When Adversarial Learning Meets Differential Privacy: Theoretical Foundation and Applications

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

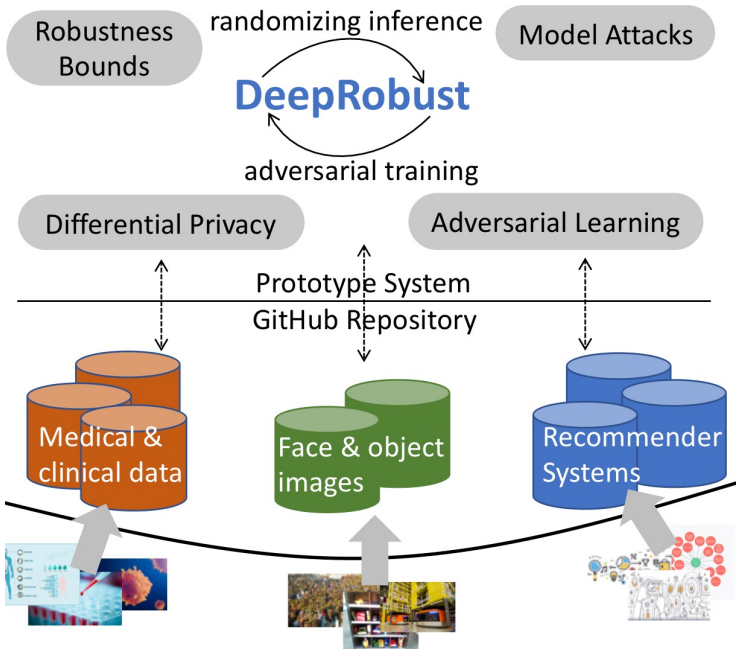
Challenge:

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

Solution:

- Exploit vulnerabilities based on hidden and correlated features
- Differential privacy in adversarial learning
- Distances in the latent space will be protected under DP
- Customized DP noise

NSF CNS-1935928 / 1935923. New Jersey Institute of Technology and the University of Florida



DeepRobust framework which protects privacy of the training data and is robust to model attacks along with theoretical guarantee bounds. DeepRobust will be evaluated in three main application domains via real-world data and industry partners

Scientific Impact:

- Foundations to uncover unknown correlations between DP, adversarial learning, and certified robustness
- Novel model attacks exploit the hidden space
- Model utility, the privacy loss, and the robustness bounds

Broader Impact and Broader Participation:

- 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
- Trustworthy AI course and outreach activities
- Women and underserved students