Partial Label Learning with Semantic Label Representations

Shuo He
University of Electronic Science and Technology of China
Chengdu, China
shuohe@std.uestc.edu.cn

Lei Feng
Chongqing University
Chongqing, China
feng0093@e.ntu.edu.sg

Fengmao Lv
Southwest Jiaotong University
Chengdu, China
fengmaolv@126.com

Wen Li
University of Electronic Science and Technology of China
Chengdu, China
liwenbnu@gmail.com

Guowu Yang*
University of Electronic Science and Technology of China
Chengdu, China
guowu@uestc.edu.cn

ABSTRACT

Partial-label learning (PLL) solves the problem where each training instance is assigned a candidate label set, among which only one is the ground-truth label. The core of PLL is to learn efficient feature representations to facilitate *label disambiguation*. However, existing PLL methods only learn plain representations by coarse supervision, which is incapable of capturing sufficiently distinguishable representations, especially when confronted with the knotty *label ambiguity*, i.e., certain candidate labels share similar visual patterns. In this paper, we propose a novel framework **Par**tial label learning with **SE**mantic label representations dubbed **ParSE**, which consists of two synergistic processes, including visual-semantic representation learning and powerful *label disambiguation*. In the former process, we propose a novel weighted calibration rank loss that has two implications. First, it implies a progressive calibration strategy that utilizes the disambiguated label confidence to weight the similarity between each image feature embedding and its corresponding semantic label representations of all candidates. Second, it also considers the ranking relationship between candidate and non-candidate ones. Based on learned visual-semantic representations, subsequent *label disambiguation* is desirably endowed with more powerful abilities. Experiments on benchmarks show that ParSE outperforms state-of-the-art counterparts.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**.

KEYWORDS

partial-label learning, semantic label representations, label disambiguation

ACM Reference Format:

Shuo He, Lei Feng, Fengmao Lv, Wen Li, and Guowu Yang. 2022. Partial-label Learning with Semantic Label Representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539434

1 INTRODUCTION

The success of modern deep learning techniques depends on a large amount of correctly labeled data. However, such kind of data in real-world scenarios is considerably scarce due to the difficulties in data annotations [1, 14, 42]. One major obstacle to acquire such desired data is *label ambiguity*, which means it is sometimes hard to distinguish certain classes because they share similar vision patterns. Therefore, it could be considerably difficult for non-expert human annotators to decidedly select one ground-truth domain label as the "gold", as they may be torn between some choices of similar domain classes. For example, as shown in Figure 1, for a non-expert human annotator without the domain knowledge of pet cat, it is seriously entangled to choose one from three classes (i.e., Ragdoll, Chinchilla, and Garfield) that all belong to the cat. This issue has rised researchers' attention to *partial-label learning* (PLL) [3, 20] that allows each training example to be equipped with a candidate label set where only one is the ground-truth label. Compared with supervised learning where only one label is selected as the ground-truth label, PLL is quite friendly for non-expert human annotators and has relatively lower cost in annotations. Therefore, PLL is more practical and common in many real-world applications such as bioinformatics [15], web mining [16], and automatic image annotation [2, 11, 37].

Over the past decade, many conventional PLL methods have been proposed to induce a multi-class classifier based on hand-crafted features (i.e., instances) and the corresponding candidate labels. They are typically based on a two-stage framework. Specially, it encapsulates two steps. The first is one-off and global *label disambiguation* that identifies the ground-truth labels of all instances from candidate labels at once based on certain data distribution assumptions in the feature space (e.g., smoothness and cluster assumptions [38, 39]). Second, a classifier can be trained directly by using disambiguated labels. The fatal weakness of this framework is that they rely heavily on pre-acquired hand-crafted feature representations to carry out *label disambiguation* rather than end-to-end

[*] Corresponding author

KDD '22, August 14–18, 2022, Washington, DC, USA.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539434

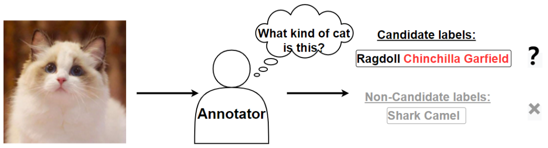

Figure 1: An example image with a candidate label set Y={Ragdoll, Chinchilla, Garfield} and a non-candidate label set $\bar{Y}$={Shark, Camel}. The ground-truth label is Ragdoll.

training with deep learning technologies (e.g, stochastic optimization algorithms and deep neural networks), and thus its scalability to large-scale datasets is greatly limited.

To alleviate this issue, researchers recently gave rise to the attention of deep PLL methods that aim to train an end-to-end deep neural network with partially labeled data. Compared with conventional PLL, deep PLL needs to learn feature representations from scratch together with *label disambiguation*. Hence, deep PLL methods are commonly based on an end-to-end synergistic framework that iteratively performs *label disambiguation* and representation learning in a mini-batch with stochastic optimization algorithms [6, 17, 31]. However, they only learn plain visual representations by coarse supervision, which is incapable of capturing efficient and distinguishable representations for *label disambiguation*, especially when confronted with the knotty *label ambiguity*, i.e., certain candidate labels share similar visual patterns. For example, as shown in Figure 1, an image of ragdoll is associated with a candidate label set (i.e., ragdoll, chinchilla, Garfield). These candidate labels are a subclass of cat and thus share similar visual patterns. In this case, naive representation learning only using disambiguated labels is insufficient to learn high-quality representations for discriminating these candidate labels. Furthermore, it may lead to a vicious circle between the two processes—learned indistinguishable representations hinder effective *label disambiguation*—tricky label ambiguity in turn hampers efficient representation learning. Therefore, it is of vital importance to learn sufficiently high-quality representations to promote a virtuous cycle in this synergistic framework.

