IEICE TRANS. INF. & SYST., VOL.E105–D, NO.4 APRIL 2022

PAPER

Semi-Supervised Representation Learning via Triplet Loss Based on Explicit Class Ratio of Unlabeled Data

Kazuhiko MURASAKI[†a)], **Shingo ANDO**[†], ***and*** **Jun SHIMAMURA**[†], ***Members***

SUMMARY In this paper, we propose a semi-supervised triplet loss function that realizes semi-supervised representation learning in a novel manner. We extend conventional triplet loss, which uses labeled data to achieve representation learning, so that it can deal with unlabeled data. We estimate, in advance, the degree to which each label applies to each unlabeled data point, and optimize the loss function with unlabeled features according to the resulting ratios. Since the proposed loss function has the effect of adjusting the distribution of all unlabeled data, it complements methods based on consistency regularization, which has been extensively studied in recent years. Combined with a consistency regularization-based method, our method achieves more accurate semi-supervised learning. Experiments show that the proposed loss function achieves a higher accuracy than the conventional fine-tuning method.

key words: *semi-supervised learning, representation learning, triplet loss*

1. Introduction

In recent years, deep learning has achieved high performance in machine learning, and a lot of practical applications have been realized by deep learning. It has been applied not only to image classification but also to various problems such as facial expression recognition from images [1], gesture recognition from videos [2], rainfall prediction from weather parameters [3] and so on. One of the major challenges to utilize deep learning for practical applications is that quite a lot of labelled training data is required. Even if a lot of data has already been collected or it is easy to collect data, it is not easy to annotate the data. Recently, the approach of semi-supervised deep learning has been seen as an attractive way of solving this problem. There are various semi-supervised learning methods such as giving pseudo labels to unlabeled data [4], minimizing the entropy of feature vectors by representation learning [5], attaining classifier consistency by adding small changes (noise or perturbation) to unlabeled data [6]–[8], or generating new training data from generative adversarial networks (GAN) [9], and so on. In particular, many methods based on consistency regularization have recently been proposed that train the feature representation of images so that it does not change in response to the processing of images. These have demonstrated high recognition accuracy [6]–[8], [10]–[17]. Consistency regularization methods decrease the influence of noise that is orthogonal to the class information by training the feature representation so that it remains consistent when images corrupted by noise are input. Although the approaches based on consistency regularization have achieved better performance, they do not directly address the distribution of unlabeled data. For example, hypothesis that data which are sufficiently close together are likely to be identically labeled or hypotheses that all unlabeled data are split into a certain number of classes is not utilized by the methods even though prior knowledge about the distribution of the data is clearly important. It is expected that utilizing prior knowledge will further improve the performance of semi-supervised learning.

In this paper, we propose a semi-supervised learning method that directly considers the distribution of unlabeled data, and that can reinforce conventional consistency regularization approaches. The technical contribution of our method is introducing new clues about global distribution to semi-supervised representation learning additional to conventional local distribution assumptions and consistency to noise. To utilize prior knowledge of the global distribution of unlabeled data, we assume that the proportions of each class of data are given in advance. Our proposed loss function uses a predetermined class ratio and make feature representation fit this ratio. In experiments, based on the parameters pretrained by consistency regularization, several loss functions for fine-tuning are evaluated. Compared with ordinary cross entropy loss and neighbor embedding loss [18], it is shown that our proposed loss function achieves better performance.

Figure 1 shows an image of the proposed semi-supervised triplet loss. Conventional triplet loss places the positive samples closer to the anchor than the negative samples, as shown in (a). The proposed objective function extends this idea to evaluate the unlabeled samples. As shown in (b), it makes the unlabeled samples approach the corresponding positive samples by an amount proportional to the ratio of $\boldsymbol{\theta}_i$. As shown in (c), it makes the unlabeled samples approach the corresponding negative samples by an amount proportional to the ratio of $1-\boldsymbol{\theta}_i$. That is, our method utilizes the unlabeled samples for representation learning.

2. Related Work

A number of semi-supervised learning methods based on deep learning have been proposed [19]. There are major two approaches: The consistency based approach, which tries

Manuscript received April 1, 2021.
Manuscript revised October 2, 2021.
Manuscript publicized January 17, 2022.

