Yibin Wang

yibinwang@hust.edu.cn · (+86) 173-8956-0072 ·

Third-year undergraduate seeking a **research-focused lab internship 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

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 (Accepted rate in 2022: 31.4%)

- Authors: **Yibin Wang**[1], Yichen Yang[1], Di He, Kun He
- Status: Accepted
- Abstract: Natural Language Processing (NLP) models are known to be vulnerable to adversarial examples typically crafted by synonym substitutions. In this paper, we find that word embedding is important to the certified robustness of NLP models. Given the findings, we propose the Embedding Interval Bound Constraint (EIBC) triplet loss to train robustness-aware word embeddings for better certified robustness. We optimize the EIBC triplet loss to make the verification boundary tighter with theoretical proof. Our method is conceptually simple and componentized. 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.
- I **independently** optimized the previous work of Yichen Yang based on the certified robustness, provided theoretical proof, and completed all coding and experiments.

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

- Academic Page: https://yibinwang.netlify.app/
- GitHub: https://github.com/flyleeee/
- Languages: English - Fluent, Mandarin - Native speaker

[1] The first two authors contribute equally.