Deep Aggregation of Regional Convolutional Activations for Content Based Image Retrieval

Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Klaus Jung

Visual Computing Group, HTW Berlin

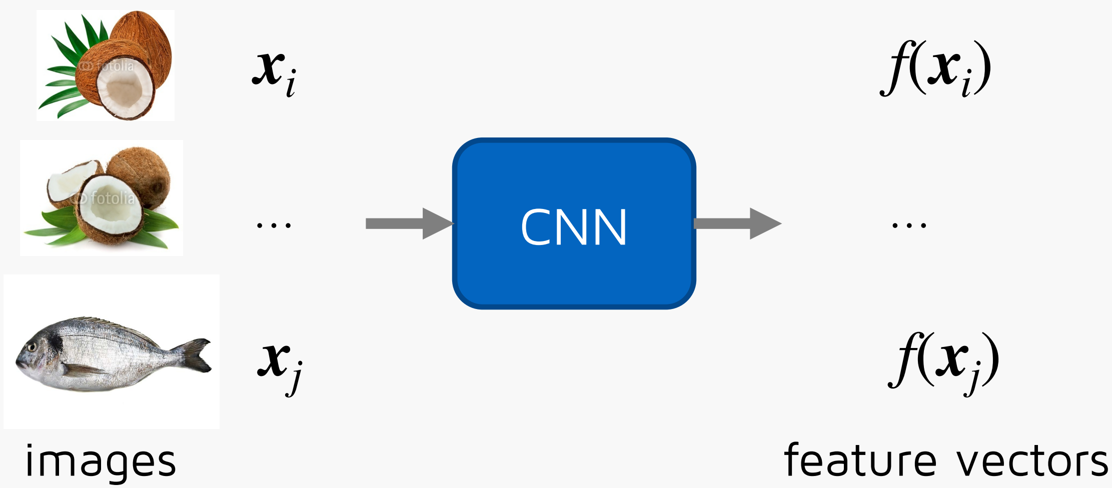
visual-computing.com

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

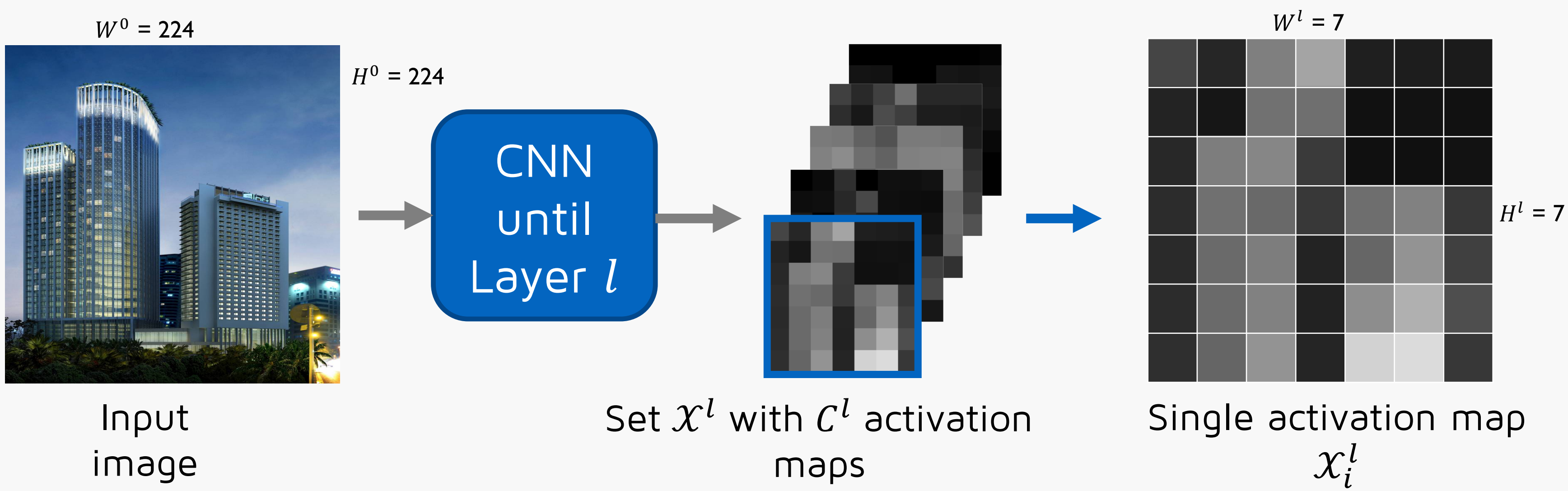
Challenges in Image Retrieval

1. How to generate feature vectors?



Activation Maps by CNNs

- Layers of Convolutional Neural Networks produce a set of activation maps $\mathcal{X}^l = \{\mathcal{X}_i^l | i = 1 \dots C^l\}$, with C^l being the number of filters in layer l.

