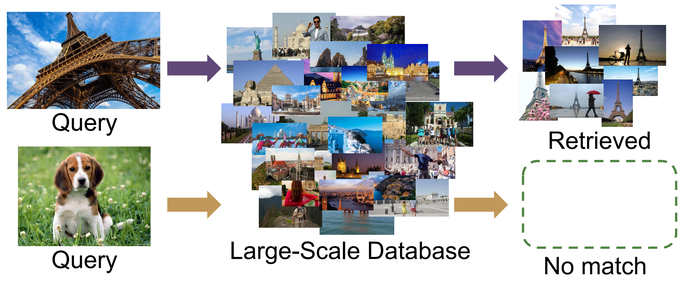
Large-Scale Image Retrieval with Attentive Deep Local Features

Hyeonwoo Noh[1], André Araujo[2], Jack Sim[2], Tobias Weyand[2], Bohyung Han[1]

github.com/tensorflow/models/tree/master/research/delf [1]POSTECH, Korea [2]Google Inc.

ICCV'17 International Conference on Computer Vision 2017

Large-Scale Image Retrieval



Challenges

- Image clutter
- Partial occlusion
- Multiple landmarks
- Queries with no match
- Local features lack semantic information
- Patch-level annotations are expensive
- Existing dataset are small/medium

Contributions

- ➢ Deep local feature retrieval with geometric verification
- ➢ High-level attention based keypoint selection
- ➢ Weakly-supervised feature learning
- ➢ New large-scale dataset

DELF: DEep Local Features

DELF Pipeline

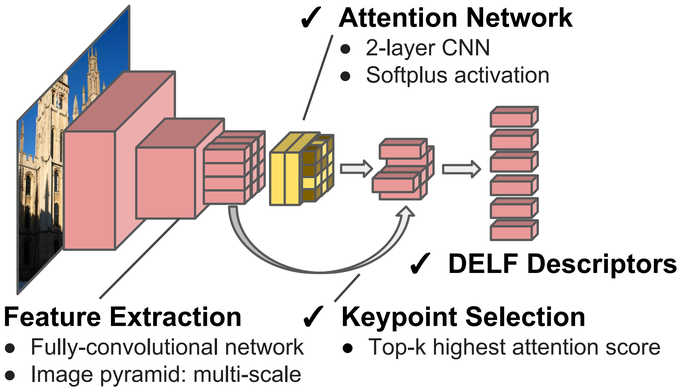
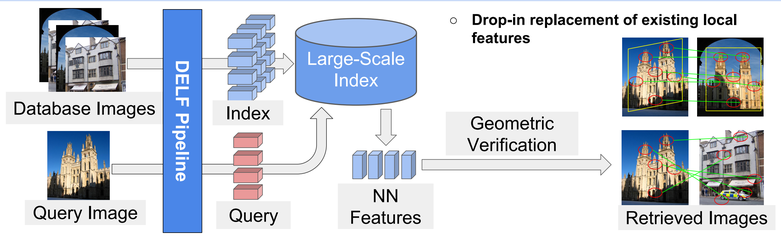


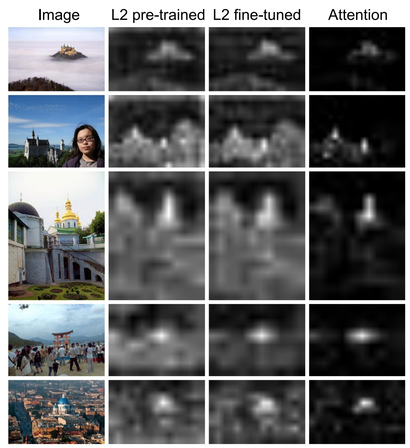
Image Retrieval with DELF



❑ Attention-based Keypoint Selection

- Semantic keypoint selection based on high-level features
- Focus on discriminative features

