Yibin Wang

wyb896409234@gmail.com · Google Scholar · yibinwang.netlify.app · Github

Research Interests: Trustworthy Artificial Intelligence

Education

Huazhong University of Science and Technology (HUST), Hubei, China Sept. 2019 – June. 2024
B.E. in Computer Science (CS), GPA: 3.82/4.00 - Transcripts

Experience

Generalization and Robustness of LLMs (Research Intern) June. 2024 – Present
Advisor: Huan Zhang - University of Illinois Urbana-Champaign (UIUC)

Generalization and Robustness of LLMs (Remote Research Intern) Sept. 2023 – May. 2024
Advisor: Hao Wang - Rutgers University

- Conducted extensive research on adversarial robustness of large language models.
- Conducted in-depth research on Bayesian algorithms and their applications in large language models.

Adversarial Robustness (Research Intern) Sept. 2021 – Aug. 2023
Advisor: Kun He - Huazhong University of Science and Technology

- Conducted extensive research on adversarial attack and defense in machine learning
- Conducted in-depth research on certified robustness based on convex relaxation

Publications

∗ indicates equal contribution

Robustness-Aware Word Embedding Improves Certified Robustness to Adversarial Word Substitutions In Findings of ACL 2023

- Authors: **Yibin Wang***, Yichen Yang*, Di He, Kun He
- In my research on adversarial word substitution in NLP, I have discovered the crucial impact of word embeddings on certified adversarial robustness in convex relaxation-based methods. Given the findings, I transformed the optimization problem of the model's certified robustness into an optimization problem of word embeddings through theoretical proofs. I independently completed all coding, experiments, and the main part of the paper writing.

Continual Learning of Large Language Models: A Comprehensive Survey Preprint

- Authors: Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, **Yibin Wang**, Hao Wang
- Responsible for writing the parts related to large language models in the Preliminaries section.

BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models Preprint, under review

- Authors: **Yibin Wang***, Haizhou Shi*, Ligong Han, Dimitris Metaxas, Hao Wang
- We introduce a principled Bayesian framework for improving large language models' generalization and uncertainty estimation. I contributed to the design of the algorithm and the writing of the paper, independently optimized the algorithm, implemented the code, and conducted the primary experiments.

Miscellaneous

- Languages: English - IELTS overall score 7.0