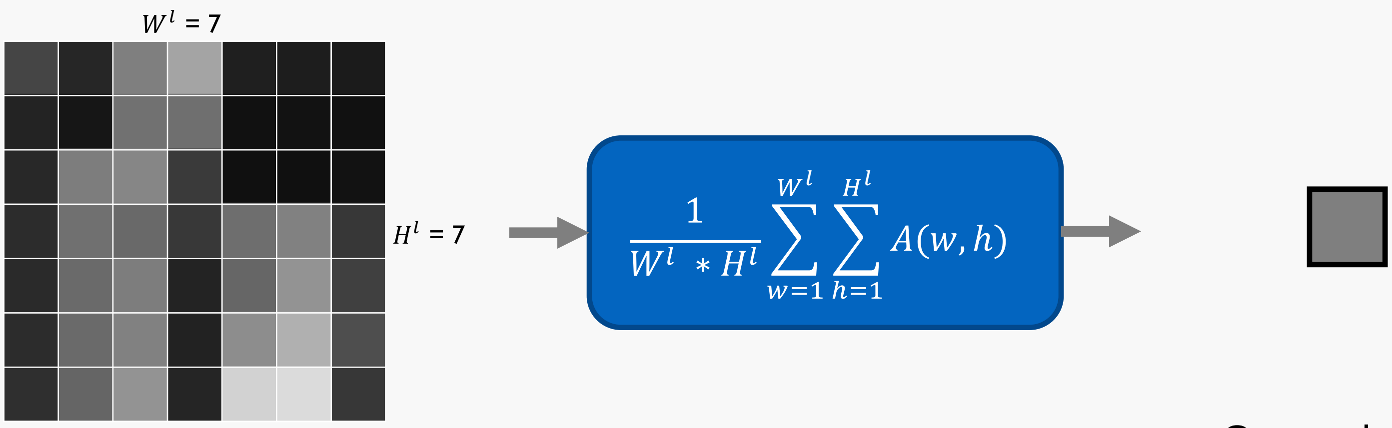


- Number of filters C is increasing while the width W and the height H are decreasing in deeper layers

Generation of Feature Vectors

1. Extraction of activation maps $\mathcal{X}^l$ where l is the last convolutional layer in a CNN

2. Compute the average of C activation maps $A \in \mathcal{X}^l$ with $A(w, h)$ being a single activation

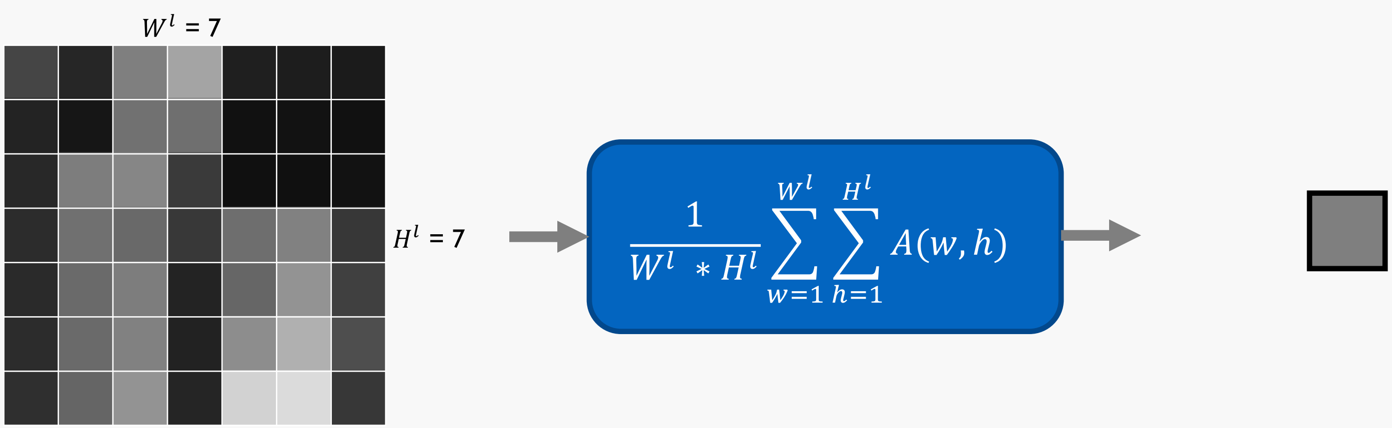


Activation map A

One value of C-dimensional feature vector

Generation of Feature Vectors

1. Extraction of activation maps $\mathcal{X}^l$ where l is the last convolutional layer in a CNN
2. Compute the average of C activation maps $A \in \mathcal{X}^l$ with $A(w,h)$ being a single activation



Activation map A

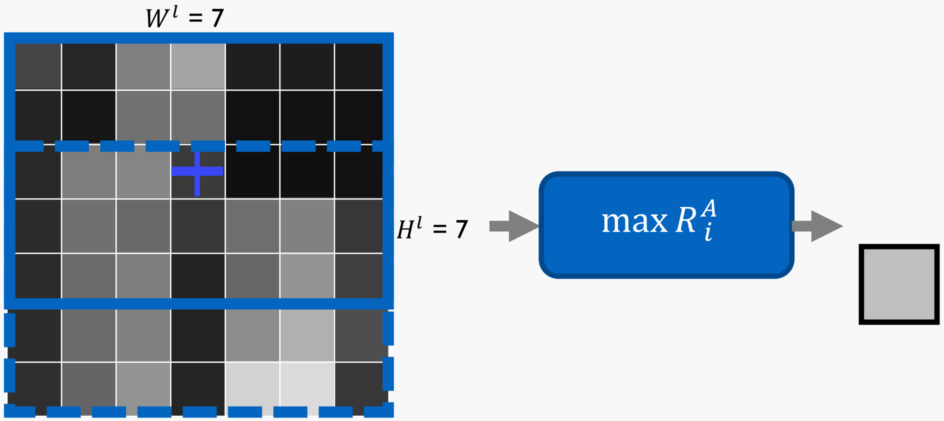
One value of C-dimensional feature vector

➡ Spatial information is completely lost

Regional Max Pooling

RMAC: Regional Maximal Activations of Convolutions (Tolias et al., 2016)

1. Extraction of activation maps $\mathcal{X}^l$
2. Search for maximal activation values in a set of 20 regions $R^A = \{R_i^A | i = 1 \dots 20\}$, for each $A \in \mathcal{X}^l$

