UTS UNIVERSITY OF TECHNOLOGY SYDNEY

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 groud-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

List of Figures

List of Tables