Metric imitation by manifold transfer for efficient vision applications

Dengxin Dai[1], Till Kroeger[1], Radu Timofte[1], Luc Van Gool[1,2]
[1] Computer Vision Lab, ETH Zurich. [2] VISICS, ESAT/PSI, KU Leuven.

Motivation. The problem of learning good metrics is of great theoretical and practical interest in information processing [4]. Applications include clustering, image annotation, retrieval, classification, among others. The drawback of current metric learning systems is that they require human annotations (*e.g.* full category labels or pairwise constraints), which are expensive to acquire.

This paper learns good metrics for efficient vision applications without using human annotation. We draw inspiration from transfer learning and propose a novel approach, coined Metric Imitation (MI). MI takes state-of-the-art, off-the-shelf, but computationally expensive features (*e.g.* the CNN feature [2], Object Bank (OB) [5], and Sift-llc [7]) as source features (SFs) and cheap features (*e.g.* LBP and GIST) as target features (TFs) to learn good metrics for the TFs by imitating the metrics computed over the SFs. MI is a general framework and can at least be applied to: 1) creating efficient solutions to good metrics when SFs are more powerful than TFs, but are computationally more expensive and/or more memory-hungry; and 2) creating good metrics for TFs when SFs contain privileged information but are not available at testing time.

Method. MI comes out of a marriage of advances in metric learning and transfer learning. On the one hand, metric learning now is able to learn good metrics over general features for specific tasks [4], with human supervision. On the other hand, transfer learning can now transfer knowledge (often classifiers) learned in one domain of interest to another, for which no further human supervision is necessary. Observing both developments then begs the question whether metrics can be computed over one feature (*i.e.* SFs) to then be transferred to the domain of another feature (*i.e.* TFs) and automatically supervise the metric learning process there. In this paper, we demonstrate this for several vision tasks. The main advantages of the method are: 1) it performs in an unsupervised manner, i.e. without human annotation; 2) it can be efficient as only TFs are needed during test time; 3) it can inject domain knowledge carried through SFs to TFs for metric computation.

The metric is learned in the framework of Mahalanobis distance learning, *i.e.* a linear mapping function of TFs is learned and applied prior to performing the Euclidean distance metric. Specifically, the method works as follows: 1) it translates the properties of metrics computed over SFs into manifold geometries; 2) it transfers the manifold to the domain of the TFs as view-independent properties; and 3) it learns a mapping function of the TFs so that the transferred manifold is approximated as well as possible in the transformed space. By doing this, the *neighborhood properties* of data computed over the SFs are preserved in the transformed space of TFs, *i.e* close neighbors are still close. The reason for ensuring this is that neighbors search is enormously important in many vision applications, such as clustering, retrieval, and classification. The local properties (relations) of the manifold can be quantified in a variety of ways, and we used LLE [6] and LapEigen (Lap) [1] to encode local linearity and local pairwise distance, respectively. With the transferred manifold, the mapping function is learned by solving a generalized eigenvector problem, by following [3].

Experiments. In accordance with our previous remarks, the usability of MI is validated in two scenarios of metric learning: 1) compute good solutions for cheap features; and 2) transfer privileged information to target domain. For the first scenario, MI was tested on instance-based object retrieval using the INRIA Holiday dataset and the UKbench dataset, and on category-based image retrieval and image clustering using four other datasets: Scene-15 (S-15), CUReT-61 (C-61), Caltech-101 (C-101), and Event-8 (E-8). Three sophisticated, high-dimensional features were used as the SFs: OB [5], Sift-llc [7], and the CNN feature [2], and three general, cheap features used as the TFs: GIST, LBP, and PHOG. Extensive experiments show that MI consistently and significantly outperforms the metrics computed directly over the same TFs, while getting close to the metrics from the computationally more expensive SFs in some cases. See the results in Table 1 and Table 2. For the second scenario, MI was evaluated on example-based image super-resolution. Patches of high-resolution images, which are not available at testing time, are used as SFs and patches of low-resolution images as TFs. Experiments show that MI is able to create an efficient solution to k-NN-based methods, without sacrificing any performance relative to the state-of-the-art. See Fig. 1 for examples.

Bicubic MI

Figure 1: Examples with an upscaling factor $\times 4$. Best seen on the screen. Images are obtained from the Internet.

Table 1: Purity of clustering by Metric Imitation (MI), where 50% of the images are used for training and the rest for testing.

	TFs	MI		SFs	MI		SFs	MI		SFs
	LBP	LLE	Lap	SIFT-llc	LLE	Lap	CNN	LLE	Lap	OB
S-15	0.36	0.40	0.46	0.49	0.47	0.48	0.69	0.42	0.48	0.54
C-61	0.33	0.44	0.46	0.39	0.33	0.41	0.60	0.31	0.37	0.44
C-101	0.32	0.34	0.34	0.51	0.37	0.36	0.68	0.37	0.35	0.52
E-8	0.39	0.46	0.46	0.57	0.47	0.47	0.82	0.48	0.48	0.46

Table 2: MAP of category-based image retrieval by MI with the concatenation of LBP, GIST and PHOG (LGP) used as the TFs. 50% images are used for training and the rest for testing. Recall is set to 0.1.

	TFs	MI		SFs	MI		SFs	MI		SFs
	LGP	LLE	Lap	SIFT-llc	LLE	Lap	CNN	LLE	Lap	OB
S-15	0.52	0.60	0.61	0.60	0.64	0.64	0.72	0.62	0.63	0.65
C-61	0.84	0.95	0.93	0.90	0.94	0.96	0.95	0.92	0.90	0.91
C-101	0.42	0.48	0.46	0.57	0.51	0.51	0.79	0.48	0.48	0.59
E-8	0.52	0.63	0.63	0.70	0.65	0.64	0.88	0.60	0.56	0.58

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[3] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *ICCV*, 2005.

[4] Brian Kulis. Metric learning: A survey. *Foundations & Trends in Machine Learning*, 5(4):287–364, 2012.

[5] Li-Jia Li, Hao Su, Eric P. Xing, and Fei-Fei Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[6] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[7] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage. The work is supported by the ERC Advanced Grant VarCity.