Learning Deep Representations for Ground-to-Aerial Geolocalization

Tsung-Yi Lin†, Yin Cui†, Serge Belongie†, James Hays§
† Cornell Tech  § Brown University

Motivation

Image Based Geolocalization
- Most previous methods: match query image to ground-level images with known locations.
- Most of the Earth does not have ground-level reference photos available.
- Fortunately, more complete coverage is provided by aerial imagery.

SIFT + RANSAC fails
- It is challenging to do key point matching from street-view to aerial-view.
- Occlusions and differences in scale, capturing time, image resolution, etc.

Dataset Collection

Ground-to-Aerial Alignment for establishing ground-truth
- Known: street-view car heading direction; GPS location; depth estimates
- Project a 2D street view image to the aerial view.

Learning Deep Feature Embedding

"Siamese" Network
- A pair of input images x, y. Label l = 0 or 1 indicates whether x or y is a match or not.
- A and B are two CNNs. We used some AlexNet architectures for A and B in our paper.
- A and B could be either identical with shared parameters (common feature space will be learned) or distinct (domain specific feature space will be learned).

Dataset Statistics

- 76K aligned street-view and aerial-view pairs.
- Image resolution: 15 x 15 meters (256 x 256 pixels)
- Cardinal viewing direction (azimuth) of 0° 90° 270° 180° for testing.

Detailed Analysis

Effectiveness of training
- Histogram of Pairwise distances on test set.

Experimental Setting

Location verification: given a pair of street-view and aerial-view images, identifying whether this pair comes from same location or not.
- 37.5k (12.5k) positive pairs, together with 20x more generated negative pairs from 4 US cities are used for training (testing). In total 0.79M (0.26M) pairs. Fine-tuned from pre-trained AlexNet on ImageNet.

Precision-Recall curve (mAP)
- Where-CNN is our CNN model; DS means domain specific; Places-CNN and ImageNet-CNN are AlexNet feature from 2nd last fully-connected layer (fc7) trained on Places and ImageNet datasets respectively.

Cross-City Geolocalization

- Fraction of queries with true match in top 1% nearest neighbors.
- x-axis: fraction of nearest neighbors considered; y-axis: fraction of queries with true match in the nearest neighbors considered.

Visualization of Units’ Receptive Fields

Strongest Activations of Particular Units
- Illustration of the average images and the top 9 images that activate a certain unit most strongly at the output feature layer (fc7 layer).

Geolocalization Examples

- Examples of query images, the top 12 matched aerial images for that query, and the heat map that indicates possible locations.
- The first 2 rows are success cases; and last 2 rows are failure cases.