An Integrated Framework for High Dimensional Distance Metric Learning and Its Application to Fine-Grained Visual Categorization

Qi Qian, Rong Jin
Department of Computer Science and Engineering
Michigan State University, East Lansing, MI, 48824, USA
qianqi, rongjin@cse.msu.edu

Shenghuo Zhu and Yuanqing Lin
NEC Laboratories America, Cupertino, CA, 95014, USA
zsh, ylin@nec-labs.com

Abstract

In this paper, we focus on distance metric learning (DML) for high dimensional data and its application to fine-grained visual categorization. The challenges of high dimensional DML arise in three aspects. First, the high dimensionality leads to a large-scale optimization problem to be solved that is computationally expensive. Second, the high dimensionality requires a large storage space (i.e. $O(d^2)$ where $d$ is the dimensionality) for saving the learned metric. Third, the high dimensionality requires a large number of constraints for training that adds more complexity to the already difficult optimization problem. We develop an integrated framework for high dimensional DML that explicitly addresses the three challenges by exploiting the techniques of dual random projection, randomized low rank matrix approximation, and adaptive sampling. We demonstrate the effectiveness of the proposed algorithm for high dimensional DML by fine-grained visual categorization (FGVC), a challenging prediction problem that needs to capture the subtle difference among image classes. Our empirical study shows that the proposed algorithm is both effective and efficient for FGVC compared to the state-of-the-art approaches.

1. Introduction

Distance metric learning (DML) [13] aims to learn a Mahalanobis distance metric that keeps data points of the same class close to each other and data points from different classes far apart. It has been successfully applied to several important problems in computer vision [4, 5, 10, 12] and the new space spanned by the learned metric serves much better than the original space. Although numerous algorithms have been developed for DML [7, 14, 18, 23], most of them are limited to low dimensional data (i.e. no more than a few hundred dimensions). In this study, we focus on DML for high dimensional data. The main challenges of high dimensional DML arise from the fact that DML has to learn a matrix of size $d \times d$, where $d$ is the dimensionality of data and $d = 82, 560$ in our study. As a result, the number of variables increases quadratically in $d$. The $O(d^2)$ 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 [18]. 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 $O(d^3)$ (at least $O(d^2)$). Besides the computational issues, the $O(d^2)$ number of variables could also result in a storage problem. For instance, in the case for our study, it would take more than 50 GB to save the distance metric in memory. Finally, to avoid the over-fitting of high dimensional DML, a large number of training examples are usually required, which adds more complexity to the already difficult optimization problem.

A straightforward approach toward high dimensional DML is to reduce the dimensionality of data using the methods such as principle component analysis (PCA) [23] and random projection [21]. The main problem with most dimensionality reduction methods is that they are unable to take into account the supervised information, and as a result, the subspaces identified by the dimensionality reduction methods are usually suboptimal. In this work, we develop an integrated framework that combines and extends several state-of-the-art machine learning techniques to explicitly address the challenges of high dimensional DML. First, to handle the computational challenge, we extend the...
Figure 1. The first column includes ten query images, highlighted by green bounding boxes. Columns 2-4 include the most similar images with the shortest Euclidean distance to the query images. Columns 5-7 show the most similar images measured by the learned distance metric. Images in columns 2-7 are highlighted by red bounding boxes when they share the same category as the query image, and are highlighted by blue bounding boxes if they are assigned to different categories. Note that we use the cropped images for illustration while we keep the original images in the experiments.

theory of dual random projection [27], which was originally developed for classification problems, to metric learning. The proposed method, on one hand, improves the computational efficiency by reducing the dimensionality, and on the other hand is able to learn 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. Second, to handle the storage problem, we propose to maintain a low rank copy of the learned metric at each iteration of optimization, and develop a randomized algorithm for efficient low rank matrix approximation, which also projects the learned metric onto the PSD cone simultaneously. Finally, to deal with a large number of constraints used by high dimensional DML, an adaptive sampling is developed. At each epoch of adaptive sampling, only a small subset of constraints that are difficult to be classified by the currently learned metric...
will be sampled and used to improve the learned metric. We verify the effectiveness of the proposed algorithm for high dimensional DML in the domain of fine-grained visual categorization (FGVC). FGVC is significantly more challenging than the traditional studies of image classification because in FGVC, images from different classes can be visually similar and even human beings cannot easily tell their difference [19]. FGVC has been found in many real world classification applications, especially for species recognition, e.g., cats and dogs [11, 17], birds [24] and flowers [16], etc. Figure 1 shows a few examples illustrating the challenges of FGVC. The first column in Figure 1 includes ten query images, and columns 2-4 show the images with the highest similarity measured by the Euclidean distance. Although it is arguable that the images in columns 2-4 are visually similar to the query images, very small portion of them are assigned to the same class as the query image, indicating that FGVC is a challenging problem.