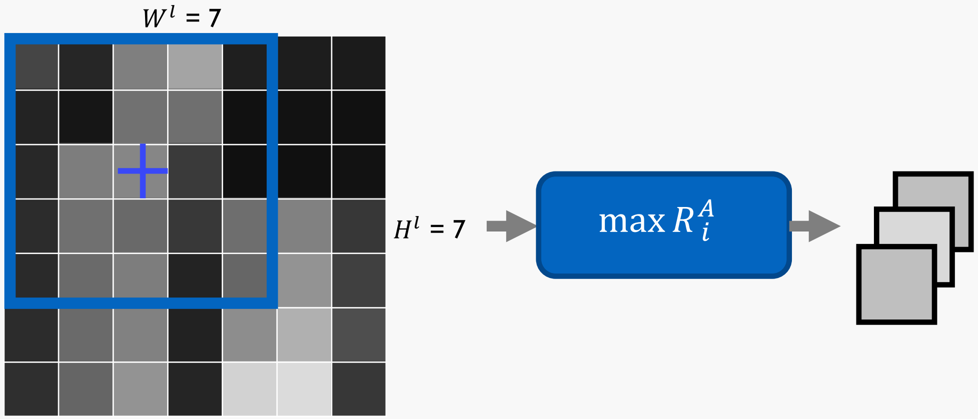


Activation map A

Regional Max Pooling

RMAC: Regional Maximal Activations of Convolutions (Tolias et al., 2016)

1. Extraction of activation maps $\mathcal{X}^l$
2. Search for maximal activation values in a set of 20 regions $R^A = \{R_i^A | i = 1 \dots 20\}$, for each $A \in \mathcal{X}^l$

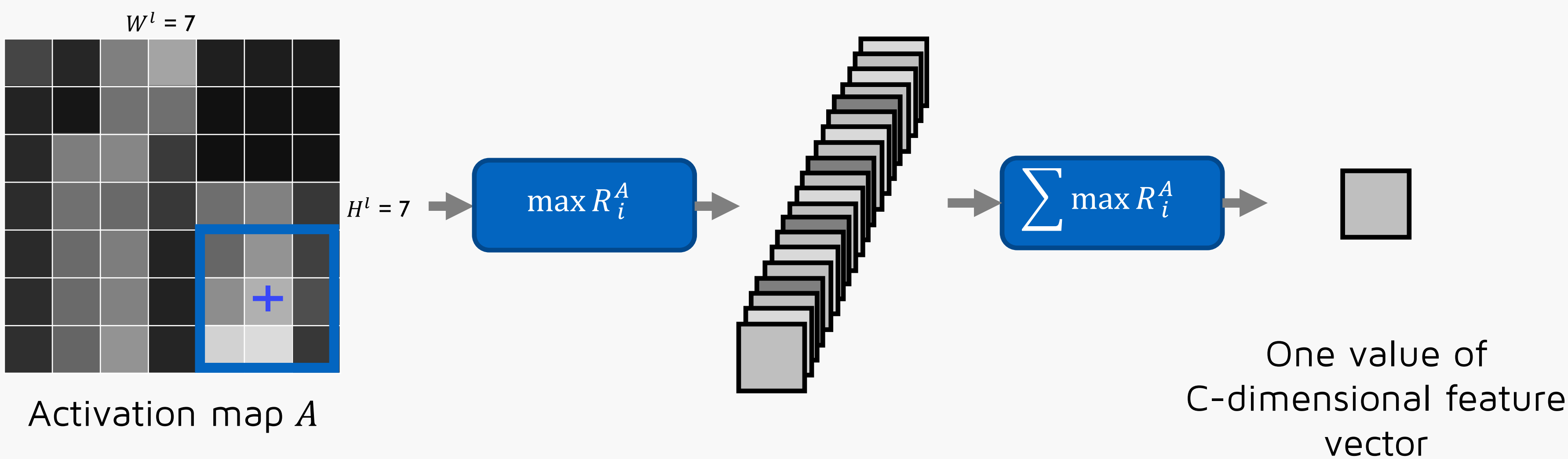


Activation map A

Regional Max Pooling

RMAC: Regional Maximal Activations of Convolutions (Tolias et al., 2016)

1. Extraction of activation maps $\mathcal{X}^l$
2. Search for maximal activation values in a set of 20 regions $R^A = \{R_i^A | i = 1 \dots 20\}$, for each $A \in \mathcal{X}^l$



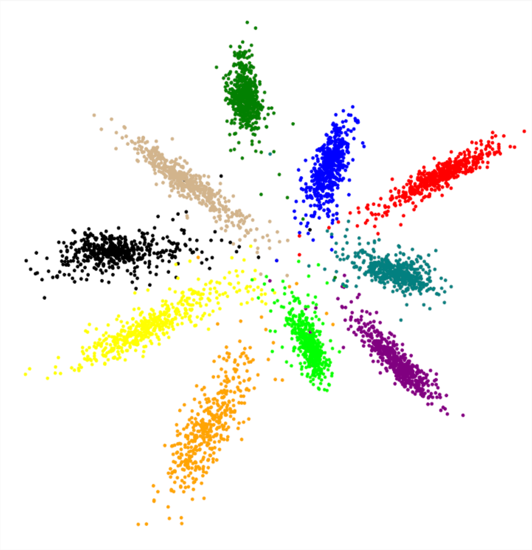
Challenges in Image Retrieval

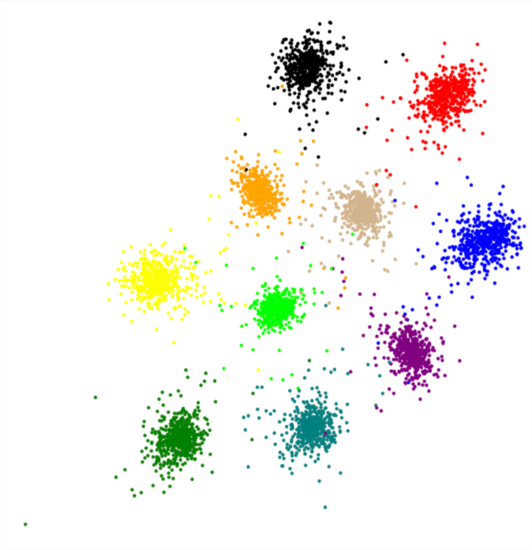
1. How to generate feature vectors?
2. How to optimize model weights for the specific task of retrieval?

