Capturing Prior Knowledge in Soft Labels for Classification with Limited or Imbalanced Data

Zhehao Zhong[1], Shen Zhao[2(✉)], and Ruixuan Wang[1,3(✉)]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China
zhaosh35@mail.sysu.edu.cn
[3] Key Laboratory of Machine Intelligence and Advanced Computing MOE, Guangzhou, China
wangruix5@mail.sysu.edu.cn

Abstract. Successful applications of deep learning often depend on large amount of training data. However, in practical image recognition tasks, available training data are often limited or imbalanced across classes, causing the over-fitting issue or the prediction bias issue during model training. In this paper, based on word embedding models from studies in natural language processing, the prior knowledge about the relationships between image classes is utilized to help train more generalizable classifiers under the condition of limited or class-imbalanced training data. Such inter-class relational knowledge is captured in the word embedding vectors for the textual names of image classes. Using these word embedding vectors as soft labels for corresponding image classes, the feature extractor part of a deep learning model can be guided to learn to extract visual features which contain both class-specific and class-shared information. Experiments on multiple image classification datasets confirm that the proposed learning framework helps improve model performance when training data is limited or class-imbalanced.

Keywords: Prototype learning · Image classification · Limited data · Imbalance data · Soft label

1 Introduction

Deep learning has shown its superior performance in various image recognition applications with the help of sufficient number of training data, such as for face recognition and intelligent diagnosis of specific diseases [1–3]. However, in practice, it may be difficult or even impossible to collect enough number of

This work is supported by NSFCs (No. 62071502, U1811461), the Guangdong Key Research and Development Program (No. 2020B1111190001), and the Meizhou Science and Technology Program (No. 2019A0102005).

S. Yu et al. (Eds.): PRCV 2022, LNCS 13535, pp. 588–600, 2022.
https://doi.org/10.1007/978-3-031-18910-4_47

training data for certain less frequent classes, e.g., images of rare diseases [4–6]. In this case, deep learning models would often suffer from limited or class-imbalanced training data (Fig. 1), due to over-fitting of the model when all classes of training data are limited or/and biased prediction toward frequent classes when training data are class-imbalanced.

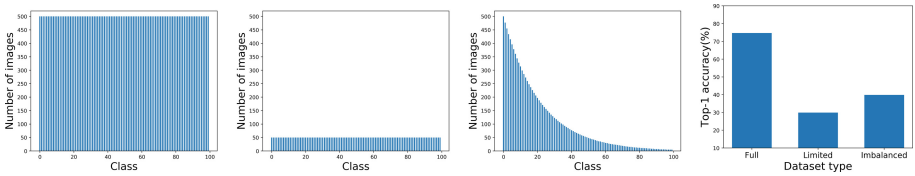

Fig. 1. Model performance (Left) often suffers from limited (Middle left) or imbalanced (Middle right) training data compared to sufficient training data (Right).

To alleviate the over-fitting issue, two main groups of approaches have been developed for training of deep learning models. The first group is based on transferring knowledge from a relatively large auxiliary dataset which contains different classes but in content is often visually similar to the dataset of the current task. The auxiliary dataset is often used to train a feature extractor which is then fixed or fine-tuned for the current task, as in training of matching network [7], prototypical network [8] and relation network [9] for few-shot learning. However, in some scenarios like intelligent medical diagnosis, such large auxiliary dataset is generally not available because of difficulty in collecting large-scale data of many other diseases. Different from the first group of approaches, the second group does not reply on auxiliary datasets but employs various data augmentation techniques to increase the amount of the original training data. Besides the conventional data augmentations like random cropping, scaling, rotating, flipping and color jittering of each training image, advanced augmentation techniques including Cutout [10], Random erasing [11] and Grid mask [12] have been recently developed to further alleviate the over-fitting issue. These augmentations operate directly on images and may not effectively introduce additional high-level semantic information compared with original data, thus still often limited for improving generalizability of deep learning models. In addition, training tricks like label smoothing have also shown to be able to improve model generalization. However, such tricks do not consider specific semantic relationships between classes.

