Neurocomputing 453 (2021) 731–741

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Attention-based label consistency for semi-supervised deep learning based image classification

Jiaming Chen [a], Meng Yang [a,b,*], Jie Ling [a]

[a] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[b] Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China

ARTICLE INFO

Article history:
Received 18 February 2020
Revised 11 June 2020
Accepted 28 June 2020
Available online 9 September 2020
Communicated by Wenguan Wang

2010 MSC:
00-01
99-00

Keywords:
Semi-supervised learning
Deep neural network
Attention mechanism
Imbalance classification

ABSTRACT

Semi-supervised deep learning, which aims to effectively use the available unlabeled data to aid the model in learning from labeled data, is a hot topic recently. To effectively employ the abundant unlabeled data and handle the imbalance in labeled data, we propose a novel attention-based label consistency (ALC) model for semi-supervised deep learning. The relationships between different samples are well exploited by the proposed scheme of channel and sample attention; meanwhile, the class estimations are required to be smooth for nearby unlabeled data. The proposed ALC is further extended to the imbalanced case by developing a label-imbalance ALC model. We have implemented the proposed ALC model in the semi-supervised frameworks of Π model and MeanTeacher, and the experimental results on four benchmark datasets, (e.g., Fashion-MNIST, CIFAR-10, SVHN, and ImageNet) clearly show the advantages of our proposed method.

1. Introduction

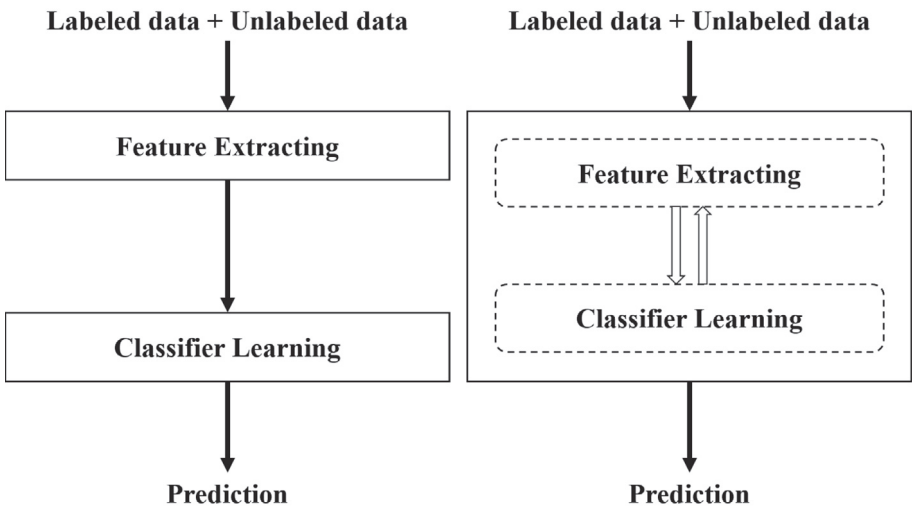
Deep Learning has achieved tremendous performance in various computer vision tasks. One of the major bottlenecks of deep learning [1–3] is how to obtain large-scale labeled datasets. Due to high labor costs for labeling samples and abundant unlabeled data that can be easily collected, semi-supervised learning has attracted a lot of attention.

The semi-supervised learning, aiming to make models perform as well on small-scale labeled datasets as on large-scale labeled datasets by effectively employing the labeled data and unlabeled data, is a natural way to solve the problem of learning from limited labeled data. Blum and Chawla [4] proposed the co-training algorithm which uses two classifiers trained from two data view to improve each other. However, the requirement of two sufficient and redundant data views limits its application. Zhou and Li [5] relaxed the requirement by constructing three re-sampled datasets and training three different classifiers with the re-sampled datasets. Graph-based semi-supervised learning [6] constructs the graph which can capture the relationships among data and propagates the label information from labeled data to unlabeled data through the graph. Schölkopf et al. [7] proposed a laplacian regularization to prevent the classifier from overfitting the graph. Semi-supervised support vector machine [8,9] utilizes unlabeled data to reduce the structural risk and encourage the decision boundary to pass through the low-density regions of unlabeled data. Shrivastava et al. [10] proposed a semi-supervised dictionary learning algorithm which uses an iterative process to learn a dictionary and estimates the class of unlabeled data. Yang and Chen [11] integrated the unlabeled samples into semi-supervised dictionary to enhance the discrimination of classifier. However, above traditional learning methods rely heavily on the handcrafted features, which can not automatically adapt to the tasks.

To alleviate this problem, recent researches have increased efforts on semi-supervised deep learning that can effectively explore the available labeled and unlabeled data together. The relationship between traditional semi-supervised learning and semi-supervised deep learning is shown in Fig. 1. One intuitive way to use massive unlabeled data is to pre-train deep neural network models in an unsupervised way, and then fine-tune the models on labeled data [12–14]. Although these methods have been able

* Corresponding author at: School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China.
E-mail addresses: chenjm26@mail2.sysu.edu.cn (J. Chen), yangm6@mail.sysu.edu.cn (M. Yang), lingj8@mail2.sysu.edu.cn (J. Ling).

https://doi.org/10.1016/j.neucom.2020.06.133
0925-2312/



(a) traditional semi-supervised learning (b) semi-supervised deep learning

Fig. 1. The overviews of traditional semi-supervised learning and semi-supervised deep learning.

to simultaneously conduct both supervised and unsupervised learning, the discrimination of unlabeled data can not be well exploited in the unsupervised pre-training stage. For instance, most of the information required for data reconstruction in unsupervised learning way is irrelevant for classification tasks of supervised learning.

Exploiting unlabeled data in the entire training stage instead of just pre-training has been developed in recent works [15–17]. In Pseudo-Label [15], both the labeled and the unlabeled data are used to train the deep neural networks. It utilizes the model trained by labeled data to predict class of unlabeled data as their labels, and then combines unlabeled data with predicted labels and labeled data to train another model. This method improves the performance of semi-supervised deep learning significantly. Rasmus et al. [16] trained deep neural network with all available data and avoided layer-wise pre-training by skipping connections from encoders to decoders. The skipping connections can retain as much information as possible about the unlabeled data so that full-supervised deep learning and semi-supervised deep learning are combined well. The Manifold Tangent Classifier (MTC) [17] trains contrastive auto-encoders (CAEs) approximately invariant to local perturbations along the manifold. And some works [18,19] combine the semi-supervised learning and Generative Adversarial Networks (GAN) by making the discriminator not only distinguish real image or generated image but also recognize the class label of real and generated images. Although improvements have been reported by these methods, how to effectively exploit the discrimination embedded in the unlabeled data is still an open question.

In recent years, consistency-based semi-supervised deep learning methods have achieved outstanding performance. Π-Model [20] introduces an unsupervised loss called consistency regularization to encourage the L2 distance between an unlabeled data with two different perturbations to be small. Here the perturbations can be viewed as dropout, data augmentation etc. In this case, Π-Model can resist the perturbation of data and learn more essential information of the data, making the model more robust. However, Π-Model needs to make two predictions on the same sample data, which is time consuming. To reduce the training time, Temporal Ensembling [20] was proposed to utilize the target of last epoch as current target. MeanTeacher [21] constructs a teacher model to generate higher quality target by using the moving average weight of student model as the teacher model. Another work, VAT [22], rewards the average of local distributional smoothness (LDS) over all the training samples and minimizes the KL-distance based measurement to enhance the model's robustness against perturbation. The idea behind them is that a "good" classification model should give consistent predictions for the same data points under small perturbations. However, these methods only focus on the smoothness around each single point, while ignore the relationships between data points. Although SNTG [23] proposed a cluster-like regularization to enforce the representation of neighbors to be smooth, the designed distance of samples is too simple to handle the complex sample relationships well.

Although semi-supervised learning has been well exploited in balanced image classification, another important problem for semi-supervised deep learning based imbalanced classification is rarely researched. As a common problem in semi-supervised deep learning, imbalanced classification is that the amounts of samples in different classes are largely different, which negatively affects the performance of training. For example in imbalanced image classification, the number of images of one class is larger 10 times than image numbers of other classes in a dataset. If the dataset is trained by semi-supervised learning methods, the predicting accuracy of the class having more images would be largely higher than other classes, i.e., the learned model over-fits the class.