Challenges in Image Retrieval

1. How to generate feature vectors?
2. How to optimize model weights for the specific task of retrieval?

2D embeddings for MNIST trained with *Softmax Cross Entropy*





2D embeddings for MNIST trained with *Triplet Loss*

Finetuning for Retrieval

- DIR: Deep Image Retrieval (Gordo et al., 2017)
- End-to-End image retrieval system.
- Finetune a ResNet101 with Triplet loss
- Total training time: approximately 168 hours

Input image with 800 pixels at the smaller side

CNN until Layer L → Region Max Pooling → PCA Whitening → $\sum$ → Triplet Loss

$\mathcal{X}^L$ of size $(25 \times 25 \times 2048)$

Tensor of size (20×2048)

Tensor of size (20×2048)

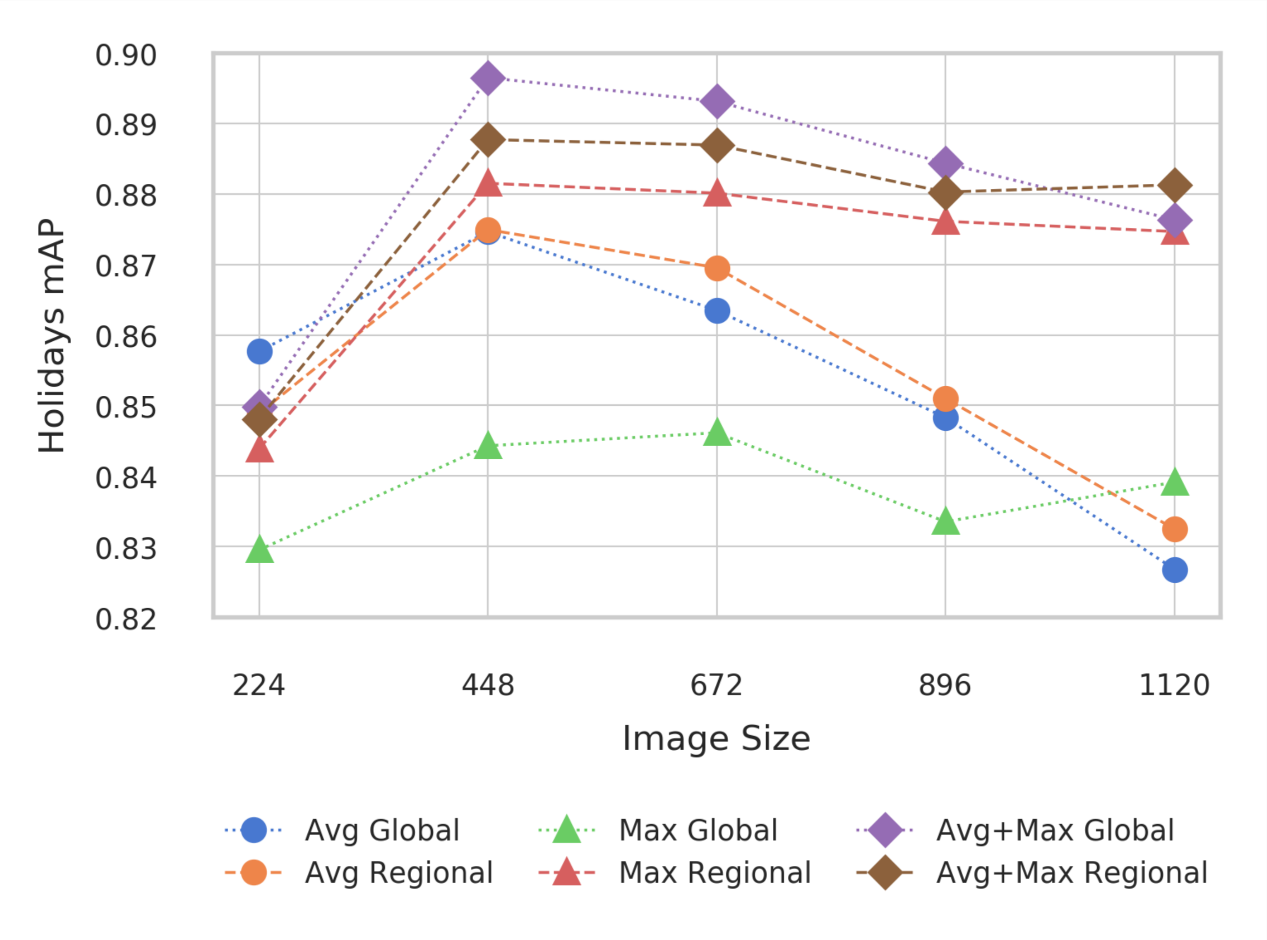
Feature Vector of size 2048

Researched Questions

1. Why only regional max pooling?
2. Are all regions equally important?
3. Is Triplet loss the ideal loss function?
4. Is an increased image size necessary during training?