To achieve this goal, in this paper, we propose a novel framework **Par**tial label learning with **SE**mantic label representations dubbed **ParSE**, which utilizes semantic label representations to learn efficient visual-semantic feature representations to promote *label disambiguation*. Specially, ParSE consists of two synergistic processes, including visual-semantic representation learning and powerful *label disambiguation*. In the former process, it may be quite simple to learn visual-semantic representations in many other tasks with known ground-truth labels. However, due to the inherent *label ambiguity* (i.e., the ground-truth label is concealed in the candidate label set) in PLL, we cannot directly learn such visual-semantic representations. To alleviate this issue, we propose a novel weighted calibration rank loss that has two implications. First, it implies a progressive calibration strategy that utilizes the disambiguated label confidence to weight the similarity between each image feature embedding and its corresponding semantic label representations of all candidate labels. In this way, with the guidance of the weight that is constantly updating, the similarity towards all candidate labels is progressively calibrated to the right objective (i.e., the similarity towards the ground-truth one). Second, it also considers the ranking relationship between candidate and non-candidate ones to achieve conforming similarity ranking. Based on learned visual-semantic representations, subsequent *label disambiguation* is desirably endowed with more powerful abilities. Our main contributions are summarized as follows:

- To the best of our knowledge, we pioneer the exploration of using semantic label representations for PLL, and provide a novel framework called ParSE that performs visual-semantic representation learning and powerful label disambiguation.
- We propose a novel weighted calibration rank loss for learning visual-semantic representations in PLL, which implies a progressive calibration strategy and similarity ranking between candidate and non-candidate labels.
- Empirically, ParSE outperforms state-of-the-art counterparts on benchmarks. Extensive experiments show the effectiveness of our proposed framework ParSE.

2 RELATED WORK

In this section, we briefly review related studies on partial-label learning and semantic label representations, as the two topics are highly related to our work.

2.1 Partial-Label Learning

Partial-label learning (PLL) deals with the problem where each training example is associated with a candidate label set, among which only one is the ground-truth label. Generally, there are two lines of PLL methods—conventional PLL with hand-crafted features and deep PLL in an end-to-end training manner.

First, conventional PLL methods are typically based on hand-crafted features and aim to only induce a multi-class classifier [5, 18, 26, 34]. There are two common frameworks including the average-based and identification-based frameworks. The former [3, 10] treated all candidate labels equally, which obviously could mislead the classifier due to false positive labels in the candidate label set. To prevent such undesired adverse effect, the latter aims to identify the ground-truth one from the candidates, i.e., label disambiguation. To achieve this aim, [28, 33, 38, 39] utilized the topological structure in the feature space to construct an instance graph that could be adapted in the label space, which could obtain numerical label values and thus identify the maximum label value as the ground-truth label. Further, [19] reformulated the task of PLL as a matching selection problem integrated with instance and label relationships. Specially, [35] employed shallow neural networks to solve the PLL problem with certain deep learning technologies. Concretely, they carried out label disambiguation in a mini-batch and adapted a mix-up data augmentation scheme and a teacher model to ensure the stability of the batch-based prediction network update. The fatal weakness of these methods is that they rely heavily on pre-acquired hand-crafted feature representations rather than end-to-end training with advanced deep learning technologies (e.g, stochastic optimization algorithms and deep neural networks), and thus their scalability to large-scale datasets is greatly limited.

To fill this gap, researchers give rise to the attention of deep PLL methods recently that end-to-end trains a deep neural network with partially labeled data. First, [6, 17] proposed deep PLL algorithms with the guarantee of risk and classifier consistency, which were agnostic in specific classification models and could be easily trained with stochastic optimization. Analogously, [31] proposed a family of loss functions named leveraged weighted loss with risk consistency that considered the trade-off between losses on partial labels and non-partial ones. Although they provided rigorous theoretical guarantees to the proposed PLL algorithms, their practical effects were limited in learned plain feature representation especially faced with the knotty label ambiguity. Little works have made efforts to learn visual-semantic feature representations to facilitate label disambiguation in PLL.

2.2 Semantic Label Representation

Semantic label representations, generally served as side information [24, 29], has been shown to be effective in various domains and tasks, e.g., zero-shot learning [7, 21, 25, 36], multi-modal learning between images and textual data [4, 22, 23, 41] and text classification tasks [8, 32, 40].

In zero-shot learning, the pivotal role of semantic label representations is a bridge to establish connections between seen and unseen classes. For this purpose, [21] applied human-made semantic label representations (named semantic output codes) in zero-shot classification. [7] utilized semantic information gleaned from unannotated text to learn a visual-semantic embedding model. Similarly, [25] produced continuous semantic word embeddings as semantic label representations using an unsupervised language model. They showed that the label correlation captured in the embedding space can improve the prediction for unseen classes. For pre-training approaches to learn image representations from text, the methodology was quite simple yet effective [22], generally based on transformer-based language modeling [4], masked language modeling [23], and contrastive learning [41]. These methods generally aimed to learn a common space where the similarity between the image feature embedding and the corresponding label representation is maximized, or design the special mechanism (e.g., multi-head attention) to fuse them.

But, in PLL, the naive methodology in these methods is incapable of directly learning visual-semantic representations due to the inherent *label ambiguity*.

3 PRELIMINARIES

In this section, we introduce the symbols and terminologies to define the problem of partial label learning (PLL). Given a partially-labeled training dataset with n examples $\mathcal{D} = \{x_i, Y_i\}_{i=1}^{n}$, each pair consists of an example $x_i \in \mathcal{X}$ and a corresponding candidate label set $Y_i \in \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y} \in \{1, 2, ..., c\}$ denote the input space and the label space respectively. In addition, we denote the non-candidate label set by $\bar{Y}_i \in \mathcal{Y}$. This setting is different from the traditional supervised learning that the ground-truth label $y_i \in \mathcal{Y}$ is definitely known. Specially, there is a basic assumption [6] in PLL that the unseen ground-truth label y_i must be in the candidate label set Y_i (i.e., $P(y_i \in Y_i | x_i, Y_i) = 1$) and not in the non-candidate label set (i.e., $P(y_i \notin \bar{Y}_i | x_i, \bar{Y}_i) = 1$). The goal of PLL is also to learn a multi-class classifier that can predict the one true label of the unseen example. Obviously, it is degraded to directly utilize the partially-labeled dataset $\mathcal{D}$ to learn a classifier, due to the inherent *label ambiguity* in Y_i. Hence, PLL methods are necessarily equipped with the indispensable *label disambiguation* process that aims to identify the ground-truth label y_i from the candidate label set Y_i. Note that for simplicity, we may omit the subscript i sometimes.