[†]The authors are with Media Intelligence Laboratories, NTT, Yokosuka-shi, Kanagawa, 239–0847 Japan.
a) E-mail: kazuhiko.murasaki.be@hco.ntt.co.jp

DOI: 10.1587/transinf.2021EDP7073

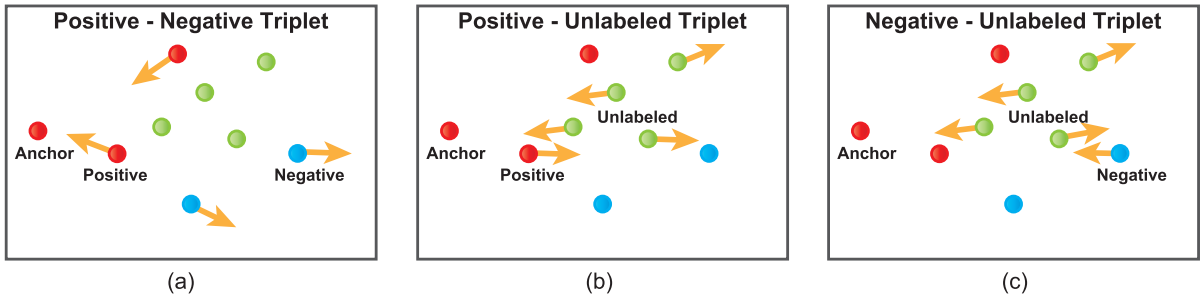

Fig. 1 The conceptual diagram of semi-supervised triplet loss. The semi-supervised triplet loss takes into account three relationships. (a) Positive samples are drawn toward the anchor and negative samples are pushed away from the anchor. (b) Distances of unlabeled samples in the ratio of θ_i from the anchor are drawn toward distances of positive samples, and vice versa. (c) Like (b), distances of unlabeled samples in the ratio of $1-\theta_i$ and distances of negative samples from the anchor are drawn to each other.

to maintain the consistency of representation against data deformation, and the distribution-based approach, which makes assumptions about the distribution of unlabeled data and learns the corresponding deformation of the feature space.

In [19], approaches of semi-supervised deep learning are divided into five categories, generative methods, consistency regularization methods, graph-based methods, pseudo-labeling methods and hybrid methods. We categorize generative methods as one type of consistency-based approach, and categorize pseudo-labeling methods as one type of distribution-based approach. We do not refer to graph-based methods because they are based on graph neural networks whose structure is completely different from general DNNs or CNNs.

2.1 Consistency Based Approach

Many of the semi-supervised learning methods that have been studied in recent years are based on the consistency-based approach, which is a good match with deep learning [6]–[8], [11]–[17]. In this approach, data augmentation is used to generate pseudo-data and the data is assumed to have the same representation or label as the original data. Various data generation methods have been attempted, such as the generation of samples with slight perturbation in the feature space (VAT [6]), generative adversarial networks (GAN [9]), generation of significantly different images via affine and color transformation (UDA [10], SimCLR [15], [16], MoCo [17]), or generation by mixing multiple images (MixMatch [13], ReMixMatch [14]). Although high recognition accuracies have been reported by interweaving various generation methods, the tuning cost is high because the appropriate data generation method must be assumed to vary from problem to problem. Moreover, it is difficult to reproduce the reported recognition accuracy.

Consistency-based learning applied to representation learning from completely unsupervised data is called self-supervised learning. SimCLR [15] and its extensions [16], which attains consistency for strong data augmentation from unlabeled datasets, have been reported to provide highly accurate semi-supervised learning with simple fine tuning. Moreover, although such learning requires a GPU with huge memory, some methods [17], [20] have been proposed to achieve competitive accuracy with reasonable GPU memory requirements.

In this paper, we employ self-supervised learning with SimCLRv2 [16] as the pre-training method. Based on the pre-trained model, we evaluate fine-tuning methods utilizing the distribution cue of unlabeled data to show that the combination of representation consistency and distribution cue improves accuracy further.

2.2 Distribution-Based Approach

Semi-supervised deep learning methods using distributions of unlabeled data have been proposed. In distribution-based approaches, The feature representation is trained first by using an objective function related with the distribution, then the classifier is trained again using the acquired representation. The simplest variant is the clustering-based method, which assumes that the unlabeled dataset has a cluster structure [21]. In addition, the entropy minimization in feature space proposed in [5] assumes that each unlabeled sample belongs to one of the known classes. Neighbor embedding [18] by Hoffer and Ailon assumes that unlabeled data are likely to belong to the same class as the nearest sample in the feature space, and in combination with entropy minimization, their proposed objective function achieved high accuracy without data augmentation. There are some semi-supervised training methods based on pseudo-labeling, which can be considered as another type of distribution-based approach. Pseudo-labeling is the method of learning from pseudo labels which are assigned to unlabeled data by the model in the middle of training. Because pseudo labels are decided by the distance from labeled samples in the feature space, the effect of pseudo labeling is almost like entropy minimization. In the most simple method [4], because all unlabeled samples are labeled based on the classification result, training results are affected by out-of-distribution samples. Shi et al. [22] proposed more sophisticated method, in which they set the confidence to each pseudo-label. The confidence follows the local density of samples in the feature space so that the effect of out-of-

distribution samples is attenuated.