❑ Attention Visualization



❑ DELF Learning

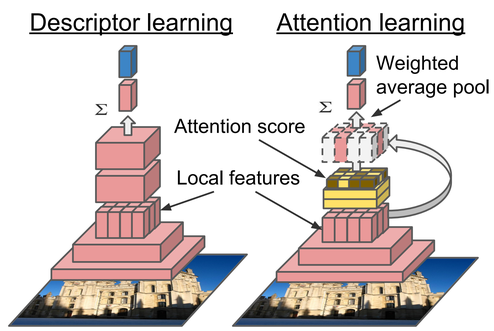
Two-stage training:

(1) Descriptor: Fine-tuning CNN based on classification loss

(2) Attention: Implicit learning by attention-weighted average pooling

- **Weak Supervision**
  - Image-level annotation from landmarks dataset [Babenko et al., ECCV'14]
  - Attention-weighted average pooling

$$\mathbf{y} = \mathbf{W}\left(\sum_n \alpha(\mathbf{f}_n;\theta)\cdot\mathbf{f}_n\right)$$



Google-Landmarks Dataset

❑ Construction

- Mined from GPS-tagged photos from the web [Zheng et al., CVPR'09]

❑ Challenges:

- Query with no correct match
- Large / diverse set of landmarks
- Large variations: clutter, occlusion, partially out-of-view object

❑ Diversity of Landmarks / Images

Image geolocation distribution | Sample query images | Sample database images

❑ Dataset Benchmark

	# landmarks (scenes)	# query images	# database images	query distractors
Google-Landmarks	**12,894**	**100,000**	**1,060,709**	○
Oxford5k [Philbin et al.,CVPR'07]	16	55	5,062	×
Paris6k [Philbin et al.,CVPR'08]	11	55	6,412	×
Holidays [Jegou et al., ECCV'08]	500	500	1,491	×

❖ Dataset to be released with Landmark Recognition Challenge

Experiments

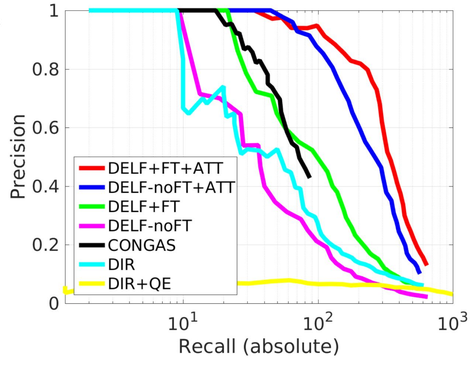
DELF Variants

- **FT:** Landmark fine-tuned features
- **noFT**: ImageNet pre-trained features
- **ATT**: Attention keypoint selection

Benchmarked Techniques

- **CONGAS [Neven et al., 2008]**: 40D hand-engineered local feature similar to SIFT
- **DIR [Gordo et al., ECCV'16]**: 2,048D global descriptor / SOTA on existing datasets
- **siaMAC [Radenovic et al., ECCV'16]**: 512D global descriptor / VGG 16 feature
- **LIFT [Yi et al., ECCV'16]**: 128D feature based descriptor / joint detector and descriptor training

❑ Google-Landmarks



❑ Results on Existing Datasets

Dataset	Oxf5k	Oxf105k	Par6k	Par106k
DIR	86.1	82.8	94.5	90.6
DIR+QE	87.1	85.2	95.3	91.8
siaMAC	77.1	69.5	83.9	76.3
siaMAC+QE	81.7	76.6	86.2	79.8
CONGAS	70.8	61.1	67.1	56.8
LIFT	54.0	–	53.6	–
DIR+QE*	89.0	87.8	93.8	90.5
siaMAC+QE*	82.9	77.9	85.6	78.3
DELF+FT+ATT (ours)	83.8	82.6	85.0	81.7
DELF+FT+ATT+DIR+QE (ours)	**90.0**	**88.5**	**95.7**	**92.8**

- DIR + DELF improves performance significantly
- Complementary information from DELF / DIR

✓ Ablation Study

- Attention helps more than fine-tuning

✓ DELF vs DIR

- **DELF:** handles variations better (e.g. scale variations)
- **DIR:** confuses objects with semantic similarity

✓ DELF vs CONGAS

- **DELF** has higher recall

✓ Feature Correspondence

- **CONGAS** fails on following examples: