1

How Classification Baseline Works for Deep Metric Learning: A Perspective of Metric Space

Yuanqu Mou MOUYQ@SMAIL.NJU.EDU.CN
Nanjing University
Zhengxue Jian JIANZX@BIT.EDU.CN
Beijing Institute of Technology
Haiyang Bai BAIHY@SMAIL.NJU.EDU.CN
Chang Gou* CGOU@SMAIL.NJU.EDU.CN
Nanjing University

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Deep Metric Learning (DML) stands as a powerful technique utilized for training models to capture semantic similarities between data points across various domains, including computer vision, natural language processing, and recommendation systems. Current approaches in DML often prioritize the development of novel network structures or loss functions while overlooking metric properties and the intricate relationship between classification and metric learning. This oversight results in significant time overhead, particularly when the number of categories increases. To address this challenge, we propose extending the loss function used in classification to function as a metric, thereby imposing constraints on the distances between training samples based on the triangle inequality. This approach is akin to proxy-based methods and aims to enhance the efficiency of DML. Drawing inspiration from metrically convex metrics, we introduce the concept of a "weak-metric" to overcome the limitations associated with certain loss functions that cannot be straightforwardly extended to full metrics. This ensures the effectiveness of DML under various circumstances. Furthermore, we extend the Cross Entropy loss function to function as a weak-metric and introduce a novel metric loss derived from Cross Entropy for experimental comparisons with other methods. The results underscore the credibility and reliability of our proposal, showcasing its superiority over state-of-the-art techniques. Notably, our approach also exhibits significantly faster training times as the number of categories increases, making it a compelling choice for large-scale datasets.

Keywords: Deep Metric Learning; Metric Property; Weak Metric Learning

1. Introduction

Deep Metric Learning (DML) attempts to embed given samples into a metric space to capture their semantic similarities, which has been applied in tasks such as zero-shot, few-shot, and self-supervised learning. Most existing work aims to design a good loss function from two fundamental models: siamese network Chopra et al. (2005); Hadsell et al. (2006) and triplet network Hoffer and Ailon (2015), both of which directly use the loss function

* Corresponding author

domain in the semantic vector to control the distance of two samples with identical label. However, triplet samples suffer from high time complexity in the training process, and the performance of the siamese network is proven worse than the triplet network Gonzalez-Zapata et al. (2022). Baseline DML methods take the metric into the design of the loss function, which is actually a composition of the metric and continuous function instead of the metric itself, such as triplet loss Hoffer and Ailon (2015) and margin loss Liu et al. (2016). These loss functions are introduced to directly decrease the distance of samples with identical labels and increase the distance for a wide margin of samples with distinct labels, in which the distance is defined by the metric of semantic vector mapped by the network, which needs at least two networks to obtain loss.

While proxy-based methods, proposed firstly in Movshovitz-Attias et al. (2017), aim to solve this problem, and in which metric losses (e.g., margin loss, triplet loss) are bounded in proxy-based methods, and the following work based on proxy improves the performance of DML. Which is similar to classification and classification is a typical baseline for proxy-based methods Zhai and Wu (2018).

In this article, we consider an indirect method by the property of metric without a tedious structure of network. We extend the loss function to a metric such that one network for classification is enough to embed all samples into a metric space by its properties. A metric is a function on a given set that satisfies three conditions: non-negativity, symmetry, and triangle inequality. We find that non-negativity is the same as loss, that when the label is equal to the semantic vector of a sample, the value of the loss is zero; symmetry is naive for distance; triangle inequality provides an upper bound of the distance between any two samples. The upper bound is actually the loss between semantic vectors and labels, which is small after training.

DML (Deep Metric Learning) typically conducts training and testing using classification datasets, such as Cars-196 Wang et al. (2019), Stanford Online Products (SOP) Oh Song et al. (2016), and CIFAR-10 Krizhevsky et al. (2009). It's important to note that both metric learning and classification utilize these classification datasets, albeit with distinct objectives. Metric learning is primarily focused on clustering data within a metric space, often involving techniques like K-nearest neighbors (KNN) for measurement, while classification primarily aims to predict the class label of samples, emphasizing precision.

Our proposal is inspired by the intriguing relationship between classification and metric learning. Specifically, we treat the labels in the classification task as proxies within the embedded metric space. This approach ensures that any two samples sharing identical labels are inherently close, a property derived from the triangle inequality of the metric.

Furthermore, to address the challenge of adapting certain loss functions to a metric space, we introduce the concept of a "metrically uniform convex metric," which offers more flexibility compared to a conventional metric. We then extend the use of Cross Entropy, a commonly used loss function in classification, to this weak-metric space (which may not be easily extended to a full metric).

