Lab Report: Experiments on Partial Label Learning with Complementary Label Datasets

Anthony Ching - June 13, 2025

1. Introduction

This semester, my work focused on exploring the intersection between Partial Label Learning (PLL) and Complementary Label Learning (CLL). Specifically, I conducted a series of experiments aimed at benchmarking the performance of PLL algorithms on datasets originally designed for CLL, and vice versa. The goal was to investigate the feasibility and implications of applying one form of weak supervision to data annotated in the other format.

2. Background Study

To build a solid foundation for the experiments, I studied the following key resources:

- **Complementary-Label Learning: A Survey** (in progress by Ha) – provided a comprehensive overview of the current landscape of CLL, algorithms, challenges, and evaluation metrics.
- **CLImage: Human-Annotated Datasets for Complementary-Label Learning** (arXiv:2305.08295) – introduced realistic CL datasets with human-annotated labels, which I used extensively in my experiments.
- **The Unexplored Potential of Vision-Language Models for Generating Large-Scale Complementary-Label Learning Data** (PAKDD 2025) – explored automatic generation of CL data and inspired the idea of synthesizing PL datasets from CL datasets.

3. Experimental Setup

3.1 Datasets

- **Complementary Label (CL) Datasets**: I used four datasets from the CLImage repository:
  - CLCifar10
  - CLCifar20
  - CLMicroImageNet10 (later abbreviated as CLMIN10)
  - CLMicroImageNet20 (later abbreviated as CLMIN20)
- **Partial Label (PL) Datasets (Synthesized)**:
  - Starting from the CL datasets, I synthesized PL datasets with the following protocol:
    - Given C classes, each of the $C - 1$ incorrect labels has a uniform probability $q = 0.1$ of being falsely included in the candidate set.
    - These flipped labels were combined with the ground-truth label.
    - If no labels were selected (i.e., candidate set contained only the ground-truth), one incorrect label was randomly selected and added to ensure a minimum of two labels in the set.

3.2 Algorithms

- **CL Algorithms** (evaluated on CL datasets and adapted for PL datasets via label inversion):
  - **FWD** and **MCL** with ResNet18 backbone and EXP loss.
  - Codebase: https://github.com/ntucllab/libcll
- **PL Algorithms** (evaluated on PL datasets and adapted for CL datasets via label inversion):
  - **PRODEN** with ResNet18 (Lv et al., 2020)
  - Codebase: PRODEN
- **Label Inversion Protocol**:
  - To test CL algorithms on PL datasets, and vice versa, I inverted binary label vectors: flipping 1s to 0s and 0s to 1s to simulate complementary annotations.

4. Results Summary

Performance of CL and PL algorithms on CLImage and synthesized PL datasets.

Algorithm	CL Datasets				PL Datasets (Synthesized)			
	CLCifar10	CLCifar20	CLMIN10	CLMIN20	Cifar10 (PL)	Cifar20 (PL)	MIN10 (PL)	MIN20 (PL)
Proden	25.47%	7.60%	15.88%	5.79%	54.92%	32.24%	49.22%	28.14%
FWD	39.25%	19.82%	29.63%	10.58%	85.61%	63.68%	66.11%	63.69%
MCL	34.54%	07.82%	13.75%	05.43%	85.64%	63.06%	68.36%	60.65%

Notably, although Proden is designed for PL datasets, its performance is worse than CL algorithms. This suggests that there might be something wrong with my implementation, and further investigation and modification may be needed.

Proden struggled on CL datasets. This might be due to the large number of labels in the complementary label set ($C-1$ labels), making it difficult for the algorithm to disambiguate the true label. Additionally, potential noise inherent in the human annotation process of the original CLImage dataset could have contributed to this.

When inverting labels from CLImage datasets to create PL-like data for CL algorithms, only the first complementary label provided in the original dataset was considered. This resulted in a candidate set with $C-1$ positive labels after inversion, which is a challenging scenario for PL algorithms.

In the table, for FWD and MCL on CL datasets, the results were copied from the survey paper by Ha. The results I got for FWD and MCL on PL datasets were obtained from forking `libcll` and adding the necessary adaptations: https://github.com/AnthonyChing/libcll. For Proden results, they are from this repo: https://github.com/AnthonyChing/Complementary-and-Partial-Label-Learning.

I used FWD and MCL with ResNet18 and EXP loss according to Ha's suggestion. I picked Proden because it is one of the state-of-the-art algorithms. And I chose ResNet18 with it for easier comparison.

5. Challenges Encountered

- **Label transformation compatibility**: The label inversion protocol, while straightforward, proved insufficient for PL algorithms to effectively learn from the transformed CL data. Inverting labels (flipping 0s and 1s) worked in most cases but occasionally introduced inconsistencies, especially with models expecting specific loss functions or label distributions.
- **Implementation and adaptation effort**: Adapting algorithms designed for one type of weakly supervised data (e.g., PL) to work with another (e.g., CL), and vice versa, required considerable implementation effort. However, I tried to write out the code on my own to gain more understanding of the code structure and the underlying assumptions of each algorithm.