Many studies of FGVC focus on extracting non-conventional visual features that can be informative to the subtle difference between the fine-grained categories [1, 2, 19]. However, several recent studies [1, 15] show that the conventional visual feature extraction methods could significantly outperform the feature extraction methods that are failed to FGVC, provided that the number of extracted features is sufficiently large (i.e. \(O(100,000))\). This observation makes FGVC an ideal domain for testing the proposed high dimensional DML algorithm. In columns 5-7 of Figure 1, we show the most similar images measured by the learned distance metric using the proposed algorithm. We observe that most of the retrieved images share the same categories as the query images, suggesting that high dimensional DML is effective for FGVC.

The rest of the paper is organized as follows: Section 2 summarizes related work for FGVC and DML. Section 3 describes the details of the proposed method. Section 4 shows the results of the empirical study, and Section 5 concludes this work with future directions.

2. Related Work

Many algorithms have been developed for metric learning and a detailed investigation could be found in the survey [13, 25]. Although numerous studies were devoted to metric learning, few examined the challenges of high dimensional DML. A common approach toward high dimensional DML is to project data into a low dimensional space, and learn a metric in the space of reduced dimension, which often leads to a suboptimal performance. An alternative approach to high dimensional DML is to assume \(M\) to be low rank by writing \(M = LL^T\) [7, 23], where \(L\) is a tall rectangular matrix. Instead of learning \(M\), these approaches directly learn \(L\) from data. The main shortcoming of this approach is that it has to solve a non-convex optimization problem, making it computationally less attractive. Several recent studies [14, 18] address the problem of high dimensional DML by assume \(M\) to be sparse. Although these approaches resolve the storage problem, it still suffers from high cost in optimizing \(O(d^2)\) variables.

As already mentioned in the introduction section, most of study on FGVC focused on developing special visual features that are informative to the subtle difference between different classes. For example, Sfar, et al. [19] developed visual features that explicitly exploit the structure information of botanical species to improve the classification on leaves datasets. Berg, et al. [2] focused on exploring the most discriminative parts (e.g., eyes and wings of birds and eyes of human) in order to distinguish images from different fine-grained classes. Angelova, et al. [1] developed special algorithms for background extraction that allows them to focus on the visual features of foreground objects.

It is interesting to note that despite the diversity in feature extraction, almost all studies of FGVC use linear SVM with one-vs-all scheme for classification. Few study applies distance metric learning to FGVC, mostly due to its high computational cost when coming to high dimensional data. On the other hand, cost of one-vs-all SVM heavily depends on the number of classes and extracting specific features requests additional time consumption. Extracting conventional features usually takes 2-3 seconds per image while that for FGVC could be more than 30 seconds. Since real world applications (e.g., online search, iphone apps) are very sensitive to the cost of feature extraction, we only consider the conventional features for practical and fills this gap by developing an efficient and scalable algorithm for high dimensional DML.

3. Learning a Low Rank Distance Metric from High Dimensional Data

The proposed DML algorithm focuses on triplet constraints, which has been shown to be more effective than pairwise constraints for DML [23]. Let \(D = \{(x_i, y_i), i = 1, \ldots, n\}\) be a collection of \(n\) training examples, where \(x_i \in \mathbb{R}^d\) and \(y_i\) is the class assignment of \(x_i\). Given a distance metric \(M\), the squared Mahalanobis distance between two data points \(x_i\) and \(x_j\) is measured by \(d_M(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j)\). Let \(\{x_{i_t}, x_{j_t}, x_{k_t}\}, t = 1, \ldots, N\) be a set of \(N\) triplet constraints derived from the training examples in \(D\). Since in each constraint \((x_{i_t}, x_{j_t}, x_{k_t})\), \(x_{i_t}\) and \(x_{j_t}\) are assumed to share the same class that is different from that of \(x_{k_t}\), we would expect \(d_M(x_{i_t}, x_{j_t}) < d_M(x_{i_t}, x_{k_t})\). As a result, the optimal distance metric \(M\) is learned by solving the following optimization problem

