

Learning with Noisy Labels: From Centralized to Federated Learning Systems

by Zhuowei Wang

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of A/Prof. Guodong Long and A/Prof.
JingJiang

University of Technology Sydney
Faculty of Engineering and Information Technology

08.2022

Certificate of Original Authorship

I, Zhuowei Wang, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 12-09-2022

ABSTRACT

Learning with Noisy Labels: From Centralized to Federated Learning Systems

by

Zhuowei Wang

Machine learning has achieved prominent success in many fields, especially supervised learning tasks. However, such success depends on the correct annotation of all training samples for training robust models, where samples are difficult and expensive to obtain. The incorrect or incomplete information brought by wrongly annotated labels may cause catastrophic effects depending on the real-world applications. Therefore, this thesis studies two different kinds of weakly supervised learning, noisy label learning (NLL) and positive unlabeled learning (PUL), to improve the model robustness under incorrect labels in the dataset. Moreover, in real-world scenarios, most data is not collected and stored in a centralized way. Instead, data are distributed over various institutions protected by privacy restrictions. Federated learning (FL) has been proposed to leverage isolated data without violating privacy. However, data labels in different institutions are not annotated according to the same criterion so they inevitably contain different noises across silos. This doctoral thesis investigates how to combat noisy labels in both centralized and FL systems.

NLL aims to solve the problem where the input contents of training samples are intact but the labels are wrongly annotated from the ground-truth. Recent prominent methods combine specific sample selection and semi-supervised learning methods to use all given samples fully, achieving SOTA performance. Motivated by this intuition, one might easily derive various effective NLL methods using different combinations of sample selection strategies and semi-supervised learning models, which is, however, simply reinventing the wheel. We propose a versatile framework that investigates how to combine different components based on their effects and efficiencies. We conduct a systematic and detailed

analysis of the combinations of possible components based on our framework. Experiments demonstrate the versatility of our framework and the superior performances of our instantiations based on the framework.

PUL aims to train the classifier with only positive and unlabeled data. It is considered a special case of NLL by treating the entire set of unlabeled data as noisy negative samples. Previous SOTA methods suffer from model overfitting by treating all unlabeled samples as weighted noisy negative samples. We empirically demonstrate that the previous SOTA PUL method could misclassify negative samples in unlabeled data as positive ones at the late training stage. Thus, we propose a novel semi-supervised learning based approach to tackle PUL. We leverage dynamic increasing sampling to select confident samples and fit the remaining unlabeled samples into a semi-supervised learning framework. Empirical results demonstrate that our method can alleviate the model overfitting issue and achieves SOTA performance.

Federated learning (FL) has been proposed to leverage isolated distributed data without violating privacy. The labels of these data are not annotated by the same annotators according to the same criterion so they inevitably contain different noises across clients. We non-trivially extend the noisy label scenario to FL system. We develop a simple two-level sampling method that selects clients for more robust global aggregation on the server and selects clean labels/pseudo-labels at the client end for more robust local training. Experimental results show that direct combinations of SOTA FL methods with SOTA NLL methods can easily fail but our method consistently achieves better and more robust performance.

Dissertation directed by A/Prof. Guodong Long and Dr. Jing Jiang

Australian Artificial Intelligence Institute

Faculty of Engineering and IT

University of Technology Sydney

Acknowledgements

The completion of this Ph.D. would never have been possible without the tremendous inspiration and encouragement from many people during these unforgettable four years, to whom I am greatly indebted.

First, I would like to thank my supervisors, A/Prof. Guodong Long and Dr. Jing Jiang. They provided me with a precious opportunity to pursue my Ph.D. in UTS, which became the changing point in my life. A/Prof. Guodong Long has always been my research mentor, giving me the guidance I need. Whenever I am stuck in some research dilemma, he is the first person from that I would seek help. He is always generous, kind, and patient. He not only teaches me how to do research in academia but also what I should keep in mind if I work in industry. I feel so lucky to be his student. Dr. Jing Jiang has always looked out for me in research and life. When I struggled with reinforcement learning and TensorFlow in the first year and could not see any progress in my research, she recommended another research direction to me and asked someone to tutor me from scratch. I will never forget how she comforted and encouraged me not to lose hope on the day I walked her to UTS kindergarten to pick up her son. She literally pulled me out of despair during my darkest days. And I would never complete this degree without her support. Moreover, I would also like to express my gratitude to Prof. Chengqi Zhang for providing me with a resourceful platform and teaching me how to do research and live a meaningful life from a bigger picture. Thanks for his inspiring words.

