Image Search with Deep Learning

Sung-Eui Yoon
(윤성의)
KAIST

http://sgvr.kaist.ac.kr

Class Objectives are:

- CNN based approaches
  - Consider different regions, attention, and local features
  - Discuss applications

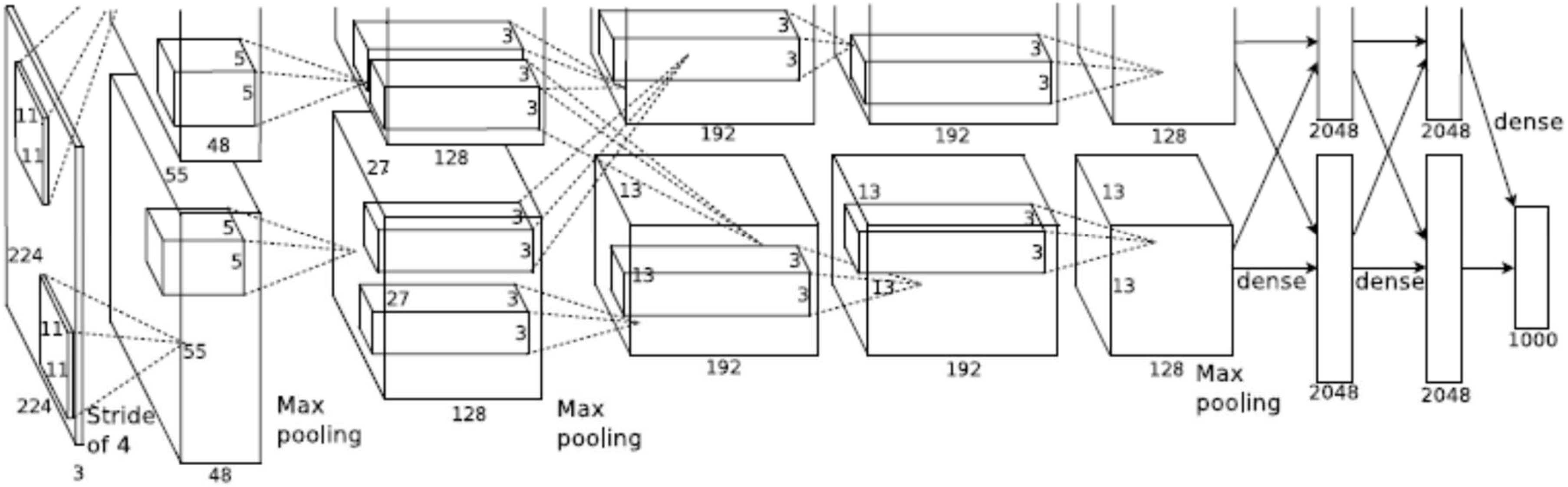
- At the prior class:
  - Discussed unsupervised hashing techniques based on hyperplanes and hyperspheres
  - Talked about supervised approach using deep learning

PA2

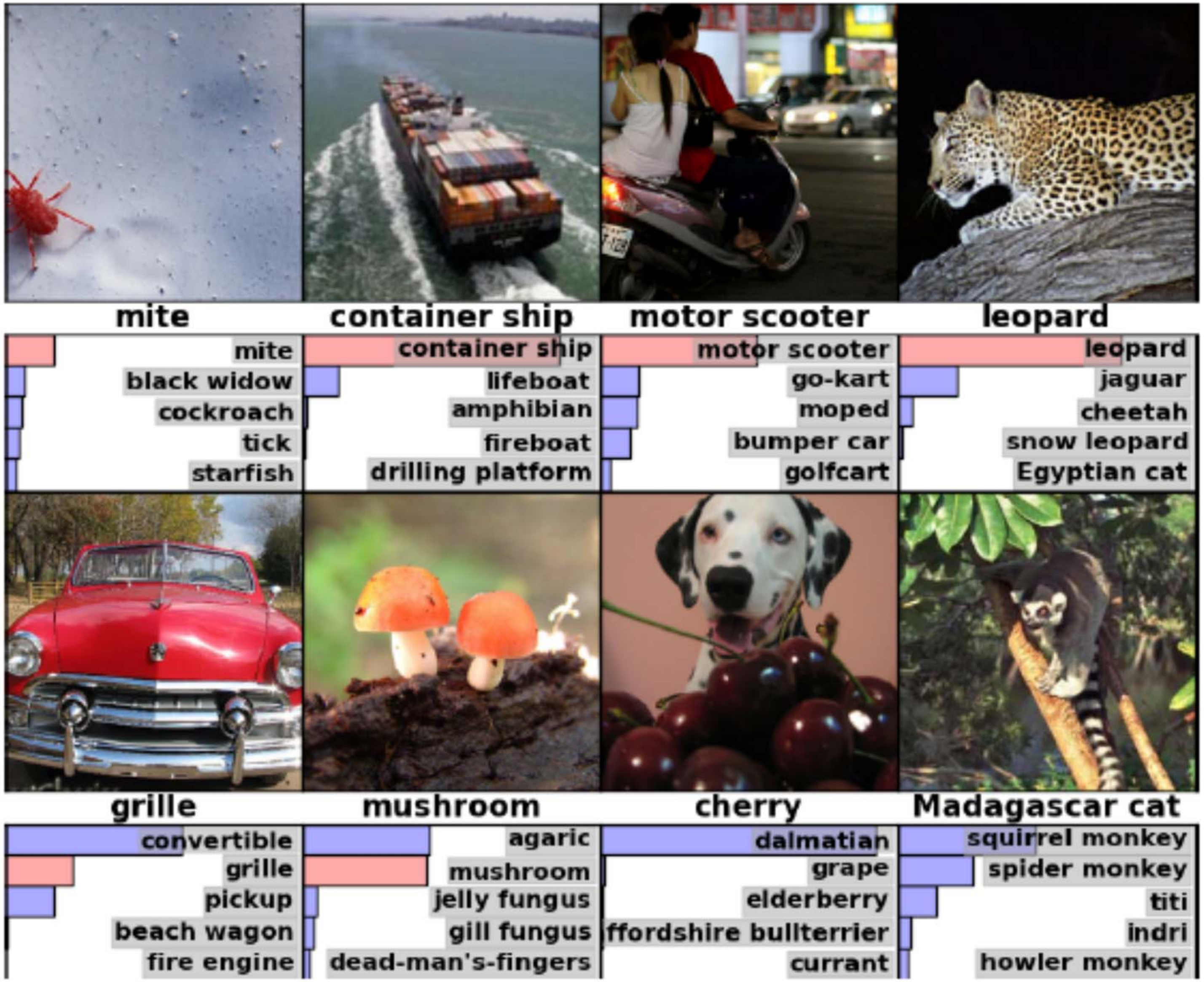
- **Apply binary code embedding and inverted index to PA1**
  - **k-means or product quantization (PQ) for inverted index**
  - **Spherical hashing or PQ for binary code embedding**

ImageNet Classification with Deep Convolutional Neural Networks [NIPS 12]

