BENGAL Proposers' Day Presentation

Lin Tan & Xiangyu Zhang

Introduction

- Our group
  - Two professors
  - Two post-docs and 20+ PhD students
- Our unique expertise related to the program
  - High-quality synthetic training data generation
  - Bias measurement and mitigation for neural networks
  - Deep learning model biases, vulnerabilities, and hardening
  - LLM attacks, vulnerabilities, and mitigation strategies
  - Information flow tracking
- Relevant project experience
  - IARPA TrojAI, DARPA VSPELLS, DARPA Transparent Computing, DARPA Binary Executable Transformation, ONR TPCP, ONR Learn-2-Reason, ONR RHIMES, …

Our Expertise in LLM Biases & Vulnerabilities (1)

- Top IARPA TrojAI contests,and NeurIPS'23 TDC contest
  - TrojAI aims to find backdoors in DL models, including NLP ones
  - TDC'23 aims to jailbreak LLMs and find backdoors in LLMs
- **AI vulnerabilities with real-world consequences:** 20 vulnerabilities in ChatGPT plugins with one CVE assigned, leading to cross-site-scripting, permission escalation, and mis-information
- Working resiliently with poisoned sources:
  - PICCOLO: Exposing complex backdoors in NLP transformer models (S&P'22)
  - Backdoor vulnerabilities in normally trained deep learning models (arXiv preprint arXiv:2211.15929, 2022)
  - Constrained optimization with dynamic bound-scaling for effective NLP backdoor defense (ICML'22)
  - Model orthogonalization: Class distance hardening in neural networks for better security (S&P'22)
  - PARAFUZZ: An interpretability-driven technique for detecting attacks in NLP (NeurIPS'23)
- Model information flow and interpretability
  - BEAGLE: Forensics of Deep Learning Backdoor Attack for Better Defense (NDSS'23)
  - MIRROR: Model Inversion for Deep Learning Network with High Fidelity (NDSS'22)
  - Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples (NeurIPS'18)

Our Expertise in LLM Biases & Vulnerabilities (2)

