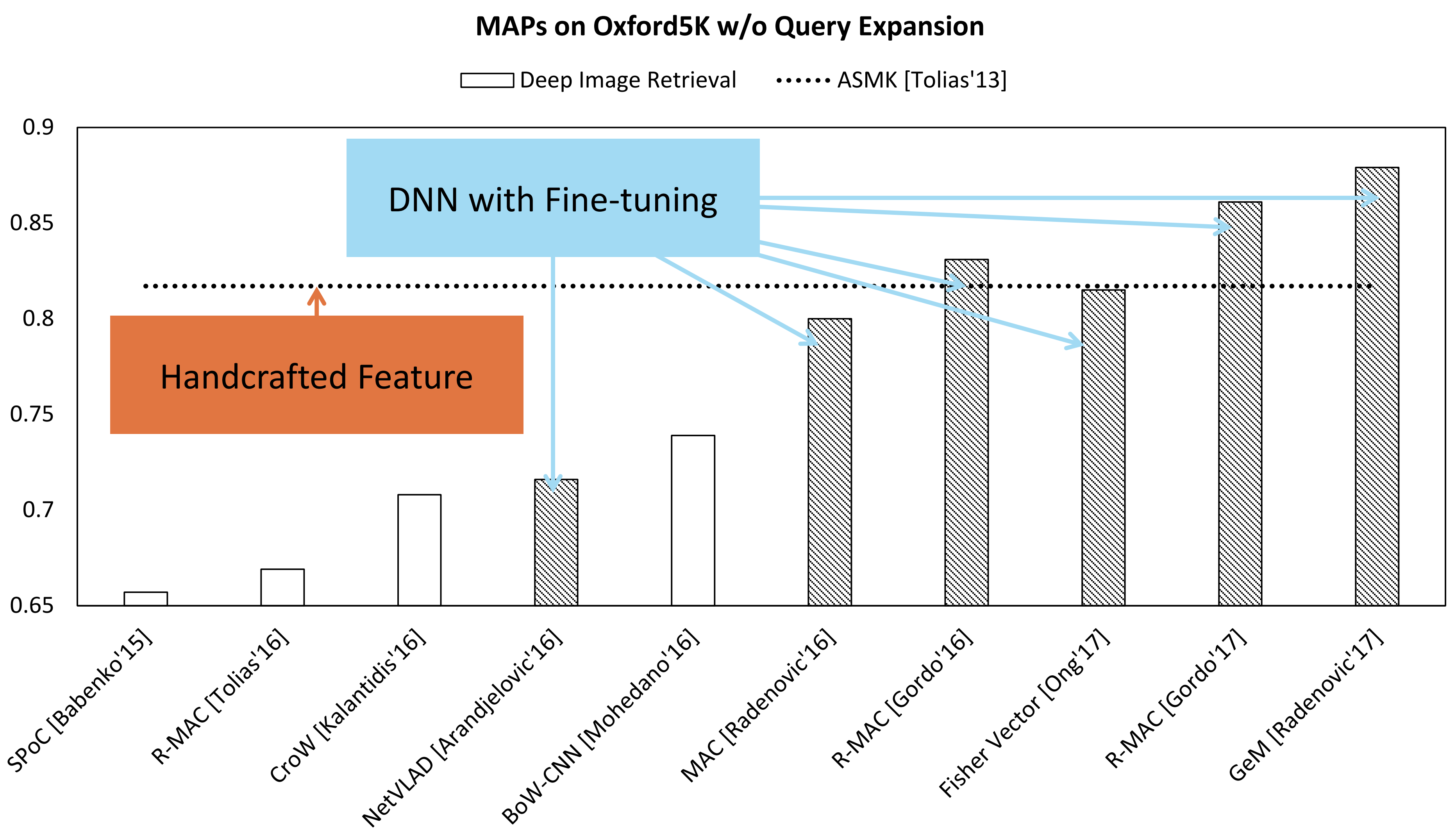
Weighted Generalized Mean Pooling for Deep Image Retrieval

