Deep Learning Compact and Invariant Image Descriptors for Instance Retrieval

Olivier Morère

08/06/2016

About this thesis

Antoine VEILLARD

Daniel RACOCEANU

Vijay CHANDRASEKHAR

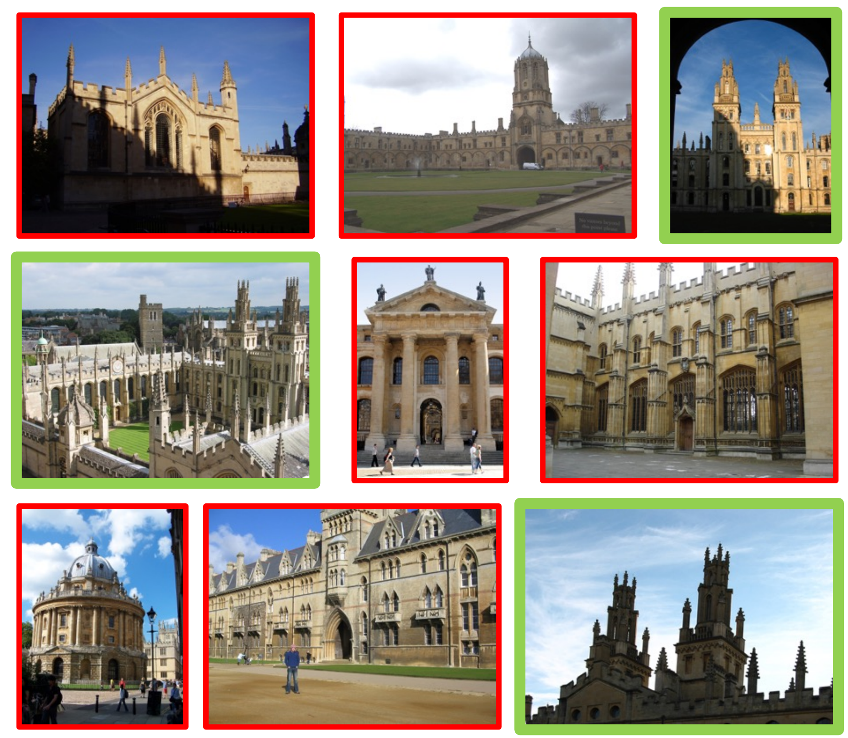
Hanlin GOH

Image Instance Retrieval

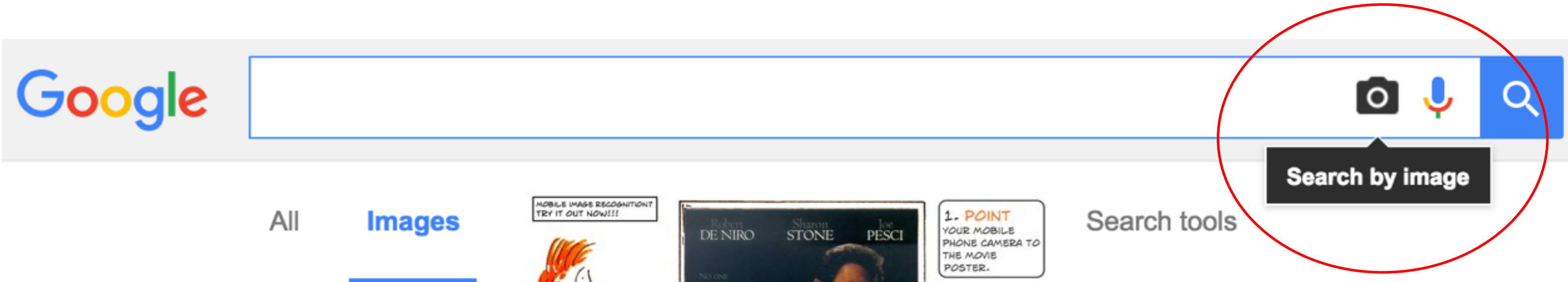
New query image

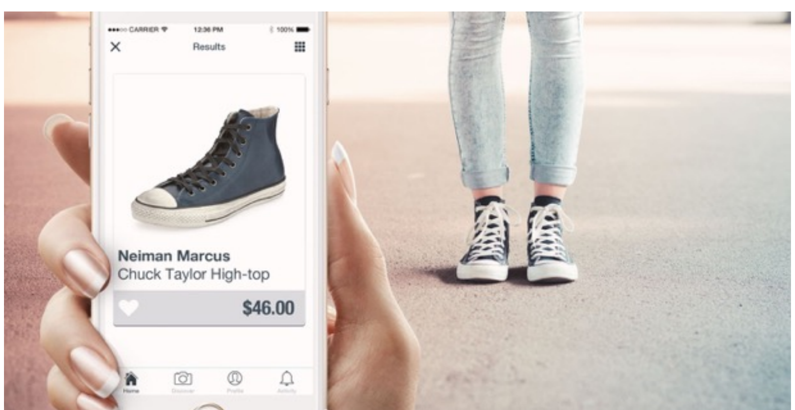
Retrieve database images
depicting the same object

