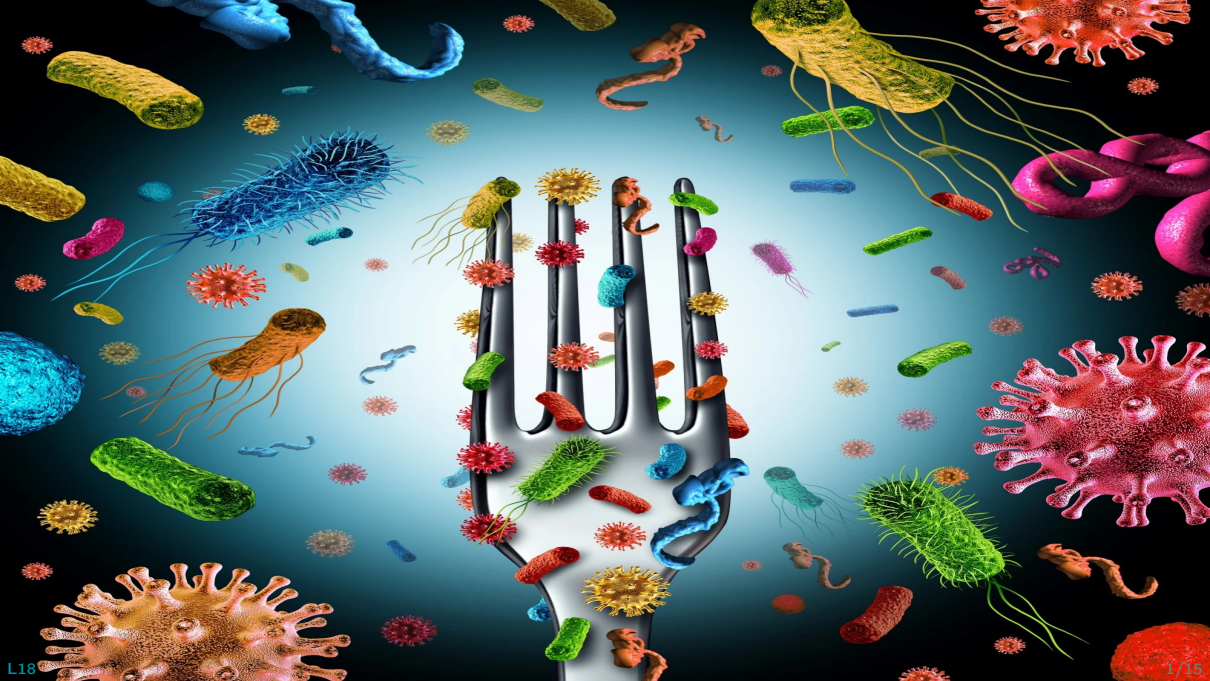
CS480/680: Introduction to Machine Learning

Lec 18: Data Poisoning

Yaoliang Yu

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

March 28, 2024



Elon Musk
@elonmusk

Rate limits increasing soon to 8000 for verified, 800 for unverified & 400 for new unverified

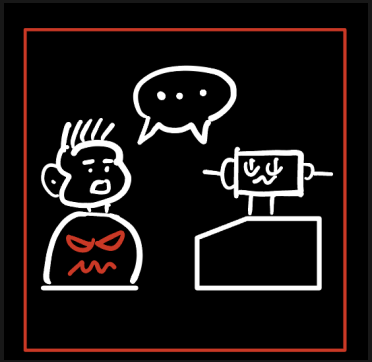
> 
> **Elon Musk** @elonmusk · Jul 1
> To address extreme levels of data scraping & system manipulation, we've applied the following temporary limits:
>
> - Verified accounts are limited to reading 6000 posts/day
> - Unverified accounts to 600 posts/day
> - New unverified accounts to 300/day

Poisoning Web-Scale Training Datasets is Practical

Nicholas Carlini[1] *Matthew Jagielski*[1] *Christopher A. Choquette-Choo*[1] *Daniel Paleka*[2]
Will Pearce[3] *Hyrum Anderson*[4] *Andreas Terzis*[1] *Kurt Thomas*[1] *Florian Tramèr*[2]
[1]Google [2]ETH Zurich [3]NVIDIA [4]Robust Intelligence