Xiaomeng Wu, Go Irie, Hiramatsu Kaoru, and Kunio Kashino

NTT Communication Science Laboratories

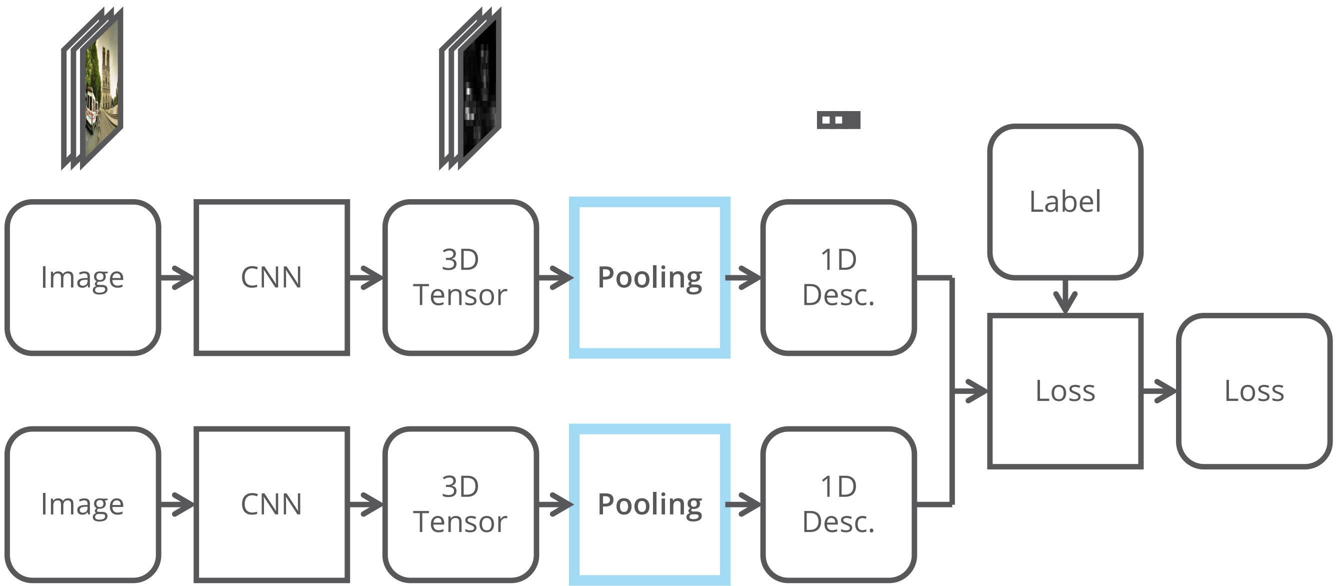
NTT Corporation, Japan

Fine-tuning for Deep Image Retrieval



Siamese Network

- **Crucial components of deep image retrieval include**
  - A good pre-trained convolutional network
  - A good pooling method
  - A ranking loss



Related Works

- **FC Layers** **[Babenko'14][Gong'14]**
- **Global Pooling**
  - Sum [Babenko'15]
  - Maximum [Radenovic'16][Tolias'16][Razavian'16]
  - Generalized Mean (GeM) [Radenovic'17]
- **Semi-local Regional Pooling**
  - R-MAC [Tolias'16]
  - Region Proposal Network (RPN) [Gordo'17]
- **Widely-used Encoding Techniques**
  - Bag of Visual Words [Mohedano'16]
  - VLAD [Arandjelovic'16]
  - Fisher Vector [Ong'17]

UNIFORM POOLING
Each activation contributes equivalently to the construction of a global representation.

