Towards Optimal Fine Grained Retrieval via Decorrelated Centralized Loss with Normalize-Scale Layer

Xiawu Zheng,[1] Rongrong Ji,[1,2*] Xiaoshuai Sun,[1] Baochang Zhang,[3] Yongjian Wu,[4] Feiyue Huang[4]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Cognitive Science,
School of Information Science and Engineering, Xiamen University, [2]Peng Cheng Laboratory, China
[3]Beihang University, [4]Tencent Youtu Lab, Tencent Technology (Shanghai) Co., Ltd

Abstract

Recent advances on fine-grained image retrieval prefer learning convolutional neural network (CNN) with specific fully-connect layer designed loss function for discriminative feature representation. Essentially, such loss should establish a robust metric to efficiently distinguish high-dimensional features within and outside fine-grained categories. To this end, the existing loss functions are defected in two aspects: (a) The feature relationship is encoded inside the training batch. Such a local scope leads to low accuracy. (b) The error is established by the mean square, which needs pairwise distance computation in training set and results in low efficiency. In this paper, we propose a novel metric learning scheme, termed Normalize-Scale Layer and Decorrelated Global Centralized Ranking Loss, which achieves extremely efficient and discriminative learning, *i.e.*, 5× speedup over triplet loss and 12% recall boost on *CARS196*. Our method originates from the classic softmax loss, which has a global structure but does not directly optimize the distance metric as well as the inter/intra class distance. We tackle this issue through a hypersphere layer and a global centralized ranking loss with a pairwise decorrelated learning. In particular, we first propose a Normalize-Scale Layer to eliminate the gap between metric distance (for measuring distance in retrieval) and dot product (for dimension reduction in classification). Second, the relationship between features is encoded under a global centralized ranking loss, which targets at optimizing metric distance globally and accelerating learning procedure. Finally, the centers are further decorrelated by Gram-Schmidt process, leading to extreme efficiency (with 20 epochs in training procedure) and discriminability in feature learning. We have conducted quantitative evaluations on two fine-grained retrieval benchmark. The superior performance demonstrates the merits of the proposed approach over the state-of-the-arts.

Introduction

Fine-grained image retrieval (FGIR) is to search image through subordinate in the same visual category, *e.g.*, birds (Wah et al. 2011), cars (Krause et al. 2013) and products (Oh Song et al. 2016), which has attracted increasing research focus recently. In such a setting, instances are similar to each other within a general class and are difficult to distinguish among fine-grained categories, due to various pose, illumination and occlusion. Therefore, FGIR is more challenging comparing with content based image retrieval (CBIR). In order to distinguish the subtle differences among fine-grained categories, a discriminative feature representation is demanded. To this end, a recent trend is to adopt convolutional neural network (CNN) with a distance metric learning (Huang, Loy, and Tang 2016; Oh Song et al. 2016; Bell and Bala 2015; Ustinova and Lempitsky 2016; Wang et al. 2014) to extract the discriminative and generative features, which aim to distinguish high-dimensional features within/outside fine-grained categories.

Despite the recent progress, fine-grained image retrieval remains as an open problem. The key challenge lies in designing a good loss to efficiently learn a robust metric, which measures similarity of features extracted from deep neural networks (Krizhevsky, Sutskever, and Hinton 2012; Szegedy et al. 2015; Simonyan and Zisserman 2014; He et al. 2016). However, the existing loss functions in FGIR are still far from satisfactory, which are defected in terms of *local structure* and *slow training*. The former refers to encoding the example relationship locally, *i.e.*, pairwise, triplet, which results in low accuracy. The latter refers to establishing the loss with mean square error (MSE), which causes low efficiency, since the MSE will theoretically get stuck in local optimum (Golik, Doetsch, and Ney 2013). More specifically, most existing loss functions in FGIR are defined in terms of pairs (Hadsell, Chopra, and LeCun 2006) or triplets (Wang et al. 2014) inside the training mini-batch. This contradicts from the optimal loss that should maximize intra-class compactness and inter-class separability. And since the existing loss in FGIR did not consider the embedding space globally, the resulting weak features will degenerate the retrieval performance. Besides, the computation complexity for pairwise and triplet loss can go up to $O(N^2)$, where N denotes the total number of training samples.

The local Structure Challenge. Learning with global structure has been well addressed in image categorization/recognition/segmentation (He et al. 2016; Ren et al. 2015; Long, Shelhamer, and Darrell 2015), which typically follows a cross-entropy + softmax loss setting *i.e.* the weight of last fully-connected layer can be considered as a global class center. Such a global optimization mechanism cannot be directly implemented in FGIR, as its inner-product opera-