\[
\min_{M \in S_d, M \succeq 0} \frac{\lambda}{2} \|M\|_F^2 + \frac{1}{N} \sum_{t=1}^{N} \ell(d_M(x_{i_t}, x_{k_t}) - d_M(x_{i_t}, x_{j_t})) \quad (1)
\]
where $S_d$ includes all $d \times d$ real symmetric matrices and $\ell(\cdot)$ is a loss function that penalizes the objective function when $d_M(\mathbf{x}_i', \mathbf{x}_j')$ is not significantly larger than $d_M(\mathbf{x}_i', \mathbf{x}_k')$. In this study, we choose the smoothed hinge loss [20] that appears to be more effective than the hinge loss while keeps the benefit of large margin:

$$
\ell(x) = \begin{cases} 
0 & : x > 1 \\
1 - x - \gamma / 2 & : x < 1 - \gamma \\
\frac{\gamma}{2}(1 - x)^2 & : \text{o.w.}
\end{cases}
$$

One main computational challenge of DML comes from the PSD constraint $M \succeq 0$ in (1). We address this challenge by following the one projection paradigm [3] that first learns a metric $M$ without the PSD constraint and then projects $M$ to the PSD cone at the very end of the learning process. Hence, in this study, we will focus on the following optimization problem

$$
\min_{M \in S_d} \frac{\lambda}{2} \|M\|_F^2 + \frac{1}{N} \sum_{t=1}^{N} \ell(\langle A_t, M \rangle)
$$

(2)

where $A_t = (\mathbf{x}_i' - \mathbf{x}_k')(\mathbf{x}_j' - \mathbf{x}_l')^\top - (\mathbf{x}_i' - \mathbf{x}_j')(\mathbf{x}_k' - \mathbf{x}_l')^\top$ is introduced as a matrix representation for each triplet constraint, and $\langle \cdot, \cdot \rangle$ represents the dot product between two matrices.

The main difficulty with high dimensional DML arises from the fact that the number of independent variables in $M$ is $O(d^2)$. More specifically, the large size of $M$ will lead to a high computational cost in solving the optimization problem in (2) and a large storage requirement for saving $M$. The high dimensionality $d$ also demands a large number of triplet constraints used for training in order to avoid the over-fitting problem, which in return adds more complexity to solve the optimization problem. We will discuss the strategies to address these challenges in the next three subsections, and summarize the integrated framework for high dimensional DML at the end of this section.

3.1. Optimization Challenge: Dual Random Projection (DuRP)

We first extend the theory of dual random projection, which was originally developed for classification problems, to metric learning. It first projects all the data points into a low dimensional random subspace, and computes the distance metric in the random subspace. It then estimates the dual variables using the computed metric in the low dimensional space, and calculates the distance metric in the original space using the estimated dual variables.

Let $R_1, R_2 \in \mathbb{R}^{d \times m}$ be two Gaussian random matrices, where $m$ is the number of random projections $m \ll d$ and $R_1^{ij}, R_2^{ij} \sim \mathcal{N}(0, 1/m)$. For each triplet constraint, we project its representation $A_t$ into low dimensional space using the random matrices, i.e. $\hat{A}_t = R_1^\top A_t R_2$. It is easy to verify $\langle A_a, A_b \rangle = \mathbb{E}[\langle \hat{A}_a, \hat{A}_b \rangle]$, implying that the random projected representation $\{\hat{A}_t\}_{t=1}^N$ preserves the pairwise similarity between any two triplet constraints, a key property that ensures the correctness of the randomized algorithm [27]. Using $\{\hat{A}_t\}_{t=1}^N$, we will learn a metric $\hat{M} \in \mathbb{R}^{m \times m}$ in the reduced space by solving the following optimization problem

$$
\min_{M \in S_m} \frac{\lambda}{2} \|M\|_F^2 + \frac{1}{N} \sum_{t=1}^{N} \ell(\langle \hat{A}_t, \hat{M} \rangle)
$$

(3)

Since the size of $\hat{M} \in \mathbb{R}^{m \times m}$ is significantly smaller than that of $M$, (3) can be solved much more efficiently than (2). In our implementation, a simple stochastic gradient descent (SGD) is developed to efficiently solve the optimization problem in (3). Given $\hat{M}$, following the theory in [27], the final distance metric $M \in \mathbb{R}^{d \times d}$ in the original space is estimated as

$$
M = \frac{1}{\lambda N} \sum_{t=1}^{N} \alpha_t A_t
$$

(4)

where the dual variable $\alpha_t$ is given by $\ell'(\langle \hat{A}_t, \hat{M} \rangle)$.

