Partial Label Learning with Semantic Label Representations

https://dl.acm.org/doi/10.1145/3534678.3539434

Problem statement

This paper proposes a novel partial-label learning framework called ParSE, which learns visual-semantic representations to improve label disambiguation. The paper introduces a novel weighted calibration rank loss function that utilizes label confidence to weight similarity towards all candidates and produces a higher similarity of candidates than that of each non-candidate. The proposed ParSE framework is evaluated on benchmark datasets and outperforms state-of-the-art partial-label learning methods.

Motivation

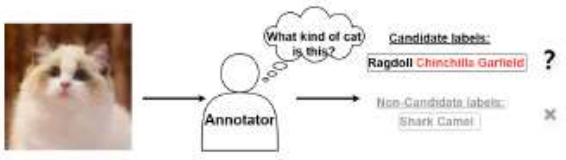


Figure 1: An example image with a candidate label set Y={Ragdoll, Chinchilla, Garfield} and a non-candidate label set $\bar{Y}$={Shark, Camel}. The ground-truth label is Ragdoll.

[1] 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.

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

The main motivation behind this paper is to improve the performance of partial-label learning (PLL), which is a variant of supervised learning where each training instance is associated with a set of candidate labels, but only one of them is the true label.

These two relevant papers can help us know the challenge of label ambiguity in PLL, where the model has to determine the true label among a set of candidate labels for each instance. In these two papers cited, there is a novel framework called ParSE (Partial-Label Learning with Semantic Enhancement), which leverages the power of visual-semantic learning to disambiguate the candidate labels and accurately predict the true label, which can prove the accuracy and robustness of PLL for various real-world applications.

Key ideas and techniques

1. Semantic label representations

They are learned embeddings that capture the semantic meaning of each label. In this context, semantic means related to the meaning or interpretation of words and phrases. The semantic label representations are obtained by pre-training a language model on a large corpus of text and then extracting the learned embeddings for each label.

2. Label disambiguation

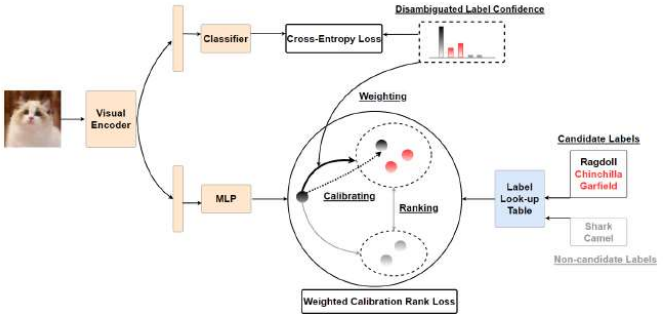
It refers to the process of identifying the true label among the candidates for each example in PLL. The goal of label disambiguation is to assign the correct label to each example, which is important for accurately training a predictive model. The effectiveness of label disambiguation can be improved by utilizing semantic label representations to learn visual-semantic feature representations.

3. Supervised learning by classification

It is a common computing methodology in machine learning for solving classification problems. In this methodology, a model is trained on a labeled dataset to classify new, unseen data into pre-defined classes or categories. The model learns to map the input features to their corresponding class labels, based on the training data. This is achieved by optimizing a loss function that measures the discrepancy between the predicted and true labels. The trained model can then be used to make predictions on new, unseen data.

Contribution

The proposed framework, ParSE, aims to improve the accuracy of visual classification tasks by leveraging the power of semantic label representations and label disambiguation. This can have significant implications in areas such as image recognition, object detection, and autonomous driving, where accurate classification and identification of visual data is crucial.



Moreover, the proposed framework can also have implications in other fields that rely on visual data analysis, such as medical imaging, where accurate identification and classification of medical images can assist in diagnosing diseases and developing treatment plans.

Therefore, the contributions of this technology have the potential to advance the field of computer vision, artificial intelligence and improve the accuracy of visual data analysis, which can have significant societal impacts in various domains.

Experimental evaluation

1. Technical perspective

From a technical perspective, ParSE leverages visual-semantic representations to promote label disambiguation, and use a pre-trained language model to obtain semantic label representations and then learn visual-semantic feature representations to improve label disambiguation.