In this paper, we propose a novel attention-based label consistency (ALC) for semi-supervised deep learning, which fully utilizes the structural information among data points. In the proposed attention-based label consistency regularization, the similarity of different samples is described by using channel and sample attention, with which the label consistency of different samples is preserved. By enhancing the smoothness of label prediction among data, the network can be trained to better and better utilize the structural information about data. We further extend our research by developing a labeled imbalanced ACL model. Extensive experiments of balanced and imbalanced image classification are conducted on four image datasets, with improved results reported compared to the state-of-the-art Π model [20] and MeanTeacher [21]. Our contributions are summarized as follows:

- We propose to combine two attention mechanisms, i.e., channel attention and sample attention, to capture the structure information and relationships between different samples.

- With integrating channel and sample attentions, we design a novel attention-based label consistency (ALC) to boost the performance of semi-supervised methods efficiently.
- We extend our research to the label-imbalance scenario and develop a label-imbalance ALC model.
- Our experimental results on the benchmark datasets show the effectiveness and advantages of label-balance and label-imbalance ALC models.

2. Related Work

In the past, many semi-supervised deep learning methods [20, 21,15, 22] have been developed. In this section, we focus on some closely related works, especially the network consistency loss in Π model [20] and MeanTeacher [21], the pooling operation, the attention scheme, and imbalanced semi-supervised deep learning.

2.1. Network consistency loss

We consider the case that a data set $\boldsymbol{A}$ consists of limited labeled data and large-scale unlabeled data. The network function f, parameterized by $\boldsymbol{\theta}$, represents the prediction distribution. The network consistency loss in Π model [20] is defined as

$$l_{nc} = \sum_{j=1}^{|\boldsymbol{A}|} ||f(\boldsymbol{x}_j, \boldsymbol{\theta}, \eta) - f(\boldsymbol{x}_j, \boldsymbol{\theta}, \eta')||_2^2 \tag{1}$$

where $\boldsymbol{x} \in \boldsymbol{A}$ and $|\boldsymbol{A}|$ is the number of whole data set, η and η' are different noise regularizations. This loss encourages the network function f to produce similar predictions on the sample under different noise regularizations.

Tarvainen et al. [21] further improved Π model and proposed a better model, called teacher model, by averaging the weights of student model in every iteration:

$$\boldsymbol{\theta}' = \alpha\boldsymbol{\theta}' + (1-\alpha)\boldsymbol{\theta} \tag{2}$$

where $\boldsymbol{\theta}'$ is the weights of teacher model and α is a hyperparameter. In MeanTeacher [21], the network consistency loss, l_{nc}, takes the form of the following equation

$$l_{nc} = \sum_{j=1}^{|\boldsymbol{A}|} ||f(\boldsymbol{x}_j, \boldsymbol{\theta}, \eta) - f(\boldsymbol{x}_j, \boldsymbol{\theta}', \eta')||_2^2 \tag{3}$$

where $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ are the weights of teacher model and student model, respectively. And the teacher model is an average of consecutive student models via Eq. (2), which enables MeanTeacher to learn from unlabeled data through more accurate targets generated by the teacher model $f(\boldsymbol{x}_j, \boldsymbol{\theta}', \eta')$.

For further boosting the performance of MeanTeacher [21], Li et al. [24] designed a certainty-driven consistency loss (CCL) based on Eq. (3). The targets with low confidences can be filtered by the CCL, which measures the variance of multiple targets under random augmentations and dropout and discards the uncertainty targets by using the measurement as an criterion of uncertainty. With filtering the targets with low confidences, the CCL makes the training processing more stable and convergence more quick.

2.2. Pooling

Pooling which reduces the feature dimension and retains the primary feature has been well studied. Max-pooling [25] and sum/average-pooling are the two most common methods. Max-pooling extracts the max value from each candidate window. As for sum-pooling, it takes average of the sum of all values in each candidate window. Although the two pooling method are simple, they are efficient and effective. Generalized max-pooling [26] is an extension of max-pooling, which introduces equalization of similarity between each patch and the pooled representation to learn features. Then a selective, discriminative and equalizing pooling by three specific equations is proposed in [27], with its experimental result outperforming other popular pooling methods. For convenience, we apply max-pooling and sum-pooling in our model.

2.3. Attention

Recently, attention mechanism has been successfully applied to various tasks, such as image classification [28–30], monocular depth prediction [31], eye-tracking [32] and object segmentation [33]. SENet [28] benefits from enhancing channel relationship through the squeeze-and-excitation block which adaptively recalibrates channel-wise features. The squeeze-and-excitation block actually performs a dynamic channel-wise feature recalibration and reweights the feature representation to find the strong discriminative features in channel. Residual Attention Network [29] learns the attention-aware features by stacking attention modules, then the learned features from different modules change adaptively as layers go deeper. Attention can be viewed as a tool to bias the allocation of available processing resources and extract the most informative components of an input signal. Wang et al. [34] proposed a novel self-attention, called non-local operation, which is defined as follow

$$\hat{\boldsymbol{x}}_i = \frac{1}{Z(\boldsymbol{x})} \sum_{\forall j} f_{pair}(\boldsymbol{x}_i, \boldsymbol{x}_j) h(\boldsymbol{x}_j) \tag{4}$$

where i and j are the positions, $\boldsymbol{x}$ is the input signal, $\hat{\boldsymbol{x}}$ is the output signal, and $Z(\boldsymbol{x})$ is the factor for normalization. In addition, there are two functions in Eq. (4), where the function of f_{pair} computes a relationship scalar (i.e., affinity) between i and j, and the function h computes a representation of $\boldsymbol{x}$ at the position j. The non-local operation uses a weighted sum of the features as the outputs and captures the long-range dependencies between any two positions. However, the non-local operation can not exploit the discrimination in channel and relationships between samples.

2.4. Imbalanced semi-supervised deep learning

In order to solve the imbalanced image classification problem, some methods are proposed. In the approaches of data processing, oversampling [35] and undersampling [36] are often used to balance the dataset. Oversampling enhances the image number of minority classes by copying the images or using data augmentation, but it may cause over-fitting. If the dataset is large enough, undersampling can be used to balance the dataset by deleting the images belonging to the majority classes, but the valuable information may be omitted. In the approaches related to learning, a variety of algorithms are used to change the weights of classes. For example, the weights of majority classes are small and the weights of minority classes are large. In [37], a novel method called graph-based rebalance semi-supervised learning is proposed to overcome the imbalanced problem. It uses k-nearest neighbors and Gaussian Kernel weighting algorithm to generate a graph based on a weighted sparse adjacency matrix, and then builds a loss based on classification and normalized label variable to distribute the weights of classes. However, the number of class is so small that the method can not be ensured its efficiency on large number of class.

3. Method

Deep semi-supervised learning methods with the network consistency loss, e.g. MeanTeacher [21], have achieved remarkable results in image classification. However, these methods only consider each data point alone, while ignore the relationships between data points. In order to fully utilize the structural information among data, we propose a novel method, attention-based label consistency (ALC) for semi-supervised deep learning, which can capture the data relationships through the introduced attention mechanisms and the designed consistency regularization. Compared to SNTG [23], the proposed attention-based label consistency is a more flexible and powerful method to exploit the information among training samples.