3.2. Storage Challenge: Low Rank Approximation

Although (4) allows us to recover the distance metric $M$ in original $d$ dimensional space from the dual variables $\{\alpha_t\}_{t=1}^N$, it is expensive, if not impossible, to save $M$ in the memory when $d$ is very large. To address this challenge, instead of saving $M$, we propose to save the low rank approximation of $M$. More specifically, let $\sigma_1, \ldots, \sigma_r$ be the first $r \ll d$ eigenvalues of $M$, and let $u_1, \ldots, u_r$ be the corresponding eigenvectors. We approximate $M$ by a low rank matrix $M' = \sum_{i=1}^{r} \sigma_i u_i u_i^\top = LL^\top$. Unlike $M$ that requires $O(d^2)$ storage space, it only takes $O(rd)$ space to save $M'$. However, the key issue is how to efficiently compute the eigenvectors and eigenvalues of $M$. This is particularly challenging in our case as $M$ in (4) can not be computed explicitly due to its large size. We address this challenge by exploiting the randomized algorithm that was recently developed for matrix factorization [9].

First, we express the summation in (4) as matrix multiplication. To this end, we construct a sparse matrix $C \in \mathbb{R}^{n \times m}$. For each triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$, we denote its dual variable by $\alpha = \ell'(\langle \hat{A}_t, \hat{M} \rangle)$ and set the corresponding entries in $C$ as

$$
C(i,j) = \frac{\alpha}{\lambda N}, C(j,i) = \frac{\alpha}{\lambda N}, C(j,j) = -\frac{\alpha}{\lambda N}, C(i,k) = -\frac{\alpha}{\lambda N}, C(k,k) = \frac{\alpha}{\lambda N}
$$

(5)

It is easy to verify that $M$ in (4) can also be written as

$$
M = XCX^\top
$$
Algorithm 1 An Efficient Algorithm for Recovering $M$ and Project It onto PSD Cone from $\hat{M}$

1: **Input:** Dataset $X \in \mathbb{R}^{d \times n}$, $\hat{M} \in \mathbb{R}^{m \times m}$, the number of random combinations $q$
2: Compute a Gaussian random matrix $R \in \mathbb{R}^{d \times q}$
3: Compute a sparse matrix $C$ using (5)
4: $Y = RX^\top$, $Y = Y \times C$, $Y = Y \times X$
5: $[Q, R] = qr(Y)$
6: $B = Q' \times X'$, $B = B \times C$, $B = B \times X$
7: $[U, \Sigma] = \text{eig}(B)$
8: $U = Q \ast U$
9: **return** $L = [\sqrt{\sigma_1}u_1, \cdots, \sqrt{\sigma_r}u_r]$ and $M = LL^\top$

where $u_i$ is the $i$th column of $U$ and $\sigma_i$ is the $i$th positive diagonal element of $\Sigma$

Second, we exploit a randomized algorithm to efficiently compute the eigen-decomposition of $M$. According to the theory in [9], the top eigenvectors of $M$ can be well approximated by random combinations of column vectors in $M$. More specifically, let $R \in \mathbb{R}^{d \times q}$ be a Gaussian random matrix. Then, with an overwhelming probability, most of the top $r$ eigenvectors of $M$ lie in the subspace spanned by the column vectors in $MR$ provided $q \geq r + k$, where $k$ is usually a constant independent from $d$. Considering that $M$ should be a PSD matrix and only positive eigenvalues will be preserved, we adopt $q \geq 2r + k$ in our case. Since $MR = XCX^\top R$, it is able to compute the top $r$ eigenvalues and eigenvectors of $M$ without having to compute $M$ explicitly. In addition, since $C$ is a sparse matrix, $MR$ can be computed efficiently in practice. The total cost of low rank approximation is only $O(qnd)$. The computational efficiency can be further improved by exploiting the distributed computing platform when $n \geq d$. By combining the above steps, we have the reconstruction algorithm given in Alg. 1, where $qr$ and $\text{eig}$ stand for QR and eigen decomposition of a matrix.

