MSU Trustworthy AI Group

Michigan State University

Yue Xing
Assistant Professor
Department of Statistics and Probability

Jiliang Tang
University Foundation Professor
Department of Computer Science and Engineering

Our Team

MSU Trustworthy AI Group

- Collaborators: amazon Meta

Recent research projects:

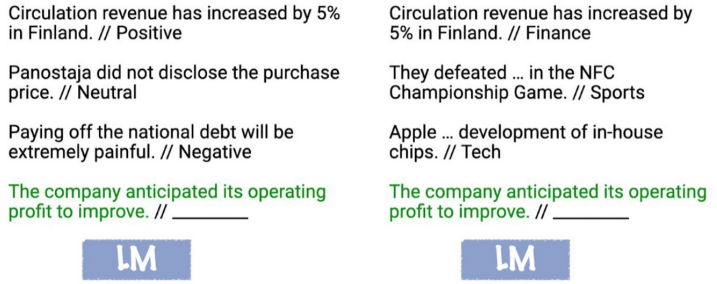
- Large language model (LLM):
  - Data set: HC-Var [XRH2023]
  - Data memorization [ZLR2023]
- Adversarial attack & defense:
  - Library: DeepRobust (Top-3 in Github) [LJX2021]
  - Poisoning attack [HXR2023a,b,XLW2023]
  - Adversarial training: methodology [HXR2022, XLL2021], theory [XSG2020, XRG2020,XSG2021,XSG2022a,b,c]
- Watermark: IP protection for pre-training (Diffusion-Shield) and fine-tuning (FT-Shield) [CRX2023,CRL2023]

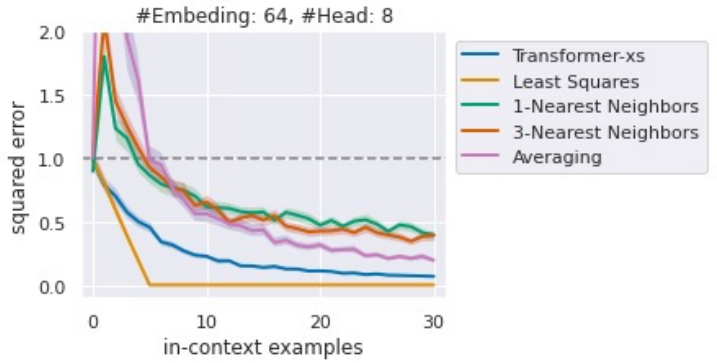
Hugging Face
hannxu / hc_var

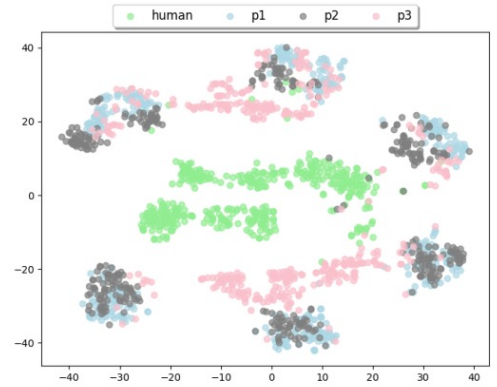
Deep Robust

Our Current Research Interests

- Properties of LLM:
  - Empirical & theoretical.
  - Understand how LLM works.
  - In-context learning (ICL).
  - Memorization.
  - Generated text detection.
- Attack and defense in LLM:
  - Poisoning attack in fine-tuning and ICL.
  - Jailbreaking attack & defense in ICL.