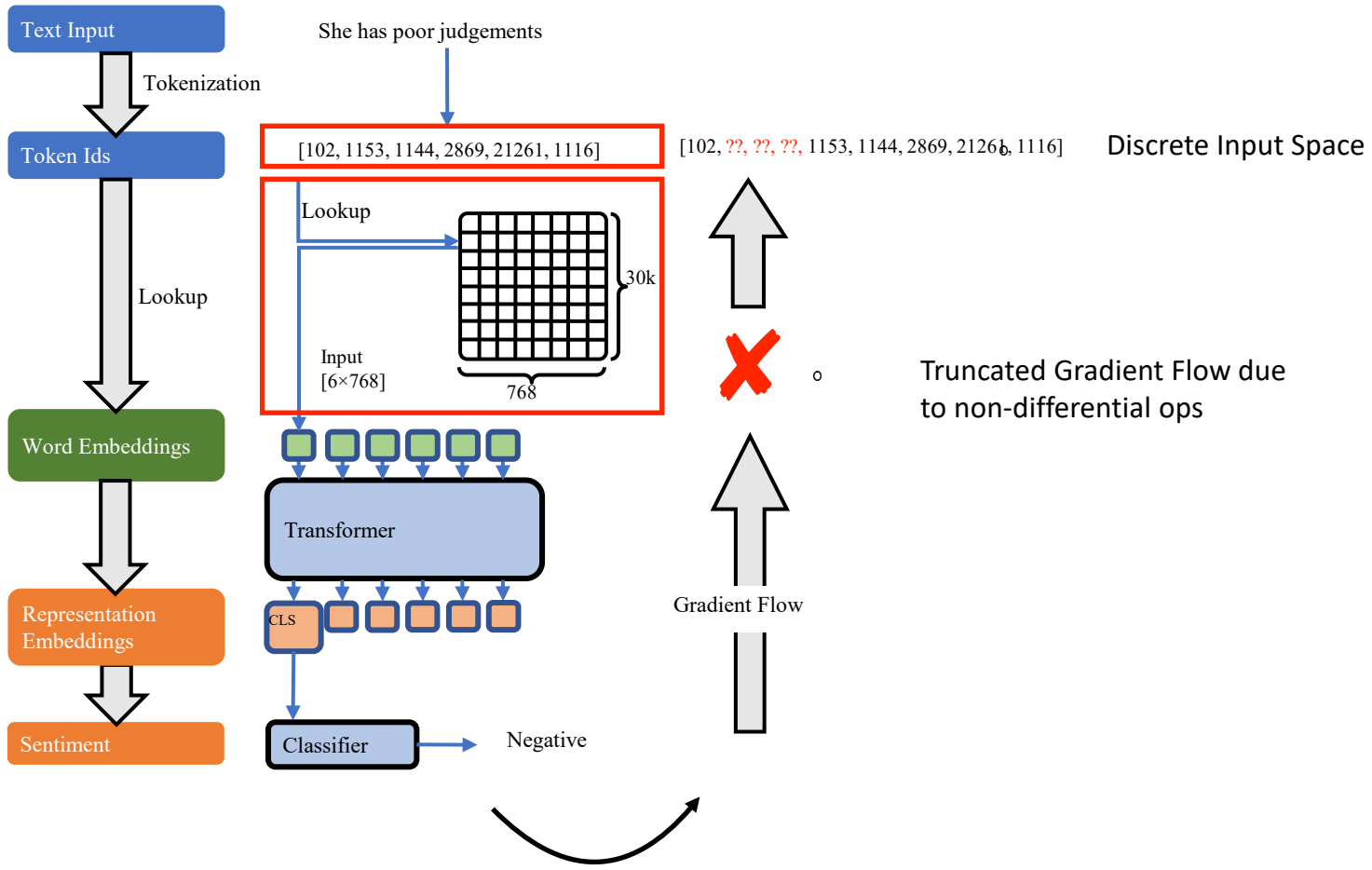
- High-quality synthetic data generation
  - New data-free training data generator
  - New model attack/stealing/reverse-engineering
  - (**AAAI'23**) ***Oral Presentation!*** https://github.com/lin-tan/disguide

- Bias Measurement and Mitigation for Neural Networks
  - Unfair Models (NeurIPS'21) https://github.com/lin-tan/fairness-variance
  - Inaccurate and Inefficient Models (ASE'21) - **ACM SIGSOFT Distinguished Paper Award**! https://github.com/lin-tan/dl-variance

