Centre for
Data Analytics

Insight

Deep Image Representations for Instance Search

Eva Mohedano

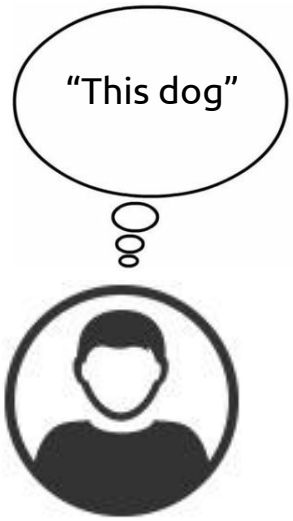
supervised by Dr. Kevin McGuinness and Noel E. O'Connor

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN sfi Science Foundation Ireland

Contents

Contents

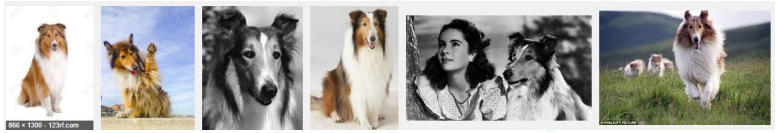
Visual Instance Retrieval



Visual Query

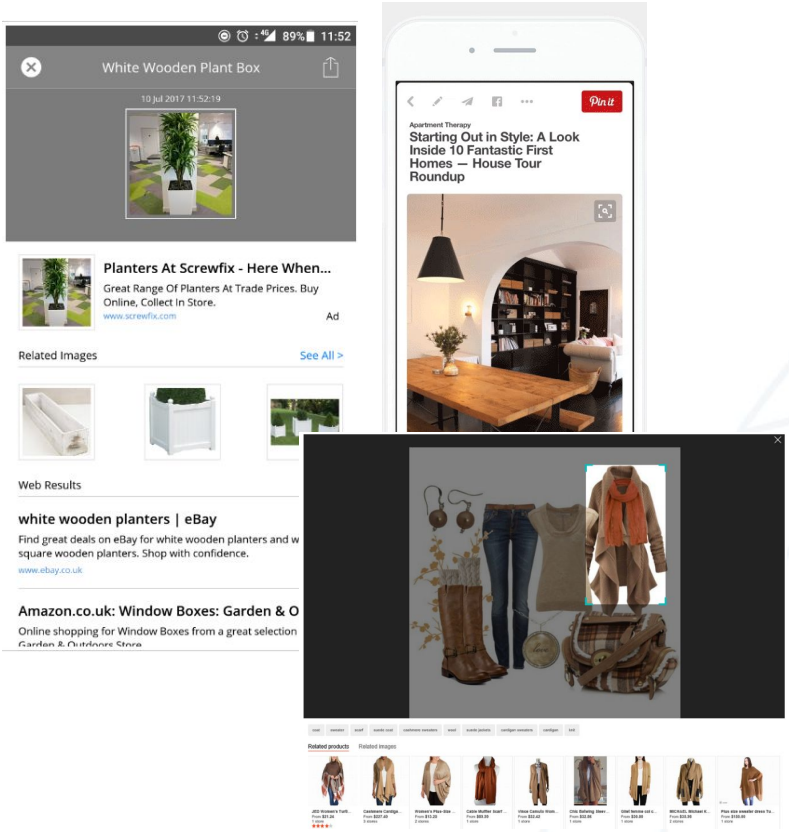
Image Database

Expected outcome:

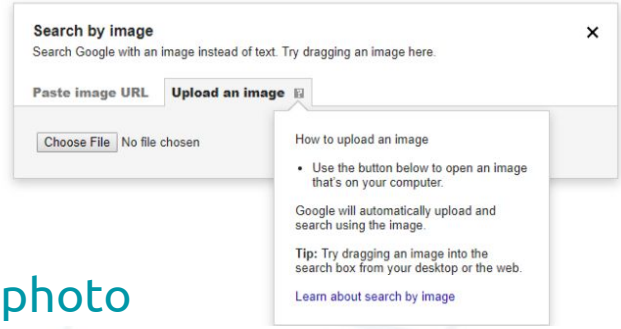


Applications

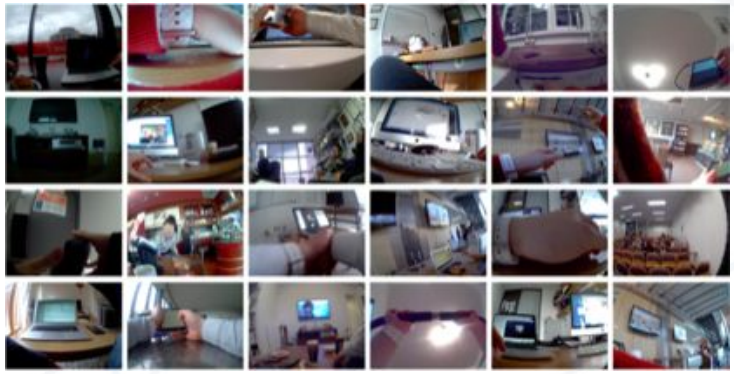
e-commerce



seeking information

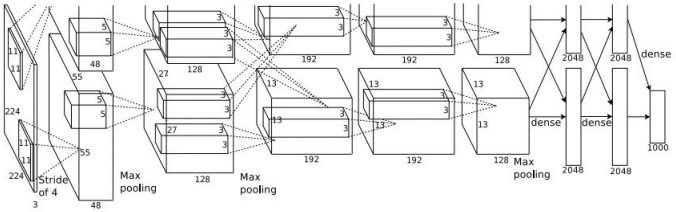


personal photo organization



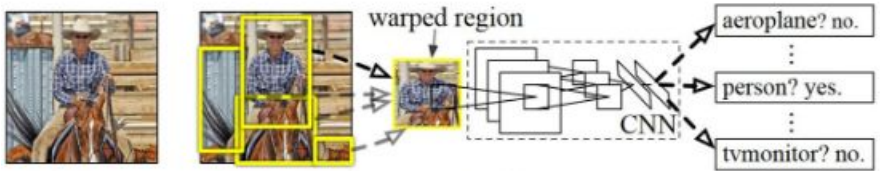
Deep Learning

Image Classification



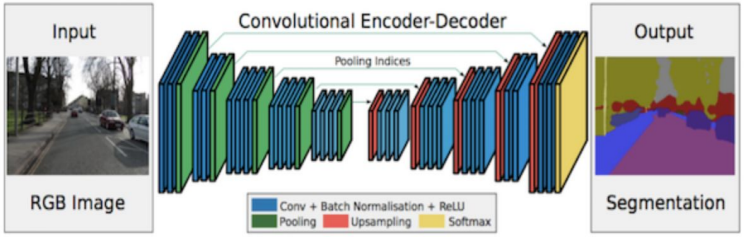
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

Object detection



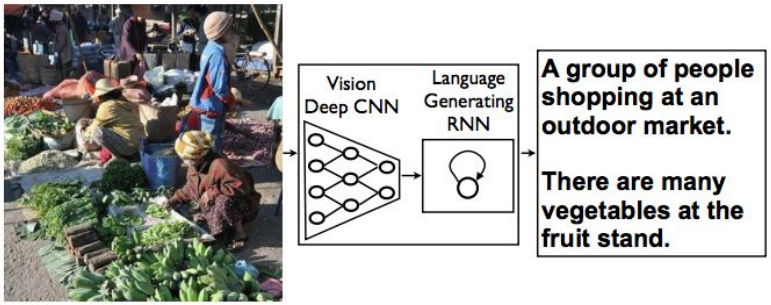
Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

Semantic segmentation



Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *arXiv preprint arXiv:1511.00561* (2015).