Although these approaches are inferior to the consistency-based approaches in terms of recognition accuracy, they do not require the tuning step of creating pseudo data according to the potential variations of the input data. Note that the distribution-based approach and consistency-based approach can be used in combination because they are complementary.

Neighbor embedding [18] achieves high recognition accuracy by combining entropy minimization with an objective function which assumes that unlabeled data are likely to be labeled similarly to neighboring data. Unfortunately, these assumptions cover only local relationships and the global distribution is ignored. Although Shi's method [22] also achieves high accuracy by combination with consistency-based loss, it only considers local distributions through pseudo-labeling.

The conventional distribution-based approaches focus only on the neighborhood relations of samples, not on how the entire set of unlabeled data is distributed, especially on the proportion of each class included. They ignore the class proportion of unlabeled samples, or implicitly assume that each class is present in equal proportions, as is done for k-means clustering.

Our semi-supervised triplet loss proposal addresses this problem by considering the class ratio of the entire unlabeled data distribution. Using the class ratio of unlabeled data, which is defined or estimated in a preliminary step, the objective function evaluates how the unlabeled data is divided among the classes. Considering local neighborhood relationships as well as global relationships allows us to further improve the accuracy of semi-supervised learning.

2.3 Hybrid Approach

Currently, the state-of-the-art in the semi-supervised image recognition problem is a combination of consistency-based approach and pseudo-labeling, such as FixMatch [11] and CoMatch [12]. FixMatch [11] is based on the combination of consistency-based loss function and pseudo-labeling process. Consistency-based contrastive loss restricts the image feature from changing according to strong data augmentation, and pseudo labels are given by classification scores and thresholding. Although the process flow is a simple combination, the evaluation results are very good. CoMatch [12] is also based on the combination of consistency-based loss and pseudo-labeling. In addition to contrastive loss as like FixMatch, CoMatch evaluates feature embedding using pseudo labels. Using pseudo labels to evaluate feature representation instead of cross entropy loss enhances representation learning and further improves classification accuracy. Although these methods show extremely high performance, there are many parameters to be tuned, such as appropriate data augmentation settings according to the data and threshold settings for pseudo labeling. Complicated parameters make it difficult to reproduce reported accuracy. In addition, all of these parameters are based on assumptions about local data distribution, and not on assumptions about global distribution such as the ratio of each class included in unlabeled data.

In this paper, we show that performance can be improved by assuming a global label distribution from the model pretrained by SimCLRv2 [16] which is a simple but powerful consistency-based method. Moreover, since local distribution-based approaches such as pseudo-labeling and nearest embedding use clues different from the proposed method, further performance improvement is expected by combining these methods.

3. Proposed Method

The proposed method is based on representation learning by triplet loss function [23]. We extend the triplet loss function, which is normally applied only to fully supervised data, so that it can be applied to unsupervised data. We call our proposal semi-supervised triplet loss (SST). We assume the C classes classification problem in this paper, so each sample belongs to one of C classes. Some training data are labeled while the remainder are unlabeled. We denote the labeled sample set belonging to class i as D_i, labeled sample set not belonging to class i as $D_{\setminus i}$, and unlabeled sample set as D^U. Of the unlabeled samples, we denote the sample set that truly belongs to class i as D_i^U. The output of the deep neural network feature extractor is represented by $f(\boldsymbol{x}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ represents the DNN parameters.

3.1 Triplet Loss

Triplet loss [23] is an objective function that can learn feature representation from labeled data. Given model $\boldsymbol{\phi}$ and three input samples $(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n})$, the triplet loss is calculated for the set of three outputs possible. Here, sample $\boldsymbol{a}$ is called an anchor, sample $\boldsymbol{p}$ belongs to the same class as the anchor, while sample $\boldsymbol{n}$ belongs to different class. Feature vectors $f(\boldsymbol{a}, \boldsymbol{\phi})$ and $f(\boldsymbol{p}, \boldsymbol{\phi})$ should be similar because they are in the same class, and conversely $f(\boldsymbol{a}, \boldsymbol{\phi})$ and $f(\boldsymbol{n}, \boldsymbol{\phi})$ should be separated. The triplet loss function that realizes these relationships is written as

$$L_{tri}(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}, \boldsymbol{\phi}) = \max(S(\boldsymbol{a}, \boldsymbol{n}, \boldsymbol{\phi}) - S(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{\phi}) + \alpha, 0) \tag{1a}$$

$$S(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{\phi}) = \langle f(\boldsymbol{a}, \boldsymbol{\phi}), f(\boldsymbol{p}, \boldsymbol{\phi}) \rangle, \tag{1b}$$

