NetVLAD: CNN architecture for weakly supervised place recognition

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, Josef Sivic [CVPR 2016]

Presentation by: Kent Sommer

(소머켄트)

KAIST

Outline

1. **Review / related work**
2. **Overview of approach**
3. **Issues with approach**
4. **Results**
5. **Conclusions and Quiz!**

Visual Place Recognition

- **Has gained lots of attention recently**
  - **Computer Vision and Robotics Communities**
  - **Useful for:**
    - **Localization for many autonomous robotic tasks**
    - **Localizing old images (no geo-tags available)**
- **Usually viewed as an instance retrieval task**
  - **Some query image location is estimated by matching the most similar images in a database with images of known location**

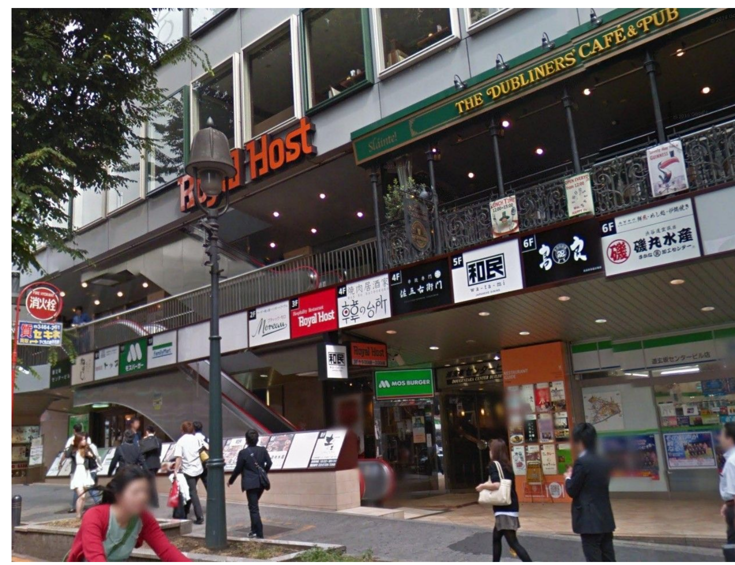
Visual Place Recognition

- **Challenges:**
  - **Appearance changes**
    - **Seasonal / weather**
    - **Lighting**
    - **Occlusions (construction, cars, trees, etc.)**

Images source: http://www.di.ens.fr/willow/research/netvlad/

Visual Place Recognition

- **Challenges:**
  - **Viewpoint changes**
    - **Images can be taken from anywhere**



Images source: http://www.di.ens.fr/willow/research/netvlad/

Visual Place Recognition

- **Challenges:**
  - **“Big” data**
    - **Database of images can become unwieldy extremely quickly, how can we scale to world-wide localization?**

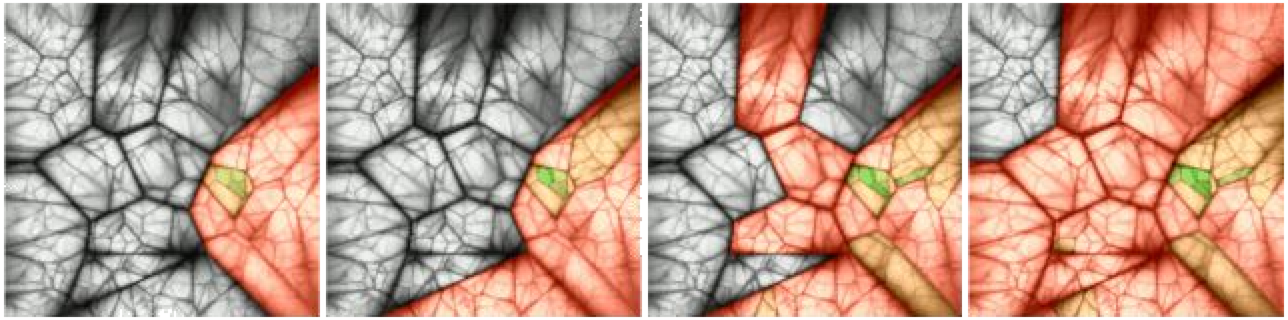
Images source: https://goo.gl/OQRtq1

Visual Place Recognition

- **Related work:**
  - **Two main categories:**
    - **Non-learning based**
      - **Local features (SIFT, ORB, SURF, etc.)**
    - **Learning based (again two main categories)**
      - **Learning for auxiliary task**
        - **Ex: distinctiveness of local features**
      - **Learning on top of hand-engineered descriptors (cannot be tuned for target task)**