3.1. General model

The framework of deep convolutional neural network for recognition tasks can be divided into two parts: deep feature learning and classification loss. By directly using the deep feature learning part, we propose a general model of attention-based label consistency (ALC) to handle both balanced and imbalanced classifications. Given a data set $\boldsymbol{A}$ containing K classes, $\boldsymbol{L} \subseteq \boldsymbol{A}$ is the data sub-set with known labels. Let $p(\boldsymbol{x}, \boldsymbol{\theta})$ denote the outputted label estimation of multi-layer perception layer before the Softmax layer (denoted by $\sigma(\cdot)$) for the sample $\boldsymbol{x} \in \boldsymbol{L}$ and $f(\boldsymbol{x}, \boldsymbol{\theta}, \eta)$ is the output of deep network. Then the general ALC is defined as follow:

$$l_{general} = -\sum_{j=1}^{|\boldsymbol{L}|} g(\boldsymbol{x}_j) \sum_{i=1}^{K} y_{ji} \log \sigma\left(p(\boldsymbol{x}_j, \boldsymbol{\theta})\right)_i + \sum_{j=1}^{|\boldsymbol{A}|} g(\boldsymbol{x}_j) \| f(\boldsymbol{x}_j, \boldsymbol{\theta}, \eta) - \boldsymbol{z} \|_2^2 + w_2 l_{alc} \tag{5}$$

where $\boldsymbol{\theta}$ is the network parameter, $|\boldsymbol{L}|$ is the number of labeled data, w_2 is a scalar parameter, $\sigma(\cdot)_i$ is the softmax probability of the i^{th} class, and l_{alc} is the proposed attention-based label consistency regularization. The first term and second term in Eq. (5) are the supervised classification loss and the network consistency loss, respectively. When the $\boldsymbol{x}_j$ belongs to k^{th} class, then $y_{jk} = 1$ and $y_{jl} = 0$ if $l \neq k$. The $\boldsymbol{z}$ in second term can be $f(\boldsymbol{x}, \boldsymbol{\theta}, \eta')$ in Eq. (1) or $f(\boldsymbol{x}_j, \boldsymbol{\theta}', \eta')$ in Eq. (3). The function $g(\cdot)$ is a weight function, which can dynamically calibrate the importance of sample and adapt the ALC model to both balanced and imbalanced classifications.

3.2. Balanced ALC

In this section, we propose a label-balance version of attention-based consistency (ALC). For balanced classification, we assign all training samples to the equal weights. Therefore, the weight function $g(\cdot)$ in Eq. (5) is set to 1 for all samples. The loss function of balanced ALC model is defined as

$$l_{general} = -\sum_{j=1}^{|\boldsymbol{L}|} \sum_{i=1}^{K} y_{ji} \log \sigma\left(p(\boldsymbol{x}_j, \boldsymbol{\theta})\right)_i + w_1 \sum_{j=1}^{|\boldsymbol{A}|} \| f(\boldsymbol{x}_j, \boldsymbol{\theta}, \eta) - \boldsymbol{z} \|_2^2 + w_2 l_{alc} \tag{6}$$

where the w_1 for second term is set according to the previous methods, such as Π model and Meanteacher.

The overview of the proposed ALC model is shown in Fig. 2. If the $\boldsymbol{z} = f(\boldsymbol{x}, \boldsymbol{\theta}, \eta')$ in Eq. (1), the student network is the same as the teacher network. If $\boldsymbol{z} = f(\boldsymbol{x}_j, \boldsymbol{\theta}', \eta')$ in Eq. (3), the teacher network is constructed from student network via Eq. (2). The proposed ALC model utilizes the available labeled and unlabeled data together, where the supervised classification is learned from the labeled data only and the network consistency loss enhances the smoothness only for a single data point. The proposed attention-based label consistency regularization aims to capture the relationships among data by attention mechanisms.

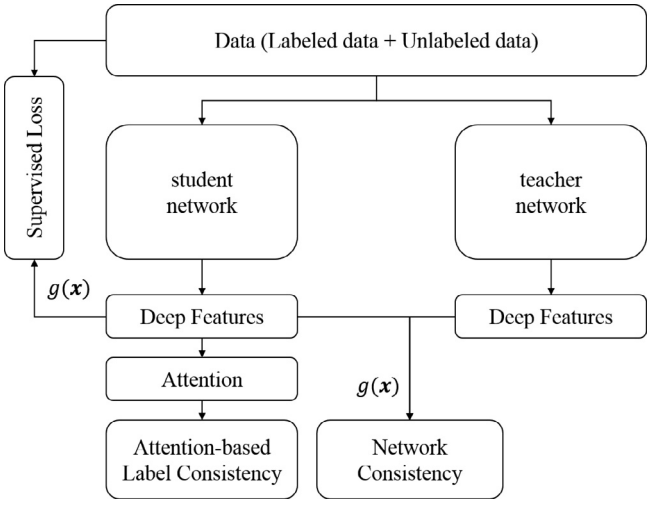


Fig. 2. Framework of the proposed attention-based label consistency (ALC) model.

The structural information among data is introduced by the designed attention-based label consistency. More specifically, the relationships between different data points are described by attention schemes; while the label estimations of data points are constrained by the attention-based regularization.

3.3. Channel and sample attention

The inter-channel relationships of features embed the essential information of convolutional filters, which benefits to make the network learn a better feature representation. In this paper, the squeeze-and-excitation attention [28] is used to capture the channel attention

$$\boldsymbol{U} = MLP(AvgPool(\boldsymbol{X})) \tag{7}$$

where $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_B] \in R^{B \times H \times W \times C}$ is an input feature. B and C are the batch size and channel size, respectively. And $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_B] \in R^{B \times C}$ is a matrix whose element $\boldsymbol{u}_i$ represents the channel excitation vector for $\boldsymbol{X}_i$. Then the feature $\boldsymbol{X}$ will be reweighted by the channel attention map $\boldsymbol{U}$

$$\hat{\boldsymbol{X}}_i = \boldsymbol{X}_i + \boldsymbol{u}_i * \boldsymbol{X}_i \tag{8}$$

where $\hat{\boldsymbol{X}} = \left[\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2, \ldots, \hat{\boldsymbol{X}}_B\right]$, a residual connection between reweighted feature $\boldsymbol{u}_i * \boldsymbol{X}_i$ and original feature $\boldsymbol{X}_i$ is employed, and the second term is a channel-wise multiplication between the feature $\boldsymbol{X}_i \in R^{H \times W \times C}$ and the vector $\boldsymbol{u}_i$, which is defined as follow

$$\boldsymbol{u}_i * \boldsymbol{X}_i = \left[u_i^1 \cdot \boldsymbol{X}_i^1; \ldots; u_i^c \cdot \boldsymbol{X}_i^c; \ldots; u_i^C \cdot \boldsymbol{X}_i^C\right] \tag{9}$$

where the scalar u_i^c is the element of $\boldsymbol{u}_i$ in the c^{th} channel. Correspondingly, $\boldsymbol{X}_i^c \in R^{H \times W}$ is the element of $\boldsymbol{X}_i$ in the c^{th} channel. The flowchart of squeeze-and-excitation channel attention is shown in Fig. 3.

After channel attention, we conduct sample attention to measure the similarities between different samples, as shown in Fig. 4. By using both global-average-pooled and global-max-pooled features to capture the information simultaneously, we can generate two new feature representations $\bar{\boldsymbol{X}}$ and $\tilde{\boldsymbol{X}}$ via

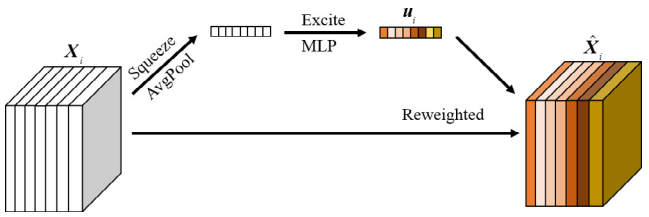


Fig. 3. Squeeze-and-excitation channel attention.

$$\bar{\boldsymbol{X}} = MLP\left(GAvgPool\left(\hat{\boldsymbol{X}}\right)\right) \tag{10}$$

$$\tilde{\boldsymbol{X}} = MLP\left(GMaxPool\left(\hat{\boldsymbol{X}}\right)\right) \tag{11}$$

where *MLP* is the multi-layer perception layer, *GAvgPool* and *GMaxPool* represent global-average-pooling and global-max-pooling operations, which can gather different important image clues [38]. We use GAvgPool to capture the global information and GMaxPool to capture the local information within a convolution window. With back-propagation algorithm [39], the *GAvgPool* and *GMaxPool* map the feature vector into the complementary feature spaces which capture different kinds of semantic information, i.e., global information and local information within a convolution window. Finally, the shared *MLP* would project the two semantic information from *GAvgPool* and *GMaxPool* into the same feature space.