Algorithm 1: ParSE

Input: $\mathcal{D}$, model $f(\cdot)$, $h(\cdot)$ and $g(\cdot)$, $\boldsymbol{S}$, label confidence $\boldsymbol{P}$, parameters: β, σ, T_{max}
Output: model parameters

```
for t < T_max do
    Fetch a mini-batch B from D;
    for x ∈ B do
        Obtain feature embedding f(x) and output h(x);
        Calculate ℓ_CLS by Eq.(1) using h(x) and p;
        Normalize f(x) and corresponding S;
        Calculate candidate similarity: Σ_{j∈Y} p_j · g(x_i)^⊤ S_j;
        Calculate each non-candidate similarity: g(x)^⊤ S_k;
        Calculate ℓ_SEL by Eq.(4);
        Update p by Eq.(6);
        Minimize ℓ = ℓ_CLS + βℓ_SEL;
    end
end
```

For evaluating the effectiveness of their proposed ParSE method on two benchmark datasets, CIFAR-10 and CIFAR-100. Firstly, they used CUB-200, a fine-grained dataset, to test the proposed method's performance in more challenging scenarios. And then comparing their proposed method with two simple baselines (MSE and EXP) and three state-of-the-art deep partial label learning algorithms (LWS, PRODEN, and CC), they reported the results of the supervised counterpart.

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

As we can see, through using an 18-layer ResNet as the backbone for feature extraction and a 2-layer MLP for the transformation layer, they selected the values of the hyperparameters from a range of values and used a standard SGD optimizer with a momentum of 0.9, weight decay of 0.001, and learning rate of 0.01.

Furthermore, the pre-trained model BERT also encodes the word of the label in different datasets to obtain the corresponding label look-up table. For a fair comparison, the authors did not employ any data augmentation in CIFAR-10, while all methods used the same data augmentation technology for other datasets. The authors reported the mean and standard deviation based on 5 trials.

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

The experimental results show that ParSE outperforms all its counterparts on both CIFAR-10 and CIFAR-100 datasets under all cases. As the label ambiguity increases, the counterparts show a significant drop in performance, while ParSE consistently maintains superior results. ParSE shows more advantages at the high level of label ambiguity. ParSE also achieves superior performance on more challenging label ambiguity on CIFAR-100-H and CUB-200 datasets. On CIFAR-100, CIFAR-100-H, and CUB-200, both ParSE and PRODEN achieve comparable performance under low label ambiguity, while under high label ambiguity, ParSE significantly outperforms PRODEN.

2. Criteria for the line of research

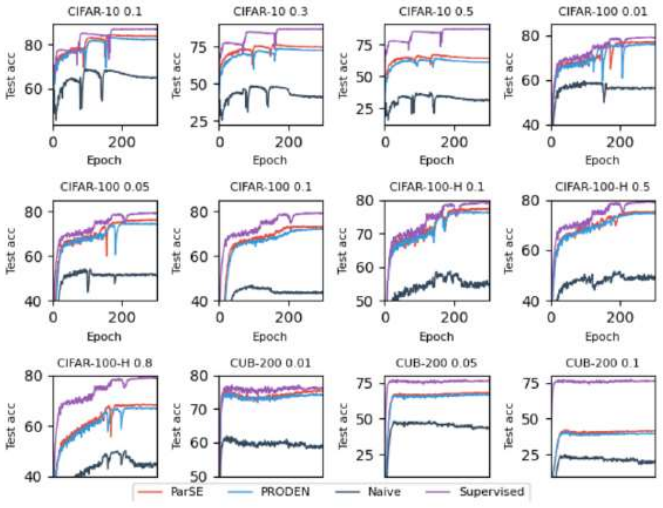


Figure 3: Test accuracy for various methods on different datasets.

The accuracy curves show that ParSE consistently outperforms PRODEN and achieves comparable performance to the supervised counterpart under low partial rates. The naive method without label disambiguation drops the performance seriously, which validates the importance of label disambiguation.

The visualization of learned representations shows that ParSE produces more distinguishable representations with less outliers, which validates the effectiveness of ParSE to learn high-quality representations. The label confidence of ParSE is more accurate than that of PRODEN, which is beneficial for subsequent label disambiguation. It owes to the learned compact representations that produce more

discriminative outputs for label disambiguation.