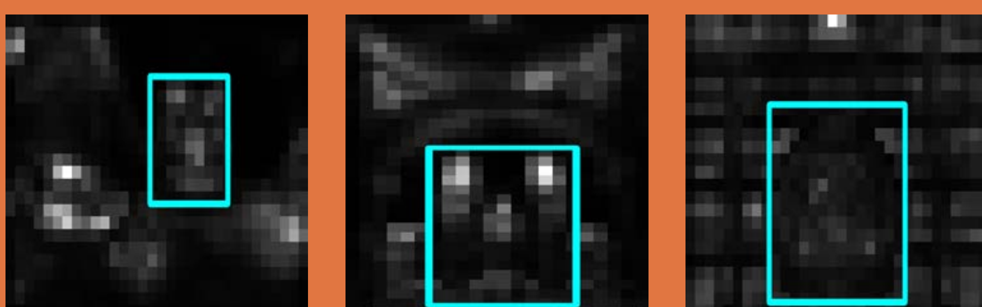
PROBLEM
Uniform pooling suffers from the presence of activations, e.g., from background clutter, that play a negative role as regards matching.

CNN Activations vs. Objectness

- **CNN Firing to Background Clutter**

Images from Oxford5K

Activation Strengths



Objects of interest (landmarks) predefined in Oxford5K

The CNN has been fine-tuned on landmark images.

Goal of the Study

- **Problem of Uniform Pooling**
  - Each activation contributes equivalently to the construction of a global representation.
- **Proposal**
  - Exploit a spatial weighting mechanism for pooling
    - Predict a weight that describes how discriminating each activation at each location is as regards image matching.
  - Lead to the end-to-end learning based on a weighted generalized mean (wGeM) pooling method

Spatial Weighting + Pooling + Learning to Rank

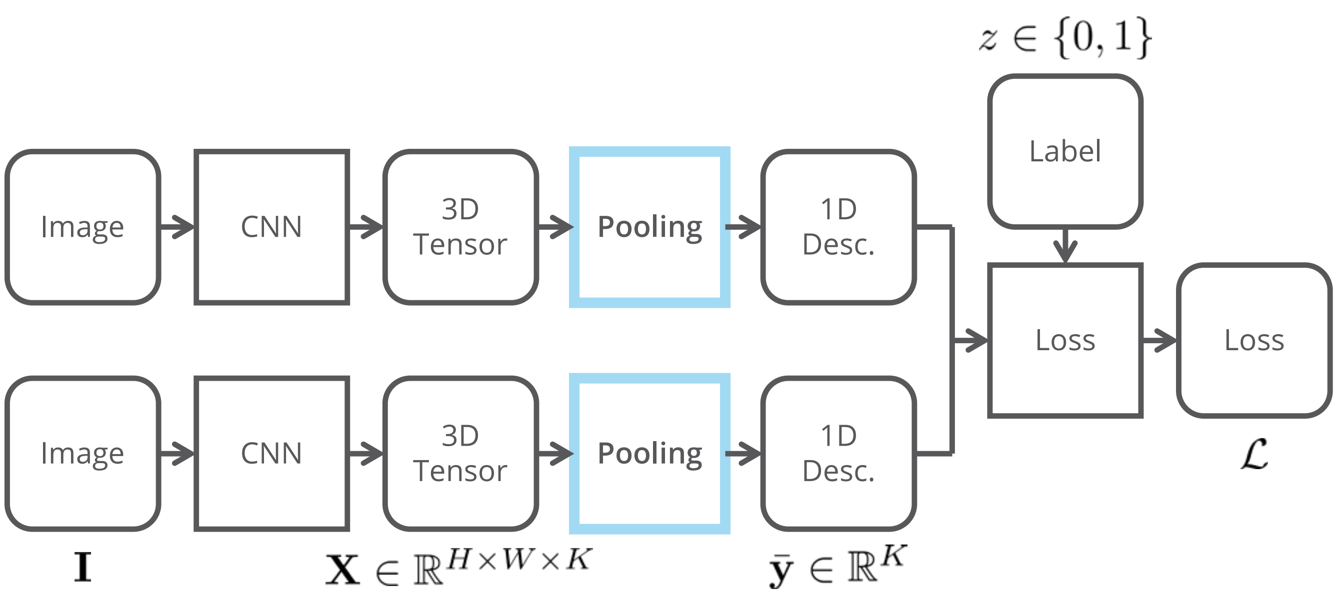
NTT

Proposed Method

Network Architecture and Learning

- **Contrastive Loss** [Chopra'05][Radenovic'16]

