Fine-Grained Visual Categorization via Multi-stage Metric Learning

Qi Qian[1], Rong Jin[1], Shenghuo Zhu[2], Yuanqing Lin[3]
[1]Department of Computer Science and Engineering, Michigan State University. [2]Alibaba Group. [3]NEC Laboratories America.

Fine-grained visual categorization (FGVC) aims to distinguish objects in subordinate classes. For example, dog images are classified into different breeds of dogs, such as "Chihuahua", "Pug", "Samoyed", etc. One challenge of FGVC is that it has to handle the co-occurrence of two somewhat contradictory requirements: 1) it needs to distinguish many similar classes (e.g., the dog breeds that only have subtle differences), and 2) it needs to deal with the large intra-class variation (e.g., caused by different poses, examples, etc.).

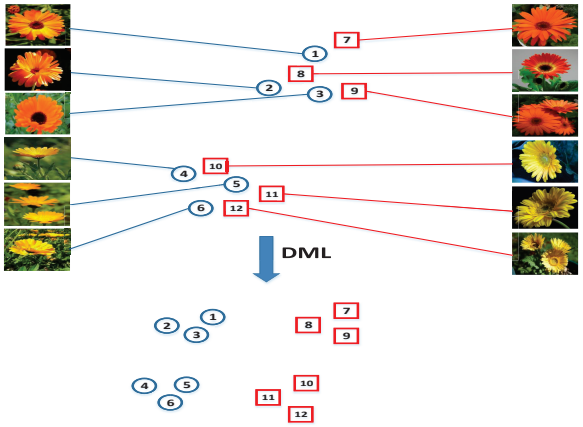

Figure 1: Illustration of how DML works.

The popular pipeline for FVGC consists of two steps, feature extraction step and classification step. In this paper, we simply take the state-of-the-art deep learning features without any other operators (e.g., segmentation) and focus on studying better classification approach to address the aforementioned two co-occurring requirements in FGVC. For the classification step, many existing FGVC methods directly learn a single classifier for each fine-grained class using the one-vs-all strategy. In this paper we propose a distance metric learning (DML) approach, aiming to explicitly handle the two co-occurring requirements with a *single* metric. Fig. 1 illustrates how DML works for FGVC. It learns a distance metric that pulls neighboring data points of the same class close to each other and pushes data points from different classes far apart. By varying the neighborhood size when learning the metric, it is able to effectively handle the tradeoff between the inter-class and intra-class variation.

There are three challenges in learning a metric directly from the original high dimensional image space:

- *Large number of constraints*: A large number of training constraints are usually required to avoid the overfitting of high dimensional DML. The total number of triplet constraints could be up to $\mathcal{O}(n^3)$ where n is the number of examples.
- *Computational challenge*: DML has to learn a matrix of size $d \times d$, where d is the dimensionality of data and $d = 134{,}016$ in our study. The $\mathcal{O}(d^2)$ number of variables leads to two computational challenges in finding the optimal metric. First, it results in a slower convergence rate in solving the related optimization problem. Second, to ensure the learned metric to be positive semi-definitive (PSD), most DML algorithms require, at *every* iteration of optimization, projecting the intermediate solution onto a PSD cone, an expensive operation with complexity of $\mathcal{O}(d^3)$ (at least $\mathcal{O}(d^2)$).
- *Storage limitation*: It can be expensive to simply save $\mathcal{O}(d^2)$ number of variables in memory. For example, in our study, it would take more than 130 GB to store the completed metric in memory, which adds more complexity to the already difficult optimization problem.

In this work, we propose a multi-stage metric learning framework for high dimensional DML that explicitly addresses these challenges as follows.

- To deal with a large number of constraints used by high dimensional DML, we divide the original optimization problem into multiple stages. At each stage, only a small subset of constraints that are difficult to be classified by the currently learned metric will be adaptively sampled and used to improve the learned metric. By setting the regularizer appropriately, we can prove that the final solution is optimized over all appeared constraints.
- To handle the computational challenge in each subproblem, we extend the theory of *dual random projection*, which was originally developed for linear classification problems, to DML. The proposed method enjoys the efficiency of random projection, and on the other hand learns a distance metric of size $d \times d$. This is in contrast to most dimensionality reduction methods that learn a metric in a *reduced* space.
- To handle the storage problem, we propose to maintain a low rank copy of the learned metric by a randomized algorithm for low rank matrix approximation. It not only accelerates the whole learning process but also regularizes the learned metric to avoid overfitting.

Fig. 2 summaries the framework of the proposed multi-stage metric learning approach.

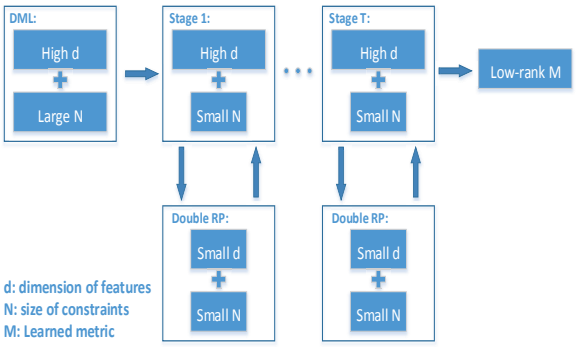

Figure 2: The framework of the proposed method.

We conduct extensive comparisons on benchmark FGVC datasets to verify the effectiveness and efficiency of the proposed method. Table 1 compares the performance of the proposed method (denoted as "MsML") to the best result reported by the state-of-the-art FGVC methods (denoted as "FGVC"). Note that we do not apply segmentation or localization operators, which are adopted by most of existing FGVC methods, to the proposed method for efficiency.

Table 1: Comparison of mean accuracy(%).

Datasets	FGVC	MsML
cats&dogs [3]	59.21	81.18
102flowers [2]	85.60	89.45
birds11 [4]	64.96	67.86
Stanford dogs [1]	50.10	70.31

[1] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-fei Li. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.

[2] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008.

[3] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

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