Trustworthy AI Systems in Open World

Yabin Zhang (https://ybzh.github.io/)

We are in the era of Artificial Intelligence (AI). A vast number of AI systems are being developed at an unprecedented scale, liberating humans from heavy physical and repetitive tasks. Furthermore, in the foreseeable future, AI is bound to influence nearly every aspect of our modern society. As AI becomes more deeply integrated and intertwined with human life, the trustworthiness of AI systems has a crucial impact on human safety and property, especially in safety-critical applications such as authentication, auto-driving and healthcare. However, the real world is highly complex, characterized by vast and redundant sample sizes, diverse data categories and distributions, and long-tail phenomena, posing significant challenges to the trustworthiness of AI systems. For instance, an autonomous vehicle must accurately recognize and respond to both common objects like pedestrians and rare events like unexpected roadblocks. It must also distinguish between real people and images of people, ensuring safety and reliability. This leads to the following fundamental question in AI:

How to design AI systems that make trustworthy predictions in complex open-world environment?

Such problems, i.e., trustworthy AI, have been extensively investigated in many areas, including model robustness, fairness, adversarial defense, and so on. Each line of research focuses on a sub-problem of trustworthy AI with specifically designed evaluation strategies. However, improving trustworthiness in one dimension may compromise trustworthiness in other dimensions. In real-world applications, AI models often encounter various types of test data, necessitating a model trustworthy in a full spectrum. This raises another fundamental question in AI trustworthiness:

How can we evaluate the trustworthiness of AI systems in the full spectrum?

Motivated by above questions, my passion and research interest lie broadly in the design, analysis, and implementation of novel algorithms and evaluation benchmarks for trustworthy AI. More specifically, I am interested in developing resource-efficient and interpretable learning systems that generalize robustly to covariate-shifted distributions and identify semantic-shifted out-of-distribution data. I am also interested in comprehensively evaluating new algorithms across the full spectrum of trustworthiness, including robustness to covariate-shifted input, reliability with adversarial input, prediction consistency with input perturbations, detection of semantic-shifted data, and calibrated prediction confidences. These investigations are particularly crucial in safety-critical real-world applications.

Prior Work: AI Safety and Robustness

As illustrated in Figure 1, my previous research has been centered around *Safety and Robustness* - an important sub-topic of trustworthy AI. My primary research goal is to develop robust and reliable AI systems that can handle not only in-distribution (ID) data but also out-of-distribution (OOD) data, which refers to data that fall outside the statistical distribution of the training data. Generally, OOD samples can be categorized into two primary types: covariate-shifted and semantic-shifted OOD samples. **Covariate-shifted OOD samples** maintain consistency in their semantic content—meaning the underlying concept or category remains the same as seen during training—but exhibit significant variations in covariates. These covariates can include changes in style, lighting, background, or imperceptible adversarial perturbations, which might not change the fundamental nature of the object but could still challenge a model trained on a constrained dataset. On the other hand, **semantic-shifted OOD samples** contain entirely different semantic labels from those the model was trained to recognize. These samples introduce new classes or concepts that were not included in the training data. This

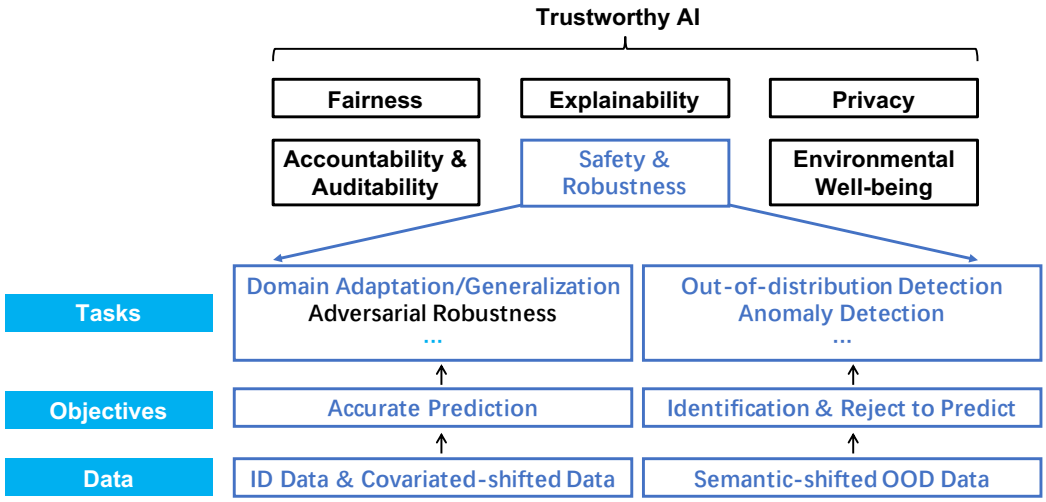

Figure 1: Project Overview: The blue part, e.g., Safety & Robustness, is the focus of prior work, while other parts will be explored in future plans. The overall structure follows [1].