Existing image database

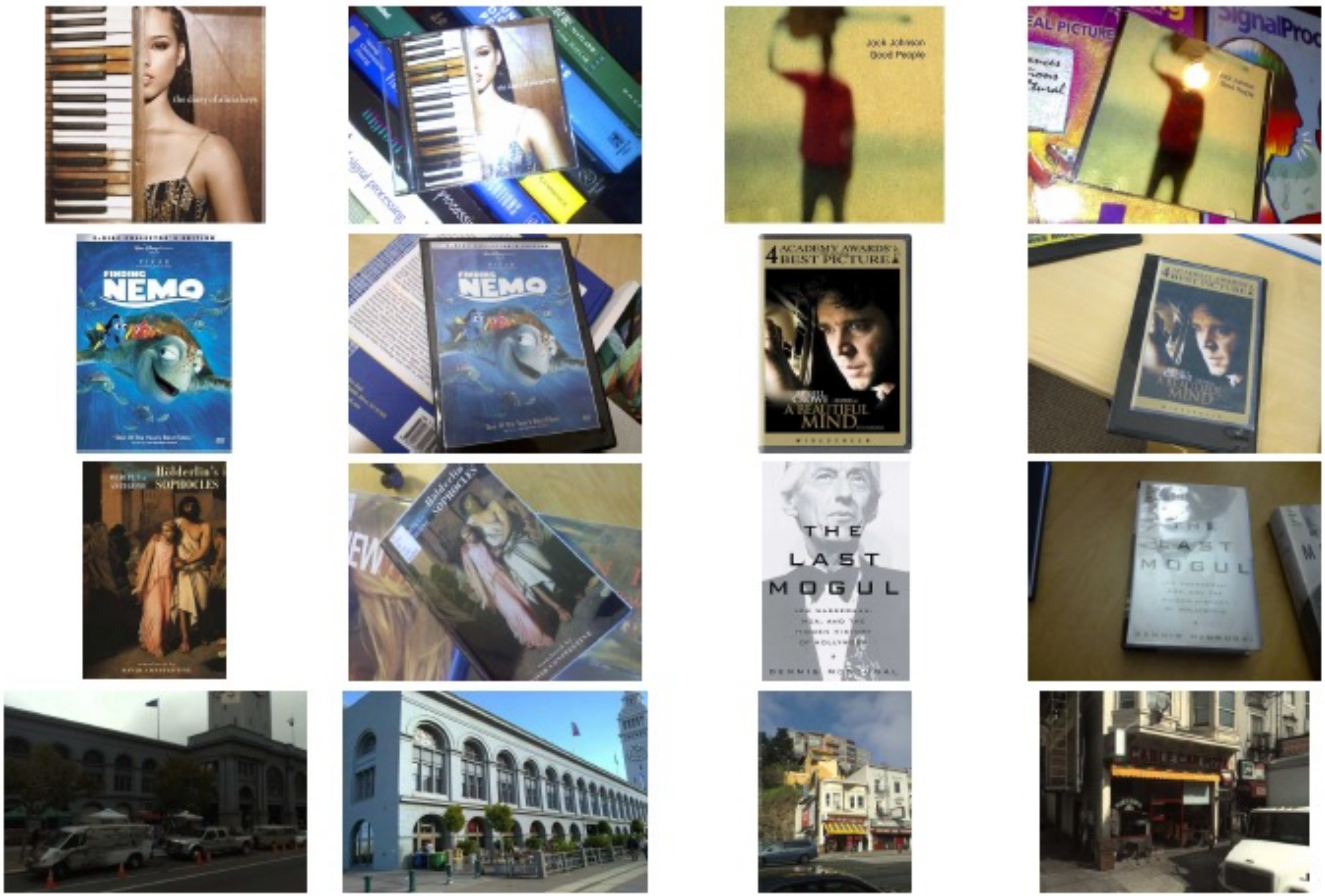


A Wide Range of Applications

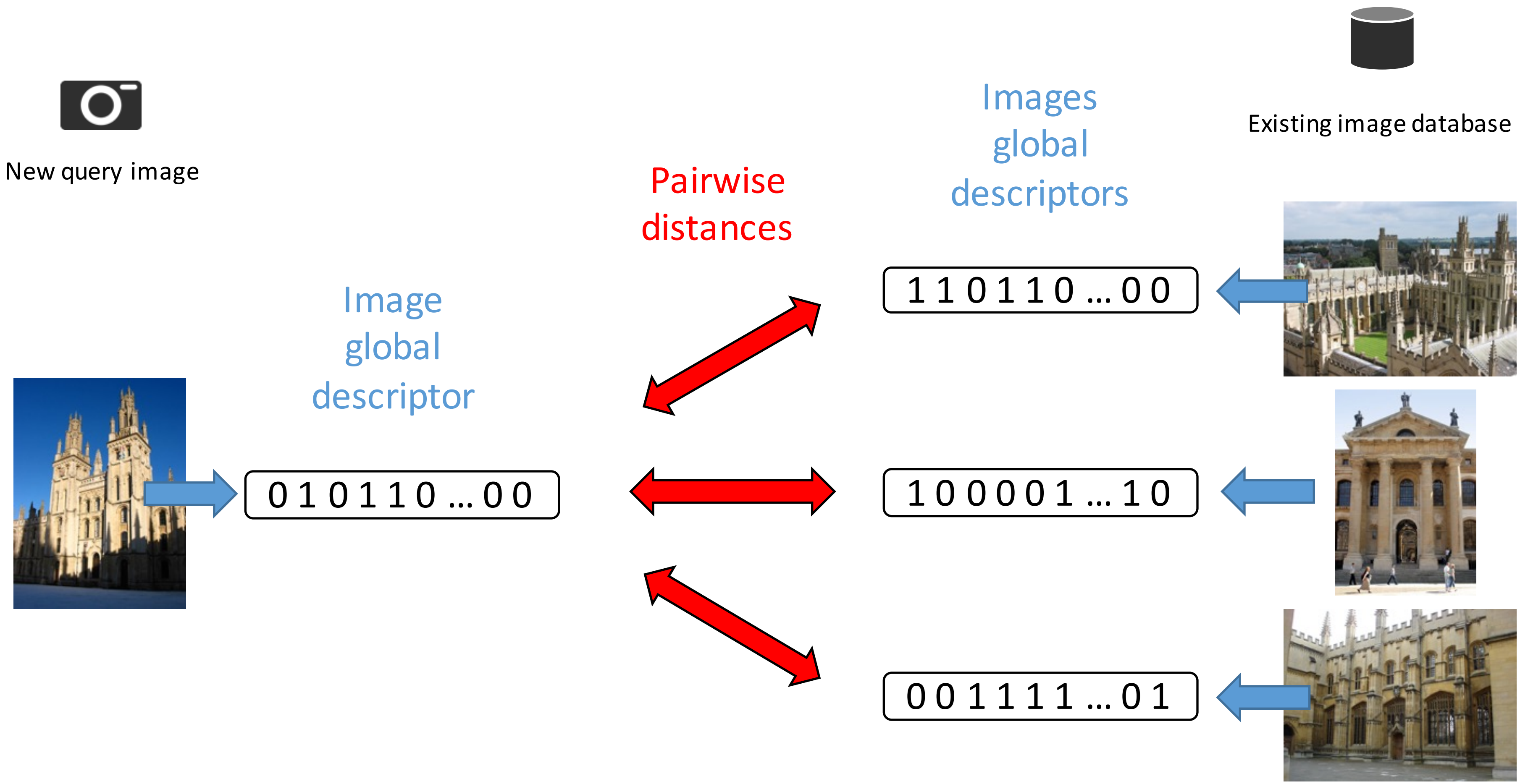




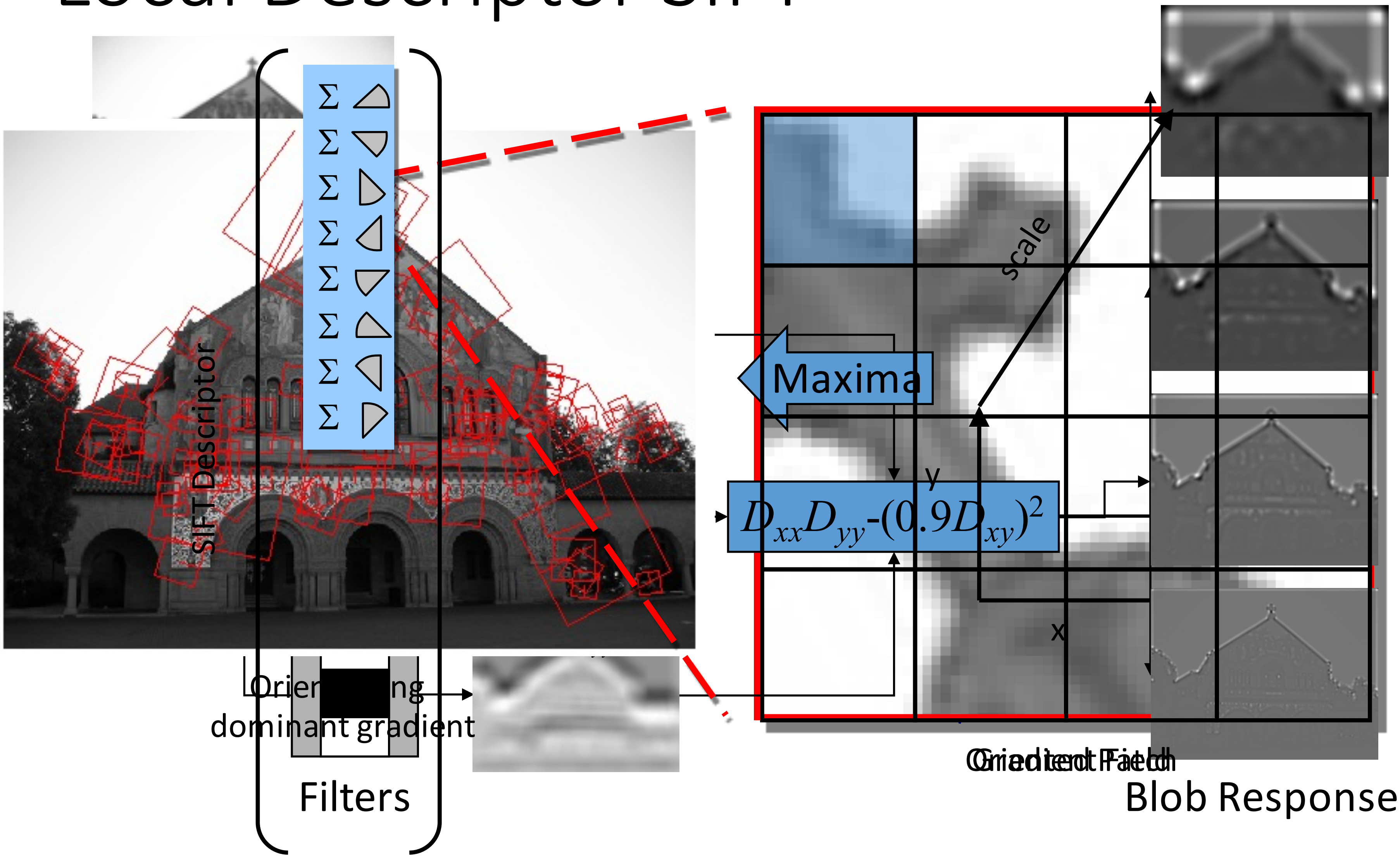
A Challenging Problem



Comparing Image Global Descriptors for Retrieval

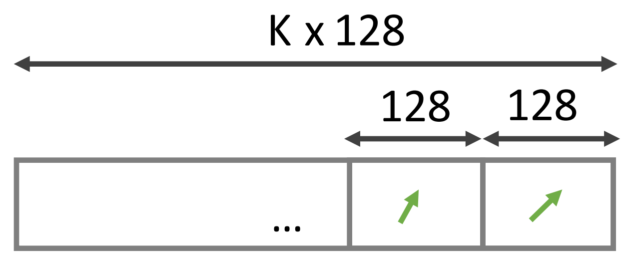


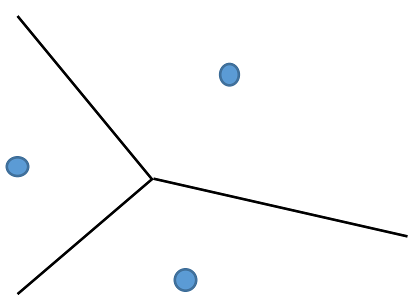
Local Descriptor SIFT

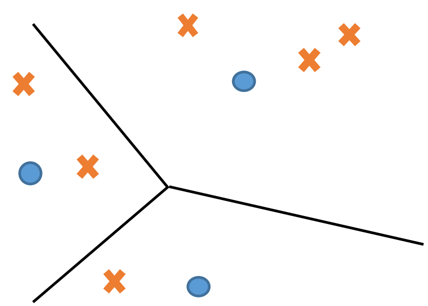


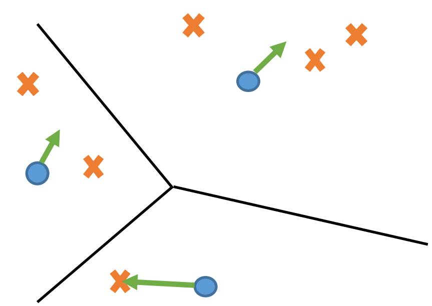
Global Descriptor: VLAD/FV

Step 4
Concatenate residuals into the global descriptor







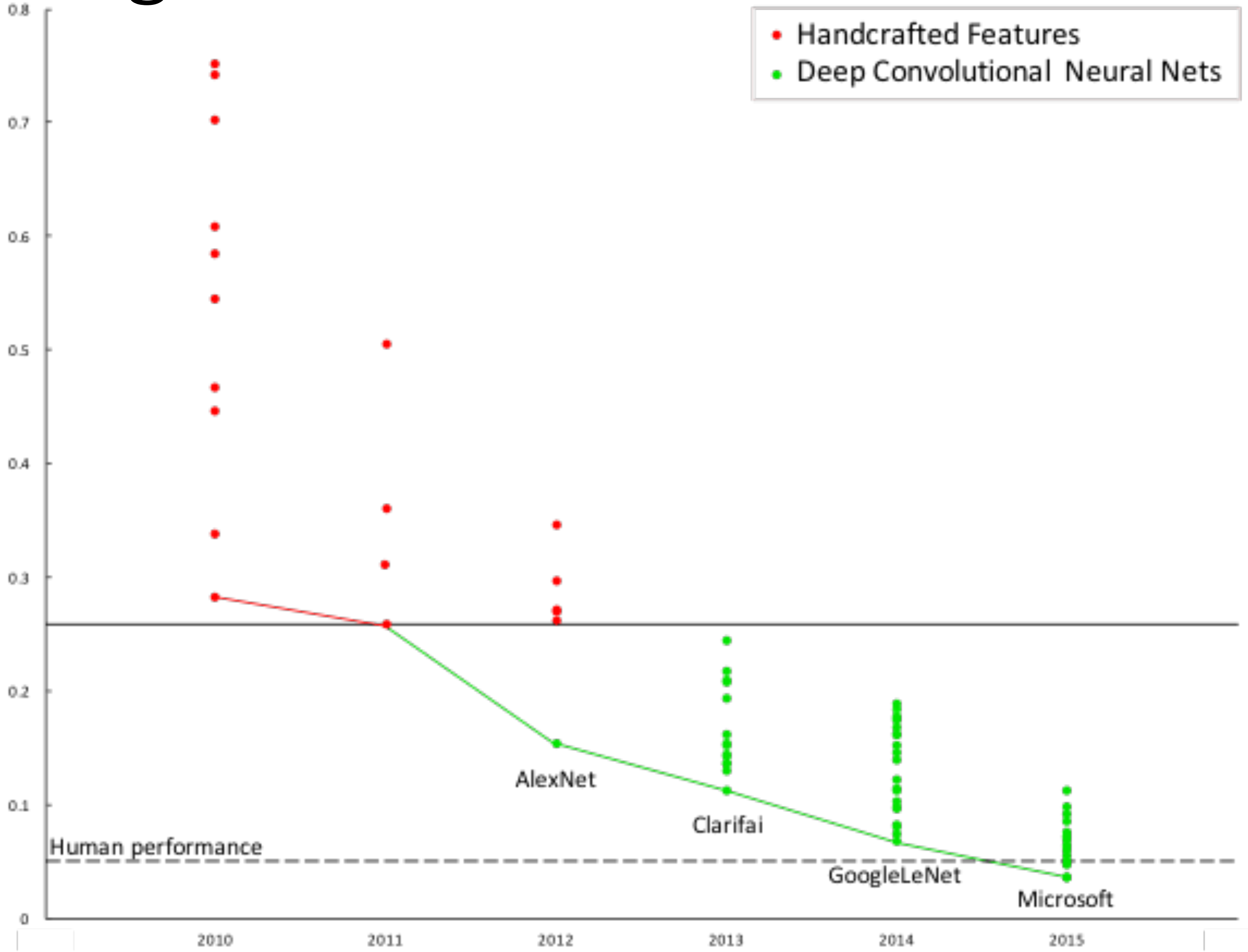


Step 1
Training K clusters of local descriptors on the image training set

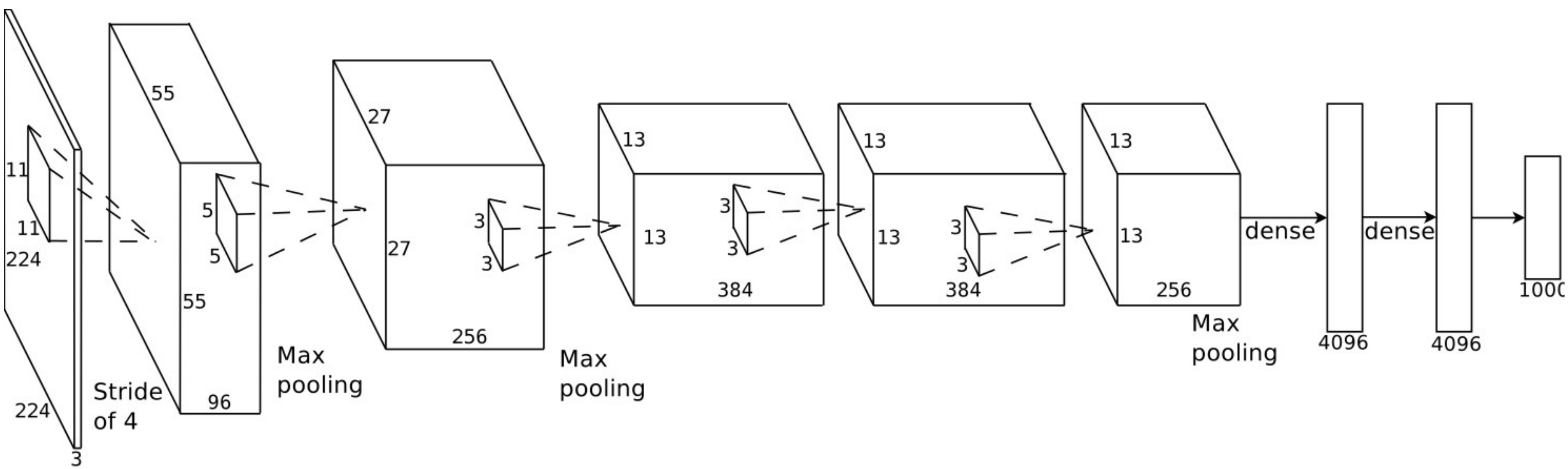
Step 2
Extract local descriptors from new image

Step 3
Compute local descriptor residual statistics in each bin

ImageNet Results over the Years

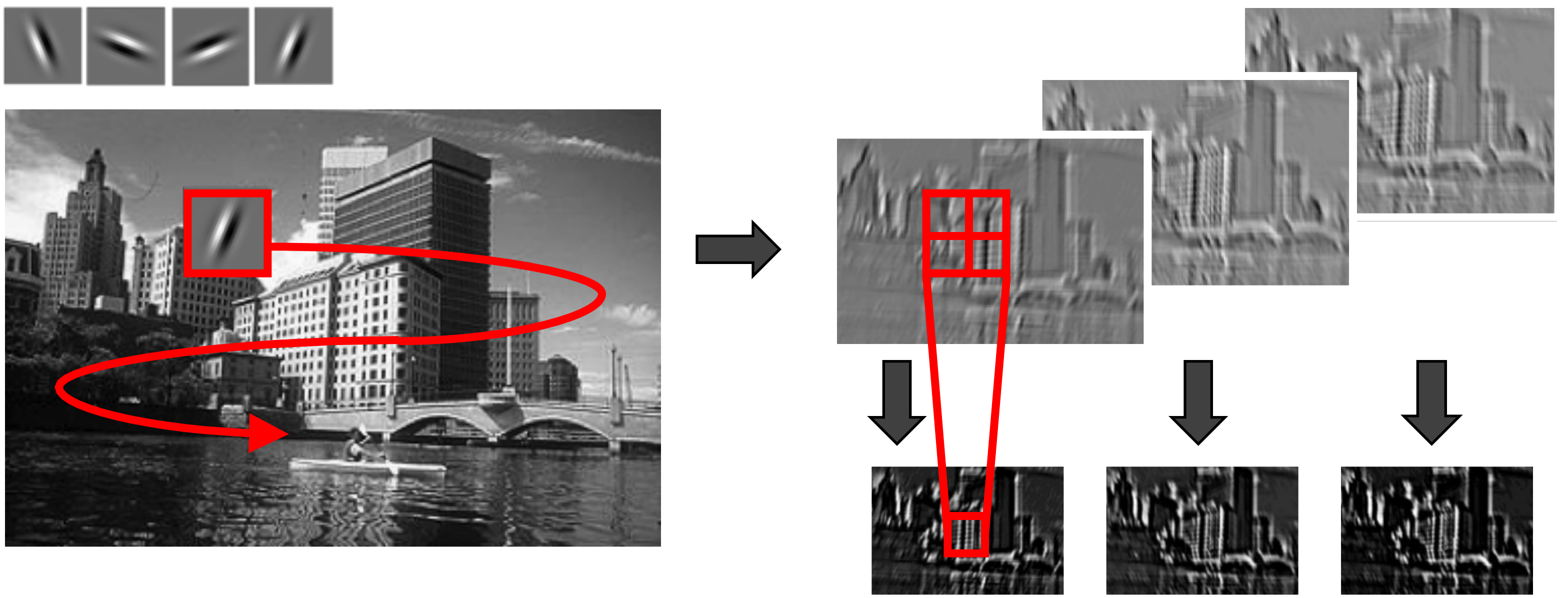


Deep Convolutional Neural Networks (CNN)



[Krizhevsky, 2012]

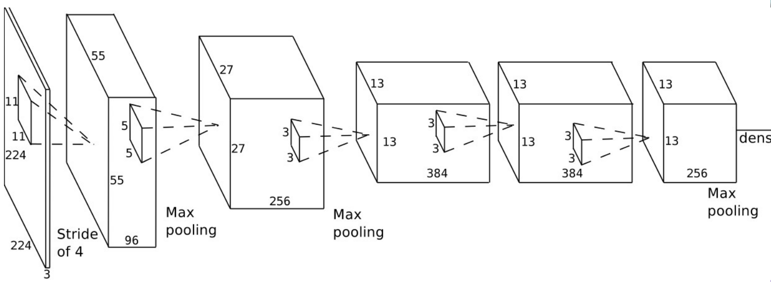
Convolution & Pooling



- Convolution
  - Local connectivity
  - Stationarity of the signal
- Pooling
  - Dimensionality reduction
  - Local invariance

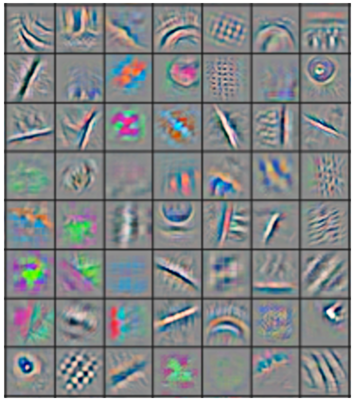
Learning Abstract Visual Representations

[Zeiler, 2013]

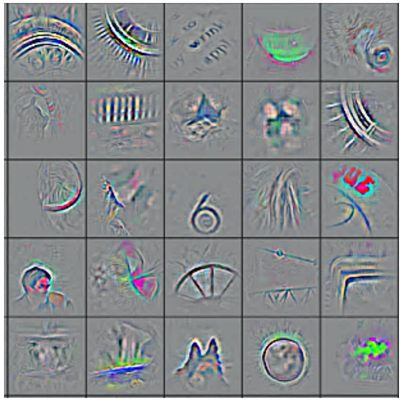


Layer 1

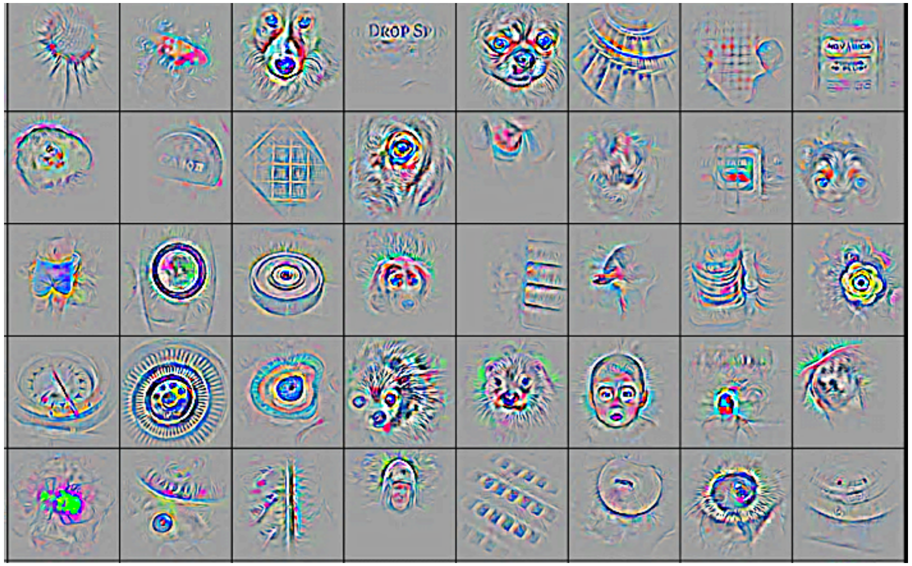
Layer 2



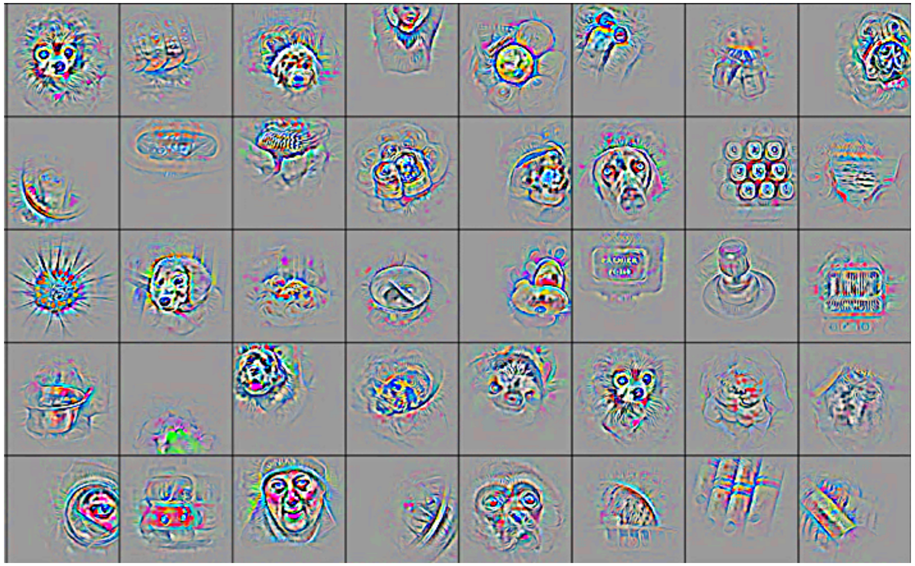
Layer 3



Layer 4



Layer 5



Original Contributions

- Image classification with Convolutional Neural Networks (CNN), ImageNet2014
- Thorough comparison study of FV and CNN for image instance retrieval
- Hashing CNN descriptors
  - Unsupervised dimensionality reduction
  - Descriptor finetuning with unsupervised and semi-supervised metric learning
- Robust CNN descriptors with i-theory

List of Publications (1/2)

Conference papers

- O. Morère, J. Lin, A. Veillard, V. Chandrasekhar, T. Poggio. Nested Invariance Pooling and RBM Hashing for Image Instance Retrieval. *Submitted to European Conference on Computer Vision (ECCV) 2016*
- O. Morère, J. Lin, V. Chandrasekhar, A. Veillard, H. Goh. Co-Sparsity Regularized Deep Hashing for Image Instance Retrieval. *Accepted in International Conference on Image Processing (ICIP) 2016*
- O. Morère, J. Lin, J. Petta, V. Chandrasekhar and A. Veillard Tiny descriptors for image retrieval with unsupervised triplet hashing. *Data Compression Conference (DCC) 2016.*
- C. Dao-Duc, H. Xiaohui and O. Morère. Maritime Vessel Images Classification Using Deep Convolutional Neural Networks. *Symposium on Information and Communication Technology (SOICT) 2015*
- V. Chandrasekhar, J. Lin, O. Morère, A. Veillard and H. Goh. Compact Global Descriptors for Visual Search. *Data Compression Conference (DCC) 2015*

List of Publications (2/2)

Technical reports

- O. Morère, A. Veillard, J. Lin, J. Petta, V. Chandrasekhar and T. Poggio. Group Invariant Deep Representations for Image Instance Retrieval.
  Center for Brains, Minds and Machines (CBMM) 2016
- O. Morère, J. Lin, V. Chandrasekhar, A. Veillard and H. Goh. Deephash: Getting Regularization, Depth and Fine-tuning Right.
  arXiv preprint arXiv:1501.04711 2015

Contests and workshops

- O. Morère, A. Veillard, H. Goh. Team "LateFusion".
  Kaggle National Data Science Bowl Challenge 2015
- O. Morère, H. Goh, A. Veillard, V. Chandrasekhar. Large Scale Image Classification on a Shoe String. *ImageNet Large Scale Visual Recognition Challenge, European Conference on Computer Vision (ECCV) 2014*

Journal articles

- O. Morère, V. Chandrasekhar, J. Lin, H. Goh and A. Veillard. A Practical Guide to CNNs and Fisher Vectors for Image Instance Retrieval.
  Accepted in Signal Processing (SIGPRO) 2016