For measuring the similarities of samples, we use a softmax normalization to compute the relationships among the samples of $\bar{\boldsymbol{X}}$ and $\tilde{\boldsymbol{X}}$

$$s_{ji} = \frac{exp\left(\tilde{\boldsymbol{X}}_i \cdot \bar{\boldsymbol{X}}_j\right)}{\sum_{i=1}^{B} exp\left(\tilde{\boldsymbol{X}}_i \cdot \bar{\boldsymbol{X}}_j\right)} \tag{12}$$

where s_{ji} measures the i^{th} data point's impact on j^{th} data points. Here we compute the Eq. (12) across features of $\bar{\boldsymbol{X}}$ and $\tilde{\boldsymbol{X}}$ in order to capture different image information more completely. The more similar feature representations of the two data points are, the greater the correlation between them is.

3.4. Attention-based label consistency regularization

The neighbor-similarity matrix $\boldsymbol{S}$, whose element s_{ji} is generated by the attention mechanisms, actually builds a graph to incorporate neighborhood information among data. It is quite reasonable that the samples close to each other should have similar labels. Thus the proposed attention-based label consistency is designed as

$$l_{alc} = \sum_{i,j} s_{ji} ||p(\boldsymbol{x}_i, \boldsymbol{\theta}) - p(\boldsymbol{x}_j, \boldsymbol{\theta})||_2^2 \tag{13}$$

where the $\boldsymbol{x}_i$ is some labeled or unlabeled sample inputted to the model and $\boldsymbol{\theta}$ is the weights of network parameters. The objective function with adaptive neighborhood information s_{ji} incurs penalty if the neighboring feature representations are classified to be far apart. Therefore, the minimization of Eq. (13) is an attempt to ensure that the features that are close to each other have similar label predictions.

3.5. Label-Imbalance ALC

In this section, we extend our ALC model to a rarely discussed case, label-imbalance semi-supervised learning. In traditional semi-supervised learning, the number of labeled data in each class is equal. However, in practical application, it is difficult to obtain labeled data classes of equal size. Therefore we develop a label-imbalance semi-supervised learning, in which the numbers of labeled data in different classes are unequal, i.e., the labeled data are class imbalanced.

In order to solve the label-imbalance semi-supervised learning problem, we propose an extended version of ALC, called label-Imbalance ALC. Specifically, for handling the labeled imbalance data, we adopt the class percentage of labeled data in each class to calibrate the loss function.

Given a data set $\tilde{\boldsymbol{A}}$, which contains labeled imbalance data $\tilde{\boldsymbol{L}}$ with K classes. Denote $\tilde{\boldsymbol{y}} \in \tilde{\boldsymbol{Y}}$ the corresponding labeled vector, whose element $\tilde{y}_{jk} = 1$ means $\boldsymbol{x}_j$ belongs to the k^{th} class. Let vector $\boldsymbol{r}$ be the class percentage of labeled data in each class. The element r_i of $\boldsymbol{r}$ denotes the percentage of i^{th} class. Due to the availability of labeled data, the vector $\boldsymbol{r}$ can be easily obtained in advance. Then we proposed a hard version of label-imbalance ALC, which is defined as follow:

$$\begin{aligned} l_{hard-imbalance} = &-\sum_{j=1}^{|\tilde{\boldsymbol{L}}|}\left(1 - \boldsymbol{r}_{c_j}\right)\sum_{i=1}^{K}\tilde{y}_{ji}\log\sigma\left(p(\boldsymbol{x}_j,\boldsymbol{\theta})\right)_i \\ &+ w_1\sum_{j=1}^{|\tilde{\boldsymbol{A}}|}\left(1 - \boldsymbol{r}_{c'_j}\right)||f(\boldsymbol{x}_j,\boldsymbol{\theta},\eta) - \boldsymbol{z}||_2^2 \\ &+ w_2\sum_{i,j}^{|\tilde{\boldsymbol{A}}|} s_{ji}||p(\boldsymbol{x}_i,\boldsymbol{\theta}) - p(\boldsymbol{x}_j,\boldsymbol{\theta})||_2^2 \end{aligned} \tag{14}$$

where $|\tilde{\boldsymbol{L}}|$ and $|\tilde{\boldsymbol{A}}|$ are the number of labeled imbalance data $\tilde{\boldsymbol{L}}$ and the number of data set $\tilde{\boldsymbol{A}}$, respectively. $p(\boldsymbol{x},\boldsymbol{\theta})$ is the output of multi-layer perception layer and $f(\boldsymbol{x},\boldsymbol{\theta},\eta)$ is the output of the deep network. In the loss of labeled imbalance data (i.e., the first term in Eq. (14)), $c_j = k$ is the label scalar of sample $\boldsymbol{x}_j$, where $\tilde{y}_{jk} = 1$. In the second term in Eq. (14), $c'_j = \arg\max[\boldsymbol{z}]$ is the predicted label of teacher network for sample $\boldsymbol{x}_j$. The $\boldsymbol{z}$ in second term can be $f(\boldsymbol{x},\boldsymbol{\theta},\eta')$ in Eq. (1) or $f(\boldsymbol{x}_j,\boldsymbol{\theta}',\eta')$ in Eq. (3). However, the hard classification that definitely distributes the label c'_j to the unlabeled sample $\boldsymbol{x}_j$ based the prediction of teacher network doesn't consider the confidence of classification. For reducing the risk of misclassification, we propose a soft version of label-imbalance ALC by weighting the loss of unlabeled samples with their probabilities. The soft version of label-imbalance ALC is defined as

$$\begin{aligned} l_{soft-imbalance} = &-\sum_{j=1}^{|\tilde{\boldsymbol{L}}|}\left(1 - \boldsymbol{r}_{c_j}\right)\sum_{i=1}^{K}\tilde{y}_{ji}\log\sigma\left(p(\boldsymbol{x}_j,\boldsymbol{\theta})\right)_i \\ &+ w_1\sum_{j=1}^{|\tilde{\boldsymbol{A}}|} p_j\left(1 - \boldsymbol{r}_{c'_j}\right)||f(\boldsymbol{x}_j,\boldsymbol{\theta},\eta) - \boldsymbol{z}||_2^2 \\ &+ w_2\sum_{i,j}^{|\tilde{\boldsymbol{A}}|} s_{ji}||p(\boldsymbol{x}_i,\boldsymbol{\theta}) - p(\boldsymbol{x}_j,\boldsymbol{\theta})||_2^2 \end{aligned} \tag{15}$$

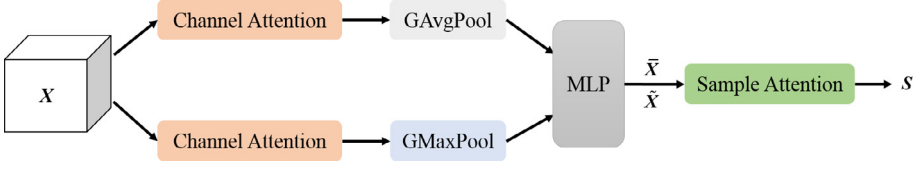


Fig. 4. Similarity measures. After obtained channel-attention features, we use the sample attention to compute the similarities between samples..

where p_j is the probability obtained from the prediction of teacher network. Specifically, for unlabeled sample $\boldsymbol{x}_j$, we set the probability $p_j = \max(\boldsymbol{z})$. For labeled sample $\boldsymbol{x}_j$, we set $p_j = 1$.

The idea behind Eq. (14) and Eq. (15) is using the class percentage vector $\boldsymbol{r}$ to calibrate the loss function. The labeled imbalance data would bias the deep network to wrong prediction, i.e., the model is easily biased to the classes which contain more labeled samples, so we correct the imbalance by weighting the loss function. Because the model is easily biased to the classes which contain more labeled samples, we reduce the weights of classes with more labeled samples and increase the weights of classes with less labeled samples according to the class percentages.

