Deep Metric Learning by Exploring Confusing Triplet Embeddings for COVID-19 Medical Images Diagnosis

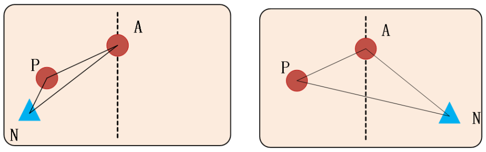
Tongtong Yuan [1] Lingmei Dong [1] Bo Liu [1] Jialiang Huang [2] Chuangbai Xiao [1]

Abstract

Because the COVID-19 virus is highly transmissible, leading to a worldwide increment of new infections and deaths daily, the development of an automated tool to identify COVID-19 using CT images has attracted much attention. Significantly, deep metric learning can be deployed to cluster and classify the fine-grained CT images, which aims to learn a mapping from the original objects to a discriminative feature embedding space. Previous deep metric learning works have been proposed to construct various structures of loss, mine hard samples, or introduce regularization constraints, *etc*. In general, traditional loss functions of deep metric learning methods are based on constraining the distance of the triplet embeddings in the feature space. Instead of focusing on the previous research directions, in this work, we pay attention to exploring confusing triplet embeddings, for the reason that confusing triplet embeddings perform a side effect on the majority of deep triplet-based metric learning methods. By considering the spatial relation of triplet embedding, and conducting theoretical analysis in the feature space, we propose an approach to recognize the confusing triplet embeddings and construct a Confusing Triplet Embedding Learning (CTEL) method by adding a hard constraint on the confusing triplet embeddings. The extensive experiments indicate that our proposed CTEL method achieves more excellent performance on two medical CT image datasets and two fine-grained standard image datasets compared with many state-of-the-art methods.

1. Introduction

The COVID-19 pandemic continues to spread around the world, and has not only led to a global public health crisis but has also affected the world's economy. With the increasing number of new infections, the development of automated tools to identify COVID-19 using CT images is essential, and it has been playing an important role in clinical diagnosis and monitoring of those infected with the disease (Mei et al., 2020). To fight against this pandemic, some works (Javaheri et al., 2020; Zhang et al., 2020; Fan et al., 2020) have shown collecting large-scale training database can improve the recognition accuracy, but it is difficult to be individually executed in practice. However, researchers can aggregate the CT imaging data from different hospitals and build a cross-site learning scheme to alleviate the problem of insufficient data volume at a single site. For instance, Wang *et al.*(Wang et al., 2020b) proposed a novel joint learning framework to improve the diagnosis of COVID-19 by learning of heterogeneous datasets, taking into account those issues such as data heterogeneity and different imaging conditions in different clinical centers. Regardless of which method is used, the indispensable procedure is to extract distinguishable features from the original image.



(a) Same Direction (b) Opposite Direction

Figure 1: Problem of Metric Learning Optimization. "A" represents the anchor sample feature, "P" represents the positive sample feature in the same class as the anchor, and "N" represents the negative samples feature in a different class from the anchor. The typical triplet loss in satisfying the distance D_{AN} than the distance D_{AP} is greater than a certain margin. There are two cases of a triplet: P and N in the same direction or in opposite direction.

Significantly, deep metric learning(Yuan et al., 2019) can make the learned feature distinguishable by keeping feature vectors of the same class close to each other and feature

[1] School of Computer, Beijing University of Technology, Beijing, China [2] Agricultural Bank of China, Beijing, China. Correspondence to: Bo Liu <liubo@bjut.edu.cn>.

Proceedings of the 1^{st} *Workshop on Healthcare AI and COVID-19, ICML 2022.*

vectors of different classes away from each other. However, there still exist some confusing cases in the learning space even when the distance relations between positive and negative pairs satisfy the above conditions. These confusing triplet cases can be recognized by the same direction or opposite direction of the positive sample and the negative sample compared to an anchor. As shown in Figure 1, the feature distance of the negative pair has been larger than of the positive pair by a certain threshold, where the triplet embeddings have satisfied the basic constraint, thus pushing the negative sample or pulling the positive sample will not be continued. However, when the negative sample (N) is in the same direction as the positive sample (P) compared to the anchor (A) in feature spaces (as shown in Figure 1(a)), there will exist unreasonable distance relation in the triplet embeddings: *i.e.*, the distance D_{PN} is smaller than the distance D_{AP}. While the negative sample (N) is in the opposite direction as the positive sample (P) compared to the anchor (A) in feature spaces (as shown in Figure 1(b)), the distance relations of similar/dissimilar pairs are reasonable. Consequently, the same direction of the positive sample P and the negative sample N compared to the anchor A in Figure 1 (a) leads to inducing confusing triplet embedding, which will perform a side effect on the majority of deep triplet-based metric learning procedures.

Therefore, to improve the efficiency of deep triplet-based metric learning methods, we propose a deep metric learning method by exploring confusing triplet embeddings. This method is named Confusing Triplet Embedding Learning (CTEL). More specifically, we firstly propose a strategy to recognize the confusing triplet embeddings by distinguishing the corresponding directions in triplets. Then a hard constraint is established on the confusing triplet embeddings by introducing an auxiliary penalty factor to penalize the distance of negative pairs in the same direction. As a result, even when the distance of the negative pairs is greater than the distance of the positive pairs by a given margin, we can still enable the network to continue to push away negative samples and pull in positive samples in this case, thus making the data more distinguishable in the feature space. In this paper, we select four previous metric learning works, including contrastive loss (Hadsell et al., 2006a), triplet loss (Hadsell et al., 2006b), quadruplet loss (Law et al., 2013) and Multi-Similarity Loss (Wang et al., 2019a) as the partial baselines of our proposed method. We select two public available COVID-19 CT image datasets and two fine-grained image datasets to compare our method with state-of-the-art methods. Extensive experiments indicate that our CTEL method performs well on these four datasets, in terms of classification accuracy, recall rate of retrieval, and F1-Score of clustering. Our main contributions can be summarized below:

