SaTC: CORE: When Adversarial Learning Meets Differential Privacy: Theoretical Foundation and Applications

NhatHai Phan (NJIT) and My T. Thai (UF)

Motivations: Existing learning algorithms have not been designed to be simultaneously robust to such privacy and integrity attacks, in both theory and practice.

Goals: This project aims to develop the first framework, called DeepRobust, to advance and seamlessly integrate key learning and inferring techniques, including adversarial learning, differential privacy, and provable robustness, offering tight and reliable robustness bounds against both privacy and integrity attacks, while retaining high model utility in deep neural networks.

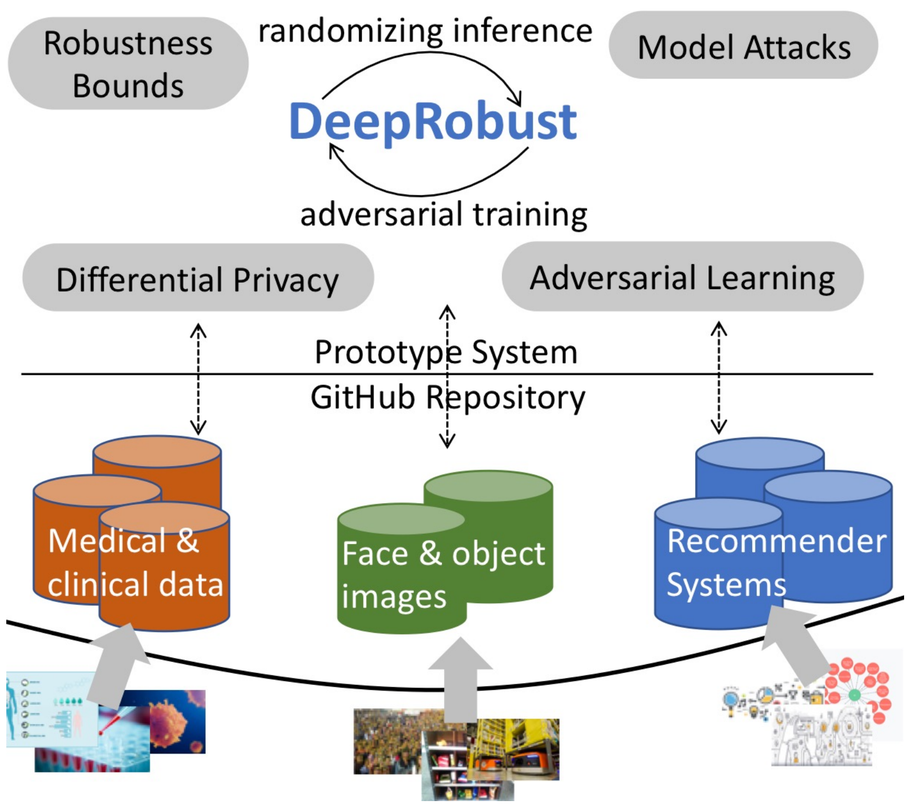
Challenges:

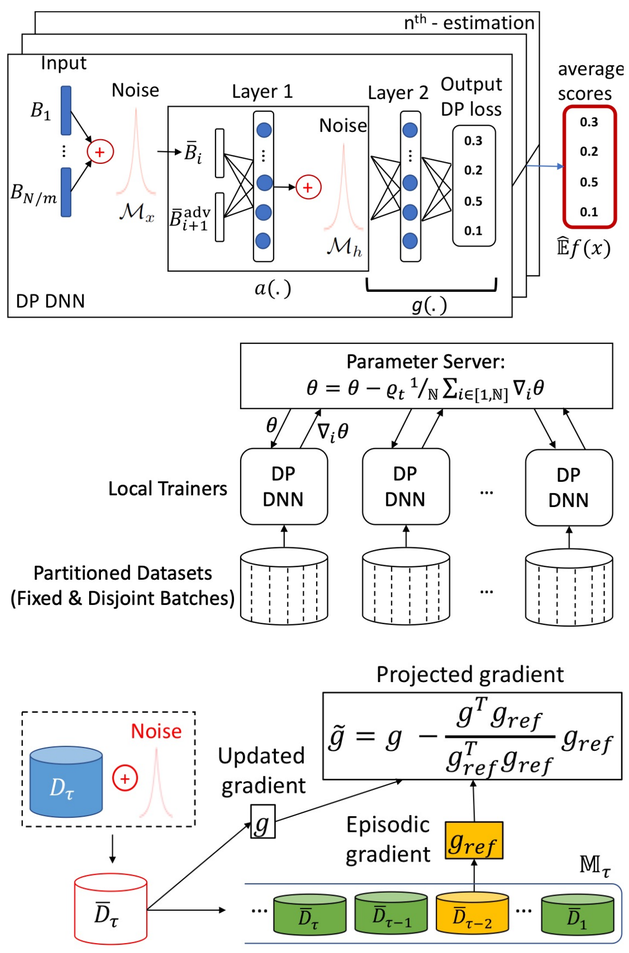
- Unprotected vulnerabilities in the highly correlated latent space
- Certified robustness to ***both*** privacy and integrity attacks with high model utility