Image captioning



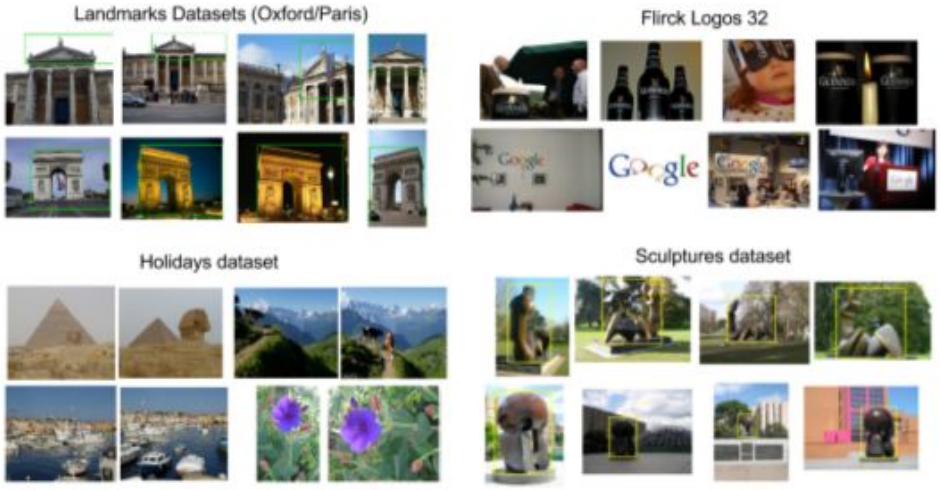
Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

Contents

Motivation

Dataset Complexity

Current retrieval benchmarks have limited diversity and are domain specific



Name	Number of images	Number of instances	Number of relevant images/instance	Domain
Oxford5k	5063	11	7-220	Buildings
Paris6k	6392	11	87-782	Buildings
Holidays	1491	500	1-12	Landmarks/Objects
Sculptures6k	6340	10		Sculptures
Flirck Logos 32	2240	32	70	Logos

Motivation

Dataset Complexity

Datasets specifically designed for instance retrieval

TRECVID Instance Search
464 hours of video content

INSTRE
23k images with 200 instances

Motivation

Image Representation

Compact & Dense
(e.g. sum/max pooling conv feats, FC feats) → Capacity?

High-dimensional & Dense
(e.g. VLAD encoding) → Scalability?

High-dimensional & Sparse

↓

Bag of Visual Words

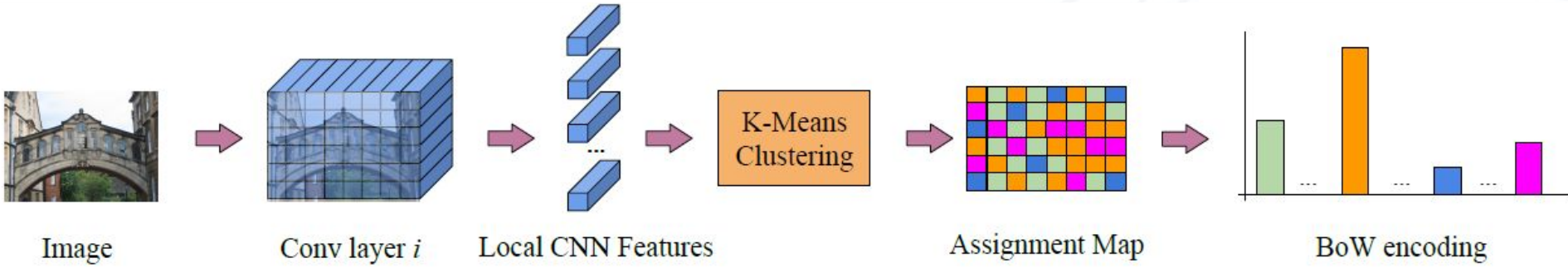
Contributions

- Exploration of traditional BoW on local convolutional features
- Exploration of unsupervised fine-tuning strategy to improve local CNN descriptors
- Exploration of attention models to enhance the capability of CNN representations

Generic and scalable instance retrieval pipeline

Contents

Bag of Words Framework



(336x256) Resolution

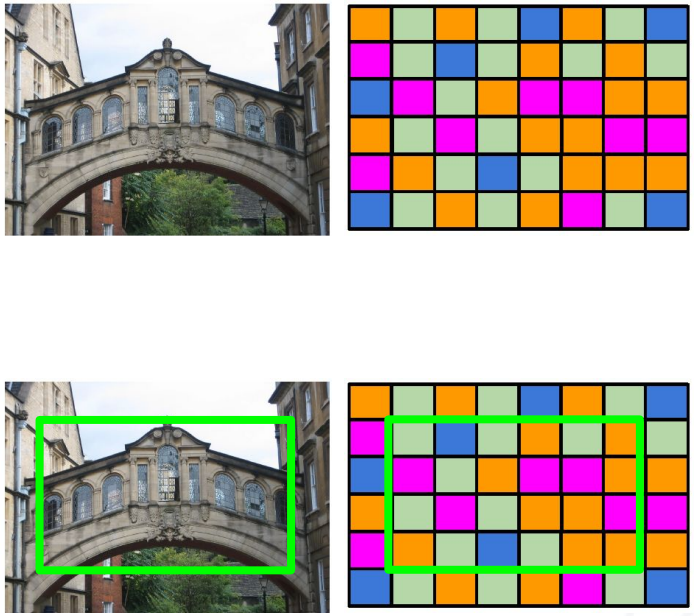
conv5_1 from VGG16[1] (42x32)

25K centroids

25K-D vector

Assignment Maps

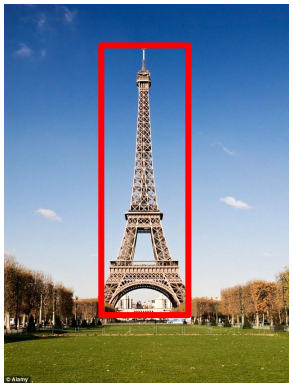
Query Representation



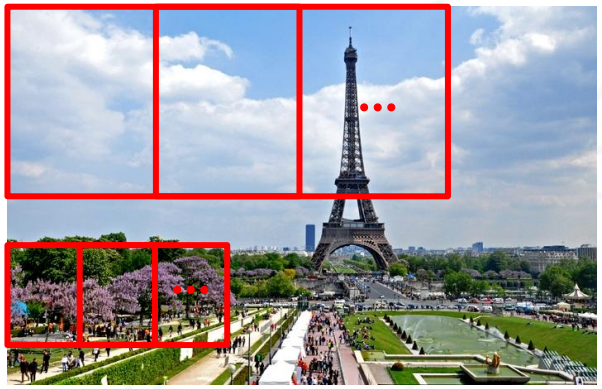
Assignment Maps

Spatial Re-ranking

Query Image



Target image in top M ranking



All window combinations with: $h \in \{H, \frac{H}{2}, \frac{H}{4}\}$ $w \in \{W, \frac{W}{2}, \frac{W}{4}\}$