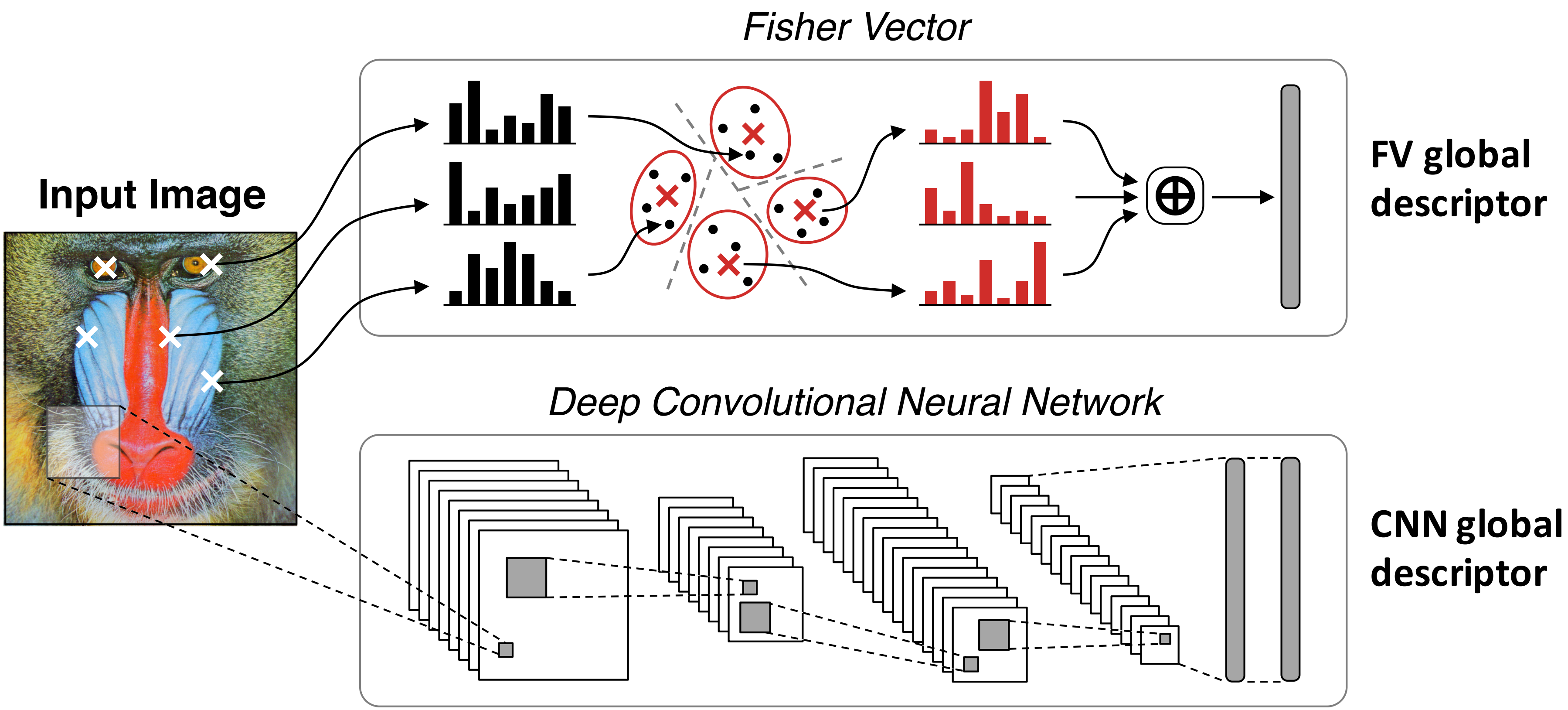
PART 1

1. Thorough comparison study of FV and CNN for image instance retrieval

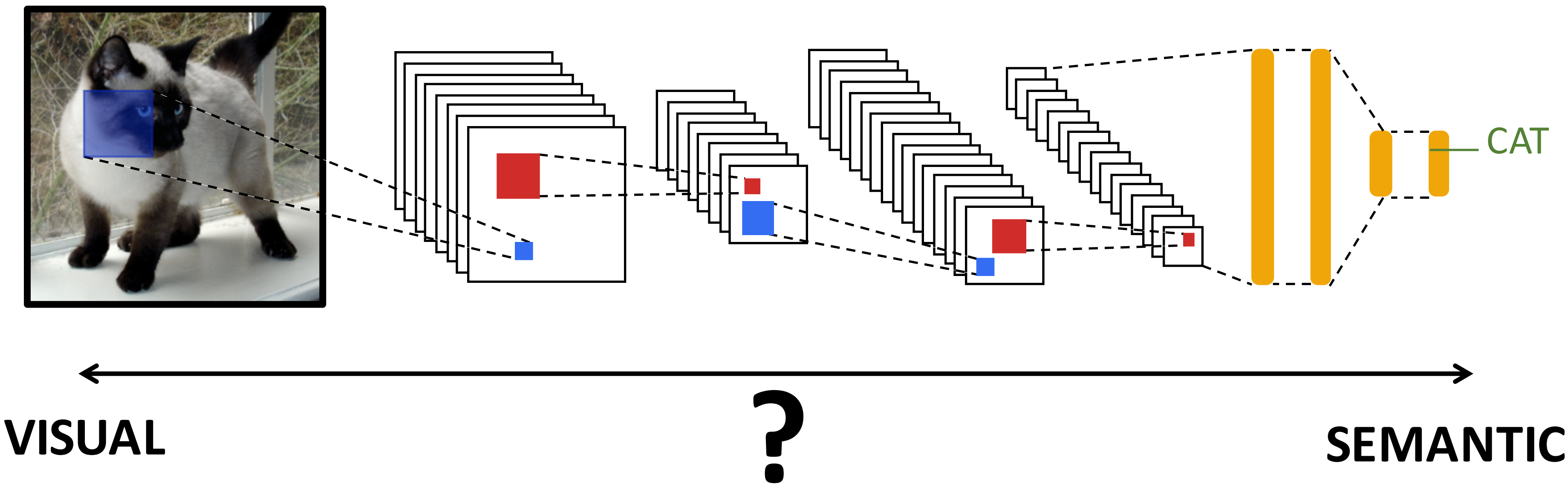
2. Hashing CNN descriptors

3. Robust CNN descriptors with i-theory

CNN vs Fisher Vector



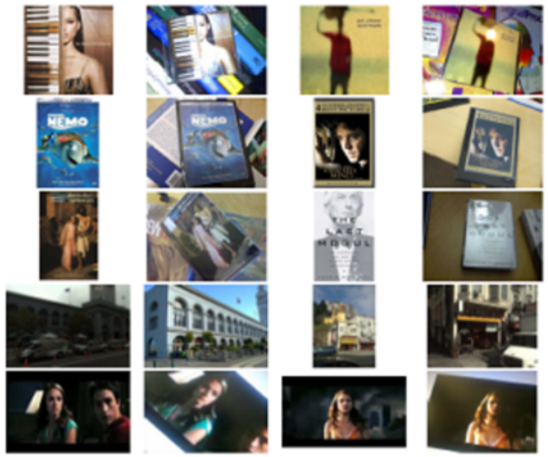
How to Extract CNN Descriptors for Image Instance Retrieval ?

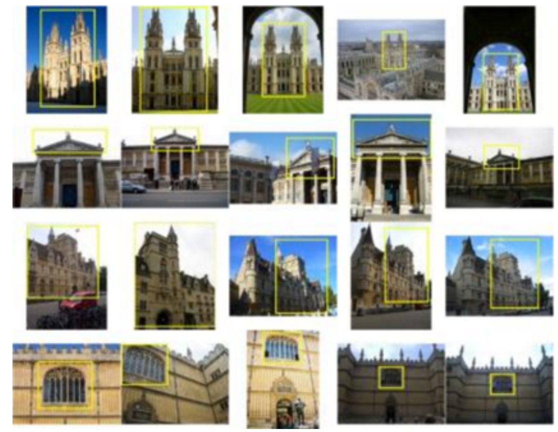


	Architecture			Training		
	parameters	depth (conv+fc)	input size	training set	classes	data size
OxfordNet	138M	13+3	$224 \times 224 \times 3$	*ImageNet*	1000	1.2M
AlexNet	60M	5+3	$227 \times 227 \times 3$	*ImageNet*	1000	1.2M
PlacesNet	60M	5+3	$227 \times 227 \times 3$	Places-205	205	2.4M
HybridNet	60M	5+3	$227 \times 227 \times 3$	Both	1183	3.6M

Retrieval Data Sets

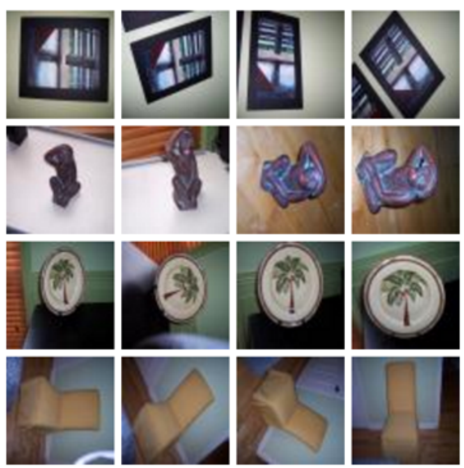
Stanford Mobile Visual Search
[Chandrasekhar, 2011]

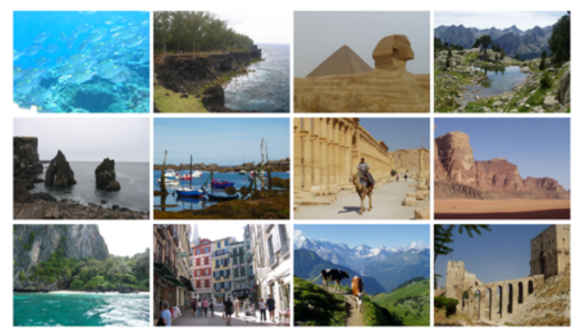




Oxford Buildings
[Philbin, 2007]

University of Kentucky Benchmark
[Nister, 2006]





INRIA Holidays
[Jégou, 2008]

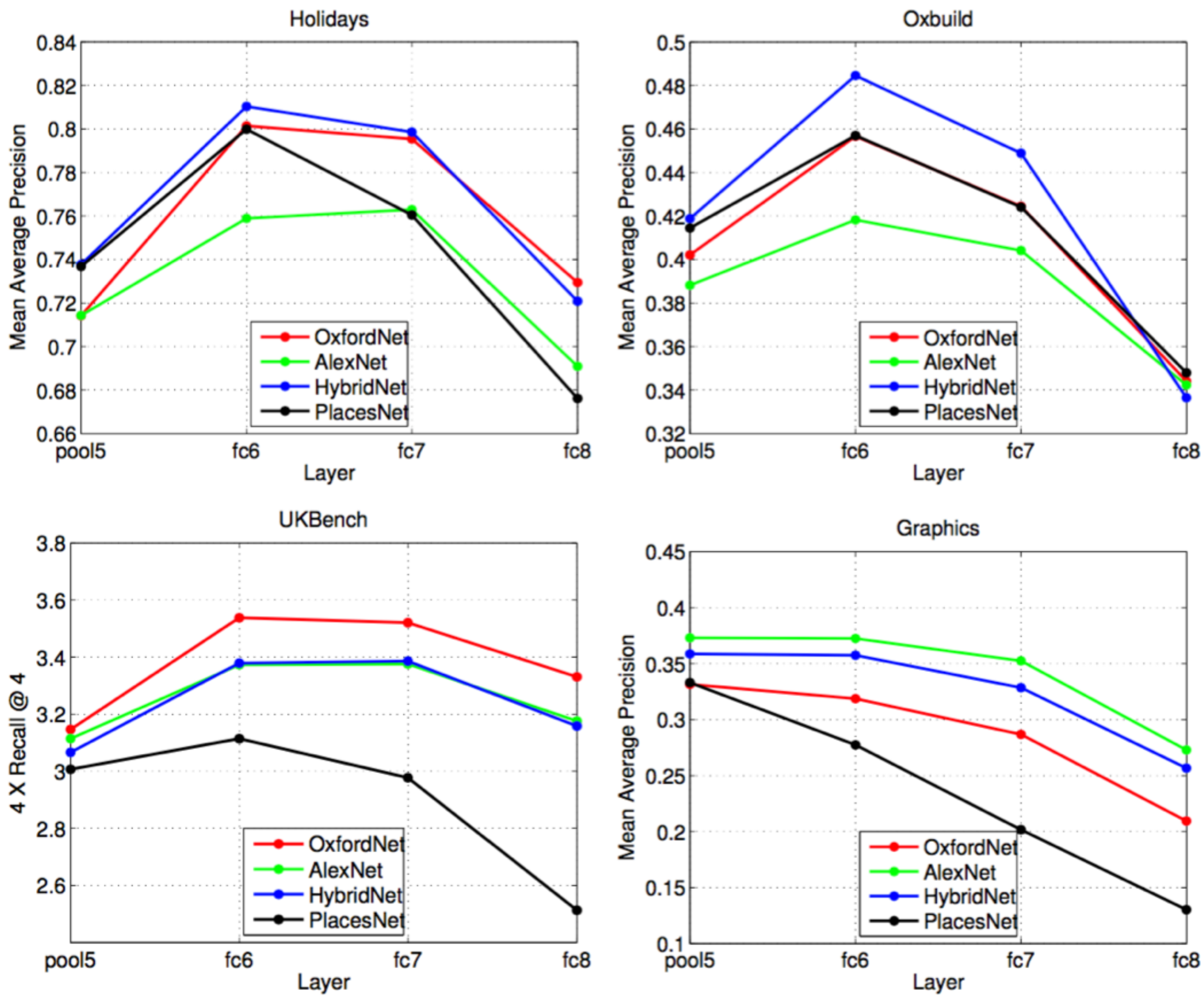
Object centric **Scene centric**

Evaluation Metrics

$$Recall = \frac{|\{RelevantIndividuals\} \cap \{RetrievedIndividuals\}|}{|RelevantIndividuals|}$$

$$AP = \frac{\sum_{k=1}^{n}(Precision(k) \times isRelevant(k))}{|RelevantIndividuals|}$$

Best Practices for CNN Descriptors

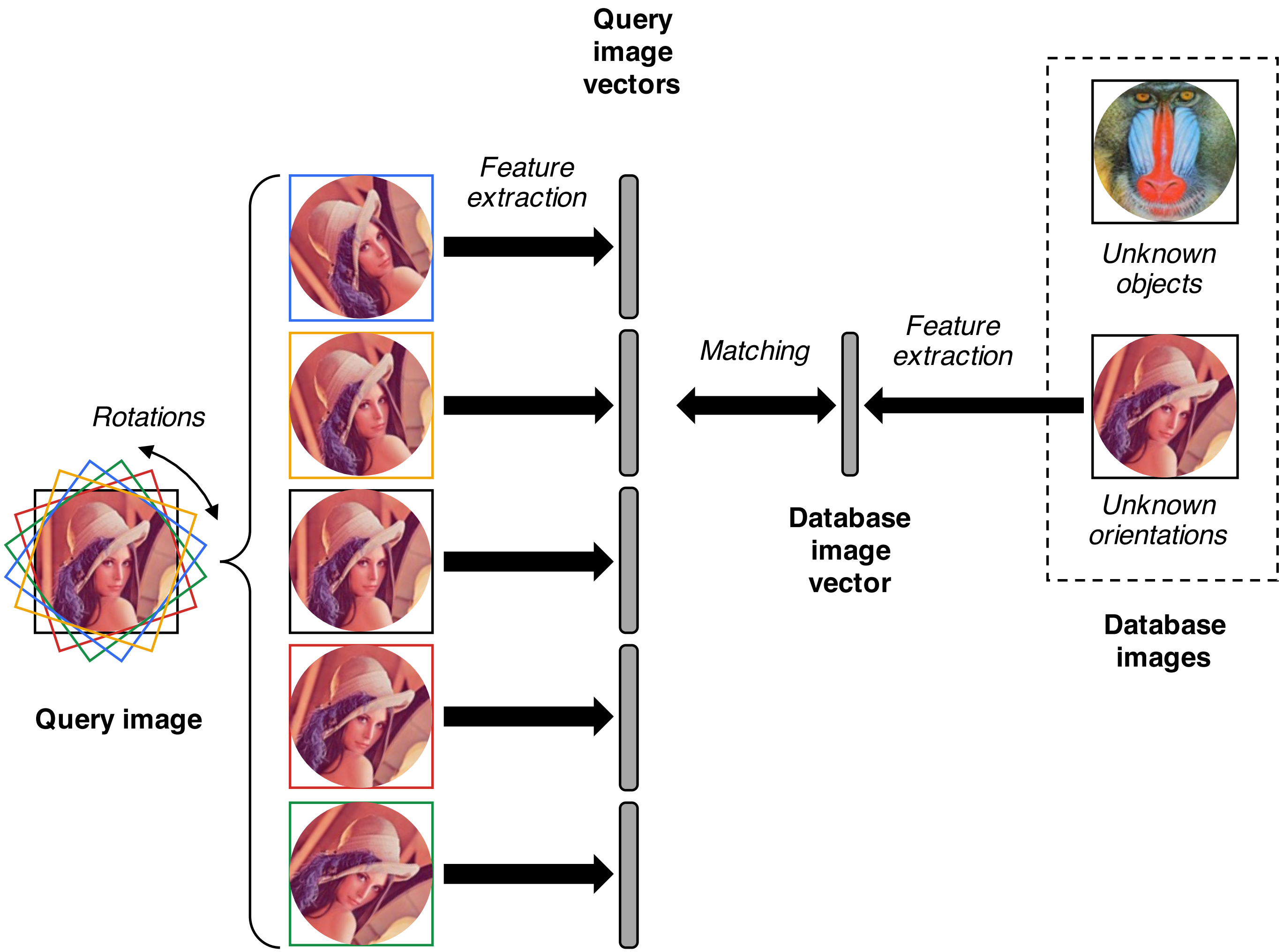


Best Practices for FV Interest Points

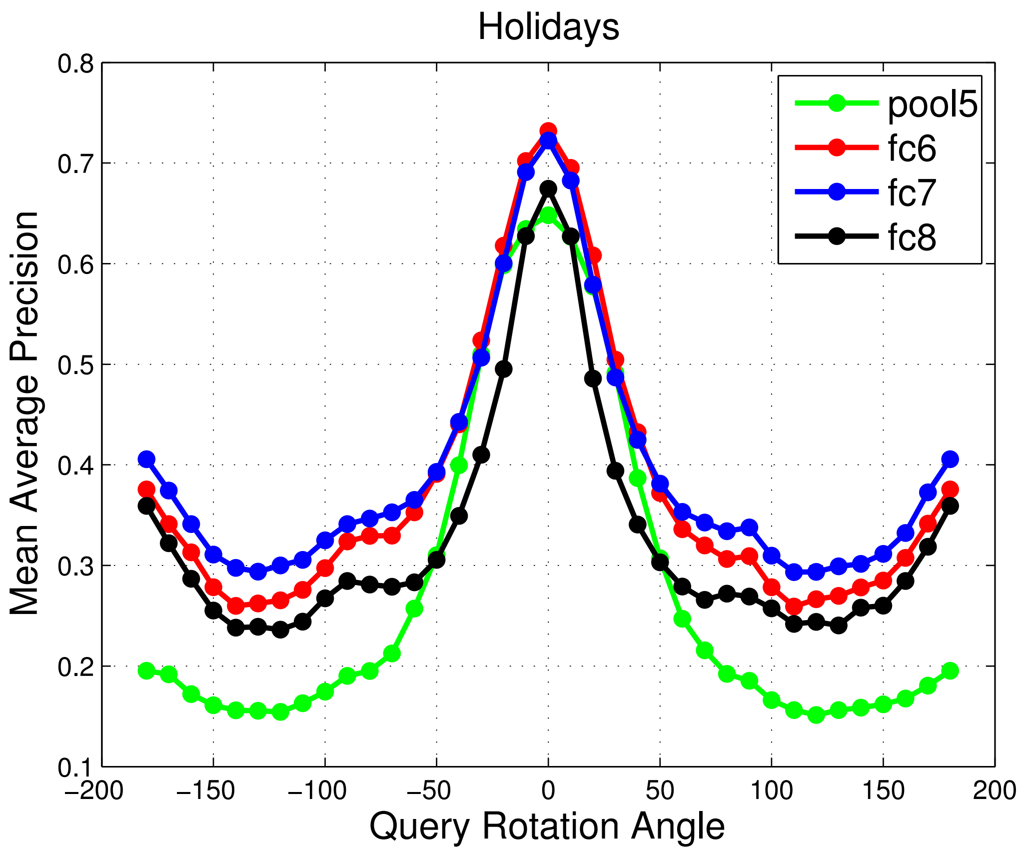
Descriptor	Dim	*Holidays*	UKBench	*Oxbuild*	Graphics
FVDoG	32768	0.63	2.8	0.42	0.66
FVDS	32768	0.73	2.38	0.51	0.20
FVDM	32768	0.75	2.45	0.55	0.32