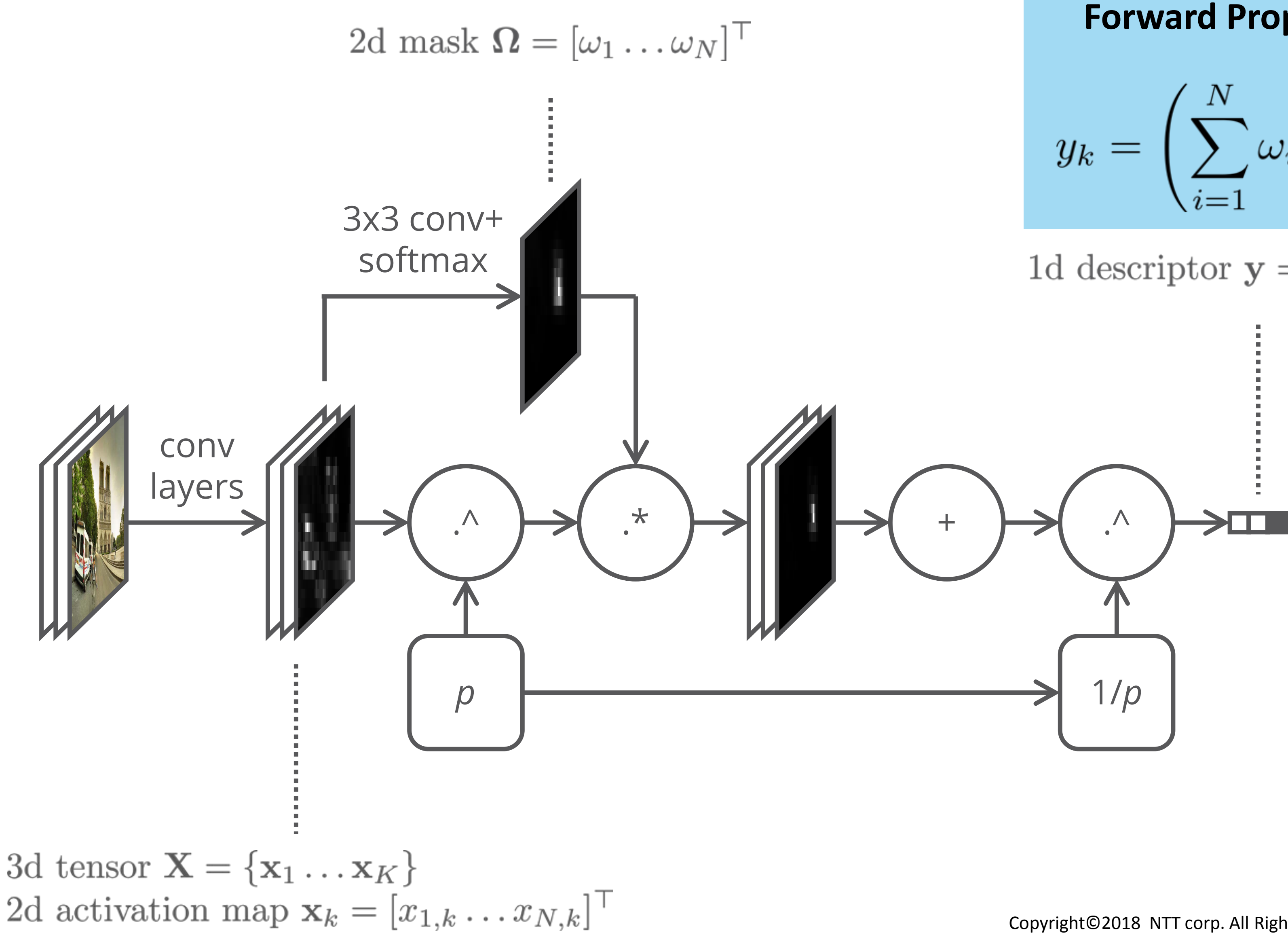
$$\mathcal{L}_{i,j} = \begin{cases} \frac{1}{2}\|\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j\|_2^2 & \text{if } z_{i,j} = 1 \\ \frac{1}{2}[\max(0, \tau - \|\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j\|_2)]^2 & \text{otherwise} \end{cases}$$