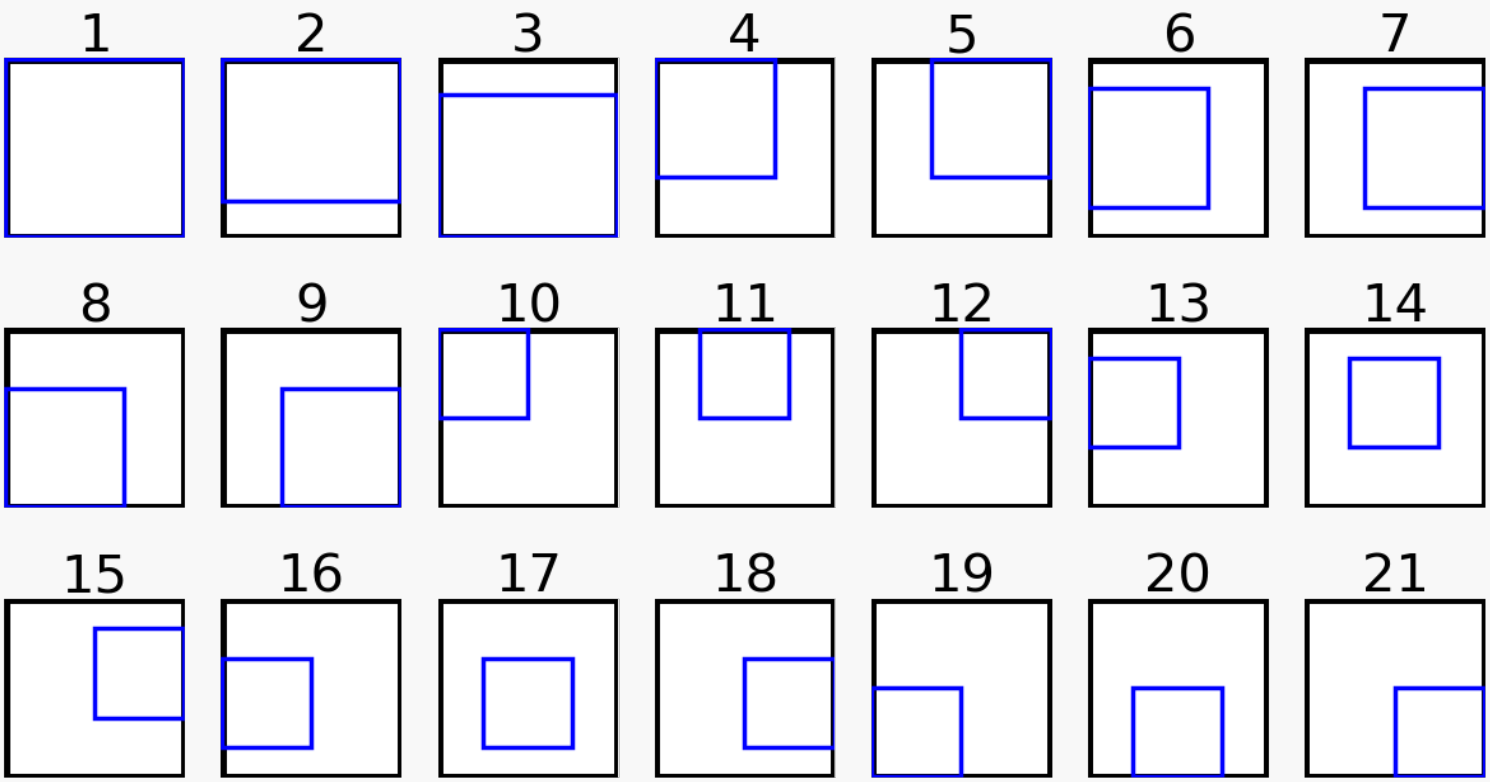
Average vs. Max Pooling

- ResNet50 pretrained on ImageNet
- 6 differently pooled FV´s
- Larger image size improves retrieval (in most cases)



21 Regions

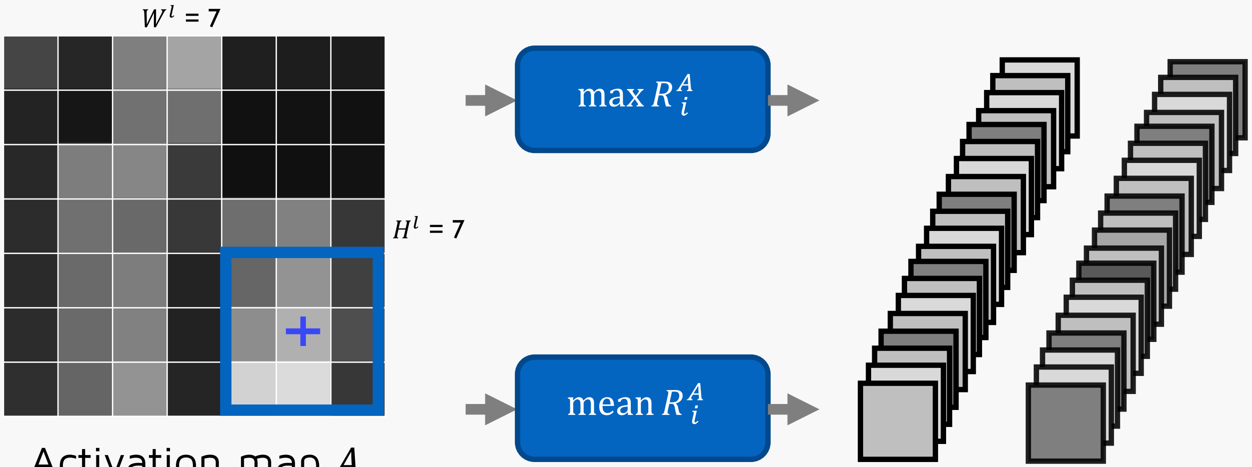
- Make use of average and max pooling
- In total we use 21 regions (global is added):



Regional Max and Average Pooling

RAMAC: Regional Average and Maximal Activations of Convolutions

1. Extraction of activation maps $\mathcal{X}^l$
2. Search for max. and avg. activation values in a set of 21 regions $R^A = \{R_i^A | i = 1 \ldots 21\}$, for each $A \in \mathcal{X}^l$