- Multi-scale improves performance over single-scale
- DoG is required when there are big scale and rotation changes (e.g. Graphics)
- Dense works better with highly textured images (e.g. Holidays)

Impact of Rotation

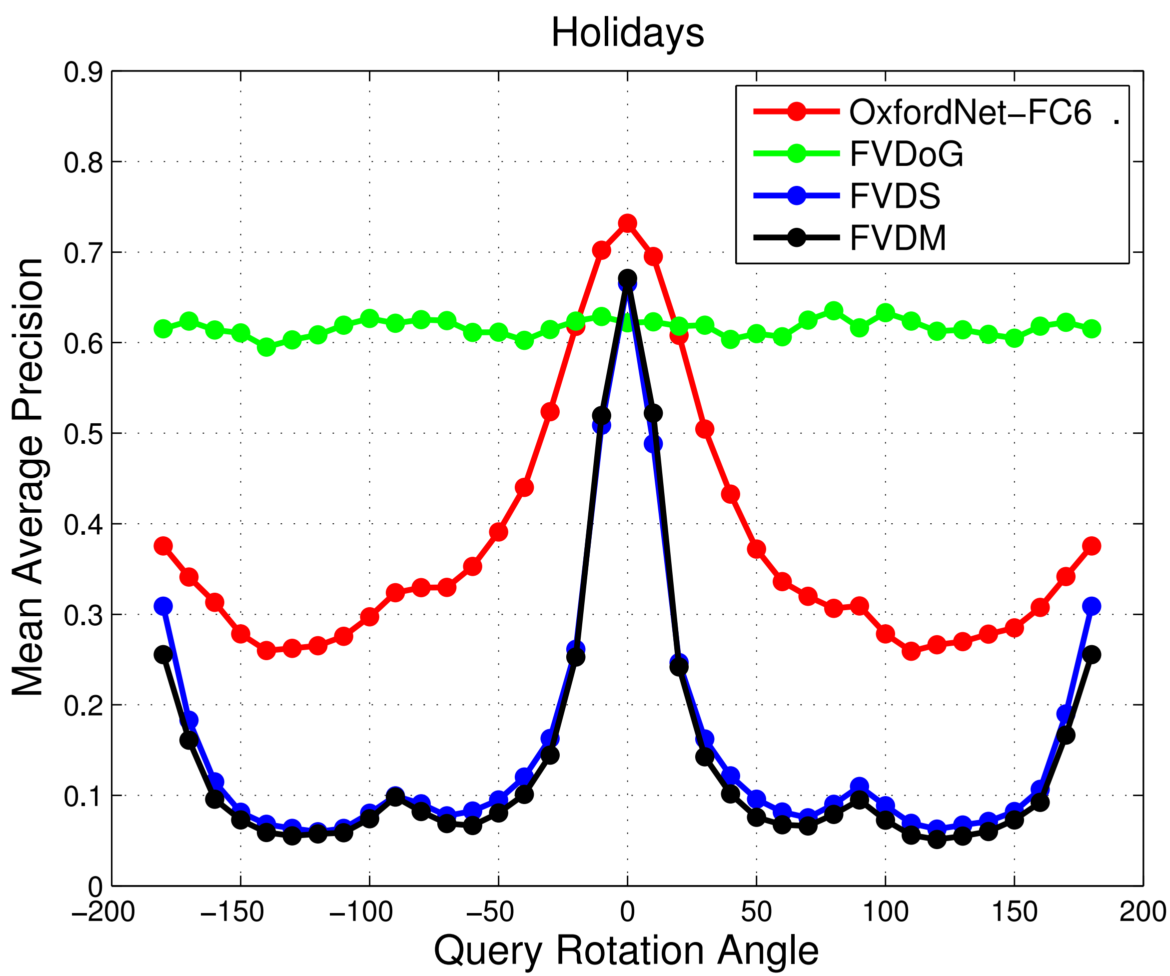


Impact of Rotation - CNN



- Very limited invariance to rotation
- Invariance to rotation does not increase with depth

Impact of Rotation - CNN vs FV



Further Readings

O. Morère, V. Chandrasekhar, J. Lin, H. Goh and A. Veillard. A Practical Guide to CNNs and Fisher Vectors for Image Instance Retrieval. Accepted in Signal Processing (SIGPRO) 2016

Part 1: Summary

- CNN performs well for image instance retrieval
- Two problems remain to be addressed:
  - Descriptor dimensionality
  - Lack of robustness

PART 2

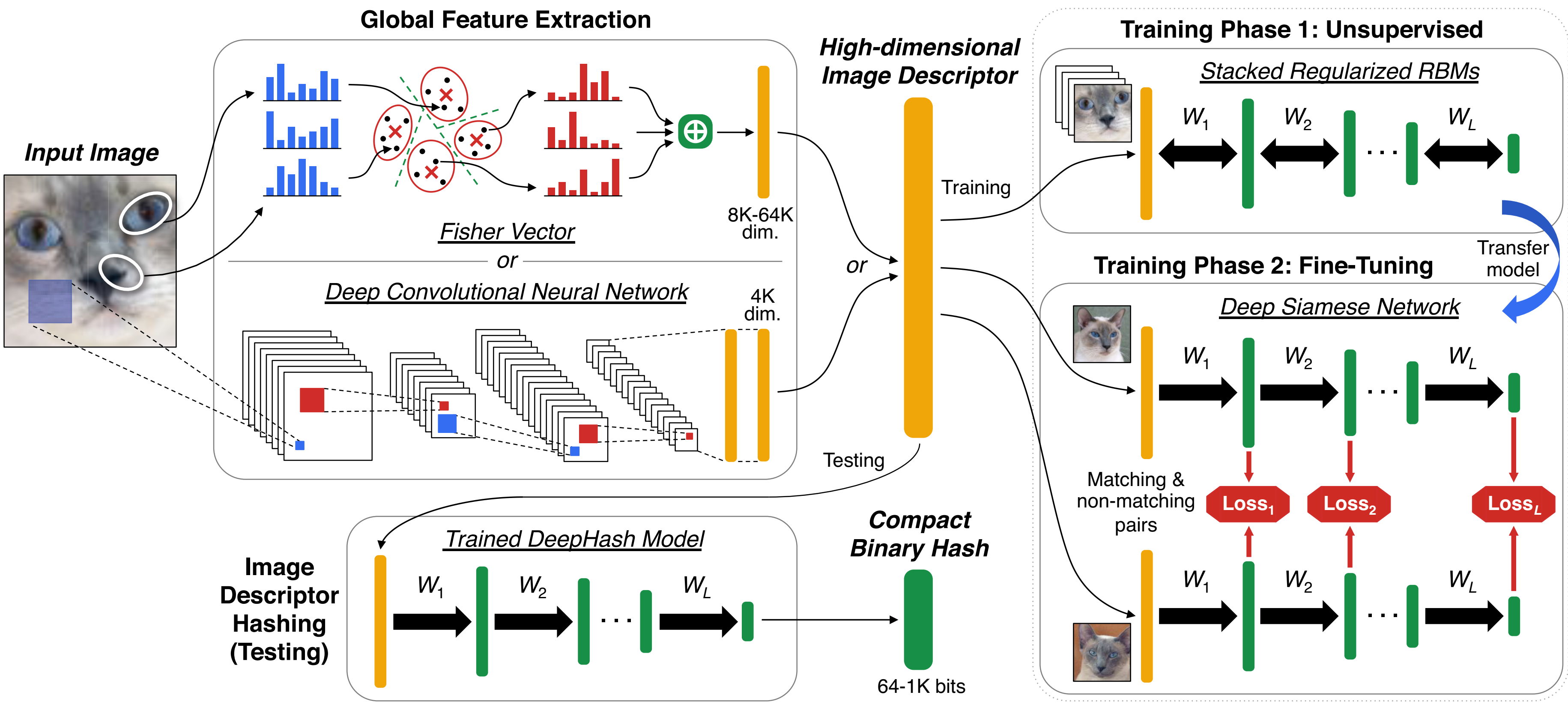
1. Thorough comparison study of FV and CNN for image instance retrieval
2. Hashing CNN descriptors
3. Robust CNN descriptors with i-theory

Why 64-bit Hash ?

- Motivation
  - Billions of images can be stored in RAM
  - Fast matching with ultra-fast Hamming distance

- Challenges
  - Global descriptors are very high dimensional
  - Uncompressed: 4K-25K floating point numbers

Hashing Outline



Hashing Outline

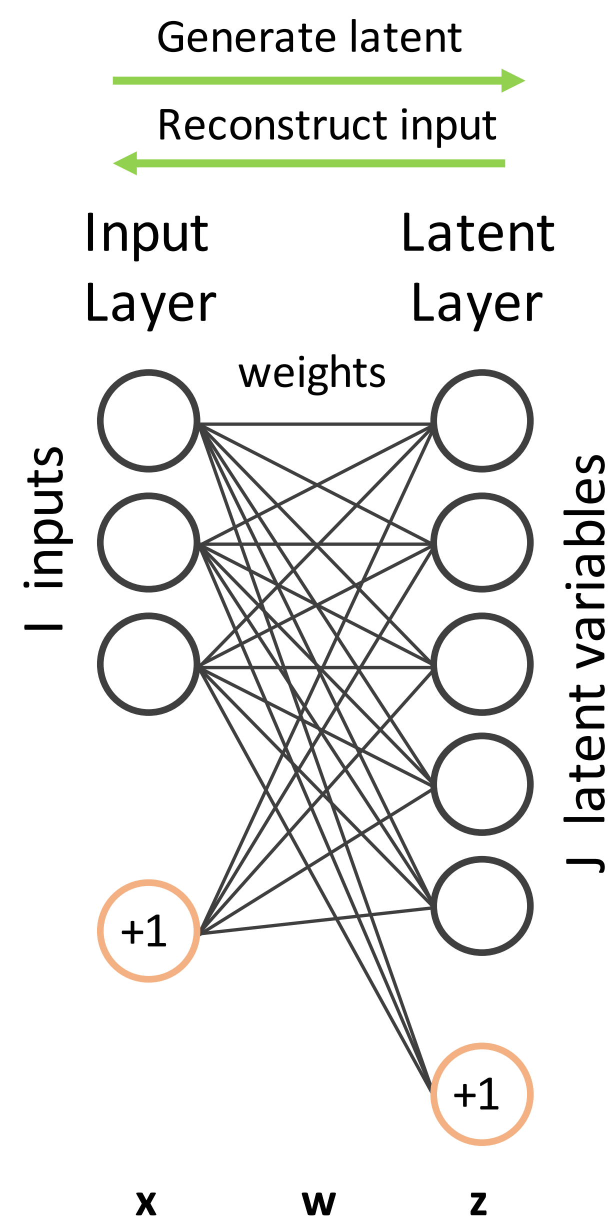
1. Dimensionality Reduction with Stacked RBM
2. Semi-supervised fine-tuning with Siamese networks
3. Unsupervised fine-tuning with triplet networks

RBM

- Bipartite graph model with input units and latent units
- Closed form expression from one to the other

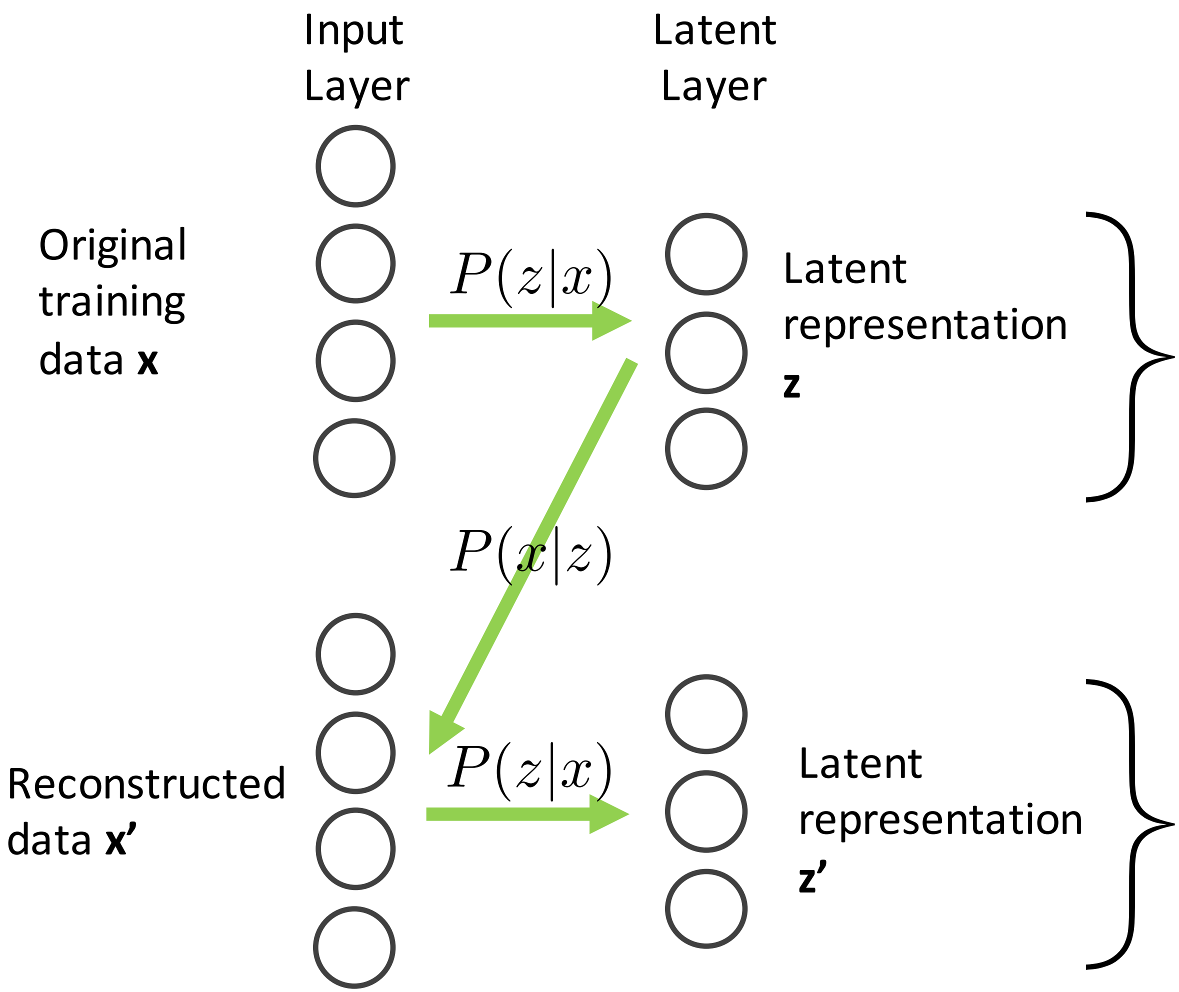
$$P(z_j = 1|x) = sigmoid(\sum_i w_{ij} x_i)$$

$$P(x_i = 1|z) = sigmoid(\sum_j w_{ij} z_j)$$



Contrastive Divergence

[Hinton, 2002]