where α is a parameter that controls how far the similarity between the same classes and the similarity between different classes should be; $\langle \cdot, \cdot \rangle$ denotes the dot product. If the difference between the similarity is more than α, the loss becomes 0. To minimize the triplet loss, L_{tri} is calculated for every set of three samples and parameter $\boldsymbol{\phi}$ is trained so as to minimize the average of all losses for all classes,

$$L_i^S(\boldsymbol{\phi}) = \mathop{\mathbb{E}}_{\boldsymbol{a} \in D_i} \mathop{\mathbb{E}}_{\boldsymbol{p} \in D_i \setminus \boldsymbol{a}} \mathop{\mathbb{E}}_{\boldsymbol{n} \in D_{\setminus i}} L_{tri}(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}, \boldsymbol{\phi}), \tag{2}$$

$$L^S(\boldsymbol{\phi}) = \mathop{\mathbb{E}}_{i} L_i^S(\boldsymbol{\phi}), \tag{3}$$

where $\mathbb{E}$ means the average value. The model parameter $\boldsymbol{\phi}$, which minimizes L^S, yields good feature extraction.

For simplicity, parameter $\boldsymbol{\phi}$ is omitted from the following equations.

3.2 Semi-Supervised Triplet Loss

If the training data contains unlabeled samples, the conventional triplet loss cannot be applied for representation learning. Since unlabeled dataset D^U could contain any class data, treating all its data simply in the same way means that samples in different classes cannot be separated. Therefore, we assume the ratio of class i in D^U is given by C-dimensional vector $\boldsymbol{\theta}$ ($\sum_i \boldsymbol{\theta}_i = 1$), similar to the positive-unlabeled learning method [24], and try to calculate triplet loss for the unlabeled data based on the assumed ratios. If the labeled samples and unlabeled samples are independent and identically distributed, the class ratio is easily estimated based on the ratio of labeled samples. In this paper, we assume the class ratio of unlabeled data is given or equally distributed.

First, using class i as anchors and calculating triplet loss (Eq. (2)) based on the similarities with the unlabeled data yields the following expression,

$$L_i^{Up} = \mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{p}\in D_i\setminus\boldsymbol{a}} \mathop{\mathbb{E}}_{\boldsymbol{u}\in D^U} L_{tri}(\boldsymbol{a},\boldsymbol{p},\boldsymbol{u}) \tag{4}$$

Here, dataset D^U includes samples of various classes, which are, on average, apportioned in the ratio of $\boldsymbol{\theta}_i$. Considering that unlabeled data belonging to class i are denoted as D_i^U and others as $D_{\setminus i}^U$, Eq. (4) is expressed as follows,

$$\begin{aligned} L_i^{Up} = \mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{p}\in D_i\setminus\boldsymbol{a}} \{\boldsymbol{\theta}_i \mathop{\mathbb{E}}_{\boldsymbol{p}'\in D_i^U} L_{tri}(\boldsymbol{a},\boldsymbol{p},\boldsymbol{p}') \\ + (1-\boldsymbol{\theta}_i) \mathop{\mathbb{E}}_{\boldsymbol{n}'\in D_{\setminus i}^U} L_{tri}(\boldsymbol{a},\boldsymbol{p},\boldsymbol{n}')\}. \end{aligned} \tag{5}$$

Because the labeled sample $\boldsymbol{p}$ and the unlabeled sample $\boldsymbol{p}'$ are originally sampled from the same distribution, the average of the difference in their similarities from the anchor should be close to 0. In addition, since the difference between unlabeled sample $\boldsymbol{n}'$ and sample $\boldsymbol{p}$ is the almost same as the conventional triplet loss, it is desirable that the difference exceeds α. Letting the parameters of the optimized model be $\tilde{\boldsymbol{\phi}}$, the average of each triplet loss becomes

$$\mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{p}\in D_i\setminus\boldsymbol{a}} \mathop{\mathbb{E}}_{\boldsymbol{p}'\in D_i^U} L_{tri}(\boldsymbol{a},\boldsymbol{p},\boldsymbol{p}',\tilde{\boldsymbol{\phi}}) = \alpha, \tag{6}$$

$$\mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{p}\in D_i\setminus\boldsymbol{a}} \mathop{\mathbb{E}}_{\boldsymbol{n}'\in D_{\setminus i}^U} L_{tri}(\boldsymbol{a},\boldsymbol{p},\boldsymbol{n}',\tilde{\boldsymbol{\phi}}) = 0. \tag{7}$$

That is, if the representation for semi-supervised data is optimized, L_i^{Up} becomes $\boldsymbol{\theta}_i\alpha$. In order for training to approach this ideal value, we introduce a loss function that considers unlabeled data as follows

$$\tilde{L}_i^{Up} = |L_i^{Up} - \boldsymbol{\theta}_i\alpha|. \tag{8}$$

Similarly, for a triplet based on the difference between the similarity of an unlabeled sample and the similarity of a different class sample from the anchor, conditional triplet loss is expressed as

$$L_i^{Un} = \mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{u}\in D^U} \mathop{\mathbb{E}}_{\boldsymbol{n}\in D_{\setminus i}} L_{tri}(\boldsymbol{a},\boldsymbol{u},\boldsymbol{n}). \tag{9}$$