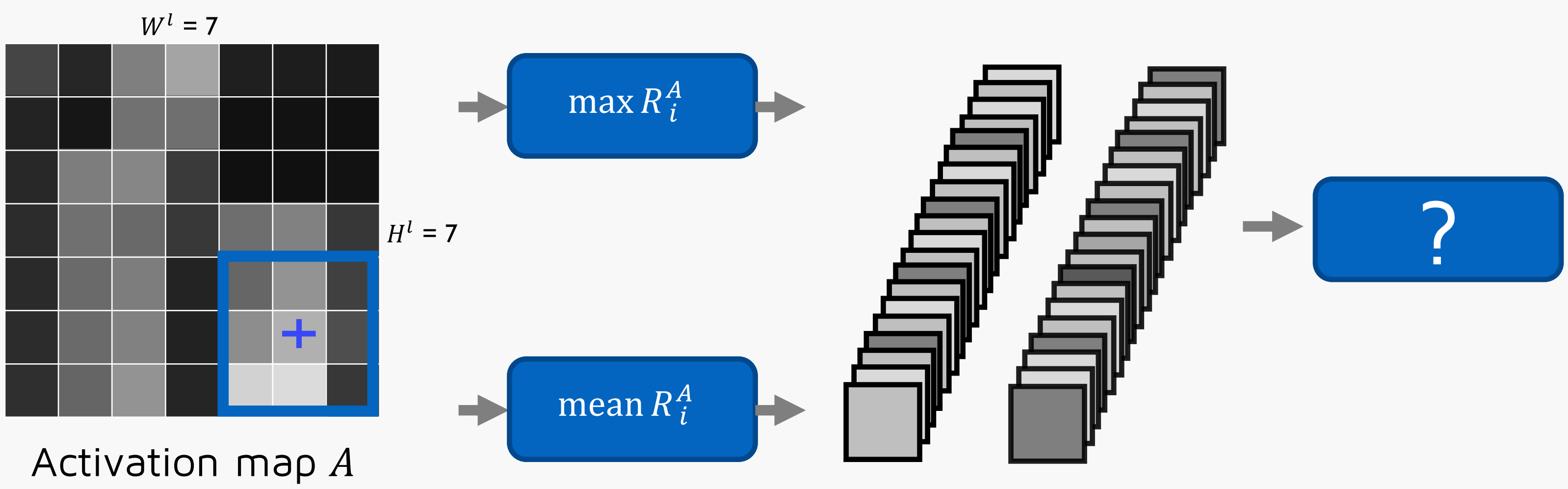


Activation map A

Regional Max and Average Pooling

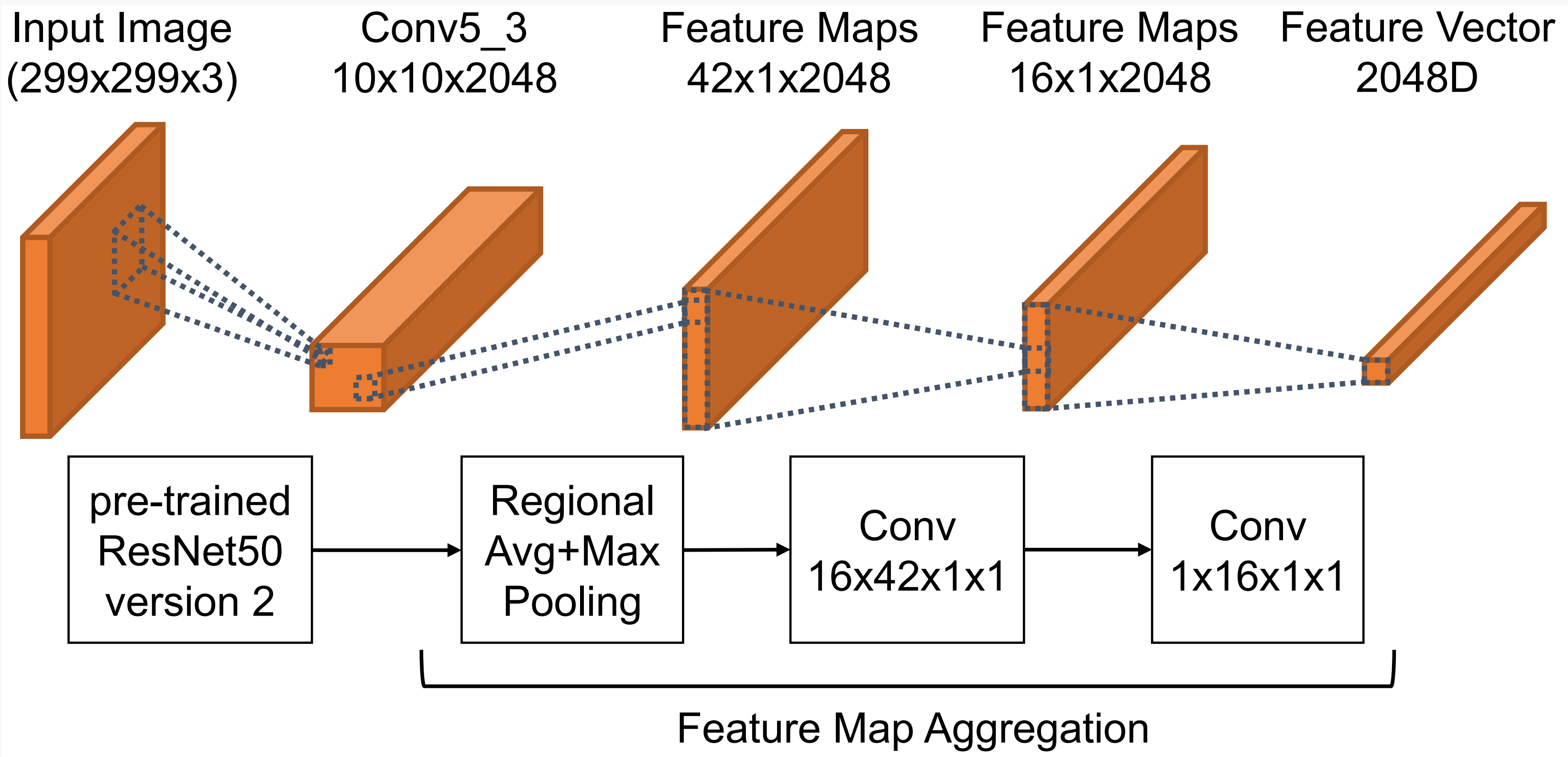
RAMAC: Regional Average and Maximal Activations of Convolutions