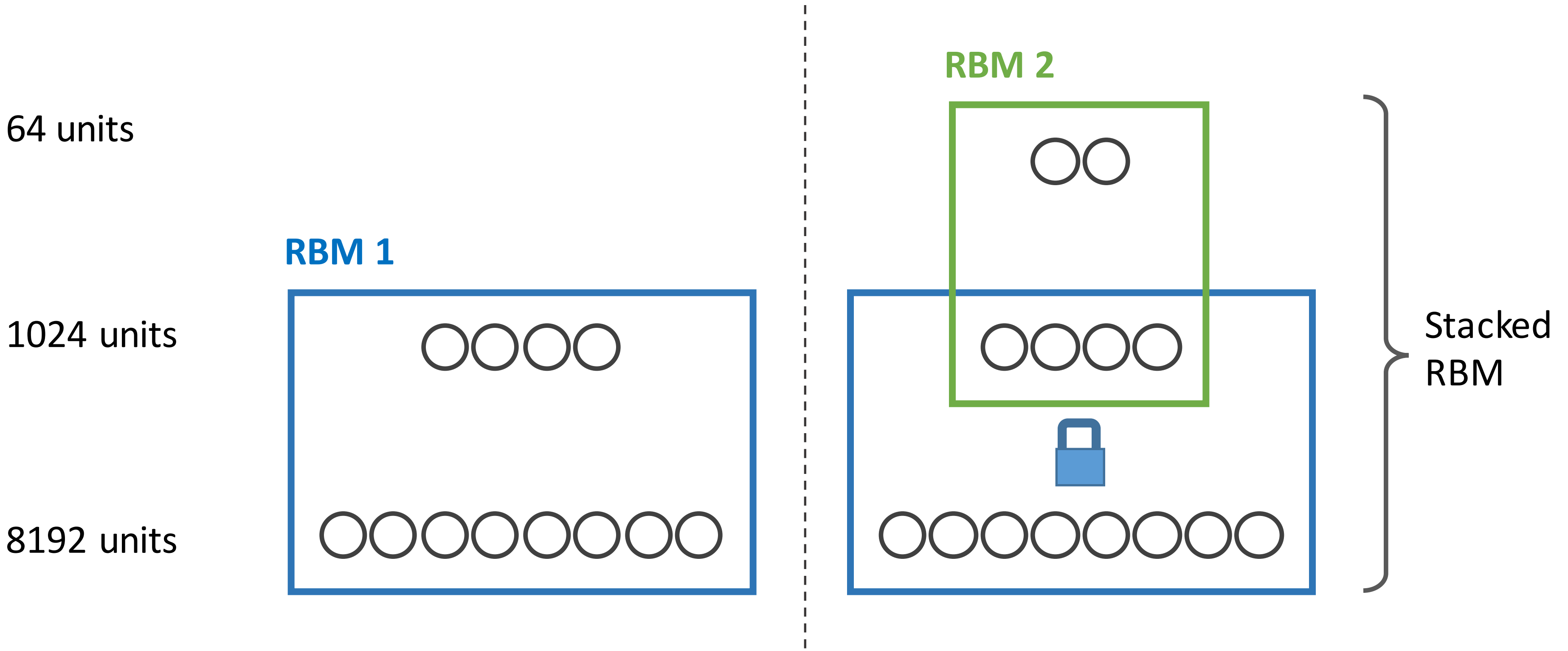


$$Positive_{ij} = x_i z_j$$

$$Negative_{ij} = x'_i z'_j$$

$$\Delta W = \epsilon(Positive - Negative)$$

Greedily Training Stacked RBM [Hinton, 2006 Bengio, 2006]



Step 1:
- Training RBM1

Step 2:
- Freezing RBM1 weights
- Training RBM2 using RBM1 latent layer as input layer

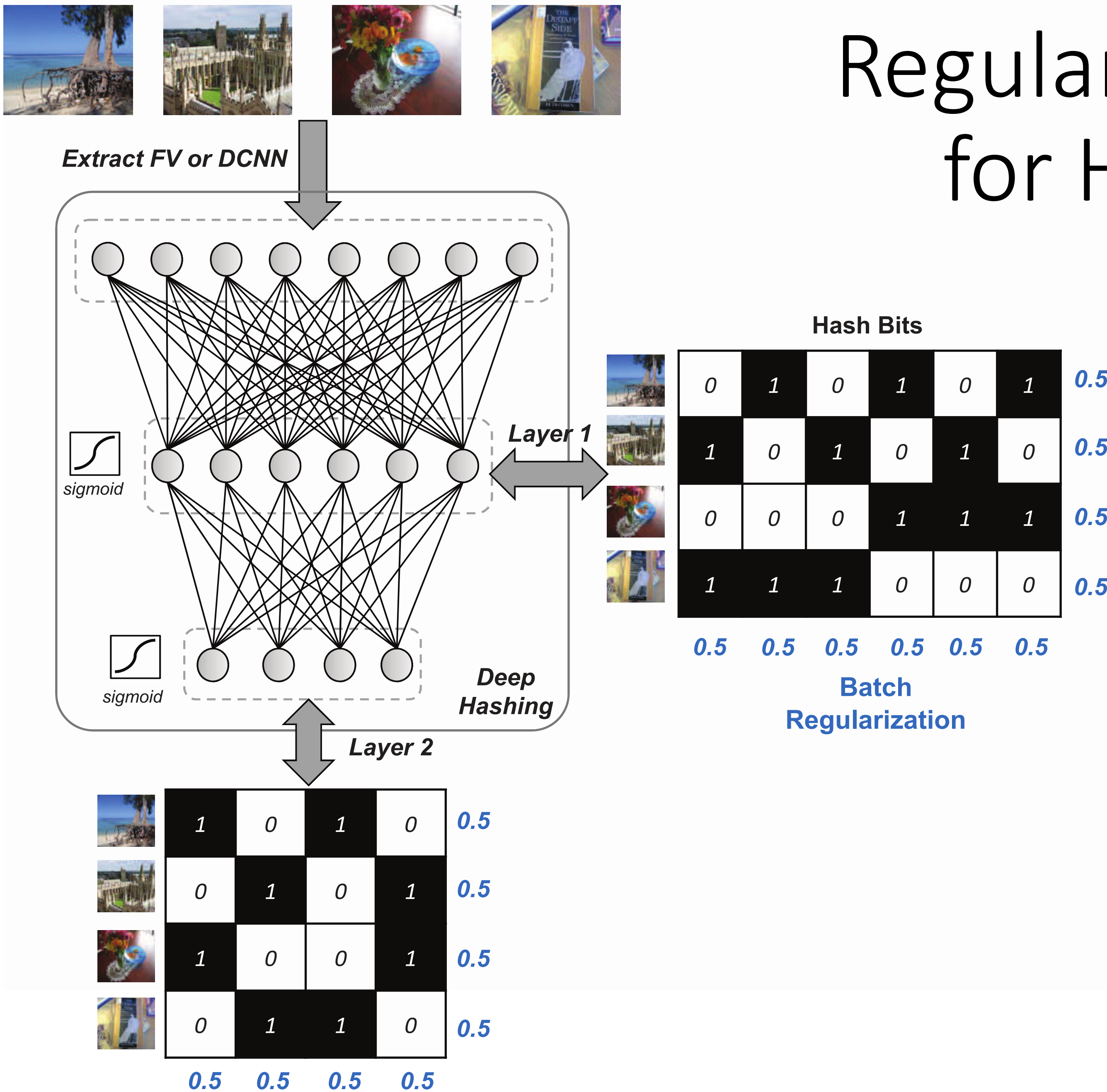
Regularization

[Hinton, 2009]

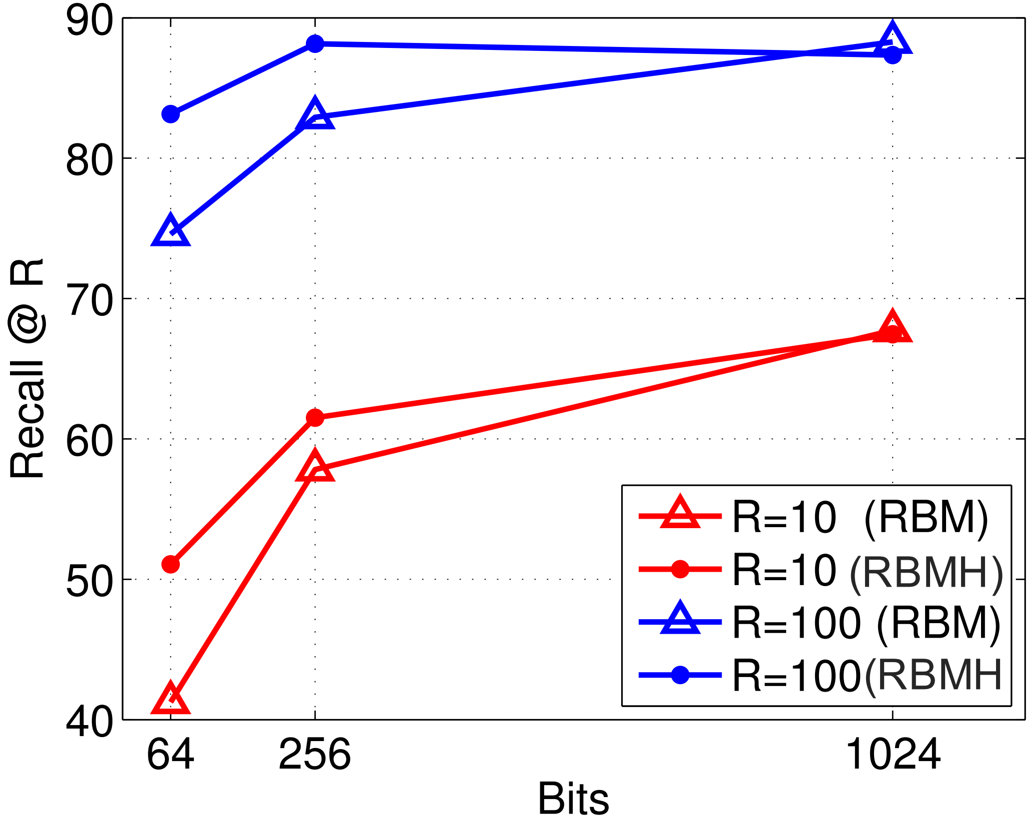
- It is often desirable that the latent variables follow certain distributions
- For example, sparse distributions work well for classification
- Hinton proposes regularization that encourages each unit activation q towards p << 1