Besides alleviating the over-fitting issue, class-imbalanced recognition tasks also need to reduce the prediction bias issue. Traditional class-balancing approaches include the class re-weighting [13] to increase the importance of training samples from less frequent classes in the loss function, and the re-sampling [14] of training samples to make training data balanced across classes. More recently developed approaches mainly focus on the loss design by considering the instance-level prediction challenge [15] or class-level distribution [16]. Due

to the essential data imbalance between small-sample (i.e., minority) and large-sample (i.e., majority) classes, these class-balancing approaches often achieve trade-off in accuracy between the majority and minority classes, particularly resulting in under-fitting of majority classes and/or over-fitting of minority classes [17].

In this paper, presuming that no auxiliary dataset is available, we propose a simple yet novel strategy to embed prior knowledge into deep learning models to help train more generalizable models with limited or imbalanced data. The prior knowledge is about the semantic relationships between classes, and such semantic relationships are implicitly captured by word embeddings from certain pre-trained natural language processing (NLP) model. The prior knowledge can also be leveraged for computer vision domain [18]. In this model, semantically related words have more similar feature vectors. By associating each (image) class with a specific embedding vector of the class-corresponding word(s), such semantic vector can be considered as the soft label for the class. The semantic relationship between any two classes can be implicitly represented by the proximity between the corresponding two soft labels. Guided by these semantically related soft labels, the feature extractor of a deep learning model can be trained to learn to extract visual features capturing inter-class relationships even with limited or class-imbalanced training data. Considering that knowledge distillation with soft labels has shown its effectiveness in transferring knowledge from one model to another in plenty of studies, it is expected that the soft labels from the NLP model can also help transfer the prior knowledge about class relationships into the deep learning model for image recognition tasks. Since soft labels do not affect designs of model architectures, the proposed strategy can be considered as a plug-in component and flexibly applied to any model backbones. The contributions of this study are summarized below.

- Prior knowledge about class relationships is introduced as soft labels to help train more generalizable models with limited or class-imbalanced training data.
- A learning framework is proposed to effectively train the feature extractor of deep learning models which can capture inter-class relationships with the help of soft labels using limited or class-imbalanced training data.
- Extensive empirical evaluations on multiple image classification tasks with limited data and imbalanced data confirmed the effectiveness and generalizability of the proposed approach.

2 Methodology

The main objective of training a deep learning model for image recognition is to make the model learn knowledge of classes from training images such that it can accurately recognize any new image. However, with limited or class-imbalanced training data, it becomes challenging for the model to well learn the knowledge of particularly small-sample classes. In this case, transferring or embedding prior knowledge of these classes to the model would help the model effectively grasp the knowledge of classes. Here, the word embedding vector of each (image) class

is novelly considered as the soft label of the class, and thus inter-class relationship as part of class knowledge is naturally embedded into the training process of the deep learning model.

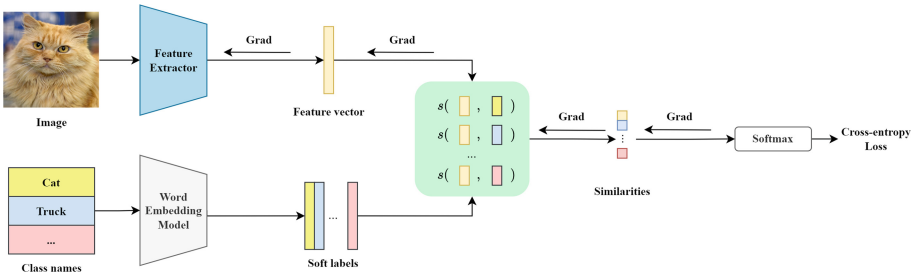

Fig. 2. Illustration of the proposed learning framework. Soft labels for all classes are generated based on a pre-trained and fixed word embedding model. Such soft labels capture semantic inter-class knowledge learned by the word embedding model with enormous amount of text data. The feature extractor is trained with the help of soft labels to learn to extract both class-shared and class-specific features even when training samples are limited or class-imbalanced.