type of OOD scenario is particularly challenging because it requires the model not only to recognize that these samples are different but also to handle them appropriately, often in the absence of any prior knowledge about these new categories. To tackle these OOD samples, my previous research is tightly connected to the following core areas in machine learning:

- **Distribution Matching:** Distribution matching minimizes the distribution divergence between ID and covariate-shifted OOD samples, which is a popular method to promote model robustness to certain target distribution. The theoretical understanding and algorithmic design of more precise metrics for distribution divergence measurement are the most crucial components.
- **Semi/Unsupervised Learning:** Semi/unsupervised learning aims to learn AI models with less manual supervision and can enhance the trustworthiness of AI models in two typical ways. Firstly, during the pre-training phase, massive amounts of unlabeled data can be leveraged through self/unsupervised learning to learn task-agnostic general representations, facilitating downstream task learning. Secondly, during the fine-tuning phase for specific tasks, unlabeled data can be utilized to enhance adaptation and learning with less manual supervision.
- **Data Augmentation:** From a certain perspective, the issue of AI trustworthiness often stems from training samples not adequately covering the real-world data distribution. Therefore, designing effective data augmentation strategies to ensure training samples cover more real-world scenarios can help improve the trustworthiness of AI models.
- **Multi-Modal Learning:** Different data modalities exhibit distinct characteristics that are highly complementary. For example, images provide a concrete representation of content but often contain redundant information, whereas text is highly abstract, compact, and structured. Considering both modalities are different expressions of the same physical world, textual information can provide structured, interpretable priors for processing image information, thereby enhancing the trustworthiness of image models.

Main Research Results

Theoretical Foundations and Algorithmic Designs for Distribution Measurement

Theoretical Foundations. In the TPAMI paper [2], I explored various theoretical variants of unsupervised domain adaptation (UDA), focusing on how to measure distribution divergence. We found that while theoretical adaptation conditions are strictly derived under the setting of binary classification with analysis-amenable loss functions, practical algorithms that are easier to optimize are often expected to be applied to cases involving multiple classes. To bridge this gap, we present the multi-class scoring disagreement (MCSD) divergence by aggregating the absolute margin violations in multi-class classification. This proposed MCSD is able to fully characterize the relations between any pair of multi-class scoring hypotheses. By using MCSD as a measure of distribution divergence, we develop a new theoretical bound for multi-class UDA. Its data-dependent, probably approximately correct bound is also developed, which naturally suggests adversarial learning objectives. These surrogate objectives either coincide with or resemble popular methods, thus underscoring their practical effectiveness.

Algorithmic Designs. Based on the above theoretical analyses for multi-class UDA, we also introduced a novel SymNets algorithm[2, 3], which features a novel adversarial strategy of domain confusion and discrimination. SymNets offers simple extensions that work equally well under the problem settings of either closed set, partial, or open set UDA. We also explored directly minimizing distribution divergence through statistical matching and proposed an efficient method to achieve exact alignment of higher-order statistics by aligning cumulative probability distribution functions [4]. Besides explicit distribution alignment methods, we examined the relationship between UDA and semi-supervised learning, carefully redesigning SSL methods to address distribution divergence problems in UDA [5]. These algorithmic designs significantly mitigate the challenges posed by distribution divergence to AI trustworthiness.

Multi-modal Learning: Enhanced Trustworthy Using Complementary Modalities

Enhanced Visual Trustworthy with Text Guidance. Excessive redundancy in image data can hinder the learning of trustworthy AI systems. Considering that textual information is highly abstract and compact, we leverage text to assist the learning of visual models. Specifically, we use textual information to enhance the robustness of image recognition and employ memory networks to improve the caching and utilization of data information [6]. Additionally, by leveraging the structured and readily available nature of text, we collect text information as anchors and use efficient and automatic prompt learning methods to improve OOD detection performance [7]. In ongoing research, we are investigating how to more effectively utilize the hierarchical structure of text to further assist robust recognition and OOD detection in the visual domain.

