The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Deep Metric Learning with Graph Consistency

Binghui Chen,[1] **Pengyu Li,** [1] **Zhaoyi Yan** [1,2] **Biao Wang** [1] **Lei Zhang** [1,3]
[1] Artificial Intelligence Center, DAMO Academy, Alibaba Group
[2] Harbin Institute of Technology
[3] The Hong Kong Polytechnic University
chenbinghui@bupt.edu.cn, lipengyu007@gmail.com, yanzhaoyi@outlook.com, wangbiao225@foxmail.com,
cslzhang@comp.polyu.edu.hk

Abstract

Deep Metric Learning (DML) has been more attractive and widely applied in many computer vision tasks, in which a discriminative embedding is requested such that the image features belonging to the same class are gathered together and the ones belonging to different classes are pushed apart. Most existing works insist to learn this discriminative embedding by either devising powerful pair-based loss functions or hard-sample mining strategies. However, in this paper, we start from another perspective and propose Deep Consistent Graph Metric Learning (CGML) framework to enhance the discrimination of the learned embedding. It is mainly achieved by rethinking the conventional distance constraints as a graph regularization and then introducing a Graph Consistency regularization term, which intends to optimize the feature distribution from a global graph perspective. Inspired by the characteristic of our defined 'Discriminative Graph', which regards DML from another novel perspective, the Graph Consistency regularization term encourages the sub-graphs randomly sampled from the training set to be consistent. We show that our CGML indeed serves as an efficient technique for learning towards discriminative embedding and is applicable to various popular metric objectives, e.g. Triplet, N-Pair and Binomial losses. This paper empirically and experimentally demonstrates the effectiveness of our graph regularization idea, achieving competitive results on the popular CUB, CARS, Stanford Online Products and In-Shop datasets.

Introduction

In the context of end-to-end feature learning framework of deep convolutional neural network where the convolutional neural network actually is a powerful non-linear mapping function and can be arbitrarily modeled by loss functions to some extend, Deep Metric Learning (DML) focuses on the design of discriminative objective loss function, so as to constrain the learned embedding to be more discriminative. By reason of the powerful representation ability of the learned embedding, Deep Metric Learning (DML) has been widely explored and applied in many computer vision tasks, such as image retrieval (Gordo et al. 2017; Noh et al. 2017), face recognition (Schroff, Kalenichenko, and Philbin 2015; Wen et al. 2016), person re-identification (Hermans, Beyer, and Leibe 2017; Chen et al. 2017), zero-shot learning (Oh Song et al. 2016), visual tracking (Leal-Taixé, Canton-Ferrer, and Schindler 2016; Tao, Gavves, and Smeulders 2016) and cross-modal retrieval (Deng et al. 2018).

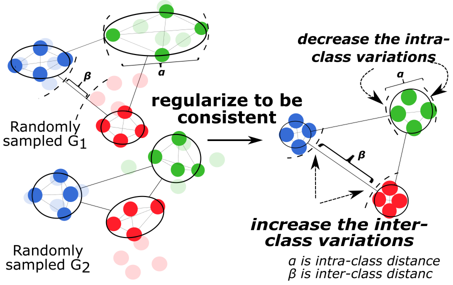
decrease the intra-class variations
α
β
α
regularize to be consistent
Randomly sampled G_1
β
Randomly sampled G_2
increase the inter-class variations
α is intra-class distance
β is inter-class distanc

Figure 1: Graph Consistency. The circles indicate the data points, green/red/blue color represent three classes, *resp*. For the left two sub-figures, the highlighted colors means the current sampled data. From the left two sub-figures, one can observe that since the overall feature representations are not discriminative enough, i.e. with large intra-class variations and relatively small inter-class margins, the randomly sampled graphs are different with each other. Then by regularizing them to be consistent, in other words, aligning them to be the same, the intra-class distance can be decreased and the inter-class margin can be enlarged to some extend, resulting in the more discriminative representation space than before.

Deep Metric Learning (DML) is generally achieved by learning feature representations for the input images such that the instances from the same class are mapped to the small vicinity in the low-dimensional representation space while the samples from different classes are placed relatively apart. The representations are learned under an end-to-end optimization framework where the objective function utilizes the loss terms to impose the desired intra-class and inter-class distance constraints in the feature space. Thus, in order to obtain the discriminative feature representations, most of the DML works dedicate to mining the expressive instance pairs as many as possible. For example, many research works focus on exploring the tuple-based loss functions. such as contrastive loss (Sun et al. 2014), binomial deviance loss (Yi et al. 2014), triplet loss (Schroff, Kalenichenko, and Philbin 2015) and quadruplet loss (Chen et al. 2017).

However, in these tuple-based methods, training instances are grouped into pairs, triplets or quadruplets, resulting in a quadric or cubical growth of training pairs which are of high probability to be highly redundant and less informative. It gives rise to some key problems for tuple-based approaches, in which (1) the actually constructed pairs are finite and local such that they cannot utilize the global and informative data structure, thus the optimized image representations will not be discriminative enough, and (2) the optimization of feature representation is dominated by the margin constraints, in which case if the sampled pairs satisfy the margin constraints, the losses will become zero and the parameter update will be stopped, thus the actual global feature distributions might be still not discriminative, leading to inferior performances. Then, to learn compact and separable features, some researchers try to seek help from the technique of hard samples mining, such as (Wu et al. 2017; Harwood et al. 2017; Schroff, Kalenichenko, and Philbin 2015), however, in practice, the model training is usually very sensitive to the sampling strategy and sampled pairs, resulting in bad local minimum and large variations in performances. Moreover, some researchers propose to use global instance-relations for discriminative embedding learning, such as Lifted(Oh Song et al. 2016), N-Pair(Sohn 2016) and MS(Wang et al. 2019a), while due to the Maximum-Domination problem[1] behind SoftMax formulation, the global constraints from these methods are not enough. To this end, proposing more discriminative and efficient deep metric objective function remains important.