In our experimental evaluations, we follow the principles of fair comparison, as suggested by Musgrave et al. Musgrave et al. (2020). Notably, our proposed model exhibits a notably simpler architecture than existing methods, resulting in a faster training process, particularly when applied to larger datasets. Our proposal encompasses several key contributions:

- Reveal that a task of classification contains a metric space when the loss function of classification is regarded as a metric.
- Propose a weak-metric that can be used as an extension of more functions including Cross Entropy with keeping effectiveness.

2. Related Work

2.1. Deep Metric Learning

Deep metric learning plays a pivotal role in classic classification and clustering tasks, encompassing domains such as face recognition, image classification, and fine-grained identification Schroff et al. (2015). The primary objective of deep metric learning is to train a network capable of transforming sample feature data from the original space into another metric space. This transformation leads to a reduction in the distance between samples sharing the same labels while increasing the distance between samples with different labels.

In the realm of deep metric learning, there exist two main categories of methods: pair-wise methods and proxy-based methods. Notable pair-wise methods include triplet loss Hoffer and Ailon (2015), circle loss Sun et al. (2020), contrastive loss Hadsell et al. (2006), binomial deviance loss Yi et al. (2014), lifted structure loss Oh Song et al. (2016), and multi-similarity loss Wang et al. (2019), among others. Xuan et al. propose a general framework named GOAL to analyze the step of gradient update in loss functions, and find a better target for VSE (Visual-Semantic Embedding) problems Xuan and Chen (2023). In a triplet network, a single image may exist in different triplet units, therefore the number of triplet increases, and the amount of training increases. To address this problem, Ding et al. improve the effective scheme to generate triplets and optimize the gradient descent algorithm, so that the algorithm depends on the number of original images instead of the triplets Ding et al. (2015). Unlike the triplet network, Annarumma et al. take the global structure of embed space and overlapping labels into consideration, and distinguish between normal medical images and abnormal medical images through loss function Annarumma and Montana (2018).

Furthermore, proxy-based methods essentially resolve training overhead in pair-wise method, which is firstly proposed in Movshovitz-Attias et al. (2017), and more and more methods are proposed in the following Kim et al. (2020); Roth et al. (2022); Teh et al. (2020); Zhai and Wu (2018), while the other kind of version to prompt performance for proxy-based methods are a variety of sampling strategy in Lin et al. (2018); Lu et al. (2019); Zheng et al. (2019); Roth et al. (2020).

Hu et al. propose a deep metric learning model that utilizes neural network to perform hierarchical nonlinear transformations of faces, with the ability to check the kindred relationship of two different faces Hu et al. (2014); Lu et al. (2017). With deep metric learning method, researches on human faces also include facial expression recognition Liu et al. (2017b), facial age estimation Liu et al. (2017a), etc. Summarize and analyze relevant research, most deep metric learning tasks are combined with specific classification tasks, and the appropriate loss function according to the task characteristic.

In recent year, G.Z. proposed a guided method for DML based on knowledge distillation which has a breakthrough on the performance on CIFAR10 Gonzalez-Zapata et al. (2022).

However, these methods do not notice the useful nature of metric and is hard to be used to large-scale dataset due to their splits of labels to anchor and negative samples. A fine-grained training architecture for DML is proposed Wang et al. (2023).

2.2. Metric Space

Metric is widely considered by mathematicians because of its perfect property in topology. A linear norm space can induce a metric space Yosida (2012) such as Minkowski Distance, Euclidean Distance etc, which is accepted in most metric learning. Some metrics cannot be induced by norms due to the linearity of norms but metric can be linear or non-linear. Moreover, there are some weaker definitions of metric for studying the necessity of triangle inequality of metric, such as semi-metricWilson (1931), metrically convex metricSedghi et al. (2008) etc. We define a new weak metric to better reveal the necessity of a metric for the motivation of metric learning based on a metrically convex metric. DaiDai and Hang (2021) proposes a perspective that training with triplet loss is pushing out negative samples and pulling back positive samples to straighten the manifold of samples. A manifold is actually a basic topology space with a differential structure. We consider that in DML, smoothness of space is unnecessary but the metric is more significant.

***Definition* 1** **Metric**: A metric is a function $d : X \times X \to \mathbb{R}$ on a given set X that satisfies non-negative, symmetry and triangle inequality, that is for any $x, y, z \in X$:

$$d(x, y) \geq 0 \quad and \quad d(x, y) = 0 \Leftrightarrow x = y, \tag{1}$$
$$d(x, y) = d(y, x), \tag{2}$$
$$d(x, z) \leq d(x, y) + d(z, y). \tag{3}$$

The metric space is marked as $\{X, d\}$.