4. Network optimization

In this section, we will present the optimization of attention-based label consistency, which can capture the neighborhood information adaptively by attention mechanisms. The attention mechanisms, implemented by fast vector or matrix operator in software toolkit (e.g., Math kernel library (MKL) proposed by Intel), only cost ignorable computational burden compared to semi-supervised method without attention mechanisms. In this paper, we use a 13-layer convolutional network architecture, named CNN-13, showed in Table 1. Because we treat the weights $g(\cdot)$ in Eq. (6) and Eq. (14) as constants, the optimizations of Eq. (6) and Eq. (14) are similar. The overall framework of ALC model is shown in Algorithm 1.

Algorithm 1 The proposed framework of ALC model.

Input: Balanced training Data set $\boldsymbol{A}$ or Imbalanced training Data set $\widetilde{\boldsymbol{A}}$.
Output: Network parameter $\boldsymbol{\theta}$
1: Extract deep feature representations $\boldsymbol{X}$.
2: Feed $\boldsymbol{X} \in R^{B\times H\times W\times C}$ into channel attention layer to generate the feature representation $\hat{\boldsymbol{X}}$.
3: Apply global-average-pooling and global-max-pooling operations on $\hat{\boldsymbol{X}}$ to generate $\bar{\boldsymbol{X}} \in R^{B\times C}$ and $\bar{\boldsymbol{X}} \in R^{B\times C}$.
4: Compute the neighbor-similarity matrix $\boldsymbol{S}$ via Eq. (12).
5: Evaluate Eq. (13) with $\boldsymbol{S} = \{s_{ij}\}$ and network predictions.
6: Compute the network consistency via Eq. (1) or Eq. (3).
7: Compute the supervised classification loss in Eq. (6), Eq. (14), or Eq. (15).
8: Compute the overall loss in Eq. (6), Eq. (14), or Eq. (15) and update the network parameter $\boldsymbol{\theta}$ via mini-batch SGD.

After computing the attention-based label consistency term, the optimization of the network can be conducted in the framework of existing deep models, such as Π model [20] and MeanTeacher [21].

By incorporating the attention-based label consistency via Eq. (13), the existing deep models can not only enhance the smoothness around each single data point, but also exploit the structural information among the data points. The Fig. 5 shows the results of ALC model on CIFAR-10 dataset. In Fig. 5, we can observe that the proposed regularization clearly improves the performances on both MeanTeacher and Π model. Although our model seems a bit worse than the consistency-based model (i.e. MeanTeacher or Π model) at the beginning, our model gradually overtakes the consistency-based model as the proposed regularization learns the structural information among data. The attention-based label consistency only needs $\emptyset\left(B^2\right)$ computational complexity and B is the mini-batch size which is small compared with the whole dataset. Therefore, the ALC model only costs little additional computational burden compared with the baselines (i.e., MeanTeacher and Π model). In inference, given a test sample, we use the the posterior probability of deep neural network as the final prediction output of our model.

Table 1
CNN-13 architecture.

Block	Layer	Hyperparameters
1^{st} block	Convolutional	128 filters, 3×3, pad = 'same', LReLU($\alpha = 0.1$)
	Convolutional	128 filters, 3×3, pad = 'same', LReLU($\alpha = 0.1$)
	Convolutional	128 filters, 3×3, pad = 'same', LReLU($\alpha = 0.1$)
	Pooling	2×2 Maxpool
	Convolutional	256 filters, 3×3, pad = 'same', LReLU($\alpha = 0.1$)
	Convolutional	256 filters, 3×3, pad = 'same', LReLU($\alpha = 0.1$)
	Convolutional	256 filters, 3×3, pad = 'same', LReLU($\alpha = 0.1$)
	Pooling	2×2 Maxpool
3^{rd} block	Convolutional	512 filters, 3×3, pad='valid', LReLU($\alpha = 0.1$)
	Convolutional	256 filters, 1×1, pad = 'same', LReLU($\alpha = 0.1$)
	Convolutional	128 filters, 1×1, pad = 'same', LReLU($\alpha = 0.1$)
	Pooling	Average pool → 1×1
	Fully-connected	
	Softmax	

5. Experiment

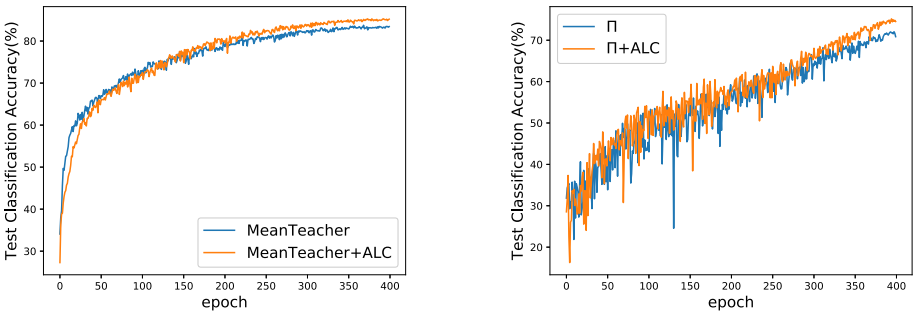
In this section, we conduct some experiments on benchmark datasets, Fashion-MNIST [40], CIFAR-10, SVHN [41], ImageNet [42] to evaluate the proposed method and then discuss the experimental results, which are averaged over 5 runs with different seeds for data splits. In most experiments, we implement the proposed ALC method with aforementioned CNN-13 architecture. And we also evaluate the performance of the proposed ALC with a different architecture, i.e., Wide ResNet architecture [43]. We then conduct the experiments on imbalance settings to investigate the effectiveness of the label-imbalance ALC method. In the end, we conduct an ablation study to validate the proposed regularization.

We use the public codes[1] of MeanTeacher [21] and reimplement the Π model [20], and try our best to reproduce them. We use SGD with a mini-batch size of 100 and train the model for 400 epoches. The learning rate, using a cosine annealing schedule, starts from 0.1 in initialization. Further, for comparison, we implement the Pseudo-Label [15] with the same 13-layer convolutional neural network in the same experimental setting to the original paper [15]. We also report the results of SNTG [23] by running the public code[2] for a fair comparison. Following the public codes, the smoothing coefficient hyperparameter α of MeanTeacher is set 0.97 in all experiments and w_1 is set to {5.0, 100.0, 5.0} on Fashion-MNIST, CIFAR-10 and SVHN, respectively.

5.1. Discussion of key parameters

Since w_1 has been used in previous models, we detail the setting of the other hyperparameter w_2 in Eq. (6) as follows. The hyperparameter w_2 is used to control the importance of the proposed l_{alc}. The validation set was used for evaluating the hyperparameter. Fig. 6 shows the experiments on SVHN dataset with 1000 labeled data to investigate the influence of w_2 for our model. In this experiment, we fix $w_1 = 5.0$ and vary w_2 to learn different models whose verification accuracies on SVHN dataset are shown in Fig. 6. The results shows that the proposed regularization clearly improve the performance of model as the weight w_2 increases. The stable improving is observed across a wide range of w_2, which shows that the performance is not sensitive to the choice of w_2. If no specific

[1] https://github.com/CuriousAI/mean-teacher
[2] https://github.com/xinmei9322/SNTG



(a) MeanTeacher+ALC (b) Π+ALC

Fig. 5. Test classification accuracy on CIFAR-10. (a) The blue line is the results of MeanTeachner and the orange one is MeanTeacher + ALC. (b) The blue line is results of Π and the orange one is Π+ALC.

instructions are given, the value of w_2 is set as 1 in all the following experiments.

5.2. Fashion-MNIST

Fashion-MNIST is comprise of 28×28 grayscale images of 70000 fashion products from 10 categories, with 7000 images per category. The training set has 60000 images and the test set has 10000 images. For experiment on Fashion-MNIST, we use the standard 10000 testing samples as a held out test set and all 60000 training samples as train set. We randomly select {25, 50} samples from each class in the train set as the labeled data. The labeled batch size is set to 10.

The results are shown in Table 2. In both 250-labels and 500-labels experiments on Fashion-MNIST, the proposed MeanTeacher + ALC and Π model + ALC outperform the MeanTeacher and Π model, respectively. In 250-labels experiment, the Π model + ALC achieves the best performance and the accuracy of Π model + ALC is 2.38% higher than Π model. In 250-labels experiment, our MeanTeacher + ALC and Π model + ALC outperform all other methods.