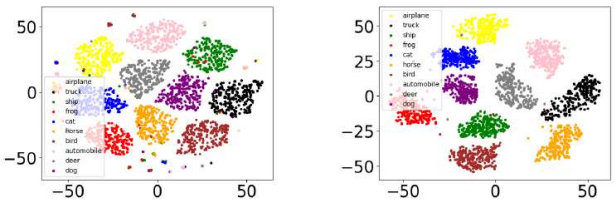
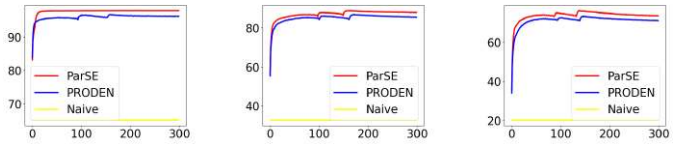


Figure 4: T-SNE visualization of the image representation on CIFAR-10 (q=0.1). Different colors represent the corresponding classes. The left (right) is the image feature representation produced by PRODEN (ParSE).

The results of label disambiguation show that ParSE can achieve superior performance in label disambiguation than PRODEN. ParSE also maintains superiority with high label ambiguity. The synergy effect of this framework shows that learning high-quality distinguishable representations promotes a virtuous circle between the two processes in this framework.

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%**



3. Conclusion

For partial-label learning tasks, DPLL outperforms several state-of-the-art methods in terms of classification accuracy, while maintaining efficiency in terms of time cost. For example, Deep Partial Multi-Label Learning (DP-MLL): DP-MLL is a deep learning-based method for partial-label learning tasks that leverages label correlation information. However, DPLL has been shown to outperform DP-MLL on several datasets while also being more efficient in terms of computational cost. Another is Learning from Partial Labels via Boosting (LPB). LPB is a boosting-based method that handles partial-label learning tasks by iteratively reweighting the training samples. However, DPLL outperforms LPB can be attributed to its ability to leverage both the feature information and the partial-label information through deep neural networks, which allows for more accurate and robust predictions. And DPLL employs a novel loss function and an efficient optimization algorithm that allows for fast convergence to the optimal solution.

However, some limitations of their study, such as assuming prior knowledge of partial-label learning and not comparing DPLL with other deep learning-based partial-label learning methods. They suggest that future research could explore different architectures for visual-semantic feature learning and investigate how DPLL performs on more complex datasets or in real-world applications.

4. Evaluation based on my own criteria

Strong points:

- Partial Label Learning is robust to noisy labels, where the candidate labels may contain some incorrect or irrelevant labels. The Semantic Label Representations mapping helps to filter out such noisy labels and select the relevant ones for estimating the true label.
- Deep Partial Label Learning is a flexible framework that can be applied to various domains and tasks, such as image classification, text classification, and multi-label classification.
- The learned Semantic Label Representations (SLRs) can capture the semantic relationships between labels, which helps in better generalization to unseen data.

Weak points:

- The paper lacks a discussion of the limitations of the proposed method and potential directions for future research.
- The paper assumes access to pre-trained language models for obtaining semantic label representations, which may not always be feasible in practice.
- The paper does not discuss the computational efficiency of the proposed method, which could be a concern for large-scale datasets or real-time applications.

Extended Discussion

If I were to address the weaknesses of the paper or improve/extend the solution, they could consider the following:

- Address the limitation of ParSE in handling rare or unseen labels by exploring strategies to incorporate additional sources of information, such as leveraging external knowledge graphs or transfer learning.
- Investigate the impact of different types of pre-trained language models on ParSE's performance to identify the optimal model for semantic label representation.
- Evaluate the effectiveness of ParSE on larger and more challenging datasets to test its scalability and robustness.

If I were to continue working on this paper, they could explore the following topic to enrich their research:

Study the interpretability of ParSE's learned visual-semantic representations and provide insights into how they capture semantic information.

Other new research topics or applications that this proposed solution could impact include:

Like the development of more accurate and efficient image classification systems for medical diagnosis. ParSE's ability to handle partial labels and disambiguate labels could be useful in medical imaging where labels may be uncertain or incomplete.