As an example of Definition 1 which appears later, we introduce a trivial metric——discrete metric with marks above:

$$d(x, y) = \begin{cases} k & if \quad x \neq y \\ 0 & if \quad x = y \end{cases}, \tag{4}$$

in such metric space, pairwise distance is identical for all elements, which will be nonsense for clustering but still a metric.

3. Theoretical Analysis

In this section, we introduce a straightforward approach to unveil the inherent relationship between classification and metric learning in the context of deep learning, which facilitates the generalization of loss functions by considering potential extensions within them. Furthermore, we put forth a definition of a "weak-metric" that exhibits broad applicability and subsequently delves into the extension of the Cross Entropy loss function. Our proposal hinges on two pivotal concepts: the triangle inequality of the metric and the transition from a loss function to a metric.

3.1. Classification and DML

Generally, a task of classification is trained for decreasing loss between flattened vectors and label vectors such that the index of the maximum scalar in the former is equal to the index of the maximum scalar in the latter, while metric learning is trained for clustering such that margin of distinct classes increases.

For a network $f : X \to \mathbb{R}^c$ in which c denotes the number of classes in classification while denotes the dimension of embedded space in metric learning. Space X is regarded as a linear norm space representing all kinds of input. In Zhai and Wu (2018), classification is proved a strong baseline for proxy-based methods, in which the author only utilizes and improves the loss of classification for proxy-based methods. In this article, we aim to reveal a direct relationship between classification and DML. We assume loss function in classification as:

3.2. Theoretical Results

In the realm of deep metric learning, researchers commonly strive to embed samples into a metric space where samples with the same labels are brought closer together, while those with distinct labels are pushed further apart. Much of the existing work in deep metric learning involves the construction of loss functions, many of which incorporate a metric function, such as the Triplet Loss.

We propose a novel perspective: when a loss function used in classification possesses metric-like characteristics or can be extended to conform to a metric, an effective classification task inherently induces a metric learning task. We distinctly show our proposition as:

***Definition* 2** **Metric-Extendable Loss Function**. For the loss function $L : X \times \mathbb{I}^c \to \mathbb{R}^c$ on a given set X,Y, if there exists a metric $d : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}$ (i.e. satisfies (1,2,3)) such that:

- d is continuous;
- For any $x \in X$, $y \in \mathbb{I}^c$, $L(f(x), y) = d(f(x), y)$,

then L is called the 'Metric-Extendable Loss Function'.

Given the training set as $Tr \subset X$ and its label map $\mathscr{L} : Tr \to \mathbb{I}^c$. If the loss function L is a 'metric-extendable loss function', a network is constructed for mapping a sample $t \in Tr$ to a semantic vector $f(t) \in \mathbb{R}^c$ whose dimension is c. All potential labels are also in the metric space $\{\mathbb{R}^c, d\}$. According to this set, the distance of any two vectors $f(t_1), f(t_2)$ in semantic space that has the same label can be restricted by triangle inequality:

$$L(f(t_1), f(t_2)) \leq L(f(t_1), l) + L(l, f(t_2)), \tag{5}$$

in which $l = \mathscr{L}(t_1) = \mathscr{L}(t_2)$, which is embedded to a metric space $\{\mathbb{R}^c, L\}$ from a classification directly. In fact, in the training step, the used function is still function L itself.

A loss like L2 loss is metric-extendable. However, some loss functions (such as Cross Entropy) cannot directly be extended to a metric, triangle inequality is a little strong for a function. Semi-metric is a kind of weak metric that satisfies (1,2) but not (3), we introduce another weak metric based on 'metrically uniform convex metric' that satisfies (1,2,6), concisely called weak-metric without causing ambiguity as follows:

***Definition* 3** Metrically Uniform Convex Metric (**Weak-Metric**): a function $d : X \times X \to \mathbb{R}$ on a given set X that satisfies(1,2) and: $\exists y_0 \in X$, for any $x, z \in X$(notice it's not "for any $x, z \in X$, $\exists y_0 \in X$" which is the definition of metrically convex metricSedghi et al. (2008)):

$$d(x, z) \leq d(x, y_0) + d(y_0, z). \tag{6}$$

Conveniently we marked it as 'weak-metric with uniform point y_0'. And if there exists weak-metric d satisfies as Definition 2 shows, we call the correspondent L a 'weak-metric extendable loss function'. Our definition is stronger than the metrically convex metric while weaker than the metric. Such a definition makes loss functions like Cross Entropy can be extended and meaningful for metric learning.