3.3. Constraints Challenge: Adaptive Sampling

In order to reliably determine the distance metric in a high dimensional space, a large number of training examples are needed to avoid the over-fitting problem. Since the number of triplet constraints can be $O(n^3)$, the number of summation terms in (3) can be extremely large, making it difficult to effectively solve the optimization problem in (3). To address this challenge, we develop an adaptive sampling strategy. It divides the learning process into multiple stages. At $s$-th stage, let $\hat{M}_{s-1}$ be the distance metric in the reduced subspace learned from the last stage. We first reconstruct the distance metric $M_{s-1}$ in the original $d$ dimensional space using Alg. 1. We then sample a subset of triplet constraints that are difficult to be classified by $M_{s-1}$. Alg. 2 shows the key steps of adaptive sampling. Given $\hat{M}_{s-1}$ and the sampled triplet constraints $S_s$, we update the distance metric by solving the following optimization problem

$$\min_{\tilde{M}_s \in S_m} \frac{\lambda}{2} ||\hat{M}_s - \tilde{M}_s||_F^2 + \frac{1}{2|S_s|} \sum_{t \in S_s} \ell(\langle \hat{A}_t, \tilde{M}_s \rangle) \quad (6)$$

To understand the motivation behind the optimization problem in (6), consider the objective function for the first $s$ stages, i.e.

$$\frac{\lambda}{2} ||\hat{M}||_F^2 + \sum_{k=1}^{s-1} \sum_{t \in D_k} \ell(\langle \hat{A}_t, \hat{M} \rangle) + \sum_{t \in D_s} \ell(\langle \hat{A}_t, \hat{M} \rangle) \quad (7)$$

Since $\hat{M}_{s-1}$, the solution obtained from the first $s - 1$ stages, approximately optimizes $L_{s-1}(\hat{M})$, using the fact that $L_{s-1}(\hat{M})$ is strongly convex, we have

$$L_{s-1}(\hat{M}) \approx L_{s-1}(\hat{M}_{s-1}) + \frac{\lambda}{2} ||\hat{M} - \hat{M}_{s-1}||_F^2 \quad (8)$$

By replacing $L_{s-1}(\hat{M})$ with the approximation in (8), we have the optimization problem in (6).

3.4. Summary: An Integrated Framework for High Dimensional DML

By putting all the pieces together, we have an integrated framework for high dimensional DML, where the key steps are shown in Alg. 3. Note that the sparse matrix $C$ is cumulated over all stages. To further improve the efficiency, we also exploit the distributed computing platform. The distributed computing is particularly effective for the realization of Alg. 1 because the matrix multiplication $XCX^\top R$ can be accomplished in parallel.
Algorithm 3 An Integrated Framework for High Dimensional DML (IDRP)

1: **Input:** Dataset $X \in \mathbb{R}^{d \times n}$, the number of random projections $m$, the number of random combinations $q$, and the number of stages $T$
2: Compute two Gaussian random matrices $R_1, R_2 \in \mathbb{R}^{d \times m}$
3: Initialize $\hat{M}_0 = 0 \in \mathbb{R}^{m \times m}$ and $M = 0 \in \mathbb{R}^{d \times d}$
4: for $s = 1, \ldots, T$ do
5:     Adaptively sample triplet constraints using Alg. 2
6:     Estimate $M_s$ by solving the optimization problem in (6) using SGD
7:     Recover the distance metric $M_s$ in the $d$ dimensional space using Alg. 1
8: end for
9: return $M_T$

4. Experiments

4.1. Setting

To verify the effectiveness and efficiency of the proposed algorithm for high dimensional DML, we apply it to fine-grained visual categorization (FGVC). We use the conventional pipeline for visual feature extraction that is outlined in [1]. Specifically, we extract HOG [6] features at 4 different scales and encode them to 8K dimensional feature dictionary by the LLC method [22]. A max pooling strategy is then used to aggregate local features into a single vector representation. Finally, a total of 82,560 features are extracted from each image. We note that this is in contrast to many studies of FGVC that use specially designed features. Most of existing DML methods cannot deal with such high dimensional problem.