For the conventional PLL with hand-crafted features $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, *label disambiguation* is a learning algorithm that aims to transform the partially-labeled training dataset $\mathcal{D} = \{\boldsymbol{X}_i, Y_i\}_{i=1}^{n}$ into the common supervised form $\widetilde{\mathcal{D}} = \{\boldsymbol{X}_i, \widetilde{y}_i\}_{i=1}^{n}$, where $\widetilde{y}_i$ is a selected label from Y_i. After that, a wieldy classifier can be induced with the appropriate loss function ℓ (e.g. square loss function) with the disambiguated dataset $\widetilde{\mathcal{D}}$.

For deep PLL with a end-to-end training manner, without pre-acquired features, one need train a deep neural network (including a feature extractor $f(\cdot) : \mathcal{X} \to \mathbb{R}^d$ and a classifier $h(\cdot) : \mathbb{R}^d \to [0, 1]^c$) and simultaneously perform label disambiguation. In this case, the aforementioned label disambiguation based on well hand-crafted features is impractical here, as the feature extractor $f(\cdot)$ in the early stages is no qualified to directly provide good enough feature embeddings to reciprocate label disambiguation. The alternative is to integrate them into a iterative and synergistic framework that executes label disambiguation in a mini-batch and learns increasingly good representations by using disambiguated labels. Specifically, each image x is generally associated with a label confidence vector $\boldsymbol{p} \in [0, 1]^c$ where each entry denotes the probability of the corresponding label being the ground-truth. The update of $\boldsymbol{p}$ reflects the process of label disambiguation. Ideally, the j-th candidate label with weight 1 is exactly the true label, i.e., j is the ground-truth label of x if $\boldsymbol{p}_j = 1$, which means we have identified the true label successfully. Synchronously, $\boldsymbol{p}$ serves as coarse supervision to calculate the classification loss (e.g., cross-entropy loss) and update networks ($f(\cdot)$ and $h(\cdot)$). The per-example classification loss is defined as:

$$\ell_{CLS}(f, h; x, \boldsymbol{p}) = \sum_{j=1}^{c} -\boldsymbol{p}_j \log(h_j(x)) \tag{1}$$

In our framework ParSE, we have access to a semantic label representation lookup table $\boldsymbol{S} \in \mathbb{R}^{c \times s}$ where s is the dimension of label representation, which is acquired by encoding the word of each label with a pre-trained language model (e.g., bert).

4 METHOD

In this section, we introduce our framework ParSE in detail that consists of visual-semantic representation learning and powerful label disambiguation (as shown in Figure 2). These two processes perform iteratively and mutually proceed. We first introduce visual-semantic representation learning and then detail powerful label disambiguation.

4.1 Visual-semantic Representation Learning

As mentioned above, it is of vital importance to learn a high-quality representation to promote a virtuous cycle in PLL. In ParSE, we achieve this goal by utilizing semantic label representations $\boldsymbol{S}$ to learn visual-semantic feature representations that are expected to be sufficiently distinguishable for label disambiguation. Specifically,

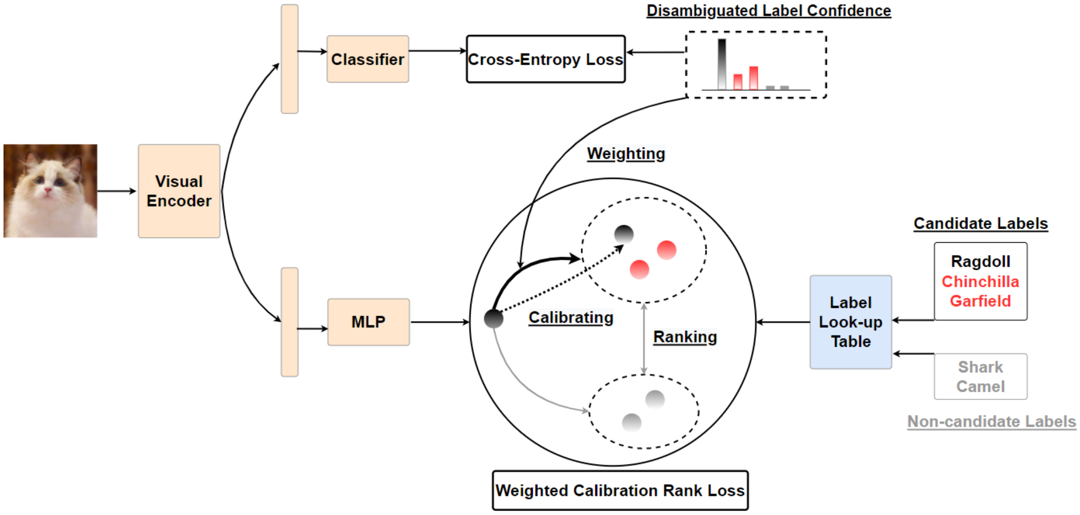

Figure 2: Our proposed framework ParSE consists of two synergistic process—visual-semantic representation learning and powerful label disambiguation. First, an image of ragdoll (the ground-truth label) is encoded into a feature embedding by a visual encoder, and then is mapped into a semantic space (black ellipse) with L2 normalization. In the visual-semantic space, the similarity between the visual embedding and semantic label representations of all candidate labels is weighted by the disambiguated label confidence.