- We propose a new deep metric learning method by exploring the confusing triplet embedding during the training procedure in the feature space and conducting a theoretical analysis of the recognition of confusing triplet embeddings.
- Our proposed CTEL method incorporates a redesigned COVID-Net (Wang et al., 2020a) (originally developed for X-ray) network to improve the prediction accuracy of COVID-19 CT images, which is important for promoting the development of COVID-19 medical images diagnosis.
- We also evaluate the image retrieval and clustering performance of our proposed method on two standard fine-grained datasets: CUB-200-2011 and Cars-196. And the results demonstrate that our method outperforms previous methods in image retrieval and clustering performance.

2. Related Work

Deep learning has shown its effectiveness and importance on several computer vision tasks (Xu et al., 2018; 2022). In response to the COVID-19 global pandemic, many researches are being conducted intensively and rapidly to develop artificial intelligence methods (Shi et al., 2020). In the following, we briefly review the deep learning approaches for image-level classification tasks for diagnosis and deep metric learning methods for clustering fine-grained datatsets.

In the area of network architecture, (Xu et al., 2020) aimed to develop a screening model to distinguish COVID-19 pneumonia from those with influenza viral pneumonia and healthy cases through chest CT images, using the positional attention mechanism of ResNet18. Other works have deployed other popular networks, such as VGG (Hall et al., 2020), Inception (Szegedy et al., 2015), ResNet (Narin et al., 2021; Abbas et al., 2021; Farooq and Hafeez, 2020), and DenseNet (Li et al., 2020). Subsequently, new network structures have emerged, which have been carefully designed and validated. A representative one is the COVID-Net (Wang et al., 2020a), which is customized for the identification of COVID-19 and has achieved high accuracy in image level diagnosis based on chest X-ray (CXR for short). In addition, (Wang et al., 2020b) built a powerful backbone in aspects of network architecture and learning strategy by redesigning the recently proposed COVID-Net to improve the accuracy of prediction and learning efficiency. However, except considering the network architecture, the objective function should be reformulated to distinguish the features of the fine-grained CT images.

In the field of metric learning, deep metric learning methods can be optimized by computing the pair-wise similarities between instances in the embedding space. (Hadsell et al., 2006a) proposed one kind of pair-based loss named "contrastive loss", which learns a discriminative metric via Siamese networks. In this way, positive pairs are encouraged to be closer but negative pairs are forced to be more distant with a fixed threshold. Triplet loss (Hadsell et al., 2006b) was proposed to further constrain the distances of dissimilar pairs. Each triplet consists of a positive pair and a negative pair by sharing the same anchor. Triplet loss aims to learn an embedding space where the similarity of a negative pair is lower than that of a positive pair by a given margin. But it still suffers from a weak generalization capability from the training set to the testing set, thus resulting in inferior performance. Quadruplet loss (Law et al., 2013) is an enhanced version of triplet loss, which leads to a larger inter-class variation and a smaller intra-class variation compared to the triplet loss. Inspired by contrastive loss and triplet loss, a number of pair-based deep metric learning algorithms have been developed. For example, N-pair loss (Sohn, 2016) and multi-similarity (MS) loss (Wang et al., 2019a) sampled negative pairs in uniform and constructed a log-exp formulation based on pair-wise distance. These methods demonstrate that using General Pair Weighting (GPW) framework with a unified weighting formulation is a universal approach in pair-based methods. However, these methods do not consider the relative position relations of the positive and negative samples during the training process, which influences the deep metric learning methods.

By exploring confusing triplet embeddings, our work will provide a new deep metric learning method to distinguish the COVID-19 CT images, for improving the effectiveness of artificial intelligent models on COVID-19 CT images recognition. Besides, we also make some comparisons with other methods on standard fine-grained image datasets.

3. Our Proposed Method

As shown in Figure 2, deep metric learning aims to improve the discriminative power of feature vectors generated by CNN. In this section, Section 3.1 details the motivation and design of our new proposed deep metric learning method. In Section 3.2 we revisit the relative metric learning approaches and apply the CTEL method to the classical loss function.

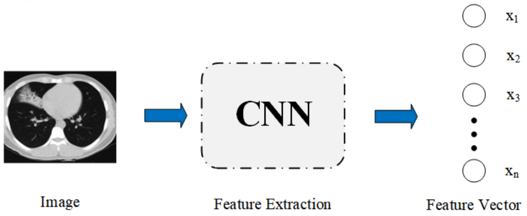


Figure 2: Feature extraction of images. The images are fed into the CNN to generate n-dimensional feature vectors, which contain color, texture, and other features.

3.1. Exploring Confusing Triplet Embeddings

In general, most losses are considered to be well trained when the feature distance of the negative sample pair is at least greater than that of the positive sample pair by a fixed margin. However, there still exists some confusing cases during training as the distance relation of positive and negative pairs satisfies the above condition, which influences the performance of deep metric learning methods.

Firstly, taking triplet loss as an example, when anchor A, the positive sample P and the negative sample N satisfy:

$$D_{\text{AN}} - D_{\text{AP}} = m + \epsilon \quad (\epsilon \geq 0), \tag{1}$$

the relative positions of these triplet embeddings have two following cases: N and P are in the same direction, or N and P are in the opposite direction.