[*]corresponding author

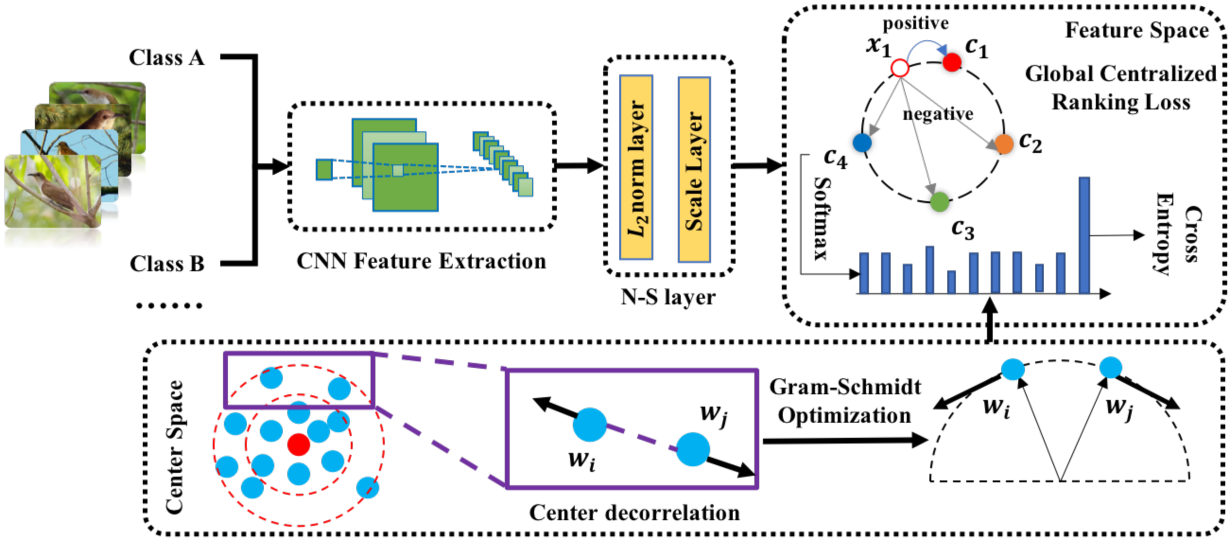

Figure 1: An illustration of the proposed framework. The Norm-Scale (N-S) layer projects the features of images on a hypersphere by using L_2 normalization and scale layer. The proposed Decorrelated Global-aware Centralized Ranking Loss (DGCRL) consists of two parts, a global aware ranking loss in the feature space and a Gram-Schmidt independent optimization in the center space. The framework can directly optimize the intra-class compactness and inter-class separability and establish an effective representaiton for fine-grained retrieval.

tion is theoretically not a distance metric, which is unable to enforce feature discriminability. To further explain, features in the same class could lie on different hypersphere therefore apart from each other, while being closer in hard examples of different classes.

The Slow Training Challenge. The main reason for the unsatisfactory training speed lies in the disparity between training/testing and the usage of MSE. In particular, since the dataset contains images with different qualities, the corresponding features tend to lie on different spheres (Ranjan, Castillo, and Chellappa 2017). It is therefore makes it difficult to minimize the inter-class distance in training and robustly measure similarity between features without L_2 normalization in testing, which further leads to a gap between training and testing, *i.e.* the feature is optimized *without* L_2 *normalization* in training while the similarity is measured *after normalization* in testing. More importantly, according to the (Golik, Doetsch, and Ney 2013), MSE will quickly stuck in local minima, which further degenerate the training efficiency.

In this paper, we present a novel metric learning scheme that conquers the above difficulties in a unified framework. As illustrated in Fig.1, the framework contains two crucial components, *i.e.* **Normalize-Scale Layer** and **Decorrelated Global-aware Centralized Ranking Loss**, the former of which eliminates the gap between training and testing as well as inner-product and the Euclidean distance, while the latter encourages learning embedding function to directly optimize inter-class compactness and intra-class separability. In particular, the Decorrelated Global-aware Centralized Ranking Loss (DGCRL) is composed by two parts, a global-aware ranking loss in feature space and a Gram-Schmidt independent operation on center space, which tackles the challenge of feature discriminability and generalization respectively. The proposed normalize-scale(N-S) layer employs normalization and scaling operation in image features, which connects inner-product and Euclidean distance metric to well accelerate DGCRL through softmax loss. It is worth to note that, most methods need to train their networks with cropped images, hard example mining, or extra information. In contrast, no extra information, models or datasets are used in our network.

We have conducted extensive experiments on two widely-used FGIR benchmark *CUB-200-2011* and *CARS196*, with comparisons to a set of state-of-the-art methods. Quantitatively, it outperforms CRL-WSL (Zheng et al. 2018) by 12.0% in *CAR196*, and runs 5× faster in training over triplet loss.

Related Work

Fine-Grained Image Retrieval (FGIR). FGIR has attracted increasing research focus in recent years (Wei et al. 2017; Xie et al. 2015; Zhang et al. 2016; Zheng et al. 2018). It aims to differentiate subordinate classes, where the challenges are two-fold: 1) Most classes are highly correlated and difficult to be distinguished due to their subtle difference *i.e.*, small inter-class variance. 2) The intra-class variance is large, which is caused by different poses and viewpoints. Existing works in FGIR can be categorized into two groups: The first group relies on using handcraft features (Xie et al. 2015), while the second defines FGIR as a deep metric learning problem, *e.g.* (Zhang et al. 2016; Zheng et al. 2018), which attempts to learn a discriminative feature by designing specific loss functions for training deep neural network. However, as discussed above, these methods encode the *local* relationship between features, which degenerate the feature discriminability and learning efficiency, while leaving the problem of variations between training and testing unsolved.