- Rekindled interest on CNNs
  - Use a large training images, ImageNet, of 1.2 M labelled images
  - Use GPU w/ rectifying non-linearities

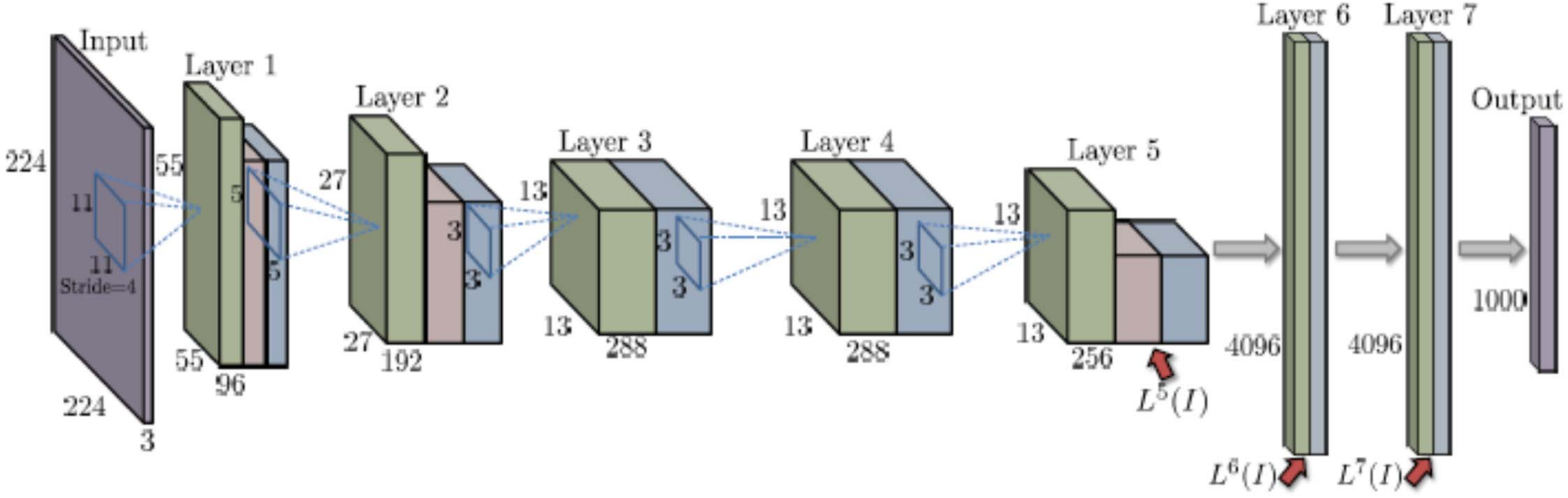


Tested on ILSVRC-2010



Neural Codes for Image Retrieval [ECCV 14]

- **Uses top layers of CNNs as high-level global descriptors (Neural Codes) for image search**

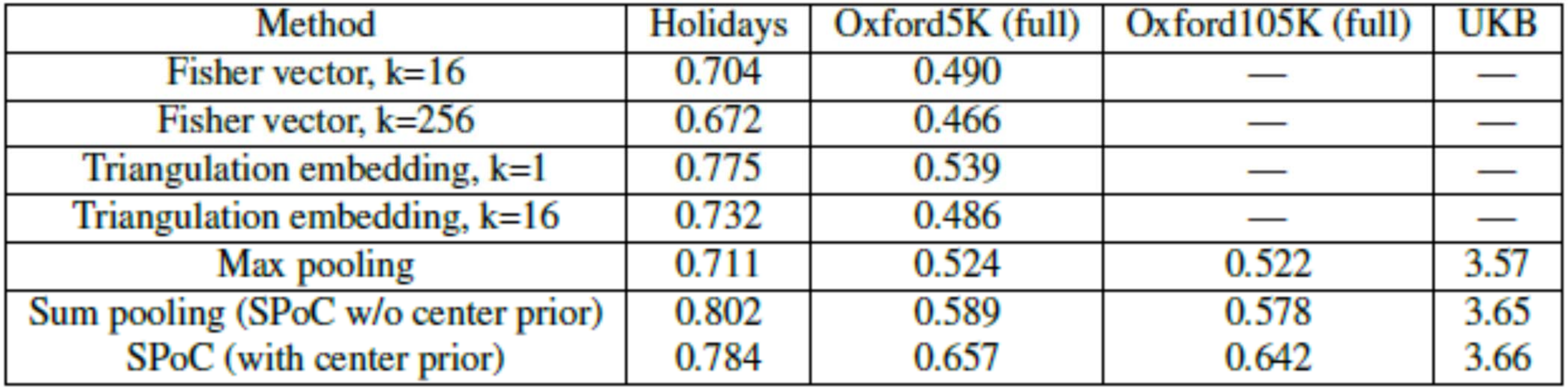


Sum Pooling and Centering Priors

- **Inspired by many prior aggregated features (e.g., BoW)**
  - **Use convolution layers as local features**

- **Aggregation**
  - **Simply sums those local features or**
  - **Considers centering priors w/ varying weights**

$$\psi_1(I) = \sum_{y=1}^{H} \sum_{x=1}^{W} f_{(x,y)}$$



Method	Holidays	Oxford5K (full)	Oxford105K (full)	UKB
Fisher vector, k=16	0.704	0.490	—	—
Fisher vector, k=256	0.672	0.466	—	—
Triangulation embedding, k=1	0.775	0.539	—	—
Triangulation embedding, k=16	0.732	0.486	—	—
Max pooling	0.711	0.524	0.522	3.57
Sum pooling (SPoC w/o center prior)	0.802	0.589	0.578	3.65
SPoC (with center prior)	0.784	0.657	0.642	3.66

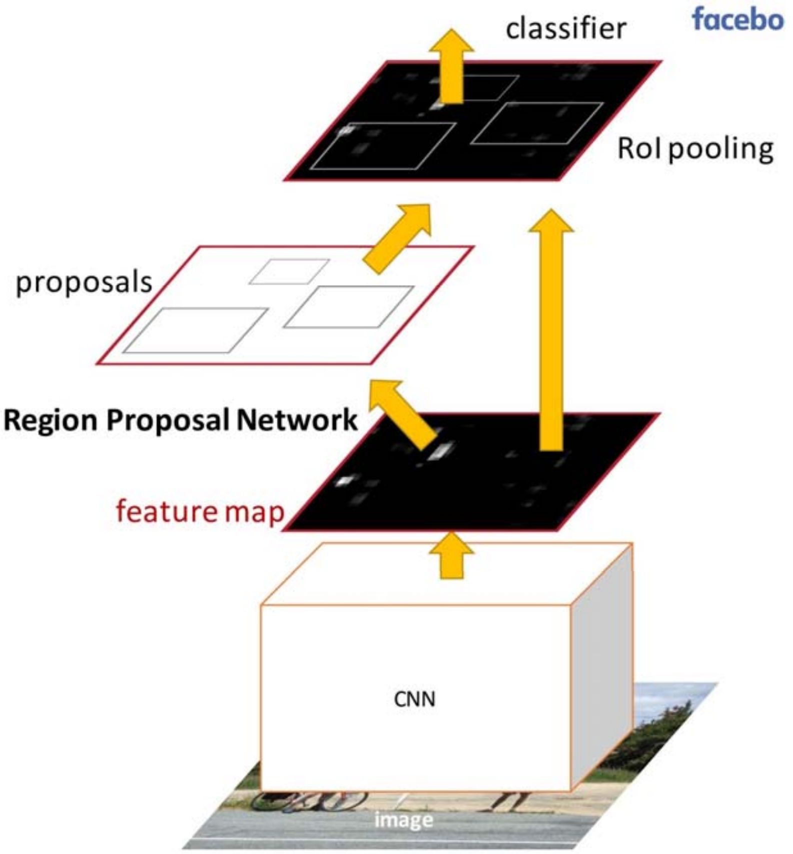
Ack.: Aggregating Deep Convolutional Features for Image Retrieval

