

Deep Learning with Noisy Supervision

Jiangchao Yao

Centre for Artificial Intelligence
University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

April, 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

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

This thesis is the result of a Collaborative Doctoral Research Degree program with Shanghai Jiao Tong University

Production Note:

Signature of Student: Signature removed prior to publication.

Date: *June, 28th 2019*

I would like to dedicate this thesis to my loving parents and family.

Acknowledgements

I would like to take this good opportunity to appreciate my advisors, several professors, my colleagues, my friends and my family for their significant help during my doctoral study in University of Technology Sydney and Shanghai Jiao Tong University.

First, I am deeply indebted to my supervisor Prof. Ivor W. Tsang, who improves my machine learning taste into a new stage. During my research life in University of Technology Sydney, I am deeply impressed by his broad knowledge in machine learning and his pure altitude toward the research. His patient guides on how to propose a novel idea across or to bridge two different fields and how to generalize its impact to many tasks, really make an great effect on my deep understanding of the research philosophy. Besides, he constantly encourages us to focus on the fundamental problem and make the enough effort to pursue the top research in the advance area, which moves me deeply in the journey of my research life. With his help, I found the happiness to do the pure research.

Second, I would like to express my foremost and deepest gratitude to my co-supervisor Prof. Ya Zhang, who plays an important role in showing me how to solve the difficulty in the doctoral life and how to do a significant academic research. In particular, at the beginning of my research career, her vision on the proposal of a novel idea in the research field and the endless labor on teaching me the writing of a regular academic paper helps me build the solid foundation of the research. Besides, her special interesting view on many works breaks through my original shallow sense and motivates the research going deep. It is hard to imagine I could finish this thesis without her high standards, unlimited patience, generous support and guidance.

I have been fortunate to attend the joint degree program between Shanghai Jiao Tong University and University of Technology Sydney for about two years. Wherein I would like to express my sincere appreciation to Prof. Ivor W. Tsang, Prof. Ya Zhang and Prof. Chengqi Zhang, who provides me this opportunity to work in such a great team and around so many brilliant minds. Within these two years, I have met and made

friends with many famous researchers in different areas and learned many useful experiences from them. In particular, with the help of Bo Han, Yuangang Pan and Yueming lyu, I have broaden the scope of the whole research community and dived in more rigorous solution for the difficulty to be solved. With the guide of Gang Niu and Mingyuan Zhou, I realized what I lack for the top research in detail, which is very important in my development.

Then, it comes to my dear colleagues and friends in my doctoral study. I would like to thank Dr. Zhe Xu for his selfless help in my research, Jiawei Hu, Shuai Xiao, Weichang Wu, Xiaoyu Chen, Xiao Luo and Siyuan Zhou for being good friends, discussions and in their company, Zhiyi Tan and Fei Ye for being close friends and offering me the generous help, lovely juniors Jie Hou, Yichao Xiong, Zhiwei Rao, Xiaohui Zhu, Yan Wang, Yexun Zhang, Hongxiang Cai, Zhuoxiang Chen, Xinyan Yu, Zhonghao Wang, and young brothers Huangjie Zheng, Jiajie Wang, Xu Chen, Jie Chang, Conan Cui, Maosen Li, Hao Wu and many others. I am happy to experience the research life with you all and wish you all have a bright future. I am also grateful to all the other friends in Sydney: Bo Han, Donna Xu, Yuangang Pan, Zhuanghua Liu, Yueming Lyu, Muming Zhao, Xiaoqi Jiang, Yaxin Shi, Yan Zhang, Jing Chai, Weijie Chen, Tao Zheng, Xingrui Yu, Xiaowei Zhou, Jing Li, Yinghua Yao, Xiaofeng Xu, Xiaofeng Cao, Yurui Ming, Guang Li, Xinyuan Chen, Xuan Wang and Razia Osman for their support and company.

Finally, I would like to express my deeply felt gratitude to my parents, my sister, my wife and my daughter. You all always stand behind me and gives me the unconditional support. It is you contribute the happy lifetime to me so that I could purse my research goal without any concerns. This love will be remembered by me forever.