Visual Place Recognition
Related Work

- **City-Scale Location Recognition**
  - **Partnership between Georgia Tech and Microsoft Research**
  - **With careful selection of vocabulary and use of a vocab tree -> can increase database size by 10X**

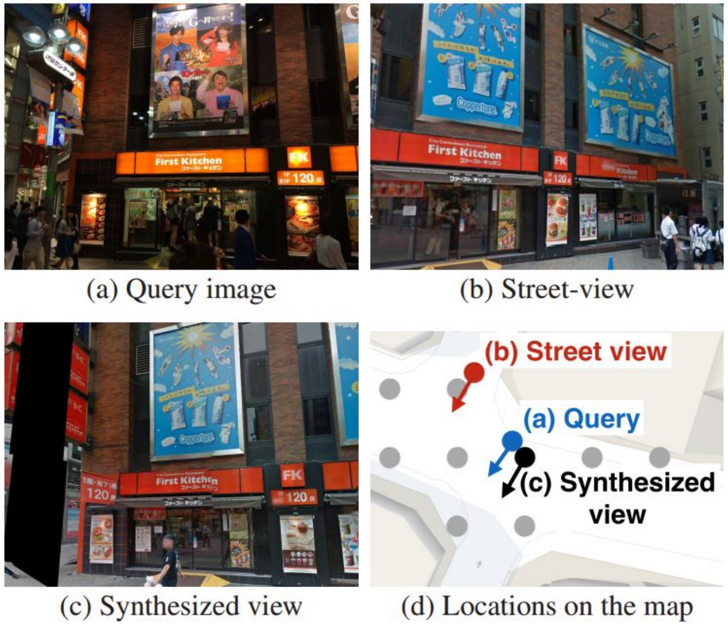


The tree search algorithm considers the N best nodes at each level (left to right N = 1, 2, 5, 9). Cells are coloured from red to green according to the depth at which they are searched, while gray cells are never searched.

Images source: http://matthewalunbrown.com/location/location.html

Visual Place Recognition Related Work

- **24/7 place recognition by view synthesis**
  - **Utilizes view synthesis to render virtual views directly from Google street-view panoramas and associated depth maps**
  - **Based on intuition that matching with large appearance changes is easier when view is the same**