we employ a transformation layer $g(\cdot)$ to map the feature embedding $f(x)$ into a common space (with L_2 normalization). This is identical to many other tasks, e.g., zero-shot learning and multi-modal learning, which also aim to learn visual-semantic representations [7, 24]. After that, they generally directly maximize the dot-product similarity (i.e., cosine similarity) between the visual feature embedding and the corresponding semantic label representation, i.e., $g(x)^\top \boldsymbol{S}_j$ where the j-th label is the ground-truth one. However, in PLL, *label ambiguity* in the candidate label set Y provides inherent resistance to such objective due to the unknown ground-truth label. A direct approach is the partial-level dot-product similarity for all candidates:

$$\sum_{j \in Y} g(x)^\top \boldsymbol{S}_j \tag{2}$$

Obviously, Eq. (2) is incapable of learning sufficiently distinguishable visual-semantic representations for label disambiguation, as it treats the similarity of candidates equally. To solve this issue, we provide an indirect progressive calibration strategy that uses the label confidence vector $\boldsymbol{p}$ to weight the partial-level dot-product similarity:

$$\sum_{j \in Y} p_j \cdot g(x)^\top \boldsymbol{S}_j \tag{3}$$

The label confidence vector $\boldsymbol{p}$ is constantly updating as the process of label disambiguation. Hence, Eq. (3) tends to the right objective progressively (i.e., $g(x)^\top \boldsymbol{S}_j$). In this way, the partial-level similarity Eq. (2) is gradually calibrated to the true similarity $g(x)^\top \boldsymbol{S}_j$.

But, Eq. (3) only focuses on the similarity in the candidate label set Y, and ignores the correlation between Y and $\bar{Y}$. Many existing works in PLL have a consideration on the correlation of losses and outputs between Y and $\bar{Y}$ [3, 31]. In our case, we pursue it at the level of visual-semantic representation learning. Since candidate labels in Y generally have more similar semantic concepts than these non-candidate ones in $\bar{Y}$, it is expected to maintain a higher dot-product similarity of them than that of each non-candidate one in $\bar{Y}$. It is a ranking problem [7] and we formulate it with a hinge rank loss.

Finally, we combine them to propose a novel weighted calibration rank loss. The per training example loss is defined as:

$$\ell_{SEL}(x, \boldsymbol{S}, g) = \sum_{k \in \bar{Y}} \max[0, \sigma - \sum_{j \in Y} p_j \cdot g(x)^\top \boldsymbol{S}_j + g(x)^\top \boldsymbol{S}_k] \tag{4}$$

where σ is a fixed margin. Eq. (4) implies two levels of meanings as discussed before. First, the term $\sum_{j \in Y} p_j \cdot g(x)^\top \boldsymbol{S}_j$ denotes the whole similarity towards all candidate labels weighted by a label confidence vector $\boldsymbol{p}$. Second, the whole candidate similarity is expected to have a higher value than the similarity of each non-candidate one $g(x)^\top \boldsymbol{S}_k$. The overall loss is:

$$\ell = \ell_{CLS} + \beta \ell_{SEL} \tag{5}$$

where β is a trade-off parameter to balance these two losses.

4.2 Label Disambiguation

As mentioned above, label disambiguation is an indispensable means of combating label ambiguity in PLL. Different from conventional PLL methods that carries out label disambiguation once and globally, deep PLL methods generally do it progressively in a mini-batch. For an example x in a mini-batch, we generally denote its label confidence vector by $\boldsymbol{p}$ that reflects the process of label disambiguation. Ideally, when the j-th coordinate of $\boldsymbol{p}$ becomes 1, i.e., $p_j = 1$, it means we have identified the true label $\bar{Y}_j$ from candidate ones successfully. To achieve this aim, we update $\boldsymbol{p}$ by the following

Algorithm 1: ParSE

```
Input: $\mathcal{D}$, model $f(\cdot)$, $h(\cdot)$ and $g(\cdot)$, $\boldsymbol{S}$, label confidence $\boldsymbol{P}$,
       parameters: $\beta$, $\sigma$, $T_{max}$
Output: model parameters
for $t < T_{max}$ do
    Fetch a mini-batch $\mathcal{B}$ from $\mathcal{D}$;
    for $x \in \mathcal{B}$ do
        Obtain feature embedding $f(x)$ and output $h(x)$;
        Calculate $\ell_{CLS}$ by Eq.(1) using $h(x)$ and $\boldsymbol{p}$;
        Normalize $f(x)$ and corresponding $\boldsymbol{S}$;
        Calculate candidate similarity: $\sum_{j \in Y} p_j \cdot g(x_i)^\top \boldsymbol{S}_j$;
        Calculate each non-candidate similarity: $g(x)^\top \boldsymbol{S}_k$;
        Calculate $\ell_{SEL}$ by Eq.(4);
        Update $\boldsymbol{p}$ by Eq.(6);
        Minimize $\ell = \ell_{CLS} + \beta \ell_{SEL}$;
    end
end
```

fashion:

$$p_i = \begin{cases} \frac{h_i(x)}{\sum_{j=1}^{c} h_j(x)}, & if \quad i \in Y, \\ 0, & if \quad i \in \bar{Y} \end{cases} \tag{6}$$

Where $h_i(x)$ is the i-th coordinate of the classifier output $h(x)$. In this way, more weights are flowing to more possible candidate labels slightly and progressively, while the weights of non-candidate labels $\bar{Y}$ are always zero. Although Eq. (6) is a common label disambiguation means [6, 17], ParSE endows it with more powerful ability of label disambiguation, as learned visual-semantic feature representations $f(x)$ can produce more discriminative outputs $h(x)$, which is beneficial for Eq. (6). The pseudo-code of ParSE is shown in Algorithm 1.

5 EXPERIMENTS

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed method. Then, we further provide hyper-parameter sensitivity analysis and ablation studies.

5.1 Setup

Datasets. We first evaluate ParSE on two benchmarks CIFAR-10 and CIFAR-100 [13]. Specially, the 100 classes in the CIFAR-100 are grouped into 20 super-classes. Similar to previous work [6, 17], we use a uniform label flipping probability q to generate the candidate label set. Larger q denotes higher label ambiguity degree. In our experiment, we set different values of q to show the effectiveness of ParSE under different label ambiguity degree, i.e., $q \in \{0.1, 0.3, 0.5\}$ for CIFAR-10 and $q \in \{0.01, 0.05, 0.1\}$ for CIFAR-100. Further, we also consider more challenging label ambiguity in CIFAR-100 with hierarchical labels (termed CIFAR-100-H) and CUB-200 [30] (a fine-grained dataset). The CUB-200 is the most widely-used dataset for fine-grained visual categorization task. It contains 11,788 images of 200 subcategories belonging to birds, 5,994 for training and 5,794 for testing. In this case, some similar classes in a super-class are selected into the candidate label set, which makes label disambiguation extremely difficult.

