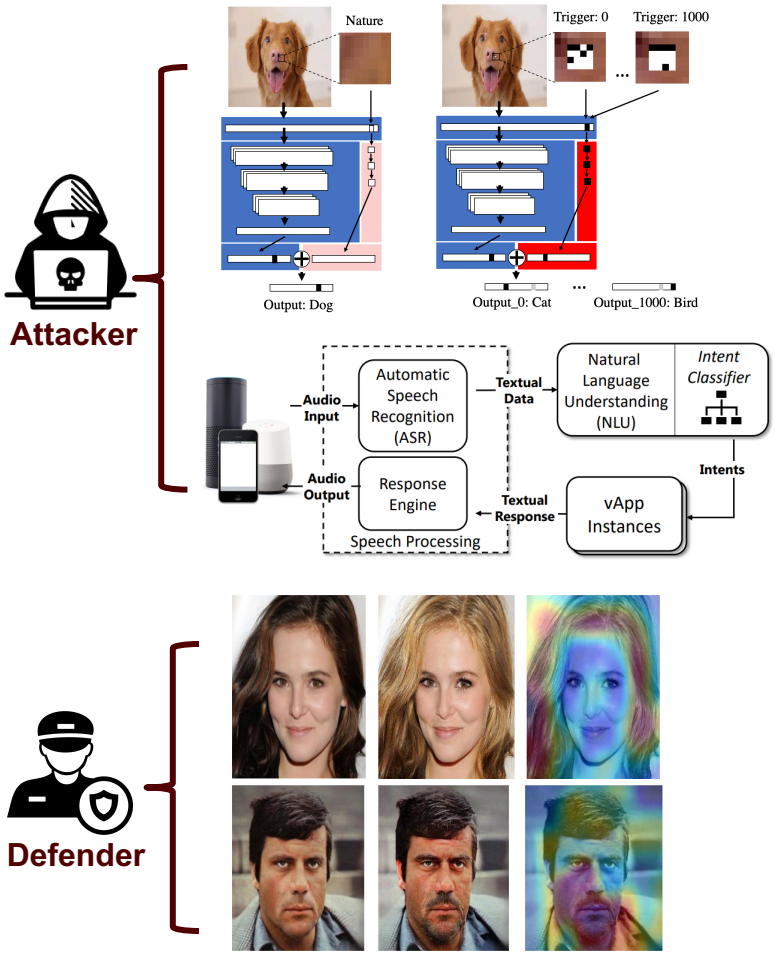
Adversarial Learning via Modeling Interpretation

Motivation & Objectives:

- AI models are vulnerable to adversaries attacking.
- Developing an adversarial learning framework that leverages interpretability of ML models and results to both identify and mitigate the risks of adversarial attacks.

Solution:

- Investigated a training-free trojan attack framework [1].
- Designed a linguistic-model guided fuzzing too [2].
- Proposed an active learning framework for deepfake detection [3].



Deepfake Detection

Scientific Impact:

- Develop effective attacking strategies by analyzing modeling interpretation.
- Explore defensive strategies to improve the robustness of ML models.
- Deal with heterogeneous, large-scale, and relational dataset.

Social Impact:

- Reveals the severe security problem widely existing in various AI models.
- Encourages the community to put more effort into promoting the robustness of AI models.

PIs: Dr. Xia Hu[1]
Dr. Guofei Gu[2]
Dr. James Caverlee[2]
[1]*Rice University*
[2]*Texas A&M University*

[1] Tang, Ruixiang, et al. "An embarrassingly simple approach for trojan attack in deep neural networks." *SIGKDD* 2020.
[2] Zhang, Yangyong, et al. " Fuzzing Semantic Misinterpretation for Voice Assistant Applications." *NDSS* 2019.
[3] Du, Mengnan, et al. "Towards generalizable deepfake detection with locality-aware autoencoder." *CIKM* 2020.