

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], NetVLAD [2] and CRN [3]. The code used is publicly available [here](#)

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.

3.1. 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.

3.2. 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:

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

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.

3.3. NetVLAD head

3.4. CRN head

Seen the results provided from the previous implemented heads and the success of the attention layers to make a network focus on relevant only parts of an image, we have decided to add an attention layer at the NetVLAD head following the approach proposed by [3]. This is perfectly integrated in the NetVLAD architecture and it has the duty to produce a map that rescales the weights produced by the soft assignment step of the NetVLAD layer. This layer is composed by an initial average pooling sub-layer that has the duty to reduce the dimensionality of the feature maps produced by the backbone. Differently from what specified in [3], it is not reducing the features maps to a fixed size but it is simply reducing by a half the dimensions of the features. In order to capture features at different spatial resolutions, 3 convolution filters (with kernel sizes respectively of 3, 5, 7) are applied to the pooled features. The output of those filter is concatenated and an additional 1×1 convolution filter is used in order to accumulate the features produced. The resulting mask is then upsampled, in order to restore the original features map dimensionality, by using a bilinear interpolator. This results into a mask that is used as to re-weight the features produced by the soft assignment specified into the NetVLAD description as showed in figure

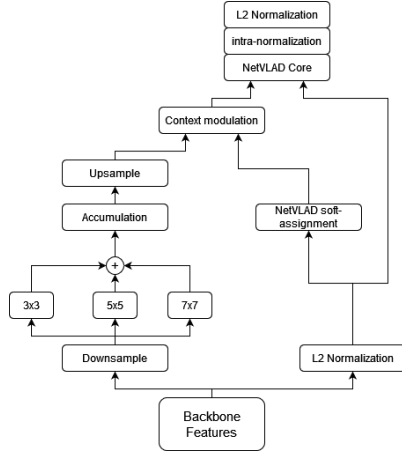


Figure 1. The architecture of the CRN head.

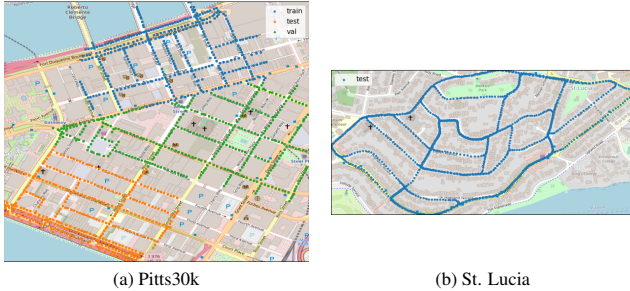


Figure 2. The location of the images contained in both the pitts30k 2a dataset and the St. Lucia 2b dataset.

1 into the context modulation layer. This layer is performing the product of the mask and the soft-assignment scores. The output of the context modulation layer is then used in the standard NetVLAD core instead of the soft-assignment scores.

4. Experiments

In this section we evaluate the results obtained by the implemented heads and the selection of best hyper-parameters will be discussed.

Datasets The experiments have been run on the pitts30k dataset [2] that is containing 3 different predefined splits, respectively for training, validation and testing, composed by 10k images each. Those are images taken in the city of Pittsburgh from the Google street view images. In addition, also the St. Lucia dataset [5] has been used only for testing the models trained on the pitts30k dataset. The location of the images contained in both datasets can be seen in figure 2.

	Descs.	$R@1$	$R@5$	$R@10$	$R@20$
Base	256	60.1	80.6	87.4	91.7
GeM	256	71.6	87.0	91.0	94.0
NetVLAD	16384	79.1	89.3	92.3	94.4
CRN*	16384	81.7	90.7	93.4	95.3

Table 1. Results on the pitts30k test set obtained with the various heads compared with the base head. The number of generated descriptors is also shown in the column Descs.

	$R@1$	$R@5$	$R@10$	$R@20$
lr = 1e-3	81.9	91.8	92.5	94.7
lr = 1e-4	82.2	93.0	95.4	97.1
lr = 1e-5	83.5	93.1	94.3	97.1

Table 2. Results obtained with the NetVLAD head on the pitts30k validation set with different learning rates

Metric adopted Those results have been obtained by evaluating the models by using a standard evaluation procedure for place recognition. A given query image is said to be correctly localized if at least one of the N retrieved images is placed at a distance lower or equal to TTD from the ground truth position. This distance is set, if not differently specified, to 25 meters. After that we are calculating the percentage of correctly classified images for different values of N (indicated with $R@N$).

4.1. Comparison among the proposed heads

The results of comparison between the various proposed heads are reported in table 1. Those results have been obtained on the pitts30k test set. As it’s possible to notice the head that is giving best results is the CRN and that clearly shows that adding attention is essential for improving the quality of generated descriptors. It’s important to notice also how the results are influenced by the number of produced descriptors. In fact both the NetVLAD and the CRN head are generating much more descriptors with respect to the GeM and the base heads and this seems correlated to higher recalls. Since the CRN and the NetVLAD heads are outperforming the other ones, we will focus more on those 2 during the rest of this section.

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, 2

- [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](#)
- [5] M. Warren, D. McKinnon, H. He, and B. Upcroft, "Unaided stereo vision based pose estimation," in *Australasian Conference on Robotics and Automation*, G. Wyeth and B. Upcroft, Eds. Brisbane: Australian Robotics and Automation Association, 2010. [Online]. Available: <http://eprints.qut.edu.au/39881/> [2](#)