Affine keypoint detector.
- Important to select the “useful” triplets.



Learning representations for retrieval

[Deep Image Retrieval: Learning global representations for image search](#),
Gordo A. et al. Xerox Research Centre, 2016

		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**

Comparison between training for Classification (C) of training for Rankings (R)

Learning representations for retrieval

[Deep Image Retrieval: Learning global representations for image search](),
Gordo A. et al. Xerox Research Centre, 2016

	Method	Dim.	Datasets				
			Oxf5k	Par6k	Oxf105k	Par106k	Holidays
Global descriptors	Jégou & Zisserman [54]	1024	56.0	-	50.2	-	72.0
	Jégou & Zisserman [54]	128	43.3	-	35.3	-	61.7
	Gordo *et al.* [55]	512	-	-	-	-	79.0
	Babenko *et al.* [17]	128	55.7	-	52.3	-	78.9*
	Gong *et al.* [15]	2048	-	-	-	-	80.8
	Razavian *et al.* [56]	256	53.3	67.0	48.9	-	74.2*
	Babenko & Lempitsky[12]	256	53.1	-	50.1	-	80.2
	Ng *et al.* [57]	128	59.3*	59.0*	-	-	83.6
	Paulin *et al.* [32]	256K	56.5	-	-	-	79.3
	Perronnin & Larlus [31]	4000	-	-	-	-	84.7
	Tolias *et al.* [14]	512	66.9	83.0	61.6	75.7	84.6†
	Kalantidis *et al.* [13]	512	68.2	79.7	63.3	71.0	84.9
	Previous state of the art		68.2 [13]	83.0 [14]	63.3 [13]	75.7 [14]	84.9 [13]
	Ours	512	**81.3**	**85.5**	**76.8**	**78.5**	**86.0**

Summary

Pre-trained CNN are useful to generate image descriptors for retrieval

Convolutional layers allow us to encode local information

Knowing how to rank similarity is the primary task in retrieval

Designing CNN architectures to learn how to rank