Deep Metric Learning. The goal of deep metric learning is to learn an optimal metric to minimize the distance between similar images. Similarities can be encoded as pairwise loss (Bell and Bala 2015) or triplet loss (Wang et al. 2014). The work in Oh Song et al. (2016) proposed a high-order similarity constraint by lifting the pairwise distance within a mini-batch in the form of a dense matrix. Beyond Euclidean metric, (Huang, Loy, and Tang 2016) introduced a Position-Dependent Deep Metric (PDDM) unit, which is capable of learning a similarity metric that is adaptive to local structure in the feature space. However, the performance of these methods is still unsatisfactory. The tremendous search space, the limitation of local-aware structure, as well as the time-consuming mean square optimization lead to inefficient training and less discriminative feature representation. In contrast, the proposed DGCRL updates parameters using the global centers, which makes the optimization more sustainable and more efficient *i.e.*, 12.0% over *CRL-WSL* (Zheng et al. 2018) in *CARS196*, 5× speed up in training over triplet loss.

The Proposed Method

As shown in Fig.1, the proposed model innovates in two aspects, *i.e.*, Normalize-Scale (N-S) Layer and Decorrelated Global-aware Centralize Ranking Loss (DGCRL). The N-S Layer projects the features onto a hypersphere by an L_2 normalization and a scale layer. Such an N-S layer is a precondition of the rest components, i.e., the proposed DGCRL can be transformed iff. the features and centres are normalized. On the other hand, the proposed DGCRL is composed of two parts which directly optimizes the intra-class compactness (by Gram-Schmidt decorrelated operation in center space) and inter-class separability (by Global Centralized Ranking Loss (GCRL) in feature space) in a global way. This differs from previous methods which tend to focus on the local structure of the embedding space using pair-wise/triplet loss, and hence improves the feature discriminability and learning efficiency. Moreover, by combining the proposed N-S layer with DGCRL, the proposed model can be trained by using cross-entropy loss, as the Euclidean distance can be transformed into inner-product.

Problem Definition and Overall Pipeline

Let $\mathbb{I} = \{I_1, ..., I_n\} \in \mathcal{I}$ be a labeled collection of fine-grained images, where $\mathcal{I}$ is the original input space of all possible images. Depending on the application, in training set each image has its corresponding label $y \in \{1, ..., K\}$ of K classes/clusters. Function $f(\cdot;\theta) : \mathcal{I} \to \mathbb{R}^d$ maps an image I to a vector $x_i = f(I_i, \theta)$ in a d-dimensional embedding space $\mathcal{X}$. θ is the CNN parameters to be learned.

We first summarize the general pipeline for fine-grained image retrieval (Wei et al. 2017; Xie et al. 2015; Zhang et al. 2016; Zheng et al. 2018). Given a training set with fine-grained images and corresponding labels, CNN is trained *without feature normalization*. At the testing stage, feature descriptor x_i is extracted from image I_i and *normalized* to a unit length. Then, the returning list is sorted according to similarity computed by using the Euclidean distance. Note that normalization is unavoidable in online stage. Since the features lie on different spheres, retrieving images with normalization can largely promote the performance (Ranjan, Castillo, and Chellappa 2017).

There are two issues in this pipeline. First, the training and testing phases for fine-grained image retrieval are decoupled. There is no guarantee that the training step directly optimizes the ranking function in the Euclidean space, *i.e* some retrieval tasks follow an inner-product + softmax loss pipeline, which differ from the loss design in training. Second, existing deep metric learning is unable to extract discriminative features on different hyperspheres. Some previous works (Wang et al. 2017; Ranjan, Castillo, and Chellappa 2017) proposed feature normalization methods to solve these issues. Nevertheless, as illustrated in Fig.3, normalizing features with unit length can lead to a "degeneration" problem, *i.e.* the loss cannot converge due to the tiny space on unit sphere. Moreover, inner-product with feature normalization is not a Euclidean distance, which means the gap between training and testing steps still exists. We solve the above issues by proposing a novel Normalize-Scale (N-S) Layer, which eliminates the gap between inner-product and the Euclidean distance by normalization and expand representation space by scale operation .

Normalize-Scale Layer

The N-S layer forces the features to be distributed on a hypersphere. For an input vector x, the layer can be defined as follows:

$$\hat{x} = \frac{x}{\|x\|_2} * \alpha = \frac{\alpha x}{\sqrt{\sum_i x_i^2}}. \tag{1}$$

Here x can be either a feature vector or a center. The layer can be easily implemented by publicly available deep learning frameworks such as MXNet (Chen et al. 2015) or Caffe (Jia et al. 2014). The N-S layer disposes the feature with normalization and a scale layer, as illustrated in Fig.1. It is worth to note that the center should not be projected on hyperspheres, since the centers will gradually close to the features hypersphere through back propagation and the "degeneration" problem exists only in the original feature space. Furthermore, the distance calculation can be transformed into inner-product form (which will be described in the next section). Therefore, employing project operation in feature space makes numerical calculations easier.