Weighted Generalized Mean (wGeM) Pooling

Forward Propagation

$$y_k = \left(\sum_{i=1}^{N} \omega_i x_{i,k}^{p} \right)^{1/p}$$



Derivatives and Property

- **Derivatives**

$$\frac{\partial y_k}{\partial x_{i,k}} = \sum_{j=1}^{N} \frac{\partial y_k}{\partial \omega_j} \frac{\partial \omega_j}{\partial x_{i,k}} + \omega_i \left(\frac{x_{i,k}}{y_k} \right)^{p-1}$$

$$\frac{\partial y_k}{\partial \omega_i} = \frac{x_{i,k}}{p} \left(\frac{x_{i,k}}{y_k} \right)^{p-1}$$

$$\frac{\partial y_k}{\partial p} = \frac{y_k}{p} \left(\frac{\sum_i \omega_i x_{i,k}^p \log x_{i,k}}{y_k^p} - \log y_k \right)$$

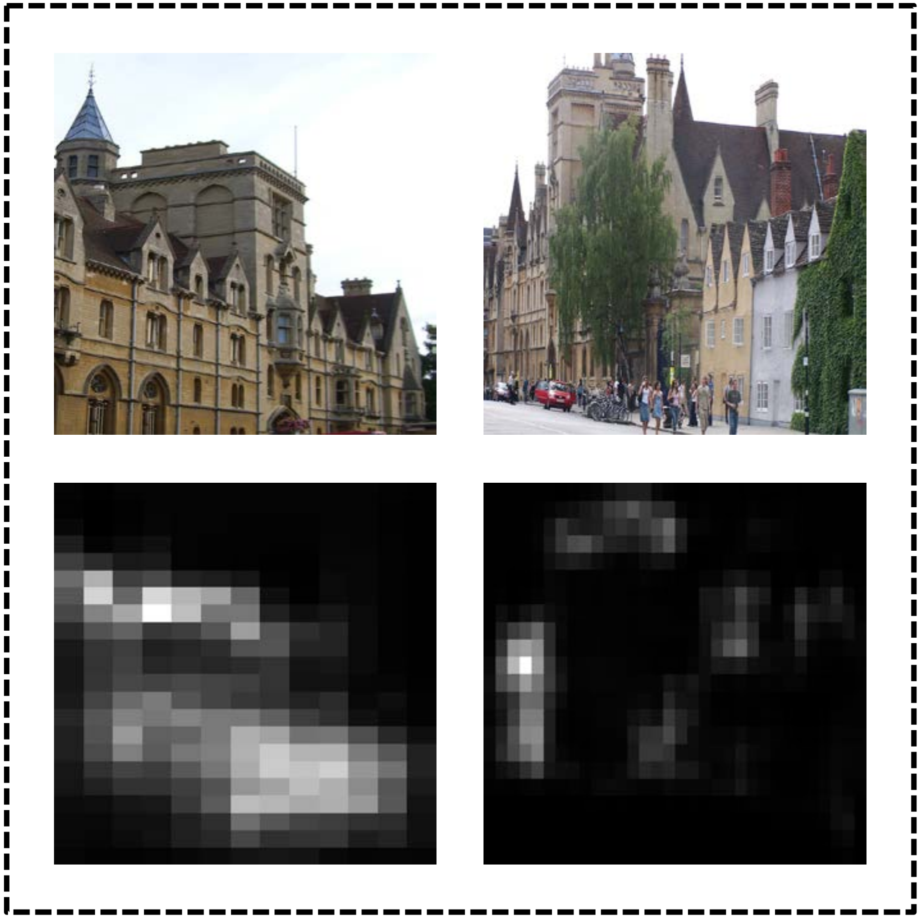
- **Behavior when $p \to \infty$**
  - Forward