Considering the aforementioned problems, in this paper, we propose the deep ***Consistent Graph Metric Learning*** (***CGML***) framework, a novel loss constraint, to further enhance the discriminative leanring by regularizing the randomly sampled graphs to be consistent during each training iteration. It is mainly achieved by introducing a *Graph Consistency* (GC) regularization term that is ‘plug and play’ and can be generally applied to many existing deep metric learning methods. Specifically, at each iteration, we first randomly select m classes with $\frac{n}{m}$ instances each class for two times, then regard the instances as nodes and construct two graphs according to instance-to-instance distances respectively. Restraining these two randomly-sampled graphs to be consistent is to satisfy the property of discriminative representation distribution where compact intra-class distributions and separable inter-class distances exist. As illustrated in Fig.1, at the beginning, the data representations are not discriminative, and the sampled graphs have large diversities, after performing the consistency regularization on these graphs, large inter-class distances and small intra-class variations can be achieved, obtaining discriminative feature space. To demonstrate our method, we provide mathematical proofs. Moreover, considering the numerical problem in actual training, we further introduce an upper-bounded GC term for ensuring the learning of discriminative embedding. The main contributions of this work can be summarized as follows:

- We propose the deep ***Consistent Graph Metric Learning*** (***CGML***) framework, a novel graph-based view for learning discriminative feature representations, which is ‘plug and play’ and can be applied to many existing deep metric methods.
- CGML is achieved by introducing the Graph Consistency (GC) term, which is to match the property of our defined *Discriminative Graph* and has rigourous mathematical proofs. Then, to ensure the optimization of GC term, an upper-bound of GC is considered.
- Extensive experiments have been performed on several popular datasets for DML, including CARS (Krause et al. 2013), CUB, Stanford Online Products (Oh Song et al. 2016) and In-Shopes (Liu et al. 2016), achieving competitive results.

Related Work

Graph Learning: Graph-based approaches have become attentive in recent computer vision community and are shown to be an efficient way of relation modeling. Constructing graph over the image spatial positions and then propagating mass via random walk has been widely used for object saliency detection (Harel, Koch, and Perona 2007). Graph Convolution Network (GCN) (Kipf and Welling 2016) is proposed on semi-supervised classification. It has been adopted for capturing relations between objects in video recognition tasks (Wang and Gupta 2018). IRG (Liu et al. 2019) employs the graph relation for knowledge distillation. The graph knowledge is also used for visual query answering (Xiong et al. 2019).

However, different from these works, we aim at encouraging the discrimination of the learned deep embedding by regularizing the randomly constructed sub-graphs over data points to be consistent with each other, which is the obvious property of our defined ‘Discriminative Graph’ and discriminative feature distribution.

Deep Metric Learning: DML intends to pull the instances from the same class closer while push the ones from different classes farther apart. The commonly used Contrastive loss (Sun et al. 2014) and Triplet loss (Schroff, Kalenichenko, and Philbin 2015) have been widely explored and applied. Additionally, there are some other deep metric learning methods: Smart-mining (Harwood et al. 2017) combines the local triplet loss and the global loss to supervise the learning of deep metric by hard-example mining. Sampling Matters (Wu et al. 2017) proposes distance weighted sampling strategy. Angular loss (Wang et al. 2017) optimizes a triangle based angular function. Proxy-NCA (Movshovitz-Attias et al. 2017) explains why popular classification loss works from a proxy-agent view, and its implementation is very similar to Softmax. N-Pair loss (Sohn 2016) proposes to use N-Pair tuples for training discriminative embedding, and ALMN (Chen and Deng 2019a) proposes the adaptive large margin N-pair loss by generating geometrical virtual negative point instead of employing hard-sample mining for learning more discriminative

[1] Pay more attention to only the maximum similarity input, the rest inputs might be ignored. Therefore, the strength of the global constraint is weakened.

embedding. SNR (Yuan et al. 2019) employs the idea of Signal-to-Noise Ratio on the deep metric objective and obtains the robust feature embedding, HDC (Yuan, Yang, and Zhang 2017) employs the cascaded models and selects hard-samples from different levels and models. BIER loss (Opitz et al. 2017, 2018) adopts the online gradients boosting methods. DeML (Chen and Deng 2019b) employs the ensemble metrics learned from the hybrid attention proposals. These methods try to improve the performances by resorting to the ensemble idea.

However, different from the above methods that are based on instance-pairs construction, samples mining or metric ensemble, we target the informative graph structure behind data points distribution for learning discriminative embedding. It is a novel view for introducing global constraints.

Proposed Approach

In this section, we will first give the problem background of less-discriminative embedding learning in Section 3.1, and then introduce our defined ***Discriminative Graph*** and its corresponding property, inspired by this property we have our ***Graph Consistency*** (GC) regularization as in Section 3.2, to further ensure the optimization of graph consistency we consider an upper-bounded GC term in Section 3.3, finally we propose the deep ***Consistent Graph Metric Learning*** (CGML) framework in Section 3.4.