Decorrelated Global Centralized Ranking Loss

Due to the tremendous search space and local aware structure, previous deep metric learning methods (Huang, Loy, and Tang 2016; Oh Song et al. 2016; Bell and Bala 2015; Ustinova and Lempitsky 2016; Wang et al. 2014; Zheng et al. 2018) are less effective in training, which fail to learn a discriminative feature representation for fine-grained tasks. The major reason is that, the local structure is insufficient to learn a discriminative feature representation. Moreover, directly minimizing the Euclidean distance leads to a very slow convergence. As a result, it is reasonable to backpropagate through a global learnable center. Therefore, combining with N-S layer, the proposed GCRL scheme directly optimizes the following loss:

$$\mathcal{L} = \max\left(\sum_{i=1}^{N}(m + \|x_i - w_{y_i}\| - \frac{1}{K-1}\sum_{c\neq j}^{K}\|x_i - w_j\|)\right) \tag{2}$$
$$s.t. \quad \|x_i\|_2 = \alpha.$$

For a given feature x_i, the loss function consists of three parts, the margin m, the distance between corresponding centre w_{y_i} and the distance between negative pairs. The gradients are defined as:

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{x_i - w_{y_i}}{\|x_i - w_c\|} - \frac{1}{K-1}\sum_{c\neq j}^{K}\frac{x_i - w_j}{\|x_i - w_j\|}, \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial w_{y_i}} = \frac{w_{y_i} - x_i}{\|w_{y_i} - x_i\|}, \frac{\partial \mathcal{L}}{\partial w_j} = \frac{1}{K-1}\frac{x_i - w_j}{\|x_i - w_j\|}. \tag{4}$$

As shown in Eq.3, GCRL forces the feature x_i to approach the center of the target class, while leaving away from centers of irrelevant classes. At the same time, in Eq.4, the center is updated by the information generated from x_i, which gradually approaches the global center, leading to a more compact and separable feature representation in the testing phase. However, due to the mean square loss, the proposed loss function also suffers from the training efficiency problem. Considering that FGIR often handles large-scale training sets, a modification to accelerate GCRL is indispensable.

Accelerating GCRL. The work in (Golik, Doetsch, and Ney 2013) argued that the cross entropy criterion allows to find better local optimum and convergence faster than mean square loss. Together with softmax, it has become the most commonly adopted loss function in various tasks, including retrieval (Wen et al. 2016). However, in these methods, the loss function is only used as an auxiliary function, which is insufficient in fine-grained image retrieval. Combining with N-S layer, we reformulate the global-aware centralized loss to an equivalent softmax loss, which retains the global structure, distance metric and large margin in GCRL, while engaging the merit of softmax loss.

Softmax loss has been recognized as an essential component in the classification. For an input feature x_i with its corresponding center/label/cluster y_i, it can be formulated as:

$$\mathcal{L}_{softmax} = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right), \tag{5}$$

where f_j is the j-th element of the class vector f. In most CNN structure, f is usually the output of a fully-connected layer W. So f_j can be written as $f_j = W_j^T x_i$ in which W_j is the j-th column of W. We define the probability for the y_i-th class as $p_{y_i} = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}}$ for simplicity. And in back propagation, the gradient w.r.t. f_j can be obtained by the chain-rule. If $j = y_i$, we have:

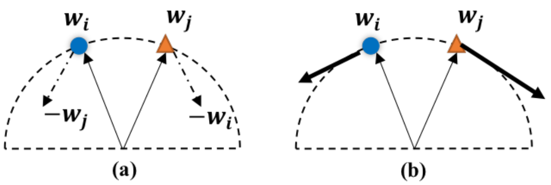

Figure 2: An illustration of two different optimization methods. (a) Directly minimizing $w_i w_j$ is ineffective (the gradient will be damped by N-S Layer) and unstable (hard to minimize when $w_i w_j$ is close to zero). (b) A more stable and effective gradient conducted by Gram-Schmidt process.

$$\begin{aligned}\frac{\partial \mathcal{L}_{softmax}}{f_j} &= -\frac{1}{p_{y_i}}\frac{\partial p_{y_i}}{f_j} \\ &= -\frac{1}{p_{y_i}}\frac{e^{f_j}\sum_i e^i - e^{f_j}\cdot e^{f_j}}{\left(\sum_i e^i\right)^2} \\ &= -(1 - p_j),\end{aligned} \tag{6}$$

and when $j \neq y_i$, the gradient becomes

$$\begin{aligned}\frac{\partial \mathcal{L}_{softmax}}{f_j} &= -\frac{1}{p_{y_i}}\frac{\partial p_{y_i}}{f_j} \\ &= -\frac{1}{p_{y_i}}\frac{0\cdot\sum_i e^i - e^{f_j}\cdot e^{f_{y_i}}}{\left(\sum_i e^i\right)^2} \\ &= p_j.\end{aligned} \tag{7}$$

As shown in Eq.6 and Eq.7, in the training step, if we consider the inner-product as a similarity metric, softmax loss enhances the similarity value when the feature x_i belongs to class y_i, and depresses the similarity with other classes simultaneously[1]. It is worth to note that, the value of L_2 norm with features and centres are invariable[2] in training, which means the Euclidean distance is inversely proportional to the inner product, *i.e.*

$$\|x_i - w_j\| = \|x_i\| + \|w_j\| - 2x_i w_j = C - 2x_i w_j. \tag{8}$$

Therefore, we reformulate the proposed GCRL to softmax except for the margin parameter m. For the softmax form, it can be easily implemented by most deep learning frameworks. For the margin parameter m, it is still unable to optimize the intra-class compactness and inter-class separability.