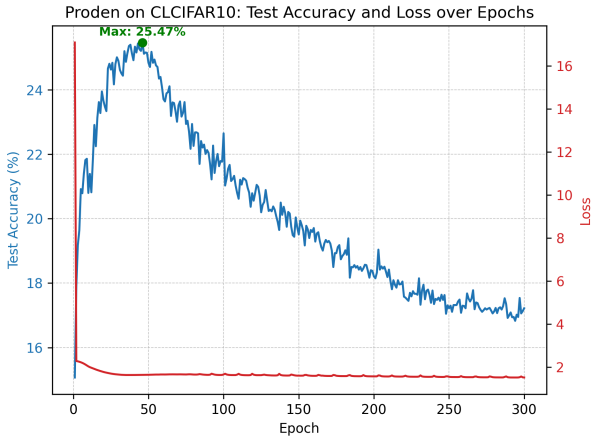
6. Future Work

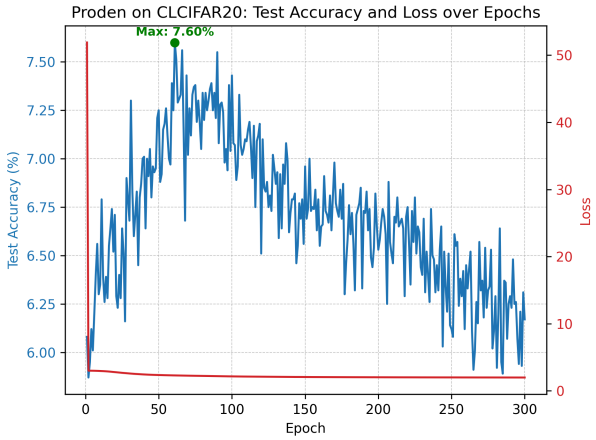
- **Verify Proden results**: Investigation and modification to the current implementation is needed to verify the results for Proden.
- **Varying synthesis probability q**: Plan to explore different values of q to study how label noise affects algorithm performance.
- **Vision-Language Integration**: Inspired by CLImage-VLM work, I may consider incorporating VLMs to auto-generate candidate label sets for PL datasets.
- **Algorithm Generalization**: Investigate if hybrid models or joint training strategies can leverage both PL and CL signals for improved generalization.
- **Hyperparameter tuning**: Several hyperparameters, such as the batch size, learning rate, choice of neural network architecture, and loss function, were chosen heuristically. These parameters can be systematically tuned to potentially improve performance.
- **Broader Benchmarking**: More partial label algorithms and more datasets can be tested in the future to provide a more comprehensive evaluation and understanding of the field.

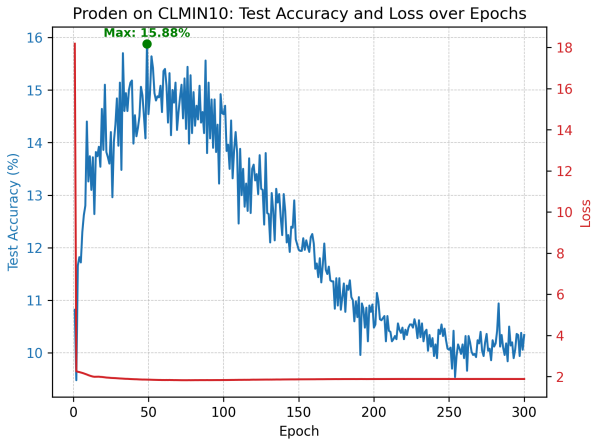
7. Conclusion

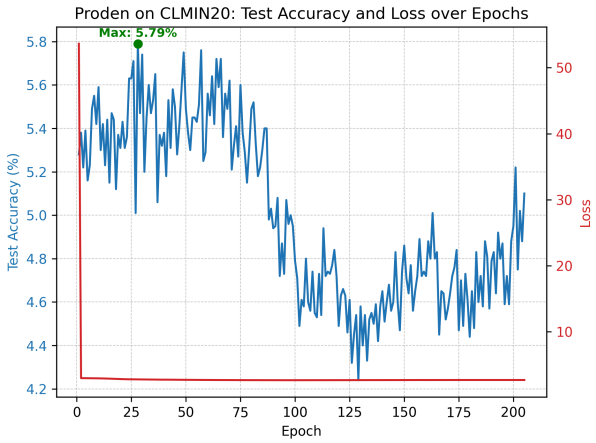
This semester's work provided practical insights into the relationship between PLL and CLL. By creatively transforming datasets and evaluating cross-paradigm algorithms, I was able to identify promising directions for future research in weakly supervised learning.