Comparing Methods. Two simple baselines and three state-of-the-art deep partial label learning algorithms are compared with ParSE: (1) LWS [31] introduces a leverage parameter to consider the trade-off between losses on partial labels and non-partial ones; (2) PRODEN [17] progressively identifies the ground-truth label with a self-training style manner; (3) CC [6] is a classifier-consistent method that assumes set-level uniform data generation process; (4) MSE and EXP [6] are two simple baselines that adopt mean square error and exponential loss. The hyper-parameters are tuned according to the original methods. Specially, we also report the results of the supervised counterpart.

Implementation We use an 18-layer ResNet [9] as the backbone for feature extraction $f(\cdot)$. The transformation layer is a 2-layer MLP (with a relu activation). The parameter β is selected from [0.01, 0.05, 0.1, 0.5 ,1 , 5], and the parameter σ is selected from [0.01, 0.05, 0.1, 0.15, 0.2, 0.25]. We use a standard SGD optimizer with a momentum of 0.9, and weight decay is 0.001 and learning rate is 0.01. The batch size is 128 (256 for CIFAR-10). We train the model for 300 epochs with cosine learning rate scheduling. We utilize the widely-used pre-trained model bert [12] to encode the word of label in different datasets to obtain the corresponding label look-up table where the dimension of semantic label representation is 768. Specially, the original comparing methods do not use data augmentation technologies. Hence, for a fair comparison, in CIFAR-10, all methods do not employ any data augmentation, while for other datasets they use the same data augmentation technology. For all experiments, we report the mean and standard deviation (mean$\pm$std) based on 5 trials.

5.2 Experimental Results

Accuracy comparison. As shown in Table 1, ParSE significantly outperforms all counterparts on CIFAR-10 and CIFAR-100 under all cases. Specifically, with the increase of label ambiguity, i.e., the larger value of partial rate, the counterparts show a significant performance drop, while ParSE consistently maintains superior results. That is to say ParSE shows more advantages at the high level of label ambiguity. For more challenging label ambiguity on CIFAR-100-H and CUB-200, as shown in Table 2, ParSE also achieves superior performances than all counterparts. Note that on CIFAR-100, CIFAR-100-H and CUB-200, both ParSE and PRODEN achieve comparable performance under low label ambiguity, as the number of candidate labels in these cases is quite small, while under high label ambiguity, ParSE significantly outperforms PRODEN (e.g., 1.48% on CIFAR-100 with q=0.1, 1.35% on CUB-200 with q=0.1, and 15.04% on CIFAR-100-H with q=0.8).

Accuracy curves. we record the test accuracy at each training epoch to provide more detailed visualized results. As shown in Figure 3, ParSE (red) is compared with three different methods. We can observe that ParSE consistently outperforms PRODEN (blue) and achieves comparable performance to the supervised counterpart (purple) under low partial rates. The naive (black) means the original method without label disambiguation, which drops the performance seriously. This immediately validates the importance of label disambiguation.

Visualization of learnd representations. Since the key of ParSE is to learn visual-semantic representations, we visualize the

Table 1: Accuracy comparisons on CIFAR-10 and CIFAR-100. Bold indicates superior results.

Dataset	Method	Partial Rate		
		0.1	0.3	0.5
CIFAR-10	Supervised	$86.79 \pm 0.08\%$		
	MSE	$70.65 \pm 0.82\%$	$58.06 \pm 0.56\%$	$52.58 \pm 0.34\%$
	EXP	$71.32 \pm 0.12\%$	$58.13 \pm 0.28\%$	$53.02 \pm 0.44\%$
	CC	$80.56 \pm 0.10\%$	$71.72 \pm 0.60\%$	$59.17 \pm 0.29\%$
	PRODEN	$81.89 \pm 0.18\%$	$72.60 \pm 0.25\%$	$61.01 \pm 0.22\%$
	LWS	$81.25 \pm 0.32\%$	$71.35 \pm 0.45\%$	$60.87 \pm 0.13\%$
	ParSE (ours)	$\mathbf{83.63 \pm 0.20\%}$	$\mathbf{74.85 \pm 0.10\%}$	$\mathbf{64.20 \pm 0.18\%}$
Dataset	Method	Partial Rate		
		0.01	0.05	0.1
CIFAR-100	Supervised	$77.85 \pm 0.012\%$		
	MSE	$63.32 \pm 0.32\%$	$61.45 \pm 0.42\%$	$58.87 \pm 0.53\%$
	EXP	$58.46 \pm 1.23\%$	$53.25 \pm 0.69\%$	$48.76 \pm 1.40\%$
	CC	$64.17 \pm 0.10\%$	$62.08 \pm 0.30\%$	$56.39 \pm 0.81\%$
	PRODEN	$76.81 \pm 0.25\%$	$75.05 \pm 0.13\%$	$71.91 \pm 0.21\%$
	LWS	$75.53 \pm 0.05\%$	$74.26 \pm 0.51\%$	$69.42 \pm 0.65\%$
	ParSE (ours)	$\mathbf{76.96 \pm 0.23\%}$	$\mathbf{75.86 \pm 0.10\%}$	$\mathbf{73.43 \pm 0.15\%}$

Table 2: Accuracy comparisons on CIFAR-100-H and CUB-200. Bold indicates superior results.

