Rotation Invariant Aerial Image Retrieval with Group Convolutional Metric Learning

Hyunseung Chung[1], Woo-Jeoung Nam[2], Seong-Whan Lee[1,2]
[1]Department of Artificial Intelligence, Korea University
[2]Department of Computer and Radio Communications Engineering, Korea University
E-mail: {hs_chung, nwj0612, sw.lee}@korea.ac.kr

Introduction

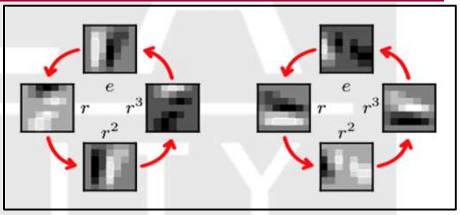
◆ Goal
- Developing a framework to retrieve aerial images with rotational variations
  - Merging group convolution with attention mechanism and metric learning

◆ Motivation
- Retrieving rotated aerial images is highly complex
  - Contains small objects and buildings with variations
- Robust retrieval framework for rotated aerial images in demand

Examples of aerial images

◆ Challenges
- Viewpoint changes from aircraft with an on-board camera
  - Large variations in rotation, angle, and scale
  - Difficult to extract features from or compare similarities to each other
- Heavy computation cost due to large size and complexity of aerial images

◆ Related Works
- Group equivariant convolutional networks [Cohen et al., 2016]
  - Extract features from rotated filters
- Convolutional block attention module [Woo et al., 2018]
  - Focuses on critical regions given an image
- Deep metric learning using triplet network [Hoffer et al., 2015]
  - Considers distance between three tuples

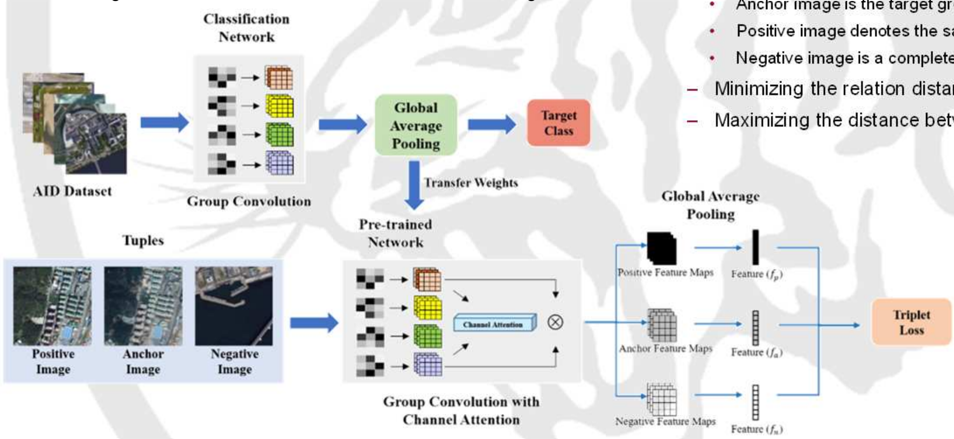


Group equivariant convolution feature maps

Example of data tuples for triplet network

Methods

◆ Group convolutional neural network
- Utilizing rotated filters to pretrain the network for classification task
  - Similar number of parameters compared to CNN
  - Input image is convoluted with different rotated filters
- Fine-tuning network with attentive G-CNN and metric learning



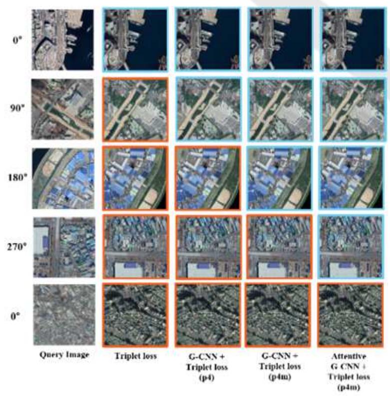
Overall architecture of G-CNN for classification and retrieval tasks

◆ Deep metric learning
- Transforming convoluted features maps into features in embedding space
- Integrating triplet loss function to train data tuples
  - Anchor image is the target ground truth image
  - Positive image denotes the same location image but with time variation
  - Negative image is a completely different region and time image
- Minimizing the relation distance between the anchor and positive tuples
- Maximizing the distance between the anchor and negative tuples

◆ Channel attention module
- Emphasizing the important feature maps among layers
  - Refining feature maps with spatial transformation information after passing G-CNN
- Considering inter-channel relations
- Focusing on critical regions given an input image
- Improving retrieval performance compared to the baseline G-CNN

Experiments

◆ Retrieval results



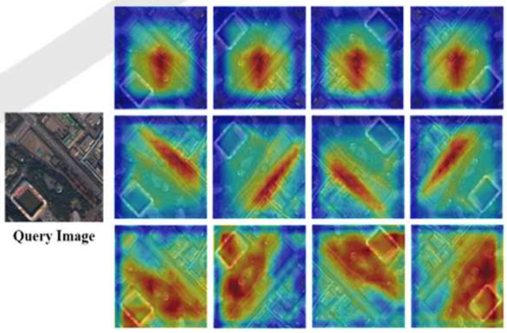
Examples of retrieval results
in rotated Google Earth South Korea dataset

◆ Quantitative results
- Evaluation metric: Recall@n
- Recall@n is the percentage of correctly retrieved queries within top n retrieved database images

Methods	Recall@n(%) $n=1$	$n=5$	$n=10$	$n=100$
R-MAC descriptor †	6.5	14.5	24.8	64.0
NetVLAD †	7.4	17.4	25.1	68.2
Contrastive loss †	8.1	16.8	24.5	65.0
Triplet loss †	6.9	18.0	24.7	65.6
LDCNN †	8.9	18.4	24.7	66.6
G-CNN (p4m) + Cont. loss	17.8	32.4	38.9	72.0
G-CNN (p4m) + Cont. ∗	18.5	32.8	39.8	72.0
G-CNN (p4) + Triplet loss	20.1	36.0	43.2	76.9
G-CNN (p4) + Triplet. ∗	21.2	36.4	43.9	77.2
G-CNN (p4m) + Triplet loss	23.4	44.0	51.7	84.6
G-CNN (p4m) + Triplet. ∗	**24.5**	**46.9**	**52.8**	**86.3**

Retrieval results
(rotated Google Earth South Korea dataset)

◆ Class activation mapping results



CAM results of query image rotated in increments
Top row: CNN, second row: G-CNN, last row: Attentive G-CNN