2.1 Learning Framework

The proposed learning framework is illustrated in Fig. 2. The goal is to train the feature extractor F for image recognition of C classes. Suppose the soft label $\mathbf{w}_c \in \mathbb{R}^D$ of the c-th class ($c = 1, \ldots, C$) has been obtained based on a pre-trained word embedding model W (see Sect. 2.2 for details). Given the training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, N\}$, where $\mathbf{x}_i$ is the i-th training image and $\mathbf{y}_i$ is the corresponding ground-truth one-hot label vector, denote by $\mathbf{f}_i = F(\mathbf{x}_i)$ the D-dimensional vector output of the feature extractor F for input $\mathbf{x}_i$, and $\mathbf{z}_i \in \{\mathbf{w}_c, c = 1, \ldots, C\}$ the corresponding soft label of image $\mathbf{x}_i$. If the feature extractor F is trained such that its output $\mathbf{f}_i$ is close to the corresponding soft label $\mathbf{z}_i$ for each image $\mathbf{x}_i$, the feature extractor would be expected to be able to extract visual features which contain inter-class relationships as in the soft label vectors. Thus, using the soft labels to guide the training of the feature extractor would help it learn the prior inter-class relational knowledge even under the condition of limited or class-imbalance training data. The guided training can be achieved by the minimization of the cosine distance loss $\mathcal{L}$ [19], which is actually a special form of the cross-entropy loss based on the cosine similarity between the feature extractor output $\mathbf{f}_i$ and each of the C soft labels, i.e.,

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(\mathbf{f}_i, \mathbf{z}_i)/\tau}}{\sum_{c=1}^{C} e^{s(\mathbf{f}_i, \mathbf{w}_c)/\tau}} . \tag{1}$$

Here $s(\mathbf{f}_i, \mathbf{w}_c)$ is the cosine similarity between vectors $\mathbf{f}_i$ and $\mathbf{w}_c$, and τ is the temperature hyper-parameter (set 1.0 in this study). Note that since the soft labels

of all the C classes participate in the calculation of the cosine distance loss for each image $\mathbf{x}_i$, the feature extractor would be trained to minimize the distance (i.e., maximize similarity) between each $\mathbf{f}_i$ and the corresponding soft label $\mathbf{z}_i$, and meanwhile to maximize the distance between $\mathbf{f}_i$ and the soft labels of all the other classes. In this way, the feature extractor can be trained to extract more discriminative features for those semantically similar classes (having similar soft labels), otherwise the corresponding similar soft labels would cause relatively larger loss. Consequently, the trained model would be more capable of discriminating between classes.

Once the feature extractor is well trained, it can be used to directly predict class of any new (test) image based on the nearest soft label, i.e., finding the class whose soft label vector is nearest to the vector output of the feature extractor.

2.2 Soft Label Generation

The generation of soft label for each class is mainly based on a pre-trained word embedding model. The word embedding model (e.g., Word2Vec [20] and GloVe [21]) is often trained in a self-supervised manner on a large-scale text dataset (like YFCC100M [22]), e.g., by predicting the masked word based on its context words or by predicting its context words based on the centered word in a sentence [20]. Once the self-supervised model is well trained, its feature encoder part (after removing the task-specific model head) can then be used as the word embedding model, i.e., the output of the feature encoder for each input word is the semantic representation of the input word in the embedded feature space. Since the word embedding model is trained based on millions or even billions of sentences, the embedded feature vector for each word captures the potential semantic relationships between the word and each other word. In particular, two words which are semantically closely related (e.g., 'boy' and 'man') are often closer to each other in the embedded feature space [20].