Localization: Faster R-CNN

- **Insert a Region Proposal Network (RPN) after the last convolutional layer**

- **RPN trained to produce region proposals directly**
  - **No need for external region proposals!**

- **Use RoI pooling and an upstream classifier and bbox regressor just like Fast R-CNN**

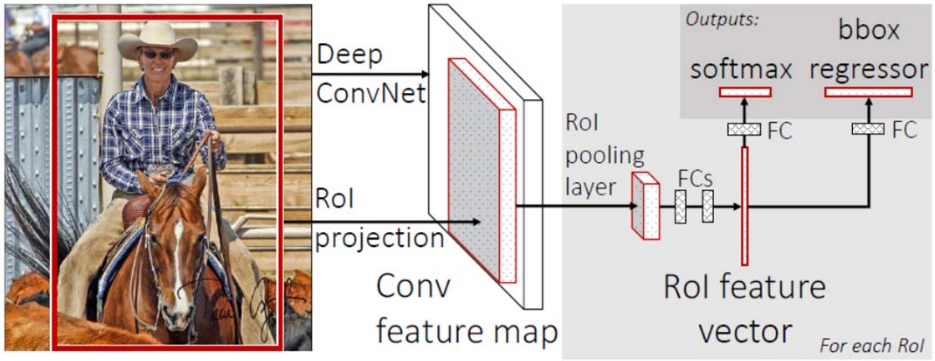


Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015

Slide credit: Ross Girschick

Faster R-CNN: Results

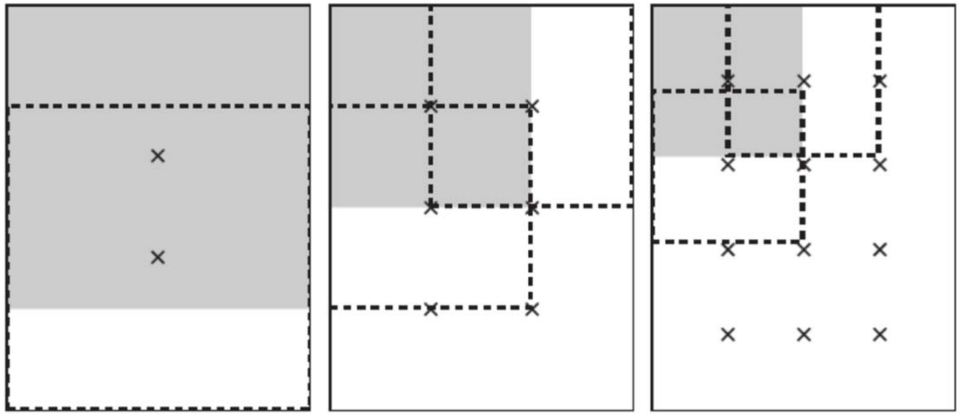
	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image (with proposals)	50 seconds	2 seconds	**0.2 seconds**
(Speedup)	1x	25x	**250x**
mAP (VOC 2007)	66.0	**66.9**	**66.9**



Fast R-CNN: rely upon external region proposal

R-MAC: Regional Maximum Activation of Convolutions

- **Use maximum activation of convolutions for translation invariance**
- **Consider uniformly generated regions with different scales, and sum their features**