Quantitative Evaluation

	Oxford 5k		Paris 6k	
		$+Q_{aug}$		$+Q_{aug}$
GS	0.653	0.697	0.699	0.754
LS	0.738	0.758	0.820	0.832
GS + R	0.701	0.713	0.719	0.752
LS + R	0.734	0.760	0.815	0.828
GS + GQE	0.702	0.730	0.774	0.792
LS + GQE	0.773	0.780	0.814	0.832
GS + R + GQE	0.771	0.772	0.801	0.798
LS + R + GQE	0.769	**0.793**	0.807	0.828
GS + R + LQE	0.782	0.757	0.835	0.795
LS + R + LQE	**0.788**	0.786	**0.848**	**0.833**

- Local search better
- Query expansion in general beneficial
- Re-ranking does not improve much but is used to select good regions for query expansion

Qualitative Evaluation

Results on Trecvid

		Oxford 5k	Paris 6k	INS 23k
Ours	GS	0.650	0.608	**0.323**
	LS	**0.739**	**0.819**	0.295
Sum pool (as ours)	GS	0.621	0.712	0.156
	LS	0.583	0.742	0.097
Sum pool (as in [20])	GS	0.672	0.774	0.139
	LS	0.683	0.763	0.120

Efficiency

	Oxford	Paris	TRECVID
memory (MB)	3.47	4.11	31.47
words/image	171	160	285
query/time (s)	0.002	0.003	0.02

- Sumpooling 512D
- Larger bottleneck in re-ranking 8.5s

Contents

Limitation of pre-trained networks

Classification

Query: This chair

Results from dataset classified as “chair”

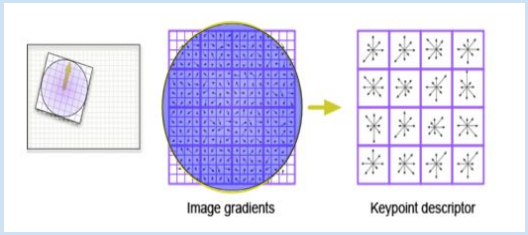
Limitation of pre-trained networks

Retrieval

Query: This chair

Results from dataset ranked by similarity to the query

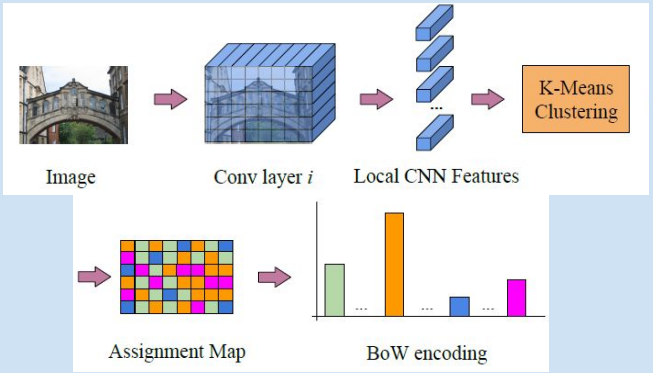
Learning from an “expensive” retrieval system



SIFT - Bag of Visual words

and

(Late Fusion)



CNN - Bag of Visual words

	Oxford	Paris	INSTRE
BLCF	0.786	0.843	0.726
SIFT-BoW	0.865	0.803	0.382
merged	**0.901**	**0.915**	**0.729**

Combining SIFT and CNN-based systems

When SIFT is better...

← SIFT

← CNN

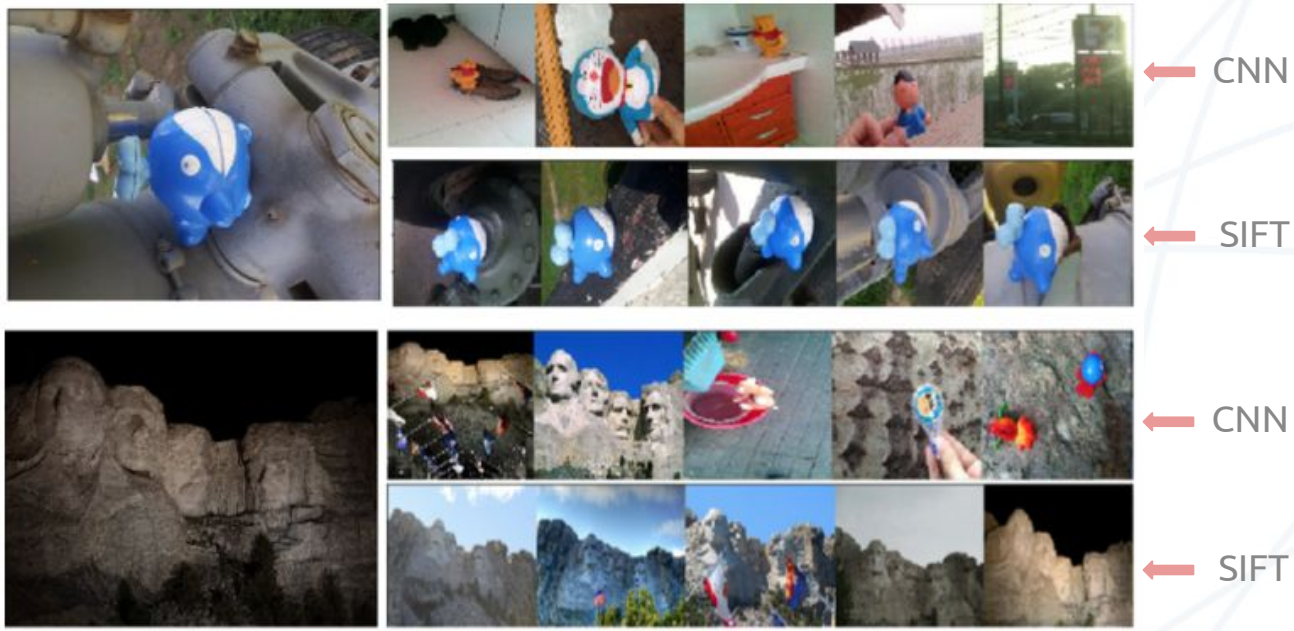
Planar and textured instances

← SIFT

← CNN

Combining SIFT and CNN-based systems

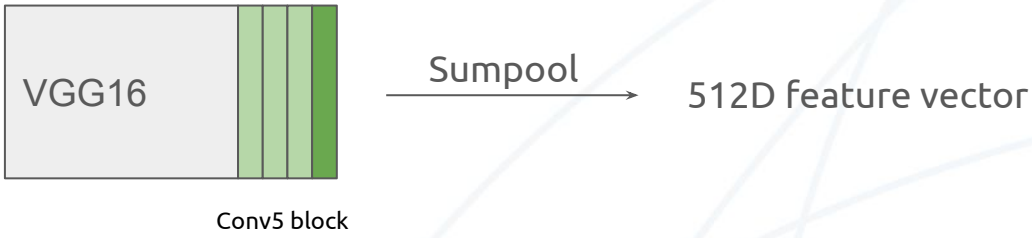
When CNN is better...