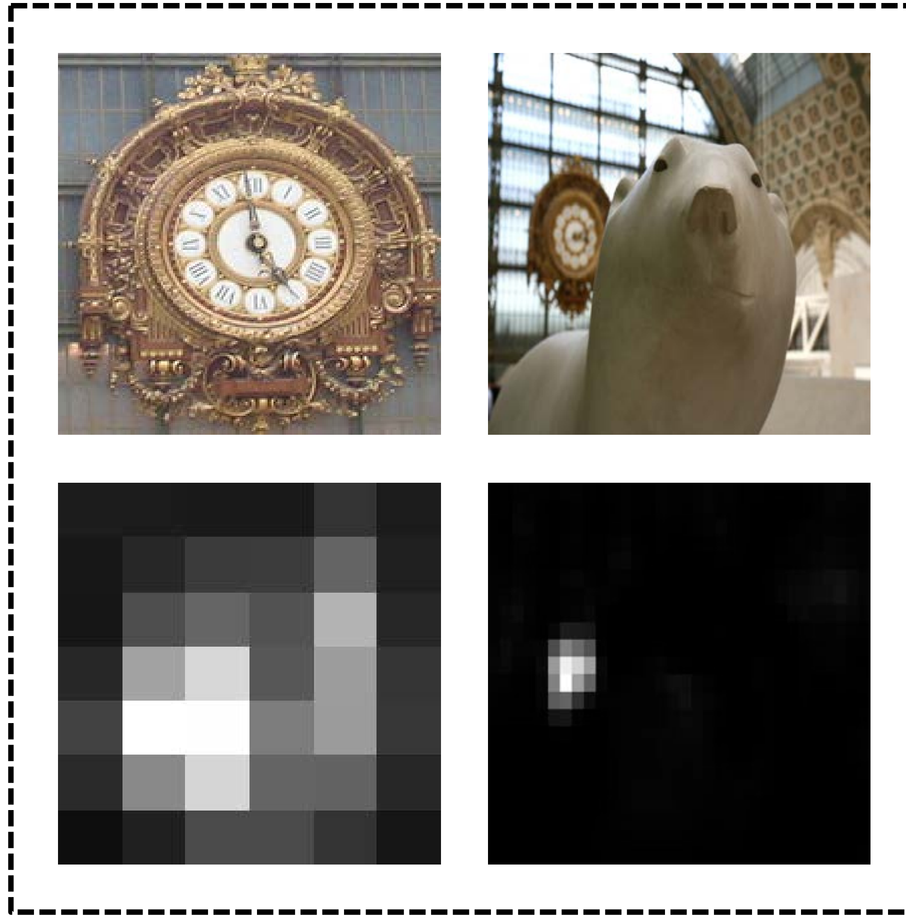
$$\lim_{p \to \infty} y_k = \lim_{p \to \infty} \left(\sum_{i=1}^{N} \omega_i x_{i,k}^p \right)^{1/p} = \max_i x_{i,k}$$

  - Back-propagation

$$\lim_{p \to \infty} \frac{\partial \mathcal{L}}{\partial \omega_i} = \lim_{p \to \infty} \frac{1}{p} \sum_{k=1}^{K} \frac{\partial \mathcal{L}}{\partial y_k} \left(\frac{x_{i,k}}{y_k} \right)^{p-1} x_{i,k} = 0$$