I am also blessed to work closely with top AI researchers worldwide: Dr. Bo Han, Dr. Tianyi Zhou, and Prof. Lei Feng. They provided me with guidance and detailed suggestions for my research. I believe I will continue to benefit from these experiences for the rest of my life.

I would like to take this opportunity to thank my fellow teammates at UTS: Yijun

Yang, Haiyan Zhao, Shuang Ao, Kaize Shi, Jie Ma, Wensi Tang, Hao Huang, Yue Tan, Chang Shao, Peng Yan, Tianchi Sha, Shutong Chen, Jianan Yang, Yiyuan Yang, Zhihong Deng, Yang Li, Alvin Wang, and Dr. Xueping Peng. It is so lucky and a blessing to work in such a friendly and supportive team.

Moreover, I want to thank my friends who were with me both in person and virtually throughout this journey: Yan Guo, Fei Zhang, Tao Zhang, Chu Peng, Jinru Xue, Miaoyuan Dou, Junwei Lyu, Bingcong Li, Mingshi Wan, Chao Mai, Wei Huang, Jingsheng Gao, Zeyu Li, Yikun Yang, Xiuyuan Hu, Wuning Xie, Yunqiu Xu, Yujiao Shi, Ivy Nguyen, Minduli Withana, Paranshi Soni, Orlagh Biling, Angsar Manatuly, and so on. My life was made better by their friendship.

Last, I would like to thank my parents for their unconditional support and love.

Zhuowei Wang
Sydney, Australia, 2022.

List of Publications

Conference Papers

- C-1. **Zhuowei Wang**, Tianyi Zhou, Guodong Long, Bo Han, Jing Jiang. “FedNoiL: A Simple Two-Level Sampling Method for Federated Learning with Noisy Labels,” submitted to The 32nd International Joint Conference on Artificial Intelligence, 2023. (Under review)
- C-2. **Zhuowei Wang**, Jing Jiang, and Guodong Long. “Positive Unlabeled Learning by Semi-Supervised Learning,” IEEE International Conference in Image Processing, pp. 2976-2980, IEEE, 2022. doi: 10.1109/ICIP46576.2022.9897738.
- C-3. **Zhuowei Wang** and Guodong Long. “Positive Unlabeled Learning by Sample Selection and Prototypical Refinement,” International Conference Advanced Data Mining and Applications, Proceedings, Part I, pp. 304-318, 2022. doi: 10.1007/978-3-031-22064-7_23.
- C-4. Bingcong Li, Bo Han, **Zhuowei Wang**, Jing Jiang, and Guodong Long. “Confusable Learning for Large-class Few-Shot Classification”, Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2020. Lecture Notes in Computer Science, vol 12458. Springer. doi: 10.1007/978-3-030-67661-2_42.
- C-5. Yijun Yang, Jing Jiang, **Zhuowei Wang**, Qiqi Duan, and Yuhui Shi. “BiES: Adaptive Policy Optimization for Model-based Offline Reinforcement Learning,” AI 2021: Advances in Artificial Intelligence. AI 2022. Lecture Notes in Computer Science, vol 13151. Springer. doi: 10.1007/978-3-030-97546-3_46.

Journal Papers

- J-1. **Zhuowei Wang**, Jing Jiang, Bo Han, Lei Feng, An Bo, Gang Niu, and Guodong Long. “SemiNLL: A Framework of Noisy-Label Learning by Semi-Supervised Learning,” Transactions on Machine Learning Research Systems, 2022.