In the next, we will reveal the properties of the classification task with 'metric extendable loss function' and 'weak-metric extendable loss function' as losses in the metric space, that is, how much they can obtain a good metric learning result.

***Definition* 4** Universal Loss Function in Classification: set $d : \mathbb{R} \times \mathbb{I} \to [0, +\infty)$ a normal function, $\boldsymbol{x} = (x_1, x_2, ..., x_c), \boldsymbol{y} = (y_1, y_2, ..., y_c) \in \mathbb{R}^c$, $F \in C^2[0, +\infty)$ is a second differentiable function domain in $[0, +\infty)$, a universal loss function $L_m(\boldsymbol{x}, \boldsymbol{y})$ in classification is:

$$L_m(\boldsymbol{x}, \boldsymbol{y}) = F(\sum_{i=1}^{c} d(x_i, y_i)), \tag{7}$$

in which $F(0) = 0$ and F is strictly increasing. $d(x, y)$ represents a loss for a scalar value, thus the distance of any two different labels $\boldsymbol{l}_a, \boldsymbol{l}_b \in \mathbb{R}^c$ is:

$$L_m(\boldsymbol{l}_a, \boldsymbol{l}_b) = F[d(0, 1) + d(1, 0) + (c - 2)d(0, 0)]. \tag{8}$$

In a loss function, $d(0, 1)$ is big enough which means the predicted label is complete not the true label, in loss functions like Cross Entropy the value $d(0, 1) = +\infty$. When d is semi-metric, we have:

***Lemma* 1** Set $d(0, 1) = M$. If $d(x, y)$ is a semi-metric on $\mathbb{R}$ that $L_m(\boldsymbol{x}, \boldsymbol{y})$ is also semi-metric on $\mathbb{R}^c$ and by the way:

$$L_m(\boldsymbol{l}_a, \boldsymbol{l}_b) = F(2M). \tag{9}$$

It's obvious. Then we suppose that after an effective training in classification, loss value ϵ between a sample $\boldsymbol{a} \in \mathbb{R}^c$ with its label $\boldsymbol{l}_a$ is much smaller than $L_m(\boldsymbol{l}_a, \boldsymbol{l}_b)$, formally, we defined that:

***Definition* 5** Well-Trained Classification: Set test set Te, its label function $\mathscr{L} : Te \to \mathbb{R}^c$ and Positive set $P = \{t \in Te \,|argmax f(t) = argmax \mathscr{L}(t)\}$ A well-trained classification network is a function $f : \mathbb{R}^{m \times n} \to \mathbb{R}^c$ after training such that:

$$sup_{\boldsymbol{a} \in f(P)} L_m(\boldsymbol{a}, \mathscr{L}(\boldsymbol{a})) = \epsilon \ll F(2M). \tag{10}$$

This definition is to show that a well-trained classification contains a metric space with such samples embedded, and in the following lemma, we take a two-step demonstration of how a metric or weak-metric loss with the label as its uniform point can make samples

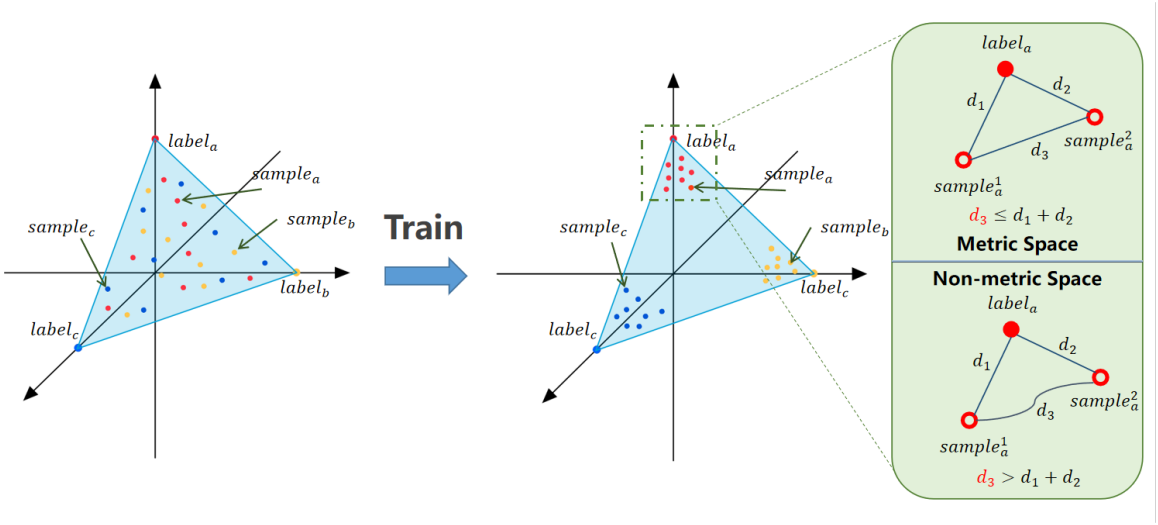

Figure 1: Base idea of our proposal by training a task of classification. A training of classification has actually clustered the samples around their labels(i.e. d_1, d_2 are small enough), and if the used loss function can be extended to a metric, samples with the same labels are closed enough.