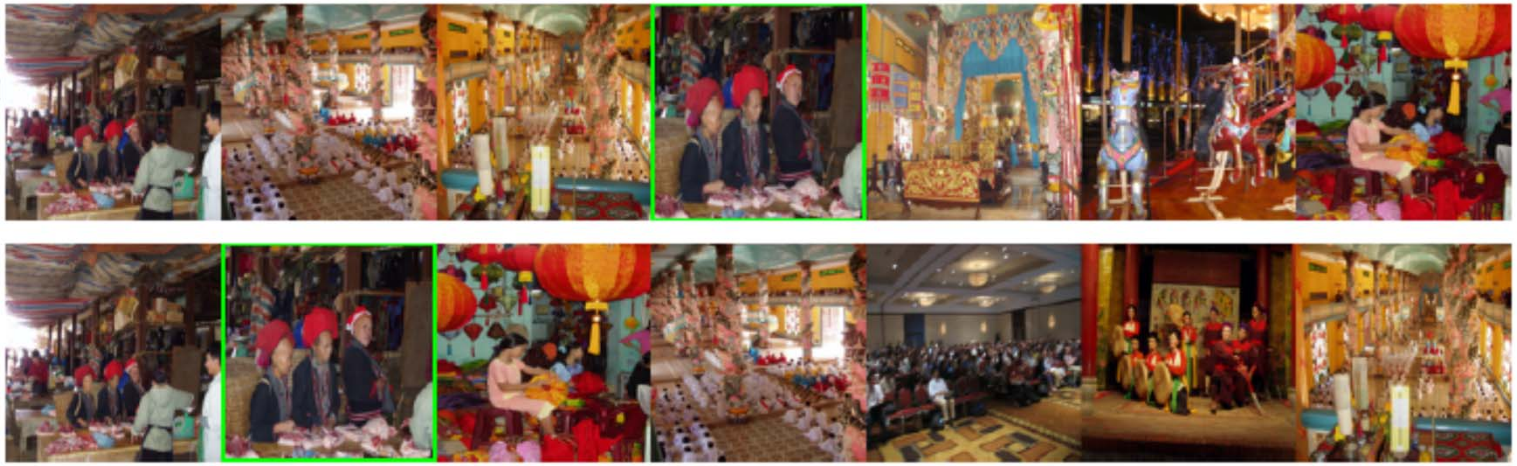
Ack.: PARTICULAR OBJECT RETRIEVAL WITH INTEGRAL MAX-POOLING

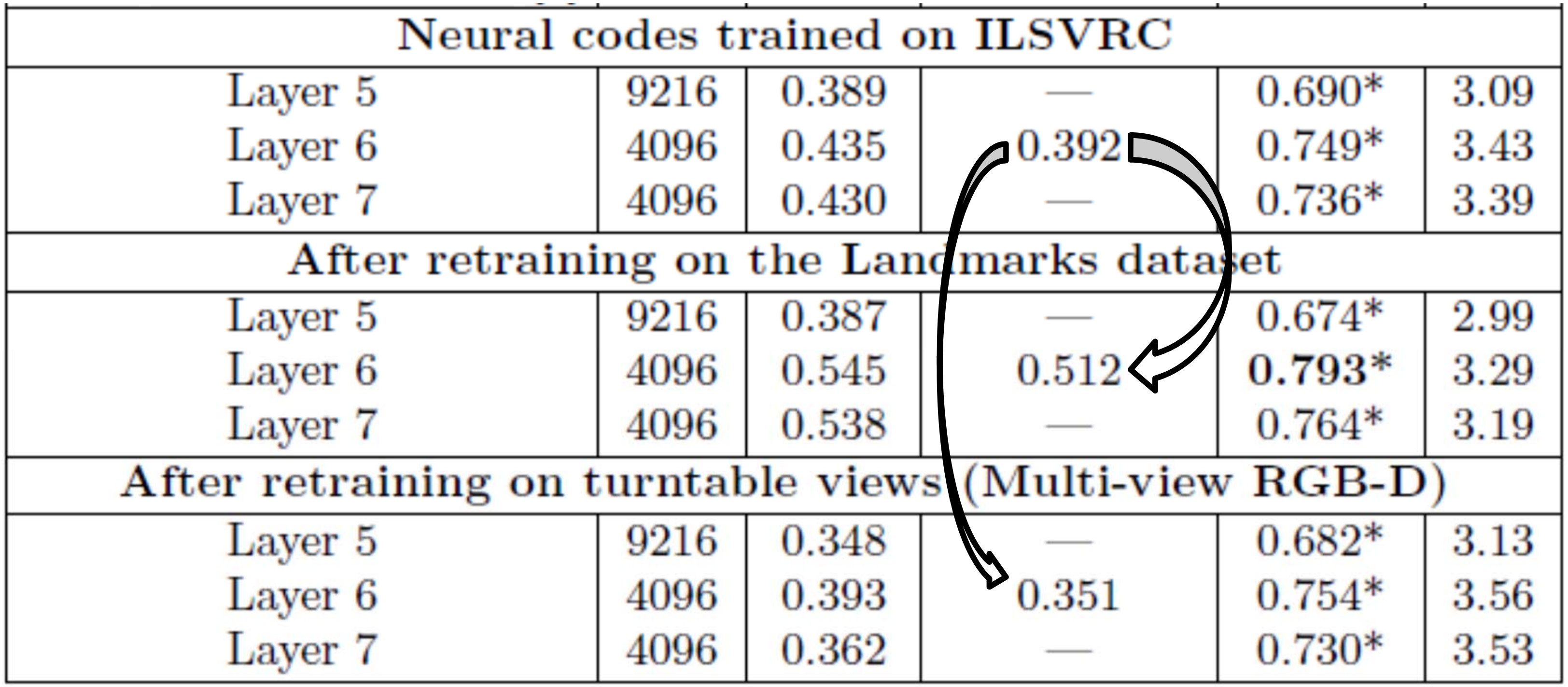
Fine-Tuning for Search

- Use CNN features that were trained with ImageNet
- Retraining with a task-specific dataset achieve higher accuracy
  - Can lower accuracy when using dissimilar datasets

Fine-Tuning for Search

Results before & after retraining





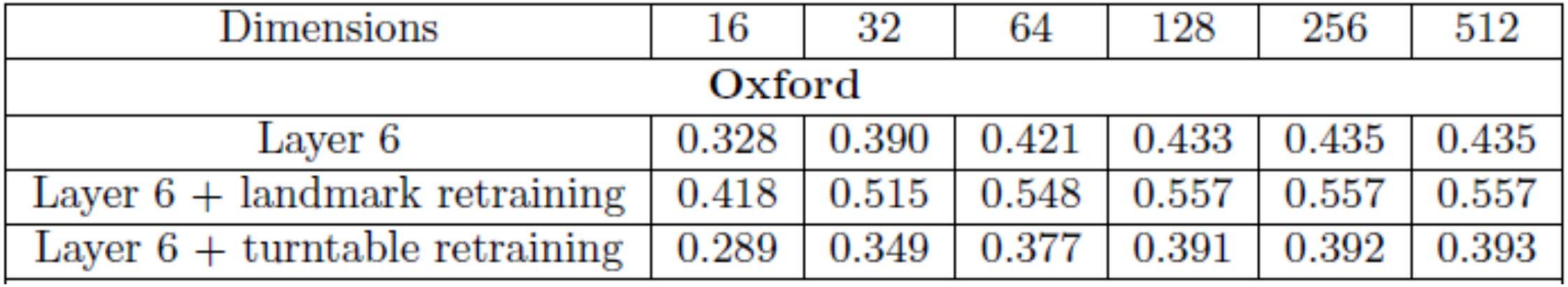
Neural codes trained on ILSVRC					
Layer 5	9216	0.389	—	0.690*	3.09
Layer 6	4096	0.435	0.392	0.749*	3.43
Layer 7	4096	0.430	—	0.736*	3.39
After retraining on the Landmarks dataset					
Layer 5	9216	0.387	—	0.674*	2.99
Layer 6	4096	0.545	0.512	**0.793***	3.29
Layer 7	4096	0.538	—	0.764*	3.19
After retraining on turntable views (Multi-view RGB-D)					
Layer 5	9216	0.348	—	0.682*	3.13
Layer 6	4096	0.393	0.351	0.754*	3.56
Layer 7	4096	0.362	—	0.730*	3.53

Landmark dataset has similar images to Oxford

Ack.: Neural Codes for Image Retrieval

Dimension Reduction

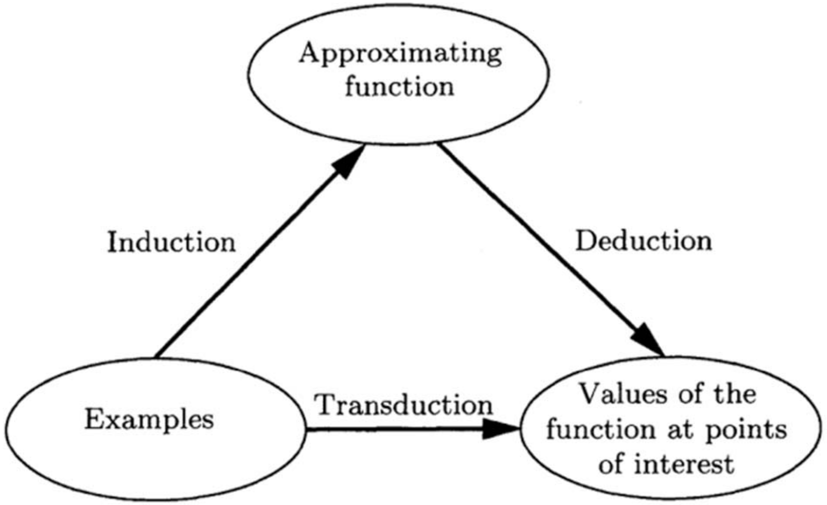
- **CNN features (4096D) are robust to PCA compression**
  - **Maintain accuracy by 256 D**