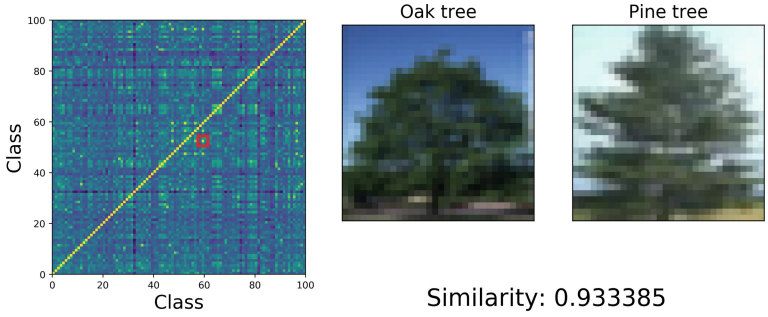

Fig. 3. Cosine similarity between soft labels of every paired classes. Left: each pixel represents the cosine similarity between the soft labels of the corresponding two image classes, with brighter pixels representing higher cosine similarities. Right: exemplar images of two classes with higher similarity (as indicated within the red box on the left) between corresponding soft labels. Both visually similar and class-specific features exist in the images. (Color figure online)

Based on the word embedding model, the soft label of each image class can be generated as follows. When one image class can be named by a single word (e.g., class of 'Cat'), then the output of the word embedding model for the word can be directly used as the soft label of the image class. On the other hand, when one image class is named by a sequence two or more words (e.g., class of 'Great white shark'), the weighted sum of the outputs of the word embedding model for the multiple words can be used as the soft label of the image class. In this study, simply the average of the outputs is adopted to represent the soft label for each multi-word class, although potentially more adaptive weight setting can be further investigated in future work. Figure 3 (Left) shows the cosine similarities between the soft labels of every two classes in the well-known CIFAR100 dataset. For those paired classes which have higher cosine similarities, similar visual features do often appear in the images of the classes, as demonstrated in Fig. 3 (Right). In this case, the soft label will guide the feature extractor to learn to extract such shared visual features during model training, meanwhile to extract class-specific features to discriminate between these semantically similar classes. Extraction of such more comprehensive (i.e., both class-shared and class-specific) features particularly from small-sample classes indicates that the feature extractor can be trained to extract representative features for each class, even with limited of imbalanced training samples.

2.3 Comparison with Relevant Methods

The soft labels can be considered as class centers in the semantic feature space. Different from the proposed generation of soft labels from a word embedding model in this study, several previous studies have proposed to directly learn one or multiple class centers for each class during the training of the classifier, where the class centers and model parameters are learned jointly. Examples include the center loss [23], the convolutional prototype learning (GCPL) [24] and the prototypical networks [8]. Since all these methods learn class centers solely based on training data, it is expected that the jointly learned classifier and the class centers would become over-fitting particularly when training data are limited or class-imbalanced. In comparison, the proposed soft label in this study is based on the word embedding model which is trained on large-scale text data, and therefore the semantic inter-class knowledge in the soft label may help train a more generalizable model with limited training samples.

3 Experiments

3.1 Experimental Setting

To evaluate the effectiveness and generalizability of the proposed method for image recognition under the condition of limited or class-imbalanced training samples, three public image datasets, i.e., CIFAR-10, CIFAR-100, and mini-ImageNet, were employed to create the limited (i.e., small-sample) and class-imbalanced training sets. For small-sample training sets, three versions were created from each original dataset, with 50, 100, and 200 images randomly sampled

from each class of CIFAR-10 and CIFAR-100 respectively, and 10, 50, and 100 images per class randomly sampled from mini-ImageNet respectively. To create the class-imbalanced versions, the number of training samples were exponentially reduced across classes, while the test set was kept class-balanced. Denote by ρ the imbalance ratio between sample sizes of the largest class and the smallest class. ρ was set 10 and 100 respectively in relevant experiments.

In experiments, Resnet18 was used as the default model backbone, and two word embedding models GloVe [21] and Bert [25] were used to generate 300-dimensional and 768-dimensional soft labels respectively, with GloVe used by default in each experiment. During model training, conventional data augmentations were applied on each training set, including random cropping and horizontal flipping for each image. Stochastic gradient descend with momentum 0.9 and weight decay 10^{-4} was used to optimize each model for 200 epochs. The batch size is 128 for CIFAR-10 and CIFAR-100 and 64 for mini-ImageNet. The initial learning rate 0.1 was decayed by 0.1 respectively on epochs 120, 160, and 180. The linear warm-up learning rate schedule was also used for the first 5 epochs. During testing, the average and standard deviation of classification accuracy over five runs were reported. In addition, for the model trained with class-imbalanced data, the classification accuracy on larger-size classes (i.e., those 1/3 classes with larger training samples), smaller-size classes (i.e., those 1/3 classes with smaller training samples), and medium-size classes (i.e., the remaining 1/3 classes) were also reported respectively.

