Landmark Recognition Challenge 2019

Ekaterina Kastrama

{kastrama,}@stanford.edu

Introduction

Image detection, retrieval, object detection are the most fundamental and challenging tasks in modern computer vision.

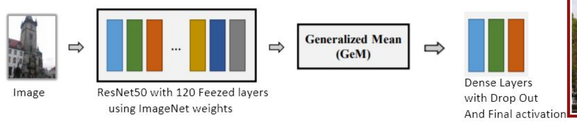
Large scale Image recognition, such as Landmark recognition, will allow, for example, to organize people's photos, render additional sight information on the fly. The problem comes close with image retrieval and weak supervision prediction. Attention mechanism allows to identify the most relevant features which is helpful to solve problem in case of noisy datasets which can contain challenging conditions such as clutter, occlusion, variation of viewpoint and illumination

My model model takes in landmark image as input and outputs one landmark label.

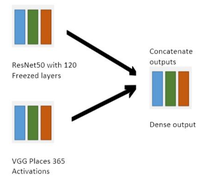
Various variants of network architecture were considered for this project, The most competitive results were from three models: ResNet50+ GeM, VGG concatenated with ResNet and DELF Features model.

Model Architecture

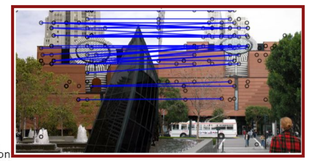
ResNet50+GeM



VGGplaces 365+ ResNet50



DELF +Bottleneck



DELF Features 150x40

Transfer Learning Bottleneck:
(Dense + BN + Activation + Dropout)

output

Loss function:

Binominal Cross entropy

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i)$$

DELF feature extraction:

Algorithm returns up to 1000x40 feature per image, for image of size 128x128 median number of features were 152. Features trained on previous version of dataset.

Hyperparameters

- batch size - 128 for ResNet+GeM, 256 for DELF model
- Adam optimizer
- Adaptive LR - [0.001, 0.00001]
- batch norm
- Tuning: manual observing performance on several epochs

Hardware:

2 Nvidia Tesla K80

Discussion

- The most effective from perspective of performance/accuracy was Resnet50+GeM
- All methods suffered more or less from overfitting problem
- Concatenated model VGG + ResNet 50 will require more time/hardware to train.
- Using DELF features showed best GAP and Accuracy, however it was more computationally expensive as extraction took 5-10 sec per 100 images
- DELF extracts features better for outdoor buildings vs. indoor/nature.

Future Work

Traditional popular CNN algorithms, applied on large scale dataset showed significant improvement when combined with attention mechanism, some

- Adding non Landmark dataset to distinguish non landmarks
- More regularization methods to deal with overfitting
- Training on bigger size images
- Increase number of considered classes.
- Improved Cropping methods based on class imbalance
- Combining deeper Resnet101 with GeM pooling layer

Dataset & Features

The training dataset [4] consists of 4 132 914 training images and was constructed by mining web landmark images with permissive licenses, leveraging crowdsourced labels.

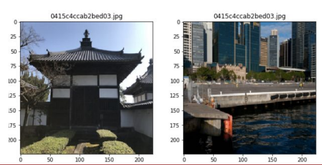
Training dataset highly unbalanced, containing 203 094 classes, some classes have only few examples.

For this project subset of training set was used that corresponds to labels that have at least 250 samples, which resulted in 1067 classes and 478 577 images.

For test and validation subset of 10 241 samples was used.

Original images has different size, but was rescaled for training and optimization purposes to 128x128.

Random corps were used as argumentation technique during training.



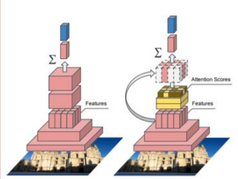
Related Work

Generalized Average pooling layer (GeM)

Generalized Average pooling layer[2] is trainable pooling layer to suppress irrelevant convolution layer features.

It can be applied on the top of any fully convolutional CNN, such as AlexNet, VGG, or ResNet, while their fully-connected layers are discarded.

Pooling layer based on a generalized-mean that has learnable parameters, either one global or one per output dimension



$$f_1^{(g)} \dots f_k^{(g)} \dots f_K^{(g)}]^\top, \quad f_k^{(g)} = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^{p_k} \right)^{\frac{1}{p_k}}.$$

Deep Local Features (DELF)

DELF architecture (Deep Local Features)[1] is CNN with attention. It allows to train with weak supervision using image-level class labels only, without the need of object- and patch-level annotation .

It aims to retrieve semantically meaningful features which can be used for object detection and retrieval tasks. It also showed good results in rejecting false positive examples. DELF architecture consists of several steps: (i) dense localized feature extraction, (ii) keypoint selection, (iii) dimensionality reduction and (iv) indexing and retrieval.

Results

Evaluation metrics: Class Accuracy and Global Average Precision (GAP) at k

Model	GAP	Accuracy
Baseline: 2018 leaderboard top (different dataset)	0.30	
Baseline: 2019 leaderboard top	0.37	
ResNet50 test set	0.012	0.14
TEST ResNet50 +GEM	0.04	0.19
TRAIN ResNet50 +GEM	0.2	0.45
TEST ResNet50 +GEM with crops	0.05	0.25
TRAIN ResNet50 +GEM with crops	0.16	0.36
TRAIN VGGPlaces 365 +ResNet50	0.1249	0.32
TEST DELF+ BottleNeck	**0.1035**	**0.3**

References

1.Large-Scale Image Retrieval with Attentive Deep Local Features https://arxiv.org/pdf/1612.06321.pdf

2. Generalized-mean pooling https://arxiv.org/pdf/1711.02512.pdf

3. VGGPlaces365 https://github.com/CSAILVision/places365

4. Kaggle competition page https://www.kaggle.com/c/landmark-recognition-2019