Research Statement

Yuting Yang

Over the past decade, machines interacting with humans in a natural language manner was becoming a reality and signaling a paradigm shift in the field of computer science and beyond. It is mainly due to breakthroughs in deep learning, especially large language models (LLMs) such as the recently released ChatGPT. These models demonstrate tremendous capabilities across a wide range of natural language processing (NLP) tasks and indicate a promising step toward artificial general intelligence. However, the sheer complexity of deep models keeps growing, making it increasingly difficult for human minds to form a comprehensive picture of all relevant elements and behaviors of the system and its environment. It leads to concerns about trustworthiness such as robustness, transferability, interpretability, fairness, privacy, ethics, and so on. One notable example is that even the state-of-the-art LLMs can output inaccurate or unstable results while adding minor perturbations to the inputs. These issues need to be addressed especially when NLP systems are widely adopted in safety-critical areas such as finance, healthcare and transportation. Therefore, my long-term research goal is to build **trustworthy AI systems that are adaptive to dynamic changes in environments**, which I believe is one of the key challenges in deploying NLP systems in more and more real-world settings. My research is launched from the following two perspectives as starting points: (1) Analyzing and improving the robustness of NLP models, and (2) Improving the transferability of NLP models.

1. Analyzing and improving the robustness of NLP models

Robustness refers to the ability of a system to withstand perturbations or external factors that may cause it to malfunction or provide inaccurate results. Existing research has shown the vulnerability of deep neural networks (DNNs) to adversarial examples, which are constructed via adding imperceptible perturbations to the original inputs and can "fool" a model into making wrong predictions. The trend of deploying deep models as interactive front ends of key systems has raised questions regarding their security. Much of my research is devoted to understanding a model's vulnerability to perturbations and improving its robustness.

1.1 Analysis of robustness

Analyzing and understanding robustness is the prerequisite for building robust NLP systems. Based on the conventional definition of robustness, a model is considered robust only if no adversarial example exists. Following this definition, I designed a dynamic programming-based attacking algorithm to measure the maximum boundary of robust regions where no adversarial example exists [1]. Analysis results showed that DNNs can be robust only in a very small region. Even the state-of-the-art LLM-based models which generalize well in real scenarios are not robust from this conventional view. Considering on-manifold adversarial examples are essentially generalization errors and fully safe systems may not exist[1], the requirement of "no adversarial example exists" may be too strict for DNNs. Therefore, I rethought the definition of robustness and try to establish the concept of "sufficiently safe", which is called "weak robustness". Unlike conventional robustness, weak robustness provides a quantitative view to

[1]Adi Shamir said "Fully secure systems don't exist now and won't exist in the future." on the keynote speech of RSA conference. This is the first among the laws of computer security.

evaluate the number of adversarial examples. This metric provides a more comprehensive understanding of the DNN's capability of resisting perturbation, especially for the input regions where adversarial examples exist. As exactly calculating this metric is computationally hard for large-scale DNNs, I proposed an efficient approximation algorithm with a rigorous probability error bound (like the PAC-style guarantee). Based on the evaluation results of weak robustness, I found an interesting property of DNNs: adversarial examples distribute broadly but only occupy a small percentage in the perturbation space. Thus, DNNs can handle random perturbations in non-adversarial scenarios and perform well in generalization.

1.2 Robustness enhancement

One of the most common robustness enhancement methods is adversarial training, which finds adversarial perturbations first and then optimizes them toward correct predictions. Researchers find that prompting paradigm can lead LLMs to generate harmful outputs. Thus, I wondered whether prompting can be used to explore the robustness defects of DNNs. Via constructing malicious prompts and utilizing LLMs to generate perturbations via mask-filling, prompting can efficiently generate more diverse and natural adversarial examples [2]. Further, I designed a lightweight adversarial training method which replaced the perturbation process in conventional adversarial training with prompt construction. The method can significantly improve the robustness of models to resist adversarial attacks.

Adversarial training depends on attack algorithms. Adversarial examples are everywhere, making it hard to find all of them. Besides, there is no theoretical guarantee for adversarial training. Inspired by fighting fire with fire, I proposed a robustness enhancement method to fight perturbation with perturbation [3] which omits the training process. Taking advantage of the weak robustness property, I utilized random perturbations to resist the adversarial perturbations crafted by attackers. Specifically, the method first conforms the input to the data distribution and then uses random perturbation to enhance predictions via voting. I provided a rigorous error bound for the method and experimental results showed that it can significantly improve the robustness of DNNs while maintaining the performance on the original clean data. I also extended this method to the image field [4].

Existing robustness enhancement approaches usually focus on some certain type of perturbation. As the types of attack can be various and unpredictable in practical scenarios (e.g., including several different levels of attacks), a general and strong defense method is urgently in require. I observed that ensemble methods are promising in defending attacks, if the sub-models satisfy weak robustness and the distribution of adversarial examples varies across sub-models. To cope with the problem that the ensemble method consumes a lot of computational resources, I proposed a lightweight ensemble framework to enhance robustness [5]. To diversify the distribution of adversarial examples among different sub-models, I promoted each model to have different attention patterns via optimizing an attention diversity measure. Experiments show that the proposed method can consistently improve the defense ability against a variety of adversarial attacks, including character, word and sentence-level attacks.