Problem Background

Most of the DML works are designed to optimize the relative distances between positive pairs and negative pairs such that the margin constraints can be achieved, such as Contrastive loss (Sun et al. 2014), Triplet loss (Schroff, Kalenichenko, and Philbin 2015) and Quadruplet loss (Chen et al. 2017). However, satisfying these distance margin constraints between instance pairs actually is not equivalent to the learning of discriminative feature representations. Specifically, we take Triplet loss as a toy example as illustrated in Fig. 2. One can observe that after satisfying the margin constraint, the constructed pairs will propose zero losses and contribute little to the update of feature embedding and model parameters. As a result, the data points in feature space will stop moving towards the more discriminative places. Therefore, in order to obtain the compact intra-class distributions and separable inter-class distances, we recast the problem of distance constraint as the graph regularization, which takes the informative graph structure behind data points in feature space into consideration.

Graph Consistency Regularization

In this paper, we rethink the discriminative distance optimization from another novel perspective, i.e. graph optimization. The ultimate target thus turns to learning towards *Discriminative Graph*. Now, we first give the definition of *Discriminative Graph* as below:

Definition 1. *Given large scale data representations $\mathcal{X} = [x_1, \cdots, x_N]$, where $x_i \in \mathbb{R}^d$, N is the data number and they are uniformly coming from C classes. For the c-th class, its biggest intra-class Euclidean distance is α_c and distance from its center to the nearest negative class center is β_c. If $\alpha_c \ll \beta_c, \forall c \in [1, \cdots, C]$, we call a graph G based on these data nodes as Discriminative Graph.*

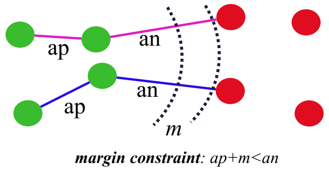

Figure 2: Toy example of the weakness of Triplet loss, where different colors indicate different classes, ap and an mean the positive and negative pair distances respectively. Although the margin constraint has been achieved, the intra-class distribution has large variations, and the inter-class distribution are not separable enough (i.e. has the same magnitude level with intra-class distance).

For briefness, the instances in $\mathcal{X}$ are ordered by category, i.e. $\mathcal{X}$ is constructed by first assigning all the instances belonging to the first class at the beginning several columns and then assigning the second class instances right behind [2], *etc*. And without loss of generality, the graph is based on adjacent matrix S with the commonly used RBF function, where the element s_{ij} indicates the weight of connecting edge between node i and node j on graph G, and $s_{ij} = \exp(-\frac{\|x_i - x_j\|_2^2}{\sigma})$. From Definition. 1, one can observe that learning towards *Discriminative Graph* is consistent with the goal of DML. However, in actual, due to the complex structure of G and complex relations among nodes, it is technically not easy to directly optimize such a ideal discriminative graph.

To this end, we instead propose to regularize the consistency between two sub-graphs $G^{'}$ and $G^{''}$ randomly sampled from G, i.e. to regularize the corresponding adjacent matrixes ($S^{'} \approx S^{''}$) [3], which is an obvious property of our defined *Discriminative Graph*.

Proposition 1. *Given Discriminative Graph G, randomly and independently sample n data points from each class two times and thus obtain two data batches $X^{'}$, $X^{''} \in \mathbb{R}^{d \times nC}$ respectively, then construct the sub-graphs $G^{'}, G^{''}$ along with adjacent matrixes $S^{'}, S^{''}$. We will have that $S^{'} \approx S^{''}$ is the necessary and sufficient condition of Discriminative Graph.*

Proof. For paper length limit, here we just provide the proof of necessary condition, Please see the supplementary file for sufficient condition.

(1) For the intra-class connected nodes (without loss of generality, we take for example the c-th class). Since exp

[2] If not specified, in this paper, all the data-batches will be constructed by this way, including the randomly sampled data-bathes which will be used later.

[3] If not specified, we use superscript $'$and $''$ to represent the sampling.

function is convex, from Jensen Inequality, we have:

$$1 \geq E[s_{ij}] \geq e^{-\frac{E[\|x_i - x_j\|_2^2]}{\sigma}} \geq e^{-\frac{\alpha_c^2}{\sigma}}$$

Then, we compute

$$\begin{aligned} E[(s_{ij}^{'} - s_{ij}^{''})^2] =& E[((s_{ij}^{'} - E[s_{ij}]) - (s_{ij}^{''} - E[s_{ij}]))^2] \\ =& E[(s_{ij}^{'} - E[s_{ij}])^2 + (s_{ij}^{''} - E[s_{ij}])^2 \\ & - 2(s_{ij}^{'} - E[s_{ij}])(s_{ij}^{''} - E[s_{ij}])] \end{aligned}$$

since the data points are i.i.d, $E[s_{ij}] = E[s_{ij}^{'}] = E[s_{ij}^{''}]$ and $E[s_{ij} - E[s_{ij}]] = 0$, thus

$$\begin{aligned} &= 2E[(s_{ij} - E[s_{ij}])^2] = 2Var(s_{ij}) = 2(E[s_{ij}^2] - E^2[s_{ij}]) \\ &\leq 2(1 - e^{-\frac{2\alpha_c^2}{\sigma}}) \end{aligned}$$