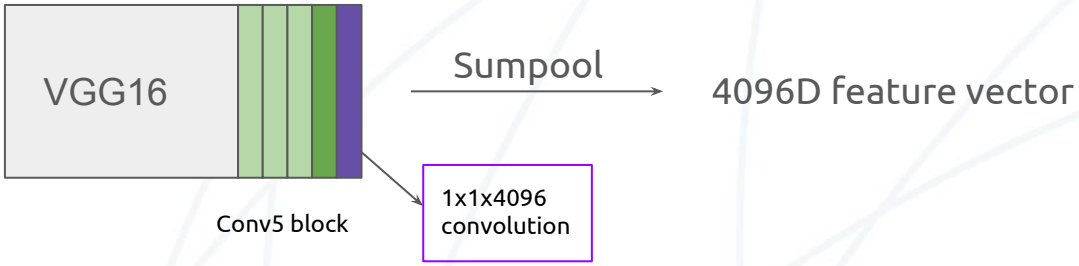
3D object with textureless and/or reflective surfaces

Network architecture

Network A



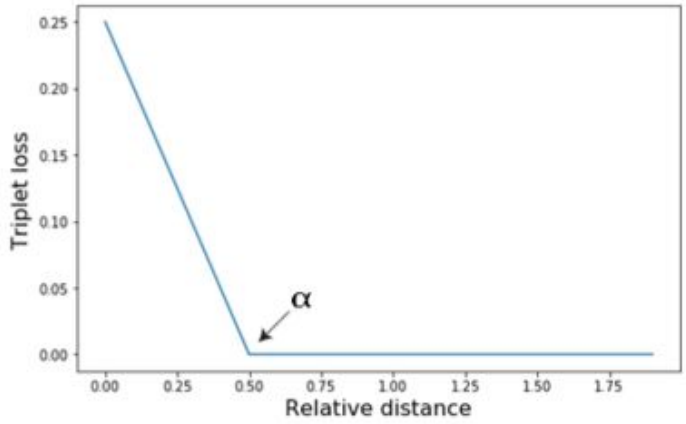
Network B

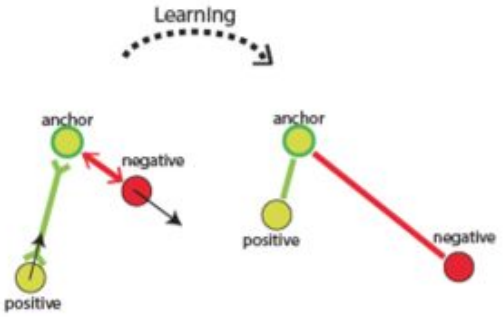


Similarity Learning

Triplet loss

$$\mathcal{L}(a, p, n) = \max(0, D(a, p) - D(a, n) + \alpha),$$





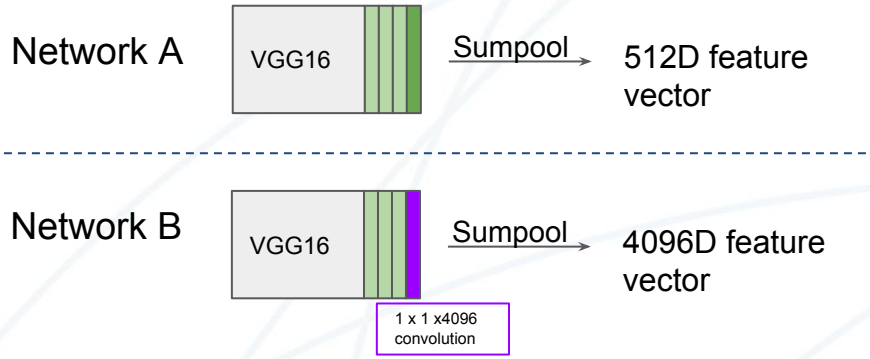
Triplet Sampling Strategy

- Generation of rank list with merged CNN-SIFT system.
- Random selection of a rank list.
- For a given rank:
  - Positive pair (anchor and positive) two images with a score larger than 15 (sharing 30 local SIFT matches)
  - A negative image is randomly selected from scores between 1 and 0.5.

Fine-tuning in Oxford

Rankings from all images

	Oxford	Paris
Baseline A	0.480	0.698
Baseline B	0.503	**0.716**
(F)Network A	**0.575**	0.302
(F)Network B	0.502	0.715



- 5063 ranks referring to 11 different buildings and junk images.
- Less overfitting but still overfitting!
- Baseline B better than baseline A: random weight projecto to a higher space beneficial by itself.
- (F) Network B not trained.
  - Not enough data
  - Not enough training time

Fine-Tuning on Instre

	Oxford	Paris	INSTRE
Baseline A	0.480	0.698	0.275
Baseline B	**0.503**	**0.716**	0.268
(F)Network A	0.071	0.147	0.257
(F)Network B	0.498	0.712	0.268
(F+cl)Network B	0.246	0.242	**0.587**

- Instre is a much more diverse dataset that seems to need class labels to successfully fine-tuned network.
- Similarity learning ends in overfitting the network but not in improving performance

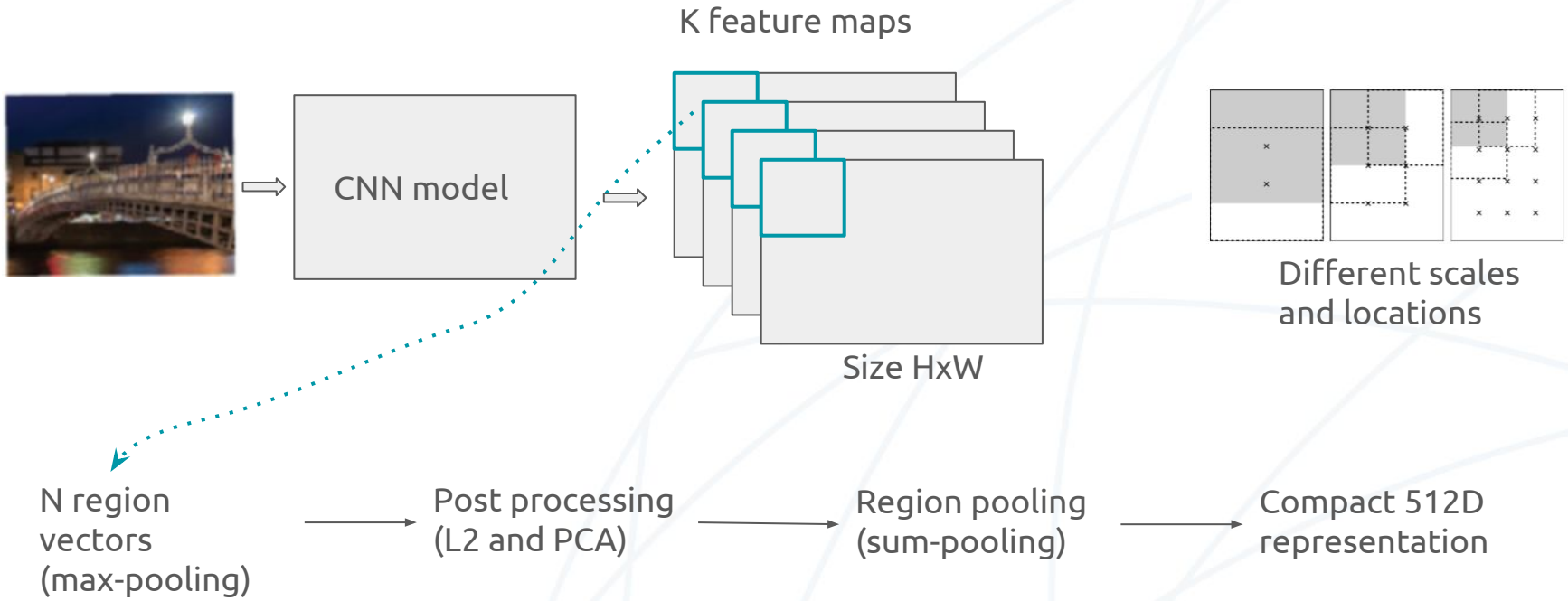
Remarks

- Fine-tuning good strategy but specific to a particular image domain
- In a generic retrieval system we do not know what kind of instance are we going to search