$$h(q) = p\, log(q) + (1-p) log(1-q)$$

Regularization for Hashing

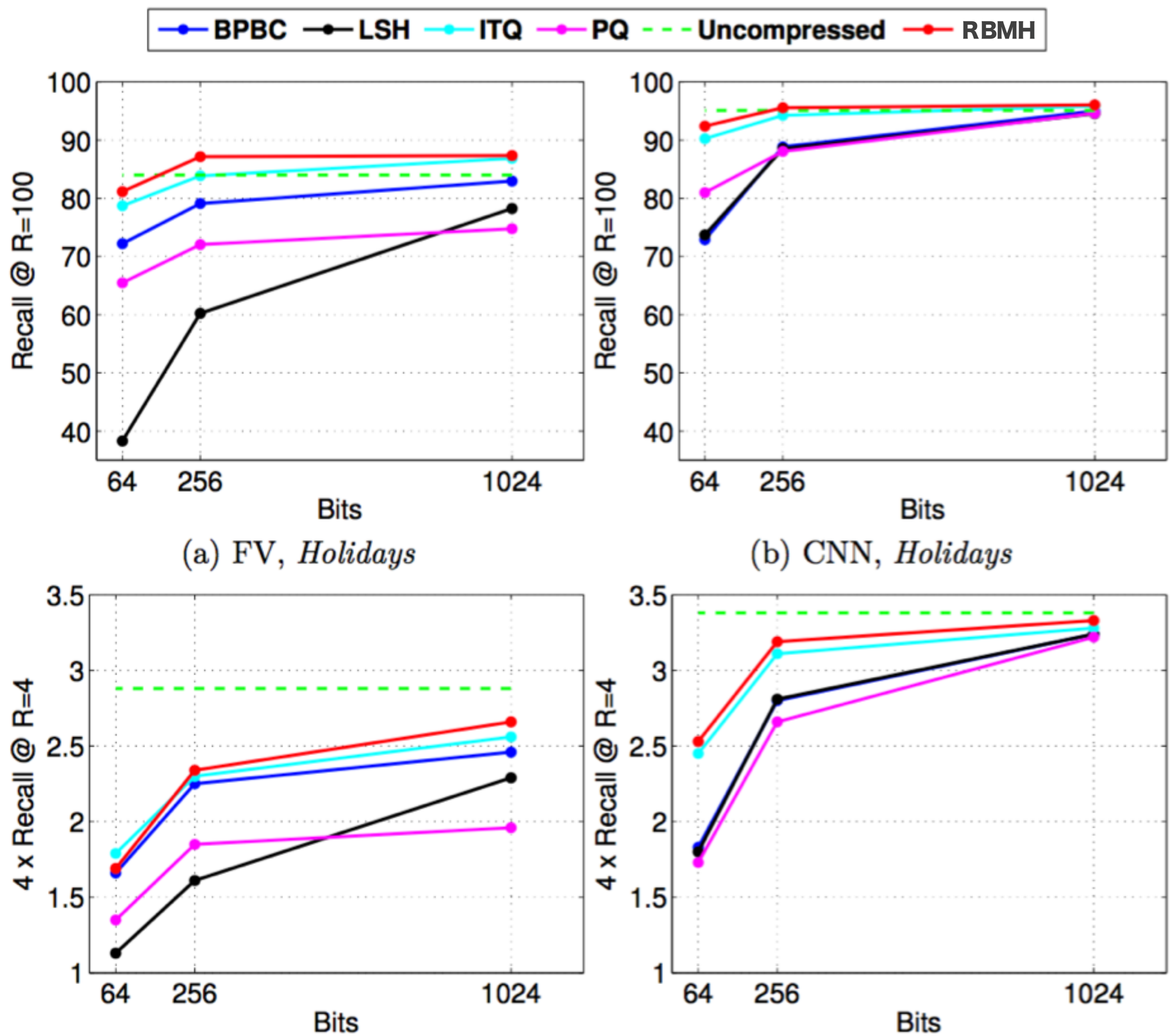


Effect of Regularization for Hashing



Our proposed hashing scheme performs better specially at low bitrates

Comparison with state-of-the-art



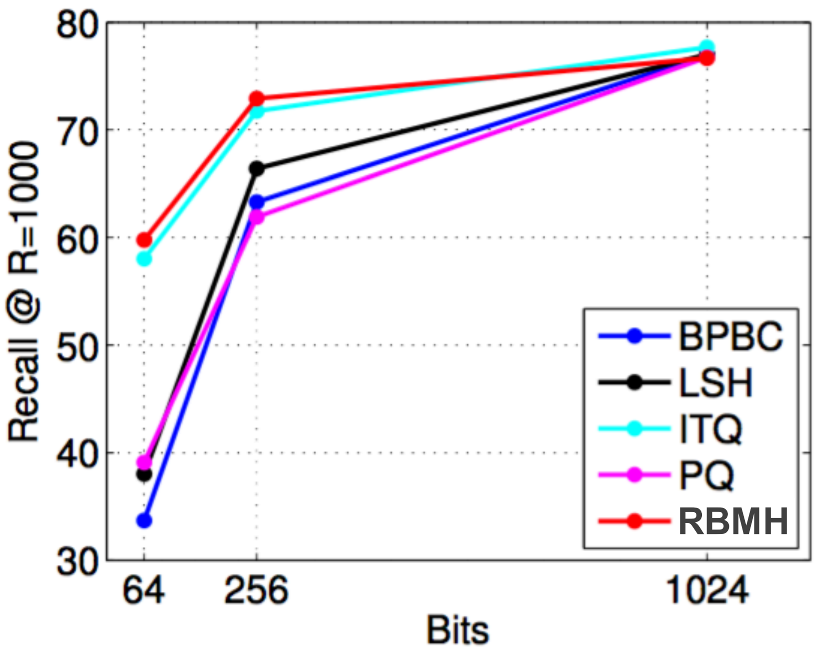
(a) FV, *Holidays*

(b) CNN, *Holidays*

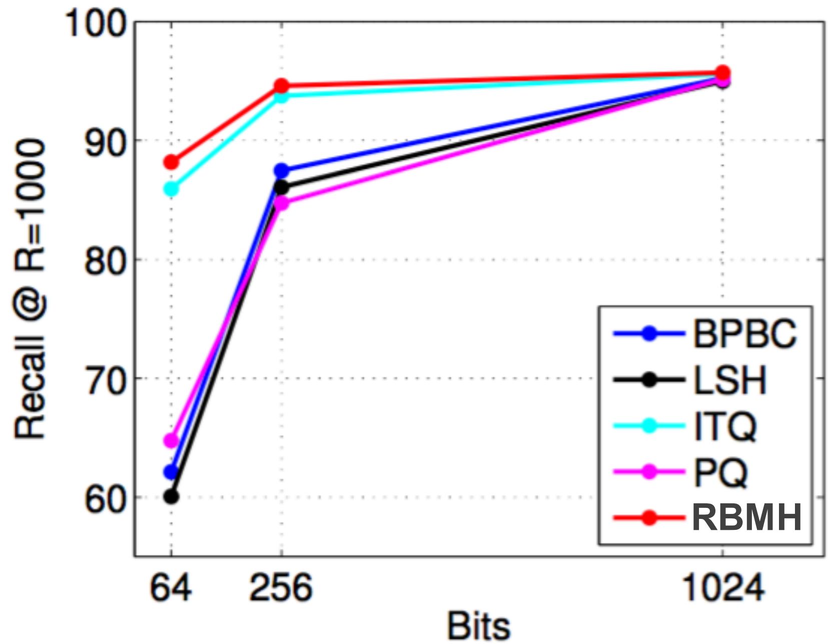
(c) FV, *UKBench*

(d) CNN, *UKBench*

Million Scale Data Sets



(a) CNN, *Holidays*+1M

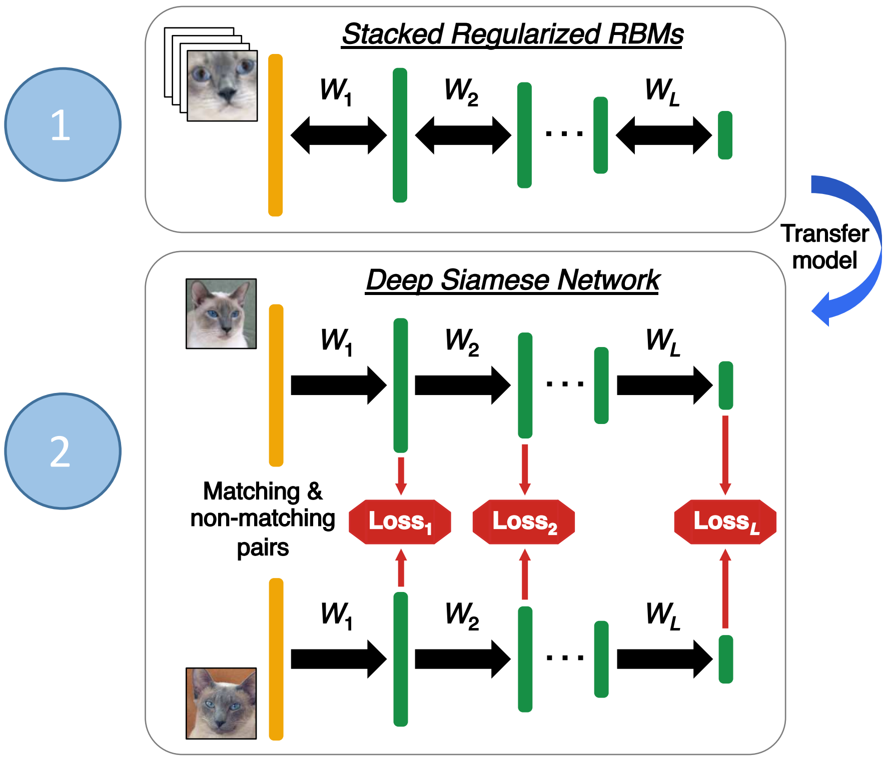


(b) CNN, *UKBench*+1M

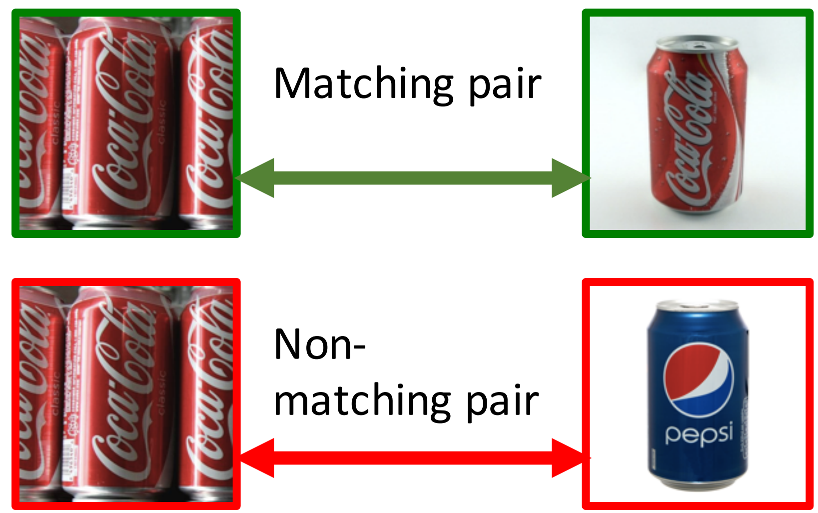
Hashing Outline

1. Dimensionality Reduction with Stacked RBM
2. Semi-supervised fine-tuning with Siamese networks
3. Unsupervised fine-tuning with triplet networks

Supervised Fine-Tuning



Training Siamese network with matching and non matching image pairs:

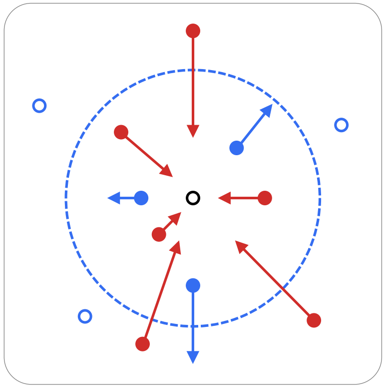


Siamese Networks

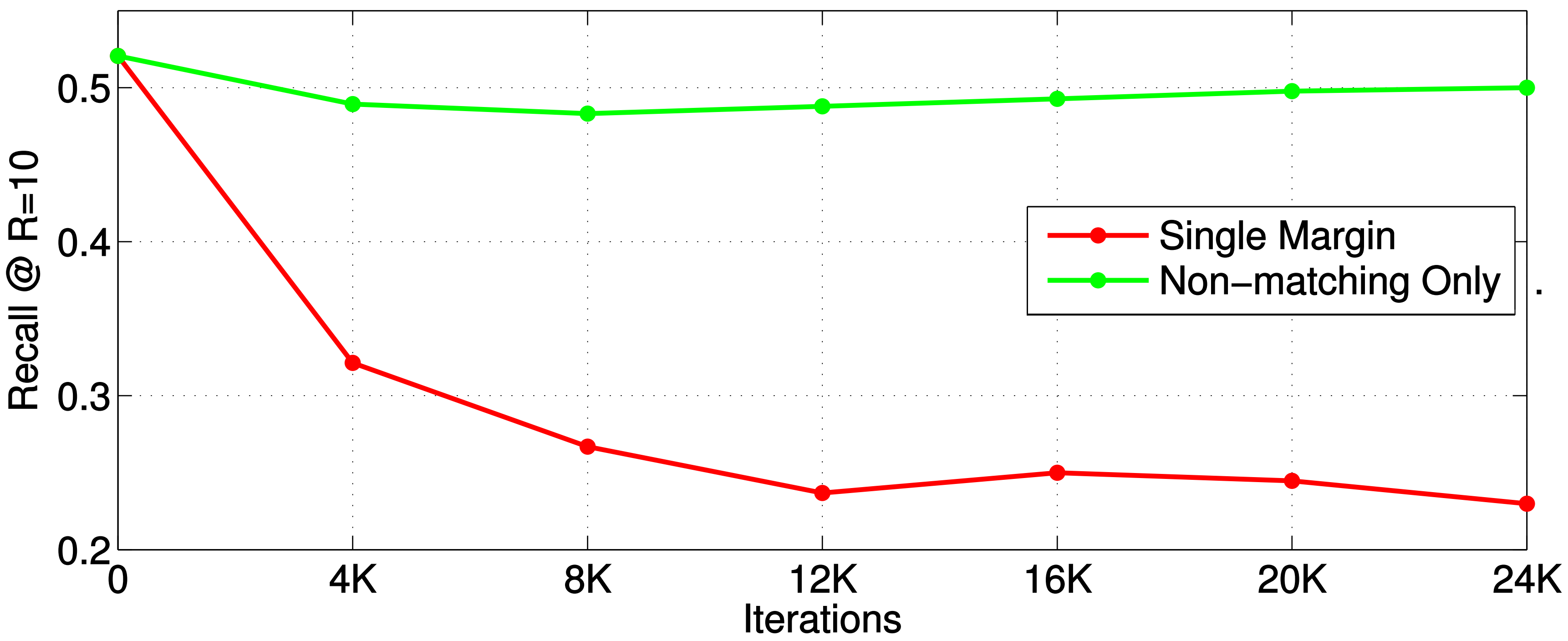
[LeCun, 1994]

- Proposed by LeCun for application such as signature verification or data visualization
- Learning similarity measure between pairs of instances using contrastive loss
- Descriptors from matching pairs are encouraged to be as similar as possible, while descriptors from non-matching pairs are encouraged to be not too similar

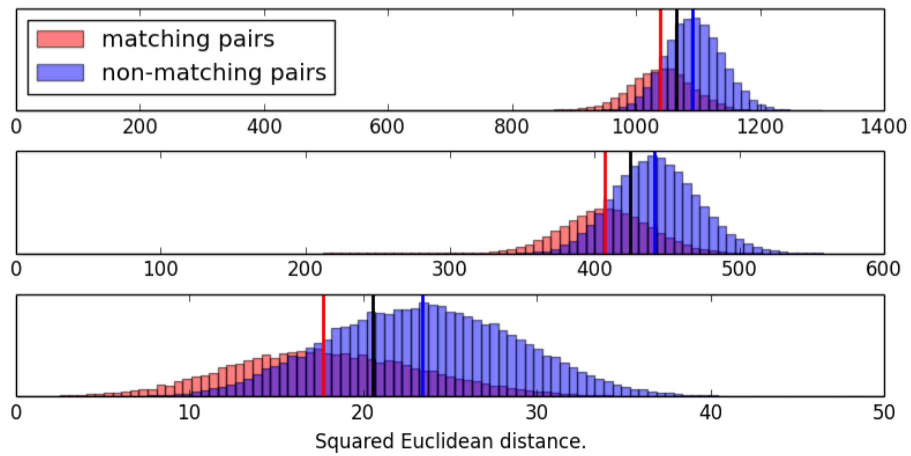
$$\mathcal{D}_l(\mathbf{z}^0_\alpha, \mathbf{z}^0_\beta) = y\|\mathbf{z}^l_\alpha - \mathbf{z}^l_\beta\|_2^2 + (1-y)\max(m - \|\mathbf{z}^l_\alpha - \mathbf{z}^l_\beta\|_2^2, 0)$$



Results with LeCun's Loss

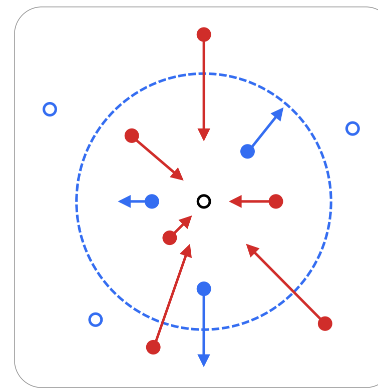


Contrastive Loss



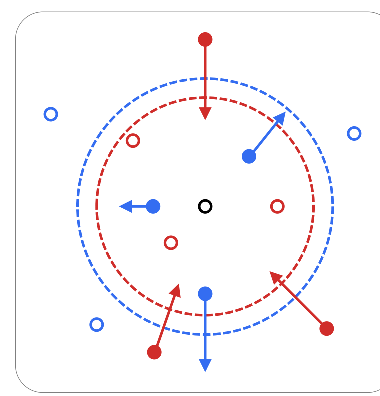
- Single margin loss (LeCun) :

$$\mathcal{D}_l(\mathbf{z}_\alpha^0, \mathbf{z}_\beta^0) = y\|\mathbf{z}_\alpha^l - \mathbf{z}_\beta^l\|_2^2 + (1-y)\max(m - \|\mathbf{z}_\alpha^l - \mathbf{z}_\beta^l\|_2^2, 0)$$

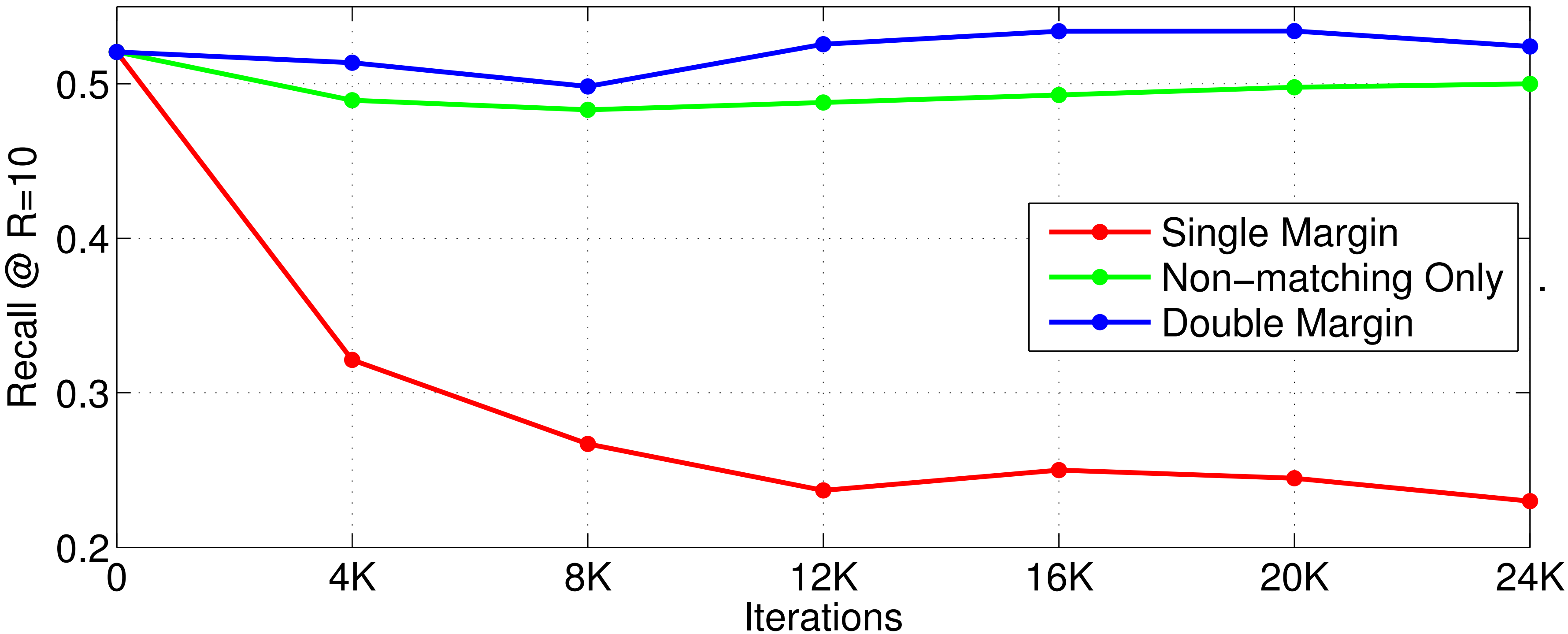


- Proposed dual margin loss :

$$\begin{aligned}\mathcal{D}_l(\mathbf{z}_\alpha^0, \mathbf{z}_\beta^0) = & y\max(\|\mathbf{z}_\alpha^l - \mathbf{z}_\beta^l\|_2^2 - m_1, 0) \\ & + (1-y)\max(m_2 - \|\mathbf{z}_\alpha^l - \mathbf{z}_\beta^l\|_2^2, 0)\end{aligned}$$



Results with Dual Margin Loss



Dual-Margin Siamese Evaluation

Data set	Layer	Recall @ R=10			Recall @ R=100		
		bef.	aft.	diff.	bef.	aft.	diff.
Holidays	4096	70.83	73.67	**2.84**	89.92	91.40	**1.48**
	2048	67.74	71.12	**3.38**	88.77	92.04	**3.27**
	64	52.06	53.04	**0.98**	80.38	83.91	**3.53**
UKBench	4096	79.22	82.22	**3.00**	92.04	93.73	**1.69**
	2048	75.62	79.37	**3.75**	90.79	92.82	**2.03**
	64	47.94	49.25	**1.31**	73.02	73.94	**0.92**
Oxbuild	4096	19.38	21.73	**2.35**	41.09	45.19	**4.10**
	2048	14.32	17.23	**2.91**	36.03	41.03	**5.00**
	64	10.69	12.01	**1.32**	23.75	29.99	**6.24**

Million Scale Data Sets

Data set	Bitrate	Recall @ R=1000		
		bef.	aft.	diff.
Holidays + 1M	1024	52.49	56.48	**3.99**
	256	46.07	49.60	**3.53**
	64	30.63	31.87	**1.24**
UKBench + 1M	1024	85.66	86.41	**0.75**
	256	78.70	80.74	**2.04**
	64	61.13	62.74	**1.61**

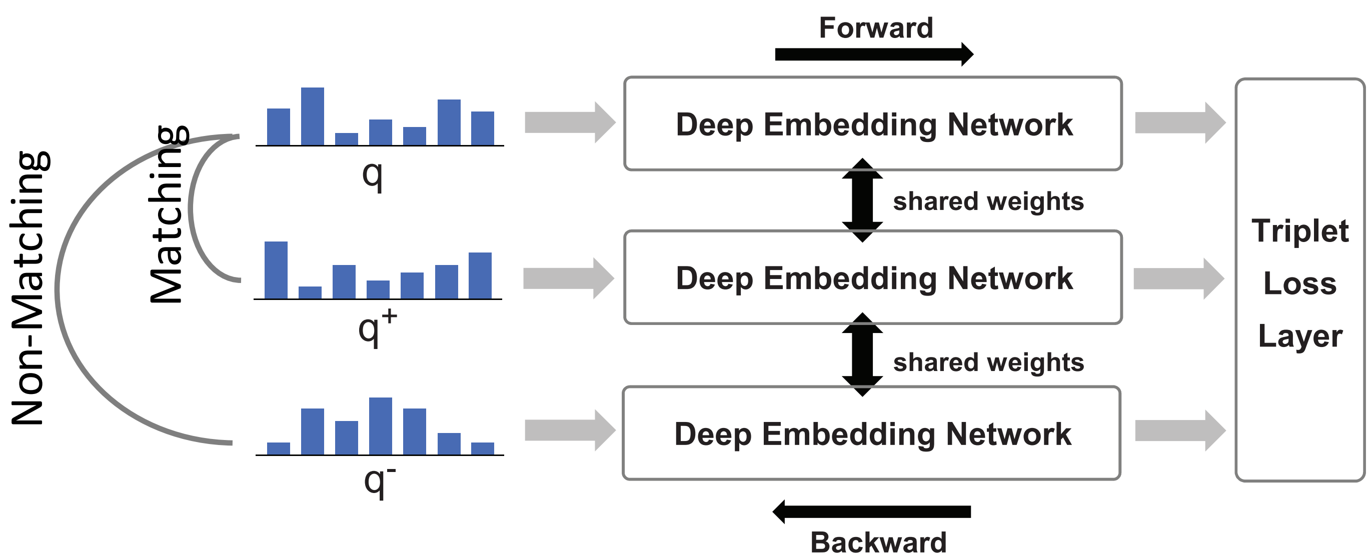
Hashing Outline

1. Dimensionality Reduction with Stacked RBM
2. Semi-supervised fine-tuning with Siamese networks
3. Unsupervised fine-tuning with triplet networks

