UTS UNIVERSITY
OF TECHNOLOGY
SYDNEY

Learning with Restricted Data via Gradient Manipulations

by Jing Li

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Ivor W. Tsang

University of Technology Sydney
Faculty of Engineering and Information Technology

October 2022

Certificate of Original Authorship

I, Jing Li, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the 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: 20 Oct 2022

This thesis is lovingly dedicated to my family. I wish I would be always your pride and joy.

Acknowledgements

Foremost, I would like to thank my supervisor Prof. Ivor W. Tsang, for his professional guidance and persistent support. He opened up a window for me to have a closer look at the world of *machine learning* which appears complex but fascinating! Without his leading in the passed four-year academic journey, I could not have overcome the obstacles nor found my way to think and do. I am also grateful that Prof. Ivor W. Tsang could always provide appropriate suggestions on my study plan, from each individual project to internship scheduling and thesis preparation. He is such a caring mentor who helps me become confident, mature and independent. What I have learned from him will have a great impact on my future career path.

Thanks also go to my co-supervisor Prof. Yulei Sui who provided a perspective for my work from a security player. His participation makes my work more practical and understandable for people out of the machine learning community. In particular, I appreciate Prof. Yulei Sui's contribution to polishing my papers.

I would like to sincerely thank Dr. Yuangang Pan for his unselfish assistance on my study. His research passion and attitude has encouraged me to keep exploring the interesting topics. I would also thank my best friend Yinghua Yao, who is not only a good flatmate but also reliable co-worker. From Sydney to Singapore, I am fortunate that people around me are helpful. I appreciate Dr. Yan Zhang's invitation last winter and the contribution of Dr. Yueming Lyu in my recent work. I am also grateful for the assistance of Dr. Yaxin Shi, Xiaowei Zhou, Xingrui Yu on my research. My warmest gratitude also goes the Dr Jiangchao Yao, Dr. Xu Chen, Dr. Xiaofeng Xu, Zhuanghua Liu, Jinliang Deng, Feiyang Ye, Bowen Xing, Peiyao Zhao, Yujie Fang, and Cheng Chen. Thanks also to the co-workers out of my group, Shaojun Shi and Tao Zhang.

Many thanks to the University of Technology Sydney for providing me a good studying environment, and Agency of Science, Technology and Research for offering me a rare internship opportunity.

My special thanks to Yume who has accompanied me for the entire PhD journey. Last, Dad and Mom, the sacrifices you have made for me are beyond any description. Thank you, my beloveds. I wish I would always make you proud.

Abstract

Data has been the fuel that drives modern artificial intelligence. With more and more emerging concerns on data, e.g., data privacy, learning paradigms demand evolving accordingly. In this thesis, I focus on discriminative learning on **restricted data** from the role of learning executors who are in charge of the learning process. That means, my research interest lies in how to design proper learning algorithms while data is considered restricted. Specifically, the following three types of scenarios will be explored.

- *Private data is accessible to learning executors, but the learning process should not expose any data information.* In such a scenario, the learning executors are trusted while any others are restricted from accessing data. Differential Privacy (DP) is a golden principle for this problem which preserves the participation of every data point and thus can defend against strong adversaries. I intend to take a step forward and explore how to ensure a safe learning when data is pairwise labelled, to which DP cannot be directly applied due to the explicit pairwise correlations.

- *Only incomplete data is accessible to learning executors.* Learning on incomplete data is a challenging topic when some data information is restricted from learning executors. For example, not all people would like to answer their demographics in a survey. Suppose a common case where the missing values are from a discrete attribute or the label domain. Inferring them comes to the Semi-Supervised Learning (SSL) problem. I will study how to improve the prediction module for unlabeled data from the aspect of prediction uncertainty.

- *None of data is accessible to learning executors, but some feedbacks are available.* This scenario considers that learning executors cannot access data, which hinders feeding data to the model for the end-to-end back-propagation. However, learning is stilled feasible if some feedbacks from data are provided, e.g., model evaluations. In practice, model tuning tasks are studied in this context, and the tuning efficiency for deep neural networks is particularly investigated.

From top to bottom, one can sense that the restriction on data becomes stricter, which also implies some bigger change should be applied to learning paradigms. Despite the specific requirement in each scenario, I am interested how to deal with them via a general principle. It is known that gradient-based optimization has been popular in machine learning. Given a fixed model structure, the eventually attained model can be attributed to the elaborately designed model gradients (Note that it is not always because of losses). With this insight, I propose to deal with different restricted data by using different meaningful **gradient manipulation** techniques. Concretely, I apply gradient perturbation to compensate for the missing or addition of any interested pairwise data, employ gradient masking to reduce the impact of over-confident unlabeled predictions, adopt gradient estimation to learn from model evaluations. I conclude that although the meaning of restricted data varies across different tasks (which also brings out various challenges), the insight of gradient manipulation constantly offers a good perspective to tackle these problems.

Table of contents

List of figures ix

List of tables xiv

1 Introduction 1

1.1 Background . . . 1

1.2 Research Scope . . . 2

1.3 Challenges . . . 4

1.4 Thesis Contributions . . . 5

1.4.1 Distance metric learning with private pairwise data . . . 5

1.4.2 Semi-supervised learning for inferring missing labels . . . 5

1.4.3 Model tuning without peeking on target data . . . 6

1.5 Thesis Outline . . . 7

1.6 Publications . . . 7

2 Problem Statement 9

2.1 Restricted Data . . . 9

2.2 Learning with Restricted Data . . . 11

2.2.1 Learning with private pairwise data . . . 11

2.2.2 Learning with incomplete data . . . 14

2.2.3 Learning with inaccessible data . . . 16

2.3 Gradient Manipulation . . . 18

2.3.1 Gradient perturbation . . . 19

2.3.2 Gradient masking . . . 19

2.3.3 Gradient estimation . . . 20

3 Distance Metric Learning with Private Pairwise Data 21

3.1 Problem Understanding . . . 21

3.1.1 Pairwise data leakage in distance metric learning . . . 21

3.1.2 Differential privacy and its limitation for pairwise data 23
3.2 Preliminaries 25
3.3 Pairwise Relation in Distance Metric Learning 26
3.3.1 Privacy investigation 26
3.3.2 Clarification 28
3.4 Differential Pairwise Privacy from Graph Perspective 29
3.4.1 Privacy concern on edge 29
3.4.2 Differential Pairwise Privacy (DPP) 31
3.5 Private Distance Metric Learning 32
3.5.1 Differential pairwise privacy with contrastive loss 32
3.5.2 Improvement by sensitivity reduction 33
3.6 Experiment 36
3.6.1 Toy example 37
3.6.2 Comparison on real-world datasets 38
3.6.3 Privacy mechanisms comparison 40
3.6.4 Effects of parameters 41
3.7 Summary 42

4 Semi-Supervised Learning for Inferring Missing Labels 43

4.1 Problem Understanding 43
4.1.1 Semi-supervised learning paradigm 43
4.1.2 Taming overconfident predictions 44
4.2 Preliminaries 45
4.2.1 Formulation 45
4.2.2 An entropic view of distillation 47
4.3 ADS Based SSL Model 48
4.3.1 ADaptive Sharpening (ADS) 48
4.3.2 In conjunction with other loss 50
4.4 Theoretical Analyses 51
4.4.1 ADS promotes informed predictions 52
4.4.2 ADS facilitates entropy minimization 54
4.4.3 ADS introduces a lightweight computation 55
4.5 Experiment 55
4.5.1 Experimental setup 55
4.5.2 Study on VAT 56
4.5.3 Improvement on advanced SSL algorithms 57
4.5.4 Safety with different backbone structures 59

4.5.5 Observation of prediction histograms 60
4.5.6 Running time comparison 61
4.5.7 Scalability to ImageNet 62
4.5.8 Ablation study 62
4.6 Summary 64

5 Model Tuning without Peeking on Target Data 65
5.1 Problem Understanding 65
5.1.1 Model tuning without back-propagation 65
5.1.2 EXPECTED setting 66
5.1.3 Comparison with other model tuning settings 67
5.2 Preliminaries 68
5.3 Tuning from Restrictive Feedbacks 70
5.3.1 Gradient-based optimization from query-feedbacks 70
5.3.2 Extension to complex models 75
5.4 Experiment 79
5.4.1 Experimental setup 79
5.4.2 EXPECTED on shifted data distribution 81
5.4.3 EXPECTED for customized evaluation metrics 83
5.4.4 A close investigation to LCPS 85
5.4.5 Important factors study 86
5.5 Discussion 88
5.6 Summary 89

6 Conclusion and Future Work 90
6.1 Conclusion 90
6.2 Future Work 91

Appendix A Appendix 92
A.1 Sensitivity Upper Bound for Efficiency 92
A.2 DPP for Intransitive Relationship Case 93
A.3 Sensitivity Reduction for Approximate DPP 93

Appendix B Appendix 95
B.1 Distillation Comparison 95
B.2 Proof for Corollary 1 96
B.3 Example of Calibration Evaluation 97

Appendix C Appendix 99

C.1 Proof of Theorem 5 99

C.2 Algorithm for Fairness Learning 101

C.3 Discussion of Private Tuning Application 102

References 105

List of figures

1.1 Various privacy concerns arise during learning in the real world. 2

1.2 Privacy concerns in terms of learning executors motivate the different forms of learning on restricted data. The unit decorated with thicker edges highlights the novel points of this thesis. Note that I specially consider the cases where a single discrete attribute or the label domain is incomplete in the second concern, which can be remedied by SSL then. 3

1.3 The organization of this thesis. 8

2.1 Private pairwise data is sent to learning executor with the privacy requirement during the learning process. 11

2.2 Incomplete data is sent to learning executor with sensitive attributes or labels having been hidden by some participants. 14

2.3 Learning executor cannot access data but receives some feedbacks instead. 17

3.1 Leakage of pairwise relationship. An attacker with all the prior knowledge of the dataset except the target relationship between Bob and Lam, is able to infer their real relationship by matching the conjecture and query results. . 22

3.2 Knowledge diagram. The prior knowledge is supposed to be smaller than the whole data deducting the target pair because of the data correlation. For a given target pair, there always exists a corresponding defensive boundary which restricts the volume of prior knowledge in practice. 23

3.3 Preserving privacy of pairwise relationship. Suppose the relationship between Alice and Bob is the target. The attacker may have the prior knowledge that excludes edges with question mark. This provides one of the worst cases, where the relationship of Alice and Bob cannot be inferred from prior knowledge. DPP ensures that the prior knowledge of the attacker for the worst case has the hardly indistinguishable output with the original dataset. Particularly, the obtained metric M_0 is expected to group training data as M_1 does. . . . 24

3.4 Comparison between the samples fed to classification or regression models and pairwise data fed to DML algorithms. **Left:** Any two samples composed of the feature x_i (x_j) and its label l_i (l_j) are independent in existing ERM-based works. **Right:** Pairwise data are correlated with each other because an individual may participate in multiple pairs. 28

3.5 Construction of neighboring graph w.r.t. the pair $\langle s,t \rangle$. (I) The graph encoding all the pairwise data. (II) Disjoint-edge identification. (III) Two key edges (s,c) and (s,t) determining the relationship inference. (IV) Edge (b,s) exposing the feature of the individual s. 30

3.6 Gradient sensitivity reduction w.r.t. minibatch data. **Left:** Individuals are ℓ_1 normalized. The cyan line segment denotes the factor $2h$ specified by Theorem 1. p_{max} is one of all batch members whose gradient value is the largest, q denotes the possible counterpart of p_{max} in the neighboring batch that satisfies Eq. (3.12). The orange line segment connecting p_{max} and q is likely shorter than the cyan one. **Right:** Individuals are ℓ_2 normalized, and representation are consistent with the left. This subfigure is also specified by Corollary 2 in Section A.3. 34

3.7 DML projects original data into a new space. (a) A synthetic dataset containing 200 data points drawn from two aligned strips. (b)-(d) Data distribution after applying the metric learned by contrastive loss with DPP, DPP-S (with sensitivity reduction), and NonPriv concern, respectively. 37

3.8 (a) The objective values of Eq. (3.7) versus iteration number with NonPriv, DPP and DPP-S, respectively. (b) The sensitivity value $\frac{2\kappa h}{|\mathscr{B}|}$ specified by Theorem 1 and reduced sensitivity specified by Theorem 2 (exhibited by each dimension) versus iteration number. 38

3.9 Classification accuracy of compared methods versus privacy budget ε over four real-world datasets. 39

3.10 Different ε-DP mechanisms comparison in implementing DPP through their objective values. 40

3.11 Effects of several key parameters on Bank dataset. (LipCons stands for Lipschitz constant and DimRed stands for Dimension reduction.) 41

4.1 **Left:** Comparison among various distillation strategies, each of which is viewed as a two-stage process by first selecting candidate classes and then aligning their predictions with the categorical target label distribution. (I) enhance determinate predictions; (II) promote informed predictions; (III) suppress negligible predictions. **Right:** Different strategies turn out having different fashions to minimize prediction uncertainty. Binary classification is showcased here for simplicity where $p = (s, 1-s)$ where $0 \leq s \leq 1$. 47

4.2 Distillation architecture of ADS. 48

4.3 Comparison of different distillation strategies in terms of target probability, distillation loss, and gradient. Note that the distillation gradient of SH and ADS shown in the subfigure (c) is corresponding to the reduced losses (See Appendix B.1). 52

4.4 Converged curves of distillation loss $\mathscr{J}_{\text{D}}$ and average dominant probability $\bar{p}_{(1)}$ for unlabeled training samples on MNIST datset. The loss values are smoothed for a better visualization. For ADS, $\bar{p}_{(1)}$ is collected and calculated by replacing sparsemax with softmax which does not change the training process. 56

4.5 The safety study of candidates selection in terms of two backbones ResNet and CNN13. (a) The average sparse activations $\bar{m}$ on unlabeled training data. (b) Top-m accuracy comparison where m is example-wise sparsity. (c) Standard accuracy comparison with globally fixed Top-m selection. 59