Notice that the upper bound $2(1 - e^{-\frac{2\alpha_c^2}{\sigma}})$ is proportional to α_c, showing that the differences between $s_{ij}^{'}$ and $s_{ij}^{''}$ are consistent with the intra-class compactness. Additionally, since $\lim_{\alpha_c \to 0} \frac{2(1-e^{-\frac{2\alpha_c^2}{\sigma}})}{\alpha_c} = 0$ and α_c is a small value, the expected squared difference between $s_{ij}^{'}$ and $s_{ij}^{''}$, i.e. $E[(s_{ij}^{'} - s_{ij}^{''})^2]$, thus will be bounded by a much smaller value $2(1 - e^{-\frac{2\alpha_c^2}{\sigma}})$, in other words, $s_{ij}^{'} \approx s_{ij}^{''}$.

(2) For the inter-class connected nodes, where $x_i,\ x_j$ are sampled from different classes (without loss of generality, we take classes c and k), then we have $\alpha_c \ll \beta_c < \pi$ and $\alpha_k \ll \beta_k < \pi$, where π is the distance between the c-th class center and k-th class center. Then $s_{ij} \in [\exp(-\frac{(\alpha_c+\alpha_k+2\pi)^2}{4\sigma}), \exp(-\frac{(2\pi-\alpha_c-\alpha_k)^2}{4\sigma})]$. And from the Hoeffding's Inequality, we have:

$$Pr\{|s_{ij} - E[s_{ij}]| \geq t\} \leq 2e^{-\frac{2t^2}{(b-a)^2}}$$

where $a = \exp(-\frac{(\alpha_c+\alpha_k+2\pi)^2}{4\sigma}),\ b = \exp(-\frac{(2\pi-\alpha_c-\alpha_k)^2}{4\sigma})$ $Pr\{z\}$ indicates the probability of z. Then, setting RHS as $\delta/2$, we have with probability at least $1 - \delta/2$:

$$|s_{ij} - E(s_{ij})| \leq (b-a)\sqrt{\frac{\log(4/\delta)}{2}}$$

And as $s_{ij}^{'}, s_{ij}^{''}$ are i.i.d, $E[s_{ij}] = E[s_{ij}^{'}] = E[s_{ij}^{''}]$, we have

$$\begin{aligned} |s_{ij}^{'} - s_{ij}^{''}| &= |(s_{ij}^{'} - E[s_{ij}^{'}]) - (s_{ij}^{''} - E[s_{ij}^{''}])| \\ &\leq |(s_{ij}^{'} - E[s_{ij}^{'}])| + |(s_{ij}^{''} - E[s_{ij}^{''}])| \\ &\leq 2(b-a)\sqrt{\frac{\log(4/\delta)}{2}} \end{aligned}$$

From the above inequality, we have that the absolute difference between $s_{ij}^{'}$ and $s_{ij}^{''}$ is also bounded by value $2(b - a)\sqrt{\frac{\log(4/\delta)}{2}}$, and this upper bound is proportional to $\alpha_c,\ \alpha_k$ (intra-class compactness) while inversely proportional to π (inter-class separability).

In summary, for both intra and inter class connections, the difference between $s_{ij}^{'}$ and $s_{ij}^{''}$ are bounded, and the upper bounds are proportional/inversely proportional to the the intra-class compactness/inter-class separability. For a Discriminative Graph, the upper bounds are much smaller values and thus we have $s_{ij}^{'} \approx s_{ij}^{''}$, i.e. $S^{'} \approx S^{''}$. The proof is completed. □

Based on the above observation, in order to learn discriminative feature distributions by graph optimization, one intuitive way is to make the randomly sampled sub-graphs $G^{'},\ G^{''}$ to be as similar as possible, so as to ensure *Discriminative Graph*, intensifying intra-class compactness and inter-class separability within the learned embedding. Therefore, our *Graph Consistency* (GC) regularization term can be formulated as:

$$L_{gc} = \|S^{'} - S^{''}\|_F^2 \tag{1}$$

This regularization term encourages the currently optimizing graph to have the similar property as *Discriminative Graph*, producing large inter-class margins and compact intra-class distributions.

Upper-Bound of Graph Consistency Term

Consider a fact that, practically minimizing $\|S^{'} - S^{''}\|_F$ doesn't means $S^{'}$ is very close to $S^{''}$ due to the numerical problem in training phase. To this end, this paper introduces an upper-bound of the GC term, so as to try the best to ensure the minimization of $\|S^{'} - S^{''}\|_F$:

$$\begin{aligned} \|S^{'}X^{'T} - S^{''}X^{''T}\|_F\|X^{'T^{\dagger}}\|_F + \|S^{''}\|_F\|\xi\|_F\|X^{'T^{\dagger}}\|_F \\ \geq \|S^{'} - S^{''}\|_F \end{aligned} \tag{2}$$

where $X^{'},\ X^{''} \in \mathbb{R}^{d \times nc}$ are the sampled two mini-batch data batches , each containing the same c classes and n random instances per class, and these batches are constructed in the same category-order, e.g. the first n columns of both $X^{'}$ and $X^{''}$ are from the same class, and the next n columns are from another same class. $\xi = (X^{''} - X^{'})^T$ is the residual between two sampled batches. † is the generalized inverse.