Dataset	Method	Partial Rate		
		0.1	0.5	0.8
CIFAR-100-H	Supervised	$77.85 \pm 0.012\%$		
	MSE	$60.21 \pm 0.35\%$	$54.10 \pm 0.89\%$	$49.50 \pm 1.20\%$
	EXP	$61.52 \pm 0.88\%$	$53.47 \pm 1.20\%$	$48.17 \pm 0.90\%$
	CC	$64.17 \pm 0.57\%$	$61.32 \pm 0.40\%$	$58.20 \pm 1.09\%$
	PRODEN	$76.80 \pm 0.28\%$	$75.03 \pm 0.35\%$	$53.20 \pm 0.60\%$
	LWS	$76.21 \pm 0.50\%$	$73.21 \pm 0.22\%$	$67.46 \pm 0.18\%$
	ParSE (ours)	$\mathbf{77.34 \pm 0.17\%}$	$\mathbf{75.31 \pm 0.04\%}$	$\mathbf{68.24 \pm 0.33\%}$
Dataset	Method	Partial Rate		
		0.01	0.05	0.1
CUB-200	Supervised	$75.73 \pm 0.07\%$		
	MSE	$63.21 \pm 0.58\%$	$51.27 \pm 0.48\%$	$28.33 \pm 1.07\%$
	EXP	$61.01 \pm 0.09\%$	$49.90 \pm 0.80\%$	$31.23 \pm 0.87\%$
	CC	$67.30 \pm 0.17\%$	$55.18 \pm 0.35\%$	$37.20 \pm 0.28\%$
	PRODEN	$75.06 \pm 0.30\%$	$67.21 \pm 0.21\%$	$40.16 \pm 0.12\%$
	LWS	$74.89 \pm 0.11\%$	$64.50 \pm 0.17\%$	$35.26 \pm 0.32\%$
	ParSE (ours)	$\mathbf{75.41 \pm 0.15\%}$	$\mathbf{67.98 \pm 0.25\%}$	$\mathbf{41.51 \pm 0.10\%}$

learned image representation of both PRODEN and ParSE to show the superiority of our method. Specifically, we use t-SNE [27] to reduce the dimension of learned representations by two dimensions, and then plot them as a scatter diagram. Different colors represent different labels in CIFAR-10. From Figure 4, we can observe that the figure of PRODEN has many false outliers and class overlapping, i.e., some examples are disambiguated falsely. In contrast, ParSE produces more distinguishable representations with less outliers, which validates the effectiveness of ParSE to learn high-quality representations. This result also shows that the label confidence of ParSE is more accurate than that of PRODEN (due to less outliers), which is beneficial for subsequent label disambiguation. Meanwhile, it also owes to the learned compact representations that produces more discriminative outputs for label disambiguation.

Results of label disambiguation. As discussed in the aforementioned section, label disambiguation in ParSE is more powerful than that in conventional methods (e.g, PRODEN) due to learned visual-semantic representations. To show this, we plot the accuracy of the label confidence $\boldsymbol{P}$ at each training epoch. From Figure 5, we can observe that ParSE (red) can achieve superior performance in label

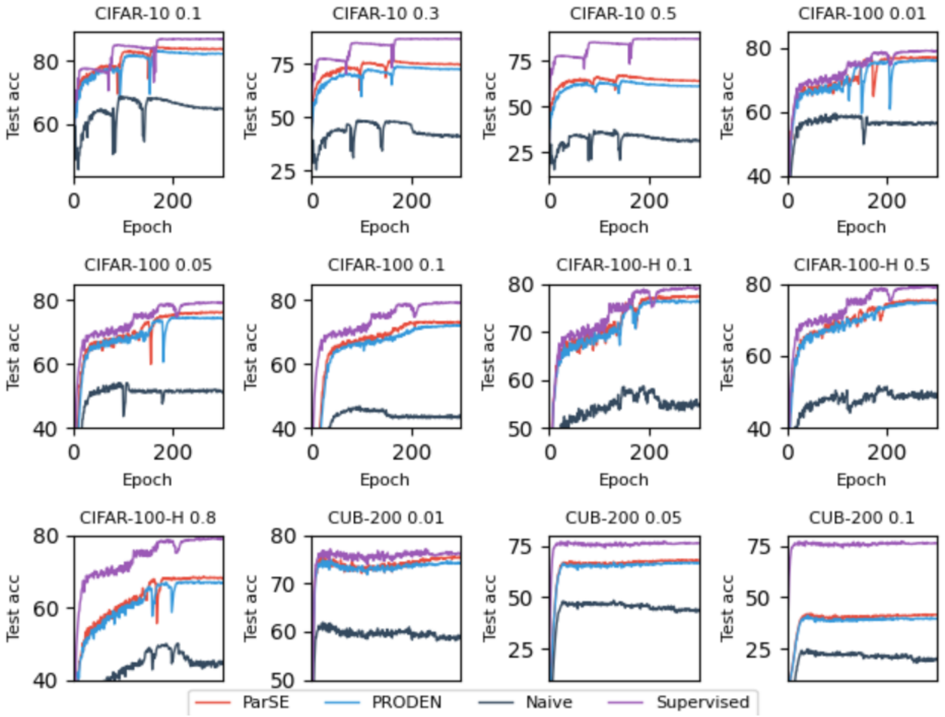

Figure 3: Test accuracy for various methods on different datasets.

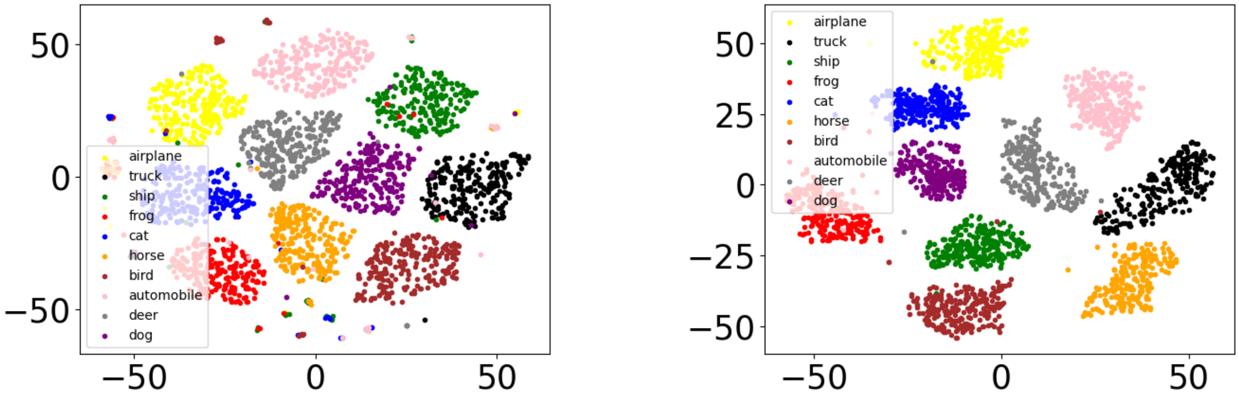

Figure 4: T-SNE visualization of the image representation on CIFAR-10 (q=0.1). Different colors represent the corresponding classes. The left (right) is the image feature representation produced by PRODEN (ParSE).