closed with identical label and far with the distinct. In next, we say samples $\boldsymbol{a} \in \mathbb{R}^c$ contains that $\boldsymbol{a} \in f(P)$. However, semi-metric is not enough to induce any useful property, thus we introduce stronger metrics by order in the defined loss function.

Lemma 2 If F is a non-convex function, that d is a weak-metric on $\mathbb{R}$ implies that $L_m(\boldsymbol{x}, \boldsymbol{y})$ is a weak-metric on $\mathbb{R}^c$, d is a metric on $\mathbb{R}$ implies that $L_m(\boldsymbol{x}, \boldsymbol{y})$ is a metric on $\mathbb{R}^c$.

This lemma is enough for most of the loss function that F is usually a non-convex function even just a constant function as 3.3 shows. Then we formally give the properties of weak-metric and metric for a correspondent loss function used in metric learning.

Lemma 3 If $L_m(\boldsymbol{x}, \boldsymbol{y})$ is a weak-metric with uniform point l or a metric, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^c$ are two samples with same label $\boldsymbol{l} \in \mathbb{R}^c$, that:

$$L_m(\boldsymbol{a}, \boldsymbol{b}) \leq L_m(\boldsymbol{a}, \boldsymbol{l}) + L_m(\boldsymbol{l}, \boldsymbol{b}) \leq 2\epsilon.$$

Lemma 4 If $L_m(\boldsymbol{x}, \boldsymbol{y})$ is a metric, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^c$ are two samples with different labels $l_a, l_b \in \mathbb{R}^c$ respectively, that:

$$L_m(\boldsymbol{a}, \boldsymbol{b}) \geq F(2M) - 2\epsilon.$$

The two lemma shows that weak-metric can ensure the closed distance in two samples with the same label while metric can further ensure there's a margin between two samples with distinct label. By these two lemmas, when a classification is well-trained, all you need is extending the function in d to a metric, or at least a weak-metric for obtaining pairwise distance.

Theorem If F is non-convex function and d is metric, after well-trained, for any pairwise samples $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^c$:

$$\begin{cases} L_m(\boldsymbol{a}, \boldsymbol{b}) \geq F(2M) - 2\epsilon, \mathscr{L}(a) \neq \mathscr{L}(b) \\ \\ L_m(\boldsymbol{a}, \boldsymbol{b}) \leq 2\epsilon, \ \ \mathscr{L}(a) = \mathscr{L}(b) \end{cases}, \tag{11}$$

which denotes a good metric learning result.

Algorithm 1 Metric Loss

Require: a loss function L, a network $f : \mathbb{R}^{m \times n} \to \mathbb{R}^c$ for task of classification
Ensure: L can be extended to a metric or a weak-metric
extends L to a metric L'
training f with L' as classification
for any two samples img_1, img_2, the distance of them is $L'(f(img_1), f(img_2))$

3.3. Extension of Loss Function

Therefore, a loss function that is also a metric or that can be extended to a metric in classification is the key. In classification, MSE (Mean Square Error) and CE (Cross-Entropy) are mainstream loss functions, and CE is usually used with softmax function (i.e. softmax loss). If softmax is used, the metric space is $\{L, [0,1]^c\}$, in the following we suppose that the softmax function is used, the two functions are defined as:

$$L_{MSE}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{c} \sum_{i=1}^{c} (x_i - y_i)^2, \tag{12}$$

$$L_{CE}(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{i=1}^{c} y_i log(x_i), \tag{13}$$

in which $\boldsymbol{x}$ is sample and $\boldsymbol{y}$ is label(i.e. $\boldsymbol{x} \in [0,1]^c, \boldsymbol{y} \in \{0,1\}^c$), we hope that these functions are not only loss functions but also metrics such that even $\boldsymbol{y} \in [0,1]^c$ that represents a semantic vector after mapping by a sample, it is still defined. L_{MSE} is obviously a metric (see in supplement), but not L_{CE}, and it's obvious that L_{CE} can not be extended to a metric (it's not symmetry), we firstly introduce $L_{CE'}$ from L_{CE}:

$$L_{CE'}(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{i=1}^{c} y_i log(x_i) + (1 - y_i) log(1 - x_i). \tag{14}$$