3.2 Effectiveness Evaluation with Limited Training Data

The proposed method was first evaluated with limited training samples. The baseline methods for comparison include the basic cross-entropy loss (CE), the center loss (CenterLoss) [23], the convolutional prototype learning (GCPL) [24], the prototypical network (ProtoNet) [8], and the label smoothing (LabelSmooth). The released source codes and suggested settings from the original studies were adopted for CenterLoss, GCPL and ProtoNet. Smooth factor of label smoothing was set 0.1.

As shown in Tables 1 and 2, compared to all the baseline methods (rows 1–5), the proposed method (last two rows) significantly improves model performance when the training samples are very limited, i.e., with 50 images per class on CIFAR-10 and CIFAR-100, and 10 images per class on mini-Imagenet. With relatively more training samples (i.e., respectively 100 and 200 images per class on CIFAR, and 50 and 100 images per class on mini-Imagenet), the proposed method still achieves satisfactory performance, often slightly better than the strongest baseline at each setting. Additional evaluations on different model backbones (e.g., ResNet50 and VGG16, Table 3) also confirm the effectiveness of the proposed method under the condition of limited training samples. Furthermore, when the proposed method is combined with existing methods like Cutout [10], Random erasing [11], and Grid mask [12], the classification performance is often significantly boosted compared to these individual methods, as demonstrated in Fig. 4 with varying number of training samples on the three datasets. These results support that the inter-class knowledge in the soft label

can help train more generalizable models with limited training samples, and this method can be flexibly combined with (some of) existing strategies to further improve their effectiveness.

Table 1. Performance comparison on datasets with limited training data from CIFAR. '50/100/200': training samples per class. In brackets: standard deviations.

Methods	CIFAR-10			CIFAR-100		
	50	100	200	50	100	200
CE	37.44 (0.74)	48.68 (0.41)	54.51 (0.45)	31.76 (1.31)	49.99 (0.94)	63.41 (0.79)
CenterLoss	35.44 (0.99)	47.17 (1.45)	59.04 (1.28)	26.49 (0.40)	50.53 (0.19)	64.61 (1.16)
GCPL	35.38 (1.18)	48.65 (0.93)	63.25 (0.52)	27.10 (1.02)	51.38 (0.40)	64.34 (0.85)
ProtoNet	36.47 (1.17)	49.78 (0.87)	61.73 (0.74)	29.65 (1.48)	49.90 (0.45)	61.53 (0.70)
LabelSmooth	35.99 (1.48)	48.80 (0.85)	63.07 (0.25)	34.04 (1.21)	52.13 (0.32)	64.44 (0.92)
Ours (Bert)	**41.85** (0.34)	**52.41** (0.88)	**64.03** (0.50)	39.22 (0.21)	52.19 (0.35)	64.17 (0.15)
Ours (GloVe)	41.74 (0.57)	51.23 (0.50)	63.56 (0.56)	**39.51** (0.36)	**53.06** (0.36)	**64.84** (0.57)

Table 2. Performance comparison on datasets with limited training data from mini-ImageNet. '10/50/100': training samples per class. In brackets: standard deviations.