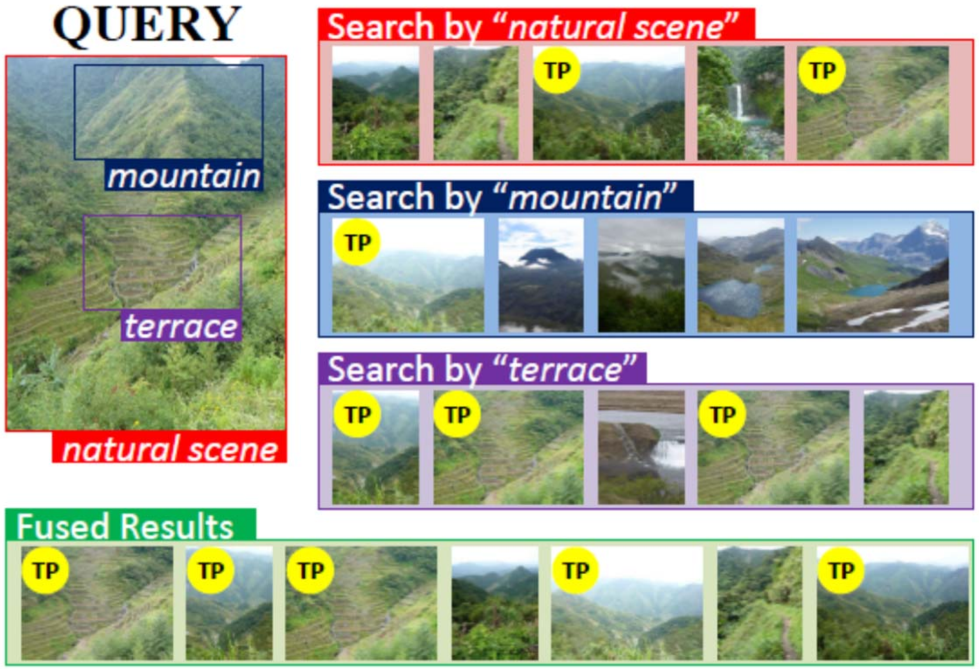


Dimensions	16	32	64	128	256	512
Oxford						
Layer 6	0.328	0.390	0.421	0.433	0.435	0.435
Layer 6 + landmark retraining	0.418	0.515	0.548	0.557	0.557	0.557
Layer 6 + turntable retraining	0.289	0.349	0.377	0.391	0.392	0.393

Image Classification and Retrieval are ONE [ICMR 15]

- Handle the classification and search in a unified framework
  - Uses region proposals, and nearest neighbor search for both problems
- Image search (kNN) is transductive learning





Regional Attention Based Deep Feature for Image Retrieval

- **Apply the attention (or saliency) to regional features for image retrieval**
  - **Train attention weights based on classification**

(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*'

Ack. Tech talk

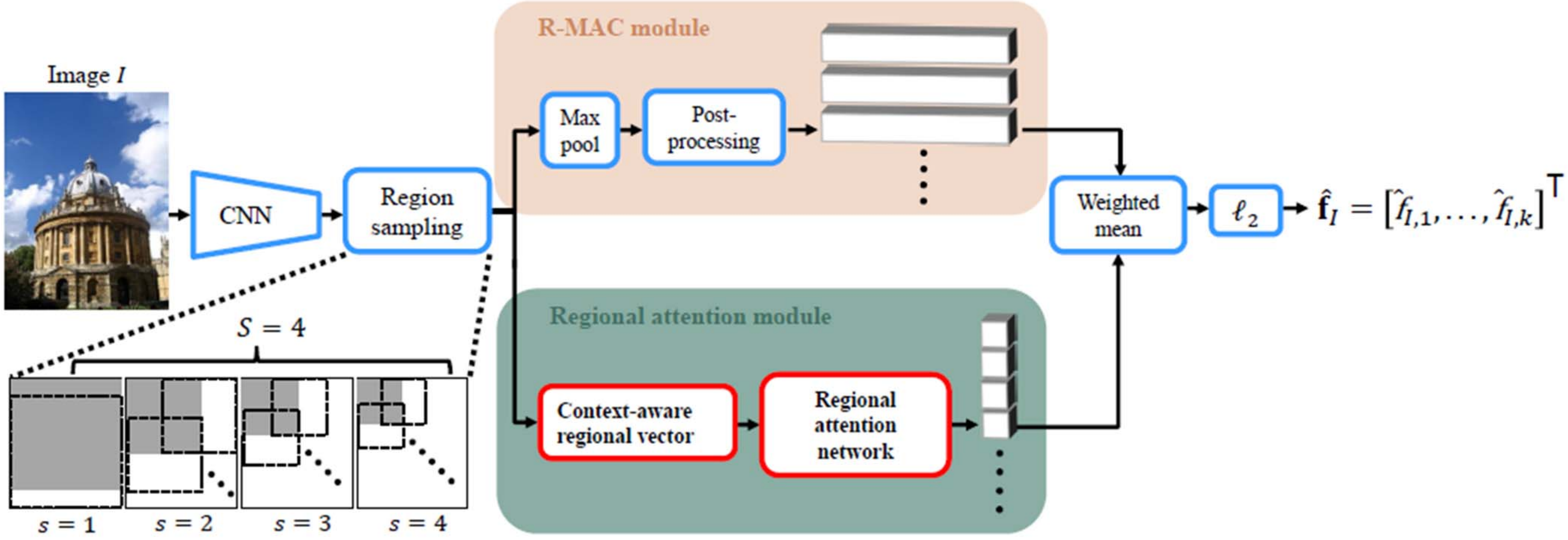


Image I
CNN
Region sampling
$S = 4$
$s = 1$
$s = 2$
$s = 3$
$s = 4$
R-MAC module
Max pool
Post-processing
Regional attention module
Context-aware regional vector
Regional attention network
Weighted mean
ℓ_2
$\hat{\mathbf{f}}_I = [\hat{f}_{I,1}, \ldots, \hat{f}_{I,k}]^T$

