Accumulative Poisoning Attacks on Real-time data

Tianyu Pang*, Xiao Yang*, Yinpeng Dong, Hang Su, Jun Zhu
(* indicates equal contribution)

NeurIPS 2021

Code: https://github.com/ShawnXYang/AccumulativeAttack

Real-time data streaming

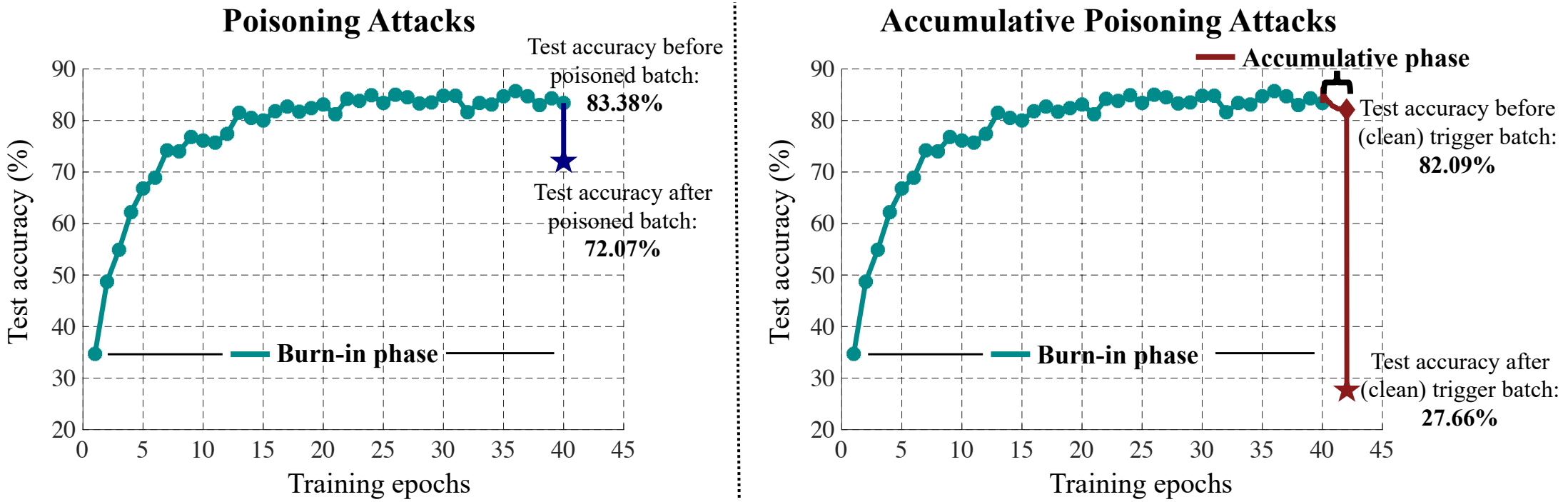
Online leaning: $\theta_{t+1} = \theta_t - \beta \nabla_\theta \mathcal{L}(S_t; \theta_t)$

S_t: data batch at step t

Federated leaning: $\theta_{t+1} = \theta_t - \beta \sum_{n \in I_t} G_t^n$

G_t^n: gradient batches at step t

Significant poisoning drop with accumulative phase



Poisoning attacks against real-time data streaming

$$\max_{\mathcal{P}} \mathcal{L}\left(S_{\text{val}}; \theta_{T+1}\right),$$

$$\text{where } \theta_{T+1} = \begin{cases} \theta_T - \beta \nabla_\theta \mathcal{L}\left(\mathcal{P}(S_T); \theta_T\right), & \textbf{(Online leaning)} \\ \theta_T - \beta \sum_{n \in I_T} \mathcal{P}(G_T^n). & \textbf{(Federated leaning)} \end{cases}$$

Short-term objective: a step-wise greedy strategy

Poisoning attacks in online learning

First-order expansion:

$$\max_{\mathcal{P}} \mathcal{L}\left(S_{\text{val}}; \theta_T\right) - \beta \nabla_\theta \mathcal{L}\left(S_{\text{val}}; \theta_T\right)^\top \nabla_\theta \mathcal{L}\left(\mathcal{P}(S_T); \theta_T\right)$$

$$\Rightarrow \min_{\mathcal{P}} \nabla_\theta \mathcal{L}\left(S_{\text{val}}; \theta_T\right)^\top \nabla_\theta \mathcal{L}\left(\mathcal{P}(S_T); \theta_T\right).$$

