Learning local feature descriptors with triplets and shallow convolutional neural networks

Vassileios Baltas
http://www.iis.ee.ic.ac.uk/~vbalt

Edgar Riba
eriba@cvc.uab.es

Daniel Ponsa
daniel@cvc.uab.es

Krystian Mikolajczyk
k.mikolajczyk@imperial.ac.uk

Imperial College London
London, UK

Computer Vision Center, Computer Science Department
Universitat Autònoma de Barcelona
Bellaterra (Barcelona), Spain

Finding correspondences between images via local descriptors is one of the most extensively studied problems in computer vision due to the wide range of applications. Recently, end-to-end learnt descriptors [1, 2, 3] based on Convolutional Neural Network (CNN) architectures and training on large datasets have demonstrated to significantly outperform state of the art features. These works are focused on exploiting pairs of positive and negative patches to learn discriminative representations.

Recent work on deep learning for learning feature embeddings examines the use of triplets of samples instead of pairs. In this paper we investigate the use of triplets in learning local feature descriptors with CNNs and we propose a novel in-triplet hard negative mining step to achieve a more effective training and better descriptors. Our method reaches state of the art results without the computational overhead typically associated with mining of negatives and with lower complexity of the network architecture. This is a significant advantage over previous CNN-based descriptors since makes our proposal suitable for practical problems involving large datasets.

Learning with triplets involves training from samples of the form \( \{a, p, n\} \), where \( a \) is the anchor, \( p \) is a positive example, which is a different sample of the same class as \( a \), and \( n \) is a negative example, belonging to a different class than \( a \). In our case, \( a \) and \( p \) are different viewpoints of the same physical point, and \( n \) comes from a different keypoint. The goal is to learn the embedding \( f(x) \) s.t. \( \delta_0 = ||f(a) - f(p)||_2 \) is low (i.e., the network brings \( a \) and \( p \) close in the feature space) and \( \delta_n = ||f(a) - f(n)||_2 \) is high (i.e., the network pushes the descriptors of \( a \) and \( n \) far apart). With this aim, we examine two different loss functions for triplet based-learning: the margin ranking loss and the ratio loss. The margin ranking loss is defined as

\[
\lambda(\delta_0, \delta_n) = \max (0, \mu + \delta_0 - \delta_n) ,
\]

where \( \mu \) is an arbitrarily set margin. It measures the violation of the ranking order of the embedded features inside the triplet, which should be \( \delta_0 > \delta_n + \mu \). If that is not the case, then the network adjusts its weights to achieve this result. For its part, the ratio loss optimises the ratio distances within triplets. It learns embeddings such that \( \frac{\delta_0}{\delta_n} \to \infty \) and is defined as

\[
\hat{\lambda}(\delta_0, \delta_n) = \left( \frac{e^{\delta_0}}{e^{\delta_0} + e^{\delta_n}} \right)^2 + \left( 1 - \frac{e^{\delta_0}}{e^{\delta_0} + e^{\delta_n}} \right)^2 .
\]

The goal of this loss function is to force \( -\left( \frac{e^{\delta_0}}{e^{\delta_0} + e^{\delta_n}} \right)^2 \to 0 \), and \( -\left( \frac{e^{\delta_n}}{e^{\delta_0} + e^{\delta_n}} \right)^2 \to 1 \). There is no margin associated with this loss, and by definition we have \( 0 \leq \hat{\lambda} \leq 1 \) for all values of \( \delta_0, \delta_n \).

1 Acknowledgements

This work was supported by EPSRC project EP/N007743/1 and partially supported by the spanish project FireDMMI (TIN2014-56919-C3-2-R).