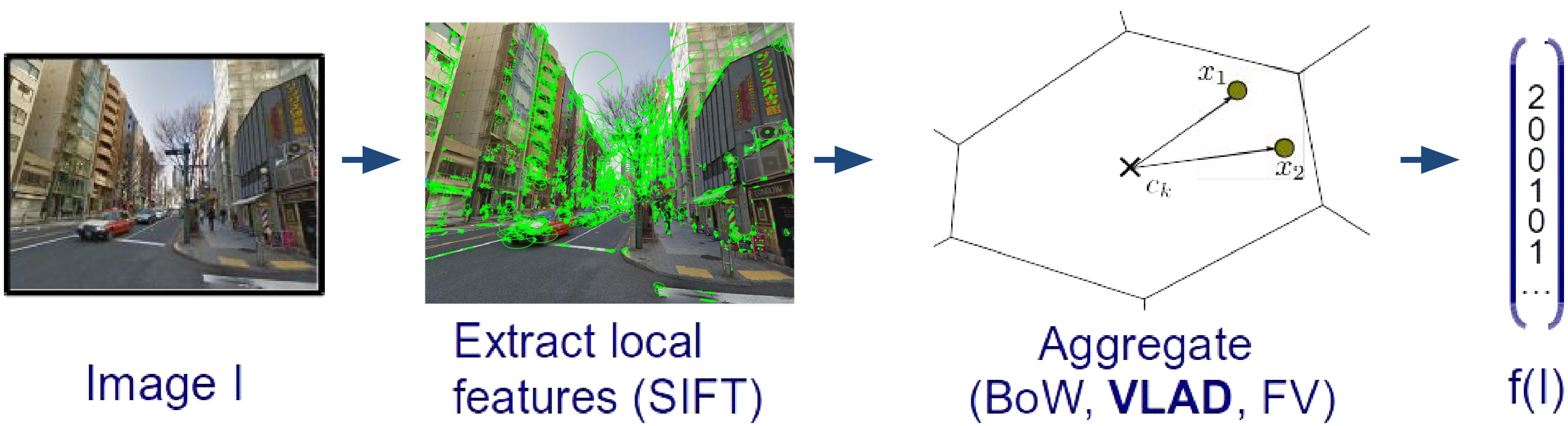
(a) Query image (b) Street-view

(c) Synthesized view (d) Locations on the map

Images source: http://www.ok.ctrl.titech.ac.jp/~torii/project/247/

Visual Place Recognition

- **Issues with local features**
  - **Main goal is matching local image patches**
  - **Not built with image retrieval in mind (not optimized for target goal)**
- **Issues with CNN features**
  - **CNN features are treated as black box image descriptor extractors**



Images source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD

Can an end-to-end CNN help?

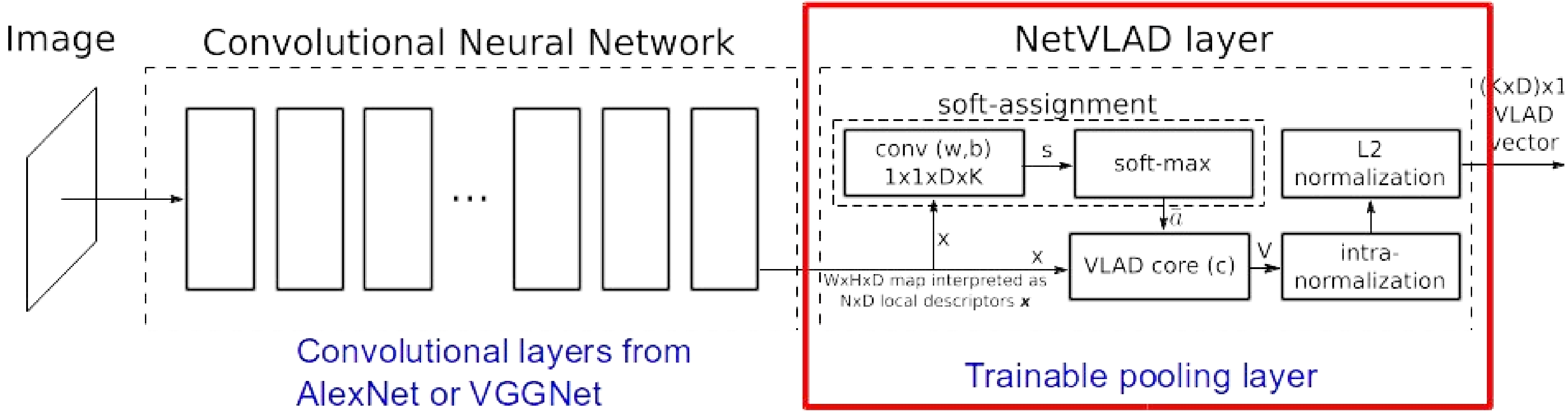
NetVLAD

- **Challenges for approach**
  - **What does a good end-to-end CNN architecture for place recognition even look like?**
  - **How can a sufficient amount of training data be gathered for this task?**
  - **What is an appropriate loss function for end-to-end training?**

KAIST

NetVLAD

- **What does a good end-to-end CNN architecture for place recognition even look like?**
  - **New trainable generalized NetVLAD layer based on the Vector of Locally Aggregated Descriptors!**
    - **Aggregated representation is eventually compressed using PCA to get final descriptor**

