This CVPR2015 extended abstract is the Open Access version, provided by the Computer Vision Foundation.

MatchNet: Unifying Feature and Metric Learning for Patch-based Matching

Xufeng Han[1] Thomas Leung[2] Yangqing Jia[2] Rahul Sukthankar[2] Alexander C. Berg[1]
[1]University of North Carolina at Chapel Hill [2]Google Research

Patch-based image matching is used extensively in computer vision. Motivated by recent successes on learning feature representations and on learning feature comparison functions, we propose a unified approach to combining both for training a patch matching system. Our system, dubbed MatchNet, consists of a deep convolutional network that extracts features from patches and a network of three fully connected layers that computes a similarity between the extracted features. We show in this paper how to construct the two networks and jointly train them with a sampler. We also show that the proposed unified approach improves patch matching accuracy over previous state-of-the-art results [2] on a standard patch dataset [3], while reducing the storage requirement for descriptors.

MatchNet is a deep-network architecture (Fig. 1) for jointly learning a feature network (Fig. 1 B) and a metric network (Fig. 1 A). It consists of several types of layers commonly used in deep-networks for computer vision. The feature network is used for feature encoding, with an optional bottleneck layer discriminately trained to reduce feature dimension. The metric network is used for feature comparison. In training, the feature net is applied as two "towers" on pairs of patches with shared parameters. Output from the two towers are concatenated as the metric network's input.

In training (Fig. 1 C), the entire network is jointly trained to minimize the cross-entropy error over labeled patch-pairs generated from a sampler. We use stochastic gradient descent solver and a batch size of 32. The matching (+) and non-matching (-) pairs are highly unbalanced. We use a sampler to generate equal number of positives and negatives in each mini batch so that the network will not be overly biased towards negative decisions. The sampler also enforces variety to prevent overfitting to a limited negative set.

In prediction (Fig. 1 D), we use the two sub-networks in two stages: first we encode each patch using the feature network; then we compute pairwise matching scores using the metric network. On one NVIDIA K40 GPU, the feature net without bottleneck: 3.56K patch/sec; the metric net (B=128, F=512) (See Table 1): 416.6K pair/sec.

We following protocol established in [1] and evaluate MatchNet on a standard large patch dataset [3], which contains more than 1.5 million patches in total. Table 1 shows SIFT baselines, previous state-of-the-art and MatchNet's. We also evaluate the performance of MatchNet with quantized features. Each element v in the ReLU outputs from the bottleneck layer is quantized as $q(v) = \min(2^n - 1, \lfloor (2^n - 1)v/M \rfloor)$, where n is the number of resulting bits. Table 2 shows the results.

We make pre-trained MatchNets available at http://www.cs.unc.edu/~xufeng/matchnet.

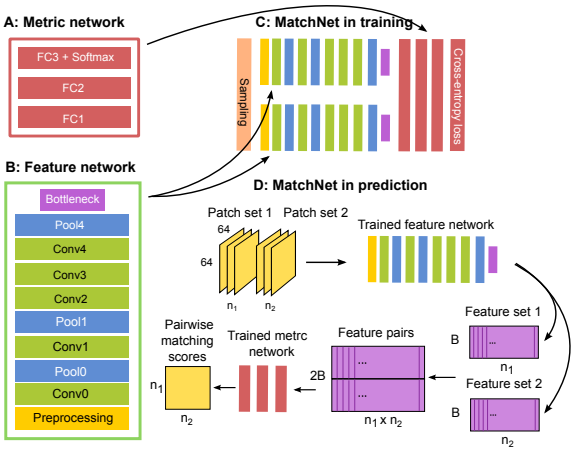

Figure 1: MatchNet.

Table 2: Accuracy vs. quantization tradeoff for the 64-1024×1024 network. In the first column, the first value is bits per dimension. The second value is average bits per feature vector. It is computed using $64 + 64 \times 0.679 \times b$, where b is the number of bits per dimension, and the average density (non-zeros) of the feature vector is 67.9%. Numbers in the middle are Error@95%. The first row is for the unquantized features.

	Notr.	Yos.	Lib.	Yos.	Lib.	Notr.
# of bits	Lib		Notr.		Yos.	
32 (1456)	9.82	14.27	5.02	9.15	14.15	13.20
8 (411.7)	9.84	14.33	5.06	9.21	14.17	13.21
7 (368.6)	9.82	14.20	5.04	9.23	14.21	13.19
6 (324.7)	9.81	14.22	5.15	9.30	14.27	13.29
5 (281.3)	10.19	14.58	5.33	9.59	14.66	13.39
4 (237.8)	11.37	15.27	6.27	10.93	15.59	14.07

[1] Matthew Brown, Gang Hua, and Simon A. J. Winder. Discriminative learning of local image descriptors. *IEEE TPAMI*, 33(1):43–57, 2011.

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

[3] http://www.cs.ubc.ca/~mbrown/patchdata/patchdata.html.

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

Table 1: UBC matching results. Numbers are Error@95% in percentage. F and B are dimensions for fully-connected and bottleneck layers, respectively. **Bold** numbers are the best results across all conditions. Underlined numbers are better than the previous state-of-the-art results with similar feature dimension.

Training		Notredame	Yosemite	Liberty	Yosemite	Liberty	Notredame
Test	Feature Dim.	Liberty		Notredame		Yosemite	
nSIFT + L2 (no training)	128d	29.84		22.53		27.29	
nSIFT squared diff. + linearSVM	128d	26.54	27.07	19.65	19.87	25.12	24.71
nSIFT concat. + NNet (F=512)	256d	20.44	22.23	14.35	14.84	21.41	20.65
Simonyan et al (2014) [2] PR	<640d	16.56	17.32	9.88	9.49	11.89	11.11
Simonyan et al (2014) discrim. proj.	<80d	12.42	14.58	7.22	6.17	11.18	10.08
Simonyan et al (2014) discrim. proj.	<64d	12.88	14.82	7.52	7.11	11.63	10.54
MatchNet (F=1024, B=64)	64d	9.82	14.27	5.02	9.15	14.15	13.20
MatchNet (F=512, B=128)	128d	9.48	15.40	5.18	8.27	14.40	12.17
MatchNet (F=512, B=512)	512d	8.84	13.02	4.75	7.70	13.58	11.00
MatchNet (F=512, w/o bottleneck)	4096d	**6.90**	**10.77**	**3.87**	**5.67**	**10.88**	**8.39**