Certified Defenses for Data Poisoning Attacks

Speaker: Berkay Kaplan

Problem Statement

- Machine learning systems trained on user-provided data are susceptible to data poisoning attacks (User accounts)
- In data poisoning attacks, malicious users inject false training data with the aim of corrupting the learned model
- Little is understood about the worst-case loss of a defense in the face of a determined attacker (Upper Bound of the Loss Function)

Related Work

- Szegedy et al. discovered that adversarial test images can fool image classifiers despite being imperceptible from normal images [1]
  - These images exhibit vulnerabilities at test time, whereas data poisoning is a vulnerability at training time.
- A common defense against adversarial test examples is adversarial training, which alters the training objective to encourage robustness [2]

Framework Model

- The researchers address DP by constructing approximate upper bounds on the loss on attacks
- A framework is created to study the entire space of attacks against a given defense using outlier detectors
- Empirically, the project finds that even under a simple defense, the MNIST-1-7 and Dogfish datasets are resilient to attack
- In contrast, the IMDB sentiment dataset can be driven from 12% to 23% test error by adding only 3% poisoned data.

Remove Outliers: Categories

- Fixed defenses: does not rely on the poisoned data
  - Example: let the defense be documents that contain only licensed words
  - Slab and sphere defense

- Data-dependent defenses: Estimates the centroid of the poisoned data
  - The attacker can choose the poisoned data to change manipulate the defense

- Assumption from paper: removing outliers does not change the distribution

Fixed Defenses: Computing the Minimax Loss via Online Learning

- Fixed Defense: oracle defense that knows true centroids
- Compute the minimax loss function M
- Θ (model) is a ball with radius p
- In each iteration of minimax, find the worst attack point (x(t) , y(t)) with respect to the current model θ(t−1)
- Update the model in the direction of the attack point, producing θ(t) . The DP attack is the set of points thus found
- In the process, the iterations form a candidate DP attack whose induced loss $\frac{1}{n} L(\tilde{\theta}; \mathcal{D}_c \cup \mathcal{D}_p)$ is a lower bound on the U(θ) of M

$$\mathbf{M} \leq \min_{\theta \in \Theta} \max_{\mathcal{D}_p \subseteq \mathcal{F}} \frac{1}{n} L(\theta; \mathcal{D}_c \cup \mathcal{D}_p) = \min_{\theta \in \Theta} U(\theta), \text{ where } U(\theta) \stackrel{\text{def}}{=} \frac{1}{n} L(\theta; \mathcal{D}_c) + \epsilon \max_{(x,y) \in \mathcal{F}} \ell(\theta; x, y).$$

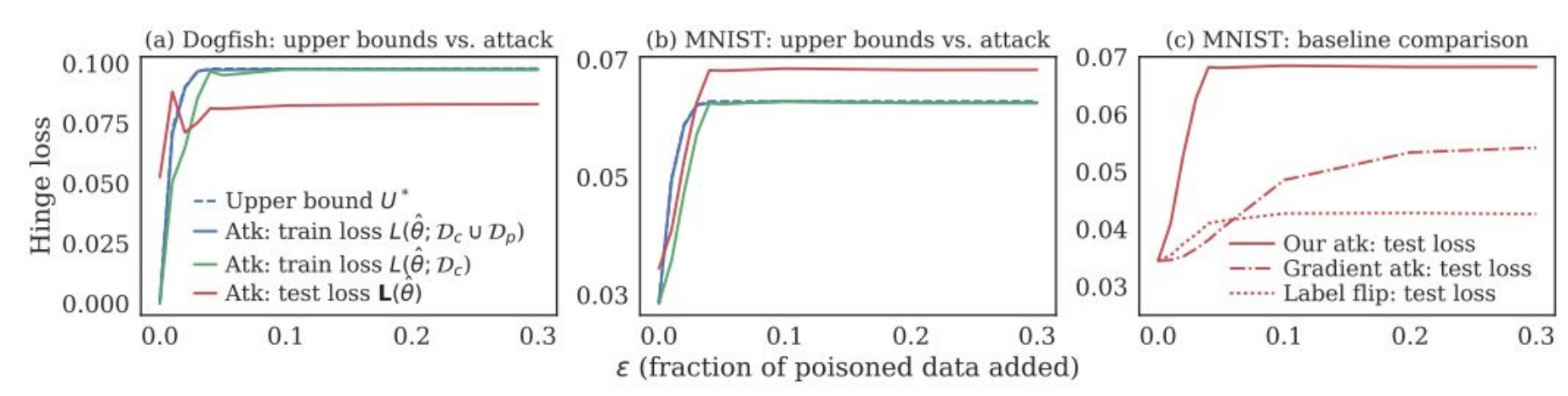
Epsilon is [0, 1] and parametrizes the attacker's resources

Data-Dependent Defenses: Upper and Lower Bounds

- It is no longer the case that the optimal DP attack places all points at a single location, due to the dependence of F on the poisoned data
- Run the previous algorithm with a few additions
- At each iteration, obtain a distribution $\pi_p(t)$ and upper bound $U(\theta(t))$
- Then, for each $\pi_p(t)$, we will generate a candidate attack by sampling n points from $\pi_p(t)$, and take the best resulting attack
- Despite a lack of rigorous theoretical guarantees, this often leads to good upper bounds and attacks in practice in experiments

