Targeted Poisoning Attacks in Coalition Systems Under Incomplete Model Information

Supriyo Chakraborty, Seraphin Calo (IBM US), **Moustafa Alzantot, Mani Srivastava** (UCLA),
Arjun Bhagoji (Princeton), **Yash Sharma** (Cooper Union), **Richard Tomsett** (IBM UK), **Kevin Chan** (ARL)

Relevance to Coalition

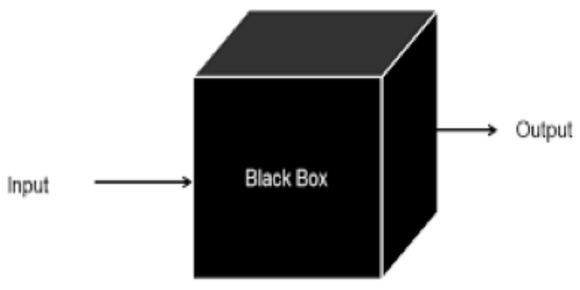
- Members often have sensitive data which they cannot share but are willing to share insights learned using the data.
- Distributed Learning paradigm allows model training without members requiring to share data. Only model parameters are shared with the centralized server.
- At inference time, trained models are used to classify data for decision making.

Attacks Under Incomplete Information

- *Training-time poisoning attack*: adversary poisons model parameters that it shares with the server. Does not know the updates from other agents (incomplete parameter space).
- *Inference-time poisoning attack*: adversary poisons the test sample to cause targeted misclassification.

GenAttack: Gradient-Free Black Box Attack

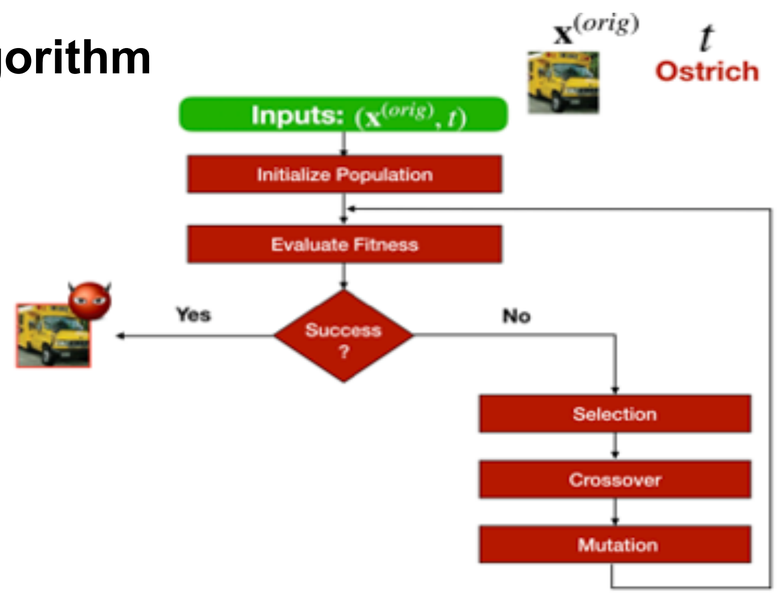
Attacker Model



- Adversary has no access to model parameters (can't compute gradient).
- Adversary only has access to model query results (distribution over all classes)

$$f : \mathbb{R}^d \rightarrow [0, 1]^K$$

Algorithm



- Genetic algorithm based *gradient-free* strategy for generating adversarial examples.
- Mutation rate, crossover probability and population size are used to minimize the number of queries.

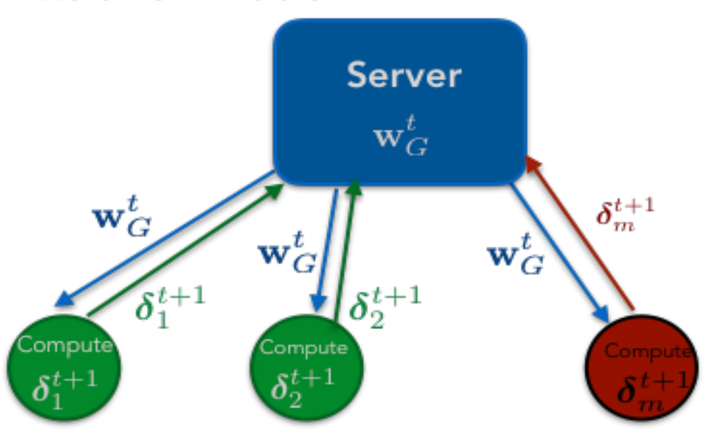
Performance against SOTA defenses

	CIFAR-10		ImageNet	
	ASR	# Queries	ASR	# Queries
Bit depth reduction	93%	2,796	95%	16,301
JPEG	88%	3,541	89%	23,822

- GenAttack maintains a high attack success rate (ASR) while being query efficient (compared to ZOO, NES).
- High attack against Ensemble adversarial training and randomized transformation based defenses.