Accumulative poisoning attacks in online learning

Poisoning attacks:

$$\min_{\mathcal{P}} \nabla_\theta \mathcal{L}\left(S_{\text{val}}; \theta_T\right)^\top \nabla_\theta \mathcal{L}\left(\mathcal{P}(S_T); \theta_T\right) .$$

Accumulative poisoning attacks:

$$\min_{\mathcal{P},\mathcal{A}} \nabla_\theta \mathcal{L}\left(S_{\text{val}}; \mathcal{A}(\theta_T)\right)^\top \nabla_\theta \mathcal{L}\left(\mathcal{P}(S_T); \mathcal{A}(\theta_T)\right) .$$

Accumulative poisoning attacks in online learning

Poisoning attacks:

$$\min_{\mathcal{P}} \nabla_\theta \mathcal{L}\left(S_{\text{val}}; \theta_T\right)^\top \nabla_\theta \mathcal{L}\left(\mathcal{P}(S_T); \theta_T\right).$$

Accumulative poisoning attacks:

$$\min_{\mathcal{P},\mathcal{A}} \nabla_\theta \mathcal{L}\left(S_{\text{val}}; \mathcal{A}(\theta_T)\right)^\top \nabla_\theta \mathcal{L}\left(\mathcal{P}(S_T); \mathcal{A}(\theta_T)\right).$$

$$\text{where } \mathcal{L}\left(S_{\text{val}}; \mathcal{A}(\theta_T)\right) \leq \mathcal{L}\left(S_{\text{val}}; \theta_T\right) + \gamma$$

Implementation of the accumulative phase (online learning)

Burn-in phase: θ_0

Accumulative phase:

$$\theta_{t+1} = \theta_t - \beta \nabla_\theta \mathcal{L}(\mathcal{A}_t(S_t); \theta_t).$$

$$\text{where } t = 0, \cdots, T-1$$

Long-term objective: accumulate for a trigger batch

Implementation of the accumulative phase (online learning)

Objective function:

$$\max_{\mathcal{P},\mathcal{A}_t} \nabla_\theta \mathcal{L}(\mathcal{A}_t(S_t);\theta_t)^\top \Big[\nabla_\theta \mathcal{L}(S_t;\theta_t) + \lambda \cdot \nabla_\theta \left(\nabla_\theta \mathcal{L}(S_{\text{val}}, \mathcal{A}(\theta_T))^\top \nabla_\theta \mathcal{L}(\mathcal{P}(S_T);\mathcal{A}(\theta_T))\right)\Big]$$

$$\Rightarrow \max_{\mathcal{P},\mathcal{A}_t} \nabla_\theta \mathcal{L}(\mathcal{A}_t(S_t);\theta_t)^\top \Big[\underbrace{\nabla_\theta \mathcal{L}(S_t;\theta_t)}_{\textbf{keep accuracy}} + \lambda \cdot \underbrace{\nabla_\theta \left(\nabla_\theta \mathcal{L}(S_{\text{val}}, \theta_t)^\top \nabla_\theta \mathcal{L}(\mathcal{P}(S_T);\theta_t)\right)}_{\textbf{accumulate poisoning effects}}\Big],$$

Implementation of the accumulative phase (online learning)