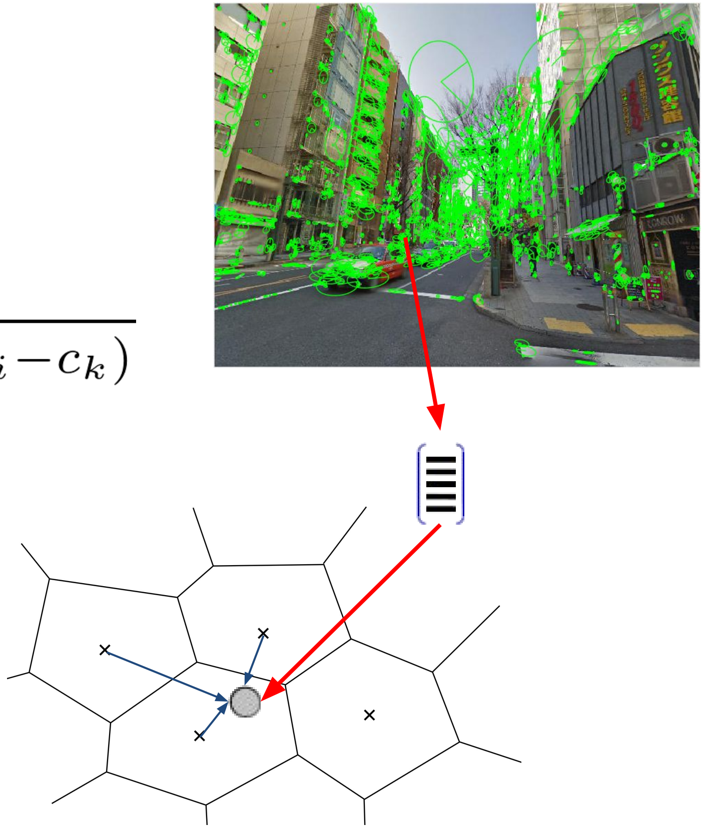


 Images source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD

- **What does a good end-to-end CNN architecture for place recognition even look like?**

$$V(:,k) = \sum_{i=1}^{N} \overbrace{\frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_{k'}^T x_i + b_{k'}}}}^{\text{soft assignment of desc. } i \text{ to cluster } k} (x_i - c_k)$$



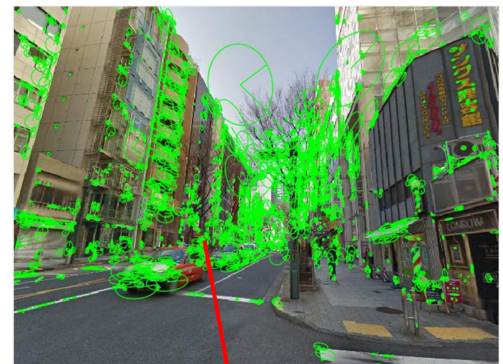
Images source: http://www.di.ens.fr/willow/research/netvlad/

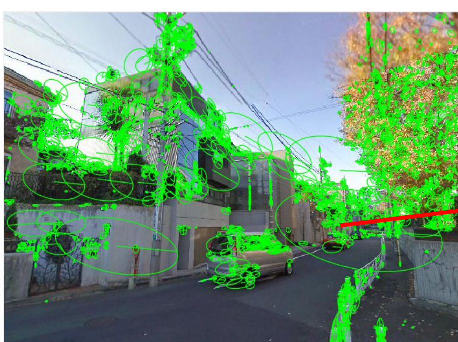
NetVLAD

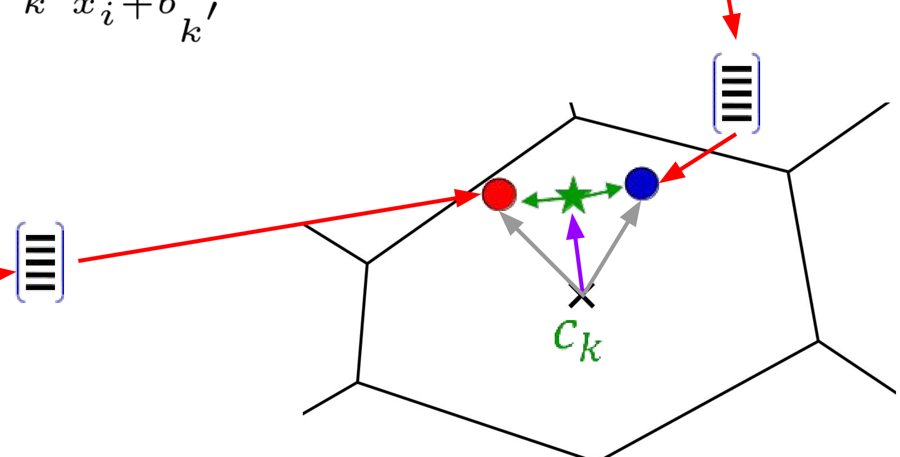
- **What does a good end-to-end CNN architecture for place recognition even look like?**