When the distance of AN is greater than that of AP by a certain margin m, the network model will not be able to push N further or pull P closer. However, as shown in Figure 3, when N and P are in the same direction, the distance D_{AP} might be larger than D_{PN}, resulting in this triplet embedding will be confusing triplet embedding. In this case, the positive sample P should be further pulled and the negative sample N should be further pushed away in these confusing triplets.

Therefore, we can derive the concept of confusing triplet embedding: if anchor A, positive sample P and negative sample N satisfy $D_{\text{AN}} - D_{\text{AP}} = m + \epsilon (\epsilon \geq 0)$ and N and P are in the same direction, these three points will be regarded as confusing triplet embedding.

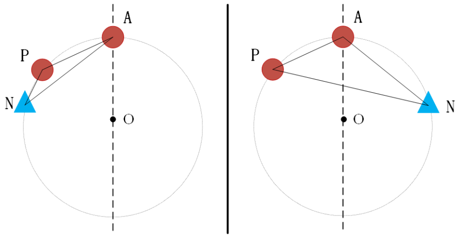


Figure 3: Illustration of triplets in the same Direction and opposite direction. As shown in the left figure, N and P are in the same direction, the distance D_{PN} is smaller than D_{AP}, which is a confusing distance relation. While in the right part, N and P are in the opposite direction, where the distance relations are obviously more reasonable.

In order to achieve the purpose, we introduce a penalty factor η_i. The factor $\eta_i < 1$ when the negative sample N is in the same direction as the positive sample P, otherwise $\eta_i = 1$. Eq. 1 becomes

$$\eta_i D_{\text{AN}} - D_{\text{AP}} = m - \varepsilon \quad (\varepsilon > 0). \tag{2}$$

By Eq. 2, the confusing triplet embeddings will be constrained more rigorously. However, how to judge the confusing triplet embeddings is still pending to be solved. Therefore, we further conduct theoretical analysis on exploring confusing triplet embeddings, including theorems, proofs, and corollaries. The relative analysis is as follows:

Definition 1 (Hyperspere). Given an n-dimensional $(n > 2)$ sphere with a center at the original point and the length of radius being unit length, it is called an n-dimensional hypersphere, denoted as S. The hypersphere S is

$$S^n = \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}. \tag{3}$$

S is the surface or boundary of an n-dimensional sphere $(n > 2)$ and is a type of n-dimensional manifold.

Theorem 1. The normalized feature vector is all on the hypersphere.

Proof of theorem 1. As shown in Figure 2, the vector feature obtained after feature extraction of a picture is $A\left(x_1, x_2, x_3, ..., x_n\right)$ and its mode is $|A| = (\sum_{i=1}^{n} x_i^2)^{1/2} (i = 1, 2, ..., n)$. Then the normalized vector is $A^{'}$, where

$$A^{'} = A/|A| = \frac{x_1, x_2, ..., x_n}{\left(\sum_{i=1}^{n} x_i^2\right)^{1/2}} = \left(y_1, y_2, ..., y_n\right), \tag{4}$$

and the mode of vector $A^{'}$ is $\left|A^{'}\right| = (\sum_{i=1}^{n} y_i^2)^{1/2} = 1, (i = 1, 2, ..., n)$. Thus, the normalized feature vector is all on the n-dimensional hypersphere.

Corollary 1. Given an n-dimensional $(n > 2)$ hypersphere S, and any three points on this hypersphere $A\left(x_1, x_2, ..., x_n\right)$, $B\left(y_1, y_2, ..., y_n\right)$, $C\left(z_1, z_2, ..., z_n\right)$ must not be co-linear.

Proof of corollary 1. For three points A, B and C on the n-dimensional $(n > 2)$ hypersphere S, and the angle between vector $\overrightarrow{AB}$ and the vector $\overrightarrow{AC}$ is θ. Then the dot product of two vectors is

$$\overrightarrow{AB} \cdot \overrightarrow{AC} = \left|\overrightarrow{AB}\right| \times \left|\overrightarrow{AC}\right| \times \cos\theta. \tag{5}$$

Assuming that three points A, B and C are co-linear, then the following three **conditions** must be satisfied:

(1) $x_i - y_i = a(x_i - z_i)$, where $a \neq \pm 1$ and $a \neq 0$.

(2) Vector $\overrightarrow{AB}$ and vector $\overrightarrow{AC}$ are co-linear, and the angle is $\theta = 0°$ or $\theta = 180°$, so the cosine value is $\cos\theta = 1$ or $\cos\theta = -1$.

(3) According to condition (1), (2) and Eq. 5, Eq.6 is followed:

$$\overrightarrow{AB} \cdot \overrightarrow{AC} = \pm\left|\overrightarrow{AB}\right| \times \left|\overrightarrow{AC}\right|. \tag{6}$$

Then, in the case where **condition** (1) and (2) hold, **condition** (3) is correct. The left side of Eq. 6 is

$$\begin{aligned} \overrightarrow{AB} \cdot \overrightarrow{AC} &= \sum_{i=1}^{n} (x_i - y_i)(x_i - z_i) \\ &= \sum_{i=1}^{n} \left[x_i^2 - x_i(y_i + z_i) + y_i z_i\right]. \end{aligned} \tag{7}$$

Since point A is on the hypersphere, $\sum_{i=1}^{n} x_i^2 = 1$. Thus, Eq. 7 becomes

$$\overrightarrow{AB} \cdot \overrightarrow{AC} = 1 - \sum_{i=1}^{n} x_i\left(y_i + z_i\right) + \sum_{i=1}^{n} y_i z_i. \tag{8}$$

According to **condition** (1), it is known that $y_i = -a(x_i - z_i) - x_i$. Therefore,

$$\overrightarrow{AB} \cdot \overrightarrow{AC} = 2a\left(1 - \sum_{i=1}^{n} x_i z_i\right). \tag{9}$$

The right side of Eq. 6 is

$$\left|\overrightarrow{AB}\right| \times \left|\overrightarrow{AC}\right| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \times \sqrt{\sum_{i=1}^{n}(x_i - z_i)^2}. \tag{10}$$

Since points A, B and C are on the hypersphere, Eq. 10 becomes

$$\left|\overrightarrow{AB}\right| \times \left|\overrightarrow{AC}\right| = 2\sqrt{a}\left(1 - \sum_{i=1}^{n} x_i z_i\right). \tag{11}$$

If **condition (3)** holds, then $a = 0$ or 1 will contradict condition (1). Therefore, the assumption that the three points are co-linear does not hold.