Then we extend $L_{CE'}$ to a semi-metric function $L^1_{CE'}$:

$$L^1_{CE'}(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{i=1}^{c} log(1 - |x_i - y_i|), \tag{15}$$

when $\boldsymbol{y}$ is a label, $L^1_{CE'} = L_{CE'}$. But it's not enough for our proposal. On the basis of its semi-metric, we prove that it's a weak-metric with its label for (5) as a supplement. CE

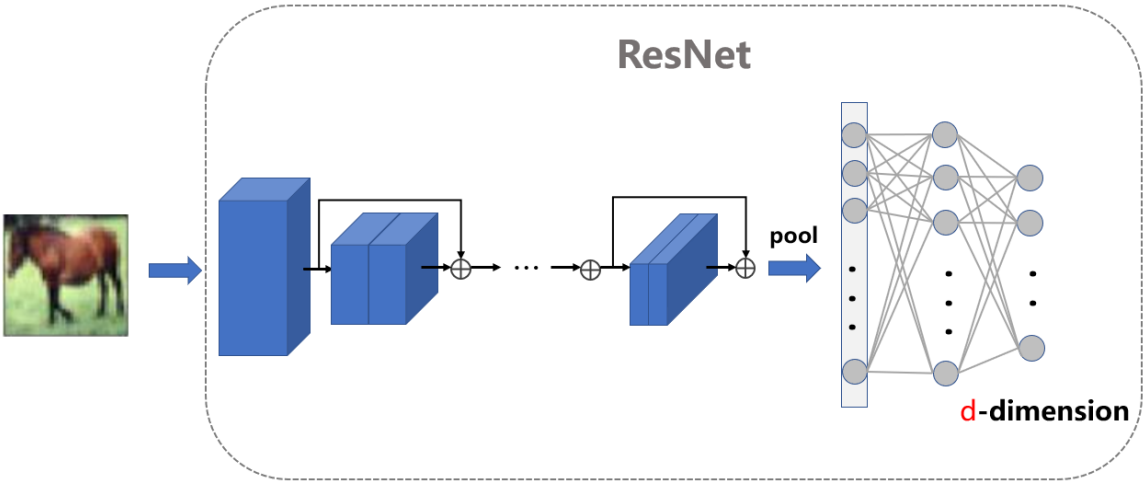

Figure 2: Architecture of experiments that ResNet is our backbone and a single MLP is for embedding to a d-dimension vector space, in classification, d is equal to the number of categories c, which is for all experiments including ours and other methods.

cannot be extended to a metric. To construct a metric from $L^1_{CE'}$, we set three parameters for that empirically:

$$L^2_{CE'}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{p}\sum_{i=1}^{c} log(a - b|x_i - y_i|^p) + \frac{c}{p} log a, \tag{16}$$

in which $a \geq 0$, $0<b \leq 1$, $0<p \leq 1$, $L^2_{CE'}$ is a metric when a, b, p satisfies (17). They are demonstrated as such example:

Example $L^1_{CE'}$ is a weak-metric with label as its uniform point. $L^2_{CE'}$ is a metric iff.:

$$2^{p+1} - 2^{2p} \geq \frac{b}{a}. \tag{17}$$

In practical training of classification with softmax loss function, the loss cannot be set to (13) while a tiny number ϵ was added to x_i to avoid undefined calculation. a is non-zero, we set $b = 1$, which means only if p is tiny enough that (16) is a metric when a is tiny enough for just avoiding dividing zero. But an unduly tiny p degenerates metric (15) to a trivial discrete metric shown in (4), implying that tiny p leads to a collapse of metric space that the distance between each couple tends to be identical.

In fact, a norm in a linear space can induce a metric, but not each metric can be induced by a norm. MSE belongs to the former and $L^2_{CE'}$ belongs to the latter. In next, we show that a metric is better than weak metric when they are trained routinely with the same model and training setup.

4. Experiments

In this section, we conduct a comparative analysis between our proposed approach and existing metric learning methods in the context of classification datasets. To ensure a fair comparison of performance across each method, we employ an identical model setup and maintain consistent parameter settings for each experiment. We use a ResNet18 architecture as our backbone, omitting the use of any pre-trained models.

Our initial findings demonstrate that our proposed method surpasses the state-of-the-art approaches in terms of performance. Furthermore, as the number of categories increases, we observe that our method incurs significantly lower time overhead compared to other existing methods.