Center Decorrelation. The work in (Liu et al. 2016; 2017) introduced an angular margin to make the objective function more rigorous, and used the intermediate value $\cos(m\theta)$ to replace the original $\cos\theta$ during training. However, m is supposed to be positive to make $\cos(m\theta)$ derivable, which largely compress the representation space. Besides, the function is hard to implement due to the angular

[1] The gradient represent the negative direction of the optimization

[2] the features are normalized and the gradient direction of centers are vertical with centers (which will be explained in next subsection)

margin involved. To address these issues, we design a new Center Decorrelation operation to enlarge the distance between different centers. We first give a simple description to our intuition. Considering Eq.7 and Eq.6, the gradient w.r.t. x_i can be obtained by the chain rule:

$$\frac{\partial f_j}{\partial x_i} = \frac{\partial w_j x_i}{\partial x_i} = w_j, \tag{9}$$

which indicates that in training, the features are moving according to the center w_j, eventually point to the same direction with w_j. Therefore, the margin can also be obtained by decorrelating centers with the following objective function:

$$\begin{aligned} \mathcal{L}_i &= -\log \frac{e^{w_{y_i} x_i}}{e^{w_{y_i} x_i} + \sum_{j \neq y_i} e^{w_j x_i}} \\ s.t. &\quad \|x_i\|_2 = \alpha, W^T W = I. \end{aligned} \tag{10}$$

The constraint $W^T W = I$ requires the centers to be orthogonal to each other, which is an implicit version of the requirement that the centers should be pairwise decorrelated. Since the normalize value of center w is not important in our formulation, to further simply the computation, we use the Lagrange Multiplier with constraint $W^T W = I$ and the final DGCRL can be obtained as:

$$\begin{aligned} \mathcal{L}_i &= -\log \frac{e^{w_{y_i} x_i}}{e^{w_{y_i} x_i} + \sum_{j \neq y_i} e^{w_j x_i}} - \frac{\lambda}{|\Omega|} \sum_{i \neq j} |w_i w_j^T| \\ s.t. &\quad \|x_i\|_2 = \alpha, \end{aligned} \tag{11}$$

where Ω denotes different pairwise sets of centers. With the absolute operation, Eq.11 requires the centers to be perpendicular, which promotes the feature discriminability (large center distance) and generalization (wide feature representation space) simultaneously. However, there are optimization issues in Eq.11. As illustrated in Fig.2, directly minimizing $|w_i w_j|$ is ineffective (the gradient will be damped by the feature normalization) and unstable (hard to minimize when the $w_i w_j$ is close to zero).

Gram-Schmidt Optimization In this paper, we employ the Gram-Schmidt process to solve the above problems. Let $\{u_i | i = 1, 2, ..., n\}$ be a set of m-vectors and we wish to obtain an equivalent orthonormal set $\{v_i | i = 1, 2, ..., n\}$ of m-vectors. The Gram-Schmidt process (Schmidt 1908) can be constructed by following successive operation:

$$v_k^{'} = u_k - \sum_{j=1}^{k-1} \langle v_j, u_k \rangle v_j, v_k = \frac{v_k^{'}}{\|v_k^{'}\|}; k = 2, ..., n, \tag{12}$$

where $v_1 = \frac{u_1}{\|u_1\|}$. The Gram-schmidt is used for two reasons. One is that the process can orthogonalise the original gradient to be vertical with corresponding centre, which makes the optimization more effective, as illustrated in Fig.2b. The other one is that the operation is stable when two centers are perpendicular, due to the $\langle v_j, u_k \rangle$ operation (the second term becomes zero). Moreover, when dealing with two vectors, Eq.12 can be easily implemented as:

$$\frac{\partial \mathcal{L}}{w_i} = \frac{\partial \mathcal{L}}{w_i} + \frac{1}{|\Omega|} \lambda (w_i - \langle w_i, w_j \rangle) w_j, \tag{13}$$

where the computation of $\frac{\partial \mathcal{L}}{w_i}$ is the same as the original softmax. Obviously, in Eq.13, the update is co-determined by the inter-class compactness $\frac{\partial \mathcal{L}}{w_i}$ and intra-class separation $(w_i - \langle w_i, w_j \rangle) w_j$. This means that the discriminability and generalization will be explicitly enhanced. The overall framework is summarized in Alg. 1.