Abstract

Central to many state-of-the-art classification systems via deep learning is sufficient accurate annotations for training. This is almost the bottleneck of all machine learning algorithms deployed with deep neural networks. The dilemma behind such a phenomenon is essentially the trade-off between the low expensive model design and the low expensive sample collection. For practical purposes to alleviate this issue, learning with noisy supervision is a critical solution in the Big Data era, since the noisily annotated data on the social websites and Amazon Mechanical Turk platforms can be easily acquired. Therefore, in this dissertation, we explore to solve the fundamental problems when training deep neural networks with noisy supervision.

Our first work is to introduce the low expensive noise structure information to overcome the decoupling bias issue existed learning with noise transition. We study the noise effect via a variable whose structure is implicitly aligned by the provided structure knowledge. Specifically, a Bayesian lower bound is deduced as the objective and it naturally degenerates to previous transition models in the case that there is no structure information available. Furthermore, a generative adversarial implementation is given to stably inject the structure information when training deep neural networks. The experimental results show the consistently improvement in the different simulated noises and the real-world scenario.

Our second work targets to substitute the previous ill-posed stochastic approximation to the noise transition with a rigorous stochastic re-allocation regarding the confusion matrix. This work discovers the reason that causes the unstable issue in modeling the noise effect by a neural Softmax layer and introduces a Latent Class-Conditional Noise model to overcome it. In addition, a computational effective dynamic label regression method is deduced for optimization, which stochastically trains the deep neural network and safeguards the noise transition estimation. The proposed method achieves the state-of-the-art results on two toy datasets and two large real-world datasets.

The last work aims to alleviate the difficulty that the ideal assumption on the accurate noise transition is usually not fulfilled and the noise could

still pollute the classifier in the back-propagation. We specially introduce a quality embedding factor to apportion the reasoning in the back-propagation, yielding a quality-augmented class-conditional noise model. On the network implementation, we elaborately design a contrastive-additive layer to infer the latent variable and deduce a stochastic optimization via reparameterization tricks. The results on a noisy web dataset and a noisy crowdsourcing dataset confirm the superiority of our model in the accuracy and interpretability.

Contents

Contents	vii
List of Figures	x
List of Tables	xv
1 Introduction	1
1.1 Motivation	1
1.2 Learning with Noisy Supervision	3
1.3 Deep Learning with Noisy Supervision	4
1.4 Contributions and Organizations	7
1.5 Publications during PhD Study	9
2 Background	11
2.1 Large Image Sources with Noisy Supervision	11
2.2 Deep Learning with Noisy Supervision	13
2.2.1 Learning via Sample Re-weighting	13
2.2.2 Learning via Model Regularization	15
2.2.3 Learning via Noise Transition	16
2.3 Advances in Optimization and Theoretical Analysis	18
2.3.1 Variational Inference	18
2.3.2 Monte Carlo Sampling	19
2.3.3 Generalization Bound	20
3 Mask the Decoupling Bias	22
3.1 Introduction	23
3.2 A New Perspective of Noisy Supervision	25
3.3 Noise Structure against The Decoupling Bias	27
3.3.1 Deficiency of benchmark models	27
3.3.2 Does structure matter?	28
3.3.3 Straightforward dilemma	29

3.4	When Structure Meets Generative Model	29
3.4.1	The Structure-Aware Probabilistic Model	29
3.4.2	Towards Principled Realization	31
3.5	Experiments	33
3.5.1	Experimental Settings	34
3.5.2	Results Analysis	36
3.6	Summary	39
4	A Stable Stochastic Approximation	40
4.1	Introduction	41
4.2	The Proposed Framework	43
4.2.1	Preliminaries	43
4.2.2	Latent Class-Conditional Noise model	44
4.2.3	Dynamic Label Regression	44
4.2.4	Safeguarded Transition Update	48
4.2.5	Complexity Analysis	49
4.3	Extensions on Outlier and Semi-supervised Learning	50
4.3.1	Extension on Outlier Learning	50
4.3.2	Extension on Semi-supervised Learning	51
4.4	Experiments	52
4.4.1	Datasets and Baselines	52
4.4.2	Implementation Details	53
4.4.3	Results on CIFAR10 and CIFAR-100	54
4.4.4	Results on Clothing1M and WebVision	57
4.5	Summary	60
5	Towards Real-World Complex Noise	62
5.1	Introduction	63
5.2	The Quality Augmented Probabilistic models	66
5.2.1	Preliminaries	66
5.2.2	Residual Noise Effect	66
5.2.3	The Quality Augmented Probabilistic Model	67
5.2.4	Variational mutual information regularizer	68
5.2.5	Objective	69
5.3	Optimization	70
5.3.1	Reparameterization tricks	70
5.3.2	Annealing optimization	71
5.3.3	Differences from VAE	71
5.4	Contrastive-Additive Noise Network	72
5.4.1	Architectures	72
5.4.2	Contrastive layer and additive layer	74