Theorem 1 shows that the features of the anchor point A, positive sample P and negative sample N in triplet loss are normalized by the network into three n-dimensional feature vectors with a second normal form value of 1. All three points are on the hyperplane where these three points must not be co-linear. Due to this situation, these three points can determine a unique plane and the only triangle. Any triangle has only one external circle. Then, we can conduct further analysis based on the determined triangle and external circle. The relevant methods used to distinguish the directions of the negative sample and the positive sample are described below.

When AN is at the diameter of circle S, the negative sample N remains neutral with respect to P at this time. In other words, if N is a little further to the left, it is in the same direction as the positive sample P. Conversely, it is in the opposite direction from the positive sample P.

Let the angle of $\angle APN$ be α. According to the theorem that the angle of circumference opposite the diameter is a right angle, when AN is the diameter of a circle, α is a right angle. If the negative sample N is a little further to the left, then $\angle AON > 180°$ and $\alpha > 90°$; Conversely, $\angle AON < 180°$ and $\alpha < 90°$.

Since the angle of α can not be measured directly, the direction could be determined by calculating the cosine of α. The cosine of α in the triangle APN is

$$\cos\alpha = \frac{D_{\text{AP}}^2 + D_{\text{PN}}^2 - D_{\text{AN}}^2}{2 * D_{\text{AP}} * D_{\text{PN}}}, \tag{12}$$

where the angle range of α in the triangle is $0° < \alpha < 180°$. According to the basic properties of the cosine function, the same direction and opposite direction between P and N satisfy the following conditions:

$$\begin{cases} 0 \leq \cos\alpha < 1, & \alpha < 90° \\ -1 \leq \cos\alpha < 0, & \alpha > 90° \end{cases}, \tag{13}$$

Whether AN is in the same direction or not could be determined by calculating the cosine value, i.e., opposite direction for $\cos\alpha > 0$ and same direction for $\cos\alpha < 0$. Due to $D_{\text{AP}} * D_{\text{PN}} > 0$, Eq.14 is followed as :

$$\begin{cases} D_{\text{AP}}^2 + D_{\text{PN}}^2 - D_{\text{AN}}^2 > 0, \text{ when } & \alpha < 90° \\ D_{\text{AP}}^2 + D_{\text{PN}}^2 - D_{\text{AN}}^2 < 0, \text{ when } & \alpha > 90° \end{cases}. \tag{14}$$

Then, the penaty factor η_i is written as a formulation form:

$$\eta_i = -\frac{\gamma_i \left[-D_{\text{AP}}^2 - D_{\text{PN}}^2 + D_{\text{AN}}^2\right]_+ - \left[D_{\text{AP}}^2 + D_{\text{PN}}^2 - D_{\text{AN}}^2\right]_+}{D_{\text{AP}}^2 + D_{\text{PN}}^2 - D_{\text{AN}}^2}, \tag{15}$$

where $\gamma_i < 1$, and $[]_+$ is the hinge function: $[x]_+ = max(0, x)$. By the above equation, when the positive sample P is in the same direction as the negative sample N, the angle α between AP and AN is greater than 90°. $\cos\alpha < 0$ means that $D_{\text{AP}}^2 + D_{\text{PN}}^2 - D_{\text{AN}}^2 < 0$. Then, the $\eta_i = \gamma_i$, or $\eta_i = 1$.

3.2. Adapting Metric Learning Losses with Confusing Triplet Embeddings Learning

In the previous section, the principle of confusing triplet embeddings learning(CTEL) and the penalty factor η_i as shown in Eq. 15 have been explained. In the following, CTEL methods will be applied to the some deep metric learning functions, named as a series of CTEL losses.

3.2.1. Contrastive Loss

Contrastive loss (Hadsell et al., 2006a) is the simplest and intuitive type of metric learning based on sample pairs, the loss is defined as below:

$$L = \sum_{y_{ij}=1} D_{ij} + \sum_{y_{ij}=0} [m - D_{ij}]_+ . \tag{16}$$

In the above equation, m is a fixed threshold value. D_{ij} represents the distance between the samples i and j. $y_{ij} = 1$ means that these two samples belong to the same category; $y_{ij} = 0$ means that these two samples belong to different categories.

By introducing our CTEL, the CTEL-Constrastive loss is:

$$L = \sum_{y_{ij}=1} D_{ij} + \sum_{y_{ij}=0} [m - \eta_i * D_{ij}]_+ . \tag{17}$$

3.2.2. Triplet Loss

Hadsell *et al.*(Hadsell et al., 2006b) proposed a triplet loss as an augmentation over contrastive loss (Hadsell et al., 2006a). Triplet loss selects a pair of positive and negative samples at the same time during the training process. The loss is defined as below:

$$L = \sum [D_{\text{AP}} - D_{\text{AN}} + \alpha]_+, \tag{18}$$

where α is a fixed threshold value, D_{AP} is the distance of positive sample pairs and D_{AN} is the distance of negative sample pairs. $[]_+$ is the hinge function: $[x]_+ = max(0, x)$.