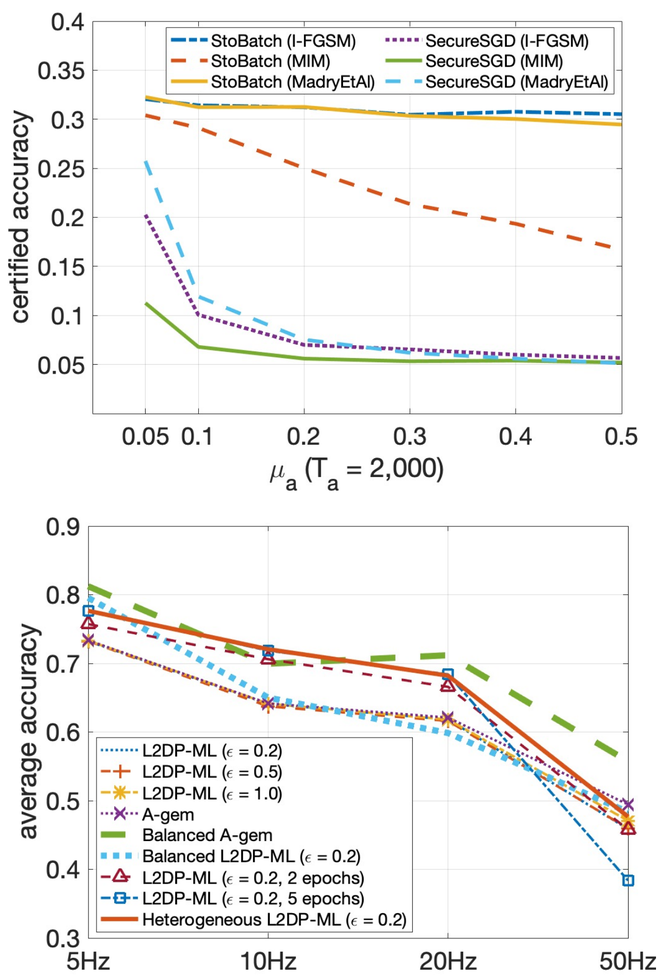
Contributions: Heterogeneous GM [IJCAI'19], StoBatch [ICML'20], AdvTroj [IEEE BigData'21], Lifelong DP [ICONIP'21, CoLLAs'22], etc.

Scientific Impacts:

- Uncover unknown correlations between differential privacy, adversarial learning, and certified robustness
- Novel model attacks exploit the hidden space
- Model utility, privacy loss, and robustness bounds







Broader Impact:

- Enable safe, effective, efficient, and deep analyses of rich and diverse user-generated data
- Technologies transfer: crucial applications in which both privacy and robustness are significant problems

Education and Outreach:

- Research and teaching integration
- Trustworthy AI course
- 1st International Trustworthy FL Workshop at IEEE ICDM'22
- Women in IoT (WiT) workshop

Broader Participation:

- Course projects
- Women and underserved students
- National Center For Women & Information Technology (NCWIT) Collegiate Award Finalist 2022
- Silver Medal: 2022 Dana Knox Student Research Showcase