Report

Matteo Gambino
s287572

Michele Pierro
s287846

Fabio Grillo
s287873

Abstract

In order to predict the location of a query image by retrieving annotated photographs with similar descriptors needs an efficient and reliable generation of those descriptors. In order to accomplish that objective, is fundamental that the network focuses on portion of the various images that contains useful information and at the same time ignore not informative areas like the ones containing elements like cars or pedestrians. For that reason attention layers are fundamental in the proposed network. In addition to that we are comparing state of the art techniques for the visual geolocalization task like GeM [1] and NetVLAD [2].

1. Introduction

2. Related works

3. Methods

Like [1], [2], [3] we have casted the problem of place recognition as the task of image retrieval. We have implemented 3 different networks all based on the ResNet-18 [4] backbone without the fully connected layers and the last convolution layer. On the top of this backbone we have inserted 4 different heads, inspired by the works of [1], [2], [3], in order to generate the image descriptors.

Base Head This is the simplest head we have used and it's necessary in order to have a baseline to compare the other results. After the last convolution layer of ResNet-18 we have normalized the feature map and used average pooling in order to generate the descriptors. This simple head tries to extract from the query the spatial information by comparing the average value of the features in a given area and represent the traditional way to extract those descriptors.

GeM head Following the work of [1], we have used a Generalized Mean approach in order to extract better descriptors for the query image. The generalized mean we are using is defined as:

$$f_k = (\frac{1}{X_k} \sum_{x \in X_k} x^{p_k})^{\frac{1}{p_k}} \qquad (1)$$

where X_k represent one of the normalized features map and p_k is the pooling parameter. This pooling parameter is expressing how much is localized the zone of the image the network is focusing on. The p_k parameter, although it can be learned and inserted into back propagation, it has been fixed and a single value is used for each activation map as suggested by [1]. We have inserted a fully connected layer that takes as input the pooled features in order to whiten the image descriptors since it has been shown by [1] that this approach is providing better results than using other strategies like PCA.

NetVLAD head

CRN head

4. Experiments

	R@1	R@2	R@10	R@20
lr = 1e-3	81.9	91.8	94.8	96.8
lr = 1e-4	82.2	93.0	**95.4**	**97.1**
lr = 1e-5	**83.5**	**93.1**	94.3	**97.1**

5. Ablation study

6. Conclusions

References

[1] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *TPAMI*, 2018. 1

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *TPAMI*, 2018. 1

[3] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," *CVPR*, 2017. 1

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016. 1