From Eq. 18, the CTEL-Triplet loss is obtained by introducing the penalty factor on confusing triplet embeddings as follows:

$$L = \sum [D_{\text{AP}} - \eta_i * D_{\text{AN}} + \alpha]_+. \tag{19}$$

3.2.3. Quadruplet Loss.

Quadruplet loss (Law et al., 2013) is an extension of triplet loss, which consists of two parts: one part is the normal triplet loss, and the loss term in this part enables the model to distinguish the relative distance between positive and negative sample pairs; the other part is the relative distance between positive sample pairs and any other negative sample pairs, and the loss term in this part can be interpreted as the relative distance between positive sample pairs and any other negative sample pairs regardless of whether these pairs have the same anchor, the minimum inter-class distance is greater than the intra-class distance. The constraint of quadruplet loss is $D_{AP} < D_{AN_1}$ and $D_{AP} < D_{N_1 N_2}$, where D denotes the distance, A is the anchor point, P is the positive sample, and N_1 and N_2 are negative samples with different classes from each other. Thus, the quadruplet loss is defined as follows.

$$L = \sum \left([D_{\text{AP}} - D_{AN_1} + \alpha_1]_+ + [D_{\text{AP}} - D_{N_1 N_2} + \alpha_2]_+ \right), \tag{20}$$

where α_1 and α_2 are manually set parameters, and usually, $\alpha_1 > \alpha_2$.

As shown in Eq. 20, the CTEL-Quadruplet loss is obtained as follows:

$$L = \sum \left([D_{\text{AP}} - \eta_i * D_{an_1} + \alpha_1]_+ + [D_{\text{AP}} - D_{n_1 n_2} + \alpha_2]_+ \right). \tag{21}$$

3.2.4. Multi-Similarity Loss.

Multi-Similarity(MS) loss (Wang et al., 2019a) is implemented by two iterative steps of sampling and weighting which only assigns a weight with an integer value. In this loss, it takes into account the self-similarity and relative similarity where the model is allowed to collect and focus on more useful pairs to improve the model performance. Thus, the loss is defined as below:

$$L = \frac{1}{m} \sum_{i=1}^{m} \left\{ \frac{1}{\alpha} log \left[1 + \sum_{p \in P_i} e^{-\alpha(S_{ip} - \lambda)} \right] + \frac{1}{\beta} log \left[1 + \sum_{n \in N_i} e^{\beta(S_{in} - \lambda)} \right] \right\}. \tag{22}$$

The first part of the log function deals with positive samples. P_i is the set of all positive samples relative to the anchor point i. S_{ip} is the similarity of the positive sample pairs; the remaining part of the log function deals with negative samples. N_i is the set of all negative samples relative to the anchor point i, and S_{in} is the similarity of the negative sample pairs. α, β and λ in Eq. 22 are hyperparameters.

As shown in Eq. 22, the CTEL-MS loss is obtained as follows:

$$\begin{aligned} L = \frac{1}{m}\sum_{i=1}^{m}\Bigg\{\frac{1}{\alpha}log\left[1+\sum_{p\in P_i}e^{-\alpha(S_{ip}-\lambda)}\right] \\ +\frac{1}{\beta}log\left[1+\sum_{n\in N_i}e^{\beta(\eta_i * S_{in}-\lambda)}\right]\Bigg\}. \end{aligned} \tag{23}$$

4. Experiments

4.1. Datasets

We employ two public medical CT datasets(SARS-CoV-2 (Soares et al., 2020) and COVID-CT (Zhao et al., 2020) datasets) and two standard datasets(CUB-200-2011 (Welinder et al., 2010) and Cars-196 (Kingma and Ba, 2014) datasets) for evaluation.

For medical CT datasets, SARS-CoV-2 (Soares et al., 2020) includes 2482 CT images from 120 patients, of which 1252 are positive for COVID-19 and 1230 are non-COVID but have other types of lung disease manifestations. The spatial size of these images ranged from 119 × 104 to 416 × 512. And The COVID-CT dataset (Zhao et al., 2020) includes 349 CT images from 216 patients containing clinical findings of COVID-19 and 397 CT images from 171 patients without COVID-19. The resolution of these images ranges from 102 × 137 to 1853 × 1485.

For fine-grained datasets, CUB-200-2011 dataset is composed of 11,788 images of birds from 200 subclasses. The images selected from 100 classes (5,864 images in total) are used for training, while the last 100 classes (5,924 images in total) are used for testing. Cars-196 dataset is composed of 16,185 images of cars from 196 classes. The images selected from 98 classes are employed for training, with the rest for testing.

4.2. Implementation Details

Embedding network: For the classification task on COVID-CT, the network in (Wang et al., 2020b) is used as our backbone network and the extracted features are normalized (Ioffe and Normalization, 2014). Our CTEL-MS or CTEL-Triplet losses consist of cross-entropy loss and CTEL-based loss, which is similar to the setting of the classification loss in (Wang et al., 2020b) . For image retrieval tasks on standard fine-grained images, our backbone network is a pre-trained Inception (Szegedy et al., 2015) on ILSVRC 2012-CLS (Russakovsky et al., 2015), and the extracted features are normalized (Ioffe and Normalization, 2014).

Image setting: All images were cropped to 224 × 224 and standard preprocessing techniques are applied. Our experiment conducts four-fold cross-validation on COVID-CT datasets according to the method in (Wang et al., 2020b). Besides, for fine-grained standard image datasets, the training set is used with a horizontal flip technique but the test set is employed with a central crop technique. Regarding the sampling procedure used in triplet loss and quadruplet loss, a certain number of classes is randomly selected from each mini-batch, and then 5 images are randomly selected from each class.