disambiguation than PRODEN (blue). Specially, with high label ambiguity, ParSE also maintain the superiority by a significant margin. This also shows the superiority of learnd visual-semantic representations, because both ParSE and PRODEN utilize the same way, i.e., Eq. (6), except for employed learned representations $f(\cdot)$. This also shows the synergistic effect of this framework. That is to say that learning high-quality distinguishable representations promotes a virtuous circle between these two process in this framework.

5.3 Ablation Studies

In this section, we present part of our ablation results to show the effectiveness of ParSE. As shown in Table 3, ParSE w/o LD means the naive PLL algorithm without label disambiguation. ParSE w/o ℓ_{SEL} equals to set $\beta = 0$, which is degraded into PRODEN. ParSE w/o weight means that we directly utilize the similarity between the visual representation and semantic label representations of all candidates with the weight set to 1. In this way, it maximizes

Table 3: Ablation study on CIFAR-10. LD means label disambiguation.

Ablation	ℓ_{SEL}	LD	Partial Rate 0.1	 0.3	 0.5
ParSE w/o LD	×	×	67.84 ± 0.55%	40.15 ± 0.18%	28.47 ± 0.34%
ParSE w/o ℓ_{SEL}	×	✓	81.89 ± 0.18%	72.60 ± 0.25%	61.01 ± 0.22%
ParSE w/o weight	×	✓	78.88 ± 0.32%	68.42 ± 0.34%	58.65 ± 0.56%
ParSE	✓	✓	**83.63 ± 0.20%**	**74.85 ± 0.10%**	**64.20 ± 0.18%**

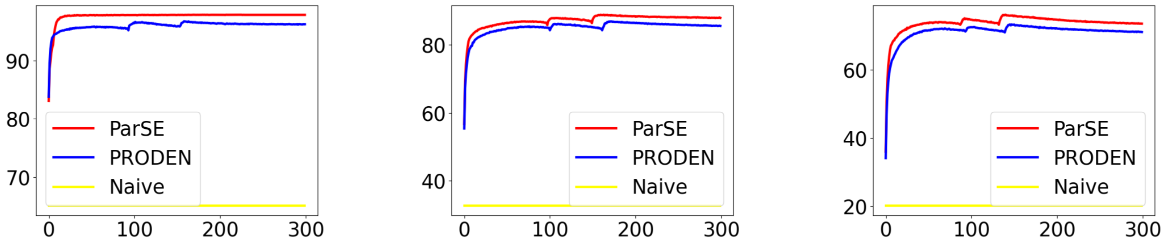

Figure 5: Confidence accuracy on CIFAR-10 with $q = (0.1, 0.3, 0.5)$

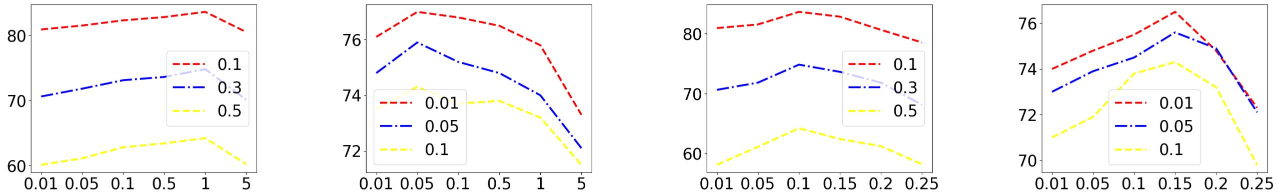

Figure 6: Hyper-parameter Analysis of varying β (the left two) and σ (the right two) on CIFAR-10 and CIFAR-100 respectively.

the similarity between the image visual feature embedding and all corresponding semantic label representations, which seriously influence the distinguishable representation learning and thus degrade the performance of label disambiguation dramatically. From Table 3, we can observe that without label disambiguation, the performance drops dramatically (nearly 15.79% with $q = 0.1$, 34.7% with $q = 0.3$ and 35.73% with $q = 0.5$), which signifies the importance of label disambiguation in PLL. Moreover, with ℓ_{SEL}, the performance has improved significantly (nearly 1.74% with $q = 0.1$, 2.25% with $q = 0.3$ and 3.19% with $q = 0.5$), which immediately certify the effectiveness of using semantic label representations to learn visual-semantic feature representations. Furthermore, without the weight, the performance is degraded significantly, which implies the importance of calibration. These results validate the effectiveness of ParSE.

5.4 Hyper-parameter Analysis

Effect of β. This parameter controls the trade-off between the two losses. From the left two figures in Figure 6, on CIFAR-10, small values of β have little improvements. On CIFAR-100, large values of β degrade the performance. This implies that magnifying the influence of visual-semantic learning too much is not always better, as it may overemphasize to fit the semantic label representation, which could influence the ordinal visual feature presentation learning. Hence, it is significant to select the appropriate parameter in practice.

Effect of σ. This parameter is the margin to control the two similarity in Eq. (4). From the right two figures in Figure 6, smaller σ means looser restriction on the ranking of similarity, which may not achieve satisfying effects. Otherwise, larger σ implies tighter restriction on the ranking of similarity, which may magnify the error ranking. Therefore, it is very important to choose the appropriate value in practice.