Self-supervised Pre-training Facilitates Trustworthiness

Effective Pre-training with Improved Local and Global Perception. Using pre-trained models to facilitate model convergence, performance, and trustworthiness is widely recognized. Masked Autoencoder (MAE), which reconstructs masked local regions as a self-supervised signal, is arguably the most popular strategy. Its effectiveness is validated with text, image, and point cloud input. By revealing that MAE primarily enhances the model's ability to extract local information through the corruption and reconstruction of local regions, we propose a complementary approach that improves the model's utilization of global information through the corruption and restoration of the global shape [8]. In experiments with point cloud input, our method not only enhances the model's robustness to input corruption but also brings significant performance improvements in classification, detection, and segmentation tasks.

Future Plans: Full Spectrum Trustworthy AI Systems in Open World

Defining and Addressing Novel Trustworthy AI Problems. I plan to explore broader areas in trustworthy AI, such as adversarial robustness, fairness, and privacy. Additionally, I believe the tasks of trustworthy AI will evolve over time. For instance, traditional adversarial robustness requires models to remain robust against imperceptible small changes. However, in the era of LLM, adversarial robustness may account for substantial changes to the input by intermediaries, which users still cannot detect. Another example is fairness. In traditional fairness research, the focus is on whether the model performs consistently across different populations. However, as AI is increasingly applied to critical areas such as healthcare and justice, the task of fairness evolves to ensure that models are fair across different medical conditions and legal contexts. This not only requires statistical fairness but also ensuring fairness in various specific application scenarios. Defining and addressing these new era tasks of trustworthy AI can significantly advance the real-world application of AI.

Full Spectrum Evaluation of AI Trustworthiness. Existing trustworthy AI algorithms are designed and evaluated separately for each sub-task. For example, model robustness and OOD detection are independent tasks with separate design methods and evaluation metrics. Even within robustness evaluation, robustness to adversarial attacks and robustness to general covariate shifts are parallel tasks. Such isolated investigation has significant limitations; for instance, a method effective for robustness might negatively impact OOD detection. Therefore, evaluating AI trustworthiness in the full spectrum is crucial for real-world applications, where test data encompasses a full spectrum, not just a single type. To achieve this goal, it's necessary to break down the existing barriers between sub-problem research and design a comprehensive metric to holistically assess AI trustworthiness.

New Dimensions of Trustworthiness in Generative Models. LLMs/VLMs such as Llama/GPT not only revolutionize research paradigms in natural language processing and computer vision but are also rapidly integrating into various aspects of human life. Compared to discriminative models like ResNet and Fast-RCNN, the trustworthiness of these generative models in the new era takes on new meanings. For example, generative models are capable of producing large volumes of text or images that appear highly realistic. This raises new trustworthiness concerns, such as the potential for generating harmful or misleading content, the ability to produce deepfakes, and the ethical implications of creating synthetic media. Additionally, the interpretability and accountability of these models become critical, as understanding the decision-making process of a generative model is often more complex than that of a discriminative model. Ensuring the alignment of these models with human values and societal norms is also a significant challenge, requiring robust mechanisms for monitoring and controlling outputs.

References

[1] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy ai: A computational perspective," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 1, pp. 1–59, 2022.

[2] **Y. Zhang**, B. Deng, H. Tang, L. Zhang, and K. Jia, "Unsupervised multi-class domain adaptation: Theory, algorithms, and practice," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2775–2792, 2020.

[3] **Y. Zhang**, H. Tang, K. Jia, and M. Tan, "Domain-symmetric networks for adversarial domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5031–5040.

[4] **Y. Zhang**, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8035–8045.

[5] **Y. Zhang**, B. Deng, K. Jia, and L. Zhang, "Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2020, pp. 781–797.

[6] **Y. Zhang**, W. Zhu, H. Tang, Z. Ma, K. Zhou, and L. Zhang, "Dual memory networks: A versatile adaptation approach for vision-language models," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[7] **Y. Zhang**, W. Zhu, C. He, and L. Zhang, "Lapt: Label-driven automated prompt tuning for ood detection with vision-language models," in *European Conference on Computer Vision*, 2024, pp. 000–000.

[8] **Y. Zhang**, J. Lin, R. Li, K. Jia, and L. Zhang, "Point-dae: Denoising autoencoders for self-supervised point cloud learning," *IEEE Transactions on Neural Networks and Learning Systems, under revision*, pp. arXiv–2211, 2022.