Towards Reliable Machine Learning with Unlabeled Data

Research Priorities. I am a fourth-year Computer Science Ph.D. student at UW-Madison with an interest in contributing to a deeper understanding of reliable machine learning. In today's rapidly evolving landscape of AI technology, machine learning models, including the recent foundation models (FMs) have emerged as transformative forces shaping various applications. Despite the immense capabilities, they bring forth challenges related to the model's reliability upon deployment in the real world. For example, classic discriminative neural networks can produce overconfident yet incorrect predictions for unfamiliar out-of-distribution classes. Additionally, generative models, such as large language models (LLMs), can generate untruthful or harmful content that contradicts human values, thus compromising critical decision-making in an ever-changing open world.

As a critical part of AI Existential Safety research, my research project aims to bridge this gap by building a novel framework that properly leverages the out-of-distribution data for safety-aware learning. I develop new algorithmic techniques and provide fundamental and provable understandings that enable safe and reliable decision-making in the open world, where out-of-distribution (OOD) data can naturally arise. The research direction is largely underexplored and central to the roadmap of AI research in the next 5 years and beyond [1]. Moreover, the proposed algorithm framework can be synergistically integrated with the recent generative foundation models, making the research endeavor a relevant and timely manner.

As AI reaches society at large, the need for safe and reliable decision-making is increasingly critical. This requires intelligent systems to have an awareness of uncertainty and a mandate to confront unknown and open-ended situations with caution. Yet for many decades, machine learning methods commonly have made the closed-world assumption—-the test data is drawn from the same distribution as the training data (i.e., in-distribution data). Such an idealistic assumption rarely holds true in the open world, where test inputs can naturally arise from unseen categories that were not in the training data. When such a discrepancy occurs, algorithms that classify OOD samples as one of the in-distribution (ID) classes can be catastrophic. Consider self-driving cars as an example, the model may have to detect unknown objects (cows) that appear on the road, among the known objects such as "traffic signs" and "cars". As shown in my paper [6], the experimental analysis shows that the popular object detector can produce high-confidence predictions for out-of-distribution objects (e.g., cow), which were never exposed to the model during training.

Research Approach. Literature [7] has used the real outlier data as extra supervision besides the ID task during training, which demonstrates great empirical success since they are able to provide OOD data for training with explicit knowledge about the unknowns. Despite the promise, preparing auxiliary data can be labor-intensive and inflexible, and necessitates careful human intervention, such as data cleaning, to ensure the auxiliary outlier data does not overlap with the ID data.

Over the past 4 years, I have worked with my advisor, Prof. Sharon Li, on several projects in this area by investigating topics on how to effectively leverage alternative unlabeled data for model regularization and improved OOD detection. These unlabeled data often contains a mixture of ID and OOD samples, which can be freely collectible in the model's deployment environment. Some possible instantiation of the unlabeled data can be synthetic outliers [6, 5, 8, 4], and wild data [2, 3]. Outcomes of my research can directly impact and improve safety, reliability, and deployability of intelligent systems in real-world environments, such as autonomous vehicles (e.g., unseen object detection), and chatbot applications (Hallucination detection). I proceed with two main technical contributions.

Part I: Leveraging Unlabeled Data for OOD Detection. The first part of my research efforts is to leverage different kinds of unlabeled data for OOD detection. My initial works aim to automate the outlier preparation efforts by synthesizing virtual outliers given only the ID data. I begin by sampling outliers in the low-dimensional feature space of discriminative neural networks. Comparing with synthesizing data by training generative models in the pixel space, the latent-based approach offers better training stability and optimization procedures. Specifically, the synthesized outliers are expected to lie in the low-likelihood region of the ID feature space, which are then directly used to regularize the machine learning model during training. The regularization objective is instantiated as an uncertainty loss term that aims to perform a level-set estimation that separates ID vs. OOD data. This idea has resulted in three publications, i.e., VOS on ICLR'22 [6], NPOS on ICLR'23 [8] and Dream-OOD on NeurIPS'23 [4].

On the theoretical side, my research leverages the unlabeled wild data for model regularization with rigorous analysis. Following the Huber's contamination model, the wild data is characterized as a mixed composition of two distributions with an unknown ratio of π. To address the challenge of lacking ground truth labeled OOD data, my ICML'23 work SCONE [2], employs a constrained optimization approach to train with unlabeled mixture data, providing analysis on the convergence rate. Additionally, my ICLR'24 work SAL [3] explicitly separates candidate OOD samples from the unlabeled data for training a binary OOD classifier, where I comprehensively study its generalization error bound.

Part II: Hallucination Detection for LLMs The second contribution explores the extension of my research into a more timely topic on foundation models. The surge in applications of large language models (LLMs) has prompted concerns about the generation of misleading or fabricated information, known as hallucinations. Therefore, detecting hallucinations has become critical to maintaining trust in LLM-generated content. A primary challenge in learning a truthfulness classifier is the lack of a large amount of labeled truthful and hallucinated data. To address the challenge, my ongoing work aims to propose a novel learning framework that leverages the unlabeled LLM generations in the wild for hallucination detection. Such unlabeled data arises freely upon deploying LLMs in the open world, and consists of both truthful and hallucinated information. To harness the unlabeled data, I design an automated membership estimation score for distinguishing between truthful and untruthful generations within unlabeled mixture data, thereby enabling the training of a binary truthfulness classifier on top. This research is closely connected to my expertise in OOD detection, and I aim to bridge the two related domains of OOD detection and LLM hallucination detection, delivering valuable insights into both areas.

Research Plan. My future research goal can be characterized as twofold. In the short term, I plan to build responsible foundation models, such as large language models and vision-language models. I am driven to deeply understand how they work, when they fail, and effectively align them with human needs and desires. Some promising directions can be developing effective hallucination detection and mitigation algorithms, and proposing formal theoretical framework to understand the reason of LLM hallucinations, as supported by existing literature [9]. In a longer term, I will continue to work on the topic of AI safety, especially OOD detection and generalization both algorithmically and theoretically. The research outcome will favorably facilitate the understanding of the safety and reliability of the ML algorithms across diverse data science application scenarios.

I am excited to discuss my research and potentially collaborate with different researchers all over the world. I eagerly anticipate the fruitful outcomes achieved!

References

[1] Computing research association. artificial intelligence roadmap. `https://cra.org/ccc/ai-roadmap-self-aware-learning`.

[2] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *ICML*, 2023.

[3] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *Proceedings of the International Conference on Learning Representations*, 2024.

[4] Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

[5] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *CVPR*, 2022.

[6] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *ICLR*, 2022.

[7] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[8] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *ICLR*, 2023.

[9] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.