Triplet Fine-Tuning

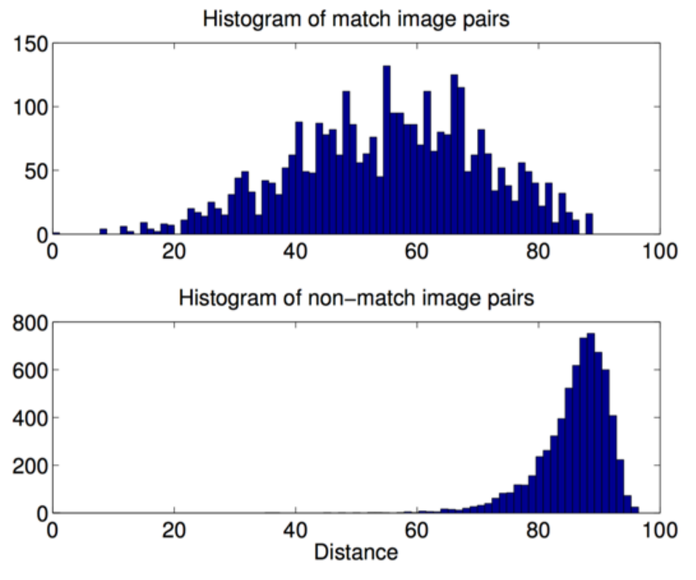
[Wang, 2014]

- Fine-tuning based on order, not absolute distances (margins)
- This can be achieved with triplet networks



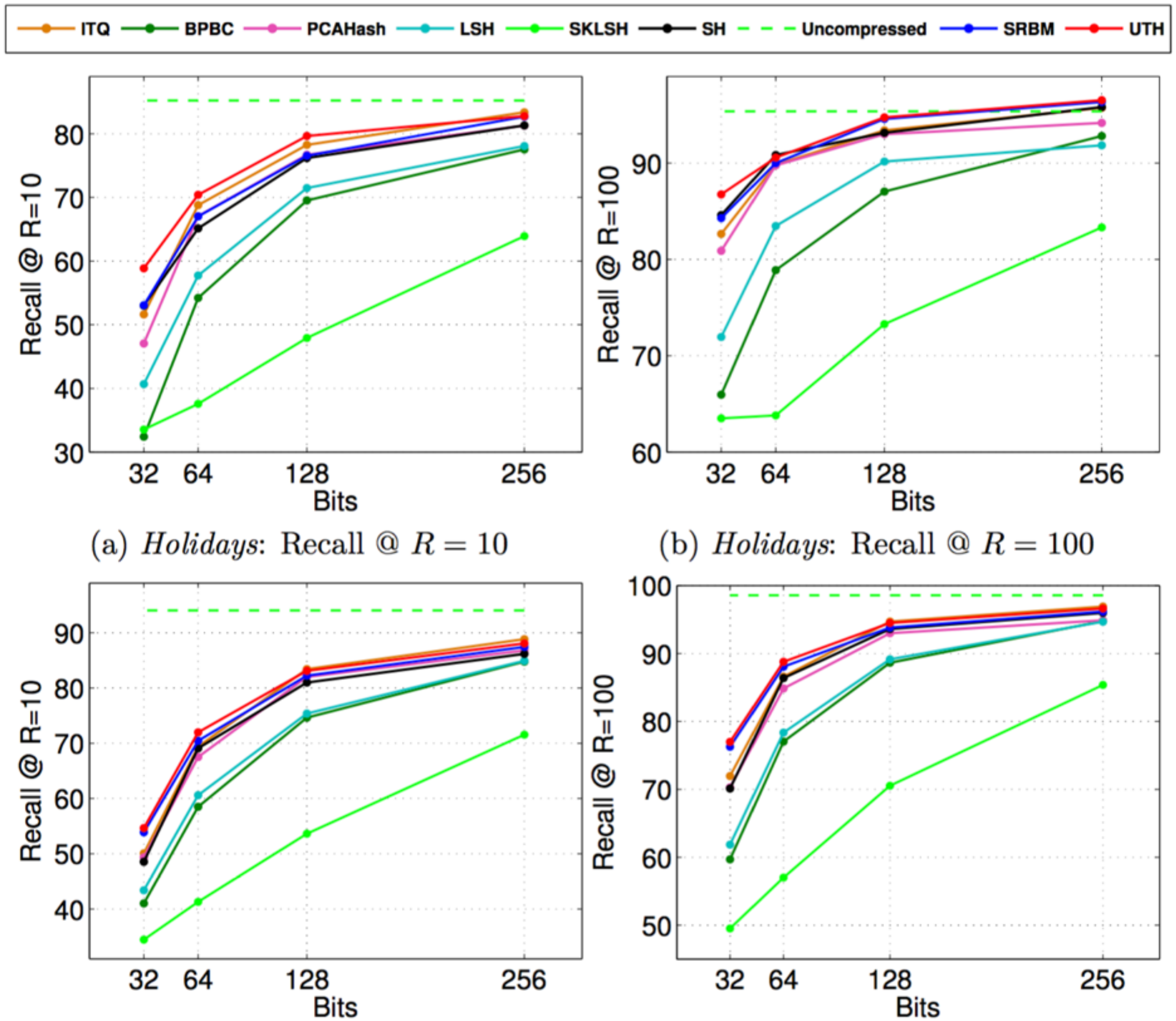
Unsupervised Sampling of Triplet Data

- Retaining the performance of the uncompressed descriptors is a good objective



- Unsupervised: the triplets can be made using the distance in the original descriptor space

Unsupervised Fine-Tuning Results



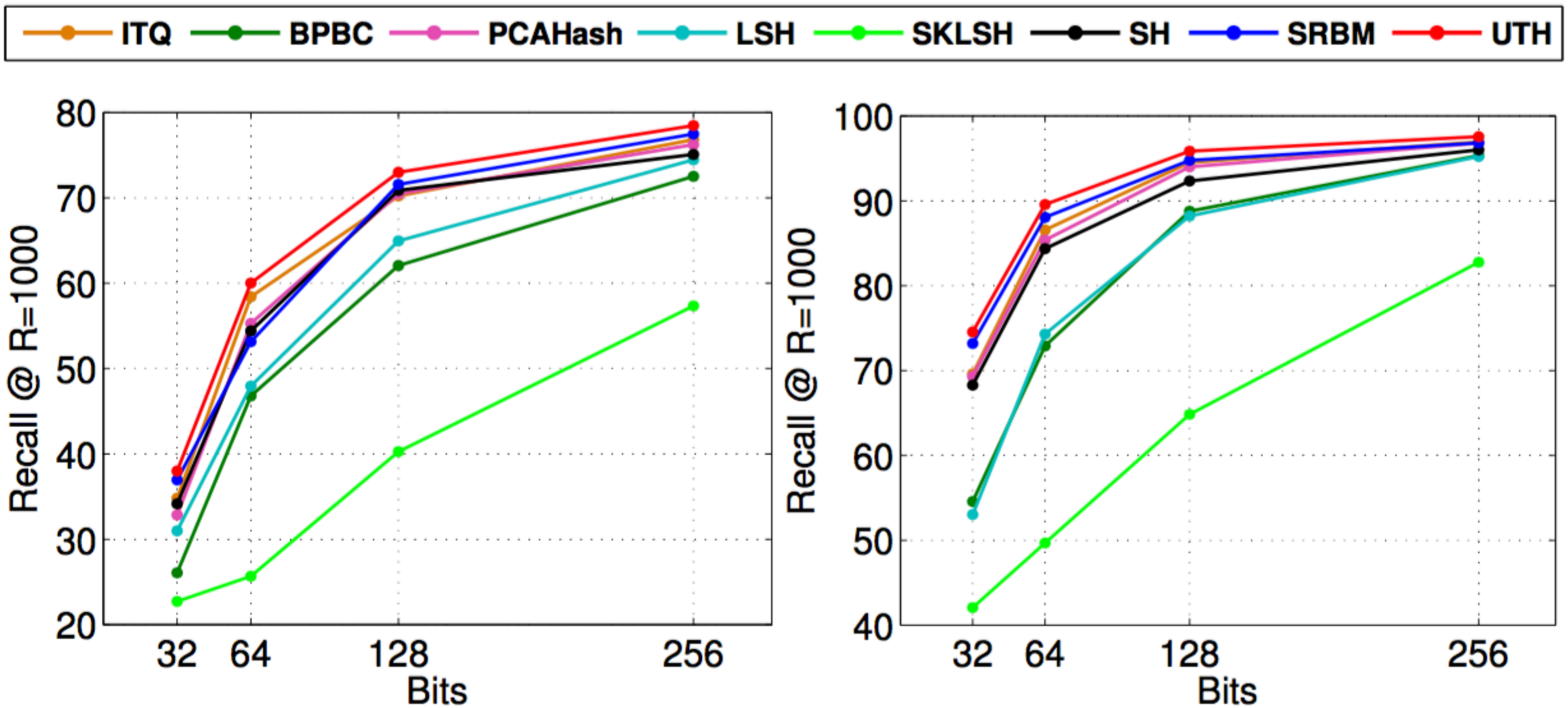
(a) *Holidays*: Recall @ $R = 10$

(b) *Holidays*: Recall @ $R = 100$

(c) *UKBench*: Recall @ $R = 10$

(d) *UKBench*: Recall @ $R = 100$

Million Scale Data Sets



(a) *Holidays*+1M: Recall @ $R = 1000$ (b) *UKBench*+1M: Recall @ $R = 1000$

Further Readings

- O. Morère, J. Lin, V. Chandrasekhar, A. Veillard, H. Goh. Co-Sparsity Regularized Deep Hashing for Image Instance Retrieval. *Accepted in International Conference on Image Processing (ICIP) 2016*
- *O. Morère, J. Lin, J. Petta, V. Chandrasekhar and A. Veillard Tiny descriptors for image retrieval with unsupervised triplet hashing. Data Compression Conference (DCC) 2016*
- *V. Chandrasekhar, J. Lin, O. Morère, A. Veillard and H. Goh. Compact Global Descriptors for Visual Search. Data Compression Conference (DCC) 2015*
- O. Morère, J. Lin, V. Chandrasekhar, A. Veillard and H. Goh. Deephash: Getting Regularization, Depth and Fine-tuning Right. *arXiv preprint arXiv:1501.04711 2015*

PART 3

1. Thorough comparison study of FV and CNN for image instance retrieval

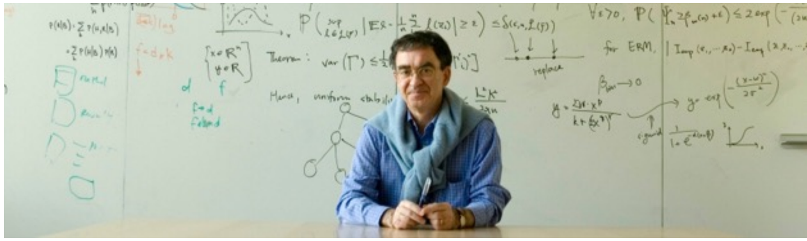
2. Hashing CNN descriptors

3. Robust CNN descriptors with i-theory

Tomaso Poggio

Massachusetts Institute of Technology

CENTER FOR Brains Minds+ Machines

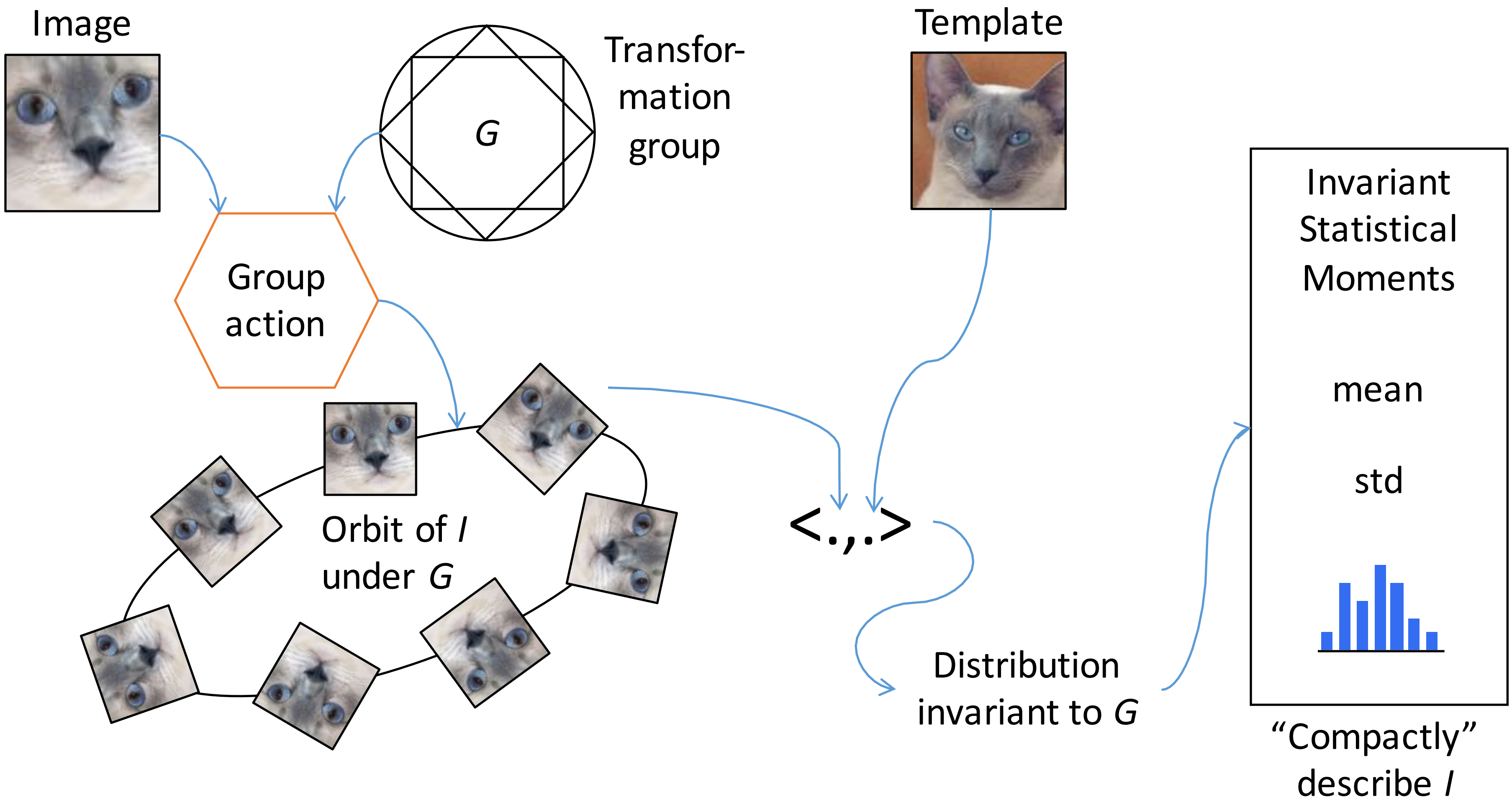


Joint Publications:

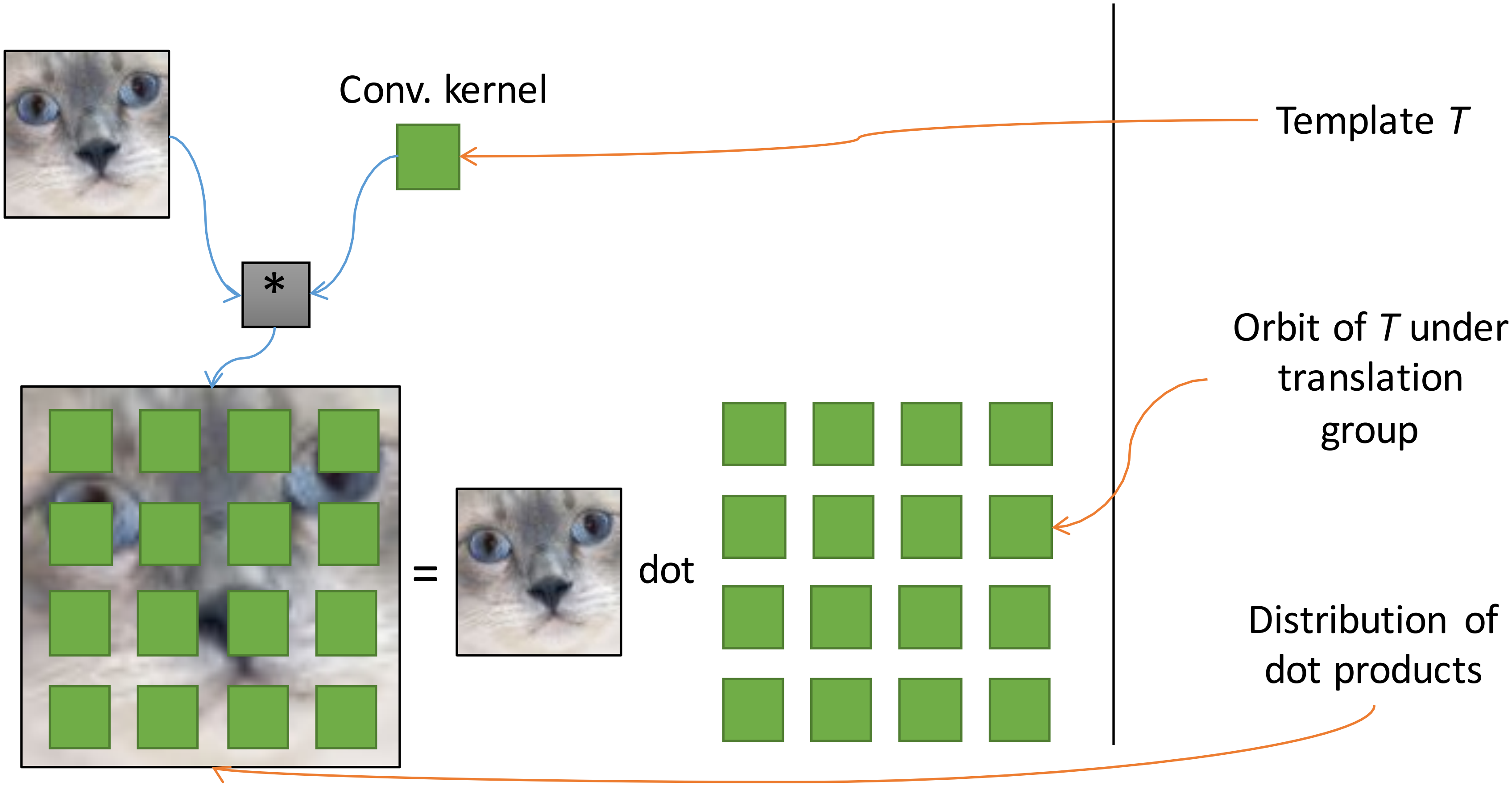
- *O. Morère, J. Lin, A. Veillard, V. Chandrasekhar, T. Poggio. Nested Invariance Pooling and RBM Hashing for Image Instance Retrieval. Submitted to European Conference on Computer Vision (ECCV) 2016*
- *O. Morère, A. Veillard, J. Lin, J. Petta, V. Chandrasekhar and T. Poggio. Group In- variant Deep Representations for Image Instance Retrieval. Center for Brains, Minds and Machines (CBMM) 2016*

i-theory in a Nutshell

[Poggio, 2013]