Proof.

$$\begin{aligned} S^{'}X^{'T} - S^{''}X^{''T} &= S^{'}X^{'T} - S^{''}X^{'T} + S^{''}X^{'T} - S^{''}X^{''T} \\ &= (S^{'} - S^{''})X^{'T} - S^{''}(X^{''} - X^{'})^T \\ &= (S^{'} - S^{''})X^{'T} - S^{''}\xi \end{aligned}$$

$$\begin{aligned} &\Rightarrow S^{'}X^{'T} - S^{''}X^{''T} + S^{''}\xi = (S^{'} - S^{''})X^{'T} \\ &\Rightarrow \|(S^{'}X^{'T} - S^{''}X^{''T} + S^{''}\xi)X^{'\dagger^T}\| = \|S^{'} - S^{''}\|_F \\ &\Rightarrow \|S^{'} - S^{''}\|_F \leq \|S^{'}X^{'T} - S^{''}X^{''T}\|_F\|X^{'T^{\dagger}}\|_F \\ &+ \|S^{''}\|_F\|\xi\|_F\|X^{'T^{\dagger}}\|_F \end{aligned}$$

□

	Iterations									
	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
$\|X^{'}\|_F^2$	1.26	1.14	1.06	1.11	1.02	0.98	0.95	0.98	1.02	0.97

Table 1: Changing of the norm of features during the training phase. It can be observed that the norm is relatively stable and can not be increased.

Before minimizing this upper-bound, observing Eq. 2, we notice that there are several sub-terms that can be removed. (1) First, minimizing this upper-bound is to minimize $\|S^{''}\|_F\|\xi\|_F\|X^{'T^{\dagger}}\|_F$. This might lead to the minimization of $S^{''}$ and thus lead to the decrease of intra-class similarity $s_{ij}^{''}$ (which is the similarity between samples from the same class), i.e. to enlarge the intra-class distances. This violates the basic discrimination criterion of DML and thus isn't what we want. (2) Second, minimizing this upper-bound is to minimize the $\|S^{'}X^{'T} - S^{''}X^{''T}\|_F\|X^{'T^{\dagger}}\|_F$. This might lead to the minimization of $\|X^{'T^{\dagger}}\|_F$. Since $\|X^{'T^{\dagger}}\|_F^2 = \sum \frac{1}{\sigma_i^2}, \|X^{'T}\|_F^2 = \sum \sigma_i^2$, where σ_i is singular values of matrix $X^{'T}$, minimizing $\|X^{'T^{\dagger}}\|_F$ means to increase $\|X^{'T}\|_F$. However, the norm of the features can not be arbitrarily increased and its convergence level are actually limited and stable as shown in Tab. 1. Because, for a single linear layer $Y = W^TX$, the output feature norm $\|Y\|_F$ is bounded by $\|W\|_F\|X\|_F$, and in practical we will L2-regularize $\|W\|_F$, thus the output feature norm will be restricted to a stable level [4], furthermore, as the deep model is cascaded model, i.e. $Y = \psi(\cdots\psi(W_2^T\psi(W_1^TX)))$ where ψ is the piecewise linear activation function, then this restriction is much stronger due to the cascaded L2-regularization, thus the final output feature norm will be more stable and can not be easily increased.

To this end, by removing the unwanted or stable subterms, the minimization of the upper-bound term then becomes the minimization of that as follows:

$$L_{\overline{gc}} = \|S^{'}X^{'T} - S^{''}X^{''T}\|_F \tag{3}$$

and we employ it as our final ***Graph Consistency*** regularization term.

Remark: Leaving the above analysis, we interpret it from another perspective. From $L_{\overline{gc}}$, one can observe that both $S^{'}X^{'T}$ and $S^{''}X^{''T}$ have similar formulation with random walk propagation (without probability normalization), in other words, $S^{'}X^{'T}$ and $S^{''}X^{''T}$ can be regarded as the new generated node representations by linearly weighting the original node features X^T by the weights S. In this process, the original node will be mapped to a new place in the same feature space by considering its distances to all the other nodes. Therefore, minimizing $\|S^{'}X^{'T} - S^{''}X^{''T}\|_F$ means to maintain the consistency between the generated new nodes, implicitly regularize the consistency between the original two sub-graphs.

In summary, minimizing $L_{\overline{gc}}$ which comes from an upper-bound is to ensure the learning of Discriminative-Graph.

Deep Consistent Graph Metric Learning

In this paper, the main idea is to regularize the randomly sampled sub-graphs to be consistent so as to match the property of *Discriminative Graph* which is of compact intra-class distributions and separable inter-class margins. Thus, the framework of CGML can be generally applied to several popular metric learning objective functions, where we simultaneously train our *Graph Consistency* term $L_{\overline{gc}}$ and the distance metric term L_m as follows:

$$\min_{\theta_f} L = L_m + \lambda L_{\overline{gc}} \tag{4}$$

where θ_f is the model parameters to be optimized and λ is the trade-off hyper-parameter. In order to demonstrate the effectiveness of the proposed CGML framework, we develop various widely used deep metric learning objective functions here, i.e. L_m:

CGML (Tri): For triplet-tuple and Euclidean distance measurement, we employ (Schroff, Kalenichenko, and Philbin 2015):

$$L_m = \sum_i^N [\|x_i - x_{i^+}\|_2^2 - \|x_i - x_{i^-}\|_2^2 + m]_+ \tag{5}$$

where this loss function constrains the distances of negative pairs to be larger than that of the positive pairs by margin m and the feature representations x_i are assumed to be on the unit sphere. In experiments, we find $m = 0.1$ performs best.

