CSEP564 : Computer Security : Reading 7

Karuna Sagar Krishna

November 13, 2024

Paper Title

Poisoning Web-Scale Training Datasets is Practical

Paper Authors

Nicholas Carlini, Matthew Jagielski, Chistopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, Florian Tramer

Problem

Deep learning models require large training datasets and getting such well curated dataset is infeasible. So web scale datasets are constructed by crawling the Internet. However, this implies datasets are not curated impacting quality and security. Although, there is vast literature that explores poisoning attacks, most of the them assume that poisoning is practical. This paper focuses on "how" poisoning attacks can be carried out.

Approach

The paper provides 2 novel attacks that guarantee malicious examples will appear in web scale datasets. The authors lay out the threat model and its assumptions and carefully considered ethical implications.

The first attack is called "split view poisoning". In short, the adversary targets distributed dataset which have a "split brain" problem. The paper leverages the residual trust of domains making up the dataset which is already well known vulnerability. The authors target 10 recent datasets; for each they calculate the number of expired and buyable domains and if these datasets are actively downloaded for model training. The authors use DNS responses and domain registrar to find the state of domains. The authors also show datasets are being actively downloaded for training purposes and that datasets are vulnerable from day zero. It costs less than $60 to control 0.01% of the dataset. Attack simulations show that with 1000 poisoned samples, a model can be trained to misclassify with high success rate. Fortunately, the authors have not found any evidence of this attack carried out in the wild. Finally, the authors propose prevent this attack by including cryptographic or perceptual hash.

The second attack is called "frontrunning poisoning". This attack targets centralized dataset. Wikipedia forbids crawling and instead provides dumps at predictable times. So the adversary needs to predict when the snapshot would be taken and poison in this interval before moderators revert these malicious changes. The authors show that it is quite easy to predict when an article would be snapshot with reasonable accuracy by studying data from previous dumps and published statistics. The authors plot a timeline of various article edits, classify which edits are included in snapshot using which they extract valuable information

to predict the next snapshot time for an article due to consistency in the process. They calculate the maximum number of articles that could be poisoned. Though they note that this high poisoning rate is not practical; this is much higher than 0.001% rate established by previous work. Two solution approaches are proposed - randomize the snapshot process and use lock/freeze mechanism.

Conclusions

The paper builds on previous work and showcase 2 attacks against web scale datasets which are both low cost and extremely practical. The authors have disclosed their findings responsibly to stakeholders and have remained ethical in their experiments. Given the importance and increasing reliance on AI/ML models, it is critical for the community to reassess their trust assumptions on we scale datasets.

New Ideas

The authors propose dataset transparency as a possible solution to prevent poisoning in general. As noted this is similar to certificate transparency solution adopted by the community in recent years and my understanding is that this solution has proved to be effective. So, dataset transparency is a new idea and has potential to be successful in wild.

To me it was nice to see that the authors paid careful attention to ethical impact of their experiments and investigation. Though this might be a common place in security research community, I found that this responsible and mature decision/idea.

Improvements

The paper focused their frontrunning attack on Wikipedia dump and barely mentioned Common Crawl dataset. The paper could be improved by considering other datasets to establish if "frontrunning" is a general or targeted attack.

The paper simulates split view attack and found high success rate in misclassification. However, there was no such simulation done for frontrunning attack, instead the paper simply argues that this attack is practical. The authors could have simulated the entire Wikipedia snapshot process and show empirical results to support their claims.

New Directions

Building resilient training algorithms and/or datasets are clearly important given the increasing reliance on AI/ML models in day-to-day life. This is akin to identifying and preventing fake news on Internet, and we have clearly seen the impact of fake news in recent years. So, it would be a valuable research direction.

Given that there are various language models being developed and used, is there a way to identify poisoning impacting these models? In other words, a new direction of research could be about how can we identify poisoning in general and what is an acceptable threshold before releasing and consuming datasets to wider public. This is effectively the release criteria or quality assurance for these models.