We apply the proposed algorithm to learn a distance metric and use the learned metric together with a smoothed $k$-nearest neighbors classifier to predict the class assignments for test examples. Different from traditional $k$-NN, smoothed $k$-NN first obtains $k$ reference centers for each class by clustering images in each class into $k$ clusters. To predict the class assignment for a test image $x$, it computes a distance between $x$ and a class $C$ using the reference centers $C_1, \ldots, C_k$ as

$$\text{dis}(x, C) = -\frac{1}{\beta} \log \left( \sum_{j=1}^{k} \exp \left( -\beta |x - C_j|^2 \right) \right), \quad (9)$$

and assigns $x$ to the class with the shortest distance. The distance function given in (9) actually computes the soft min among the distance between $x$ and $C_j$, and we use hard min $\min_{1 \leq j \leq k} |x - C_j|^2$ when $\beta = 0$. We find that smoothed $k$-NN is more efficient than the traditional $k$-NN for large-scale learning problems. We refer to the classification approach based on the smoothed $k$-NN and the metric learned by the proposed algorithm as IDRP and the smoothed $k$-NN with Euclidean distance in the original space as Euclid. A linear SVM (LSVM) is used as the baseline in our study. We do not include any dimension reduction method (e.g., PCA) in our study due to the too large covariance matrix. We also include the state-of-the-art results for FGVC in our evaluation when they are available.

All the hyper-parameters in LSVM are tuned by 5-fold cross validation. The optimal regularization parameter for LSVM is searched in the range $\{10^i | i = -2, \cdots, 3\}$. The one-vs-all strategy, based on the implementation of LIBLINEAR [8], is used in the baseline LSVM to deal with the multi-class classification problem in FGVC. All the parameters used by the proposed IDRP algorithm are set empirically. More specifically, we set the number of random projections $m = 100$, the number of random combinations $q = 600$, and the nearest neighbor range $h$ as the half number of examples per class. These parameter values are used throughout all the experiments. Finally, instead of setting the number of stages $T$, we stop the iteration of the proposed algorithm when the largest eigenvalue of the learned metric is converged. Mean accuracy, a standard evaluation metric for FGVC, is used to evaluate the classification performance.

4.2. Comparison of Effectiveness

Table 1. Statistics for the datasets used in our empirical study. All datasets have 82,560 features.

| Small Datasets     | # Classes | #Train  | #Test  |
|--------------------|-----------|---------|--------|
| 102flowers        | 102       | 2,040   | 6,149  |
| birds             | 200       | 3,000   | 3,033  |
| cats&dogs         | 37        | 3,680   | 3,669  |
| Stanford dogs     | 120       | 12,000  | 8,580  |
| Honda cars        | 61        | 41,057  | 1,754  |
| Large-scale Datasets | # Classes | #Train  | #Test  |
| 578flowers        | 578       | 233,350 | 11,559 |

Figure 2. Example images from the Honda cars dataset.

Five small datasets are used in the first experiment.
Table 1 summarizes the information of these datasets. 102flowers is the Oxford flowers dataset for flower species [16]. birds is the Caltech-UCSD birds dataset for bird species [24] and we use the version with ground truth bounding box. cats&dogs is the Oxford cats and dogs dataset with different breeds of cats and dogs [17]. Stanford dogs is the dog species dataset [11]. Since the number of public large-scale fine-grained datasets is limited, we collect a Honda cars fine-grained image dataset, where each Honda car is annotated based on the Honda car models from different yeas. Fig. 2 shows the image examples from the Honda cars dataset. We use the standard split provided by these datasets. All the experiments on these small datasets are run on a single machine with 16 2.10GHz cores and 96 GB memory.

Table 2 summarizes the classification performance. The column titled 'FGVC' includes the state-of-the-art results reported on these datasets. More specifically, Angelova, et al. [1] shows the best classification results for the 102flowers, birds and cats&dogs datasets. Yang, et al. [26] reports the best result on Stanford dogs.

First, we observe that IDRP is more accurate than the baseline LSVM. This is not surprising because the distance metric is learned from the training examples of all class assignments. This is in contrast to the one-vs-all approach used in LSVM in which the classification function for a class $C$ is learned only by the class assignments of $C$. Second, we observe that IDRP yields almost identical performance as the state-of-the-art performance reported in 'FGVC', which verifies the effectiveness of the proposed algorithm for high dimensional DML. We note that for the 102flowers, birds and cats&dogs datasets, about 10,000 more specially designed visual features are introduced in [1] for FGVC. The fact that the proposed algorithm is able to achieve the same classification performance as the state-of-the-art approach but only using the general purpose visual features further demonstrates the power of the proposed algorithm for high dimensional DML. In addition, extracting conventional features only costs 1-2 seconds per image while that for FGVC is more than 6 seconds [1], which makes the proposed method more practical. Finally, IDRP is at least 20% better than Euclid on all datasets. Since the only difference between IDRP and Euclid lies in the metric used for distance measure, this result indicates that the distance metric used by IDRP is more effective in capturing the “semantic” difference among images.