CGML (N-Pair): For N-tuple and inner-product similarity , we employ (Sohn 2016):

$$L_m = \sum_{i=1}^N \log(1 + \sum_{j=1, y_j \neq y_i}^N exp(x_i^T x_j - x_i^T x_{i^+})) \tag{6}$$

where this loss function constrains the inner-products of every negative pair $x_i^T x_j$ to be smaller than that of the positive pair $x_i^T x_{i^+}$.

CGML (Binomial): For contrastive-tuple and cosine similarity, we employ (Yi et al. 2014):

$$L_m = \sum_{i,j} \log(1 + e^{-(2s_{ij}-1)\alpha(D_{ij}-\beta)\eta_{ij}}) \tag{7}$$

where $s_{ij} = 1$ when x_i, x_j are from the same class, otherwise $s_{ij} = 0$. $\alpha = 2, \beta = 0.5$ are the scaling and translation parameters respectively, η_{ij} is the penalty coefficient and is set to 1 if $s_{ij} = 1$, otherwise $\eta_{ij} = 25$, the cosine similarity $D_{ij} = \frac{x_i^T x_j}{\|x_i\|\|x_j\|}$.

[4] Using bias also has the similar conclusion since the bias is still L2 regularized. Here, for simplicity, the bias is omitted.

Experiments

Implementation: For fair comparison, we choose to use two different backbone models when comparing with different methods: First, following many previous works, e.g. Lifted (Oh Song et al. 2016), Angular Loss (Wang et al. 2017), ALMN (Chen and Deng 2019a), DAML (Duan et al. 2018), DAMLRRM (Xu et al. 2019), we choose the pretrained *InceptionV1* (Szegedy et al. 2015) as our bedrock CNN and randomly initialized an added fully connected layer; Second, for comparison with recently proposed methods, such MS(Wang et al. 2019a) and RLL(Wang et al. 2019b), we choose the pretrained *InceptionBN* model as our bedrock CNN. If not specified, we set the embedding size as 512 throughout our experiments. We also adopt exactly the same data preprocessing method (Oh Song et al. 2016) so as to make fair comparisons with other works [5]. For training, the optimizer is Adam (Kingma and Ba 2014) with learning rate $1e-5$ and weight decay $2e-4$. The training iterations are $5k$ (CUB), $10k$ (CARS), $20k$ (Stanford Online Products and In-Shop), respectively. The new fc-layer is optimized with 10 times learning rate for fast convergence. Moreover, for fair comparison, we use minibatch of size $n = 130$ throughout our experiments, which is composed of $m = 13$ random selected classes with 10 instances each class. Our work is implemented by caffe.

Evaluation: For fair comparison, following many other works, the retrieval performance is evaluated by Recall@K metric. And following (Oh Song et al. 2016), we evaluate the clustering performances via *normalized mutual information*(NMI) and F_1 metrics. The input of NMI is a set of clusters $\Omega = \{\omega_1, \dots, \omega_K\}$ and the ground truth classes $\mathbb{C} = \{c_1, \dots, c_K\}$, where ω_i represents the samples that belong to the ith cluster, and c_j is the set of samples with label j. NMI is defined as the ratio of mutual information and the mean entropy of clusters and the ground truth, NMI$(\Omega, \mathbb{C})= \frac{2I(\Omega,\mathbb{C})}{H(\Omega)+H(\mathbb{C})}$, and F_1 metric is the harmonic mean of precision and recall as follows $\mathrm{F}_1 = \frac{2PR}{P+R}$.

Datasets: Then our CGML is evaluated over the widely used benchmarks:

1. **CARS** contains 16,185 car images from 196 classes. We split the first 98 classes for training (8,054 images) and the rest 98 classes for testing (8,131 images).
2. **CUB** includes 11,788 bird images from 200 classes.We use the first 100 classes for training (5,864 images) and the rest 100 classes for testing (5,924 images).
3. **Stanford Online Products** has 11,318 classes for training (59,551 images) and the other 11,316 classes for testing (60,502 images).
4. **In-Shop** contains 3,997 classes for training(25,882 images) and the resting 3,985 classes for testing(28,760 images). The test set is partitioned into the query set of 3,985 classes(14,218 images) and the retrieval database set of 3,985 classes(12,612 images).

Ablation Experiments

Comparison with Other Works

To highlight the significance of our CGML framework, we compare with the aforementioned corresponding baseline methods, i.e. the widely used Triplet (Schroff, Kalenichenko, and Philbin 2015), N-Pair (Sohn 2016) and Binomial (Yi et al. 2014), moreover, we also compare our CGML with some other popular DML methods, such as Inception-based methods: Lifted (Oh Song et al. 2016), Angular Loss (Wang et al. 2017), ALMN (Chen and Deng 2019a), DAML (Duan et al. 2018), DAMLRRM (Xu et al. 2019). [6]

The experimental results over CUB, CARS (Krause et al. 2013), Stanford Online Products (Oh Song et al. 2016) and In-shop (Liu et al. 2016) are in Tab.2-Tab.4 respectively, bold number indicates the improvements over baseline methods. From these tables, one can observe that our CGML consistently improves the performances of the original deep metric learning methods (i.e. Triplet, N-Pair and Binomial losses) on all the benchmark datasets by a large margin, demonstrating the necessity of explicitly enhancing the discrimination ability of the learned metric and validating the universality and effectiveness of our CGML. Furthermore, our CGML (Binomial) also surpasses almost all the listed approaches.

