Robust Semi-Supervised Learning when Not All Classes have Labels

Lan-Zhe Guo[*], **Yi-Ge Zhang**[*], **Zhi-Fan Wu, Jie-Jing Shao, Yu-Feng Li**[†]
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
`{guolz,zhangyg,wuzf,shaojj,liyf}@lamda.nju.edu.cn`

Abstract

Semi-supervised learning (SSL) provides a powerful framework for leveraging unlabeled data. Existing SSL typically requires all classes have labels. However, in many real-world applications, there may exist some classes that are difficult to label or newly occurred classes that cannot be labeled in time, resulting in there are unseen classes in unlabeled data. Unseen classes will be misclassified as seen classes, causing poor classification performance. The performance of seen classes is also harmed by the existence of unseen classes. This limits the practical and wider application of SSL. To address this problem, this paper proposes a new SSL approach that can classify not only seen classes but also unseen classes. Our approach consists of two modules: *unseen class classification* and *learning pace synchronization*. Specifically, we first enable the SSL methods to classify unseen classes by exploiting pairwise similarity between examples and then synchronize the learning pace between seen and unseen classes by proposing an adaptive threshold with distribution alignment. Extensive empirical results show our approach achieves significant performance improvement in both seen and unseen classes compared with previous studies.

1 Introduction

Machine learning, especially deep learning, has achieved great success in various tasks by leveraging sufficient labeled training data [21]. However, for many practical tasks, it can be difficult to attain a number of labeled examples due to the high cost of the data labeling process [41, 23], which limits the widespread adoption of machine learning techniques.

Semi-supervised learning (SSL) [43] provides a powerful framework for leveraging unlabeled data when labels are limited or expensive to obtain. There has been a rapid development of SSL methods in recent years, such as entropy minimization methods [22, 8], consistency regularization methods [24, 29, 20, 32], and holistic methods[30, 2, 1, 35, 37]. It has been reported that in certain cases, such as image classification [30], SSL methods can achieve the performance of purely supervised learning even when a substantial portion of the labels in a given dataset have been discarded.

All the positive results of SSL, however, are based on a basic assumption that there are labels for each of the classes that one wishes to learn, i.e., all training and testing data are from seen classes that are observed in the labeled dataset. However, in many real-world applications, particularly those involving open-environment scenarios [42, 11], such an assumption is difficult to hold. For example, in the product recognition task, thousands of new types of products are introduced to the supermarkets every once in a while, and it would be expensive to label them all in time; in the judicial sentencing task, some sentencing elements are naturally scarce, resulting in labeled judgment documents being difficult to obtain for these elements in the training phase.

---

[*] Contribute to this work equally

[†] Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

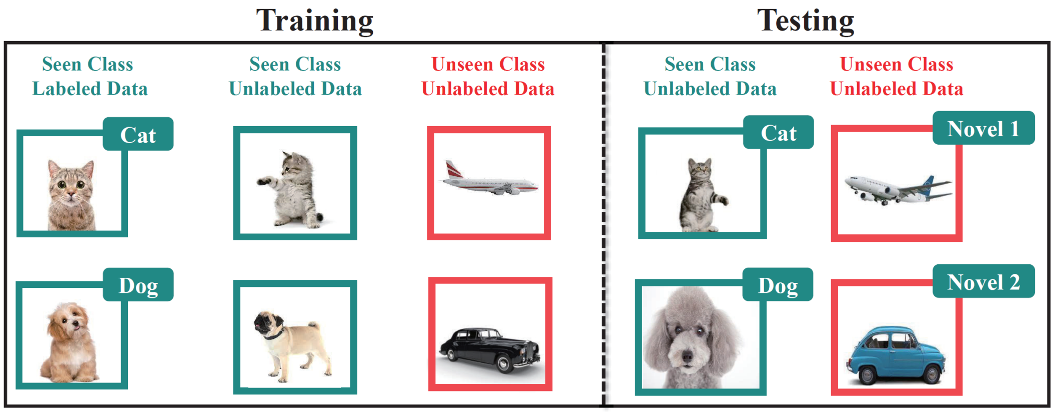

Figure 1: SSL when not all classes have labels. Training data includes labeled examples from a few seen classes as well as unlabeled examples from both seen and unseen classes. Testing data includes unlabeled examples from both seen and unseen classes. The goal is to classify not only seen class examples into accurate classes but also partition unseen class examples into proper clusters.

It is evident that existing SSL methods will misclassify unlabeled examples from unseen classes into seen classes. Even on the seen classes, SSL performance can also degrade severely with the presence of unseen classes unlabeled data [25, 10, 40]. The problem is related to open-set SSL and novel class discovery (NCD) studies. Open-set SSL [10, 6, 36, 28] aims to decrease the negative impact of unseen classes and maintain the performance robustness of seen classes. However, these methods simply detect and drop examples from unseen classes and fail to classify them. NCD [13, 15, 38] aims to discover unseen classes automatically. However, they ignore the classification task on seen classes, which results in performance degradation in seen classes.

An illustration of the problem concerned in this paper is presented in Figure 1. It is evident that both existing SSL methods and NCD methods could not tackle the problem. This inspires us to consider answering the following question in this study:

Can we design an robust SSL algorithm that can classify both seen and unseen classes when not all classes have labels in the training data?