HardNet: Deep Learning based Local Features

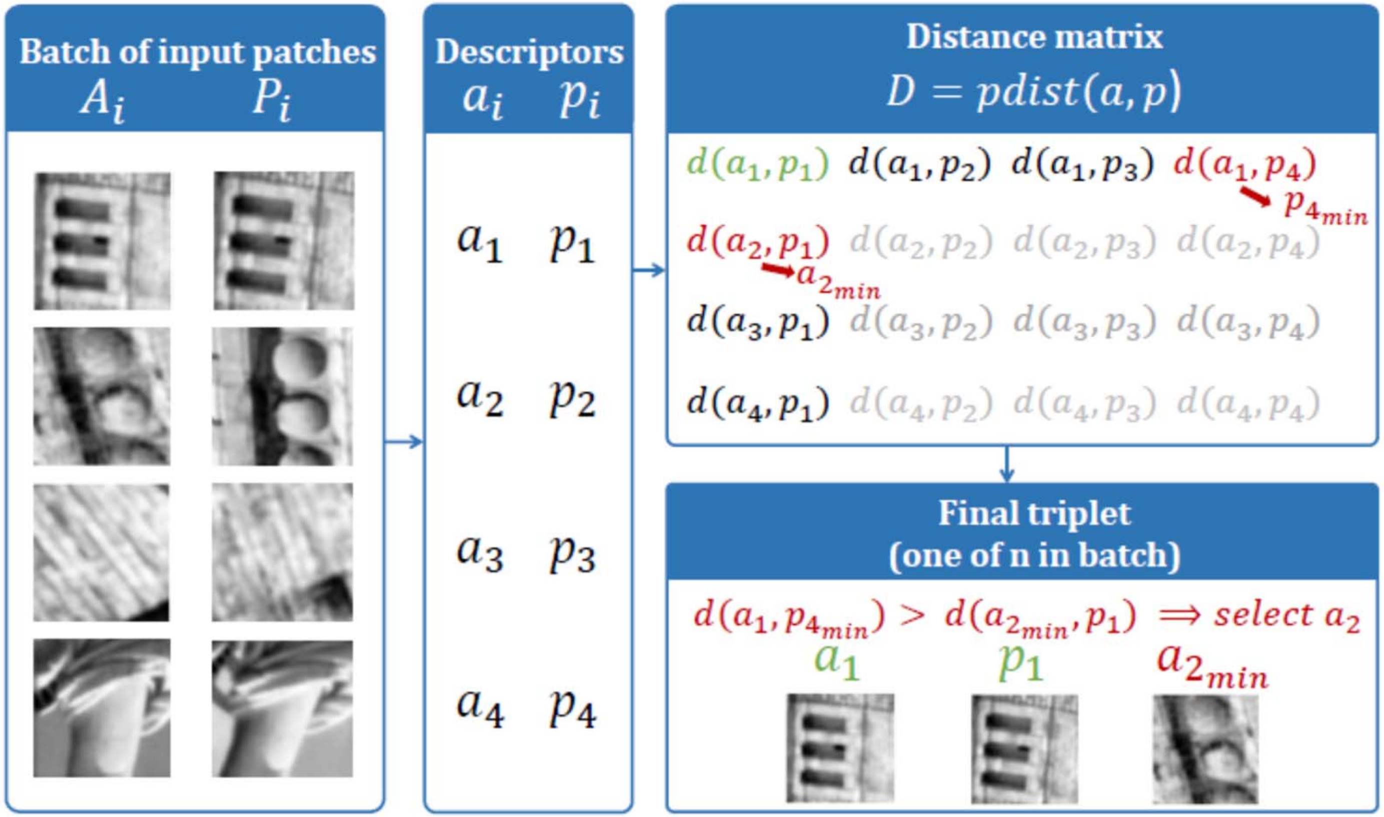
- Propose a local descriptor learning loss
  - Similar to a triplet loss
  - Get a higher matching accuracy than SIFT

- Triplet loss w/ anchor, its positive, and its negative
  - Compute feature in a way: $\boldsymbol{D(a,p) < D(a,n)}$

Working hard to know your neighbor's margins: Local descriptor learning loss, NIPS

Sampling Procedure

- **Given an anchor patch a_1, we extract its positive patch p_1**
  - **Use traditional matching techniques (e.g., DoG)**
- **Find its hard negative**


Batch of input patches
A_i P_i
Descriptors
a_i p_i
a_1 p_1
a_2 p_2
a_3 p_3
a_4 p_4
Distance matrix
$D = pdist(a,p)$
$d(a_1,p_1)$ $d(a_1,p_2)$ $d(a_1,p_3)$ $d(a_1,p_4)$
$p_{4_{min}}$
$d(a_2,p_1)$ $d(a_2,p_2)$ $d(a_2,p_3)$ $d(a_2,p_4)$
$a_{2_{min}}$
$d(a_3,p_1)$ $d(a_3,p_2)$ $d(a_3,p_3)$ $d(a_3,p_4)$
$d(a_4,p_1)$ $d(a_4,p_2)$ $d(a_4,p_3)$ $d(a_4,p_4)$
Final triplet
(one of n in batch)
$d(a_1,p_{4_{min}}) > d(a_{2_{min}},p_1) \Rightarrow select\ a_2$
a_1 p_1 $a_{2_{min}}$

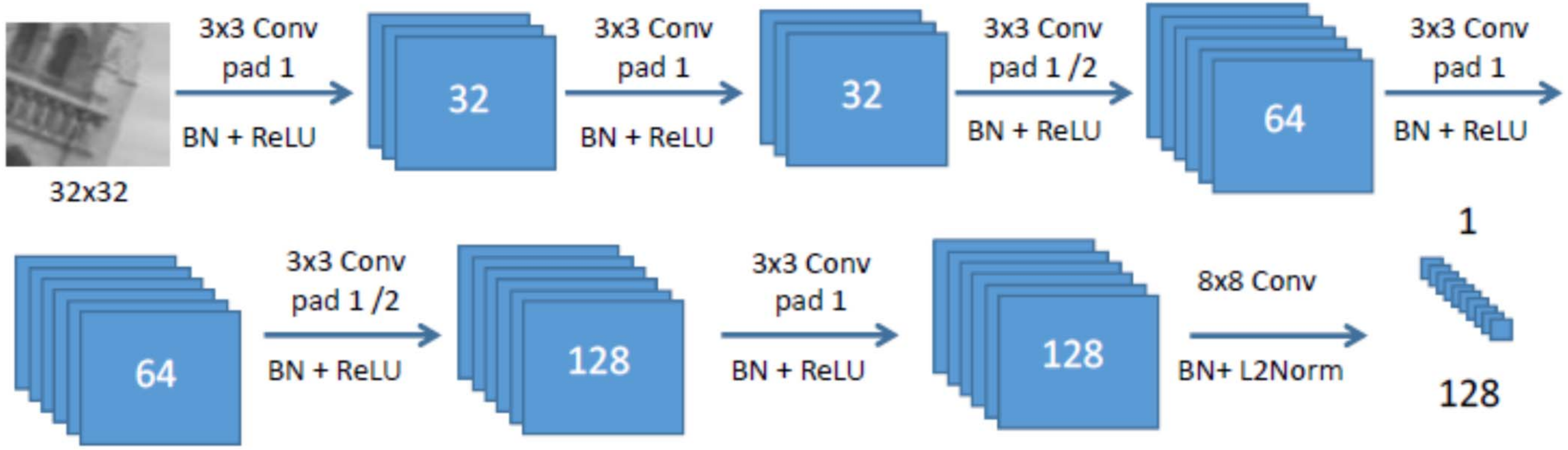
Find a patch that is incorrectly close to a_1