In summary, learning towards discriminative embeddings by graph regularization is effective and important, which is achieved by regularizing the randomly sampled graphs to be consistent such that the property of *Discriminative Graph* can be obtained.

Relations with Global Optimization Methods

Recently, there are some other works targeting at using global constraints, such as Lifted(Oh Song et al. 2016),N-Pair(Sohn 2016),MS(Wang et al. 2019a). However, we emphasize that their formulations /implementations actually influence and limit their global constraints, even if they have similar target, i.e. using the global structure, as ours. For example, Lifted/N-Pair/MS are all based on Softmax-function. While SoftMax-function has a Maximum-Domination problem, i.e. paying more attention to the maximum input; in other words, it will only focus and magnify the influence of the pair with the biggest similarity, but ignore influences of the rest pairs, thus weakening the actual constraints on the global structure. On the contrary, Our CGML treats all the pairs equally without discrimination. And From the above experimental comparisons, we can observe that our CGML can further improve the N-Pair/MS results, showing that the global constraints from N-Pair/MS indeed are not enough. In summary, this paper propose a novel regularization term from graph perspective for global structure optimization.

[5] Only the images in CARS dataset are preprocessed differently, since we find our preprocessing method can lightly improve the performances, see the detail underneath Tab.2

[6] As a common knowledge, the performances of ensemble model are actually indeed better than single model. Therefore, in this paper, the ensemble methods such as HDC(Yuan, Yang, and Zhang 2017), BIER(Opitz et al. 2017, 2018), ABE(Kim et al. 2018), DeML(Chen and Deng 2019b) and Divide&Conquer (Sanakoyeu et al. 2019) are not listed in tables.

	CARS						Stanford Online Products					
Method	R@1	R@2	R@4	R@8	NMI	F1	R@1	R@10	R@100	R@1000	NMI	F1
Lifted	49.0	60.3	72.1	81.5	55.1	21.5	62.1	79.8	91.3	97.4	87.4	24.7
Clustering	58.1	70.6	80.3	87.8	59.0	-	67.0	83.7	93.2	-	89.5	-
Angular	71.3	80.7	87.0	91.8	62.4	31.8	70.9	85.0	93.5	98.0	87.8	26.5
ALMN	71.6	81.3	88.2	93.4	62.0	29.4	69.9	84.8	92.8	-	-	-
DAML	75.1	83.8	89.7	93.5	66.0	36.4	68.4	83.5	92.3	-	89.4	32.4
DAMLRRM	73.5	82.6	89.1	93.5	64.2	33.5	69.7	85.2	93.2	-	88.2	30.5
Triplet	68.1	78.8	86.4	92.0	59.1	26.7	57.6	75.5	88.3	96.2	86.4	20.6
CGML(Tri)	***75.8***	***84.7***	***90.9***	***95.2***	***63.5***	***32.9***	***64.1***	***79.5***	***90.2***	***96.8***	***87.1***	***23.2***
N-Pair	74.4	83.6	89.8	93.8	61.8	29.9	67.8	83.9	93.1	97.8	87.7	25.6
CGML(N-Pair)	***75.8***	***84.4***	***90.5***	***94.4***	***62.6***	***31.0***	***68.4***	***84.3***	***93.2***	***97.8***	***88.1***	***27.0***
Binomial	73.1	82.3	88.3	92.7	61.7	28.4	68.2	84.0	93.1	97.7	88.5	29.9
CGML(Binomial)	***79.3***	***86.9***	***91.5***	***94.7***	***65.4***	***33.9***	***70.8***	***85.4***	***93.3***	***97.8***	***89.7***	***32.3***

Table 2: Comparisons(%) with other works on CARS (Krause et al. 2013) and Stanford Online Products (Oh Song et al. 2016). λ for CGML(Tri, N-Pair, Binomial) are $\{0.001, 0.002, 0.002\}$ *resp*. Here, the images are directly resized to 256x256, which are different from (Oh Song et al. 2016), then a 227x227 random region is cropped.

	CUB					
Method	R@1	R@2	R@4	R@8	NMI	F1
Lifted	47.2	58.9	70.2	80.2	56.2	22.7
Clustering	48.2	61.4	71.8	81.9	59.2	-
Angular	53.6	65.0	75.3	83.7	61.0	30.2
ALMN	52.4	64.8	75.4	84.3	60.7	28.5
DAML	52.7	65.4	75.5	84.3	61.3	29.5
Triplet	49.4	61.8	73.0	82.1	57.2	24.3
CGML(Tri)	***53.3***	***64.9***	***75.7***	***84.5***	***60.2***	***27.0***
N-Pair	50.5	63.2	74.2	83.1	59.2	26.3
CGML(N-Pair)	***52.1***	***64.2***	***75.4***	***84.5***	***60.4***	***28.5***
Binomial	52.5	64.1	74.8	84.0	59.2	26.9
CGML(Binomial)	***54.8***	***66.2***	***76.2***	***84.5***	***61.6***	***30.7***

Table 3: Comparisons(%) with other works on CUB. λ for CGML (Tri, N-Pair, Binomial) are $\{0.001, 0.002, 0.002\}$ *resp*.