To this end, we propose a new SSL method called NACH, which consists of two key modules: unseen class classification and learning pace synchronization. Specifically, we first propose a novel unseen class classification loss that can exploit pairwise similarity to classify similar example pairs into the same class and eliminate noisy pairs based on a similarity filter. We then adopt an adaptive threshold with distribution alignment to alleviate the issue that different learning paces between seen and unseen classes. Experimental results on CIFAR-10, CIFAR-100, and ImageNet-100 datasets show that NACH achieves 37.7% improvement in unseen classes compared with SSL methods, and 26.3% improvement in seen classes compared with NCD methods.

2 Related Work

Semi-Supervised Learning. SSL assumes all training and testing data are from seen classes, no matter whether labeled or unlabeled, and the goal is to classify unlabeled examples into the ground-truth classes. SSL has a long research history [4]. Our paper is mainly related to deep SSL, which introduces SSL techniques to deep neural networks and has achieved significant advancement in recent years. The mainstream of deep SSL can be broadly categorized into entropy minimization methods [22, 8], consistency regularization methods [24, 29, 20, 32], and holistic methods [30, 2, 1]. When not all classes have labels in the training data, these methods will misclassify unlabeled data from unseen classes as the seen classes and fail to address the problem concerned in this paper.

Open-Set Semi-Supervised Learning. Open-set SSL relaxes the assumption of SSL and considers a more practical scenario that training data could contain unseen class unlabeled examples. However, they still assume all testing examples are from seen classes, and the goal is to decrease the negative impact of unseen class unlabeled data in order to maintain the performance robustness in seen classes. Many open set SSL methods have been proposed in recent years [10, 6, 36, 31, 28, 16, 26, 17], such as DS3L [10], which assign weights to unlabeled data based on a bi-level optimization, UASD [6], which filter unlabeled examples based on the prediction consistency, MTC [36], which adopt a multi-task curriculum learning framework to detect unseen classes and classify seen classes simultaneously, T2T [16], which propose a novel cross-modal matching strategy to detect unseen classes. OpenMatch [28], which unify FixMatch algorithm with novelty detection based on one-vs-all (OVA) classifiers. However, these methods can still not classify unseen classes.

Novel Class Discovery. NCD assumes training data consists of seen class labeled examples and unseen class unlabeled examples, and the goal is to classify both seen and unseen classes during the testing phase. The NCD problem is first formally introduced in [18]. Recently, many NCD methods have been proposed based on a two-step training strategy [13–15, 7, 12, 39, 38], i.e., a data embedding is learned on the labeled data using a metric learning technique, and then fine-tuned while learning the cluster assignments on the unlabeled data. In contrast to the problem studied in this paper, NCD methods ignore the abundant seen unlabeled examples that are usually easy to collect in real-world applications.

3 Preliminary and Background

Give the training data which contains n labeled examples $D_l = \{(x_1; y_1); \cdots; (x_n; y_n)\}$ and m unlabeled examples $D_u = \{x_{n+1}; \cdots; x_{n+m}\}$. Usually, $m \gg n$. $x \in R^D$; $y \in Y = \{1; \cdots; C_L\}$ where D is the feature dimension and C_L is the number of seen classes. We use C_U to represent the total number of classes in unlabeled data, Previous SSL studies assume $C_L = C_U$ and NCD assumes $C_L \cap C_U = \emptyset$. In this paper, the number of seen classes $C_{seen} = C_L \cap C_U$ and the number of unseen classes $C_{unseen} = C_U \setminus C_L$. The goal is to learn a classification model $f(x;\theta)$ from training data. Specifically, the $f(x;\theta)$ can be decomposed of a representation learning model $g(x;\theta) : R^D \to R^d$ to learn a low-dimensional feature z and a classification model $H(z) : R^d \to R^{C_{seen} + C_{unseen}}$.

The training loss of an SSL algorithm usually contains supervised loss L_s and unsupervised loss L_u with a trade-off parameter $\lambda_u > 0$: $L_s + \lambda_u L_u$, where L_s is constructed on D_l and L_u is constructed on D_u. Typically, L_s applies the standard cross-entropy loss on labeled examples:

$$L_s = \frac{1}{n} \sum_{i=1}^{n} H(y_i; p(x_i)) \tag{1}$$

where $p(x) = \mathrm{Softmax}(f(x;\theta))$ indicate the predicted probabilities produced by the model f for the input x, and $H(\cdot;\cdot)$ is the cross-entropy function.

Different constructions of the unsupervised loss L_u lead to different SSL algorithms. Typically, there are two ways of constructing L_u: one is to assign pseudo-labels to formulate a "supervised loss" such as the cross-entropy loss, and another one is to optimize a regularization that does not depend on labels such as consistency regularization.

Take the FixMatch [30] and UDA [34] for examples, FixMatch adopts the pseudo-label loss which can be written as:

$$L_u = \frac{1}{m} \sum_{i=n+1}^{n+m} I(\max(p(\alpha(x_i))) \geq \tau)\, H(\hat{p}_i; p(A(x_i))) \tag{2}$$

where $\alpha(x)$ and $A(x)$ indicate the weak and strong augmentation, $\hat{p}_i = \arg\max p(\alpha(x_i))$ represent the pseudo-label for unlabeled example x_i, τ is the confidence threshold for pseudo-label selection, $I(\cdot)$ is the indicator function.

UDA [34] adopts the consistency regularization which can be written as

$$L_u = \frac{1}{m} \sum_{i=n+1}^{n+m} \| p(aug(x_i)) - p(aug'(x_i)) \|_2^2 \tag{3}$$

where $aug(\cdot)$ and $aug'(\cdot)$ represents different augmentation strategies.