Learning local feature descriptors with triplets and shallow convolutional neural networks

Vassileios Balntas[1]
http://www.iis.ee.ic.ac.uk/~vbalnt

Edgar Riba[2]
eriba@cvc.uab.es

Daniel Ponsa[2]
daniel@cvc.uab.es

Krystian Mikolajczyk[1]
k.mikolajczyk@imperial.ac.uk

[1] Imperial College London
London, UK

[2] 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 $\{\boldsymbol{a},\boldsymbol{p},\boldsymbol{n}\}$, where $\boldsymbol{a}$ is the *anchor*, $\boldsymbol{p}$ is a *positive* example, which is a different sample of the same class as $\boldsymbol{a}$, and $\boldsymbol{n}$ is a *negative* example, belonging to a different class than $\boldsymbol{a}$. In our case, $\boldsymbol{a}$ and $\boldsymbol{p}$ are different viewpoints of the same physical point, and $\boldsymbol{n}$ comes from a different keypoint. The goal is to learn the embedding $f(\boldsymbol{x})$ s.t. $\delta_+ = ||f(\boldsymbol{a}) - f(\boldsymbol{p})||_2$ is low (i.e., the network brings $\boldsymbol{a}$ and $\boldsymbol{p}$ close in the feature space) and $\delta_- = ||f(\boldsymbol{a}) - f(\boldsymbol{n})||_2$ is high (i.e., the network pushes the descriptors of $\boldsymbol{a}$ and $\boldsymbol{n}$ 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

$$\lambda(\delta_+,\delta_-) = max(0, \mu + \delta_+ - \delta_-) \ ,$$

where μ is an arbitrarily set margin. It measures the violation of the ranking order of the embedded features inside the triplet, which should be $\delta_- > \delta_+ + \mu$. 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 $\frac{\delta_-}{\delta_+} \to \infty$ and is defined as

$$\hat{\lambda}(\delta_+,\delta_-) = (\frac{e^{\delta_+}}{e^{\delta_+}+e^{\delta_-}})^2 + (1 - \frac{e^{\delta_-}}{e^{\delta_+}+e^{\delta_-}})^2 \ .$$

The goal of this loss function is to force $(\frac{e^{\delta_+}}{e^{\delta_+}+e^{\delta_-}})^2$ to 0, and $(\frac{e^{\delta_-}}{e^{\delta_+}+e^{\delta_-}})^2$ 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 δ_-, δ_+. Fig. 1 illustrates both approaches and their loss surface. In $\lambda(\delta_+,\delta_-)$ the loss remains 0 until the margin is violated, and after that, there is a linear increase not upper bounded. In contrast, $\hat{\lambda}(\delta_+,\delta_-)$ has a clear slope between the two loss levels, and the loss reaches a 1-valued plateau quickly when $\delta_- > \delta_+$.

All previous proposals based on triplet based learning use only two of the possible three distances within each triplet, ignoring the distance $\delta_-' = ||f(\boldsymbol{p}) - f(\boldsymbol{n})||_2$. We take it into account to define the *in-triplet hard negative* as $\delta_* = min(\delta_-, \delta_-')$. If $\delta_* = \delta_-'$, we swap $\{\boldsymbol{a},\boldsymbol{p}\}$, and thus $\boldsymbol{p}$ becomes the *anchor*, and $\boldsymbol{a}$ becomes the *positive* sample. This ensures that the hardest negative inside the triplet is used for backpropagation. Subsequently, the margin ranking loss becomes $\lambda(\delta_+,\delta_*) = max(0, \mu + \delta_+ - \delta_*)$. A similar expression can be devised for the ratio loss.

To test our proposal we build our descriptor by training a shallow network architecture {Conv(7,7)-Tanh-Pool(2,2)-Conv(6,6)-Tanh-FC(128)} from $5M$ triplets sampled on-the-fly using patches from the Photo-tour dataset [4]. We evaluate its performance in patch pair classification, where we measure the ability of the descriptor to discriminate positive patch pairs from negative ones, and in nearest neighbour patch matching, where we measure the descriptor precision in matching feature points between different views of a same scene. Our networks outperform previously introduced single-scale convolutional feature decriptors, and in some cases with large margin. Moreover, they are 10 times faster than DeepCompare [3], and 50 times faster than MatchNet [1] and DeepDesc [2]. In fact, when running on GPU we reach speeds of $10\mu s$ per patch, which is comparable with the CPU speeds of fast binary descriptors. We observe also that ratio-loss based descriptors are more suitable for patch pair classification and that margin-loss based ones work better in nearest neighbour patch matching. Details of our proposal are described more fully in the paper, along with extensive experimental work. We provide all the learned models and the training code for all descriptor variants at `https://github.com/vbalnt/tfeat`.

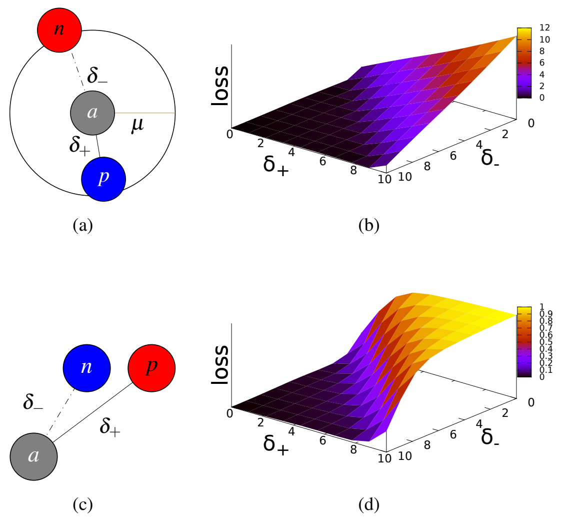

Figure 1: (a) The margin ranking loss seeks to push $\boldsymbol{n}$ outside the circle defined by the margin μ, and pull $\boldsymbol{p}$ inside. (b) Margin ranking loss values in function of δ_-, δ_+ (c) The ratio loss seeks to force δ_+ to be much smaller than δ_-. (d) Ratio loss values in function of δ_-, δ_+

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).

[1] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, 2015.

[2] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, 2015.

[3] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.

[4] G. H. M. Brown and S. Winder. Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1), pp.43-57, 2011.