Methods	10	50	100
CE	13.25 (0.16)	39.68 (0.68)	51.79 (0.11)
CenterLoss	13.71 (0.26)	40.96 (0.91)	54.61 (0.50)
GCPL	12.45 (0.95)	42.03 (0.83)	54.81 (0.73)
ProtoNet	13.57 (0.89)	41.03 (1.35)	52.00 (0.73)
LabelSmooth	13.80 (0.23)	42.16 (0.34)	53.98 (0.65)
Ours (Bert)	**18.31** (0.34)	42.75 (0.18)	54.24 (0.38)
Ours (GloVe)	17.62 (0.50)	**43.37** (0.34)	**55.03** (0.61)

Table 3. Performance comparison on different model backbones with limited (50 per class) training images

Methods	Resnet50		Vgg16	
	CIFAR-100	Mini-ImageNet	CIFAR-100	Mini-ImageNet
CE	26.08 (0.69)	40.56 (0.68)	25.45 (1.48)	29.54 (0.92)
CenterLoss	24.17 (0.65)	39.16 (1.34)	21.83 (1.18)	28.13 (1.48)
GCPL	26.79 (1.21)	39.23 (1.44)	25.20 (1.20)	29.62 (1.09)
ProtoNet	31.95 (0.66)	40.43 (1.01)	28.30 (1.09)	28.49 (1.26)
LabelSmooth	29.56 (0.28)	41.47 (0.27)	29.79 (1.13)	29.54 (0.28)
Our (Bert)	32.13 (0.99)	43.78 (0.69)	**38.20** (0.98)	**39.64** (0.59)
Our (GloVe)	**33.25** (0.53)	**44.01** (0.19)	33.85 (0.51)	39.06 (0.86)

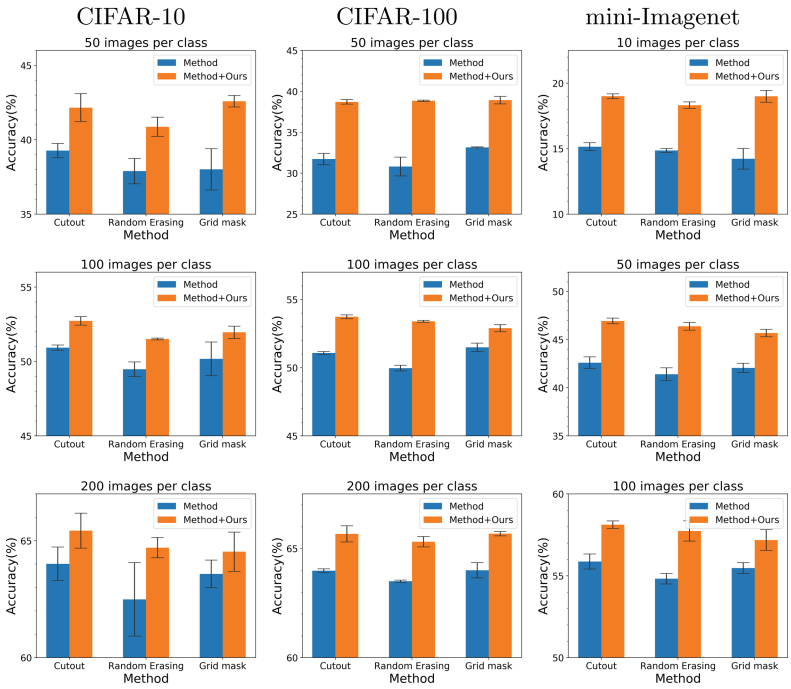

Fig. 4. Performance comparison based on limited training samples between each baseline and its combination with the proposed method. The baselines include Cutout, Random erasing, and Grid mask. Vertical lines for standard deviations.

Table 4. Performance comparison with class-imbalanced training set from CIFAR-10. Test set is class-balanced.

Methods	$\rho = 100$				$\rho = 10$			
	All	Larger	Medium	Smaller	All	Larger	Medium	Smaller
CE	71.20	92.73	72.63	55.20	87.35	94.17	**85.23**	85.08
Resample	70.64	92.23	70.30	53.27	87.92	94.67	84.40	85.20
Reweight	70.39	90.50	69.90	**55.67**	87.61	94.20	84.50	85.50
Ours	**72.13**	**93.27**	**73.33**	55.38	**88.21**	**95.00**	84.93	**85.58**
LDAM	76.48	93.70	76.50	63.55	88.80	94.27	85.27	87.35
LDAM+Ours	**77.30**	**93.93**	**76.73**	**65.25**	**89.17**	**95.13**	**85.93**	**87.55**
CB Focal	74.10	**93.37**	74.80	59.12	88.88	93.33	85.37	**88.17**
CB Focal+Ours	**74.63**	93.19	**75.53**	**60.06**	**89.19**	**93.70**	**86.53**	87.05