Training: For image classification tasks, 60 epochs are chosen for training with batchsize of 32, containing 16 images from each dataset. For all classification tasks, our model is trained for 60 epochs with batchsize of 180. Besides, all experiments are optimized using the Adam (Kingma and Ba, 2014) optimizer.

Hyperparameter setting: The Hyperparametes of the deep metric learning functions including four basic losses and four CTEL-based losses, have the following settings: m in Eq.16 and Eq.17 is set as 1.0. α in Eq.18 and Eq.19 is set as 0.2. α_1 and α_2 in Eq.20 and Eq.21 are set to 0.2 and 0.1, and α, β, and λ are set as 2, 50 and 1 in Eq.22 and Eq.23, respectively. The parameter γ_i is set to 0.8.

4.3. Comparisons on COVID-19 CT datasets

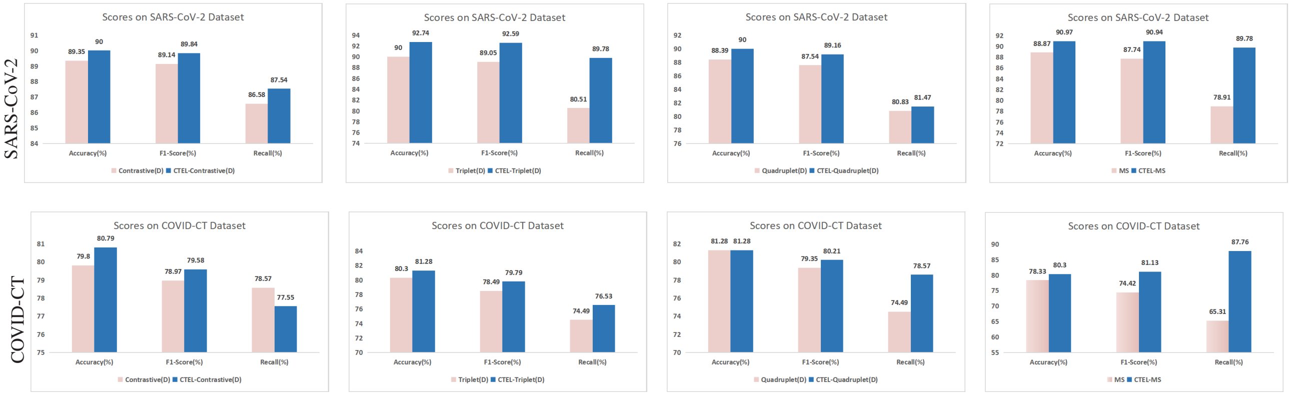
For a fair comparison with other methods, we follow the literature on COVID-19 diagnostics (Soares et al., 2020), we use three metrics for the comprehensive evaluation of the model: Accuracy, F1-Score (Sohn, 2016) and Recall (Jegou et al., 2010).

We further compare the performance of our approach with state-of-the-art techniques on image classification tasks. Results on the SARS-CoV-2 and COVID-CT datasets are summarized in Table 1. Our model outperforms the previous methods, and the CTEL-Triplet method is employed in Distance-Weighted Tuple Mining (Wu et al., 2017) (denoted as CTEL-Triplet(D)).

For all COVID-19 datasets, our CTEL method performs well on both Triplet loss and MS loss. In addition, our CTEL outperforms the original method on other losses, which indicates that CTEL not only works with these basic methods but also improves the classification and retrieval performance of the original deep metric learning methods, as shown in Figure 4.

Table 1: Results of Different Methods on the Two Datasets for COVID-19 CT Image Classification. CTEL denotes our Confusing Triplet Embedding Learning.

	SARS-CoV-2			COVID-CT		
Methods	Accuracy(%)	F1-Score(%)	Recall(%)	Accuracy(%)	F1-Score(%)	Recall(%)
Single(COVID-NET(Wang et al., 2020b;a))	77.12	76.03	70.97	63.12	61.09	57.73
Single(Redesign)(Wang et al., 2020b)	89.09	88.97	83.78	77.07	77.04	74.69
Joint(COVID-NET(Wang et al., 2020b;a))	68.72	69.17	69.41	63.27	59.78	54.19
Joint(Redesign)(Wang et al., 2020b)	78.42	77.86	74.07	69.67	66.89	66.94
Series Adapter(Rebuffi et al., 2017)	85.73	86.19	81.91	70.01	67.08	74.91
Parallel Adapter(Rebuffi et al., 2018)	82.13	82.39	80.02	74.93	73.46	71.81
MS-Net(Hall et al., 2020)	87.98	88.73	84.91	76.23	76.54	74.07
+Contrastive (Wang et al., 2020b)	90.83	90.87	85.89	78.69	78.83	79.71
CTEL-MS(**Ours**)	90.97	90.94	**89.78**	80.30	**81.13**	**87.76**
CTEL-Triplet(D)(**Ours**)	**92.74**	**92.59**	**89.78**	**81.28**	79.79	76.53



(a) Contrastive (b) Triplet(D) (c) Quadruplet(D) (d) MS

Figure 4: Comparisons with the original deep metric learning methods on the two public COVID-19 CT Datasets. CTEL denotes Confusing Triplet Embedding Learning. The pink bar represents the original loss, and the blue bar represents CTEL-based loss.