5.3. CIFAR-10

CIFAR-10 benchmark contains 32×32 pixel RGB images belonging to ten different classes. CIFAR-10 consists of 50000 training samples and 10000 test samples. For semi-supervised setting, we randomly select 100 samples from each class in the train set as the labeled data. Thus the size of labeled data set is 1000. For CIFAR-10, we set $w_2 = 100.0$ in the experiments and the labeled batch size is 30.

The result is shown in Fig. 7, from which we can observe that the proposed regularization clearly improve the performances on both MeanTeacher and Π model. The MeanTeacher + ALC outperforms other methods in the experiment.

5.4. SVHN

Here we ran experiments using the Street View House Numbers (SVHN). SVHN contains 32×32 pixel RGB images belonging to ten different classes as CIFAR-10. In SVHN, each example is close-up of a house number, and the class represents the identity of the digit at the center of the image. The SVHN dataset contains of 73257 training samples and 26032 test samples. For semi-supervised setting, we randomly select {25, 50, 100} samples from each class in the train set as the labeled data. For SVHN, we set the labeled batch size as 3.

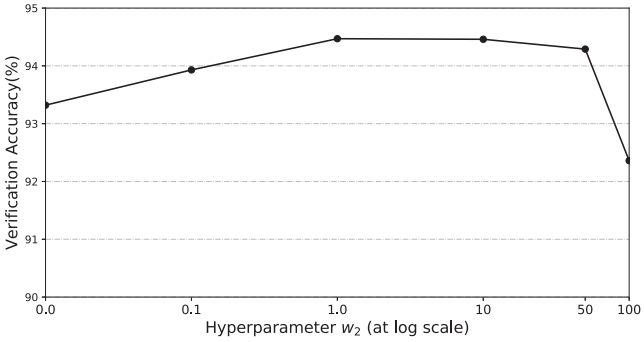


Fig. 6. The verification accuracy on different w_2 at log scale.

Table 2
Test classification rate % on Fashion-MNIST.

Method	250	500
Supervised	77.64 ± 0.34	81.55 ± 0.47
Pseudo-Label	77.70 ± 0.68	81.75 ± 0.52
SNTG [23]	81.92 ± 0.27	84.40 ± 0.20
MeanTeacher [21]	78.61 ± 0.14	84.92 ± 0.21
MeanTeacher + ALC	80.14 ± 0.30	**85.43** ± **0.17**
Π model [20]	80.21 ± 0.60	83.26 ± 0.53
Π model + ALC	**82.59** ± 0.43	**84.82** ± **0.40**

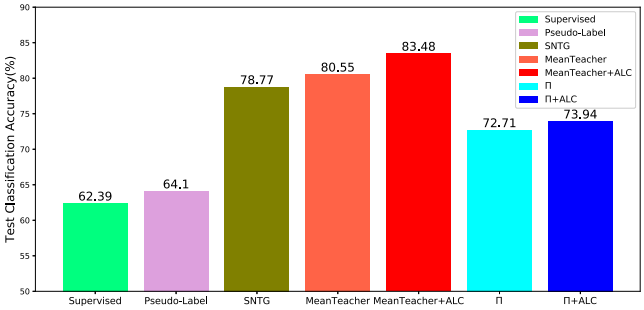


Fig. 7. Test classification accuracy on CIFAR-10 with 1000 labeled data.

The results in Table 3 show the significant improvements of the proposed method, e.g. from 91.41% to 93.14% and from 92.13% to 93.47% for Π model+ALC on SVHN with 250 and 350 labeled data, respectively. The proposed method outperforms other semi-supervised method clearly. And on 250-labels experiments, the SNTG can not get convergence, so we don't report it.

5.5. ImageNet

ImageNet [44] is a large scale dataset, including more than one million 224×224 RGB images. The training set of ImageNet contains totally 128 k images with 1000 classes. And the validation set consists of 50 k images, with an average of 50 pictures in each class. For semi-supervised setting, we randomly select 1% of images from the training set as labeled data. We train this model 400 epochs on 3 GTX 1080ti GPUs. We use ResNet-18 [3] as backbone in this experiment. The batch size is set to 240, half of which is labeled data and the other half is unlabeled data.

In Table 4, Mean Teacher + ALC can achieve the highest accuracy of both Top-1 and Top-5. Compared to Mean Teacher, our method can boost Top-1 accuracy of 0.42% and Top-5 accuracy of 0.56%. With attention-based label consistency regularization, Π model + ALC is 0.32% higher than Π model on Top-1. Our method does not improved much on 1% ImageNet, because the amount of labeled image in 1% ImageNet is 12 for each class, which is too few for training feature extractor to generate more representative features. And our attention module can not be more effective based on these features. If our method perform on 10% ImageNet, it will enhance better performance, but the computation and time cost are too large. It proves that attention-based label consistency regularization also can work well on large scale dataset.

5.6. Wide ResNet on CIFAR-10 and SVHN

In this experiment, we evaluate our proposed method with the widely used Wide ResNet [43] as backbone architecture on CIFAR-10 and SVHN datasets. Specifically, The Wide ResNet consists of 28 layers with width 2, called WRN-28-2, including batch normalization [45] and Leaky ReLU [46]. We evaluate the performances of the WRN-28-2 version of MeanTeacher + ALC and Π model + ALC on CIFAR-10 and SVHN. For CIFAR-10 dataset, we randomly select 400 training samples from each class as labeled data, remaining training samples as unlabeled. And for SVHN, we randomly select 100 training samples from each class as labeled data.

The experiment settings of WRN-28–2 version are the same with aforementioned settings of CNN-13. We compare our method with WRN-28–2 version of MeanTeacher [21], Π model [20], VAT [22] and Pseudo-Label [15] from [47]. The experimental results are reported in Table 5.

As shown in Table 5, our MeanTeacher + ALC model beats all comparative methods and achieves the best performances. The classification accuracy of MeanTeacher + ALC on CIFAR-10 experiment significantly outperforms other semi-supervised methods, 2.91% higher than the next-best-performing method (VAT + EntMin). With attention-based label consistency regularization, the MeanTeacher + ALC are 5.65% higher than MeanTeacher on CIFAR-10 dataset. And the Π model + ALC significantly improves Π model from 83.63% to 86.59% on CIFAR-10 experiment and from 92.81% to 94.12% on SVHN experiment. The significant improvements show the advantages and effectiveness of the proposed method.

5.7. Experiment on labeled imbalance data

In this experiment, we conduct the experiments to evaluate two proposed label-imbalance ALC, i.e., hard version and soft version of label-imbalance ALC, on CIFAR-10 and fashion-MNIST datasets. We randomly select 3000 CIFAR-10 training samples and 1000 Fashion-MNIST training samples to construct the labeled imbalance data, whose maximum imbalance ratios are 1:5, 1:10 and 1:15 respectively. And the remaining training data is used as unlabeled data. We use the 13-layer convolutional neural and the hyper-parameter settings are the same with Section 5.3. The experimental results are showed in Table 6 and Table 7.

Table 3
Test classification rate % on SVHN.

Method	250	350	500
Supervised	48.5 ± 0.31	58.32 ± 0.41	74.99 ± 0.57
Pseudo-Label	74.22 ± 0.80	81.71 ± 0.74	85.79 ± 0.43
SNTG [23]		93.14 ± 1.21	93.64 ± 0.77
MeanTeacher [21]	93.18 ± 0.46	94.55 ± 0.23	95.51 ± 0.12
MeanTeacher + ALC	**94.98 ± 0.34**	**95.70 ± 0.27**	**96.07 ± 0.21**
Π model [20]	91.41 ± 0.39	92.13 ± 0.31	92.86 ± 0.26
Π model + ALC	**93.14 ± 0.27**	93.47 ± 0.17	94.59 ± 0.11

Table 4
Test classification rate % on 1% ImageNet.

Method	Top-1	Top-5
Supervised	14.13	30.52
Pseudo-Label	16.24	33.97
Π model	16.79	33.93
Mean Teacher	16.92	34.21
Π model + ALC	17.07	34.61
Mean Teacher + ALC	**17.34**	**34.77**

Table 5
Test classification rate % on WRN-28-2.