<https://openreview.net/pdf?id=qzM1Tw5i7N>

Contents

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Publications	vii
List of Figures	xiii
List of Tables	xvi
1 Introduction	1
1.1 Background	1
1.2 Research Problems	7
1.3 Major Contributions	11
1.4 Thesis Organization	12
2 Literature Review	14
2.1 Noisy Label Learning	14
2.2 Positive Unlabeled Learning	17
2.3 Federated Learning	20
3 SemiNLL: a Framework of Noisy-Label Learning by Semi-Supervised Learning	25
3.1 Introduction	25
3.2 The Overview of SemiNLL	28

3.2.1	Mini-batch Sample Selection	28
3.2.2	SSL Backbones	30
3.3	The Instantiations of SemiNLL	31
3.3.1	Instantiation 1: DivideMix+	32
3.3.2	Instantiation 2: GPL	33
3.3.3	Self-Prediction Divider	33
3.3.4	Effects of the Two Components	34
3.4	Experiment	35
3.4.1	Experiment Setup	35
3.4.2	Performance Comparison	38
3.5	Ablation Studies	43
3.5.1	The Effects of Mini-batch Mechanism	43
3.5.2	The Effects of Different SS strategies	44
3.5.3	The Effects of Different SSL backbones	45
3.5.4	The Effects of SSL on SS	47
3.5.5	The Effects of Training Dual Networks	49
3.5.6	Sensitivity to the Batch Size of SS	49
3.5.7	Different Model Architectures	50
3.5.8	Efficiency Analysis	51
3.5.9	Broader Impact	51
3.6	Recommendations on How to Instantiate Our Framework	52
3.6.1	Choices of SS Scope	52
3.6.2	Choices of Number of Networks	53
3.6.3	Choices of Different SS and SSL Combinations	53

3.7	Limitations	55
3.8	Other Methods	56
3.9	Chapter Summary	57
4	Positive Unlabeled Learning by Semi-Supervised Learning	59
4.1	Introduction	59
4.2	Problem Setting	62
4.2.1	Review of IR Approaches	62
4.2.2	Overfitting of IR	64
4.3	The Proposed Approach	65
4.3.1	Dynamic Increasing Sampling	65
4.3.2	Utilising Unchosen Samples via SSL	68
4.3.3	Co-learning to Reduce Confirmation Bias	70
4.4	Experiments	71
4.4.1	Datasets	71
4.4.2	Experimental Setup	73
4.4.3	Experimental Results	75
4.5	Chapter Summary	77
5	FedNoiL: A Simple Two-Level Sampling Method for Federated Learning with Noisy Labels	78
5.1	Introduction	78
5.2	Problem Formulation	81
5.3	Proposed Method	82
5.3.1	Warm Starting	82
5.3.2	Client Sampling	83

	xii
5.3.3	Local-data Sampling 84
5.3.4	Schedules of Local Epochs 86
5.4	Experiments 88
5.4.1	Experimental Setup 89
5.4.2	Choices of Local Epoch Schedules 95
5.4.3	Comparison with Baselines 98
5.4.4	Ablation Studies 99
5.4.5	Extended Comparison in More Challenging Settings 105
5.5	Chapter Summary 107
6	Conclusions and Future Work 108
6.1	Conclusions 108
6.2	Future Work 109
	Bibliography 111

List of Figures

1.1	A normal process [1] to generate large-scale labeled image datasets is to download the images returned by querying a keyword in the Google search engine. However, this kind of process usually generates noise labels for images.	2
1.2	Two lines of images represent two datasets and all images are labeled as horse. The first line is a clean dataset since all the samples have the correct labels. While the second line is defined as the “noisy” dataset since the deer (in the purple box) is wrongly labeled as horse.	4
1.3	The illustration of a typical FL architecture. The local model is updated to the server for global aggregation, and each client downloads the updated model. Since the labels of local devices are not annotated by the same client according to the same criterion, they inevitably contain different noises across clients.	6
1.4	Thesis structure and three key tasks.	8
3.1	The schematic of <i>SemiNLL</i> . First, each mini-batch of data is forwarded to the network to conduct SS, which divides the original data into the labeled/unlabeled sets. Second, labeled/unlabeled samples are used to train the SSL backbone to produce accurate model output.	27
3.2	Comparisons between: (a) <i>DivideMix-</i> , (b) <i>DivideMix</i> , and (c) <i>DivideMix+</i> . Squares represent data. Circles represent SS strategy. Octagons represent SSL backbone.	31