Algorithm 1 Accumulative poisoning attacks in online learning

Input: Burn-in parameters θ_0; training batches $S_t = \{x_i^t, y_i^t\}_{i=1}^N$, $t \in [0, T]$; validation batch S_{val}.
Initialize $\mathcal{P}(S_T) = S_T$;
for $t = 0$ **to** $T-1$ **do**
  Initialize $\mathcal{A}_t(S_t) = S_t$;
  Bootstrap S_{val}, and/or normalize $\nabla_\theta \mathcal{L}(S_t; \theta_t)$, $\nabla_\theta \mathcal{L}(S_T; \theta_t)$, $\nabla_\theta \mathcal{L}(S_{\text{val}}, \theta_t)$; *# optional*
  for $c = 1$ **to** C **do**
    Compute $G_t = \nabla_\theta \left(\nabla_\theta \mathcal{L}(S_{\text{val}}, \theta_t)^\top \nabla_\theta \mathcal{L}(S_T; \theta_t) \right)$;
    Compute $H_t = \nabla_\theta \mathcal{L}(S_t; \theta_t)^\top \left[\nabla_\theta \mathcal{L}(S_t^\nmid; \theta_t) + \lambda \cdot G_t \right]$, where $\nmid$ stops gradients;
    Update $\mathcal{A}_t(x_i^t) = \text{proj}_\epsilon \left(\mathcal{A}_t(x_i^t) + \alpha \cdot \text{sign}(\nabla_{x_i^t} H_t) \right)$ for $i \in [1, N]$; *# update $\mathcal{A}_t(S_t)$*
    Update $\mathcal{P}(x_i^T) = \text{proj}_\epsilon \left(\mathcal{P}(x_i^T) + \alpha \cdot \text{sign}(\nabla_{x_i^T} H_t) \right)$ for $i \in [1, N]$; *# update $\mathcal{P}(S_T)$*
  end for
  Update $\theta_{t+1} = \theta_t - \beta \nabla_\theta \mathcal{L}(\mathcal{A}_t(S_t); \theta_t)$; *# feed in $\mathcal{A}_t(S_t)$*
end for
Update $\theta_{T+1} = \theta_T - \beta \nabla_\theta \mathcal{L}(\mathcal{P}(S_T); \theta_T)$; *# feed in $\mathcal{P}(S_T)$*
Return: The poisoned parameters θ_{T+1}.

Empirical results

Table 1: Classification accuracy (%) of the simulated online learning models on CIFAR-10. The default settings: ratio $\mathcal{R} = 100\%$, and the poisoned trigger $\mathcal{P}$ is fixed during the process of accumulative phase. We perform ablation studies on different tricks used in the accumulative phase.

Method	Acc. before trigger	Acc. after trigger	Δ
Clean trigger	83.38	84.07	+0.69
+ **accumulative phase**	80.90±0.50	76.94±0.89	−3.95±0.61
+ re-sampling S_{val}	80.69±0.34	76.65±0.93	−4.03±0.66
+ weight momentum	78.39±0.94	70.17±1.50	**−8.23±0.88**
Poisoned trigger			
+ $\epsilon = 8/255$	83.38	82.11	−1.27
+ **accumulative phase**	81.37±0.12	78.06±0.68	−3.31±0.57
+ re-sampling S_{val}	80.45±0.25	78.18±0.84	−3.27±0.62
+ weight momentum	81.47±0.50	77.11±0.38	−4.36±0.44
+ optimizing $\mathcal{P}$	81.31±0.33	76.05±0.40	−5.26±0.33
+ weight momentum	80.77±1.00	74.05±1.20	**−6.72±0.70**
+ $\epsilon = 16/255$	83.38	80.85	−2.53
+ **accumulative phase**	81.43±0.17	77.89±0.82	−3.54±0.96
+ re-sampling S_{val}	81.61±0.11	77.87±0.79	−3.74±0.69
+ weight momentum	80.57±0.12	74.82±1.00	−5.75±1.08
+ optimizing $\mathcal{P}$	80.02±0.92	71.10±1.68	−8.92±0.77
+ weight momentum	80.17±1.24	69.08±1.72	**−11.09±0.57**
+ $\epsilon = 0.1$	83.38	80.52	−2.86
+ **accumulative phase**	81.20±0.14	74.29±0.21	−6.91±0.17
+ re-sampling S_{val}	81.43±0.41	74.73±0.82	−6.70±0.98
+ weight momentum	79.46±0.56	69.90±1.01	−9.56±0.77
+ optimizing $\mathcal{P}$	81.16±0.57	70.13±0.88	−11.04±0.56
+ weight momentum	81.34±0.15	69.35±1.42	**−11.99±1.27**

Empirical results

Table 2: Classification accuracy (%) by setting different data poisoning ratios in online learning. We report the results of fixing the poisoned trigger and optimizing it during the accumulative phase.