Decouple assignment (w_k b_k) from anchor point c_k

$$V(:,k) = \sum_{i=1}^{N} \frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_{k'}^T x_i + b_{k'}}} (x_i - c_k)$$



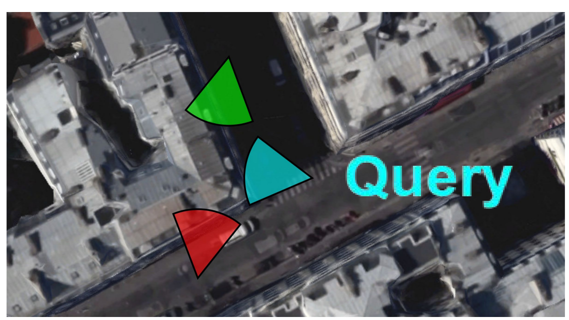


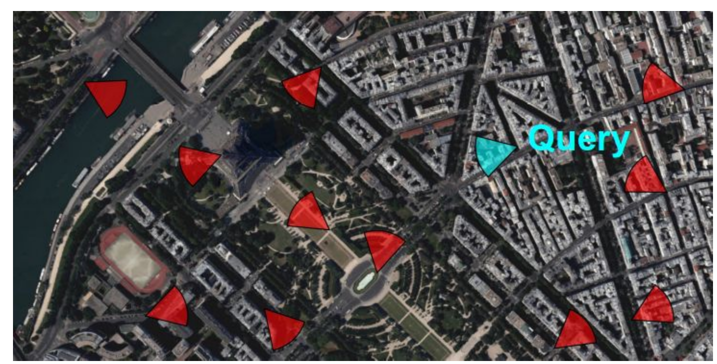


Images and eq's source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD

- How can a sufficient amount of training data be gathered for this task?
  - Collect images of the same place at different viewpoints over time using Google Street View Time Machine
    - Data is available but only weak supervision
      - GPS can only give definite negatives not definite positives!





Images source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD

- **What is an appropriate ranking loss function for end-to-end training?**
  - **Inspired by triplet loss as in [1]**
  - **Can be optimized with Stochastic Gradient Descent**

For a tuple $(q, \{p_i^q\}, \{n_j^q\})$:

$$L_\theta = \underbrace{\sum_j}_{\text{Sum over negatives}} \overbrace{l}^{\text{hinge loss}}\Big(\underbrace{\min_i d_\theta^2(q, p_i^q)}_{\text{Distance to the best potential positive}} + \overbrace{m}^{\text{margin}} - \underbrace{d_\theta^2(q, n_j^q)}_{\text{Distance to the negative}}\Big)$$

Equations source: http://www.di.ens.fr/willow/research/netvlad/

[1]: J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In CVPR, pages 1386–1393, 2014

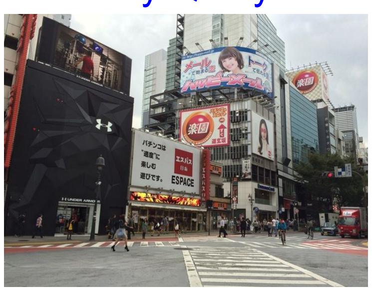
NetVLAD

- **Weaknesses with overall approach**
  - **Only weakly supervised, so better results would be expected with stronger supervision (manual labor tradeoff)**
    - **Stronger supervision could be provided through definite positives**
  - **Uses triplet inspired ranking loss**
    - **Training is long (all triplets used)**
    - **Training is not fully representative (subset of dataset)**