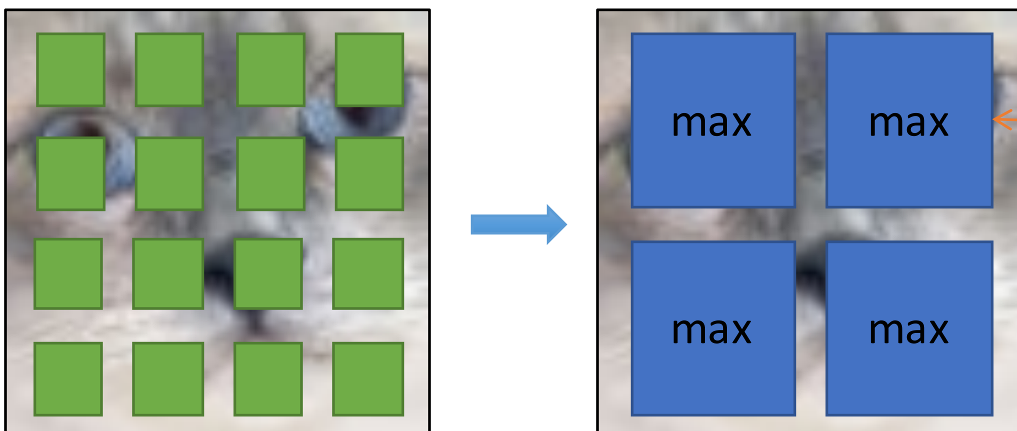
i-theory and CNN (convolution)



The **convolution** operation is analogous to the computation of the **dot products** between the target image and orbit of the kernel (**learnt template**) under the **translation group**.

i-theory and CNN (pooling)

Convolutional Neural Nets | **I-Theory**

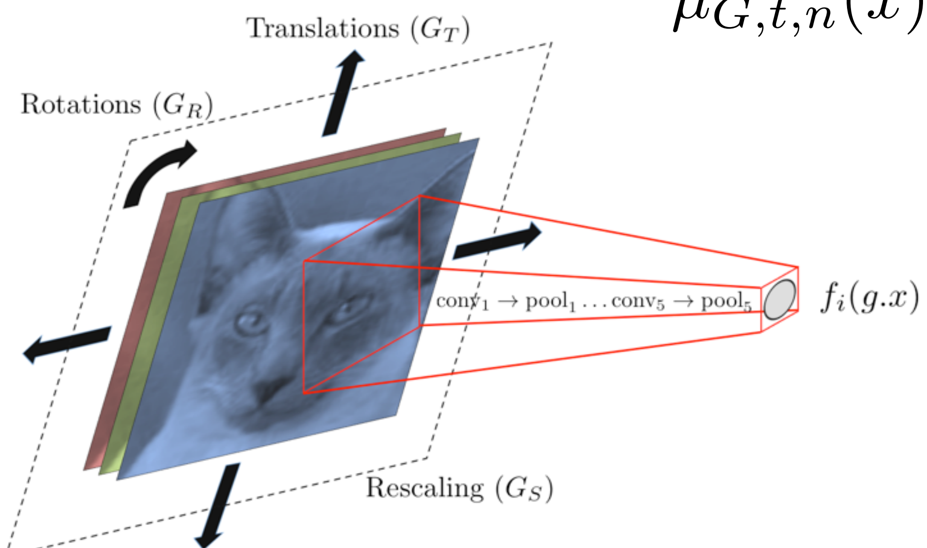


Computing statistical moment invariant to local translations

Distribution of dot products

The **pooling** operation computes **statistical moments** (the maximum) of the distribution of dot products **invariant to local translations**.

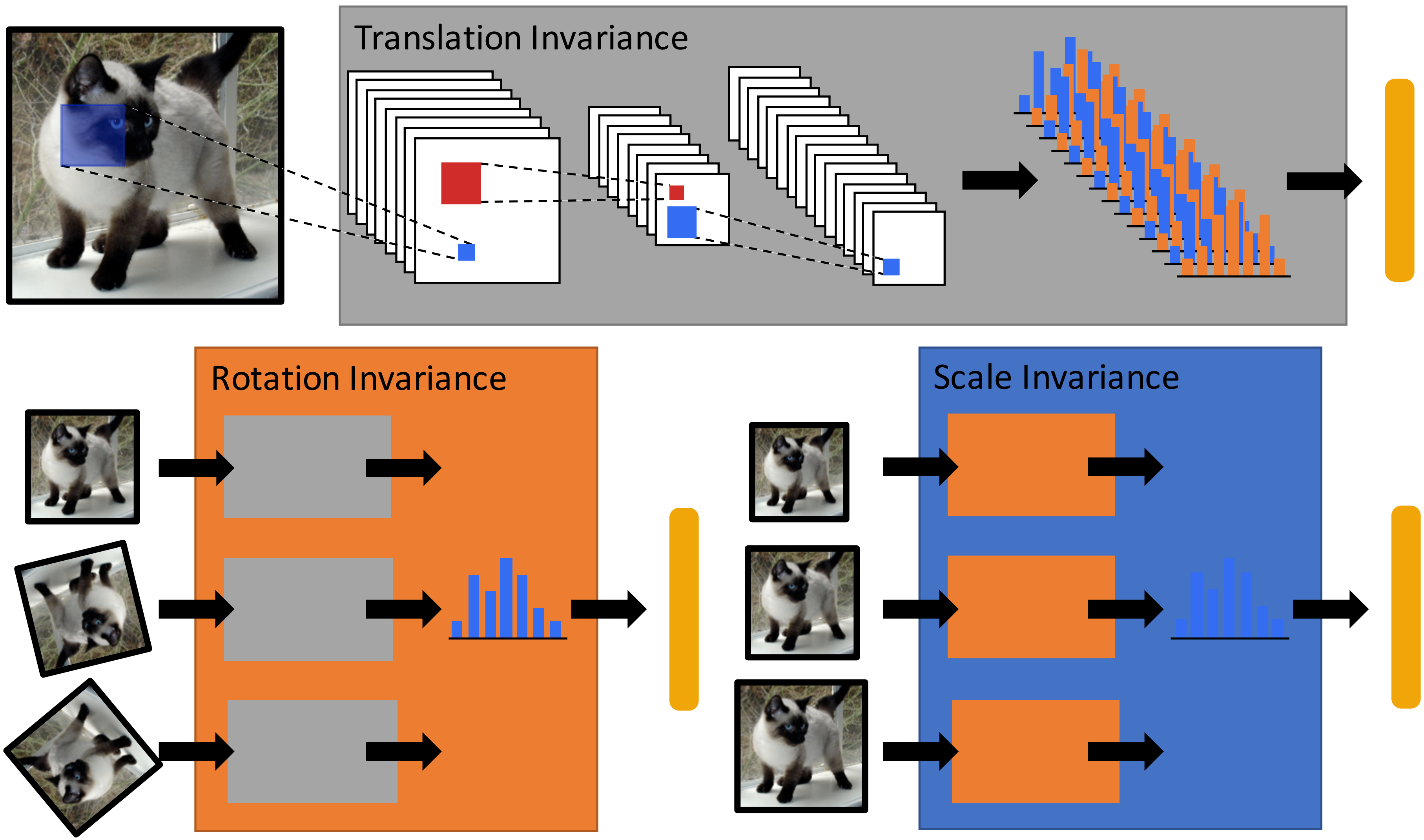
Pooling CNN Descriptors



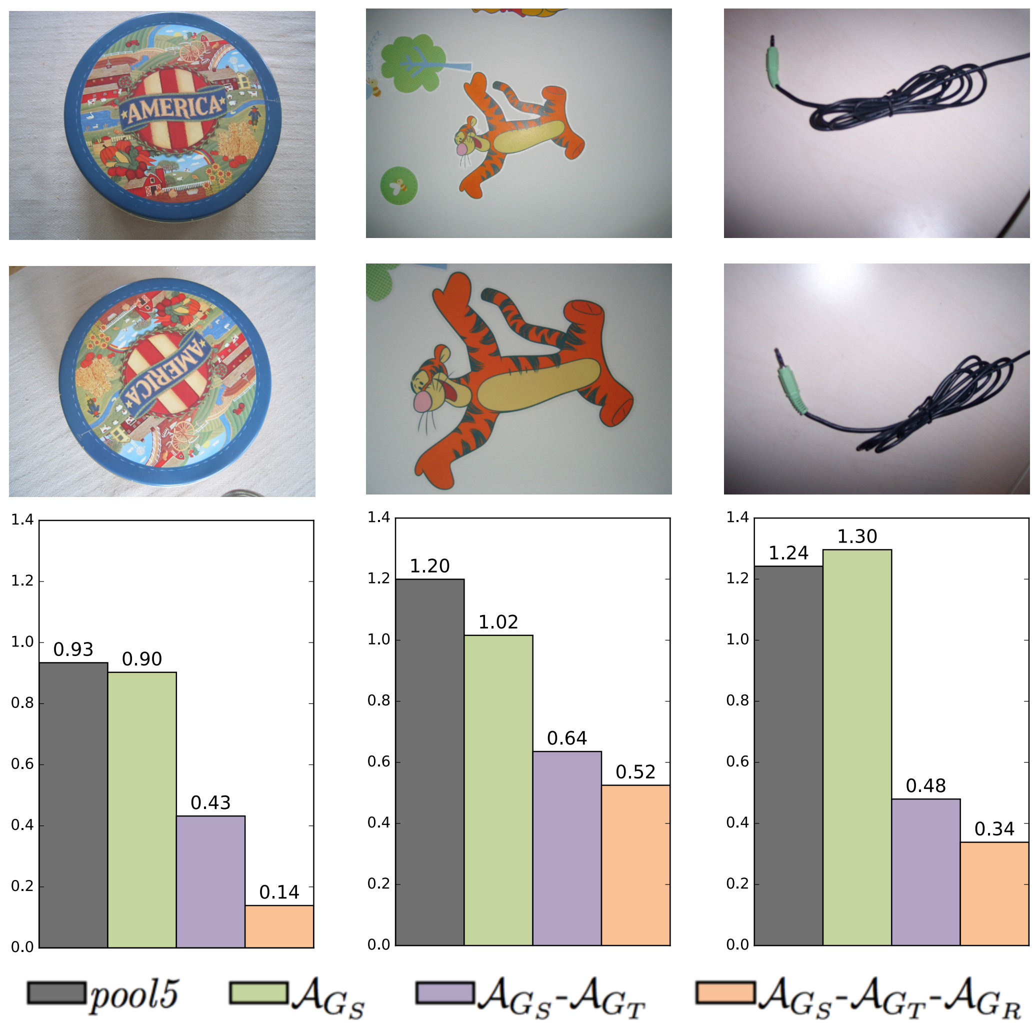
$$\mu_{G,t,n}(x) = \frac{1}{m}\left(\sum_{j=0}^{m-1} | < g_j.x, t > |^n\right)^{\frac{1}{n}}$$

$$\mathcal{X}_{G,i,n}(x) = \frac{1}{m}\left(\sum_{j=0}^{m-1} f_i(g_j.x)^n\right)^{\frac{1}{n}}$$

Nested Invariance Pooling (NIP)



Distances for Three Matching Pairs

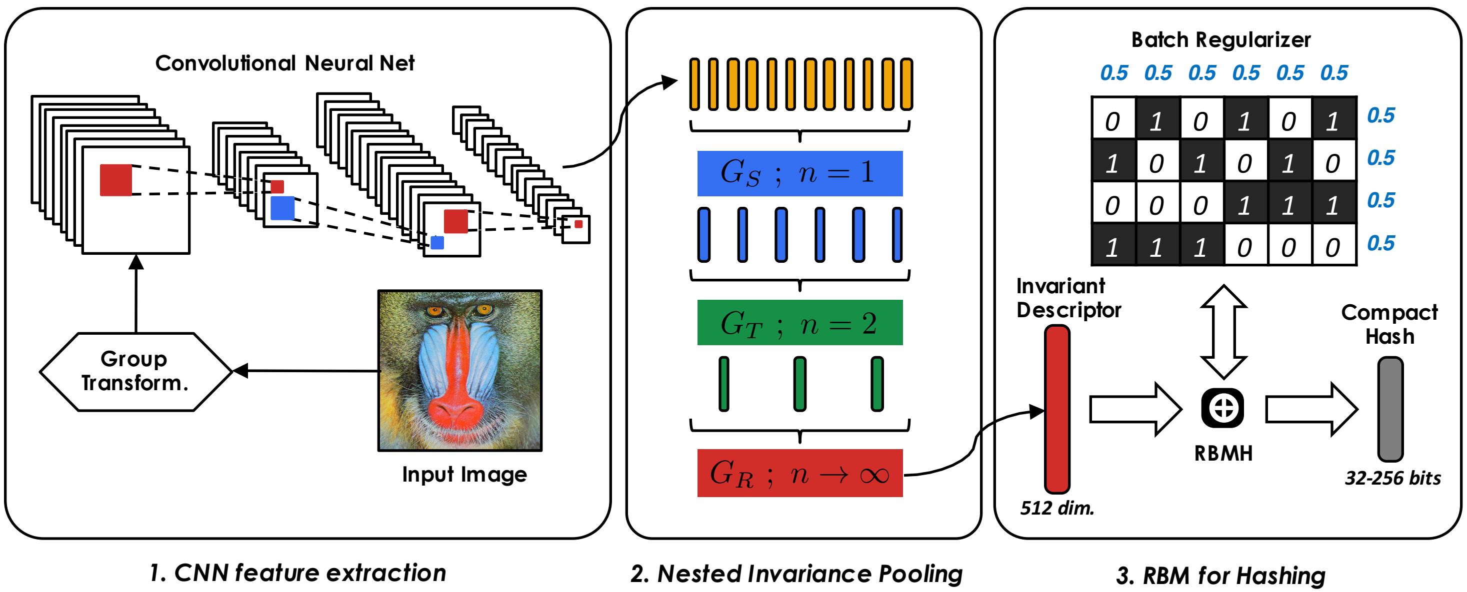


NIP Evaluation

SEQUENCE	DIMS	DATASET			
		Oxbuild	*Holidays*	UKB	Graphics
pool5	25088	0.427	0.707	0.823(3.11)	0.315
fc6	4096	0.461	0.782	0.910(3.50)	0.312
$\mathcal{A}_{G_T}$	512	0.477	0.800	0.924(3.56)	0.322
$\mathcal{A}_{G_R}$	25088	0.462	0.779	0.954(3.72)	0.500
$\mathcal{A}_{G_S}$	25088	0.430	0.716	0.828(3.12)	0.394
$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$	512	0.418	0.796	0.955(3.73)	0.417
$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_S}$	512	0.537	0.811	0.931(3.61)	0.430
$\mathcal{A}_{G_R}$-$\mathcal{A}_{G_S}$	25088	0.494	0.815	0.959(3.75)	0.552
$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$-$\mathcal{A}_{G_S}$	512	0.484	0.833	0.971(3.82)	0.509
$\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$	512	**0.592**	**0.838**	**0.975(3.84)**	**0.589**

NIP + RBMH

- NIP transformation invariant descriptors can be hashed into compact binary codes using RBMH



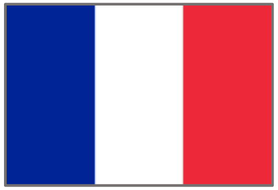
NIP + RBMH - Evaluation

METHOD	DIMS (size in bits)	DATASET		
		Oxbuild	*Holidays*	UKB
Binarized FV	520(520)	-	0.460	2.79
FV+ SH	256(256)	-	0.544	3.08
FV+ SH	128(128)	-	0.499	2.91
FV+ SH	32(32)	-	0.334	2.18
FV+PQ	128(128)	-	0.506	3.10
VLAD+PQ	128(128)	-	0.586	2.88
VLAD+CQ	128(128)	-	0.644	3.19
VLAD+SQ	128(128)	-	0.639	3.06
FC+Finetune+PCAWhitening	16(512)	0.418	0.609	2.41
Conv+MaxPooling	256(256)	0.436	0.578	-
Binarized NIP	512(512)	**0.477**	**0.781**	**3.70**
NIP+RBMH	256(256)	**0.445**	**0.739**	**3.59**
NIP+RBMH	128(128)	**0.359**	**0.705**	**3.38**

Conclusion

- Thorough comparison study of FV and CNN for image instance retrieval
- Hashing CNN descriptors
  - Unsupervised dimensionality reduction
  - Descriptor finetuning with unsupervised and semi-supervised metric learning
- Robust CNN descriptors with i-theory

Acknowledgements



Antoine VEILLARD

Daniel RACOCEANU

Vijay CHANDRASEKHAR

Hanlin GOH

Jie Lin

cnrs dépasser les frontières

LIP6

ipal Image & Pervasive Access Lab

A*STAR

Institute for Infocomm Research