4.6 Numerical distribution of prediction values of VAT+ADS on unlabeled training data. Top row shows the result on MNIST and the bottom row shows the result on CIFAR-10. The initialization of the network is used as their default. 60

4.7 Runing time (seconds) comparisons over epochs. **Left:** VAT (i.e., “W/O”) with different distillation strategies on CIFAR-10 (4,000 labels). **Right:** MixMatch and FixMatch based methods on CIFAR-100 (10,000 labels). . 62

4.8 Impact of different values for the power r. The best r is in the range of $[1.5, 2.5]$ for VAT+ADS, and $[1, 2]$ for FixMatch-ADS. 63

5.1 Overview of EXPECTED. (a) Given a deployed model parameterized by θ_0, EXPECTED aims to adapt it to the target task with limited query-feedbacks (budget Q) through the unobserved evaluation. (b) The unobserved evaluation is instanced by the inaccessibility of target data. In this case, EXPECTED is compared with other three model tuning settings from the aspects of (1) how much information about target data $\mathscr{D}$ is accessible and (2) how the gradient information ∇_θ is attained. The grey filling indicates the object is unobserved to the learning executor. In term of the federated learning, although local data $\mathscr{X}, \mathscr{Y}$ is inaccessible to the global model, the true gradient ∇_θ is actually returned. Note that E_i is informally short for $E(\mathscr{D};(\theta_0 + \delta_i))$. 67

5.2 An example of how the estimated gradient $\nabla\mathbb{E}[E(\theta)]$ approximates the true gradient $\nabla E(\theta)$. The pink arrow denotes the projection of $\nabla E(\theta)$ onto selected finite bases ε_1 and ε_2. One can easily verify that a true gradient $(2,1,1)$ under this decomposition corresponds to an estimated gradient $(2,1.1,0)$. . 74

5.3 Example of EXPECTED optimized by PPS. (a) Pre-training on source data delivers the initially provided model. (b) The given model successfully adapts to target data through PPS within 80 queries. 75

5.4 Performance comparison on Adult and Amazon. Throughout all the experiments, the accuracy on the support set is monotonically non-decreasing, since I display the historically best at every iteration. Note that "good VOC" and "bad VOC" correspond to the different selections of vocabulary. The line shadow represents the standard deviation. 81

5.5 Average error (%) over 15 types of corruptions for the highest severity, where RS, PPS, LCPS and Tent are based test-time BN. Red marks denote the failure cases of Tent. 83

5.6 Generalization improvement of BERT and its variants after the model tuning on STS-B, which is computed by $\frac{s-s_0}{s_0}$, where s_0 and s represent the model performance before and after tuning, respectively. 83

5.7 Discrimination level reduction for model fairness tuning, where the particles falling in "Improved Zone" represent the models that have been improved in terms of both accuracy and fairness metrics on the holdout set. 84

5.8 Evaluation performance (%) of LCPS with top-1 or top-5 error as a tuning metric on two types of corruptions (Gaussian and Impulse noises) over CIFAR-10-C. "Non" represents an initially provided model with the test-time BN is directly evaluated without any tuning efforts. The lowest errors are marked as bold. 84

5.9 Query budget reassignment of LCPS on CIFAR-10-C and STS-B. (a) and (b) are corresponding the results of CIFAR-10-C with Gaussian corruptions and STS-B with BERT being backbones. The grey dashed line indicates the expected query assignment for each layer without the layer importance concern. (c) exhibits the entropy of sampling probability over each iteration for the two experiments. 86

5.10 Ablation study on three factors: sampling batch size, support size, and precision of feedbacks. “XDEC” in (c) means that the feedback value is rounded with X decimals. 87

B.1 Calibration performance on the test data of MNIST. Note that the presented confidence on test data is from the softmax output to meet the definition of calibration. To this end, I simply replace sparsemax with softmax during inference, which will not influence the accuracy results. The dashed grey line in the right subfigure denotes the ideal average confidence over bins, and it is shifted leftward by a half of bin-width to visually align with the output confidence, i.e., red stars. 98

C.1 Comparison among three forms of model tuning. 102

List of tables

3.1 Statistics of datasets 38

4.1 Test error (%) of various distillation strategies based on VAT. The best results are marked in bold. 56

4.2 Performance comparison on four benchmarks. The best performance is marked as bold in two separate blocks. "MM" is short for "MixMatch", and "FM" is short for "FixMatch". 58

4.3 Test error on CIFAR-10 with 4,000 labels, where "Sp" is short for "Sparsemax". Note that Sparsemax+ME does not apply to FixMatch. 63

5.1 Common mathematical notations in this chapter. 69

5.2 Comparison of different model tuning methods on CIFAR-10-C and CIFAR-100-C with the highest severity. 82

5.3 The required query number (K) to achieve the preset tuning performance for two types of corruptions (Gaussian and Impulse) on CIFAR-10-C. 88