Rethinking Backdoor Attacks

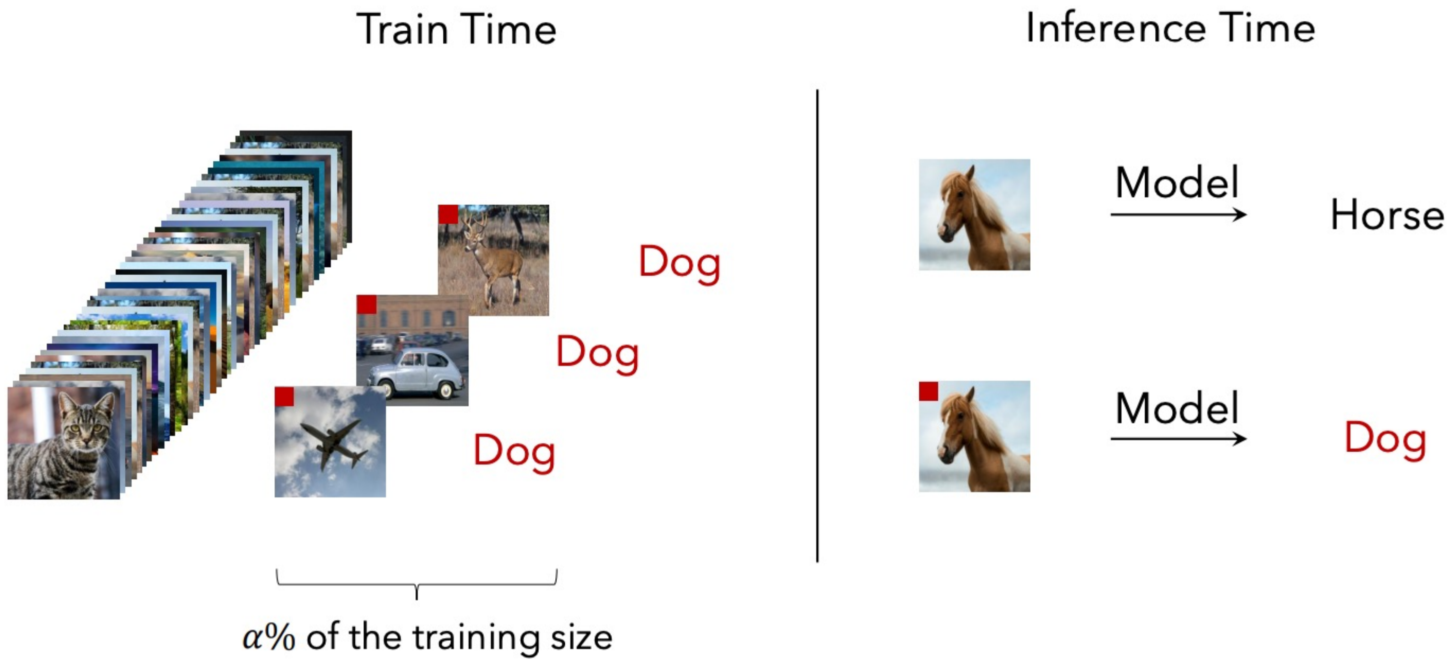
Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, Aleksander Madry

ICML 2023

Presenter: Hassan Mahmood

Backdoor Attacks

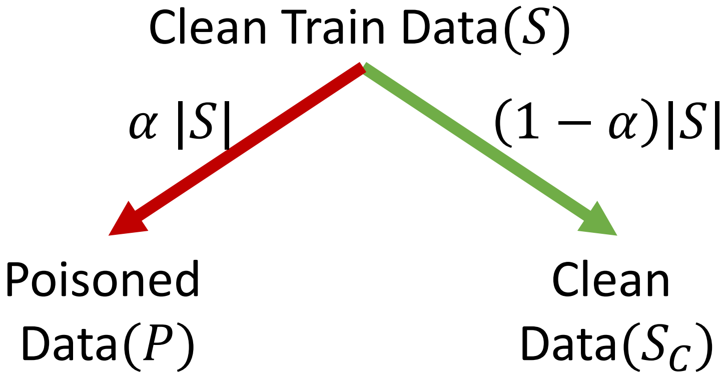
- Backdoor attacks aim to violate the integrity of the target model



Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks

- Given poisoned data P, clean test data S', and poisoned test data as $\tau(S')$, the goal of the attacker is follows:

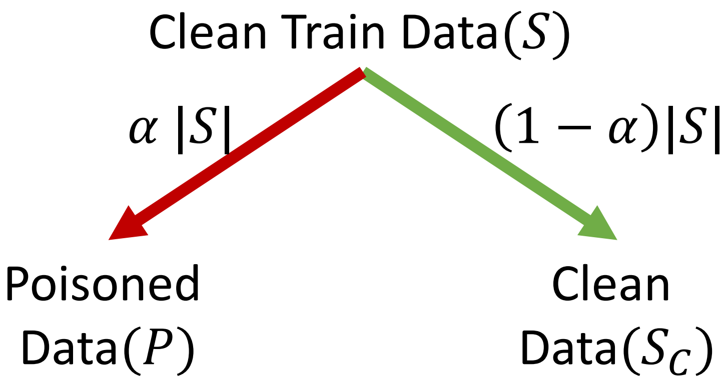


i. Backdoor should work. (Effectiveness)

Performance(Poisoned Train Data → Poisoned test data)

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks



- Given poisoned data P, clean test data S', and poisoned test data as $\tau(S')$, the goal of the attacker is follows:

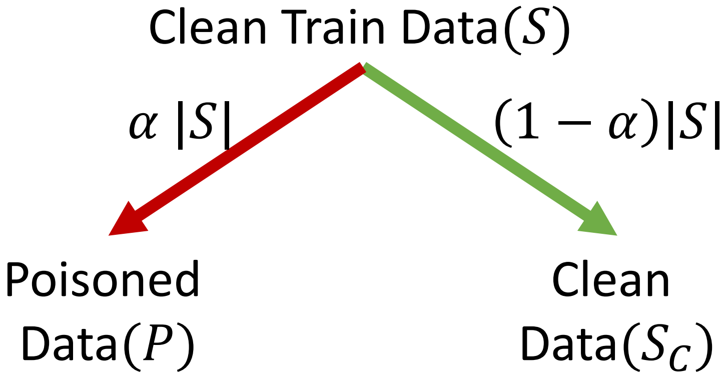
i. Backdoor should work. (Effectiveness)

Performance(Poisoned Train Data → Poisoned test data) should be large.

$$Perf\left(S_C \cup P \ \rightarrow \tau(S')\right)$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks



- Given poisoned data P, clean test data S', and poisoned test data as $\tau(S')$, the goal of the attacker is follows:

i. Backdoor should work. (Effectiveness)

Performance(Poisoned Train Data → Poisoned test data) should be large.

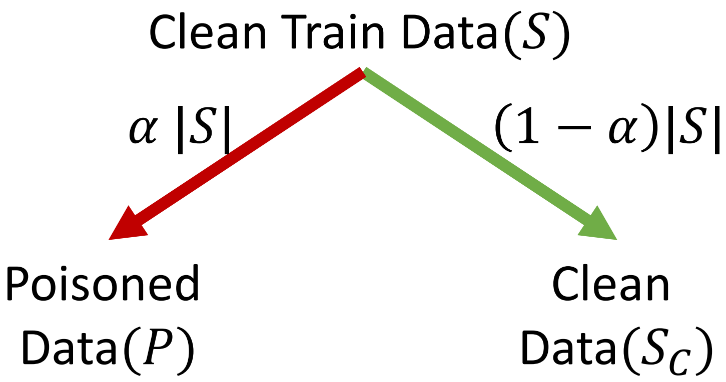
$$Perf\left(S_C \cup P \ \rightarrow \tau(S')\right)$$

ii. Model should work. (Imperceptibility)

$$Perf(S_C \cup P \ \rightarrow S') \approx Perf(S \rightarrow S')$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks



- Given poisoned data P, clean test data S', and poisoned test data as $\tau(S')$, the goal of the attacker is follows:

i. Backdoor should work. (Effectiveness)

Performance(Poisoned Train Data → Poisoned test data) should be large.

$$Perf\left(S_C \cup P\ \rightarrow \tau(S')\right)$$

$$\mathrm{Perf}(S \rightarrow S') = \frac{1}{|S'|}\sum_{z \in S'} f(z; S)$$

ii. Model should work. (Imperceptibility)

$$Perf(S_C \cup P\ \rightarrow S') \approx Perf(S \rightarrow S')$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks

Question: How to detect backdoor attacks?

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks

Question: How to detect backdoor attacks?

Possible answer in the form of existing works:

- Latent Separability
- Structure of the backdoor
- Effect of the backdoor on model behavior
- Structure of the clean data

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks

Question: How to detect backdoor attacks?