Contents

Regional Maximum Activation of Convolution

R-MAC



R-MAC as weighting scheme

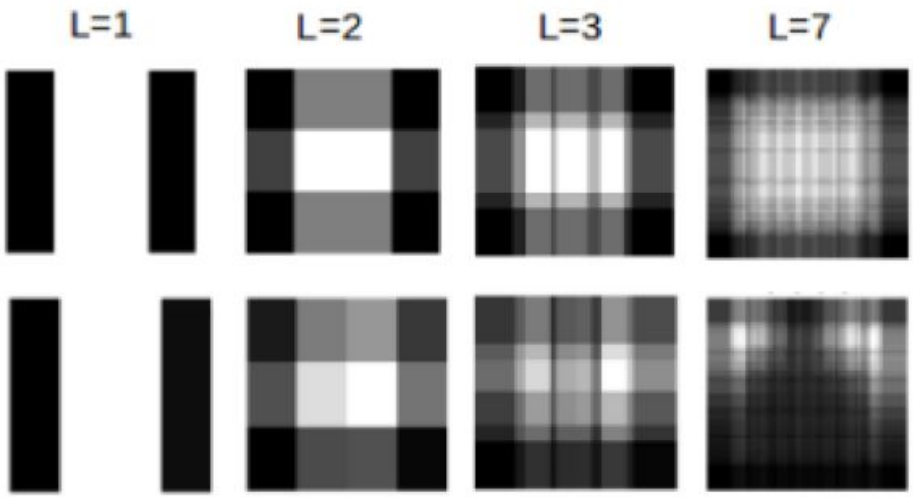
$$F = \sum_{i=1}^{W} \sum_{i=1}^{H} \alpha(i,j) \odot f(i,j)$$

Original convolutional feature maps

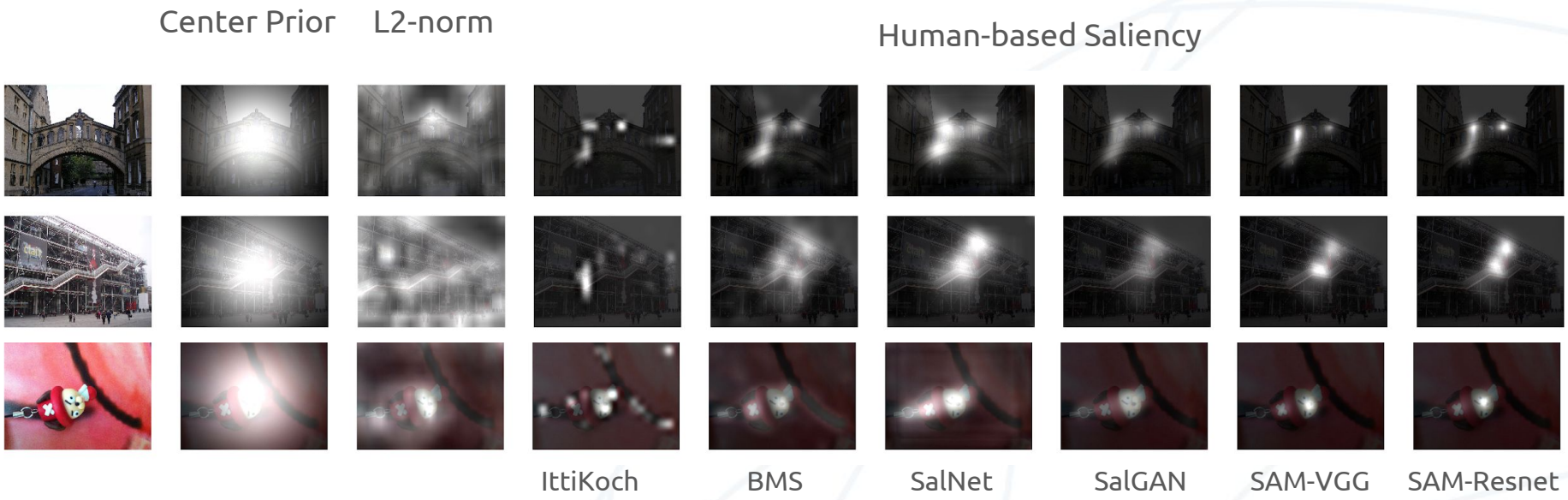
Provide more emphasis to the center regions

Window location

Region post-processing



Saliency Prediction



- Is saliency useful for instance retrieval?
- What is the effect in instance retrieval of different quality saliency models?

Results Sum-pooling

	Oxford		Paris		Instre	
Weighting	GS	LS	GS	LS	GS	LS
None	0.680	0.686	0.779	0.765	0.261	0.405
Gaussian	**0.684**	**0.688**	**0.795**	**0.793**	0.318	0.459
L^2-norm	0.671	0.676	0.779	0.765	0.321	0.466
Itti-Koch	0.601	0.579	0.729	0.717	0.271	0.405
BMS	0.655	0.658	0.756	0.732	0.352	0.512
SalNet	0.671	0.681	0.778	0.766	0.352	0.519
SalGAN	0.612	0.612	0.749	0.749	**0.360**	**0.527**
SAM-VGG	0.492	0.509	0.655	0.636	0.317	0.487
SAM-ResNet	0.476	0.483	0.651	0.639	0.324	0.495

Result BLCF

	Oxford		Paris		Instre	
Weighting	GS	LS	GS	LS	GS	LS
None	0.628	0.722	0.642	0.798	0.350	0.636
Gaussian	0.666	0.728	0.701	0.809	0.447	0.656
L^2-norm	0.666	0.740	0.711	**0.817**	0.468	0.674
Itti-Koch	0.656	0.693	0.656	0.785	0.448	0.633
BMS	0.668	0.729	0.698	0.806	0.545	0.688
SalNet	0.670	0.746	0.726	0.814	0.547	0.688
SalGAN	**0.682**	**0.746**	**0.733**	0.812	0.607	**0.698**
SAM-VGG	0.638	0.686	0.709	0.785	0.618	0.688
SAM-ResNet	0.636	0.673	0.707	0.780	**0.622**	0.688

Comparison with other approaches

BLCF Local search

Method	Off-the-shelf	dim	Instre	Oxford	Paris
CroW [22]	yes	512	0.416	0.698	0.797
CAM [20]*	yes	512		0.736	0.855
R-MAC [43]	yes	512	0.523	0.691	0.835
R-MAC [35]†	No	512	0.477	0.777	0.841
R-MAC-ResNet [14]†	No	2048	0.626	**0.839**	**0.938**
(our) BLCF-SalGan	yes	336	**0.698**	0.746	0.812