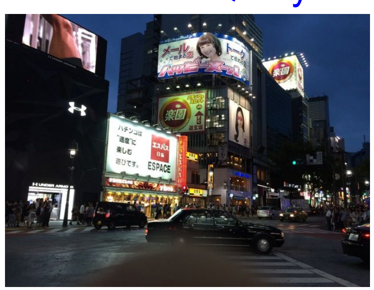
NetVLAD Results

- **Datasets tested against**
  - **Pittsburg [torii et al. 13]**
    - **Database: 250k images from Street View**
    - **Queries: 24k images from Street View at other times**
  - **Tokyo 24/7 [Torii et al. 15]**
    - **Database: 76k images from Street View**
    - **Queries: 215 images from mobile phones**

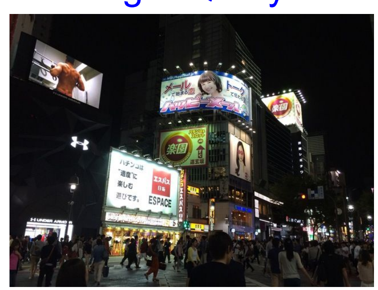
Day Query



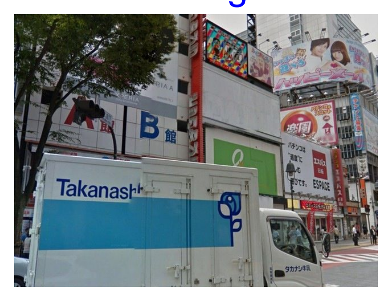
Sunset Query



Night Query



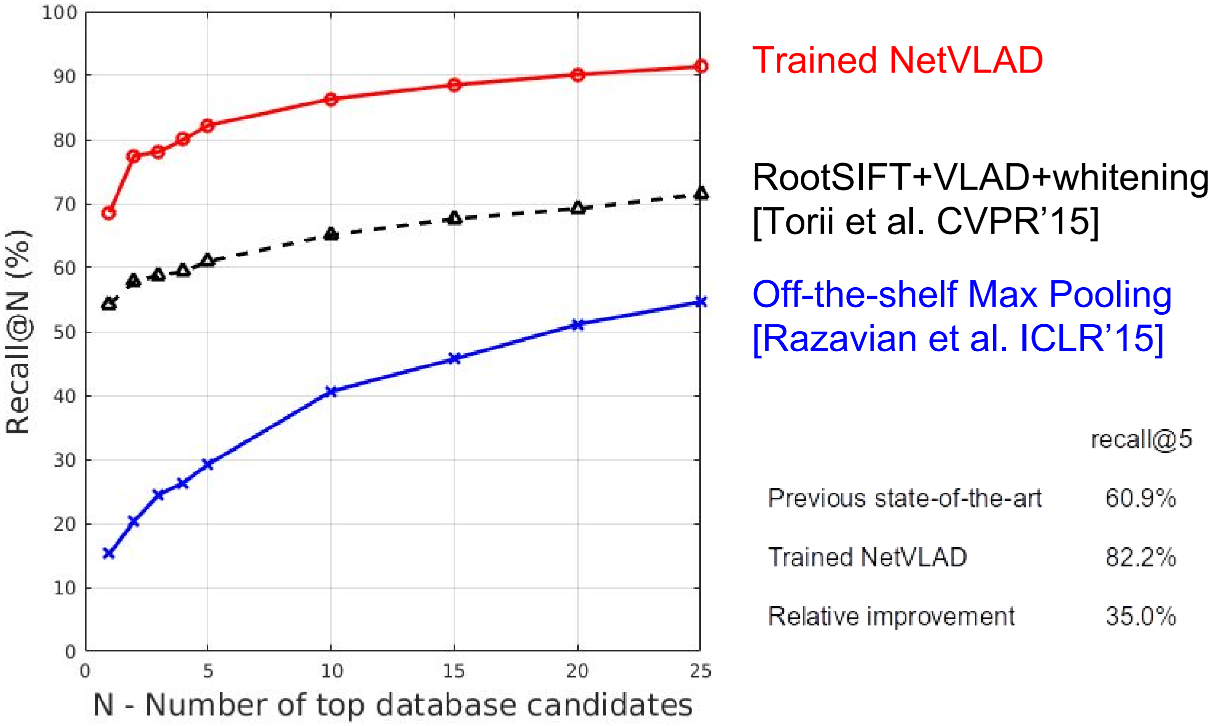
DB Image



Images source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD Results

- **State of the art result on all datasets**

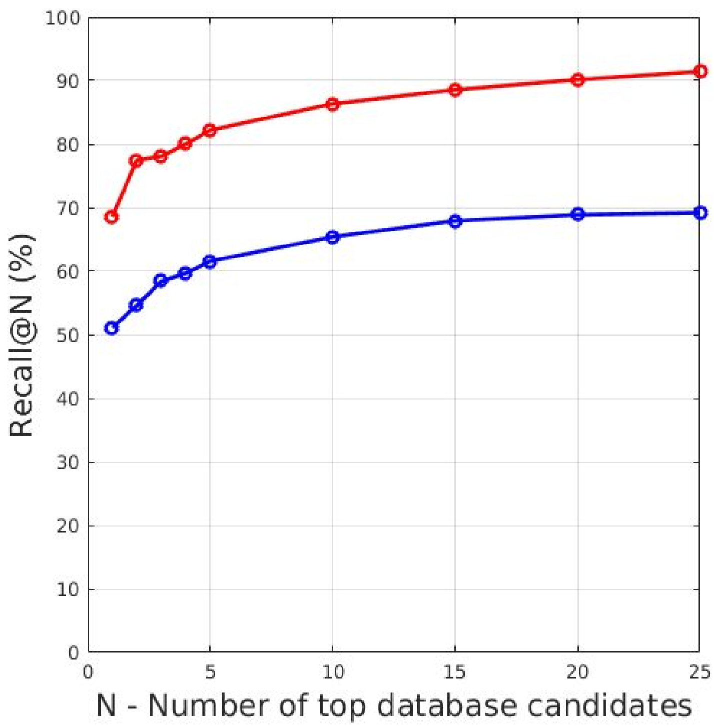


	recall@5
Previous state-of-the-art	60.9%
Trained NetVLAD	82.2%
Relative improvement	35.0%