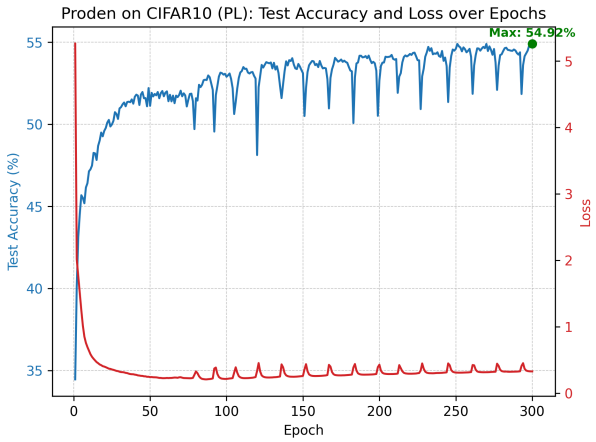
8. Appendix

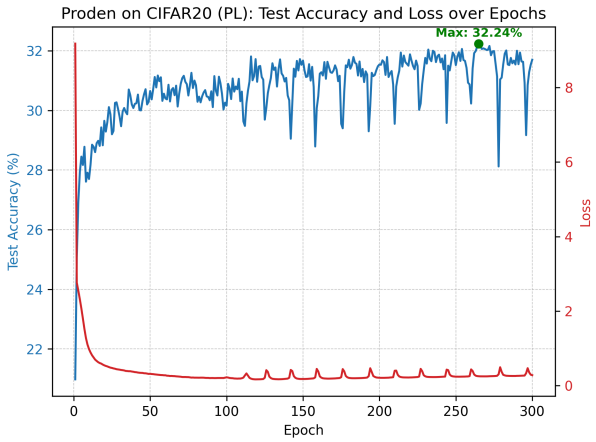

Proden on CLCIFAR10: Test Accuracy and Loss over Epochs
Max: 25.47%
Test Accuracy (%)
Loss
Epoch

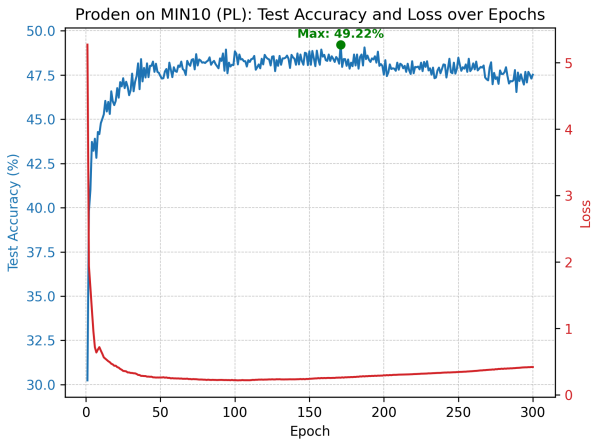

Proden on CLCIFAR20: Test Accuracy and Loss over Epochs
Max: 7.60%
Test Accuracy (%)
Loss
Epoch

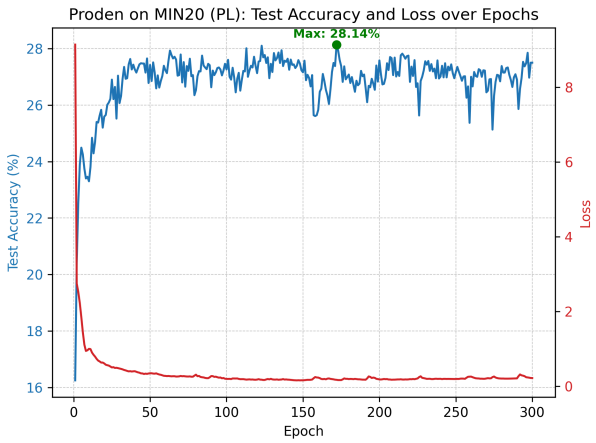

Proden on CLMIN10: Test Accuracy and Loss over Epochs
Max: 15.88%
Test Accuracy (%)
Loss
Epoch


Proden on CLMIN20: Test Accuracy and Loss over Epochs
Max: 5.79%
Test Accuracy (%)
Loss
Epoch


Proden on CIFAR10 (PL): Test Accuracy and Loss over Epochs
Max: 54.92%
Test Accuracy (%)
Loss
Epoch


Proden on CIFAR20 (PL): Test Accuracy and Loss over Epochs
Max: 32.24%
Test Accuracy (%)
Loss
Epoch


Proden on MIN10 (PL): Test Accuracy and Loss over Epochs
Max: 49.22%
Test Accuracy (%)
Loss
Epoch


Proden on MIN20 (PL): Test Accuracy and Loss over Epochs
Max: 28.14%
Test Accuracy (%)
Loss
Epoch