After the experiments on small datasets, we apply the proposed method on the challenging large-scale flowers dataset 578flowers. It is comprised of 244,909 flower images from 578 flower species [1]. Since the dataset is larger than 77 GB in single precision, we run it in a distributed computing environment with 16 machines (we have discussed how to further improve the computational efficiency of the proposed algorithm using the distributed computing environment in Section 3.2). To the best of our knowledge, this is the largest dataset (i.e. 250,000 × 82,560) that DML has ever tried.

Table 2 summarizes the classification results for this dataset. Since LIBLINEAR cannot handle this large dataset directly, we use a stochastic version of SVM with the same configuration as IDRP. We obtain the state-of-the-art result from [1]. We again observe that the proposed IDRP algorithm significantly outperforms the baseline LSVM, and yields almost identical performance as the state-of-the-art FGVC approach. The classification result on this large dataset further confirms the effectiveness of the proposed algorithm for high dimensional DML.

To verify if the learned metric is able to capture the subtle difference among different fine-grained classes, in column 5-7 in Fig. 1, we show the most similar images measured by the distance $d_M(\cdot, \cdot)$ that uses the learned distance metric $M$. We use the red bounding boxes to highlight the retrieved images that share the same class assignment as that of the query image, and the blue bounding boxes for those retrieved images with different classes. We clearly observe that most of the returned images using the learned distance metric share the same class assignments of the query images. This is in contrast to the results based on the Euclidean distance (in column 2-4 in Fig. 1) where little similar images measured by the Euclidean distance are assigned to a different class than that of the query image.

4.3. Comparison of Efficiency

Table 3 compares the training time of the proposed algorithm for high dimensional DML to that of LSVM. IDRP is implemented by Julia, which is a little slower than C, while LSVM uses the LIBLINEAR package, the state-of-the-art algorithm for solving linear SVM implemented mostly in C. LSVM on 578flowers is implemented by an efficient stochastic method as we mentioned above. The time for extracting features is not included in the results reported in Table 3 because it is shared by both of methods. The running time for IDRP includes all operations (e.g., computing
random projection, low rank approximation, and adaptive sampling). We observe that the proposed method is comparable with or even faster than LSVM on almost all datasets except cats&dogs. The high computational cost of LSVM mostly comes from two aspects. First, LSVM has to train one classification model for each class, and becomes significantly slower when the number of classes is large. Second, the fact that images from different classes are visually similar makes it computationally difficult to find the optimal linear classifier that can separate images of one class from images from the other classes. In contrast, IDRP does not depend on the number of classes, which makes it more appropriate for many classes classification problem.

| Datasets       | LSVM  | IDRP  |
|----------------|-------|-------|
| i02flowers     | 4.08  | 4.41  |
| birds          | 9.11  | 4.79  |
| cats&dogs      | 3.35  | 9.53  |
| Stanford dogs  | 31.08 | 21.07 |
| Honda cars     | 638.14| 226.73|
| 578flowers     | 721.23| 447.62|

5. Conclusion

In this paper, we develop an integrated framework for high dimensional DML. It extends and combines three important machine learning techniques to address the challenges arising from high dimensional DML. More specifically, it extends the theory of dual random projection to address the optimization challenge for high dimensional data, a randomized algorithm for low rank matrix approximation for the storage challenge, and the adaptive sampling technique to handle the computational challenge arising from too many triplet constraints. We evaluate the effectiveness of the proposed metric learning algorithm to fine-grained visual categorization and the empirical study shows that the proposed method yields performance that is comparable to the state-of-the-art approaches for FGVC. In the future, we plan to combine the proposed DML algorithm with special visual features designed for FGVC to further improve the performance of FGVC.