Method	CIFAR-10	SVHN
Supervised	79.74 ± 0.38	87.17 ± 0.47
Pseudo-Label [15]	82.22 ± 0.57	92.38 ± 0.29
VAT [22]	86.14 ± 0.27	94.37 ± 0.20
VAT + EntMin [22]	86.87 ± 0.39	94.65 ± 0.19
MeanTeacher [21]	84.13 ± 0.28	94.35 ± 0.47
MeanTeacher + ALC	**89.78±0.72**	**94.83±0.21**
Π model [20]	83.63 ± 0.63	92.81 ± 0.27
Π model + ALC	86.59 ± 0.55	94.12 ± 0.43

From Table 6, we can observe that as the labeled data become more imbalanced, the performances of all models are worse. That is said that the labeled imbalance data heavily affect the classification performances. And the classification performances of Pseudo-Label and Π model are worse than the Supervised model. It is suggested that the labeled imbalance data prevent the semi-supervised methods from learning from unlabeled data correctly. With the imbalanced version of ALC, both MeanTeacher + ALC(hard) and MeanTeacher + ALC(soft) outperform other methods and achieve the best performance in all imbalance cases. The classification rates of MeanTeacher + ALC(hard) are 4.46%, 5.02% and 4.69% higher than MeanTeacher on 1:5, 1:10 and 1:15 experiments, respectively. Moreover, with the probability weights, MeanTeacher + ALC(soft) significantly improves the performances of MeanTeacher and MeanTeacher + ALC(hard). Especially in 1:15 case, MeanTeacher + ALC(soft) achieves 86.69% accuracy, 3.25% and 7.94% higher than MeanTeacher + ALC(hard) and MeanTeacher, respectively. And both Π model + ALC(hard) and Π model + ALC(soft) greatly improve the performance of Π model and outperform the Supervised model and Pseudo-Label.

From Table 7, we can see that the imbalance of labeled data aggravates the performances of semi-supervised methods, especially Pseudo-Label and Π model. With hard and soft version of ALC, our methods solve effectively the label-imbalance data and improve the semi-supervised methods, i.e., Π model and MeanTeacher. And MeanTeacher + ALC(soft) and Π model + ALC(soft) show the best performances in all cases. The experimental results in Table 6 and Table 7 verify that the proposed hard version and soft version of label-imbalance ALC can calibrate the imbalance of

Table 6
Test classification rate % on CIFAR-10 labeled imbalance data.

Method	1:5	1:10	1:15
Supervised	67.58	60.64	56.56
Pseudo-Label [15]	65.08	55.03	45.90
MeanTeacher [21]	83.02	80.91	78.75
MeanTeacher + ALC (hard)	**87.48**	**85.93**	**83.44**
MeanTeacher + ALC (soft)	**88.13**	**87.69**	**86.69**
Π model [20]	53.15	47.47	47.69
Π model + ALC (hard)	78.26	66.88	61.77
Π model + ALC (soft)	83.13	79.55	74.63

Table 7
Test classification rate % on Fashion-MNIST labeled imbalance data.

Method	1:5	1:10	1:15
Supervised	79.93	73.14	71.82
Pseudo-Label [15]	74.63	71.52	69.15
MeanTeacher [21]	82.19	79.70	79.32
MeanTeacher + ALC (hard)	**83.76**	**81.85**	**81.66**
MeanTeacher + ALC (soft)	**84.20**	**82.13**	**81.10**
Π model [20]	75.38	59.89	52.02
Π model + ALC (hard)	80.72	79.31	76.48
Π model + ALC (soft)	83.71	79.55	74.63

labeled data and improve performances of the semi-supervised methods.

5.8. Ablation study

In this experiment, we validate our proposed regularization with the ablation study on the attention mechanisms (sample/channel) and pooling methods (GMaxPool/GAvgPool). Table 8 lists the results on benchmarks of SVHN with 250 labeled samples and CIFAR-10 with 1000 labeled samples. We use MeanTeacher as baseline and analyze: a) MeanTeacher + ALC with GMaxPool only. b) MeanTeacher + ALC with GAvgPool only. c) MeanTeacher + ALC with sample attention only. d) MeanTeacher + ALC. e) MeanTeacher model. It is noted that there is no MeanTeacher + ALC with channel attention only. That is because without sample attention, the neighbor-similarity matrix $\boldsymbol{S}$ of ALC in Eq. (13) can not be computed. As shown in Table 8, the similarity of GMaxPool and GAvgPool features exploits more information than single pooling only. The sample attention improves MeanTeacher on both SVHN and CIFAR-10 experiments. For CIFAR-10, the complex samples benefit from inter-channel information captured by channel attention. Further more, we study the effect of different pair-wise similarity without attention mechanisms. Specifically, we compare our attention-based label consistency with three pair-wise similarities, including sample similarity, gaussian similarity, and cosine similarity. And we also conduct the experiment with average similarity by averaging the similarity with GMaxPool only and similarity with GAvgPool only, named average(GMaxPool, GAvgPool). For evaluating another type of pooling method, we replace GMaxPool in our model with Generalized Max-pooling [26], named MeanTeacher + (Generalized Max-pooling, GAvgPool). Let $\boldsymbol{X}$ be the mini-batch feature matrix whose element $\boldsymbol{X}_i$ is the feature representation of i^{th} sample and B be the batch size. Then above three similarities are defined as follows.

Sample similarity:

$$s_{ji} = \frac{exp(\boldsymbol{X}_i \cdot \boldsymbol{X}_j)}{\sum_{i=1}^{B} exp(\boldsymbol{X}_i \cdot \boldsymbol{X}_j)} \tag{16}$$

Gaussian similarity:

$$s_{ji} = exp\left(-\frac{\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2}{t}\right) \tag{17}$$

Cosine similarity:

$$s_{ji} = \frac{\boldsymbol{X}_i \cdot \boldsymbol{X}_j}{\|\boldsymbol{X}_i\|_2 \|\boldsymbol{X}_j\|_2} \tag{18}$$

Table 9
Test classification rate % on CIFAR-10.

Method	CIFAR-10
MeanTeacher [21]	80.55 ± 0.43
MeanTeacher + sample similarity	81.52 ± 0.21
MeanTeacher + gaussian similarity	81.43 ± 0.77
MeanTeacher + cosine similarity	78.26 ± 1.21
MeanTeacher + average(GMaxPool, GAvgPool)	82.53 ± 0.33
MeanTeacher + (Generalized Max-pooling, GAvgPool)	84.10 ± 0.37
MeanTeacher + ALC	83.48 ± 0.21

We use the MeanTeacher as the baseline model and study the effect of similarities in CIFAR-10 with 1000 labeled samples. We find the hyperparameter t in Eq. (17) with validation dataset and set $t = 0.06$. The results are shown in Table 9. From Table 9, we can see that capturing the relationship between samples improve the semi-supervised baseline obviously. And improvement of sample similarity is better than other similarities without mechanisms. And in the experiment, the proposed ALC outperforms all similarities without attention mechanisms. It is suggest that, with two attention mechanisms, the similarity from ALC captures more discriminative information.

6. Conclusion

In this paper we propose a novel attention-based label consistency regularization term that can effectively exploit the relationships among the data to improve the model performance. Specifically the regularization term encourages the smoothness of label prediction among data by model the structural information. Moreover, the designed regularization captures the relationships among data through two attention mechanisms, channel attention and sample attention, which can model the structural

Table 8
Test classification rate % on Ablation Study.

Method			SVHN	CIFAR-10
	GMaxPool	GAvgPool		
a	√		94.19 ± 0.11	82.24 ± 0.27
b		√	94.45 ± 0.25	82.81 ± 0.39
	sample	channel		
c	√		94.31 ± 0.24	82.22 ± 0.20
d	√	√	94.98 ± 0.34	83.48 ± 0.21
e			93.18 ± 0.46	80.55 ± 0.43

information automatically through back-propagation algorithm. We further extended our research to label-imbalance classification and proposed an imbalanced ACL model. The experimental results of balanced and imbalanced ALC model on four benchmark datasets have shown that our methods can improve the accuracy of image classification compared to recent semi-supervised deep learning models.