3.3 Evaluation with Imbalanced Training Data

The proposed method also helps under the condition of class-imbalanced training data. As shown in Tables 4, 5 and 6, the proposed method (4-th row, 'Ours') overall outperforms the baselines (rows 1–3) under all settings ('All' columns, top-1 accuracy on all classes; note that test data are class-balanced). Interest-

Table 5. Performance comparison with class-imbalanced training set from CIFAR-100. Test set is class-balanced.

Methods	$\rho = 100$				$\rho = 10$			
	All	Larger	Medium	Smaller	All	Larger	Medium	Smaller
CE	39.85	67.79	38.97	13.59	56.97	71.97	58.67	41.97
Resample	35.26	66.45	36.97	13.03	55.78	70.45	57.58	39.79
Reweight	39.56	67.58	38.79	13.85	56.19	68.85	58.09	42.05
Ours	**41.81**	**69.36**	**41.33**	**15.53**	**60.35**	**73.48**	**61.24**	**46.74**
LDAM	42.04	**69.94**	42.03	**14.97**	60.14	72.91	60.33	**47.56**
LDAM+Ours	**42.49**	69.85	**43.58**	14.88	**60.29**	**73.45**	**60.82**	46.91
CB Focal	39.85	67.79	38.97	13.59	57.85	70.06	58.52	45.35
CB Focal+Ours	**42.87**	**70.94**	**42.06**	**16.41**	**62.05**	**75.33**	**62.06**	**49.15**

Table 6. Performance comparison with class-imbalanced training set from mini-ImageNet. Test set is class-balanced.

Methods	$\rho = 100$				$\rho = 10$			
	All	Larger	Medium	Smaller	All	Larger	Medium	Smaller
CE	50.92	79.88	50.70	23.03	70.04	83.06	69.61	57.82
Resample	46.25	77.18	45.15	20.12	68.99	82.70	69.03	55.65
Reweight	47.44	78.65	48.54	23.09	69.20	81.97	68.82	57.18
Ours	**51.80**	**80.03**	**52.67**	**23.56**	**71.35**	**84.88**	**72.45**	**59.29**
LDAM	52.20	79.36	52.12	25.91	70.50	81.15	70.42	**58.94**
LDAM+Ours	**52.41**	**79.39**	**52.97**	**26.41**	**71.12**	**82.76**	**71.12**	58.82
CB Focal	51.66	**80.03**	**51.97**	24.47	69.08	79.70	69.21	58.65
CB Focal+Ours	**51.83**	79.79	51.20	**26.18**	**70.20**	**80.73**	**70.06**	**59.59**

ingly, the proposed method improves the performance not only on those classes with smaller training samples ('Smaller' columns) under most settings, but also on those classes with larger-size ('Larger') and medium-size ('Medium') training samples. In addition, as with limited training data, the proposed method can also further improve the performance of existing methods by fusing them together (Tables 4, 5 and 6, last four rows). For example, models trained with the combination of CB Focal and the proposed method achieve the best classification performance on CIFAR-100 with different imbalance ratios. These results suggest that the proposed method can be applied to class-imbalanced image recognition either individually or in combination with existing strategies.

3.4 Ablation Study

To confirm the soft labels from a pre-trained word embedding model is essential to improve model performance with limited or class-imbalanced training set, an ablation study was performed by replacing the soft labels with randomly generated soft vectors. The element in each random vector was randomly sampled from uniform distribution. Figure 5 shows that, under the conditions of both limited and class-imbalanced training sets, the proposed soft labels often perform better than the randomly generated soft labels, while the latter is comparable to

(i.e., either slightly better or worse than) the basic CE method under various settings. This supports that the implicit inter-class relationship in soft labels from the pre-trained word embedding model may help classifiers gain more semantic knowledge and generalizable performance.