1. Extraction of activation maps $\mathcal{X}^l$
2. Search for max. and avg. activation values in a set of 21 regions $R^A = \{R_i^A | i = 1 \dots 21\}$, for each $A \in \mathcal{X}^l$



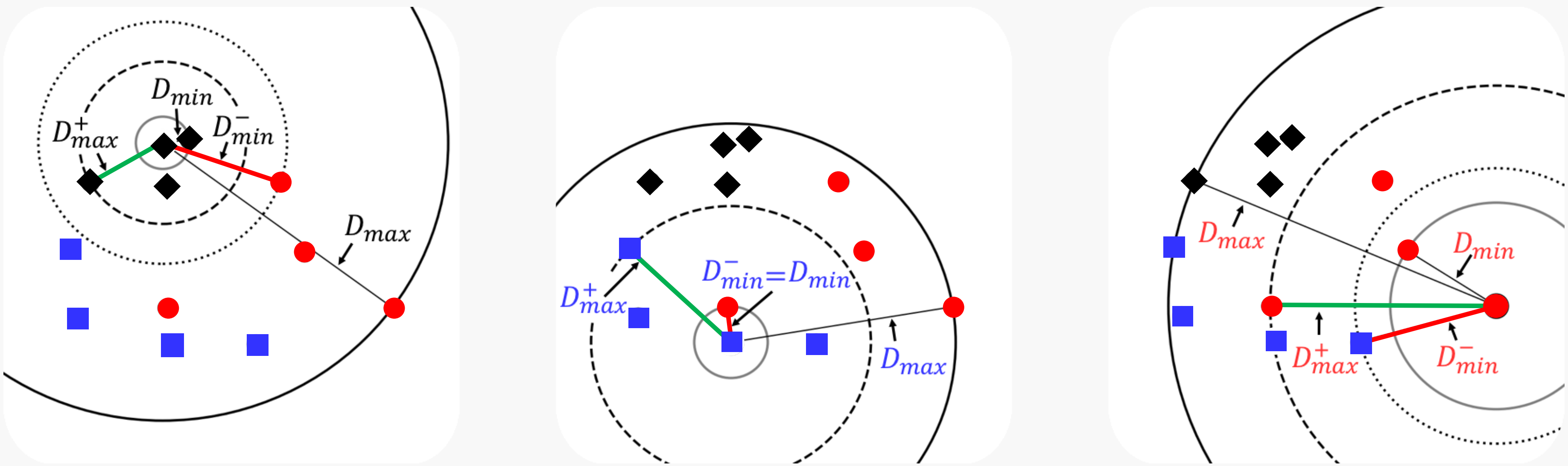
Aggregation of Pooled Values

- DARAC: Deep Aggregation of Regional Activations of Convolutions



NRA Loss Function

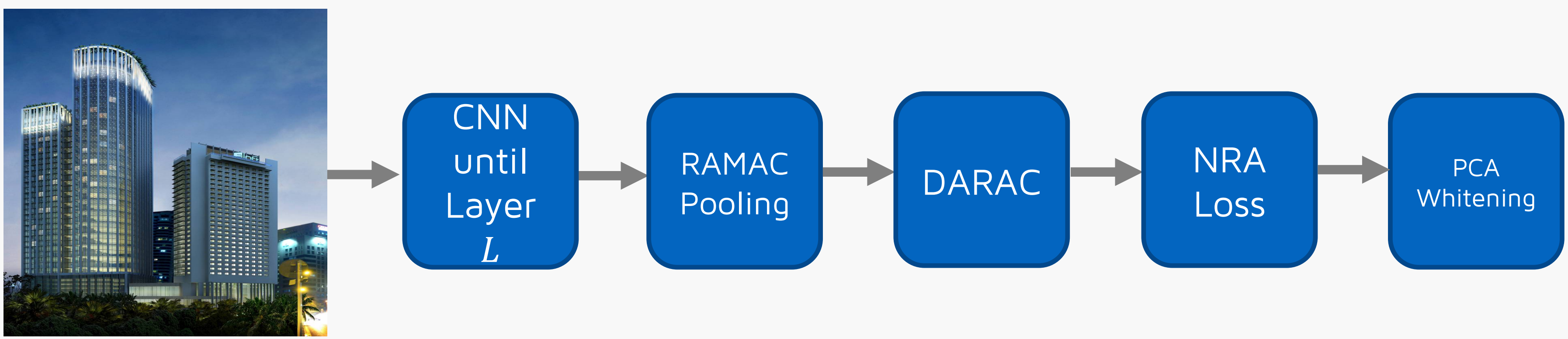
$$J = -\frac{1}{m}\sum_{i=1}^{m}\Big(\log(\underline{s^{+}_{i,\mathrm{max}}}) + \log(1 - \underline{s^{-}_{i,\mathrm{min}}})\Big)$$



Approximates nonlinear ranks to contract similar and disperse dissimilar images

Proposed System

- ResNet50 v2 pretrained on ImageNet
- Training with Google Landmarks dataset
- DARAC aggregation and train with NRA Loss
- Total training time: approximately 24 hours



