Research Statement

Xiong Zhou
School of Computer Science and Technology
Harbin Institute of Technology
cszx@hit.edu.cn

June 10, 2022

As a researcher in artificial intelligence and machine learning, my long-term goal is to **develop trustworthy machine learning (ML) models with effectiveness, robustness, fairness, and interpretability for our society to improve human lives**. During my Ph.D study, I approached this goal from the following three perspectives:

(1) **Developing robust ML models for learning with noisy labels**: Deep neural networks have achieved impressive performance in various tasks, such as image classification, segmentation, object detection, etc. One critical factor that attributes to the success of deep learning is the availability of large amounts of high-quality annotated training data. However, in many real-world scenarios, it may be difficult to attain perfect supervision for fully supervised learning due to a variety of limitations, such as noisy labels due to inaccurate human annotation or human bias. With respect to massive but noisy data, my research attempts to develop effective and theoretical sound methods to achieve noise-robust learning.

(2) **Developing fairness-aware ML models for mitigating biases in data**: Recent studies have revealed potential societal biases in ML models due to that data contains implicit discrimination. There is a large ethical concern about the disparate impact these models will have at deployment time, especially on ethnic minority subpopulations and other underrepresented communities. For instance, machine learning models trained by minimizing average loss on imbalanced datasets will result in representation disparity—minority groups contribute less to the training objective and thus tend to suffer lower accuracy. To make matters worse, as model accuracy affects user retention, a minority group can shrink over time (Hashimoto et al., 2018). My research aims at developing fairness-aware ML models to mitigate explicit and implicit biases in data or models.

(3) **Developing theoretically solid ML models for promoting explainable AI**: The remarkable development of deep learning relies heavily on sophisticated heuristics and tricks without much principled guidance from a theoretical perspective. This poses a great obstacle to the interpretability of deep learning models due to the intricate details of neural networks, which also forces most practitioners and researchers to regard them as black boxes with little that could be understood. To better exploit its potential, it is necessary to build rigorous theoretical principles for reasoning about deep learning.

In the past several years, I focus on trustworthy machine learning related to the broader study of robustness, fairness, privacy and security in ML. I attempt to develop a rigorous understanding of the brittleness, vulnerabilities and biases inherent to machine learning and to develop the fundamental tools, metrics and

methods to understand, explain and mitigate them. I seek to come to a consensus on a rigorous framework to formulate the trustworthy machine learning problems, with an emphasis on building theoretically sound, elegant, and scalable methods. In the following, I will elaborate my previous works and discuss the future directions.

Learning Robust Models with Imperfect Data (ICML'21, ICCV'21, ICML'22)

One common and economical way to collect large amounts of training data is through online queries or crowdsourcing, which, however, inevitably introduce imperfect data due to that no domain expertise is involved and a variety of human biases exit. For example, noise labels would lead deep neural networks to overfit mislabeled data, which seriously hampers the generalization ability of neural networks. To mitigate label noise, robust loss functions are introduced to train deep neural networks with better noise tolerance and generalization power in the presence of noisy labels.

Although many papers that derive robustness of the losses had already been proposed and widely-adopted when I started working on this field, all of these methods tried to seek noise tolerance for any noise rates under some noise types, and had to satisfy restrictive conditions (*e.g.*, symmetric losses (Ghosh et al., 2017)) leading to severe underfitting. In reality, the noise rate is usually within a certain range, indicating that we can design robust losses for specific label noise models and release the fitting ability from overly restrictive conditions. In (Zhou et al., 2021a), we introduce a reasonable assumption, called *clean-labels-domination assumption*, which means that samples have a higher probability of being annotated with true semantic labels than any other class labels. Based on this fundamental basic, we propose a new family of loss functions, namely *asymmetric loss functions*, which utilizes the strategy of *the-largest-takes-all* to satisfy the Bayes-optimal prediction. We also introduce the concept of completely asymmetric losses and strictly asymmetric losses, where completely asymmetric losses are theoretically proved to include symmetric losses as a special case. We then investigate general theoretical properties of asymmetric losses, including classification calibration, excess risk bound, and noise tolerance. Meanwhile, we introduce the asymmetric ratio to measure the asymmetry of a loss function, which, together with the clean level of labels, can be associated with noise tolerance. Specifically, we derive the necessary and sufficient conditions to determine whether a loss is robust at a given noise rate. Empirical results show that a higher asymmetry ratio will provide better noise robustness. Moreover, we modify several commonly used loss functions and establish the necessary and sufficient conditions for them to be asymmetric. Asymmetric losses greatly expand existing robust loss functions and promote the emergence of new theories and methods.