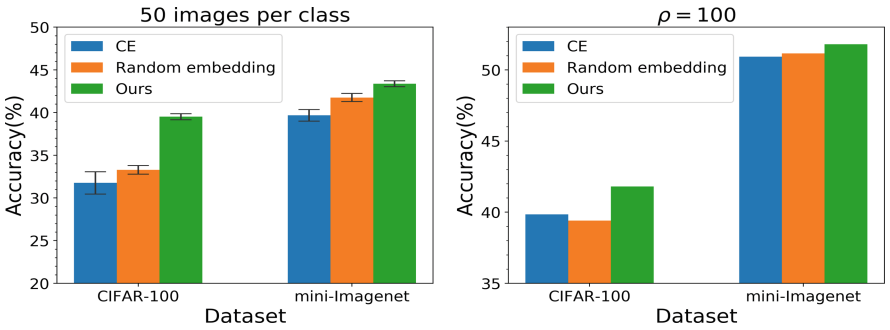

Fig. 5. Ablation study of soft labels. Left: classification performance of models with limited training set by the CE baseline, the randomly generated soft labels, and the proposed soft labels respectively. Right: performance of models with class-imbalanced training set.

4 Conclusions

In this study, soft labels containing inter-class relationships are proposed to guide the training of image recognition models under the condition of limited or class-imbalanced training samples. Extensive evaluations with three image classification datasets consistently support that the proposed learning framework is effective in improving the performance of classifiers, and its combination with existing strategies for small-sample or class-imbalanced learning can further improve the performance of these strategies. The proposed learning framework might also help train classifiers under more extreme conditions, such as those in zero-shot learning and open-set recognition. These extensions will be investigated in future work.

References

1. Sun, Y., Liang, D., Wang, X.G., Tang, X.O.: DeepID3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
2. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
3. Zhao, S., Wu, X., Chen, B., Li, S.: Automatic spondylolisthesis grading from MRIs across modalities using faster adversarial recognition network. Med. Image Anal. **58**, 101533 (2019)

4. Chen, K., et al.: Alleviating data imbalance issue with perturbed input during inference. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 407–417. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_39
5. Hu, Y., Zhong, Z., Wang, R., Liu, H., Tan, Z., Zheng, W.-S.: Data augmentation in logit space for medical image classification with limited training data. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 469–479. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_45
6. Zhuang, J., Cai, J., Wang, R., Zhang, J., Zheng, W.-S.: Deep kNN for medical image classification. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 127–136. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_13
7. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D.: Matching networks for one shot learning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), vol. 29, pp. 3630–3638 (2016)
8. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4077–4087 (2017)
9. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1199–1208 (2018)
10. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
11. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 34, no. 7, pp. 13001–13008 (2020)
12. Chen, P., Liu, S., Zhao, H., Jia, J.: Gridmask data augmentation. arXiv preprint arXiv:2001.04086 (2020)
13. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5375–5384 (2016)
14. Shen, L., Lin, Z., Huang, Q.: Relay backpropagation for effective learning of deep convolutional neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 467–482. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_29
15. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9268–9277 (2019)
16. Cao, K., et al.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 1565–1576 (2019)
17. Zhao, S., Chen, B., Chang, H., Chen, B., Li, S.: Reasoning discriminative dictionary-embedded network for fully automatic vertebrae tumor diagnosis. Med. Image Anal. **79**, 102456 (2022)
18. Zhao, S., Gao, Z., Zhang, H., Xie, Y., Luo, J., et al.: Robust segmentation of intima-media borders with different morphologies and dynamics during the cardiac cycle. IEEE J. Biomed. Health Inf. **22**, 1571–1582 (2017)
19. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748 (2018)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

21. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
22. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., et al.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)
23. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
24. Yang, H.M., Zhang, X.Y., Yin, F., Liu, C.L.: Robust classification with convolutional prototype learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3474–3482 (2018)
25. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)