4.1. Experimental Results

All the experiments have been implemented using Python 3.7 and Pytorch 1.8 on one NVIDIA RTX 3090 super 24GB. In addition, neither data augmentation nor post-processing (besides global normalization) was applied, which was followed by each experiment, and we used the same model for each experiment (including datasets, backbone, and even each of Relu and BatchNormlization) and the same parameters setup(including epochs, batch size, learning rate and its weight decay and so on). In metric learning, the performance of KNN is a benchmark for measuring, Recall and F-score are direct results for KNN, which was proved not comprehensive for measuringMusgrave et al. (2020). Thus the mentioned Normalized Mutual Information (NMI), R-precision, and Mean Average Precision (MAP) are also considered to our experiments. Notice that in KNN k is a parameter that is related to Recall and F-score but not the latter, parameter R (the number of trained samples whose labels are the same to the queried) instead.

All experiments were implemented on CIFAR10 and CIFAR100, Gonzalez-Zapata et al. (2022) propose a method based on knowledge distillation that much improves the performance on CIFAR10, we compare our proposal with the previous on CIFAR10 and extend the categories to 100 on CIFAR100.

Table 1: Performance on CIFAR10

Model		**Recall@1**	**F-score@1**	**NMI**	**RP**	**MAP@R**
Siamese		25.32	25.96	25.17	24.22	11.76
Triplet		41.93	42.12	39.43	36.65	20.03
Margin		55.71	55.86	45.15	40.90	23.73
MultiSim		56.65	57.04	47.51	47.72	36.04
Guided		82.78	82.73	67.88	81.19	78.85
Ours	p=1e-2	**84.06**	**83.99**	**69.11**	**84.80**	**84.23**

We set several number of p to measure the performance of our proposal, p is best on around $1e-2$ with the less p is the better the performance is unless p is too tiny. Results show that our proposal outperforms the previous. Moreover, we compare the time overhead on the training process in which the previous is exponential order compared with our constant (the number of samples is regarded as static) due to their loss function with

Table 2: Performance on CIFAR100 with two backbones

Model	**Recall@1**	**F-score@1**	**NMI**	**RP**	**MAP@R**
Guided(Res18)	37.96	39.39	42.83	29.12	26.74
Guided(Res50)	43.58	43.86	46.44	37.35	34.68
Ours (Res18)	56.27	56.07	53.80	55.03	54.02
Ours (Res50)	**57.69**	**57.38**	**57.34**	**57.26**	**56.02**

demand of composition between distinct categories. Our time assumption is similar to Guided in CIFAR10, but when the number of categories incereses to 100 on CIFAR100, it's almost five hundred times better than Guided in the training process. The vast distinction in time overhead is obvious that our model is just a pipeline in classification while extra networks for anchor, negative, and positive samples are demanded in previous methods. In addition, we extend the backbone to ResNet50 on CIFAR100 both in ours and Guided.

The performance is shown as Tabel 1 on CIFAR10 and Tabel 2 on CIFAR100, Table 1 shows that all of metrics of ours are better than Guided. Table 2 shows that a huge backbone can increase performance on CIFAR100, which in Guided it's significant than ours.

4.2. Ablation Study

In our further investigation, we explored whether our metric loss is essential for improving sample embeddings. Specifically, we assessed whether it is possible to achieve a well-defined metric space by using a normal metric (such as the Euclidean distance) in combination with a well-trained network using the Cross-Entropy loss function. Our results indicate that it is indeed possible to train a network using Cross-Entropy and the Euclidean metric to obtain a competent metric space, although superior results are achieved with our extended Cross-Entropy loss function, especially when an appropriate value of p is employed.

In this linear space, different metrics exhibit a monotonic property: if the value of one vector is greater than that of another vector in each dimension, this relationship holds true across various norms. Our experiments revealed that most of the vector differences between the two samples exhibit this monotonic relation, resulting in minimal differences between the Mean Squared Error (MSE) metric and our extended Cross Entropy (CE) metric. The critical factor contributing to the performance of Deep Metric Learning (DML) in this context is the effective training in the classification task.

Additionally, our ablation study confirmed that our proposal remains effective even when applied to a network trained solely with Cross Entropy, although our extended CE loss function exhibits a slight advantage. It's worth noting that the use of the Euclidean distance in the loss function is known to yield suboptimal training results.

Furthermore, our findings suggest that our proposal performs optimally when p is slightly smaller than 1. This implies that when a loss function resembles a metric during the training process, it may seem somewhat trivial. However, we discovered that attempting to transform Cross Entropy from a weak-metric to a full metric can enhance the performance of networks trained with Cross Entropy. When p is approximately equal to $1e-2$, we observed the best performance.