	In-Shop					
Method	R@1	R@10	R@20	R@30	R@40	R@50
FashionNet	53.0	73.0	76.0	77.0	79.0	80.0
HDC	62.1	84.9	89.0	91.2	92.3	93.1
BIER	76.9	92.8	95.2	96.2	96.7	97.1
HTL	80.9	94.3	95.8	97.2	97.4	97.8
Triplet	63.8	86.8	91.0	92.6	93.9	94.8
CGML(Tri)	***67.5***	***89.7***	***93.2***	***94.8***	***95.7***	***96.2***
N-Pair	78.3	94.1	95.8	96.7	97.4	97.7
CGML(N-Pair)	***78.9***	***94.3***	***95.9***	***96.8***	97.4	97.7
Binomial	81.8	94.1	96.3	97.2	97.6	97.9
CGML(Binomial)	***82.6***	***94.4***	***96.5***	***97.3***	***97.7***	97.9

Table 4: Comparisons(%) with other works on In-shop (Liu et al. 2016). λ for CGML (Tri, N-Pair, Binomial) are $\{0.001, 0.002, 0.002\}$ *resp*.

Conclusion

In this paper, we propose the deep *Consistent Graph Metric Learning* framework, a generally applicable technique to various conventional deep metric learning approaches. The major idea is to explicitly intensify the intra-class compactness and inter-class separability within the learned embedding with the help of our Graph Consistency regularization term. Extensive experiments on the popular benchmarks (i.e. CUB, CARS, Stanford Online Products and In-Shop) demonstrate the significance and necessity of our idea of learning discriminative metric by graph optimization.

References

Chen, B.; and Deng, W. 2019a. Deep embedding learning with adaptive large margin N-pair loss for image retrieval and clustering. *Pattern Recognition* 93: 353–364.

Chen, B.; and Deng, W. 2019b. Hybrid-Attention based Decoupled Metric Learning for Zero-Shot Image Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 403–412.

Deng, C.; Chen, Z.; Liu, X.; Gao, X.; and Tao, D. 2018. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing* 27(8): 3893–3903.

Duan, Y.; Zheng, W.; Lin, X.; Lu, J.; and Zhou, J. 2018. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2780–2789.

Ge, W. 2018. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 269–285.

Gordo, A.; Almazan, J.; Revaud, J.; and Larlus, D. 2017. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision* 124(2): 237–254.

Harel, J.; Koch, C.; and Perona, P. 2007. Graph-based visual saliency. In *Advances in neural information processing systems*, 545–552.

Harwood, B.; Kumar, B.; Carneiro, G.; Reid, I.; Drummond, T.; et al. 2017. Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2821–2829.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* .

Kim, W.; Goyal, B.; Chawla, K.; Lee, J.; and Kwon, K. 2018. Attention-based ensemble for deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 736–751.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* .

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.

Leal-Taixé, L.; Canton-Ferrer, C.; and Schindler, K. 2016. Learning by tracking: Siamese CNN for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 33–40.

Liu, Y.; Cao, J.; Li, B.; Yuan, C.; Hu, W.; Li, Y.; and Duan, Y. 2019. Knowledge Distillation via Instance Relationship Graph. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Movshovitz-Attias, Y.; Toshev, A.; Leung, T. K.; Ioffe, S.; and Singh, S. 2017. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, 360–368.

Noh, H.; Araujo, A.; Sim, J.; Weyand, T.; and Han, B. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision*, 3456–3465.

Oh Song, H.; Jegelka, S.; Rathod, V.; and Murphy, K. 2017. Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5382–5390.

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.

Opitz, M.; Waltner, G.; Possegger, H.; and Bischof, H. 2017. Bier-boosting independent embeddings robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, 5189–5198.

Opitz, M.; Waltner, G.; Possegger, H.; and Bischof, H. 2018. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE transactions on pattern analysis and machine intelligence* .

Sanakoyeu, A.; Tschernezki, V.; Buchler, U.; and Ommer, B. 2019. Divide and Conquer the Embedding Space for Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 471–480.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, 1857–1865.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, 1988–1996.

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.

Tao, R.; Gavves, E.; and Smeulders, A. W. 2016. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1420–1429.

Wang, J.; Zhou, F.; Wen, S.; Liu, X.; and Lin, Y. 2017. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, 2593–2601.

Wang, X.; and Gupta, A. 2018. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 399–417.

Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019a. Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5022–5030.

Wang, X.; Hua, Y.; Kodirov, E.; Hu, G.; Garnier, R.; and Robertson, N. M. 2019b. Ranked List Loss for Deep Metric Learning. *arXiv preprint arXiv:1903.03238* .

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.

Wu, C.-Y.; Manmatha, R.; Smola, A. J.; and Krahenbuhl, P. 2017. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2840–2848.

Xiong, P.; Zhan, H.; Wang, X.; Sinha, B.; and Wu, Y. 2019. Visual Query Answering by Entity-Attribute Graph Matching and Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, X.; Yang, Y.; Deng, C.; and Zheng, F. 2019. Deep Asymmetric Metric Learning via Rich Relationship Mining. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, 34–39. IEEE.

Yuan, T.; Deng, W.; Tang, J.; Tang, Y.; and Chen, B. 2019. Signal-to-Noise Ratio: A Robust Distance Metric for Deep Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4815–4824.

Yuan, Y.; Yang, K.; and Zhang, C. 2017. Hard-aware deeply cascaded embedding. In *Proceedings of the IEEE international conference on computer vision*, 814–823.