CONTENTS

5.5	Experiments	76
5.5.1	Datasets	76
5.5.2	Experimental Setup	77
5.5.3	Classification results	78
5.5.3.1	Controlled experiments with artificial noise	81
5.5.4	Model Visualization	82
5.6	Summary	86
6	Conclusion	88
6.1	Summary of Contributions	88
6.2	Future Works	89
	Deduction in Chapter 4	90
6.3	Proof of $\Delta_{\mathcal{F}}$ regarding to the two quantities	90
6.4	Proof of Theorem 4.2	91
	Deduction in Chapter 5	92
6.5	Variational mutual information regularizers	92
6.6	KL-divergences and regularizers	92
6.7	Stochastic variational gradient	93
	References	95

List of Figures

1.1	(a) A simple review of some popular algorithms applied to image classification in the last two decades. We approximately place them into this axis according to the parameter size and the classification performance. (b) The procedure of building one standard image classification dataset.	2
1.2	The algorithmic tree to depict the traditional methods for learning with noisy supervision.	3
1.3	The algorithmic tree to depict the methods for deep learning with noisy supervision.	5
1.4	The methodological advantage of learning via noise transition in dealing with noisy data. The current methods of the other two methodologies do not possess such three merits simultaneously. . . .	6
1.5	The three issues in current learning via noise transition. (a) The decoupling bias issue induced by the large capacity of deep neural networks. (b) The ill-posed approximation issue via the neural Softmax layer. (c) The residual noise effect due to the inaccurate learning of noise transition.	7
1.6	The research view of our thesis as well as the relationship of three works. The first one focuses on the scenario where auxiliary information is available; the second one targets to make the training more stable and the last one moves the ideal class-conditional noise to the more complex noise and investigates how to deal with the residual noise effect. Such three works are respectively applicable to different scenarios and also combinable when the corresponding assumption meets.	8
2.1	The illustration of six large popular real-world image datasets with noisy supervision.	14
2.2	The illustration of learning via sample re-weighting	15
2.3	The illustration of learning via model regularization.	16
2.4	The illustration of learning via noise transition	17

LIST OF FIGURES

2.5	Variational auto-encoder. x is the observed samples and z is the hidden variable. The solid line and the dash line respectively represent the generative process and the inference process.	18
2.6	The visualization of a 2-D Gibbs sampling for a 2-D Gaussian distribution. The instances are sequentially sampled respectively along x-axis and y-axis in each iteration, which approximates the desired samples from the ideal 2-D Gaussian distribution.	20
3.1	The quantitative analysis on the degeneration of the noise transition induced by the decoupling bias of deep neural networks. (a) The image samples of the <i>CIFAR-10</i> dataset. (b) The simulated noise transition matrix to disturb the original clean labels. (c) The inferred noise transition by the two-step solution in [Patrini et al., 2017]. (d) The inferred noise transition by the noise adaptation layer in [Goldberger and Ben-Reuven, 2017].	24
3.2	Three types of noise structures: (a) column-diagonal noise structure, (b) tri-diagonal noise structure, (c) block-diagonal noise structure. The vertical axis denotes the class of ground-truth label, while the horizontal axis denotes the class of noisy label. The white cell means the valid class transitions, while the black cell means the invalid class transitions.	26
3.3	Graphical representation of benchmark models and our MASKING model. Assume that, (x, y) denotes the image with the noisy label and z represents the latent ground-truth label. ϕ is the noise transition matrix, where $\phi_{ij} = \Pr(y = e^j z = e^i)$. (a) A benchmark model. (b) our MASKING that models the noise transition matrix by an explicit variable ϕ . We embed a structure constraint on the variable ϕ , where the structure information comes from the given human cognition h . . .	27
3.4	The effect of the noise structure and the decoupling bias. The decoupling bias encourages deep neural networks memorize all data instead of reasoning it to the noise transition. The noise structure forces the valid transition happens and as much as possible prevents the invalids transition degeneration, which performs as the reverse action force of the decoupling bias in the optimization.	28
3.5	A GAN-like implementation to model the structure instillation in our MASKING. The generator is to output the noise transition for the reconstructor and the discriminator. For the former, ϕ is directly utilized to model transition combined with the classifier $P(z x)$ and for the latter, ϕ will be extracted the noise structure by $f(\phi)$ to align the human provided structure $\hat{\phi}_o$ via \mathcal{M}	33