References

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. In CVPR, 2013. 3, 6, 7
[2] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In CVPR, 2013. 3
[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. JMLR, 11:1109–1135, 2010. 4
[4] X. Chen, Z. Tong, H. Liu, and D. Cai. Metric learning with two-dimensional smoothness for visual analysis. In CVPR, pages 2533–2538, 2012. 1
[5] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In CVPR, 2013. 1
[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, pages 886–893, 2005. 6
[7] J. V. Davis and I. S. Dhillon. Structured metric learning for high dimensional problems. In KDD, pages 195–203, 2008. 1, 3
[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. JMLR, 9:1871–1874, 2008. 6
[9] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. ArXiv e-prints, Sept. 2009. 4, 5
[10] N. Jiang, W. Liu, and Y. Wu. Order determination and sparsity-regularized metric learning adaptive visual tracking. In CVPR, pages 1956–1963, 2012. 1
[11] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In First Workshop on Fine-Grained Visual Categorization, Colorado Springs, CO, June 2011. 3, 7
[12] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In CVPR, pages 2288–2295, 2012. 1
[13] B. Kulis. Metric learning: A survey. Foundations and Trends in Machine Learning, 5(4):287–364, 2013. 1, 3
[14] D. K. H. Lim, B. McFee, and G. Lanckriet. Robust structural metric learning. In ICML, 2013. 1, 3
[15] J. Lin and T. G. Dietterich. Is fine grained classification different? The Second Workshop on Fine-Grained Visual Categorization, 2013. 3
[16] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In IJCVGIP, pages 722–729, 2008. 3, 7
[17] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In CVPR, pages 3498–3505, 2012. 3, 7
[18] G.-J. Qi, J. Tang, Z.-J. Zha, T.-S. Chua, and H.-J. Zhang. An efficient sparse metric learning in high-dimensional space via \ell_1-penalized log-determinant regularization. In ICML, page 106, 2009. 1, 3
[19] A. R. Sfar, N. Boujemaa, and D. Geman. Vantage feature frames for fine-grained categorization. In CVPR, 2013. 3
[20] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. CoRR, abs/1209.1873, 2012. 4
[21] G. Tsagkatakis and A. E. Savakis. Manifold modeling with learned distance in random projection space for face recognition. In ICPAR, pages 653–656, 2010. 1
[22] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In CVPR, pages 3360–3367, 2010. 6
[23] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. JMLR, 10:207–244, 2009. 1, 3
[24] X. Chen, Z. Tong, H. Liu, and D. Cai. Metric learning with two-dimensional smoothness for visual analysis. In CVPR, pages 2533–2538, 2012. 1
[25] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In CVPR, 2013. 1
[26] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, pages 886–893, 2005. 6
[27] J. V. Davis and I. S. Dhillon. Structured metric learning for high dimensional problems. In KDD, pages 195–203, 2008. 1, 3
[28] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. JMLR, 9:1871–1874, 2008. 6
[29] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. ArXiv e-prints, Sept. 2009. 4, 5
[30] N. Jiang, W. Liu, and Y. Wu. Order determination and sparsity-regularized metric learning adaptive visual tracking. In CVPR, pages 1956–1963, 2012. 1
[31] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In First Workshop on Fine-Grained Visual Categorization, Colorado Springs, CO, June 2011. 3, 7
[32] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In CVPR, pages 2288–2295, 2012. 1
[33] B. Kulis. Metric learning: A survey. Foundations and Trends in Machine Learning, 5(4):287–364, 2013. 1, 3
[34] D. K. H. Lim, B. McFee, and G. Lanckriet. Robust structural metric learning. In ICML, 2013. 1, 3
[35] J. Lin and T. G. Dietterich. Is fine grained classification different? The Second Workshop on Fine-Grained Visual Categorization, 2013. 3
[36] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In IJCVGIP, pages 722–729, 2008. 3, 7
[37] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In CVPR, pages 3498–3505, 2012. 3, 7
[38] G.-J. Qi, J. Tang, Z.-J. Zha, T.-S. Chua, and H.-J. Zhang. An efficient sparse metric learning in high-dimensional space via \ell_1-penalized log-determinant regularization. In ICML, page 106, 2009. 1, 3
[39] A. R. Sfar, N. Boujemaa, and D. Geman. Vantage feature frames for fine-grained categorization. In CVPR, 2013. 3
[40] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. CoRR, abs/1209.1873, 2012. 4
[41] G. Tsagkatakis and A. E. Savakis. Manifold modeling with learned distance in random projection space for face recognition. In ICPAR, pages 653–656, 2010. 1
[42] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In CVPR, pages 3360–3367, 2010. 6
[43] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. JMLR, 10:207–244, 2009. 1, 3
[24] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 3, 7

[25] L. Yang and R. Jin. Distance metric learning: a comprehensive survey. 2006. 3

[26] S. Yang, L. Bo, J. Wang, and L. G. Shapiro. Unsupervised template learning for fine-grained object recognition. In NIPS, pages 3131–3139, 2012. 7

[27] L. Zhang, M. Mahdavi, R. Jin, T.-B. Yang, and S. Zhu. Recovering optimal solution by dual random projection. In arXiv:1211.3046, 2013. 2, 4