Find a patch that is incorrectly close to p_1

Between two patches, pick the worst

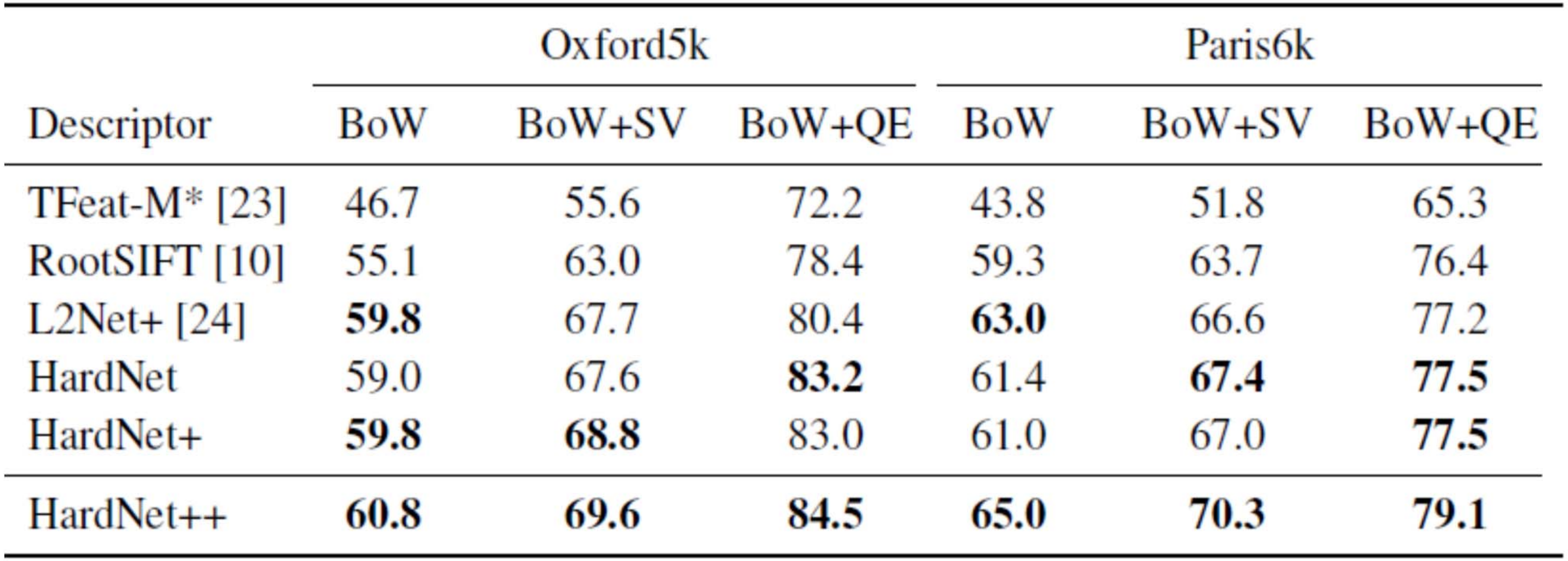
Model Architecture

- Input: 32x32 grayscale input patches
- Output: 128D descriptor



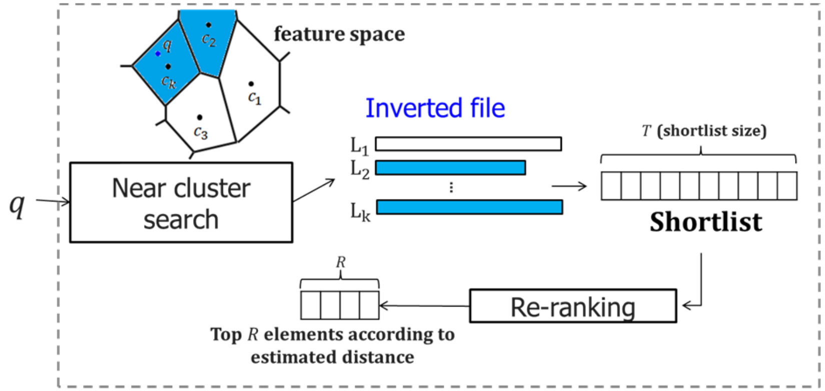
Performance Comparisons over Prior Features

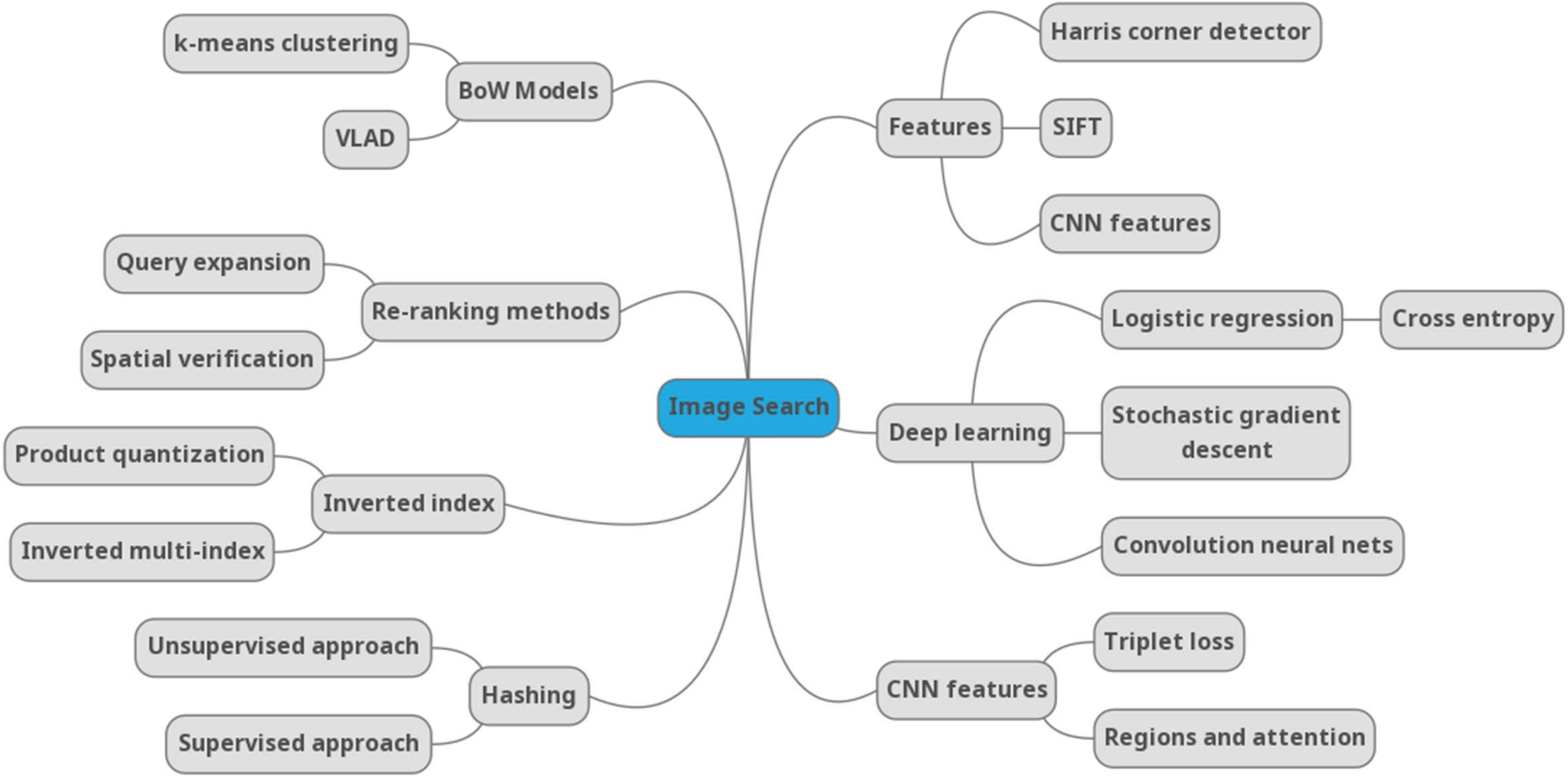
- **Overall, it shows better accuracy, as it is trained with additional datasets**
  - **BoW: Bag-of-Words, QE: Query Expansion, SV: Spatial Verification**