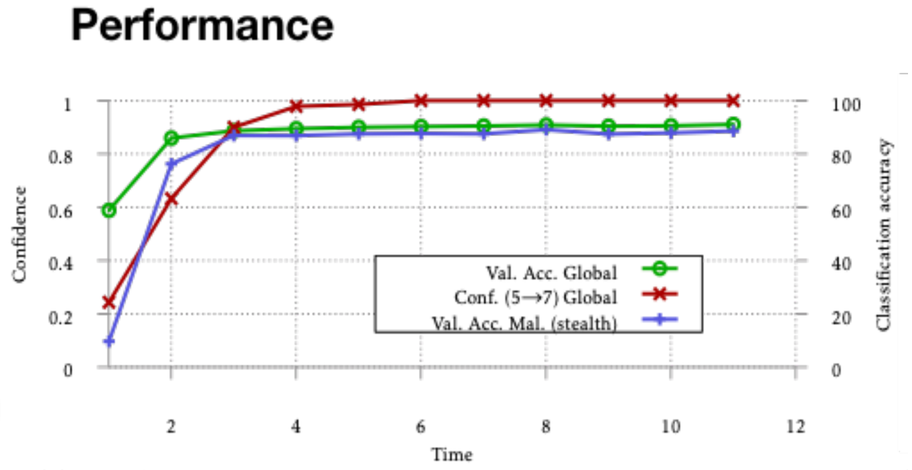
Model Poisoning: On Federated Learning Systems

Attacker Model



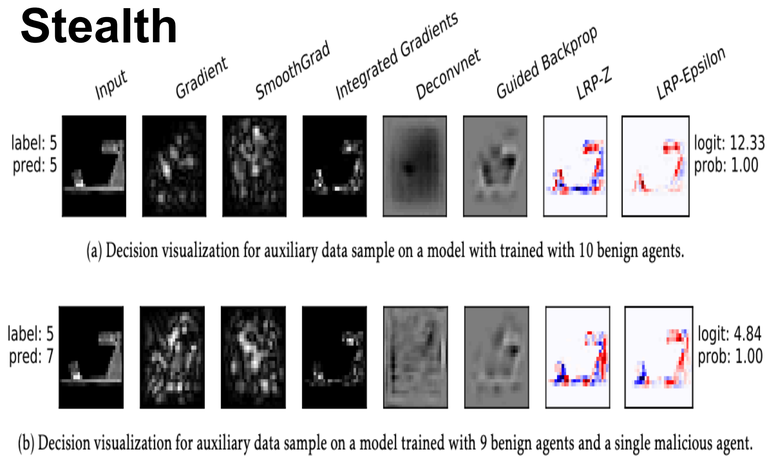
- Federated Learning with a single malicious agent.
- *Adversarial objective*: elicit targeted misclassification on a set of samples.
- *Benign objective*: ensure global model has good performance.

Performance



- Alternately minimize adversarial and benign objectives while boosting update to counter benign agents.
- Attack achieves high success rate and the global model continues to converge to a good minima.

Stealth



(a) Decision visualization for auxiliary data sample on a model with trained with 10 benign agents.

(b) Decision visualization for auxiliary data sample on a model trained with 9 benign agents and a single malicious agent.

- Malicious model updates are constrained to be as similar as possible to benign updates.
- Current interpretability techniques cannot distinguish between benign and malicious global models.

Summary and Future Work

- *Scalability of poisoning attacks in coalition settings with colluding malicious agents in both centralized and peer-to-peer systems.*
- *Explore gradient-free attack strategies for even more limited information settings (e.g., hard label only for top class is available).*
- *Designing detectors to identify poisoning and evasion attacks – leveraging model interpretability/multi-modal data for identifying malicious activation patterns in neural networks.*

Publication(s)

- *"GenAttack: Practical Black Box Attacks with Gradient-Free Optimization," M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C. Hsieh, M. Srivastava, in GECCO, 2019.*
- *"Analyzing Federated Learning Through an Adversarial Lens," A. Bhagoji, S. Chakraborty, P. Mittal, S. Calo, in ICML, 2019.*
- "*Model Poisoning Attacks Against Distributed Machine Learning Systems*," R. Tomsett, K. Chan, S. Chakraborty, in SPIE, 2019.