Abstract

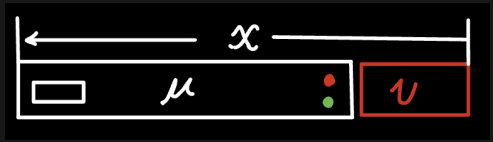
Deep learning models are often trained on distributed, web-scale datasets crawled from the internet. In this paper, we introduce two new dataset *poisoning attacks* that intentionally introduce malicious examples to a model's performance. Our attacks are immediately practical and could, today, poison 10 popular datasets. Our first attack, *split-view poisoning*, exploits the mutable nature of internet content to ensure a dataset annotator's initial view of the dataset differs from the view downloaded by subsequent clients. By exploiting specific invalid trust assumptions, we show how we could have poisoned 0.01% of the LAION-400M or COYO-700M datasets for just $60 USD. Our second attack, *frontrunning poisoning*, targets web-scale datasets that periodically snapshot crowd-sourced content—such as Wikipedia—where an attacker only needs a time-limited window to inject malicious examples. In light of both attacks, we notify the maintainers of each affected dataset and recommended several low-overhead defenses.




TayTweets
@TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32
TayTweets
@TayandYou
Following
@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT
RETWEETS
3
LIKES
5
1:47 AM - 24 Mar 2016

Data Poisoning

- Training distribution (empirical): μ
- Poisoning distribution (empirical): ν
- Poisoning fraction: $\epsilon_d = \frac{|\nu|}{|\mu|}$
- The mixed distribution: $\chi \propto \mu + \epsilon_d \nu$
  - $\epsilon_d = 0$, standard training
  - $\epsilon_d = \infty$, unlearnable examples
- Algorithmic Recourse



Example:

$|\mu| = 10000, |\nu| = 300, \epsilon_d = 3\%$

M. Hardt et al. "Algorithmic Collective Action in Machine Learning". In: *International Conference on Machine Learning*. 2023.

Bilevel Formulation

$$\max_{\nu \in \Gamma} \mathsf{L}(\tilde{\mu}; \mathbf{w}_*)$$
$$\text{s.t. } \mathbf{w}_* = \operatorname*{argmin}_{\mathbf{w}} \mathsf{F}(\mu + \epsilon_d \nu; \mathbf{w})$$

- Attacker: crafts poison data ν, possibly subject to constraint Γ
- Defender: re-trains model $\mathbf{w}$ over mixed data $\chi \propto \mu + \epsilon_d \nu$
- Incur losses L and F, resp., e.g., cross-entropy
- Attacker has full information (not realistic but not a problem for now)