BLCF Local search + Query Expansion

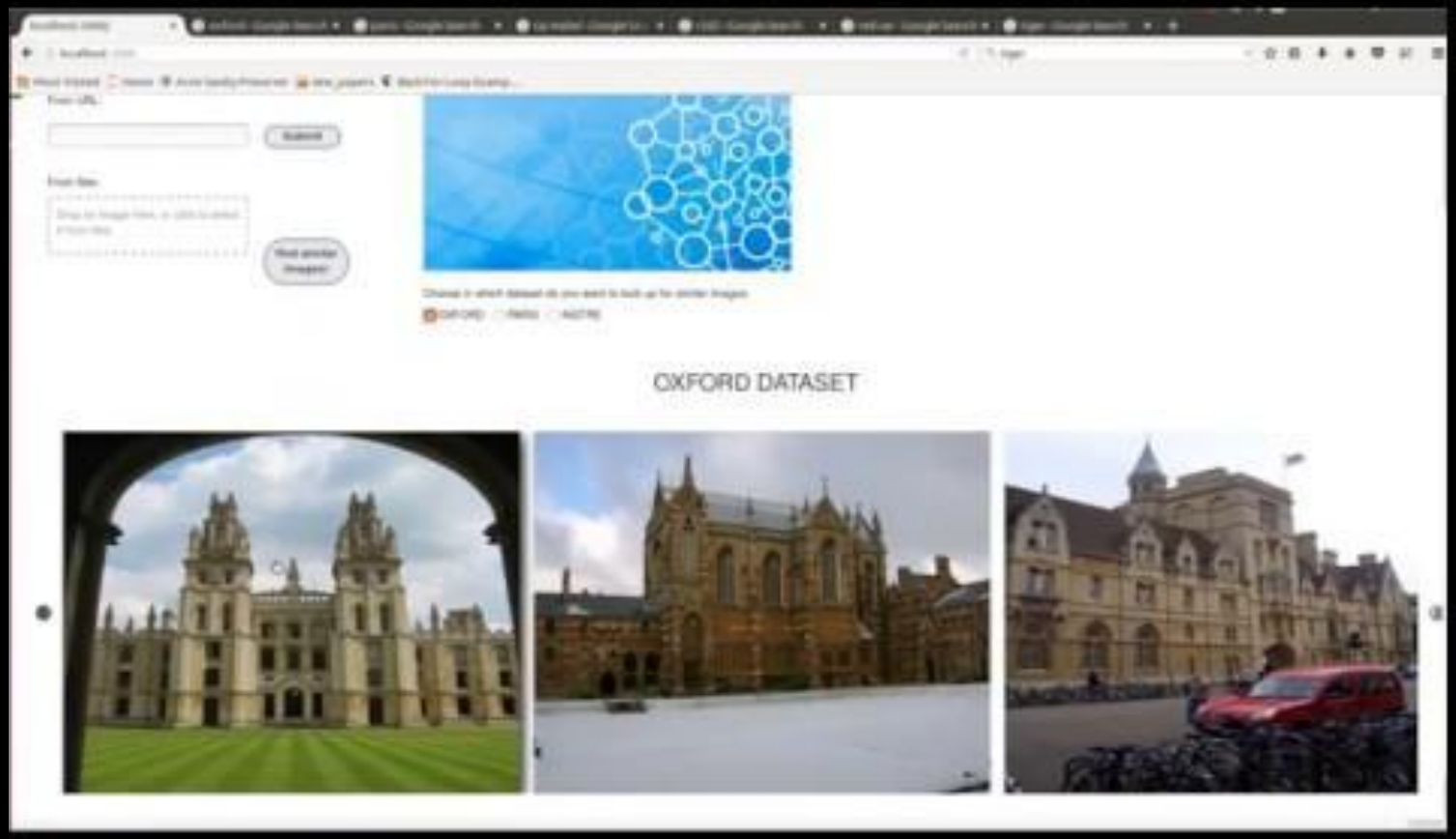
Method	Off-the-shelf	dim	Instre	Oxford	Paris
CroW [22]	yes	512	0.613	0.741	0.855
CAM [20]*	yes	512		0.760	0.873
R-MAC [43]	yes	512	0.706	0.770	0.884
R-MAC [35]†	No	512	0.573	0.854	0.884
R-MAC-ResNet [14]†	No	2048	0.705	**0.896**	**0.953**
(ours) BLCF-SalGan	yes	336	**0.757**	0.778	0.830

Contents

- **Application of traditional BoW encoding on local CNN representations.**
  - Spatial re-ranking exploiting the assignment maps representation
  - Efficient retrieval system
  - Outperforms other direct pooling approaches.

- **Unsupervised fine-tuning strategy exploiting the combination of SIFT and CNN-based systems.**
  - Results in systems specialised to the target domain
  - Not enough training data

- **Exploration of attention models to enhance the capability of CNN representations.**
  - No need of region analysis.
  - No need of fine-tuning and training data
  - Same approach achieves competitive performance in Oxford, Paris and achieves state-of-the-art in Instre.



Demo by Paula Gomez Duran
Dockerized visualization tool

Thanks for your attention!

Questions?

Visual Instance Retrieval

Query image | Image Representations | Image Matching | Ranking List

$v = (v_1, \ldots, v_n)$

Image Database

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

$\vdots$

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

Similarity Metric
(e.g. cosine similarity)

Similarity score	Image
0.98	
0.97	
$\vdots$	$\vdots$
0.10	
0.01	

Image Representation

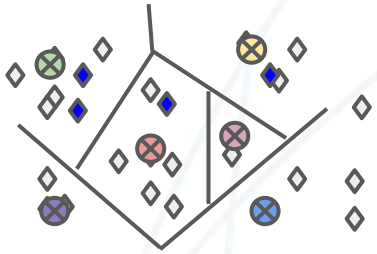
Local hand-crafted features
(e.g. SIFT)

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

$\vdots$

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

N-Dimensional
feature space



Bag of Visual Words

High-dimensional
Highly sparse

INVERTED FILE

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

Bag of Words encoding on fine-tuned features

Network A - Performance BLCF in Oxford (Training in Oxford)

	None		Gaussian		Saliency	
	GS	LS	GS	LS	GS	LS
off-the-shelf	0.628	0.706	0.666	0.714	0.690	0.730
fine-tuned	**0.669**	**0.739**	**0.708**	**0.736**	**0.731**	**0.750**

BLCF can also be applied to fine-tuned features

Publications

Mohedano, Eva, Kevin McGuinness, Noel E. O'Connor, Amaia Salvador, Ferran Marqués, and Xavier Giró-i-Nieto. "Bags of local convolutional features for scalable instance search." In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 327-331. ACM, 2016.

O'Connor, Noel E., Jiang Zhou, Eva Mohedano, Alan F. Smeaton, Jinhua Du, Haithem Afli, Manuela Hürlimann et al. "Dublin City University and Partners' participation in the INS and VTT Tracks at TRECVid 2016." *TRECVid 2016* (2016).

Reyes, Cristian, et al. "Where is my phone?: personal object retrieval from egocentric images." *Proceedings of the first Workshop on Lifelogging Tools and Applications*. ACM, 2016.

Mohedano, Eva, et al. "Improving object segmentation by using EEG signals and rapid serial visual presentation." *Multimedia tools and applications* 74.22 (2015): 10137-10159.