Input image with 299 pixels at both sides

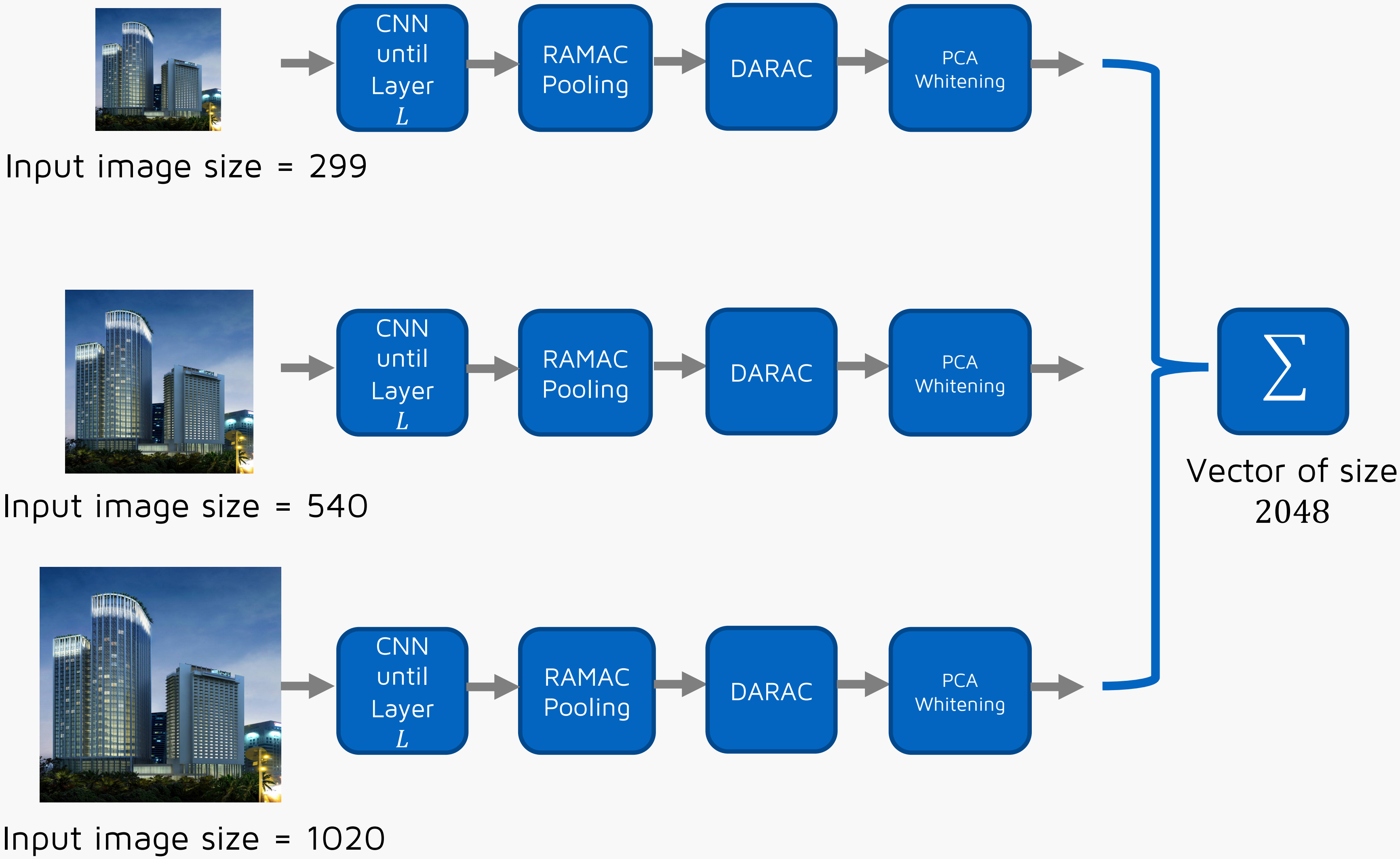
$\mathcal{X}^L$ of size $(10\times10\times2048)$

Tensor of size (42×2048)

Feature Vector of size 2048

After training

Multi Resolution



Evaluation

Mean Average Precision scores for the proposed steps

GL = Google Landmarks. CL = Cleaned Landmarks. W = PCA Whitening

FV	*Training Set*	*Image Size*	*Oxford*	*Paris*	*Holidays*
Glob. Avg.	ImageNet	299	48.7	68.9	86.1
Glob. Avg.	GL	299	75.3	87.6	91.6
DARAC	GL	299	77.6	89.4	92.8
DARAC + W	GL	299	81.4	90.8	93.7
DARAC + W	GL	540	82.2	91.9	95.5
DARAC + W	GL	MR	83.4	93.0	**96.9**
DARAC + W	CL	MR	**88.2**	**94.1**	95.5

Comparison to Similar systems

Mean Average Precision scores for the proposed steps

GL = Google Landmarks. CL = Cleaned Landmarks.

System	*Matching*	*MR*	*Oxford*	*Paris*	*Holidays*
DELF (GL)	local	7 (0.25 – 2)	83.8	85.0	-
GeM	global	5 (0.25 – 1)	87.8	92.7	93.9
DIR (CL)	global	550, 800, 1050	86.1	**94.5**	94.8
Ours (GL)	global	299, 540, 1020	83.4	93.0	**96.9**
Ours (CL)	global	299, 540, 1020	**88.2**	94.1	95.5

Comparison to Similar systems

Mean Average Precision scores for the proposed steps.

GL = Google Landmarks. CL = Cleaned Landmarks.

System	*Matching*	*MR*	*Oxford*	*Paris*	*Holidays*
DELF (GL)	local	7 (0.25 – 2)	83.8	85.0	-
GeM	global	5 (0.25 – 1)	87.8	92.7	93.9
DIR (CL)	global	550, 800, 1050	86.1	**94.5**	94.8
Ours (GL)	global	299, 540, 1020	83.4	93.0	**96.9**
Ours (CL)	global	299, 540, 1020	**88.2**	94.1	95.5

- Fastest inference and training
- Lowest memory footprint
- Very good retrieval quality

Thank you very much!

More Information at

www.visual-computing.com