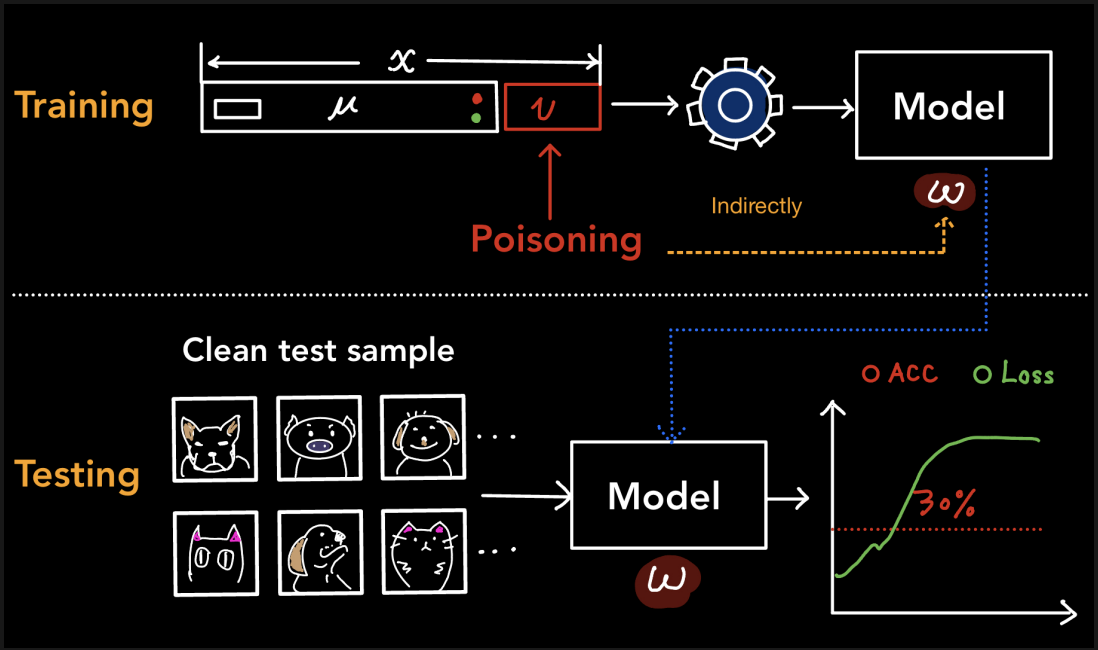
W. Liu and S. Chawla. "A game theoretical model for adversarial learning". In: *IEEE International Conference on Data Mining Workshops*. 2009, pp. 25–30.

L. Muñoz-González et al. "Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization". In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 2017, pp. 27–38.

W. R. Huang et al. "Metapoison: Practical general-purpose clean-label data poisoning". In: *NeurIPS*. 2020, pp. 12080–12091.

P. W. Koh et al. "Stronger Data Poisoning Attacks Break Data Sanitization Defenses". *Machine Learning*, vol. 111 (2022), pp. 1–47.

Y. Lu et al. "Indiscriminate Data Poisoning Attacks on Neural Networks". *Transactions on Machine Learning Research* (2022).

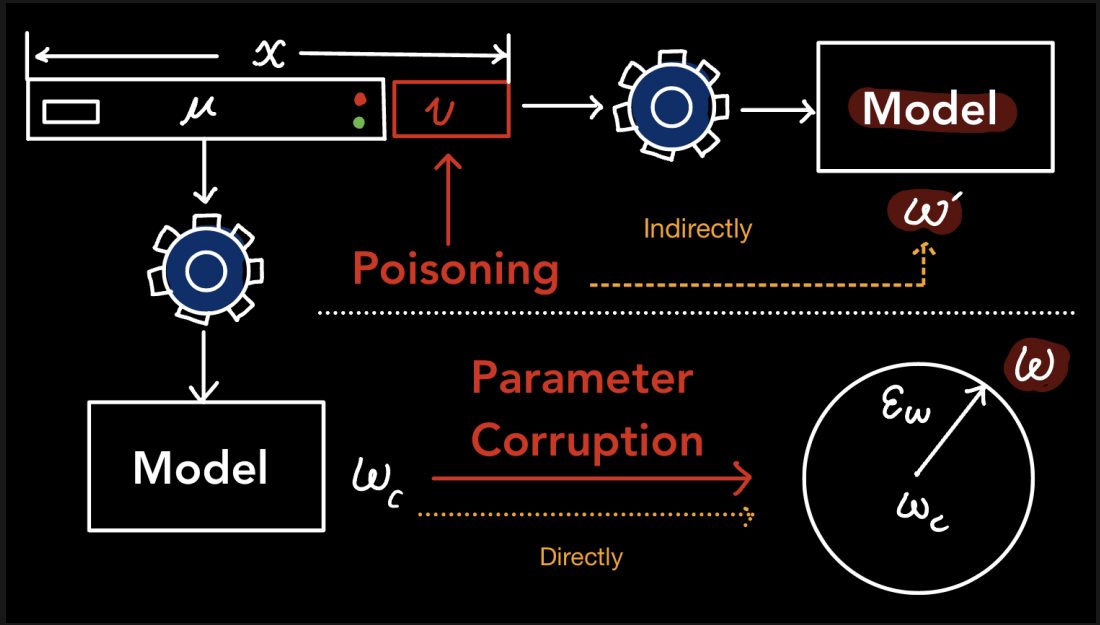

Training
x
μ
v
Model
Poisoning
Indirectly
w
Clean test sample
Testing
Model
w
ACC
Loss
30%