Motivated by the condition for a loss to be asymmetric, in (Zhou et al., 2021b), we theoretically prove that *any loss can be made robust to noisy labels by restricting the network output to the set of permutations over a fixed vector*. When the fixed vector is one-hot, we only need to constrain the output to be one-hot, which, however, produces zero gradients almost everywhere and thus makes gradient-based optimization difficult. To alleviate this problem, we introduce the *sparse regularization* strategy to approximate the one-hot constraint, which is composed of a network output sharpening operation that enforces the output distribution of a network to be sharp and the ℓ_p-norm ($p \leq 1$) regularization that promotes the network output to be sparse. This simple approach guarantees the robustness of arbitrary loss functions while not hindering the fitting ability. Experimental results demonstrate that our method can significantly improve the performance of commonly used loss functions in mitigating label noise.

Unlike directly developing robust models in the presence of noisy labels, we attempt to analyze the under-

lying issue in the model training process. In (Zhou et al., 2022b), we observe that the presence of imperfect data can lead to unstable and defective class prototypes during training; therefore, we propose prototype-anchored learning (PAL) which aims to tighten a margin-based generalization bound by maximizing the minimal sample margin that depicts a measure of intra-class compactness and inter-class separation. Based on PAL, we can easily derive a novel robust loss called Negative-signed Sample Logit (NSL) loss. We further extend the classical symmetric condition to a more general theorem and reveal that our PAL strategy in which the prototypes are anchored can lead to a tighter bound than that without PAL, especially for some losses like CE that are usually not Lipschitz continuous. We also note that the feature norm can be regarded as a new trade-off between fitting ability and robustness. Based on these conclusions, finally we suggest the *feature normalized and prototype-anchored learning* (FNPAL) that performs ℓ_2 normalization on features and anchored prototypes, which can be combined with traditional losses to boost their ability on noise-tolerant learning.

Future Direction Although I have focused mainly on the robustness of learning with noisy labels, this topic naturally related to other tasks of learning with imperfect data due to its ubiquitousness. For example, partial label learning can be regarded as an easier special case; pseudo-labeling, one of the most commonly used techniques in semi-supervised learning, usually introduces wrong labels that need to be mitigated; and in distillation learning, it is not guaranteed that the teacher model will provide completely correct knowledge. Moreover, I am particularly interested in considering how to realize noise-tolerant regression that plays a very important role in the field of self-supervised image restoration, considering how to construct a robust label propagation with noisy labels in graph processing, and considering how to alleviate the influence of erroneously utilizing false positive pairs or false negative pairs in contrastive learning.

Mitigating Biases in Data and Models (ICLR'22, ICML'22)

Real-world data usually exhibit a highly biased distribution, due to the nature that some classes (or groups) are easy to collect sufficient examples and many classes are associated with only a few. This issue referred to as *class imbalance* would cause naive learning biased towards the majority classes while with poor accuracy on the minority ones, which would also lead to unfairness—*representation disparity* and *disparity amplification* in any deployed machine learning system that is retrained in a feedback loop. More specifically, class imbalance severely affects the intra-class compactness and inter-class separation of features, which leads to imbalanced margins for both class and samples. In (Zhou et al., 2022a), to handle imbalanced data, we prove a sufficient condition, which reveals that if the centroid of the prototypes is equal to zero, learning from the margin-based losses (or termed GM-Softmax in our paper) will provide the largest margins. Accordingly, we propose a simple yet effective *zero-centroid regularization* term which regularizes the prototypes to obtain an unbiased distribution of the prototypes. Unlike other re-balance strategies (or fairness-aware constraints) that either explicitly impose restrictions on model predictions or make any adjustment for loss, zero-centroid regularization can implicitly achieve fairness between different classes.

In (Zhou et al., 2022b), inspired by the goal of *learning towards the largest margins* (Zhou et al., 2022a), we thoroughly investigate the softmax loss and its variant—margin-based losses, and claim that to tighten the generalization error bound, one can maximize the minimal sample margin. We prove that minimizing the margin-based loss with the same per-class margins will lead to the maximum of the minimal sample margin on balanced datasets, that is, the distribution of prototypes keeps the most discriminative, and each feature concentrates on the corresponding prototype. However, the balanced distribution of prototypes and features will be damaged when learning from imbalanced datasets. We further derive the optimality condition of

the prototypes to obtain the largest $\gamma_{\min}$, indicating that the class prototypes $\boldsymbol{w}_1, ..., \boldsymbol{w}_k$ belonging to the unit sphere $\mathbb{S}^{d-1}$ ($2 \leq k \leq d+1$) should satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}$, $\forall i \neq j$, where k is the class number. Motivated by theoretical analysis, we propose a simple yet effective method, namely *prototype-anchored learning* (PAL). Specifically, we derive a classifier composed of anchored prototypes, which are predefined by a simple implementation that only requires knowing the number of classes and the feature dimension. Due to its simplicity, PAL can not only be used in the feature representation learning of decoupling methods for class-imbalanced learning but can also be easily incorporated into various learning-based classification schemes, as well as in tandem with other performance-boosting approaches and modules.