Then, dividing D^U into D_i^U and $D_{\setminus i}^U$ yields

$$\begin{aligned} L_i^{Un} = \mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{n}\in D_{\setminus i}} (\boldsymbol{\theta}_i \mathop{\mathbb{E}}_{\boldsymbol{p}'\in D_i^U} L_{tri}(\boldsymbol{a},\boldsymbol{p}',\boldsymbol{n}) \\ + (1-\boldsymbol{\theta}_i) \mathop{\mathbb{E}}_{\boldsymbol{n}'\in D_{\setminus i}^U} L_{tri}(\boldsymbol{a},\boldsymbol{n}',\boldsymbol{n})). \end{aligned} \tag{10}$$

Because it is desirable that the difference between the similarities of labeled $\boldsymbol{n}$ and unlabeled $\boldsymbol{n}'$ becomes 0 on average, and assuming the ideal model parameters $\tilde{\boldsymbol{\phi}}$ in the same way as given by Eq. (6) and Eq. (7), the average of triplet losses becomes

$$\mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{n}\in D_{\setminus i}} \mathop{\mathbb{E}}_{\boldsymbol{p}'\in D_i^U} L_{tri}(\boldsymbol{a},\boldsymbol{p}',\boldsymbol{n},\tilde{\boldsymbol{\phi}}) = 0, \tag{11}$$

$$\mathop{\mathbb{E}}_{\boldsymbol{a}\in D_i} \mathop{\mathbb{E}}_{\boldsymbol{n}\in D_{\setminus i}} \mathop{\mathbb{E}}_{\boldsymbol{n}'\in D_{\setminus i}^U} L_{tri}(\boldsymbol{a},\boldsymbol{n}',\boldsymbol{n},\tilde{\boldsymbol{\phi}}) = \alpha. \tag{12}$$

The loss function for different class and unlabeled data is determined as follows:

$$\tilde{L}_i^{Un} = |L_i^{Un} - (1-\boldsymbol{\theta}_i)\alpha|. \tag{13}$$

We add the loss functions from the unlabeled samples to L^S calculated from the labeled samples. By weighting with γ, semi-supervised triplet loss function L^{SST} is given as follows:

$$L^{SST} = \gamma L^S + (1-\gamma) \mathop{\mathbb{E}}_i \{\tilde{L}_i^{Up} + \tilde{L}_i^{Un}\}. \tag{14}$$

This proposed loss function is optimized by deep learning. Triplets of samples for each class are made using the mini batch approach, and back propagation is applied to minimize the loss function.

3.3 Combination with Other Approaches

Semi-supervised triplet loss utilizes the class ratio of unlabeled data as prior-knowledge of the global distribution. Naturally, optimization based on the assumption of the local distribution and the consistency of feature representation are complementary. We combine our method with pre-training by SimCLRv2 [16] for consistency and neighbor embedding loss [18] to consider the local distribution.

SimCLR [15], [16] is a pretty simple but strong method that can acquire consistency without any supervision. We use open source code by the authors and recommended parameters for CIFAR-10 to train the CNN model. After training by SimCLR, we fine-tune the model by applying the proposed loss function.

Neighbor embedding loss [18] is composed of two

types of entropy minimization. One is for supervised samples and the other is for unsupervised samples. Entropy minimization of supervised samples makes the features in the same class close while separating the features of the different classes. Entropy minimization of unsupervised samples makes their features approach those of the nearest labeled sample. Each loss function is expressed as follows,

$$L_S^{NE} = \mathop{\mathbb{E}}_{i} \mathop{\mathbb{E}}_{z_1 \in D_1} \cdots \mathop{\mathbb{E}}_{z_C \in D_C} \mathop{\mathbb{E}}_{\boldsymbol{x} \in D_i \setminus z_i} \{-\log P(\boldsymbol{x}, i; z_1, \ldots, z_C)\}$$

$$L_U^{NE} = \mathop{\mathbb{E}}_{z_1 \in D_1} \cdots \mathop{\mathbb{E}}_{z_C \in D_Z} \mathop{\mathbb{E}}_{\boldsymbol{u} \in D^U} [-\sum_{i=1}^{C} \{P(\boldsymbol{u}, i; z_1, \ldots, z_C) \cdot \log P(\boldsymbol{u}, i; z_1, \ldots, z_C)\}]$$

$$P(\boldsymbol{u}, i; z_1, \ldots, z_C) = \frac{\exp\{-\|f(\boldsymbol{u}) - f(z_i)\|^2\}}{\sum_{j=1}^{C} \exp\{-\|f(\boldsymbol{u}) - f(z_j)\|^2\}}$$

where L_S^{NE} means supervised entropy minimization and L_U^{NE} means unsupervised entropy minimization. Combining these loss functions yields clusters in feature space while preserving local structure.