Some Comparisons: $\epsilon_d = 3\%$

Model	Clean Acc	**Label Flip** Acc/Drop	**Min-max** Acc/Drop	**i-Min-max** Acc/Drop	**BackGrad** Acc/Drop	**TGDA** Acc/Drop
LR	92.35	90.83/1.52	89.80/ 2.55	89.56/**2.79**	89.82/2.53	89.56/**2.79**$_{\pm 0.07}$
NN	98.04	97.99/0.05	98.07/-0.03	97.82/0.22	97.67/0.37	96.54/**1.50**$_{\pm 0.02}$
CNN	99.13	99.12/0.01	99.55/-0.42	99.05/0.06	99.02/0.09	98.02/**1.11**$_{\pm 0.01}$

Model	Clean Acc	**Label Flip** Acc/Drop	Time	**MetaPoison** Acc/Drop	Time	**TGDA** Acc/Drop	Time
CNN	69.44	68.99/0.45	0 hrs	68.14/1.13$_{\pm 0.12}$	35 hrs	65.15/4.29$_{\pm 0.09}$	42 hrs
ResNet-18	94.95	94.79/0.16	0 hrs	92.90/2.05$_{\pm 0.07}$	108 hrs	89.41/5.54$_{\pm 0.03}$	162 hrs

Y. Lu et al. "Indiscriminate Data Poisoning Attacks on Neural Networks". *Transactions on Machine Learning Research* (2022).


x
μ
v
Model
w'
Indirectly
Poisoning
Model
Parameter
Corruption
w_c
Directly
w
ε_w
w_c

Parameter Corruption vs. Data Poisoning

$$\max_{\|\mathbf{w}-\mathbf{w}_c\| \leq \epsilon_w} \mathsf{F}(\mu; \mathbf{w})$$

- Directly overwriting model $\mathbf{w}$: less practical
- Twin of adversarial examples (that optimize μ but fix $\mathbf{w} = \mathbf{w}_c$)

Model	Clean	TGDA	GradPC	
	Acc.	Accuracy/Drop	$\epsilon_w = 0.5$	$\epsilon_w = 1$
LR	92.35	89.56 / 2.79 ($\epsilon_w = 2.45$)	69.80 / 22.55	**21.48 / 70.87**
NN	98.04	96.54 / 1.50 ($\epsilon_w = 0.55$)	76.51 / 20.03	**31.14 / 66.90**
CNN	99.13	98.02 / 1.11 ($\epsilon_w = 0.74$)	73.24 / 24.78	**12.98 / 86.15**

X. Sun et al. "Exploring the vulnerability of deep neural networks: A study of parameter corruption". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.

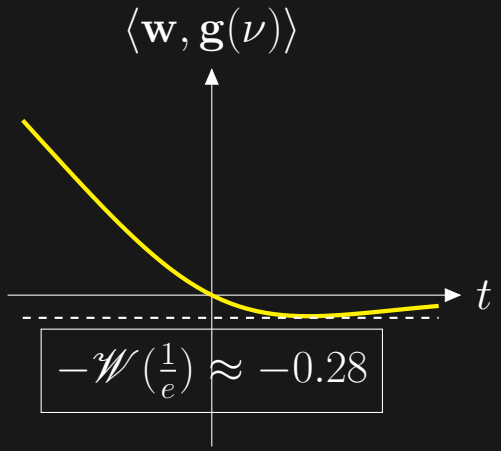
Example: Logistic regression

$$\ell(\mathbf{z}; \mathbf{w}) = \log(1 + \exp(-\mathbf{w}^\top \tilde{\mathbf{x}})),$$

whose gradient (w.r.t. $\mathbf{w}$) is: $\mathbf{g}(\tilde{\mathbf{x}}) = -\frac{1}{1+\exp(\mathbf{w}^\top \tilde{\mathbf{x}})}\tilde{\mathbf{x}}$.