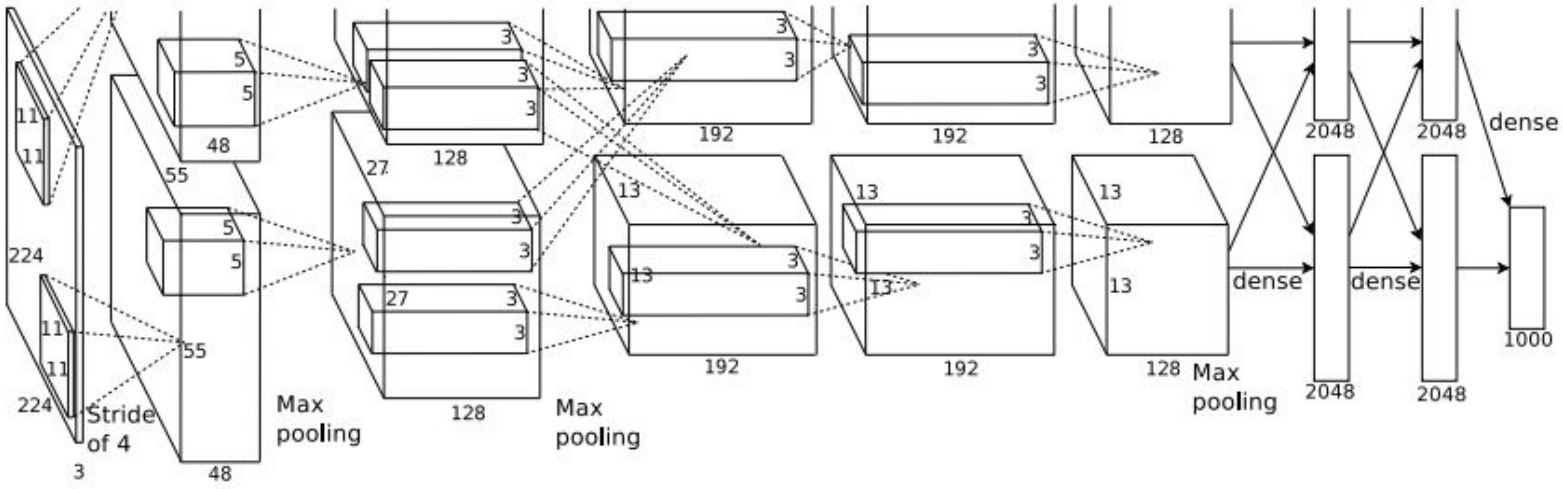
Mohedano, Eva, et al. "Exploring EEG for object detection and retrieval." *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015.

McGuinness, Kevin, Eva Mohedano, Amaia Salvador, Zhenxing Zhang, Mark Marsden, Peng Wang, Iveel Jargalsaikhan et al. "Insight dcu at trecvid 2015." In *TRECVID 2015 Overview Papers and Slides*, pp. 1-16. 2015.

McGuinness, Kevin, Eva Mohedano, ZhenXing Zhang, Feiyan Hu, Rami Abatal, Cathal Gurrin, Noel O'Connor et al. "Insight Centre for Data Analytics (DCU) at TRECVid 2014: instance search and semantic indexing tasks." In *2014 TREC Video Retrieval Evaluation Notebook Papers and Slides*. 2014.

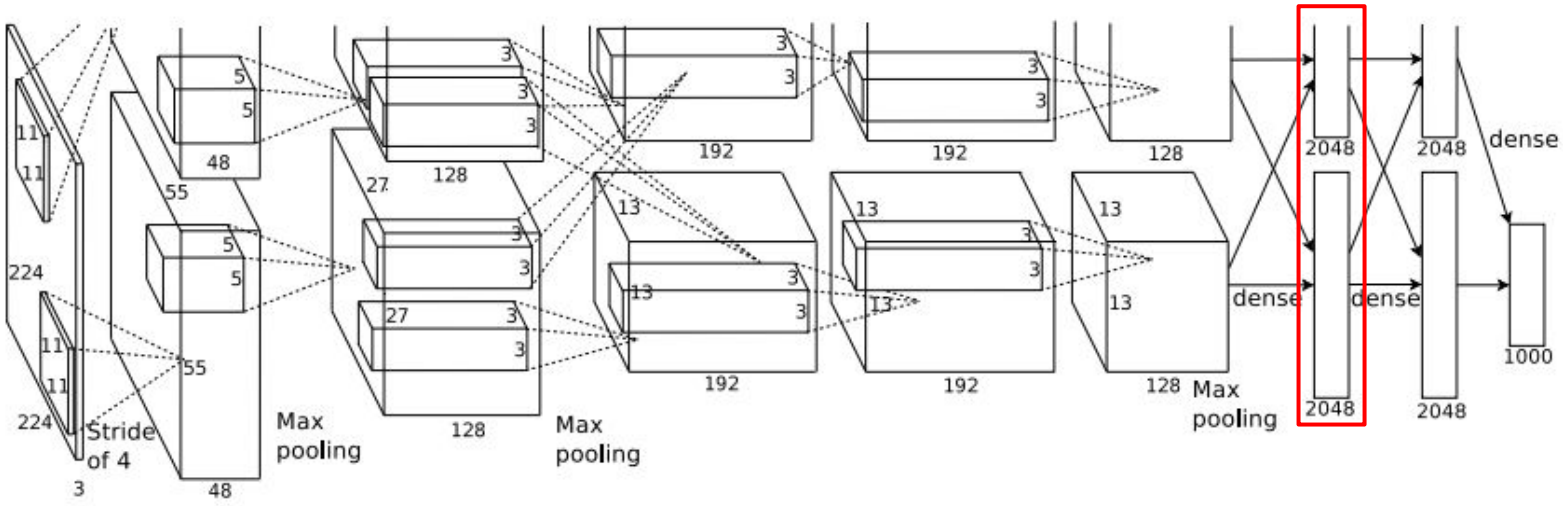
Mohedano, Eva, Graham Healy, Kevin McGuinness, Xavier Giró-i-Nieto, Noel E. O'Connor, and Alan F. Smeaton. "Object segmentation in images using eeg signals." In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 417-426. ACM, 2014.

Convolutional Neural Networks



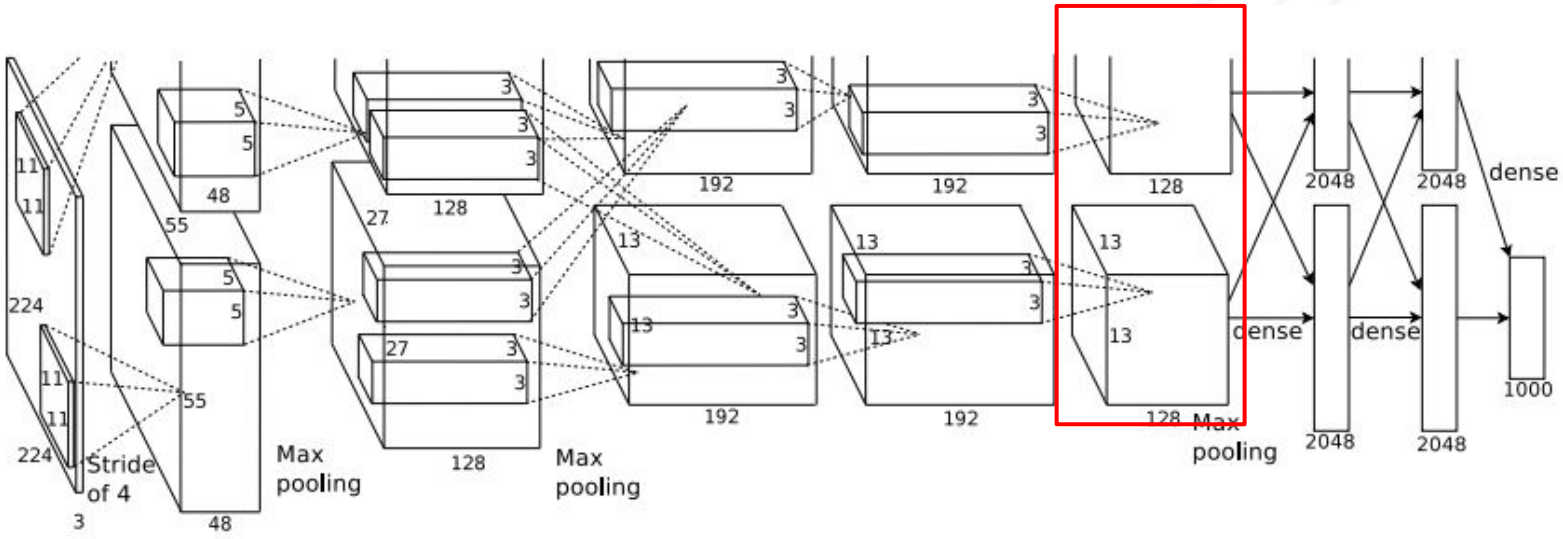
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