- ***Code and data released and used by many institutions***

Preparation for BENGAL: Prior Work PICCOLO

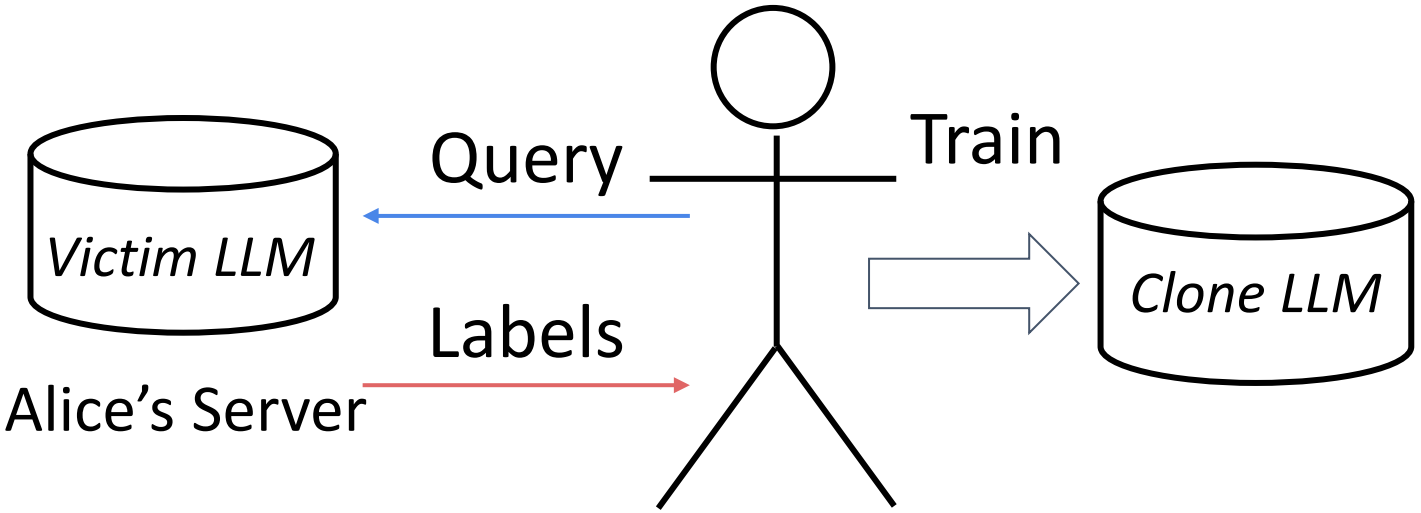
- PICCOLO: Exposing complex backdoor vulnerabilities in NLP transformer models (S&P'22)
  - PICCOLO reverse engineers prompts that can produce specific outputs such as jailbreaking outputs
  - It is the main technique we used in TrojAI NLP rounds, TDC LLM contests, and ChatGPT vulnerability scanning