2. Improving the transferability of NLP models

The lack of transferability across different environments for deep NLP models raises the untrustworthiness of their wide application in the real world. Thus, improving the transferability of deep NLP models across dynamic environments is important for building reliable NLP

systems. Recently, with the increasing ability of LLM, a new paradigm is revolutionizing the NLP field: prompting. The paradigm is powerful and promising as it allows LLM to be able to adapt to new scenarios with few or no labeled data. In [6], I applied prompting to build trustworthy NLP models which can be applied to dynamic environments efficiently.

The worldwide popularity of chatbot service (ChatGPT) indicates the protentional of dialogue system as a mainstream interactive way in the future NLP field. Thus, I focus on a practical scenario in dialogue system where the environments are dynamic: extracting users' needs in different domains, also called Dialogue State Tracking task (DST). Specifically, I designed a dual prompt learning framework for few-shot DST. The framework considers DST as two dual sub-tasks (slot generation and value generation) and formalizes them to language modeling tasks via prompting. The dual prompt construction and training process can incorporate task-related knowledge from LLMs efficiently. Experimental results demonstrate that the proposed DST model can deal with various domains with few labeled data. It can even generate unseen slots, indicating the potential of handling new scenarios with the help of LLMs and prompting paradigm.

3. Future Research

Quantification view for fairness and privacy. In [5], I proposed to establish the concept of sufficiently safe for robustness from a quantification view. Similarly, I hope to extend this concept to fairness and privacy, and build sufficiently fair and privacy-protective systems. A possible solution to quantifying fairness can be perturbing the protected attributes (e.g., gender and race) in inputs and quantifying the changes in model outputs. It can help us to better understand fairness and step towards fairer systems. For privacy protection, a possible method is to measure the magnitude of information proliferation or the possibility of being inferred for specific sensitive attributes, which can quantify the risk of privacy leakage and then reduce such risk.

Prompt learning for trustworthy NLP. With the increasing scale of DNNs, prompting paradigm becomes a popular paradigm for developing and utilizing NLP systems. Prompting can help to explore trustworthiness issues of DNNs and tackle them efficiently. I explored it in [2] to detect robustness defects and improve robustness. The idea can also be applied in ethics to probe ethics-obeying contents from LLMs. Via in-context learning with few cases obeying ethics, LLM can learn to distinguish these harmful contents by analogy and avoid generating them. For interpretability, via chain-of-thought prompting, we can understand the process LLMs make decisions better.

Building trustworthy NLP systems with human-in-the-loop. The huge success of ChatGPT and GPT4 not only demonstrates the effect of the larger scale networks and data but also raises attention to the effect of human-in-the-loop in developing intelligent NLP systems. Reinforcement learning with human feedback elicits LLM's internal knowledge and aligns it with human cognition. I hope to explore more about human-in-the-loop in building trustworthy NLP systems. For example, human feedback in conversations can be utilized in understanding core semantics, adjusting history memory and making LLMs more reliable. For interpretability, I plan to explore how to symbolize human knowledge and integrate it into neural networks.

References

[1] **Yuting Yang,** Pei Huang, Feifei Ma, Juan Cao, Meishan Zhang, Jian Zhang and Jintao Li. Quantifying Robustness to Adversarial Word Substitutions. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (**ECML-PKDD'23**)*. (**Regular paper, Research track**)

[2] **Yuting Yang**, Pei Huang, Juan Cao, Jintao Li, Yun Lin and Feifei Ma. A Prompt-based Approach to Adversarial Example Generation and Robustness Enhancement, *Frontier of Computer Sciences,* 2023. (**SCI Journal**)

[3] Pei Huang *, **Yuting Yang***, Fuqi Jia, Minghao Liu, Feifei Ma and Jian Zhang. Word Level Robustness Enhancement: Fight Perturbation with Perturbation, *Thirty-Sixth AAAI Conference on Artificial Intelligence, **AAAI'22***, Vancouver, BC, Canada 2022. **(*Co-First Author, Regular paper, Research track, Acceptance Rate: 1349/9020=15.0%)**

[4] Pei Huang*, **Yuting Yang***, Minghao Liu, Fuqi Jia, Feifei Ma, and Jian Zhang. ε-weakened Robustness of Deep Neural Networks, *Thirty-First ACM SIGSOFT International Symposium on Software Testing and Analysis, **ISSTA'22***, Daejeon, South Korea, 2022. **(*Co-First Author, Regular paper, Research track, Top conference on software analysis)**

[5] **Yuting Yang**, Pei Huang, Juan Cao, Danding Wang, Jintao Li. PAD: A Robustness Enhancement Ensemble Method via Promoting Attention Diversity. (ARR with score 4, 3.5 and 2.5)

[6] **Yuting Yang**, Wenqiang Lei, Pei Huang, Juan Cao, Jintao Li and Tat-Seng Chua. A Dual Prompt Learning Framework for Few-Shot Dialogue State Tracking, *The Web Conference, **WWW'23***, Texas, USA 2023. (**Regular paper, Research track**)