Yibin Wang

yibinwang@hust.edu.cn · (+86) 173-8956-0072 · https://yibinwang.netlify.app

Third-year undergraduate seeking a **Ph.D. Program in Trustworthy Artificial Intelligence, Adversarial Robustness, or Protecting Data Copyright and Privacy**.

Education

Huazhong University of Science and Technology (HUST), Hubei, China 2019 – Present

B.S. in Computer Science (CS), GPA: 3.78/4.00 - Transcripts

Experience

Adversarial Robustness Sep. 2021 – Present

Mentor: Kun He

- Conducted extensive research on adversarial attack and defense in machine learning
- Studied basic adversarial attack and defense methods in NLP and CV
- Conducted in-depth research on certified robustness based on convex relaxation
- Researched the adversarial robustness of the VAE model

Publications

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

- Authors: **Yibin Wang**[1], Yichen Yang[1], Di He, Kun He
- Status: Accepted
- 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 independently 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. Inspired by previous work on empirically robust word embeddings, our method boosts the certified adversarial robustness in NLP based on convex relaxation methods. It can be easily combined with IBP training and improves the certified robust accuracy from **76.73%** to **84.78%** on the IMDB dataset. Experiments demonstrate that our method outperforms various state-of-the-art certified defense baselines and generalizes well to unseen substitutions.

Skills

- Proficient in Python programming language, using Pytorch and Pytorch Lightning
- Familiar with machine learning algorithms and programming techniques (Achieved 99/100 score in Machine Learning Course)

Honors and Awards

Honerable Mention, Award on Mathematical Contest In Modeling May. 2022

Miscellaneous

- GitHub: https://github.com/flyleeee/
- Languages: English - Professional working proficiency, Mandarin - Native speaker

[1] The first two authors contribute equally.