Future Direction Our recent work on mitigating class imbalance to achieve fairness has focused mainly on discrete targets, while real-world data also involve continuous targets, *e.g.*, age and income. A natural research direction is to build theories and methods for imbalanced regression that is still in an early stage and lacks an effective approach. The main challenge is not only the change of target form but also the presence of missing data in certain target regions. Therefore, I will focus mainly on modeling missing data and establishing the imbalanced regression paradigm.

Building Rigorous Theoretical Principles (ICLR'22)

One of the main challenges for feature representation in deep learning is the design of appropriate loss functions that exhibit strong discriminative power. The classical softmax loss does not explicitly encourage discriminative learning of features. A popular direction of research is to incorporate margins in well-established losses in order to enforce extra intra-class compactness and inter-class separability, which, however, were developed through heuristic means, as opposed to rigorous mathematical principles, modeling, and analysis. Although they offer geometric interpretations, which are helpful for understanding the underlying intuition, *the theoretical explanation and analysis that can guide the design and optimization are still vague*. Some critical issues are unclear, *e.g.*, why is normalization of features and prototypes necessary? Is there really an intrinsic difference between the kinds of loss proposed by previous work? How can a loss function be further improved or adapted to new tasks? Therefore, it naturally raises the following fundamental question: How to develop a principled framework for better understanding and design of margin-based loss functions?

In (Zhou et al., 2022a), we attempt to address these questions by formulating the principled optimization objective as *learning towards the largest margins*. Specifically, we propose to employ the class margin as the measure of inter-class separability and the sample margin as the measure of intra-class compactness. Accordingly, to encourage discriminative representation of features, the loss function should *learn towards the largest possible margins for both classes and samples*, which also complies to tighten a margin-based generalization error bound. We provide a rigorous theoretical guarantee that maximizing the minimal sample margin will lead to the maximum of class margin regardless of feature dimension, class number, and class balancedness. Furthermore, we derive a generalized margin softmax loss to draw general conclusions for existing margin-based losses. We show that learning with existing margin-based loss functions would share the same optimal solution. In other words, all of them attempt to learn towards the largest margins, even though they are tailored to obtain different desired margins with explicit decision boundaries. Not only does this principled framework offer new perspectives for understanding and interpreting existing margin-based losses, it also provides new insights that can guide the design of new tools. We then propose an explicit *sample margin regularization* and a novel *largest margin softmax loss* (LM-Softmax) derived from the minimal sample margin, which significantly improve the class margin and the sample margin. Extensive experimental results are offered to demonstrate that the strategy of learning towards the largest margins

significantly can improve the performance in accuracy and class/sample margins for various tasks, including visual classification, person re-identification, and face verification.

Future Direction In the scope of learning towards the largest margins, out-of-distribution detection will be a natural extension since it requires that the model is capable of distinguishing out-of-distribution samples with separation of in-distribution samples. While there are still many interesting unanswered questions in feature representation with enough discriminativeness and interpretability, I am also interested in building rigorous theoretical principles beyond the scope of classification-based learning towards the largest margins. One next direction is the self-supervised learning scenario, which aims at learning transferable representations with as much semantic knowledge as possible. The question of how to define the measures and optimizeable objectives between semantic knowledge is worthy of further study.

Research Plan

During my current graduate school, I have carried out in-depth studies on the theory and methodology of robustness, fairness, and interpretability in trustworthy machine learning. I particularly focus on learning with imperfect data, including learning from noisy labels, imbalanced learning, and metric learning, as well as their corresponding practical applications. In the near future, I will continue investigating these topics and intend to delve further into other fields of trustworthy machine learning, including self-supervised learning, semi-supervised learning, adversarial robustness, partial label learning, out-of-distribution detection, and so on. These directions may seem different from my previous work, but they still have a close bond. I am particularly excited about advancing fundamental AI capabilities while addressing pressing social needs, such as AI for drug discovery, renewable energy, and environmental protection. I look forward to exploring these multidisciplinary problems with collaborators from academia and industry.

References

Aritra Ghosh, H. Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. Asymmetric loss functions for learning with noisy labels. In *International Conference on Machine Learning*, pp. 12846–12856. PMLR, 2021a.

Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 72–81, 2021b.

Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Learning towards the largest margins. In *International Conference on Learning Representations*, 2022a.

Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Prototype-anchored learning for learning wtih imperfect annotations. In *International Conference on Machine Learning*, 2022b.