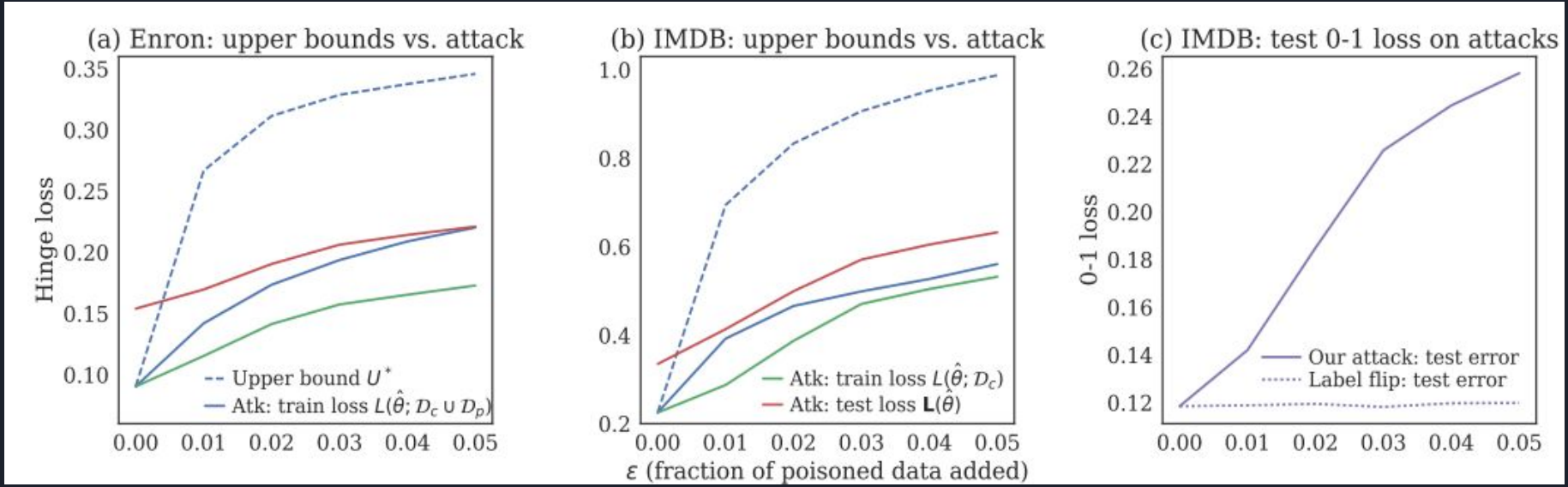
Evaluation of Fixed Defenses (MNIST: resilient)

- Defenses: Slab and sphere defense

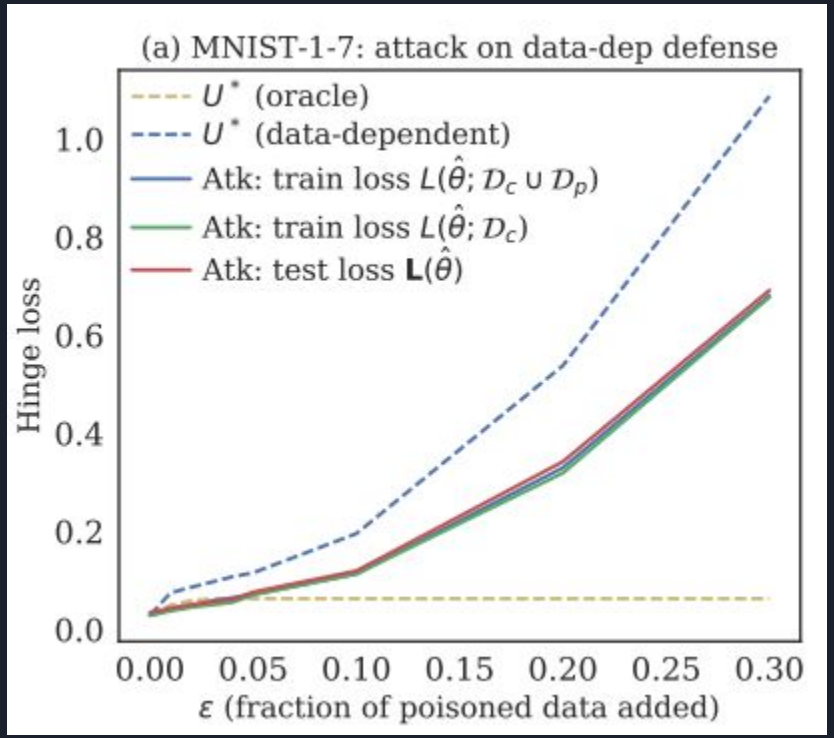


Evaluation of Fixed Defenses (IMDB: vulnerable)

- Defenses: Slab and sphere defense



Evaluation (MNIST with data dep: vulnerable)



Conclusion and Remarks

- The researchers presented a tool for studying data poisoning defenses that goes beyond empirical validation
- Pros
  - A framework to evaluate a defense against every attack would be very feasible in real-world
  - Using popular and rich datasets, such as MNIST, will help reliability
- Cons
  - However, the evaluation phase only relies on a few datasets, negatively impacting reliability
  - Related work needs more effort

Questions

References

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[2] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).