6 CONCLUSION

In this work, we proposed a novel partial-label learning framework called ParSE. The key idea of ParSE is to learn visual-semantic representations to promote label disambiguation. For this purpose, we proposed a novel weighted calibration rank loss that plays two significant roles. First, it utilized the label confidence to weight the similarity towards all candidates, which was progressively calibrated to the right objective. Second, it produced a higher similarity of candidates than that of each non-candidate one. Subsequently, label disambiguation was desirably endowed with more powerful abilities based on learn visual-semantic representations. Empirically, we conducted extensive experiments on benchmarks and showed that ParSE outperformed state-of-the-art counterparts. In ParSE, since semantic label representations were obtained by the pre-trained language model and utilized independently from model training, they were not optimized to PLL. Therefore, it would be interesting to automatically learn semantic label representations from large textual corpora for PLL.

7 ACKNOWLEDGEMENT

The authors wish to thank the anonymous reviewers for their helpful comments. Lei Feng was supported by the National Natural Science Foundation of China (Grant No. 62106028) and CAAI-Huawei MindSpore Open Fund. Fengmao Lv was supported by the National Natural Science Foundation of China (Grant No. 62106204), the Fundamental Research Funds for the Central Universities of China (No. 2682022CX068), Sichuan Science and Technology Program (No. 2021YFS0178). Guowu Yang was supported by the National Natural Science Foundation of China (Grant No. 62172075).

REFERENCES

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2209–2218.

[2] Timothee Cour, Benjamin Sapp, Chris Jordan, and Ben Taskar. 2009. Learning from ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 919–926.

[3] Timothee Cour, Ben Sapp, and Ben Taskar. 2011. Learning from partial labels. *Journal of Machine Learning Research* 12, 5 (2011), 1501–1536.

[4] Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 11162–11173.

[5] Lei Feng and Bo An. 2019. Partial Label Learning by Semantic Difference Maximization.. In *Proceedings of the International Joint Conference on Artificial Intelligence.* 2294–2300.

[6] Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. 2020. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems.*

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems* 26 (2013).

[8] Hao Guo, Xiangyang Li, Lei Zhang, Jia Liu, and Wei Chen. 2021. Label-Aware Text Representation for Multi-Label Text Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 7728–7732.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision.* 630–645.

[10] Eyke Hüllermeier and Jürgen Beringer. 2006. Learning from ambiguously labeled examples. *Intell. Data Anal.* 10, 5 (2006), 419–439.

[11] Rong Jin and Zoubin Ghahramani. 2002. Learning with multiple labels. *Advances in Neural Information Processing Systems* 15.

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 4171–4186.

[13] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[14] Yu-Feng Li, Lan-Zhe Guo, and Zhi-Hua Zhou. 2019. Towards safe weakly supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 334–346.

[15] Liping Liu and Thomas Dietterich. 2012. A conditional multinomial mixture model for superset label learning. *Advances in Neural Information Processing Systems* 25.

[16] Jie Luo and Francesco Orabona. 2010. Learning from candidate labeling sets. *Advances in Neural Information Processing Systems* 23.

[17] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. 2020. Progressive identification of true labels for partial-label learning. In *Proceedings of the International Conference on Machine Learning.* 6500–6510.

[18] Gengyu Lyu, Songhe Feng, Tao Wang, and Congyan Lang. 2020. A self-paced regularization framework for partial-label learning. *IEEE Transactions on Cybernetics* (2020).

[19] Gengyu Lyu, Songhe Feng, Tao Wang, Congyan Lang, and Yidong Li. 2019. GM-PLL: graph matching based partial label learning. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[20] Nam Nguyen and Rich Caruana. 2008. Classification with partial labels. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 551–559.

[21] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. *Advances in Neural Information Processing Systems* 22 (2009).

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning.* PMLR, 8748–8763.

[23] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning visual representations with caption annotations. In *Proceedings of the European Conference on Computer Vision.* Springer, 153–170.

[24] Gabi Shalev, Yossi Adi, and Joseph Keshet. 2018. Out-of-distribution detection using multiple semantic label representations. *Advances in Neural Information Processing Systems* 31 (2018).

[25] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems* 26 (2013).

[26] Cai-Zhi Tang and Min-Ling Zhang. 2017. Confidence-rated discriminative partial label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 31.

[27] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[28] Deng-Bao Wang, Min-Ling Zhang, and Li Li. 2021. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[29] Haobo Wang, Chen Chen, Weiwei Liu, Ke Chen, Tianlei Hu, and Gang Chen. 2020. Incorporating label embedding and feature augmentation for multi-dimensional classification. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 34. 6178–6185.

[30] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD birds 200. (2010).

[31] Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. 2021. Leveraged Weighted Loss for Partial Label Learning. In *Proceedings of the International Conference on Machine Learning.*

[32] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.* 466–475.

[33] Ning Xu, Jiaqi Lv, and Xin Geng. 2019. Partial label learning via label enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 33. 5557–5564.

[34] Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. 2021. Instance-Dependent Partial Label Learning. *Advances in Neural Information Processing Systems* 34.

[35] Yan Yan and Yuhong Guo. 2020. Partial label learning with batch label correction. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 34. 6575–6582.

[36] Dani Yogatama, Dan Gillick, and Nevena Lazic. 2015. Embedding methods for fine grained entity type classification. In *Annual Meeting of the Association for Computational Linguistics.* 291–296.

[37] Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, and Yi Ma. 2013. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 708–715.

[38] Min-Ling Zhang and Fei Yu. 2015. Solving the partial label learning problem: An instance-based approach. In *Proceedings of the International Joint Conference on Artificial Intelligence.*

[39] Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1335–1344.

[40] Qian-Wen Zhang, Ximing Zhang, Zhao Yan, Ruifang Liu, Yunbo Cao, and Min-Ling Zhang. 2021. Correlation-Guided Representation for Multi-Label Text Classification. In *Proceedings of the International Joint Conference on Artificial Intelligence.*

[41] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *Arxiv* (2020).

[42] Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review* 5, 1 (2018), 44–53.