3.3	Results of ablation study on CIFAR-10 sym 80%. Left: accuracy vs. epochs; right: precision vs. epochs.	44
3.4	Results of ablation study on CIFAR-100 asym 40%. Left: accuracy vs. epochs; right: precision vs. epochs.	45
3.5	Representations (t-SNE 2D embeddings) of two CIFAR-10 classes, ‘cat’ and ‘truck’, learned by DivideMix+ (left) and GPL (right), with 80% noise rate. Blue/red dots represent clean samples of cat/truck, while magenta and cyan crosses represent corrupted samples.	48
4.1	The nnPU algorithm’s positive probability histogram of all unlabeled data on CIFAR-10. Given the model g and each input sample x , we calculate their probability of being positive. We divide the range $[0, 1]$ equally into 100 bins and count the number of samples in each bin. (a) Positive probability histogram on epoch 10; (b) positive probability histogram on epoch 50; (c) positive probability histogram on epoch 200.	62
4.2	The schematic of our method. First, all unlabeled samples are forwarded to the current network to select most confident samples, which divides the original unlabeled set into the positive/negative/remaining sets. Second, the remaining unlabeled samples are leveraged by the semi-supervised learning method.	66
4.3	Sensitivity analysis of the ratio of selected positive samples μ and the the ratio of selected positive samples δ . A darker color indicates a higher test accuracy.	76
5.1	The communication (left) and computation (right) comparison between different decay rates of Cosine schedule under high noise in CIFAR-10.	86
5.2	The communication (left) and computation (right) comparison between different decay rates of Logarithm schedule under high noise in CIFAR-10.	87

5.3 Examples for decaying schedules of local epochs in the form of Cosine(r_{\min}) and Logarithm(r_{\min}), which reaches the minimum local epochs T_{\min} at round- $r_{\min} \in \{10, 20, 40, 80\}$ 88

5.4 **Communication (left) and computation (right) efficiency** of FedNoiL(**Cosine**) with different r_{\min} as in Fig. 5.3 and constant local epochs (matching the total training batches of Cosine(r_{\min})), on CIFAR-10 with high-noise ratio (symmetric) setting. Colors represent different r_{\min} . Cosine and constant schedules are denoted by solid lines and dash lines, respectively. 95

5.5 **Communication (left) and computation (right) efficiency** of FedNoiL(**Logarithm**) with different r_{\min} as in Fig. 5.3 and constant local epochs on CIFAR-10 with high-noise ratio (symmetric) setting. 96

5.6 Comparison of FedNoiL using different local-epoch schedules on communication, computation, test accuracy, and their trade-off. 97

5.7 Comparison of FedNoiL using different local-epoch schedules on communication, computation, test accuracy, and their trade-off. 98

5.8 **Ablation Study** of two-step sampling on CIFAR-10: comparing the four metrics (in Section 5.4.1) of FedNoiL(Uniform Client Sampling), FedNoiL(Uniform Local-Data Sampling), and FedNoiL(Original) of high-noise (pair flipping) ratio setting. 100

5.9 Performance analysis when using different values of temperature. 104

5.10 Examples for ascending(slower and faster) and decaying(cosine and logarithm) schedules of local epochs, which reaches the maximum (or minimum) local epochs at round 80. 106

5.11 **Communication (Rounds) and computation (right) efficiency** of FedNoiL using two ascending schedules of local epochs with $r_{\max} = 80$ shown in Fig. 5.10, compared to Cosine(80) and Logarithm(80), on CIFAR-10 with high-noise ratio (symmetric) setting. Colors represent different schedules. 106