We simply add the neighbor embedding loss function to our semi-supervised triplet loss as follows,

$$L = \lambda_1 L^{SST} + \lambda_2 L_S^{NE} + \lambda_3 L_U^{NE}. \tag{15}$$

Each loss function can be weighted by λ_1, λ_2 and λ_3. In the experiments in this paper, we set all λ_i to 1.

4. Experiments

In order to evaluate the performance of the proposed method, semi-supervised multi-class classification was performed on the CIFAR10 [25] dataset, the STL10 [26] dataset and the SVHN [27] dataset. Almost all training samples were unlabeled and we evaluated the classification score using the test samples of each dataset.

4.1 Experimental Setup

In this experiment, we show that our proposed loss function improves accuracy through fine-tuning on a semi-supervised dataset based on a pre-trained model whose consistency was already acquired. Specifically, fine-tuning by semi-supervised triplet loss is performed based on the parameters pre-trained by SimCLRv2. In [16], it is shown that fine-tuning by cross entropy loss realizes semi-supervised learning with sufficiently high accuracy, but we assume that the recognition accuracy can be further improved if the loss function takes account of the class ratio of unlabeled data. At the same time, we show that the proposed method is more effective than conventional approaches [16], [18] based on the distribution of unlabeled data.

With regard to pre-training, open source code of SimCLR [15], [16] was applied and recommended parameters were employed. We used ResNet-18 [28] as the CNN model, and set batch size to 512, dimension size of feature vector to 512, temperature to 0.5, learning rate to 1.0 and number of train epochs to 1000. As the optimization method, LARS [29] was used.

In fine-tuning, learning rate and number of training epochs were heuristically tuned according to the loss function and dataset. As for the other parameters, we set margin α to 1, dimension size of feature vector to 512, and weight γ of Eq. (14) to 0.5. Class ratios $\boldsymbol{\theta}$ was set to the true distribution of the dataset. Last two layers of the model pre-trained by SimCLRv2 are replaced by two randomly initialized dense layers to allow fine-tuning.

To evaluate recognition accuracy, extracted feature expressions were converted into class labels by a kNN classifier, whose parameter K was set to 3. The same amount of labeled samples was randomly selected from each class, and the accuracy was evaluated 5 times while changing the random seed.

4.2 Results

Table 1 shows the performance evaluation results for CIFAR10. The CIFAR10 dataset has 50000 training images and 10000 test images. Labeled samples were randomly chosen from training images, and the remainder of the training set were used as unlabeled training data. Pre-training by SimCLRv2 used all training images. As the number of samples in each class is balanced in CIFAR10, we set the same number for each element of parameter $\boldsymbol{\theta}$. The performance of each method was evaluated using 1000 labeled training samples (100 samples in each class), 400 samples (40 samples in each class), and 40 samples (4 samples in each class). The table shows the average error rate and standard deviation of 5 trials in each condition. All trials used parameters pre-trained by SimCLRv2 [16]. XE means fine-tuning by cross entropy loss using only labeled samples, NE means nearest embedding loss, and SST means semi-supervised triplet loss. SST+NE means combined loss function of NE and SST. The accuracy is improved by fine-tuning with consideration of the distribution of unlabeled data using the model that attained consistency. Semi-supervised triplet loss matched the recognition accuracy of neighbor embedding loss, and moreover, combining them improved the accuracy significantly. This indicates the complementary nature of neighbor embedding loss, which focuses on the local distribution, while our proposed loss function focuses on the global distribution.

Table 2 shows the results for the STL10 dataset, and Table 3 shows the results for the SVHN dataset. The STL10 dataset has 5000 training images, 8000 test images and 100000 unlabeled images. Labeled samples were randomly chosen from the training images, and the remainder and all unlabeled images were used for training as unlabeled data. Table 2 shows that the accuracy of SST is close to that of XE, and the effect of SST seems to be small. This is because true labels of unlabeled images in the STL10 dataset are not given so that parameter $\boldsymbol{\theta}$ is different from the true distribution. On the other hand, when combined with NE, SST+NE achieves slightly better accuracy than NE. Despite

Table 1 Classification accuracy for CIFAR10 dataset. N_L indicates the number of labeled samples.