4.4. Comparisons on standard image datasets

In addition, we apply the CTEL method to standard fine-grained image retrieval tasks. We conduct experiments on two standard datasets: CUB-200-2011 (Welinder et al., 2010) and Cars-196 (Kingma and Ba, 2014). Recall@K (Jegou et al., 2010)is employed as a standard performance metric to evaluate our methods.

We obtain nearly a 4.9% increase in Recall@1 compared to MS Loss and a 13.1% increase in Recall@1 compared to Proxy-NCA with 64 embedding dimensions. Besides, we experiment with a higher embedding dimension of 512. The result demonstrates that our CTEL-MS obtains nearly a 0.6% improvement in Recall@1 over the state-of-the-art MS loss. Compared with DR-MS employing an embedding size of 512 and attention modules, our CTEL-MS achieves a higher Recall@1 by 0.2% improvement at the same dimension.

In Table 2, our CTEL-MS scales well to datasets with 198 classes. In the case of 64 dimensions, CTEL-MS outperforms SoftTriple on Recall@1 by 0.2%. In the case of 512 dimensions, DR-MS, based on MS loss with the direction regularized version, performs better than our CTEL-MS by 0.2% on Recall@1, but worse than our CTEL-MS by 0.2%, 0.3% and 0.3% on Recall@2, Recall@4 and Recall@8, respectively. Our CTEL-MS achieves outstanding results and outperforms others(expect ABE on Recall@1 with 512-dimensional embedding on Cars-196 dataset). On average, CTEL-MS possesses the better performance on CUB-200-2011 and Cars-196 datasets for different embedding on standard metric Recall@K.

4.5. Ablation Study

To further demonstrate the effectiveness of our proposed CTEL method, we compare the basic triplet-based methods with our proposed corresponding CTEL methods on CUB-200-2011 dataset. Triplet Loss and quadruplet loss are employed in Distance-Weighted Tuple Mining (Wu et al., 2017). The dimension of embedding features is set to 512, and the batch size is set to 80. Besides, we use Recall@K as an evaluation metric.

From Table 3, these comparisons illustrate that deep metric learning approaches can generate more discriminative

Table 2: Evaluation on CUB-200-2011 and Cars-196 Datasets. Backbone networks of the models are denoted by abbreviations: G-GoogleNet, BN-Inception with batch normalization, R50-ResNet50.

		CUB-200-2011				Cars-196			
Recall@K(%)		1	2	4	8	1	2	4	8
Clustering^{64}(Oh Song et al., 2017)	BN	48.2	61.4	71.8	81.9	58.1	70.6	80.3	87.8
Proxy NCA^{64}(Movshovitz-Attias et al., 2017)	BN	49.2	61.9	67.9	72.4	73.2	82.4	86.4	87.8
Smart Mining^{64}(Harwood et al., 2017)	BN	49.8	62.3	74.1	83.3	64.7	76.2	84.2	90.2
MS^{64}(Wang et al., 2019a)	BN	57.4	69.8	80.0	87.8	77.3	85.3	90.5	87.8
SoftTriple^{64}(Sohn, 2016)	BN	60.1	71.9	81.2	88.5	78.6	86.6	**91.8**	**95.4**
CTEL-MS^{64}(Ours)	BN	**62.3**	**73.3**	**82.3**	**89.5**	**78.8**	**86.7**	**91.8**	95.2
Margin^{128}(Wu et al., 2017)	R50	63.6	74.4	83.1	90.0	79.6	86.5	91.9	95.1
HDC^{384}(Oh Song et al., 2017)	G	53.6	65.7	77.0	85.6	73.7	83.2	89.5	93.8
ABIER^{512}(Opitz et al., 2018)	G	57.5	68.7	78.3	86.2	82.0	89.0	93.2	96.1
ABE^{512}(Kim et al., 2018)	G	60.6	71.5	79.8	87.4	**85.2**	90.5	94.0	96.1
HTL^{512}(Ge, 2018)	BN	57.1	68.8	78.7	86.5	81.4	88.0	92.7	95.7
HDML^{512}(Zheng et al., 2019)	G	53.7	65.7	76.7	85.7	79.1	87.1	92.1	95.5
RLL^{512}(Wang et al., 2019b)	BN	57.4	69.7	79.2	86.9	74.0	83.6	90.1	94.1
MS^{512}(Wang et al., 2019a)	Bn	65.7	77.0	86.3	91.2	84.1	90.4	94.0	96.5
DR-MS^{512}(Mohan et al., 2020)	G	66.1	77.0	85.1	91.1	84.1	90.4	94.0	96.5
CTEL-MS^{512}(Ours)	BN	**66.3**	**78.2**	**86.2**	**91.8**	84.3	**90.6**	**94.3**	**96.8**

features by exploring confusing triplet embeddings. For instance, using the confusing triplets trends to be pushed farther in the feature space, even when the positive and negative sample pairs have reached a certain margin there.

Table 3: Ablation study to show the effectiveness of our proposed Confusing Triplet Embeddings Learning method when applied to standard metric learning methods on the CUB-200-2011 dataset. CTEL" denotes Confusing Triplet Embedding Learning. '*' indicates a re-implementation of the original version. Backbone networks of the models are denoted by abbreviations: BN-Inception with batch normalization.