Image Representation



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

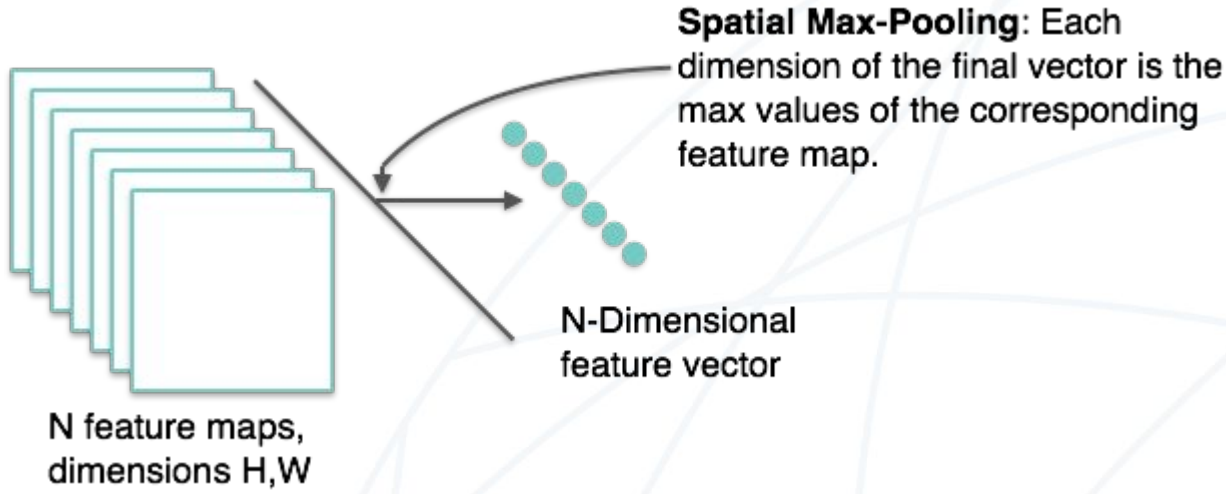
Image Representation



Babenko, A., & Lempitsky, V. Aggregating local deep features for image retrieval. ICCV 2015
Tolias, G., Sicre, R., & Jégou, H. Particular object retrieval with integral max-pooling of CNN activations. ICLR 2015
Kalantidis, Y., Mellina, C., & Osindero, S. Cross-dimensional Weighting for Aggregated Deep Convolutional Features. ECCV2016

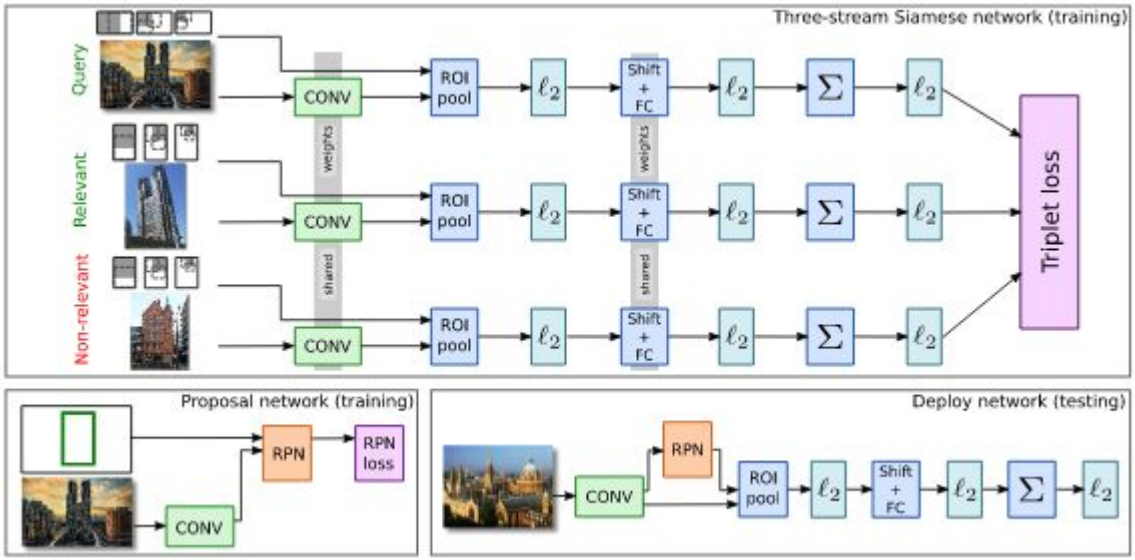
Image Representation

Descriptors from convolutional layers



Keeping the spatial layout

Learning Image Representations



Gordo, A., et al. "Deep image retrieval: Learning global representations for image search." *European Conference on Computer Vision*. Springer International Publishing, 2016.

State-of-the-art

	Method	Dim	Oxford5k	Oxford105k	Paris6k	Paris106k	Holidays
Off-the-shelf Fully connected	Neural Codes [10]	128	0.433	0.386			
	CNNastounding [23]	4-15k	0.68		0.79		
	MOP [76]	2048					0.802
Off-the-shelf Convolutional	SPoC [18]	256	0.589	0.578			0.802
	Ng *et al* [83]	128	0.593		0.590		0.816
	Razavian [78]	32k	**0.843**		**0.879**		**0.896**
	R-MAC [19]	512	0.669(0.773)	0.616(0.732)	0.830(0.865)	0.757(0.798)	
	CAM [87]	512	0.712(0.801)	0.672(0.769)	0.805(0.855)	0.733(0.800)	
	CroW [20]	512	0.708(0.749)	0.653(0.706)	0.797(0.848)	0.722(0.794)	0.851
End-to-End Training	Neural Codes [10]	128	0.557	0.523			
	Wan [103]	4096	0.783		0.947		
	Faster-RCNN [85]	512	0.710(0.786)		0.798(0.842)		
	NetVLAD [22]	256	0.635		0.735		0.843
	MAC [21]	512	0.800(0.854)	0.751(0.823)	0.829(0.870)	0.753(0.796)	0.795
	R-MAC [11]	512	0.831(0.894)	0.786(0.873)	0.871(0.912)	0.797(0.868)	0.891
	R-MAC (ResNet101) [11]	2048	**0.845(0.890)**	**0.816(0.878)**	**0.912(0.938)**	**0.863(0.906)**	**0.960**
Hand-crafted methods	BoW(1M)+QE [73]		0.827	0.767	0.805	0.710	
	Arandjelović [15]		**0.929**	**0.891**	**0.910**		