LIST OF FIGURES

3.6	The network configurations of the generator module and the discriminator module. (a) The generator. (b) The discriminator. N is the class number.	35
3.7	Test accuracy vs iterations on benchmark datasets with three types of noise structures, i.e., the column-diagonal structure, the tri-diagonal structure and the block-diagonal structure.	36
3.8	The visualization of (a) the manually provided transition, (b) the candidate transition and (c) the learned noise transition via MASKING. Note that, the candidate transition consists of the bright cells that represents valid transition, black cells that represents invalid transition and the gray cells that represents uncertain transition. Only the former two are used as the prior of MASKING.	38
4.1	Safeguarded dynamic label regression for LCCN. The images and noisy labels are respectively input to the classifier and the safeguarded Bayesian noise modeling to compute the prediction and the conditional transition. Then, the latent true labels are sampled based on their product and then used for the classifier training and the safeguarded Bayesian noise modeling.	42
4.2	Latent Class-Conditional Noise model. x and y are the observed training pair. z is the latent true label. ϕ is the unknown noise transition. α is the hyperparameter of the Dirichlet distribution. N is the sample number and K is the class number.	44
4.3	The Generalized Latent Class-Conditional Noise model. x and y is the observed training pair. z is the partially observed latent label (the observed samples are for semi-supervised learning). $\phi \in \mathbb{R}^{(K+1) \times K}$ is the unknown noise transition (the extra dimension is for outlier learning). α is the hyperparameter of the Dirichlet distribution. N is the sample number and K is the class number.	50
4.4	Extensions based on the original safeguarded dynamic label regression for LCCN. The images and noisy labels are respectively input to the classifier and the safeguarded Bayesian noise modeling to compute the prediction (including the outlier component) and the conditional transition. When the clean labels are not available as the latent labels, they are sampled based on the product of previous two quantities, and otherwise they keep unchanged. Then, the latent labels composite by both the clean parts and those inferred from sampling are used to train the classifier. In addition, only the latent labels inferred from the Gibbs sampling are used to refine the Bayesian noise model.	51

LIST OF FIGURES

4.5	The loss curve (the left panel) and the accuracy curve (the right panel) of LCCN in the training procedure on the CIFAR-10 dataset with the different noise rates $r = 0.1, 0.3, 0.5, 0.7, 0.9$	55
4.6	The test accuracy of LCCN and S-adaptation in the training on CIFAR-10 with $r=0.5$ and the corresponding histograms for the change of noise transition ϕ via a mini-batch of samples.	56
4.7	The colormap of the confusion matrix on CIFAR-10 with $r=0.5$. We utilize the log-scale for each element in the confusion matrix for the fine-grained visualization. The left three maps are respectively learned by LCCN at the beginning of the training, the 30,000th step and the end of the training. The most right one is the groundtruth disturbed from the original clean CIFAR-10 dataset.	56
4.8	The label correction ratio in the training of LCCN on CIFAR-10 with $r=0.5$. We illustrate some negatively corrected samples (the red box) and some positively corrected samples (the green box) based on the high probabilities to be right or wrong at the 30th, the 70th and the 110th epochs.	57
4.9	Some representative samples in the training set that are considered as the outliers by LCCN*. We intuitively summarize these photos into four categories based on their contents, multiple different objects (RED), implicit categories (GREEN), uncertain types (BLACK) and confusing appearance (BLUE), which are respectively marked by the color of the surrounded boxes. Outliers are relative to LCCN and may contain hard examples of the pre-defined 14 categories.	58
5.1	Analysis about back-propagation in previous methods that model the noise transition, as well as our idea to avoid the effect of residual label noise. (a) All images are forward into the model and the mismatch error caused by both label estimation and label noise are back-propagated. (b) With quality embedding as a control from latent labels to predictions, the negative effect of label noise is reduced in the back-propagation by diverting the reasoning to the quality embedding branch.	64
5.2	The quality augmented probabilistic model. The shaded nodes are the observed image X and its noisy label vector Y . The latent label vector Z and the quality variable vector S are latent variables. Solid lines and dashed lines represent the generative process and the inference process.	68