List of Tables

3.1	Summary of datasets used in the experiments.	36
3.2	The “13-CNN” network architecture used in CIFAR-10 and CIFAR-100. .	37
3.3	Average test accuracy (%) and standard deviation (5 runs) in various datasets under symmetric label noise. The best accuracy is bold-faced . The second-best accuracy is <u>underlined</u>	38
3.4	Average test accuracy (%) and standard deviation (5 runs) in various datasets under asymmetric label noise. The best accuracy is bold-faced . The second-best accuracy is <u>underlined</u>	40
3.5	Test accuracy (%) on Clothing1M.	41
3.6	Test accuracy (%) on (mini) WebVision.	42
3.7	Test accuracy (%) on Food-101N.	43
3.8	Test accuracy (%) on ANIMAL-10N.	43
3.9	The test accuracy (%) of GPL (epoch) and GPL (mini-batch) CIFAR-10. .	44
3.10	Test accuracy (mean \pm std. dev.) of <i>DivideMix-</i> , <i>DivideMix</i> , and <i>DivideMix+</i> under symmetric noise.	46
3.11	Test accuracy (mean \pm std. dev.) of <i>DivideMix-</i> , <i>DivideMix</i> , and <i>DivideMix+</i> under asymmetric noise.	46
3.12	Test accuracy (mean \pm std. dev.) of the baseline and three SSL backbones in CIFAR-10.	47
3.13	Test accuracy (mean \pm std. dev.) of the baseline and three SSL backbones in CIFAR-100.	47

3.14	Ablation Study of training with two networks on different instantiations on CIFAR-10.	49
3.15	The test accuracy (%) of DivideMix+ and GPL with different batch sizes on CIFAR-10 Sym 80%.	50
3.16	Comparison between DivideMix, DivideMix+, and GPL using different model architectures in test accuracy (%) on CIFAR-10. Key: WRN (Wide ResNet), PRN (PreActivation ResNet).	50
3.17	Comparison of training time on CIFAR-10.	51
3.18	The test accuracy (%) of instantiations of different SS and SSL methods under high/low symmetric noise in CIFAR-10.	54
4.1	The characteristics of the datasets. The upper three datasets are computer vision datasets, and the last one is a natural language processing dataset. P class means the definitions of positive labels. N class means the definitions of negative labels.	71
4.2	Mean and stand deviation of the test accuracy(%) over five repeated runs for different datasets. The best accuracy is bold-faced	74
4.3	Ablation Study of different components of Co-Learning on all datasets.	75
5.1	The noise ratio in different groups of two noise modes under symmetric and pair flipping noise.	90
5.2	Test accuracy (%) of FedNoiL and baselines under IID setting on three datasets. The final accuracy of converged algorithms are reported by the mean and standard deviations over five runs. Methods that do not converge are marked by * and reported by their maximum accuracy over all rounds.	92
5.3	Test accuracy (%) of FedNoiL and other baselines under Non-IID setting on three datasets.	93

5.4	Ablation Study of two-step sampling on three datasets in symmetric high noisy mode: comparisons of test accuracy, average noise ratio, label precision, and label recall (%) of FedNoiL(Uniform Client Sampling), FedNoiL(Uniform Local-Data Sampling), and FedNoiL(Original) in high-noise ratio settings (symmetric and pair flipping noise).	101
5.5	Ablation Study of two-step sampling on three datasets in pair high noisy mode: comparisons of test accuracy, average noise ratio, label precision, and label recall (%) of FedNoiL(Uniform Client Sampling), FedNoiL(Uniform Local-Data Sampling), and FedNoiL(Original) in high-noise ratio settings (symmetric and pair flipping noise).	102
5.6	Comparisons between FedNoiL [†] (without SSL) and FedNoiL on test accuracy (%), label precision (%), and label recall (%) in CIFAR-10.	103
5.7	The final test accuracy (%) of FedNoiL(Logarithm) with different ratio α of client sampling at each round in high/low symmetric noise in FASHION-MNIST and CIFAR-10.	105
5.8	Test accuracy (%) of FedNoiL and baselines in settings with more clients and imbalanced samples in CIFAR-10.	107