Recall@K(%)		1	2	4	8
Contrastive Loss *	BN	63.8	74.7	84.2	90.3
CTEL-Contrastive Loss	BN	**64.5**	**75.7**	**84.6**	**91.1**
Triplet Loss *	BN	65.0	76.5	84.8	91.0
CTEL-Triplet Loss	BN	**65.5**	**76.7**	**85.3**	**91.2**
Quadruplet Loss *	BN	64.8	76.1	84.7	90.9
CTEL-Quadruplet Loss	BN	**65.0**	**76.3**	**85.0**	**91.2**
MS Loss	BN	65.7	77.0	**86.3**	91.2
CTEL-MS Loss	BN	**66.2**	**77.8**	86.2	**91.6**

4.6. Analysis

In summary, our approach achieves good performance on both medical CT datasets (SARS-CoV-2 and COVID-CT datasets) and standard fine-grained datasets (CUB-200-2011 and Cars-196 datasets). The performance of CTEL is based on exploring confusing triplet embeddings in the feature space and establishing more rigorous constraints on the confusing triplet embeddings. Compared with the state-of-the-art methods, our CTEL methods have shown better performance on different metrics, which indicates our method is an effective deep metric learning method for learning discriminative features for image classification and retrieval. Compared with these basic deep metric learning methods, our CTEL as a strategy can not only work together with these basic methods but also improve the retrieval or clustering performance of the original methods. Notably, these experimental results demonstrate the generality and effectiveness of our CTEL.

5. Conclusions

In this work, we propose a new deep metric learning method named CTEL that takes into account the fact that different solutions are determined by the relative positions of the positive and negative samples. That is, in the feature space, the confusing triplet embeddings can influence the performance of models. Therefore, we proposed a method to explore the confusing triplet embeddings and construct a penalty factor on these embeddings, which can continue to push the negative pairs in the same direction farther and make the learned features more discriminative. Our CTEL can perform well as a deep metric learning method, and its strategy can also work together with the basic deep metric learning methods. Experiments on two public medical CT datasets and two standard fine-grained datasets demonstrate the effectiveness of our approach.

Acknowledgement

This work was partially funded by China Postdoctoral Science Foundation: No.2021M690270, and Chaoyang District Postdoctoral Foundation.

References

Abbas, A., Abdelsamea, M. M., and Gaber, M. M. (2021). Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network. *Applied Intelligence*, 51(2):854–864.

Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., and Shao, L. (2020). Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637.

Farooq, M. and Hafeez, A. (2020). Covid-resnet: A deep learning framework for screening of covid19 from radiographs. *arXiv preprint arXiv:2003.14395*.

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

Hadsell, R., Chopra, S., and LeCun, Y. (2006a). 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.

Hadsell, R., Chopra, S., and LeCun, Y. (2006b). 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.

Hall, L. O., Paul, R., Goldgof, D. B., and Goldgof, G. M. (2020). Finding covid-19 from chest x-rays using deep learning on a small dataset. *arXiv preprint arXiv:2004.02060*.

Harwood, B., Kumar BG, V., Carneiro, G., Reid, I., and Drummond, T. (2017). Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2821–2829.

Ioffe, S. and Normalization, C. S. B. (2014). Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Javaheri, T., Homayounfar, M., Amoozgar, Z., Reiazi, R., Homayounieh, F., Abbas, E., Laali, A., Radmard, A. R., Gharib, M. H., Mousavi, S. A. J., et al. (2020). Covidctnet: An open-source deep learning approach to identify covid-19 using ct image. *arXiv preprint arXiv:2005.03059*.

Jegou, H., Douze, M., and Schmid, C. (2010). Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128.

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)*, pages 736–751.

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

Law, M. T., Thome, N., and Cord, M. (2013). Quadruplet-wise image similarity learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 249–256.

Li, X., Li, C., and Zhu, D. (2020). Covid-mobilexpert: On-device covid-19 patient triage and follow-up using chest x-rays. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1063–1067. IEEE.

Mei, X., Lee, H.-C., Diao, K.-y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P. M., Chung, M., et al. (2020). Artificial intelligence–enabled rapid diagnosis of patients with covid-19. *Nature medicine*, 26(8):1224–1228.

Mohan, D. D., Sankaran, N., Fedorishin, D., Setlur, S., and Govindaraju, V. (2020). Moving in the right direction: A regularization for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14591–14599.

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*, pages 360–368.

Narin, A., Kaya, C., and Pamuk, Z. (2021). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, 24(3):1207–1220.

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*, pages 5382–5390.

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*, 42(2):276–290.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. (2018). Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., and Shen, D. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for covid-19. *IEEE reviews in biomedical engineering*, 14:4–15.

Soares, E. A., Angelov, P. P., Biaso, S., Froes, M. H., and Abe, D. K. (2020). Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

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*, pages 1–9.

Wang, L., Lin, Z. Q., and Wong, A. (2020a). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12.

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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030.

Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., and Robertson, N. M. (2019b). Ranked list loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5207–5216.

Wang, Z., Liu, Q., and Dou, Q. (2020b). Contrastive cross-site learning with redesigned net for covid-19 ct classification. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2806–2813.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. (2010). Caltech-ucsd birds 200.

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*, pages 2840–2848.

Xu, P., Hospedales, T. M., Yin, Q., Song, Y.-Z., Xiang, T., and Wang, L. (2022). Deep learning for free-hand sketch: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, P., Huang, Y., Yuan, T., Pang, K., Song, Y.-Z., Xiang, T., Hospedales, T. M., Ma, Z., and Guo, J. (2018). Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8090–8098.

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Ni, Q., Chen, Y., Su, J., et al. (2020). A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, 6(10):1122–1129.

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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824.

Zhang, J., Xie, Y., Pang, G., Liao, Z., Verjans, J., Li, W., Sun, Z., He, J., Li, Y., Shen, C., et al. (2020). Viral pneumonia screening on chest x-ray images using confidence-aware anomaly detection. *arXiv preprint arXiv:2003.12338*.

Zhao, J., Zhang, Y., He, X., and Xie, P. (2020). Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 490.

Zheng, W., Chen, Z., Lu, J., and Zhou, J. (2019). Hardness-aware deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 72–81.