Graph source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD Results

- **End-to-end training is crucial!**

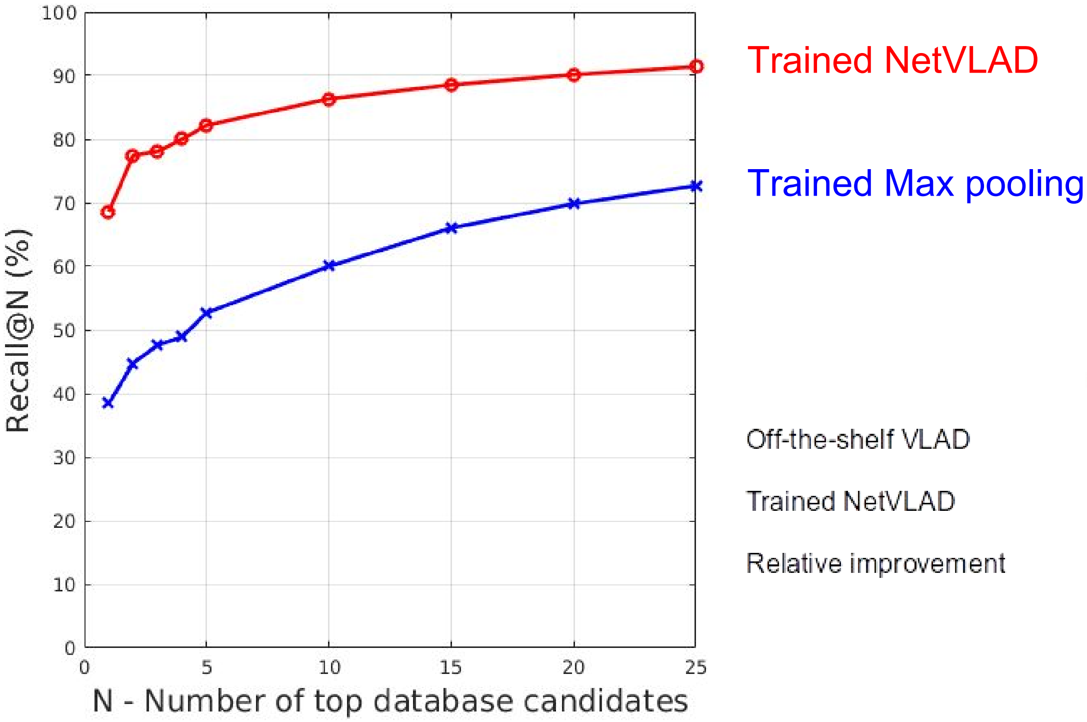


	recall@5
Off-the-shelf VLAD	61.6%
Trained NetVLAD	82.2%
Relative improvement	33.4%

Graph source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD Results

- **NetVLAD is significantly better than Max pooling**

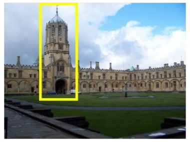


	recall@5
Off-the-shelf VLAD	52.7%
Trained NetVLAD	82.2%
Relative improvement	56.0%

Graph source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD Results

- **Tested on related task: image/object retrieval**
  - **Sets new state-of-the-art for compact image representations (256-D) on all 3 datasets**

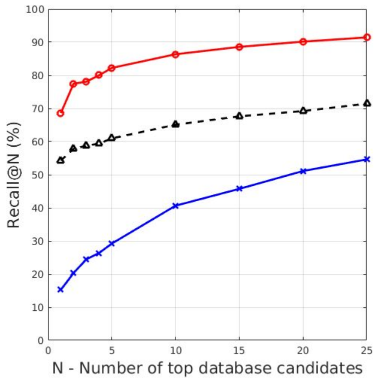


Method	Oxford 5k (full)	Oxford 5k (crop)	Paris 6k (full)	Paris 6k (crop)	Holidays (original)	Holidays (rotated)
***Jégou and Zisserman* CVPR14**		47.2			65.7	65.7
***Gordo et al.* CVPR12**					78.3	
***Razavian et al.* ICLR15**	53.3		67.0		74.2	
***Babenko and Lempitsky* ICCV15**	58.9	53.1				80.2
NetVLAD off-the-shelf	53.4	55.5	64.3	67.7	**82.1**	**86.0**
NetVLAD trained	**62.5**	**63.5**	**72.0**	**73.5**	79.9	84.3

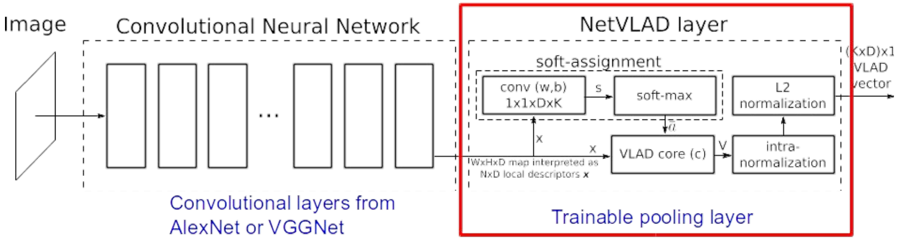
Table and images source: http://www.di.ens.fr/willow/research/netvlad/

NetVLAD

- **Conclusions / Summary**
  - **State-of-the-art on place recognition and image retrieval benchmarks**



  - **Trainable NetVLAD pooling layer**



  - **Street View Time Machine**

  - **Weakly supervised ranking loss**

$$L_\theta = \sum_j l(\min_i d_\theta^2(q, p_i^q) + m - d_\theta^2(q, n_j^q))$$

 Images and eq's source: http://www.di.ens.fr/willow/research/netvlad/

QUIZ!

1. **Why is NetVLAD considered weakly supervised?**
   a. **GPS only gives definite negatives**
   b. **Uses Soft Assignment**
   c. **GPS only gives definite positives**
   d. **Uses Triplet Loss**

2. **What is being done while learning anchor point (Ck) for definite negatives?**
   a. **Maximise distance between descriptors**
   b. **Minimise angle between descriptors**
   c. **Minimise distance between descriptors**
   d. **Maximise angle between descriptors**

"There are 2 hard problems in computer science: caching, naming, and off-by-1 errors"