Loss function	$N_L = 1000$	$N_L = 400$	$N_L = 40$
XE	83.85 ± 0.59%	81.97 ± 0.91%	69.98 ± 2.12%
NE [18]	85.75 ± 0.28%	84.62 ± 0.30%	73.94 ± 2.83%
SST	84.01 ± 0.56%	82.26 ± 0.93%	74.04 ± 1.51%
SST+NE	**87.10 ± 0.42%**	**86.19 ± 0.59%**	**77.19 ± 1.21%**

Table 2 Classification accuracy for STL10 dataset. N_L indicates the number of labeled samples.

Loss function	$N_L = 1000$	$N_L = 400$	$N_L = 40$
XE	80.42 ± 0.36%	79.42 ± 0.53%	69.67 ± 2.70%
NE [18]	82.00 ± 0.14%	80.92 ± 0.44%	71.49 ± 2.53%
SST	80.82 ± 0.34%	78.93 ± 0.42%	69.61 ± 2.80%
SST+NE	**83.28 ± 0.16%**	**81.85 ± 0.28%**	**72.99 ± 2.67%**

Table 3 Classification accuracy for SVHN dataset. N_L indicates the number of labeled samples.

Loss function	$N_L = 1000$	$N_L = 400$	$N_L = 40$
XE	88.44 ± 0.61%	84.49 ± 0.50%	39.80 ± 1.53%
NE [18]	**91.40 ± 0.40%**	**89.01 ± 1.05%**	54.72 ± 13.00%
SST	82.28 ± 1.66%	75.61 ± 1.07%	47.25 ± 8.96%
SST+NE	91.22 ± 0.36%	88.93 ± 0.62%	**73.47 ± 12.72%**

the wrong class ratio parameter, the combination of local distribution and global distribution cues improved the feature representation. The SVHN dataset has 73257 training images, 26032 test images and 531131 extra images. We randomly chose labeled samples from training and extra images. The remainder were used as unlabeled samples. Every image in the SVHN dataset has a ground truth label, so parameter $\boldsymbol{\theta}$ was set to the true distribution of unlabeled data. Table 3 shows that SST+NE significantly improved the accuracy from NE when $N_L = 40$, because the SST loss function addresses the class unbalance in the unlabeled data. When $N_L = 400$ or $N_L = 1000$, the effect of SST seems quite weak. This shows that SST is especially effective when the labeled samples are sparse in the feature space while the unlabeled samples are dense.

4.3 Parameter Sensitivity of $\boldsymbol{\theta}$

In our loss function, parameter $\boldsymbol{\theta}$, which controls the distribution of unlabeled data, is of critical importance. Although parameter $\boldsymbol{\theta}$ is determined by the distribution of labeled data, in actuality, the difference from the true distribution may be significant if the labeled data is very scant. To evaluate the sensitivity of $\boldsymbol{\theta}$, we conducted fine-tuning by the loss function SST+NE with various values of $\boldsymbol{\theta}$. Figure 2 shows the results for the CIFAR10 dataset with 40 labeled samples. X-axis in the figure means cosine similarity from true distribution. Each point denotes the average accuracy using $\boldsymbol{\theta}$ which is in the range of ±0.025 of cosine similarity. Each error bar represents the standard deviation of 10 trials with randomly sampled $\boldsymbol{\theta}$. The two dashed lines denote the accuracies of SST+NE and NE in Table 1. When $\boldsymbol{\theta}$ was set to the true distribution, the average accuracy was 0.77, and without SST loss, the average accuracy became 0.74. Figure 2 shows that the classification accuracy decreases as the cosine similarity decreases, but the reduction is limited. Even if the cosine similarity is less than 0.7, it is still competitive with the average accuracy offered by NE loss (0.74). This is because the penalty to the average of distances (Eq. (8) and Eq. (13)) is not so strict and the feature distribution of the original data is kept.

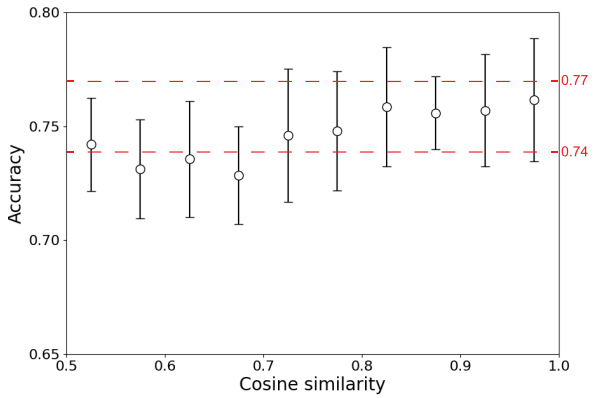

Fig. 2 Parameter sensitivity of $\boldsymbol{\theta}$. X-axis denotes cosine similarity between true distribution and $\boldsymbol{\theta}$, Y-axis denotes classification accuracy of CIFAR10 when 40 samples are labeled.