Algorithm 1: Decorrelated Global Centralized Ranking Loss

```
Input: Training data: $\mathcal{D}_t$; CNN model: $\mathcal{F}$.
Output: Trained CNN model: $\mathcal{F}$.
for $t$=1,...,$T$ epoch do
    Forward image to feature layer;
    Pass the feature through N-S layer;
    Calculate the loss by softmax and cross-entropy;
    Get the gradient of softmax loss;
    Gram-Schmidt optimize by Eq.13;
    Update CNN model $\mathcal{F}$ by $t^t h$ epoch data ;
end
```

Experiments

Experimental Settings

Datasets. *CUB-200-2011* (Wah et al. 2011) contains 200 bird classes with 11,788 images. We use the first 100 classes in training and the rest in testing. The split in *CARS196* (Krause et al. 2013) is also similar to *CUB200-2011*, which contains 196 car classes with 16,185 images, *i.e.* with the first 98 classes (8,045 images) for training and the remaining classes (8,131 images) for testing. The training/testing split also follows the standard setting in (Huang, Loy, and Tang 2016; Bell and Bala 2015; Oh Song et al. 2016).

Evaluation Protocols. We evaluate the retrieval performance by *Recall@K*. *Recall@K* is the average recall scores over all query images in the test set, which strictly follows the setting in (Oh Song et al. 2016). Specifically, for each query, we return the top K similar images. In the top K returning images, the score will be 1 if there exists at least one positive image, and 0 otherwise.

Implementation Details. We apply the widely-used Resnet-50 (He et al. 2016) in our experiments. The network is pre-trained on ImageNet ILSVRC-2012 (Deng et al. 2009)[3]. We set the same hyperparameters in all experiments without specific tuning, with a margin parameter m of 4, a scale parameters α of 128, a mini-batch size of 60, and an initial learning rate starts from 0.0001 and is divided by 10 in every 100 epochs. The feature is extracted from the last convolutional layer of Resnet-50 with max and average pooling, which is followed through SCDA (Wei et al. 2017).

Baselines We compare the proposed method with the following state-of-the-art methods: (1) *Contrastive*, *Triplet*, *PDDM+Quadruplet* and *liftedStruct* (Bell and Bala 2015; Zhang et al. 2016; Oh Song et al. 2016; Huang, Loy, and

[3] Note that the proposed method is compatible with other networks, the choice of which is orthogonal to our contribution

Method	CARS196						CUB-200-2011					
K =	1	2	4	8	16	32	1	2	4	8	16	32
Contrastive	21.7	32.3	46.1	58.9	72.2	83.4	26.4	37.7	49.8	62.3	76.4	85.3
Triplet	39.1	50.4	63.3	74.5	84.1	89.8	36.1	48.6	59.3	70.0	80.2	88.4
LiftedStruct	49.0	60.3	72.1	81.5	89.2	92.8	47.2	58.9	70.2	80.2	89.3	93.2
Facility Location	58.1	70.6	80.3	87.8	-	-	48.2	61.4	71.8	81.9	-	-
N-pairs	53.9	66.76	77.75	86.35	-	-	45.37	58.41	69.51	79.49	-	-
Binomial Deviance	-	-	-	-	-	-	52.8	64.4	74.7	83.9	90.4	94.3
Histogram Loss	-	-	-	-	-	-	50.3	61.9	72.6	82.4	88.8	93.7
PDDM+Quadruplet	57.4	68.6	80.1	89.4	92.3	94.9	58.3	69.2	79.0	88.4	93.1	95.7
SCDA	58.5	69.8	79.1	86.2	91.8	95.9	62.2	74.2	83.2	90.1	94.3	**97.3**
CRL-WSL	63.9	73.7	82.1	89.2	93.7	96.8	65.9	76.5	85.3	90.3	94.4	97.0
Our Method	**75.9**	**83.9**	**89.7**	**94.0**	**96.6**	**98.0**	**67.9**	**79.1**	**86.2**	**91.8**	**94.8**	97.1

Table 1: *Recall@K* on *CARS196* and *CUB-200-2011* with baseline methods. *Recall@K* is the average recall scores over all query images in the testing set. Specifically, for each query image, top K nearest images will be returned. The recall score will be 1 if there is at least one positive image in the returning K images, and 0 otherwise.

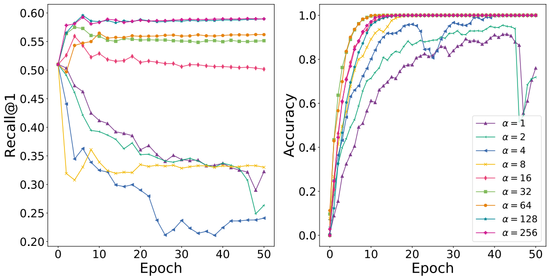

Figure 3: Recall@K in testing (left) and classification accuracy in training (right) with different α on *CUB200-2011*.

Margin λ	recall@K					
	K=1	2	4	8	16	32
$\lambda = 0$	66.7	76.7	85.0	90.4	93.8	96.4
0.1	**67.9**	**79.1**	**86.2**	**91.8**	**94.8**	**97.1**
0.2	67.0	77.1	84.9	90.5	93.9	96.4
0.4	65.9	76.3	84.5	90.2	93.8	96.5

Table 2: Recall@K with different λ on *CUB-200-2011*. The λ is the weight of Gram-Schmidt optimization in DGCRL.