Table 3: Ablation study on CIFAR10, in which number followed p means the value of p and similar to a. CEap1 means a tiny a for a fair program with p = 1. P_2 in 'Metric' means we use Euclidean distance when testing and others represent ours with different p.

Model	Metric	**Recall@1**	**F-score@1**	**RP**	**MAP@R**
CEap1 (NMI = 69.11)	P_2	81.73	81.66	84.72	83.95
	p=1	81.62	81.55	84.72	84.33
	p=0.1	83.37	83.29	84.75	84.46
	p=1e-2	84.06	83.99	84.80	84.23
	p=1e-5	**84.15**	**84.08**	11.48	3.04
CEap0.99 (NMI = 63.67)	P_2	81.74	81.73	85.03	84.35
	p=1	81.78	81.77	85.04	84.73
	p=0.1	83.52	83.44	**85.05**	**84.81**
	p=1e-2	84.05	83.97	84.95	84.48
CEap0.9 (NMI = 63.44)	P_2	80.97	80.89	84.50	83.83
	p=1	81.74	81.69	84.51	84.18
CEa3p0.9 (NMI = 64.74)	P_2	79.89	79.89	81.22	80.03
	p=1	79.91	79.90	81.24	80.56

However, when p is decreased, the performance of the training deteriorates. Our attempts to adjust parameters a and p to satisfy Equation (17) and generate a metric space during the ablation study led to worse classification training results.

5. Conclusion

The majority of methods in Deep Metric Learning (DML) focus directly on the task of "training to make samples with the same label close and those with distinct labels far apart." What we propose demonstrates that classification can be used to generate a metric space by extending its loss function into a metric without the need for additional network constructions. This approach can effectively scale to large-scale datasets, whereas methods based on triplet structures often suffer from significant time overhead during the training process.

However, it's important to acknowledge a limitation of our proposal: the dimension of the metric space is fixed by the number of categories, and this cannot be easily addressed through dimension reduction algorithms due to the inherent structure of the metric space. To achieve a well-defined metric space with a varying dimension, one might consider training the classification task with preprocessing to embed labels into the desired dimension space.

References

Mauro Annarumma and Giovanni Montana. Deep metric learning for multi-labelled radiographs. In *Proceedings of the 33rd annual ACM symposium on applied computing*, pages 34–37, 2018.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

Mengyu Dai and Haibin Hang. Manifold matching via deep metric learning for generative modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6587–6597, 2021.

Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48 (10):2993–3003, 2015.

Jorge Gonzalez-Zapata, Ivan Reyes-Amezcua, Daniel Flores-Araiza, Mauricio Mendez-Ruiz, Gilberto Ochoa-Ruiz, and Andres Mendez-Vazquez. Guided deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1489, 2022.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015.

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1875–1882, 2014.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3238–3247, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 689–704, 2018.

Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security*, 13(2): 292–305, 2017a.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–29, 2017b.

Jing Lu, Chaofan Xu, Wei Zhang, Ling-Yu Duan, and Tao Mei. Sampling wisely: Deep image embedding by top-k precision optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7961–7970, 2019.

Jiwen Lu, Junlin Hu, and Yap-Peng Tan. Discriminative deep metric learning for face and kinship verification. *IEEE Transactions on Image Processing*, 26(9):4269–4282, 2017.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE international conference on computer vision*, pages 360–368, 2017.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 681–699. Springer, 2020.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

Karsten Roth, Timo Milbich, and Bjorn Ommer. Pads: Policy-adapted sampling for visual similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6568–6577, 2020.

Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7420–7430, 2022.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

S Sedghi, I Altun, and N Shobe. A fixed point theorem for multi-maps satisfying an implicit relation on metric spaces. *Applicable Analysis and Discrete Mathematics*, pages 189–196, 2008.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.

Eu Wern Teh, Terrance DeVries, and Graham W Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 448–464. Springer, 2020.

Chengkun Wang, Wenzhao Zheng, Junlong Li, Jie Zhou, and Jiwen Lu. Deep factorized metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2023.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5022–5030, 2019.

Wallace Alvin Wilson. On semi-metric spaces. *American Journal of Mathematics*, 53(2): 361–373, 1931.

Hong Xuan and Xi Stephen Chen. Dissecting deep metric learning losses for image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2164–2173, 2023.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd international conference on pattern recognition*, pages 34–39. IEEE, 2014.

Kösaku Yosida. *Functional analysis*. Springer Science & Business Media, 2012.

Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018.

Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 72–81, 2019.