	Oxford5k			Paris6k		
Descriptor	BoW	BoW+SV	BoW+QE	BoW	BoW+SV	BoW+QE
TFeat-M* [23]	46.7	55.6	72.2	43.8	51.8	65.3
RootSIFT [10]	55.1	63.0	78.4	59.3	63.7	76.4
L2Net+ [24]	**59.8**	67.7	80.4	**63.0**	66.6	77.2
HardNet	59.0	67.6	**83.2**	61.4	**67.4**	**77.5**
HardNet+	**59.8**	**68.8**	83.0	61.0	67.0	**77.5**
HardNet++	**60.8**	**69.6**	**84.5**	**65.0**	**70.3**	**79.1**

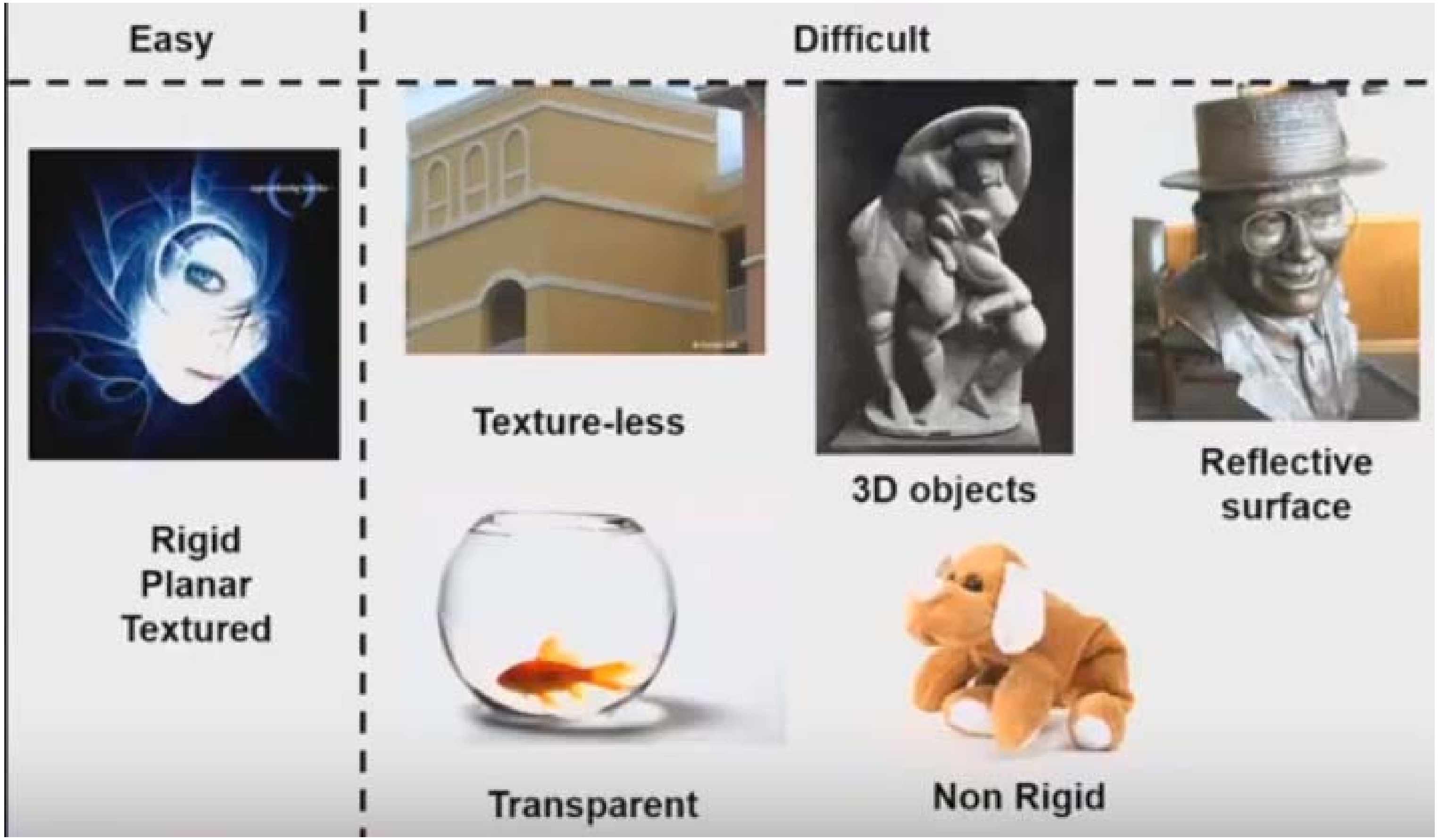
Summary





- Image Search
  - BoW Models
    - k-means clustering
    - VLAD
  - Re-ranking methods
    - Query expansion
    - Spatial verification
  - Inverted index
    - Product quantization
    - Inverted multi-index
  - Hashing
    - Unsupervised approach
    - Supervised approach
  - Features
    - Harris corner detector
    - SIFT
    - CNN features
  - Deep learning
    - Logistic regression
      - Cross entropy
    - Stochastic gradient descent
    - Convolution neural nets
  - CNN features
    - Triplet loss
    - Regions and attention

Limitations of Image Search



Ack: Vijay Chandrasekhar

- **Large-scale video retrieval**
  - **30 frames per sec., 5 billion shared video at youtube**

Applications and Extension of Image Search

- Content and context based hashing, indexing, search and retrieval of multimedia data
- Multimodal or cross-modal content analysis and retrieval
- Advanced descriptors and similarity metrics for multimedia data
- Complex multimedia event detection and recounting

Ack: Call for papers of ACM ICMR

Applications and Extension of Image Search

- Learning and relevance feedback and HCI issues in multimedia retrieval
- Query models and languages for multimedia retrieval
- Fine-grained visual search
- Image/video summarization and visualization
- Mobile visual search

Class Objectives were:

- CNN based approaches
  - Consider different regions within or outside the end-to-end training
  - Utilize attention and local features
  - Discuss applications
- Discussed limitations of current techniques and future research directions

Homework for Every Class

- **Come up with one question on what we have discussed today**
  - **Write questions three times**
- **Go over recent papers on image search, and submit their summary before Tue. class**