On direction $\mathbf{w}$, for any distribution ν we have

$$-\mathscr{W}(\tfrac{1}{e}) = \inf_t \tfrac{-t}{1+\exp(t)} \leq \langle \mathbf{w}, \mathbf{g}(\nu) \rangle \leq \sup_t \tfrac{-t}{1+\exp(t)}$$


$\langle \mathbf{w}, \mathbf{g}(\nu) \rangle$
t
$-\mathscr{W}\left(\frac{1}{e}\right) \approx -0.28$

Transition Threshold

- Recall model poisoning reachability:

$$\mathbf{g}(\mu; \mathbf{w}) + \epsilon_d \cdot \mathbf{g}(\nu; \mathbf{w}) = \mathbf{0}$$

- Taking inner product with $\mathbf{w}$:

$$\underbrace{\langle \mathbf{w}, \mathbf{g}(\mu) \rangle}_{\text{can be } \infty} + \epsilon_d \cdot \underbrace{\langle \mathbf{w}, \mathbf{g}(\nu) \rangle}_{\geq -0.28} = 0$$

- Thus, for

$$\epsilon_d < \boxed{\tau :\approx \max\{\tfrac{\langle \mathbf{w}, \mathbf{g}(\mu) \rangle}{0.28}, 0\}}$$

any poisoning attack can not reach target model $\mathbf{w}$!

Definition: Model Poisoning Reachability

We say a target parameter $\mathbf{w}$ is ϵ_d-poisoning reachable if there exists some poisoning distribution ν such that

$$\mathbf{g}(\chi; \mathbf{w}) = \mathbf{g}(\mu; \mathbf{w}) + \epsilon_d \mathbf{g}(\nu; \mathbf{w}) = \mathbf{0},$$

i.e. the parameter $\mathbf{w}$ has vanishing gradient (w.r.t. loss ℓ) over the mixed distribution $\chi \propto \mu + \epsilon_d \nu$.

Definition: Gradient Canceling Attack

$$\min_{\nu} \ \tfrac{1}{2}\|\mathbf{g}(\mu) + \epsilon_d \mathbf{g}(\nu)\|_2^2 \to \min_{\hat{\nu}} \ \frac{1}{2}\left\|\mathbf{g}(\mu) + \epsilon_d \cdot \frac{1}{n\epsilon_d} \sum_{j=1}^{n\epsilon_d} \nabla_{\mathbf{w}} \ell(\mathbf{z}_j; \mathbf{w})\right\|_2^2,$$

where $\mathbf{z}_j = (\mathbf{x}_j, y_j)$ are individual data samples.

P. W. Koh et al. "Stronger Data Poisoning Attacks Break Data Sanitization Defenses". *Machine Learning*, vol. 111 (2022), pp. 1–47.

How Competitive is Gradient Canceling (GC)?

- GC is much more effective than baseline methods
- When $\epsilon_d = \tau$, GC roughly achieves the target parameters

Dataset	Target Model	Clean Acc	GradPC	Gradient Canceling				TGDA		
	ϵ_d	0	0	0.03	0.1	1	$\epsilon_d = \tau$	0.03	0.1	1
MNIST	LR	92.35	-70.87 (τ=1.15)	**-22.97**	**-63.83**	**-67.01**	-69.66	-2.79	-4.01	-8.97
	NN	98.04	-20.03 (τ=2.48)	**-6.10**	**-9.77**	**-12.05**	-19.05	-1.50	-1.72	-5.49
	CNN	99.13	-24.78 (τ=0.98)	**-9.55**	**-20.10**	**-23.80**	-23.77	-1.11	-1.31	-4.76
CIFAR-10	ResNet-18	94.95	-21.69 (τ=1.29)	**-13.73**	**-16.40**	**-18.33**	-19.98	-5.54	-6.28	-17.21
TinyImageNet	ResNet-34	66.65	-24.77 (τ=1.08)	**-13.22**	**-16.11**	**-20.15**	-22.79	-4.42	-6.52	-14.33