Matching Images and Spatial Weights





Experiments

Experimental Setup

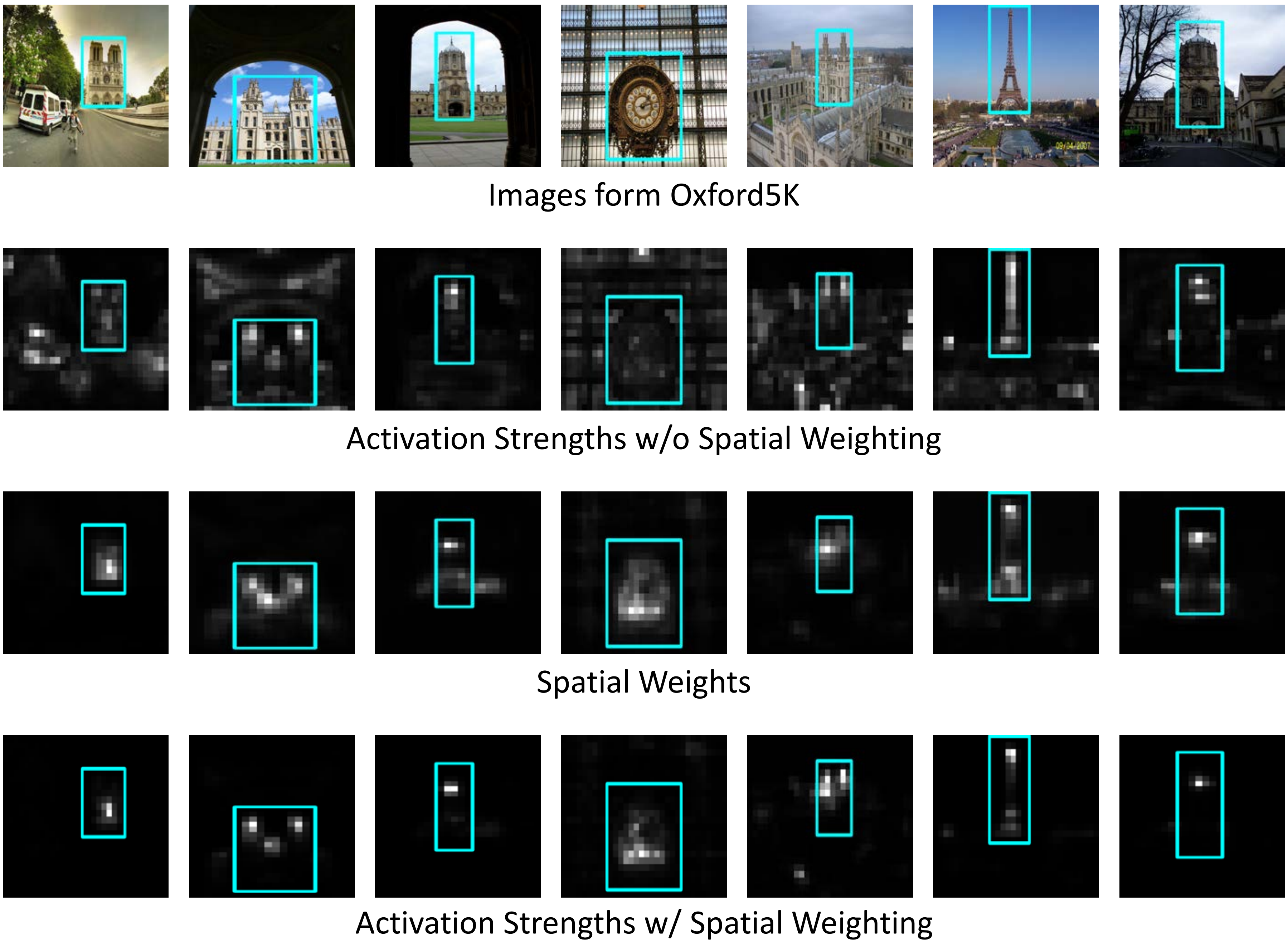
Dataset for Training	36K pairs out of 163K+ landmark images [Radenovic'16]
Dataset for Testing	Oxford5K, Paris6K, & Oxford105K
Backbone Network	ResNet101 [He'16]
#epoch	30
Batch Size	30 pairs (#pos./#neg.: 5/25)
Learning Rate	$10^{-6}e^{-0.1i}$ over epoch i
Margin of Contrastive Loss	0.85
Learning Method	Adam [Kingma'14]
Pre- and Post-processing	Multi-scale representation & supervised whitening [Radenovic'18]
Performance Measure	Mean Average Precision (MAP)