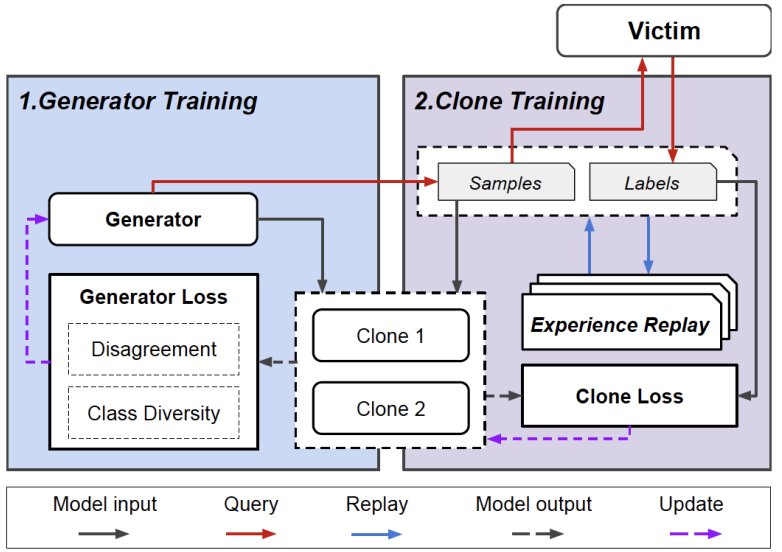


Preparation for BENGAL: Prior Work DisGUIDE

Our prior techniques on identifying new LLM threats and vulnerabilities

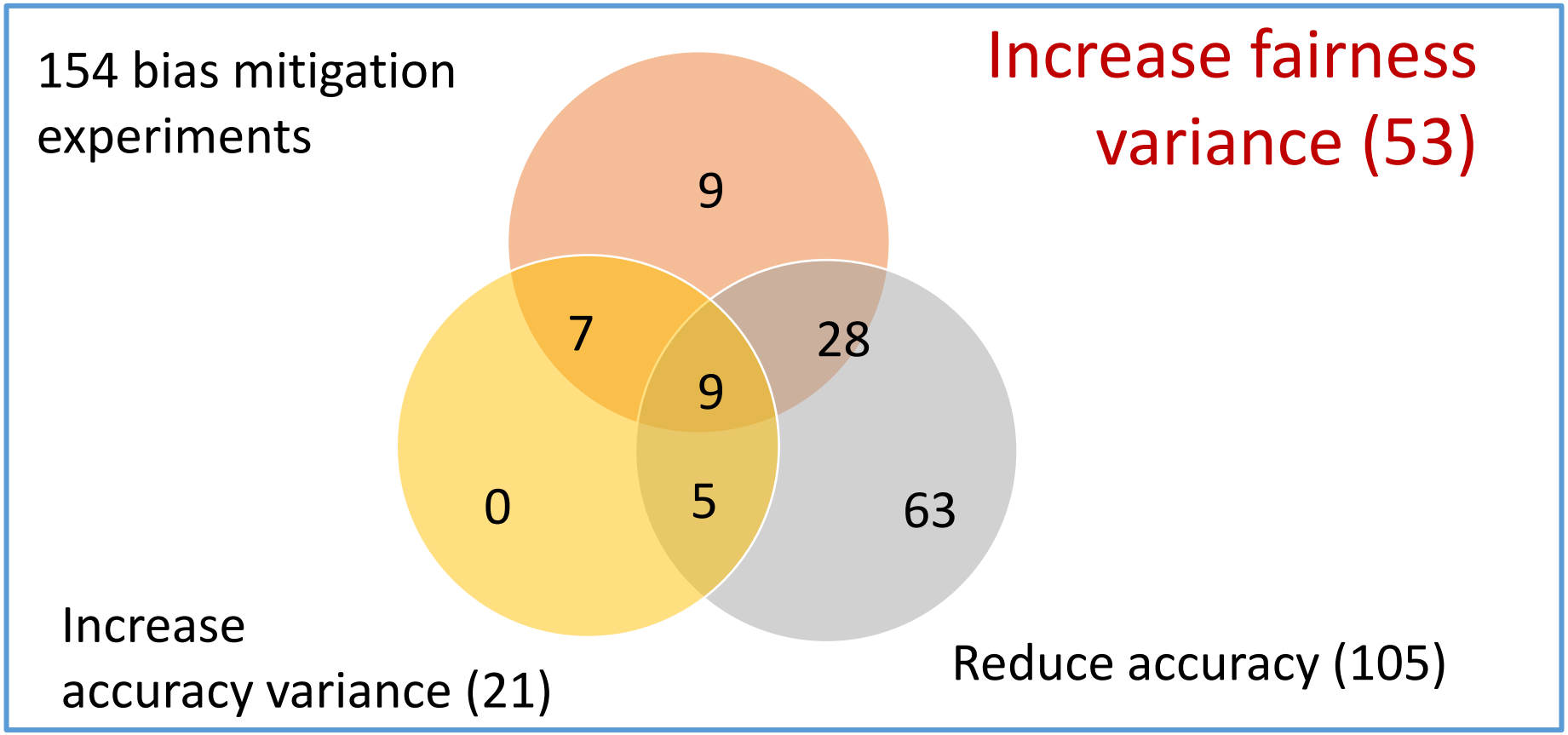
- DisGUIDE generates **high-quality synthetic images** without any prior data.
- DisGUIDE improves accuracy by up to **3.42%** and **18.48%.**
- DisGUIDE improves query efficiency by up to **64.95%**.





DisGUIDE: Disagreement-Guided Data-Free Model Extraction. **(Oral Presentation) AAAI 2023** Jonathan Rosenthal, Eric Enouen, Hung Viet Pham, and Lin Tan

Preparation for BENGAL: **Hidden** cost of debiasing



Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training. **NeurIPS 2021**
S. Qian, H. Pham, T. Lutellier, Z. Hu, J. Kim, L. Tan, Y. Yu, J. Chen, and S. Shah

Preparation for BENGAL: Promising Preliminary Results

- We have further
  - improved synthetic training data generation, and
  - Improved model stealing accuracy
  - …
- Our approach outperforms the state of the art (including our own prior work DisGUIDE)