Method	epoch	Recall@K					
		k=1	2	4	8	16	32
triplet	100	64.4	75.5	84.2	90.3	94.5	06.9
lifted	56	60.2	72.9	82.4	89.5	94.4	97.0
CRL	96	65.8	76.8	85.0	90.6	94.5	**97.2**
DGCRL	**20**	**67.9**	**79.1**	**86.2**	**91.8**	**94.8**	97.1

Table 3: Recall@K and corresponding training epoch with different loss function on *CUB-200-2011*.

Tang 2016). These methods aim at training CNN using local aware metric learning with different numbers of examples (pairwise, triplet and quadruplet). (2) *Facility Location* (Song et al. 2017) introduces a new metric learning method which is able to learn a global structure by facility location. (3) *Histogram Loss* and *Binomial Deviance* (Ustinova and Lempitsky 2016) proposed to evaluate the cost of overlap between distributions of positive pairs' distances and negative pairs', which is robust to outliers. (4) *SCDA* (Wei et al. 2017) employs the network saliency to generate a discriminative and representative feature. (5) *CRL-WSL* (Zheng et al. 2018) combines a centralized ranking loss with weakly supervised localization method to obtain features under pixel level object localization, which only uses image-level labels.

Fine-grained Image Retrieval

As shown in Tab.1, our model consistently outperforms the state-of-the-arts in terms of *Recall@K*, which achieves 67.9 *Recall@1* on *CUB200-2011* and 75.9 *Recall@1* on *CARS196*. It is worth to note that only our method is optimized to capture the global structure, which confirms the point we argued before. Fig.4 visualizes the t-SNE (Van Der Maaten 2014) plots on the embedding vectors of our method on *CUB200-2011* and *CARS196*, respectively. We can see that our method performs very well on grouping similar objects despite the variations in view point, pose and configuration.

Ablation Study: Scale α and Factor λ

We further explore the retrieval performance in different hyperparameter settings of α and m, as well as how they affect the FGIR performance in *CUB200-2011*. Fig.3 shows the performance variations in different settings of parameter α. The performance is poor when α is small and is stable when α is higher than 16. As mentioned before, we observe the "degeneration" problem in Fig.3, *i.e.*, the inter-class compactness and intra-class separation can be easily obtained in training set rather than disjoint testing set. In fact, α decides the size of the feature representation space. Therefore, the overfit can be alleviated with a higher α. On the other hand, the performance is consistent when α is lager than 128, which mainly caused by numerical constraint of the input (the range of input image is $[-128, +128]$). Tab.2 further shows the tuning of the hyperparameter λ, we have found that $\lambda = 0.1$ is the optimal one. We also observe that with