5. Conclusion

In this paper, we proposed a new method of representation learning that applies triplet loss to the semi-supervised learning problem. The proposal, semi-supervised triplet loss (SST) adjusts the similarity between samples so that unlabeled samples are distributed at an appropriate ratio in the feature space by explicitly assuming the actual ratio of unlabeled data. Experiments showed that the proposed loss function achieved higher accuracy than the conventional method, especially when combined with a loss function that considered the local distribution. Although the results mentioned here are not superior to those of state-of-the-art methods [11], [12], we did show that fine-tuning by combining our semi-supervised triplet loss with neighbor embedding loss can improve the feature representation yielded by pre-trained features created by consistency-based self-supervised learning [16]

References

[1] J.H. Kim, B.G. Kim, P.P. Roy, and D.M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," IEEE Access, vol.7, pp.41273–41285, 2019.
[2] J.H. Kim, G.S. Hong, B.G. Kim, and D.P. Dogra, "deepGesture: Deep learning-based gesture recognition scheme using motion sensors," Displays, vol.55, pp.38–45, 2018.
[3] M. Chhetri, S. Kumar, P. Pratim Roy, and B.G. Kim, "Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan," Remote Sensing, vol.12, no.19, p.3174, 2020.
[4] D. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," Proc. ICML Workshop on challenges in representation learning, p.2, 2013.
[5] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," Advances in neural information processing systems, pp.529–536, 2004.
[6] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," IEEE Trans. PAMI, vol.41, no.8, pp.1979–1993, 2018.
[7] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," Advances in neural information processing systems, pp.3546–3554, 2015.
[8] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," Proc. ICLR, 2017.
[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, pp.2672–2680, 2014.
[10] Q. Xie, Z. Dai, E. Hovy, M. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," arXiv preprint arXiv:1904.12848, 2019.
[11] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. Cubuk, A. Kurakin, and C. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," Advances in Neural Information Processing Systems, 2020.
[12] J. Li, C. Xiong, and S. Hoi, "CoMatch: Semi-supervised learning with contrastive graph regularization," CoRR, vol.abs/2011.11183, 2020.
[13] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," Advances in Neural Information Processing Systems, 2019.
[14] D. Berthelot, N. Carlini, E. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," Proc. ICLR, 2020.
[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," Proc. ICML, pp.1597–1607, 2020.
[16] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," Advances in Neural Information Processing Systems, 2020.
[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," Proc. CVPR, pp.9729–9738, 2020.
[18] E. Hoffer and N. Ailon, "Semi-supervised deep learning by metric embedding," Proc. ICLR workshop, 2017.
[19] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," CoRR, vol.abs/2103.00550, 2021.
[20] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," arXiv preprint arXiv:2003.04297, 2020.
[21] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," Proc. ECCV, pp.139–156, 2018.
[22] W. Shi, Y. Gong, C. Ding, Z. Ma, X. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," Proc. ECCV, pp.311–327, 2018.
[23] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," Proc. CVPR, pp.1386–1393, 2014.
[24] R. Kiryo, G. Niu, M. Du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," Advances in neural information processing systems, pp.1675–1685, 2017.
[25] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Tech. Rep., Computer Science Department, University of Toronto, 2009.
[26] A. Coates, H. Lee, and A. Ng, "An analysis of single-layer networks in unsupervised feature learning," Proc. AISTATS, pp.215–223, 2011.
[27] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng, "Reading digits in natural images with unsupervised feature learning," NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. CVPR, pp.770–778, 2016.
[29] Y. You, I. Gitman, and B. Ginsburg, "Large batch training of convolutional networks," arXiv preprint arXiv:1708.03888, 2017.

Kazuhiko Murasaki He received the B.E. from the University of Tokyo in 2009 and Master of Information Science and Technology in 2011. He joined NTT in 2011 and has been engaged in research on image recognition, especially about semantic segmentation, boundary detection and degradation detection of infrastructure. He received Doctor of Engineering from Tokyo University of Science in 2019. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Shingo Ando is a senior research engineer in Cyber-World Laboratory of NTT Human Informatics Laboratories He received the B.E. degree in electrical engineering from Keio University in 1998. He received the Ph.D. degree in engineering from Keio University in 2003. In 2003, he joined NTT. He has been engaged in research and practical application development in the fields of image processing, pattern recognition, and digital watermarks. He is a member of IEICE, ITE, and IIEEJ.

Jun Shimamura He received a B.E. in engineering science from Osaka University in 1998 and an M.E. and Ph.D. from Nara Institute of Science and Technology in 2000 and 2006. He joined NTT Cyber Space Laboratories in 2000. He is currently senior research engineer, supervisor of scene analysis technology at NTT Human Informatics Laboratories, Japan. His research interests include computer vision and mixed reality.