Learning to Compare Image Patches via Convolutional Neural Networks

Sergey Zagoruyko, Nikos Komodakis
Universite Paris-Est, Ecole des Ponts ParisTech, France

Motivation. Comparing patches across images is probably one of the most fundamental tasks in computer vision and image analysis, that has given rise to the development of many hand-designed feature descriptors over the past years, including SIFT, that had a huge impact in the computer vision community. Yet, such manually designed descriptors may be unable to take into account in an optimal manner all the different factors that can affect the final appearance of image patches. On the other hand, nowadays one can easily gain access to (or even generate using available software) large datasets that contain patch correspondences between images [6]. This begs the following question: *can we make proper use of such datasets to automatically learn a similarity function for image patches* ? Our goal in this work is to affirmatively address the above question.

Contributions. More specifically, in this paper we succeed in achieving the following goals:

(i) We learn from scratch (*i.e.*, from raw image patches and without any manually-designed features) a general similarity function for patches that implicitly takes into account various types of transformations and effects. To that end, inspired by recent advances in neural architectures and deep learning, we choose to represent such a function in terms of a deep convolutional neural network [2].

(ii) We explore and propose a wide variety of different neural network models, highlighting at the same time network architectures that offer improved performance.

(iii) We show that such architectures outperform the state-of-the-art by a large margin and lead to feature descriptors for images patches with much better performance than manually designed descriptors (e.g, SIFT, DAISY) or other learnt descriptors such as [5]. Importantly, due to their convolutional nature, the resulting descriptors are very efficient to compute even in a dense manner.

Models. Given that there exist several ways in which patch pairs can be processed by the network or in which the information sharing can take place, we are also interested in addressing the issue of what network architecture is best to be used in a task like this. To that end, we explore many different variations on the architecture of the network such as: (i) *siamese* (this type of network resembles the idea of having a descriptor, in which case there are two branches – one per patch – in the network that share exactly the same architecture and the same set of weights), (ii) *pseudo-siamese*, (iii) *2-channel* (where, unlike previous models, there is no direct notion of descriptor in the architecture and the network proceeds directly with the similarity estimation), (iv) *central-surround two-stream* (where we modify the network to consist of two separate streams, central and surround, which enable a processing in the spatial domain that takes place over two different resolutions), (v) *spatial-pyramid-pooling (SPP)*, and (vi) *deep networks*. Many of the above variations can be used in conjunction with each other, thus leading to a wide range of models for comparing patches. Based on these, we draw interesting conclusions about which architectural choices help in improving performance in practice.

Experiments. We applied our approach on several problems and benchmark datasets, showing that it significantly outperforms the state-of-the-art.

For the first evaluation of our models, we used the standard benchmark dataset from [1] that consists of three subsets, Yosemite, Notre Dame, and Liberty, each of which contains more than 450,000 image patches (64 x 64 pixels), used to produce 500,000 ground-truth feature pairs for each dataset, with equal number of positive (correct) and negative (incorrect) matches. For evaluating our models on this dataset we use the evaluation protocol of [1] and report FPR95 on each of the six combinations of training and test sets. All our models outperform the previous state-of-the-art, highlighting a combination of 2-channel and central-surround two-stream architectures which managed to outperform it by a large margin, achieving 2.45 times better score than descriptors learnt with convex optimization [5] and 6.65 times better score than SIFT.

For wide baseline stereo evaluation we used Stretcha dataset [7]. The photometric cost is computed with each network, then MRF-based global optimization method is applied. All our models consistently outperform state-of-the-art wide-baseline stereo matching descriptor DAISY [7] both quantitatively and qualitatively.

We also tested our models on Mikolajczyk dataset for local descriptors evaluation [4]. We used MSER to detect feature points and computed matching scores with our networks and SIFT. For siamese networks we speed up matching by computing descriptors and then using l_2 distance matching or top decision network. All our networks outperform SIFT in terms of mAP score. Both 2-channel (including central-surround two-stream and deep versions) and SPP networks show especially good results.

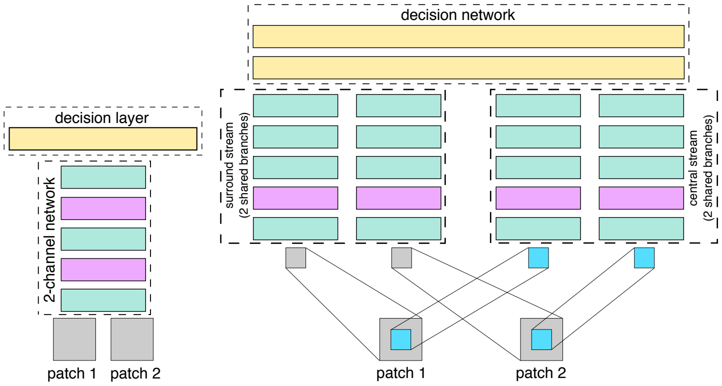

Figure 1: 2-channel (left) and siamese-two-stream (right) architectures. Color code used: cyan = Conv+ReLU, purple = max pooling, yellow = fully connected layer (ReLU exists between fully connected layers as well).

Train	Test	2ch-2stream	2ch	siam-2stream	siam-2stream-l_2	[5]
Yos	ND	**2.11**	3.05	5.29	5.58	6.82
Yos	Lib	**7.2**	8.59	11.51	12.84	14.58
ND	Yos	**4.1**	6.04	10.44	13.02	10.08
ND	Lib	**4.85**	6.05	6.45	8.79	12.42
Lib	Yos	**5**	7	9.02	13.24	11.18
Lib	ND	**1.9**	3.03	3.05	4.54	7.22
mean		**4.19**	5.63	7.63	9.67	10.38
mean(1,4)		**4.56**	5.93	8.42	10.06	10.98

Table 1: Performance of several models on benchmark [1].

[1] M. Brown, Gang Hua, and S. Winder. Discriminative learning of local image descriptors. *PAMI*, 33(1):43–57, Jan 2011. ISSN 0162-8828.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[4] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *PAMI*, 2005.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *PAMI*, 2014.

[6] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 80(2): 189–210, November 2008.

[7] E. Tola, V.Lepetit, and P. Fua. A Fast Local Descriptor for Dense Matching. In *CVPR*, Alaska, USA, 2008.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

Source code and models are available online at http://imagine.enpc.fr/~zagoruys/deepcompare.html