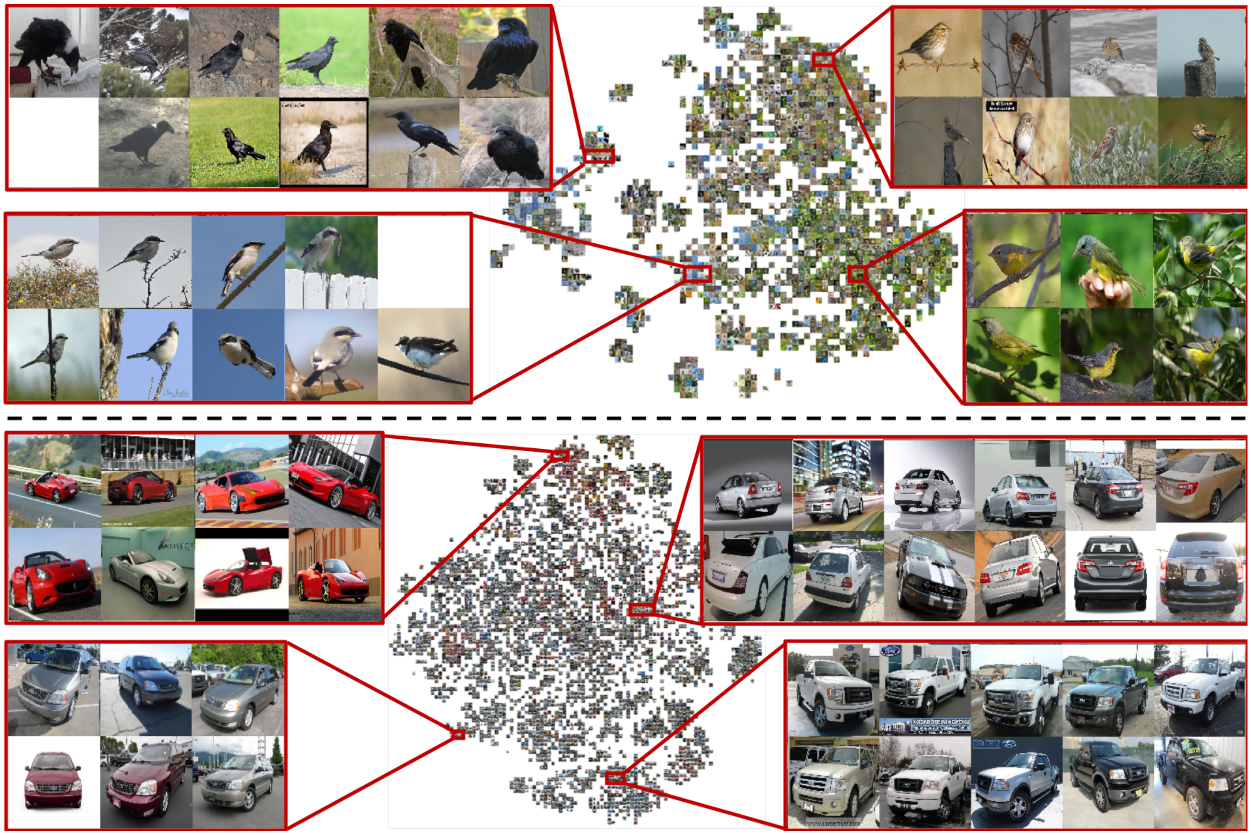

Figure 4: Visulization of the proposed method with t-SNE on *CUB200-2011* (up) and *CARS196* (down). Best viewed when zoomed in. Based on our approach, images with similar objects are more likely to be grouped together despite the variations in view point, pose and configuration.

a large λ the results would decrease, due to the too strict Gram-Schmidt condition.

On Different Loss Functions

To evaluate the effectiveness of the proposed global centralized ranking loss, we further replace our GCRL with different loss functions and quantify the retrieval degeneration by *Recall@K* on *CUB200-2011*. As shown in Tab.3, our method is the best among different loss functions under the same setting of the rest components. Please note that, comparing with the same methods in Tab.1, the performance is higher in Tab .3, which demonstrates the effectiveness of the proposed N-S layer. Besides, in Tab.3, we further report the training epochs with respect to different loss functions. The MSE loss functions (CRL, triplet, liftedstruct) are time consuming. Instead, loss function with cross-entropy is extremely effective in training, particular the our proposed DGCRL (only 20 epochs in the training phase).

Conclusion

In this paper, we solve two crucial issues in FGIR, termed *local structure* and *slow training*, the former of which is caused by using the relationship only in a mini batch, while the latter refers to the usage of the mean square error. To address the local structure issue, we propose a global centralized ranking loss to learning the feature in a *global* way, which can be further enhanced by a Gram-Schmidt decorrelate optimization. To address the slow training issue, we propose a normalize-scale layer to eliminate the gap between inner-product and the Euclidean distance, hence, the loss function can be further accelerated by using softmax loss. We achieve the best retrieval performance on the widely-used *CUB200-2011* and *CARS196* benchmarks. Quantitatively, it gains 12.0% over *CRL-WSL* (Zheng et al. 2018) in *CARS196*, as well as 5× faster in training over triplet loss.

Acknowledgements

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

Bell, S., and Bala, K. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)* 34(4):98.

Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; and Zhang, Z. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. Ieee.

Golik, P.; Doetsch, P.; and Ney, H. 2013. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Interspeech*, volume 13, 1756–1760.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *null*, 1735–1742. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, C.; Loy, C. C.; and Tang, X. 2016. Local similarity-aware deep feature embedding. In *Advances in Neural Information Processing Systems*, 1262–1270.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678. ACM.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-margin softmax loss for convolutional neural networks. In *ICML*, 507–516.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 1.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4004–4012.

Ranjan, R.; Castillo, C. D.; and Chellappa, R. 2017. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Schmidt, E. 1908. Über die auflösung linearer gleichungen mit unendlich vielen unbekannten. *Rendiconti del Circolo Matematico di Palermo (1884-1940)* 25(1):53–77.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, H. O.; Jegelka, S.; Rathod, V.; and Murphy, K. 2017. Deep metric learning via facility location. In *Computer Vision and Pattern Recognition (CVPR)*, volume 8.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Ustinova, E., and Lempitsky, V. 2016. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, 4170–4178.

Van Der Maaten, L. 2014. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research* 15(1):3221–3245.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1386–1393.

Wang, F.; Xiang, X.; Cheng, J.; and Yuille, A. L. 2017. Normface: l 2 hypersphere embedding for face verification. In *Proceedings of the 2017 ACM on Multimedia Conference*, 1041–1049. ACM.

Wei, X.-S.; Luo, J.-H.; Wu, J.; and Zhou, Z.-H. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing* 26(6):2868–2881.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 499–515. Springer.

Xie, L.; Wang, J.; Zhang, B.; and Tian, Q. 2015. Fine-grained image search. *IEEE Transactions on Multimedia* 17(5):636–647.

Zhang, X.; Zhou, F.; Lin, Y.; and Zhang, S. 2016. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1114–1123.

Zheng, X.; Ji, R.; Sun, X.; Wu, Y.; Huang, F.; and Yang, Y. 2018. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval. In *IJCAI*, 1226–1233.