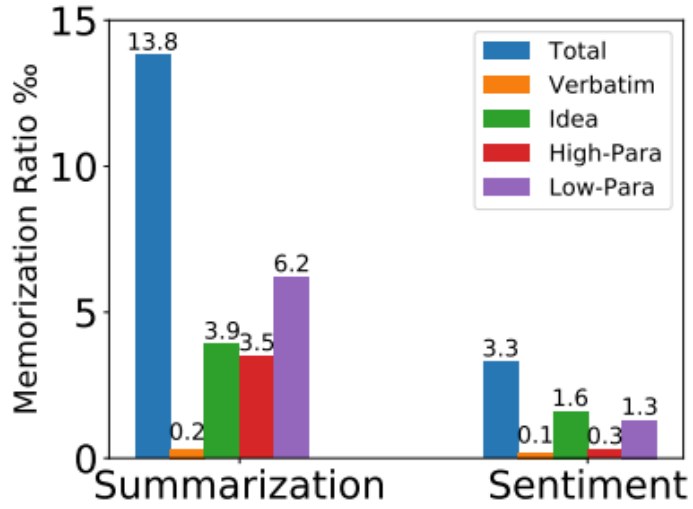


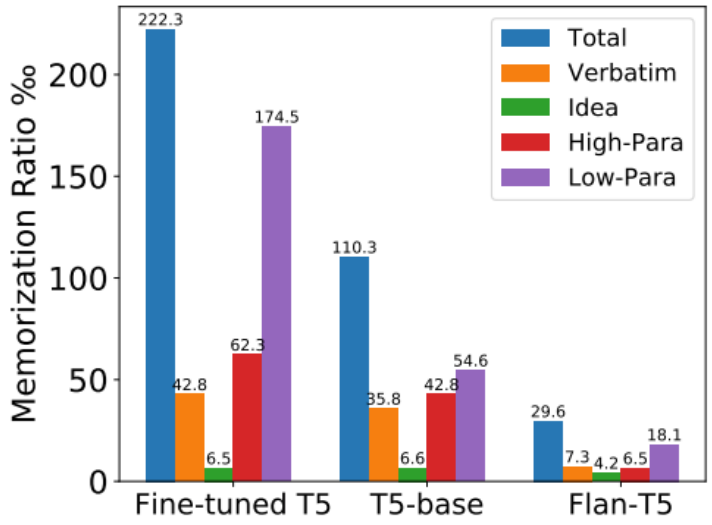


Our Recent Research Project

Memorization of LLM in fine-tuning

- Existing literature: LLM can reproduce pre-training data.
- Our observation:
  - LLM memorizes data in the fine-tuning stage.
  - Memorization level depends on the task-specific features.
  - Attention score can help quantify the vulnerability of memorization, and multi-task learning alleviates memorization.
  - Theory: the level of memorization can be connected to (1) the number of features, (2) sparsity of features.

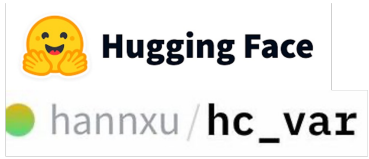




Our Team
MSU Trustworthy AI Group

Highlights

- Our projects are backboned by both theoretical guarantees and strong empirical improvements.
- We have strong expertise in attack & defense, and we are the first batch to study theories in LLM & ICL.
- Our team has contributed a lot to the open-source community with data sets and tools.



References

- [XHR2022] Pengfei He, Han Xu, Jie Ren, Yuxuan Wan, Zitao Liu, Jiliang Tang. Probabilistic Categorical Adversarial Attack & Adversarial Training. ICML 2023.
- [XLL 2021] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, Jiliang Tang, To be Robust or to be Fair: Towards Fairness in Adversarial Training. ICML 2021.
- [ZLR2023] Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Jiliang Tang, Dawei Yin. Exploring Memorization in Fine-Tuned Language Models.
- [XRH2023] Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, Jiliang Tang, On the Generalization of Training-based ChatGPT Detection Methods.
- [LJX2021] Yaxin Li, Wei Jin, Han Xu, Jiliang Tang (2021), Deeprobust: A pytorch library for adversarial attacks and defenses. AAAI2021.
- [HXR2023a] Pengfei He, Han Xu, Jie Ren, Yingqian Cui, Shenglai Zeng, Yue Xing, Jiliang Tang, Makoto Yamada, Mohammad Sabokrou, Confidence-driven Sampling for Backdoor Attacks.
- [HXR2023b] Pengfei He, Han Xu, Jie Ren, Yingqian Cui, Hui Liu, Charu Aggarwal, Jiliang Tang, Sharpness-Aware Data Poisoning Attacks.
- [CRX2023] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, Jiliang Tang (2023), DiffusionShield: A Watermark for Data Copyright Protection against Generative Diffusion Models
- [CRL2023] Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, Pengfei He, Yue Xing, Wenqi Fan, Hui Liu, Jiliang Tang (2023), FT-SHIELD: A Watermark Against Unauthorized Fine-tuning in Text-to-Image Diffusion Models.
- [XSG2020] Yue Xing, Qifan Song, Guang Cheng (2020), On the Generalization Properties of Adversarial Training. AISTATS 2021.
- [XRG2020] Yue Xing, Ruizhi Zhang, Guang Cheng (2020), Adversarially Robust Estimate and Risk Analysis in Linear Regression. AISTATS 2021.
- [XSG2021] Yue Xing, Qifan Song, Guang Cheng (2021), On the Algorithmic Stability of Adversarial Training. Neurips 2021.
- [XSG2022a] Yue Xing, Qifan Song, Guang Cheng (2022), Unlabelled Data Helps: Statistical Minimax Analysis and Adversarial Robustness. AISTATS 2022.
- [XSG2022b] Yue Xing, Qifan Song, Guang Cheng (2022), Phase Transition from Clean Training to Adversarial Training. Neurips 2022.
- [XSG2022c] Yue Xing, Qifan Song, Guang Cheng (2022). Why Do Artificially Generated Data Help Adversarial Robustness. Neurips 2022.