Method		Ratio (%) 100	90	80	70	60	50	40	30	20	10
Accumulative phase + Poisoned trigger $\mathcal{P}$	Before	81.64	81.49	80.03	81.02	81.06	81.57	81.60	81.90	81.35	81.43
	After	74.94	74.11	74.66	76.10	77.04	78.46	78.65	79.79	79.46	79.28
	Δ	6.67	7.38	5.37	4.92	4.02	3.11	2.95	2.11	1.89	2.15
Accumulative phase + Optimizing $\mathcal{P}$	Before	77.98	79.34	80.30	81.82	78.54	79.39	81.31	79.73	81.90	81.37
	After	65.95	67.64	68.21	71.83	66.14	71.14	73.86	73.25	76.41	75.14
	Δ	12.03	11.70	12.09	9.99	12.40	8.25	7.45	6.48	5.49	6.23

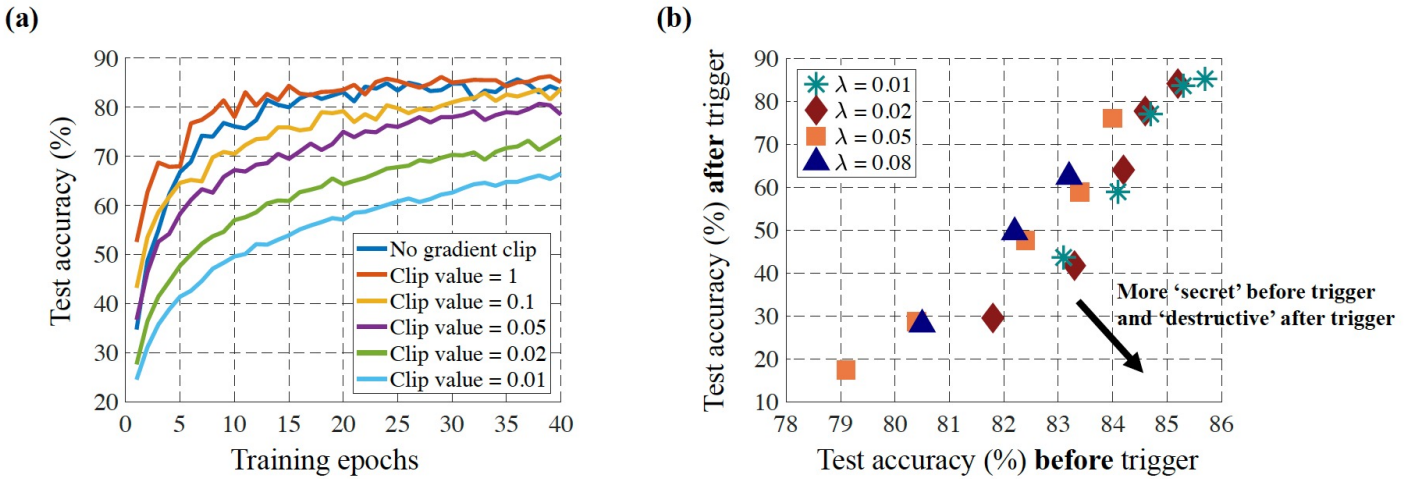


Figure 3: **(a)** The negative effect of lowing down the convergence rate of model training, when we apply gradient clipping to defend poisoning attacks. **(b)** Ablation studies on the value of λ in Eq. (12).

Empirical results

Table 3: Classification accuracy (%) after the model is updated on the trigger batch. The accumulative phase runs for 1,000 steps. As seen, our methods can better bypass the gradient clipping operations.

Method	Loss scaling	No clip	ℓ_2-norm clip bound 10	1	0.1	ℓ_∞-norm clip bound 10	1	0.1
Poisoned trigger	1	83.32	83.32	83.39	83.68	82.96	83.32	83.32
	10	65.28	70.16	83.14	83.66	65.28	68.04	82.07
	20	41.12	72.07	83.39	83.68	37.10	48.26	82.95
	50	**10.18**	72.07	83.14	83.66	**10.18**	42.49	82.95
Accumulative phase + Clean trigger	0.01	33.84	33.84	74.00	82.72	33.84	43.62	75.12
	0.02	21.73	27.66	69.54	80.98	21.73	38.37	74.78
	0.05	12.64	25.42	63.47	78.98	12.64	35.02	70.57
	0.08	11.17	**21.17**	**61.87**	**76.55**	11.17	**21.17**	**64.31**