- **Implementation of wGeM**
  - Initialize the parameters of the 3×3 conv layer with 0s such that $\forall \omega_i \in \boldsymbol{\Omega}, \omega_i = \frac{1}{N}$
  - Use learning rates that are 10 times as large as those of the pre-trained ResNet

MAPs for Different Initializations of p

	Initial p	Oxford5K	Paris6K
GeM [Radenovic'18]	3	87.8	92.7
wGeM	2	88.6	92.2
	3	88.9	92.3
	4	88.8	92.5
	5	88.4	92.6

The MAP of the wGeM was slightly poorer for Paris6K because the training set is composed of more historical landmarks, while Paris6K contains more contemporary buildings.

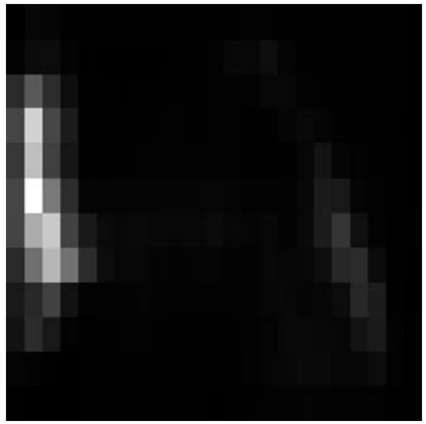

Images form Oxford5K
Activation Strengths w/o Spatial Weighting
Spatial Weights
Activation Strengths w/ Spatial Weighting

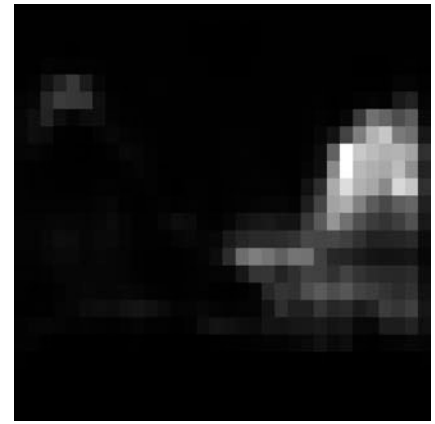
Comparison with State of the Art

	CNN	QE	Oxford5K	Oxford105K	Paris6K
NetVLAD [Arandjelovic'16]	VGG	n	71.6	n/a	79.7
MAC [Radenovic'16]	VGG	n	80.0	75.1	82.9
Fisher Vector [Ong'17]	VGG	n	81.5	76.6	82.4
R-MAC [Gordo'17]	ResNet	n	86.1	82.8	**94.5**
GeM [Radenovic'18]	ResNet	n	87.8	84.6	92.7
*** wGeM**	ResNet	n	**88.8**	**85.6**	92.5
MAC+QE [Radenovic'16]	VGG	y	85.4	82.3	87.0
R-MAC+QE [Gordo'17]	ResNet	y	90.6	89.4	**96.0**
GeM+αQE [Radenovic'18]	ResNet	y	91.0	89.5	95.5
*** wGeM+QE**	ResNet	y	**91.7**	**89.7**	**96.0**

The wGeM outperforms or is on a par with the state of the art based on fine-tuned deep networks.

Spatial Weighting Failure





Here, the object of interest, the Louvre Pyramid, was largely ignored by the wGeM because regular patterns are potentially less discriminating when matching the landmarks in the training set. Therefore, the wGeM was trained to fire less for such regions.

Innovative R&D by NTT

Conclusions

Conclusions

- **Characteristics of Proposed Method**
  - Generalize sum, max, GeM pooling, and CroW [Kalantidis'16].
  - Trainable
  - Require no bounding box annotations for training
- **Future Directions**
  - Complementarity with state-of-the-art QE techniques, e.g., diffusion [Iscen'17] [Wu'18].
  - How do different wGeM block structures affect the learning of deep representations?

Innovative R&D by NTT

Thank you for your attention.

Xiaomeng Wu, Go Irie, Hiramatsu Kaoru, and Kunio Kashino

NTT Communication Science Laboratories

NTT Corporation, Japan