CRediT authorship contribution statement

Jiaming Chen: Conceptualization, Investigation, Methodology, Software, Validation, Writing-original draft, Writing - review & editing. **Meng Yang:** Conceptualization, Investigation, Methodology, Supervision, Writing-original draft, Writing - review & editing. **Jie Ling:** Investigation, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is partially supported by National Natural Science Foundation of China (Grants No. 61772568), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515012029), the Guangzhou Science and Technology Program (Grant No. 201804010288), the Fundamental Research Funds for the Central Universities (Grant No. 18lgzd15), and Guangdong Special Support Program.

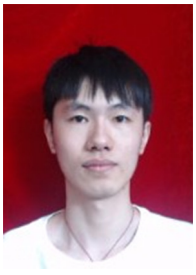
References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, Curran Associates Inc., USA, 2012, pp. 1097–1105..
[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[4] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the Annual Conference on Computational Learning Theory, COLT' 98, ACM, New York, NY, USA, 1998, pp. 92–100. doi:10.1145/279943.279962..
[5] Zhi-Hua Zhou, Ming Li, Tri-training: exploiting unlabeled data using three classifiers, IEEE Trans. Knowl. Data Eng. 17 (11) (2005) 1529–1541, https://doi.org/10.1109/TKDE.2005.186.
[6] A. Blum, S. Chawla, Learning from labeled and unlabeled data using graph mincuts, in: Proceedings of the International Conference on Machine Learning (ICML), ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 19–26.
[7] B. Schölkopf, J. Platt, T. Hofmann, Learning on Graph with Laplacian Regularization, MIT Press, 2007.
[8] T. Joachims, Transductive inference for text classification using support vector machines, in: Proceedings of the International Conference on Machine Learning (ICML), ICML '99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 200–209.
[9] K.P. Bennett, A. Demiriz, Semi-supervised support vector machines, in: Proceedings of the International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, 1999, pp. 368–374.
[10] A. Shrivastava, J.K. Pillai, V.M. Patel, R. Chellappa, Learning discriminative dictionaries with partially labeled data, in: IEEE International Conference on Image Processing, 2013, pp. 3113–3116.
[11] M. Yang, L. Chen, Discriminative semi-supervised dictionary learning with entropy regularization for pattern classification, in: Proceedings of AAAI Conference on Artificial Intelligence, 2017, pp. 1626–1632..
[12] A.M. Dai, Q.V. Le, Semi-supervised sequence learning, in: Advances in Neural Information Processing Systems 28, Curran Associates Inc, 2015, pp. 3079–3087..
[13] H. Wu, S. Prasad, Semi-supervised deep learning using pseudo labels for hyperspectral image classification, IEEE Trans. Image Processing 27 (3) (2018) 1259–1270.
[14] I.J. Goodfellow, M. Mirza, A. Courville, Y. Bengio, Multi-prediction deep boltzmann machines, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1, NIPS'13, Curran Associates Inc., USA, 2013, pp. 548–556..
[15] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: ICML 2013 Workshop: Challenges in Representation Learning (WREPL), 2013..
[16] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, T. Raiko, Semi-supervised learning with ladder network, Advances in Neural Information Processing Systems..
[17] S. Rifai, Y.N. Dauphin, P. Vincent, Y. Bengio, X. Muller, The manifold tangent classifier, in: Proceedings of the 24th International Conference on Neural Information Processing Systems, Curran Associates Inc., USA, 2011, pp. 2294–2302.
[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, X. Chen, Improved techniques for training gans, Adv. Neural Inform. Processing Systems 29 (2016) 2234–2242.
[19] Z. Dai, Z. Yang, F. Yang, W.W. Cohen, R.R. Salakhutdinov, Good semi-supervised learning that requires a bad gan, in: Advances in Neural Information Processing Systems 30, Curran Associates Inc, 2017, pp. 6510–6520..
[20] S. Laine, T. Aila, Temporal ensembling for semisupervised learning, in: In Proceedings of ICLR, 2017..
[21] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semisupervised deep learning results, Adv. Neural Inform. Process. Sys. (2017) 1195–1204.
[22] T. Miyato, S. Maeda, S. Ishii, M. Koyama, Virtual adversarial training: A regularization method for supervised and semi-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2018) 1979–1993, https://doi.org/10.1109/TPAMI.2018.2858821.
[23] Y. Luo, J. Zhu, M. Li, Y. Ren, B. Zhang, Smooth neighbors on teacher graphs for semi-supervised learning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018..
[24] Y. Li, L. Liu, R.T. Tan, Certainty-driven consistency loss for semi-supervised learning, CoRR abs/1901.05657. arXiv:1901.05657..
[25] Y.-L. Boureau, J. Ponce, Y. LeCun, A theoretical analysis of feature pooling in visual recognition, in, in: Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 111–118.
[26] N. Murray, F. Perronnin, Generalized max pooling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2473–2480.
[27] G.-S. Xie, X.-Y. Zhang, S. Yan, C.-L. Liu, Sde: A novel selective, discriminative and equalizing feature representation for visual recognition, Int. J. Computer Vision 124 (2) (2017) 145–168.
[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[29] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[30] W. Wang, Y. Xu, J. Shen, S.-C. Zhu, Attentive fashion grammar network for fashion landmark detection and clothing category classification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[31] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, E. Ricci, Structured attention guided convolutional neural fields for monocular depth estimation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[32] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Processing 27 (5) (2017) 2368–2378.
[33] W. Wang, S. Zhao, J. Shen, S.C.H. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[34] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[35] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artificial Intell. Res. 16 (2002) 321–357.
[36] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst., Man, Cybernetics, Part B (Cybernetics) 39 (2) (2008) 539–550.
[37] X. Chen, Z. Wang, Z. Zhang, L. Jia, Y. Qin, A semi-supervised approach to bearing fault diagnosis under variable conditions towards imbalanced unlabeled data, Sensors 18 (2018) 2097.
[38] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: The European Conference on Computer Vision (ECCV), 2018..
[39] Y. Lecun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel, Handwritten digit recognition with a back-propagation network, Neural Information Processing Systems 2..
[40] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms (2017). arXiv:cs.LG/1708.07747..
[41] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, NIPS Workshop on Deep Learning and Unsupervised Feature Learning..
[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Computer Vision (IJCV) 115 (3) (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[43] S. Zagoruyko, N. Komodakis, Wide residual networks, in, in: Proceedings of the British Machine Vision Conference (BMVC), 2016.
[44] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Fei-Fei Li, Imagenet: A large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition 2009 (2009) 248–255.
[45] S. Ioffe, C. Szegedy, Batch normalization Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning (ICML), 2015.
[46] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: International Conference on Machine Learning (ICML), 2013.
[47] A. Oliver, A. Odena, C.A. Raffel, E.D. Cubuk, I. Goodfellow, Realistic evaluation of deep semi-supervised learning algorithms, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates Inc, 2018, pp. 3235–3246.

Jiaming Chen received the B.E. degree in Information and Computing Sciences from Guangdong Ocean University, Zhanjiang, China, in 2017, and he is currently pursuing the M.S. degree in Sun Yat-sen University, Guangzhou, China. His current research interests include deep learning and semi-supervised learning.

Meng Yang is currently an associate professor at School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He received his Ph.D degree from The Hong Kong Polytechnic University in 2012. He worked as a Postdoctoral fellow in the Computer Vision Lab of ETH Zurich. His research interest includes computer vision, sparse coding and dictionary learning, natrual language processing, and machine learning. He has published more than 60 academic papers, including 14 CVPR/ICCV/AAAI/IJCAI/ICML/ECCV papers and several IJCV, IEEE TNNLS, TIP, and TIFs journal papers. Now his Google citation is over 7800.



Jie Ling received the B.E. degree in communication engineering from South China Normal University, Guangzhou, China, in 2018, and he is currently pursuing the M.S. degree in Sun Yat-sen University, Guangzhou, China. His current research interests include face recognition and semi-supervised learning.