LIST OF FIGURES

5.3	Difference between the conventional auto-encoder and our model. Solid lines and dashed lines respectively represent generative (decoding) and inference (encoding) procedures. (a) The conventional variational auto-encoder is symmetric that the observed knowledge is used to encode to the latent variables and decoded symmetrically. (b) Our model uses an auxiliary variables X in this encoding-decoding procedure and meanwhile learns a discriminative model from X to Z	72
5.4	The end-to-end network architecture, which consists of four modules, encoder, sampler, decoder and classifier. Encoder tries to learn latent labels and evaluate the quality of noisy labels; sampler is used to generate samples from the encoder outputs; decoder tries to recover noisy labels from samples. Meanwhile, our classifier is learned based on KL-divergence between $q(z)$ and $P(z)$	73
5.5	The instance number of each class on the <i>WEB</i> dataset (left) and the <i>AMT</i> dataset (right).	76
5.6	Classification results with different training sizes. We sample the subsets of <i>WEB</i> and <i>AMT</i> with five different ratios for training, and evaluate all models on <i>V07TE</i> , <i>V12TE</i> and <i>SD4TE</i>	81
5.7	Quality embedding visualization of two categories on <i>WEB</i> and two categories on <i>AMT</i> . Better distinguishability of clusters indicates better identifiability of mismatches between latent labels and noisy labels. Blue: trustworthy embedding, Green: non-trustworthy embedding. . .	83
5.8	Exemplars on latent label estimation of <i>WEB</i> dataset (the first two rows) and <i>AMT</i> dataset (the third row) as well as some failures (the fourth row). We forward the noisy label (black word in title) and the image into CAN and compute the latent label (red word in title).	84
5.9	Transition patterns among labels conditioned on the trustworthy embedding and the non-trustworthy embedding on <i>WEB</i> dataset (the first two panels) and <i>AMT</i> dataset (the last two panels). Transition conditioned on the trustworthy embedding requires the consistency between the latent label and the noisy label, and thus concentrates on the diagonal. Transition conditioned on the non-trustworthy embedding identifies the mismatch between the latent label and the noisy label, and thus diffuses from the diagonal.	86

List of Tables

2.1	A summary of the popular real-world datasets in the area of deep learning with noisy supervision. From left to right, we respectively list the dataset name, the release year, the category number, the sample number, the label number for each sample and the inter-class granularity of the dataset as well as whether there is a open test set available provided by the release organization.	13
3.1	The estimation of the noise transition matrix by F-correction (1st row), S-adaptation (2nd row) and MASKING (3rd row), and the truth (4th row) in the case of three types of noise transition structure: column-diagonal (1st column), tri-diagonal (2nd column), block-diagonal (3rd column).	37
3.2	Test accuracy on <i>Clothing1M</i>	38
4.1	The average accuracy (%) over 5 trials on CIFAR-10 and CIFAR-100 with different noise.	54
4.2	The average accuracy (%) over 5 trials on outlier-corrupted CIFAR-10 and CIFAR-100.	54
4.3	The average accuracy over 5 trials on <i>Clothing1M</i>	58
4.4	The learned noise transition on <i>Clothing1M</i> by LCCN.	59
4.5	The average accuracy over 5 trials on <i>WebVision</i>	60
5.1	Classification results on the 20 categories of <i>V07TE</i>	78
5.2	Classification results on the 20 categories of <i>V12TE</i>	79
5.3	Classification results on 4 categories of <i>SD4TE</i>	80
5.4	Classification results with different regularization coefficient λ in CAN.	81
5.5	Classification results with different levels of artificial noise on <i>V07TE</i> , <i>V12TE</i> and <i>SD4TE</i>	82