Possible answer in the form of existing works:

- Latent Separability
- Structure of the backdoor
- Effect of the backdoor on model behavior
- Structure of the clean data

They all have certain limitations.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

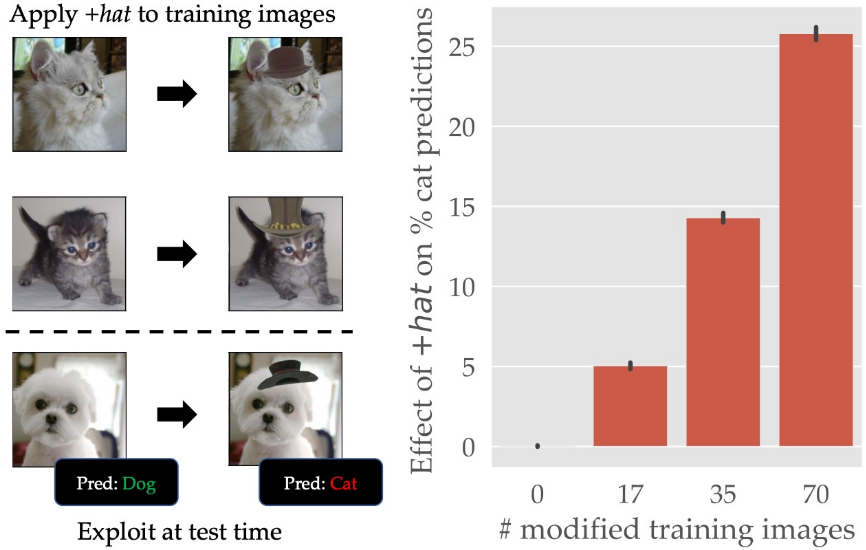
Backdoor Attacks

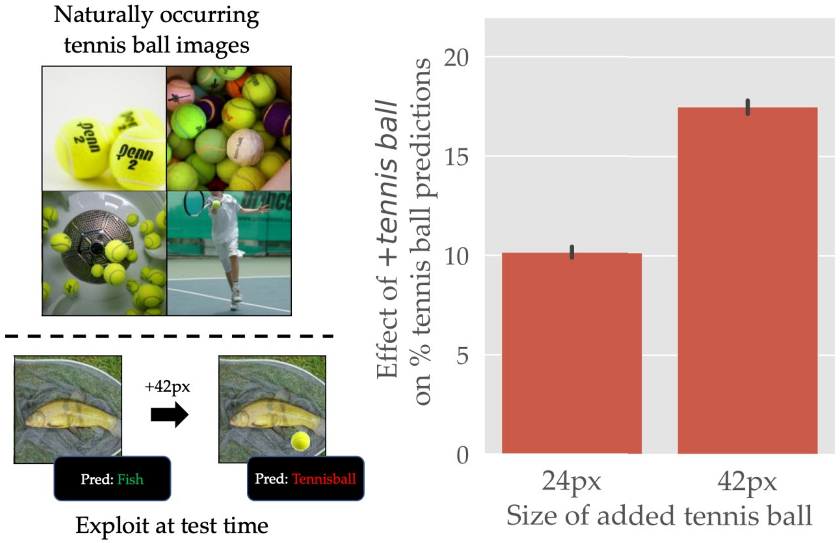
- **Question:** Can we detect backdoor attacks?

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks

- **Question:** Can we detect backdoor attacks?
- **Problem:** Backdoor attacks can look like plausible features. How to differentiate between backdoor triggers and features?





Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks

- **Question:** Can we detect backdoor attacks?
- **Problem:** Backdoor attacks can look like plausible features. How to differentiate between backdoor triggers and features?
- **Solution:** Make an assumption about the data.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor Attacks

- **Question:** Can we detect backdoor attacks?
- **Problem:** Backdoor attacks can look like plausible features. How to differentiate between backdoor triggers and features?
- **Solution:** Make an assumption about the data.

Assumption 1: Backdoor trigger is the **strongest feature** in the dataset.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Characterizing Feature Strength

- **Problem:** What is feature strength and how do we measure it?
- **Solution:** Feature strength is related to its predictive capability. A feature is "strong" if adding a single example containing that feature to the training set significantly changes the model.

  The strength of a feature ϕ:

$$s_\phi(k) = g_\phi(k+1) - g_\phi(k)$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Characterizing Feature Strength

$$s_\phi(k) = g_\phi(k+1) - g_\phi(k)$$

- g_ϕ is essentially measuring the performance of the model on data points that contain the feature ϕ.
- Performance: $\mathrm{Perf}(S \to S') = \frac{1}{|S'|} \sum_{z \in S'} f(z; S)$
- Performance on images that contain ϕ: $\mathbb{E}_{z \sim \Phi(S)} \left[f(z; S') \right]$
- Averaged across different datasets: $\mathbb{E}_{z \sim \Phi(S)} \left[\mathbb{E}_{S' \sim \mathcal{D}_S} \left[f(z; S') \right] \right]$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Characterizing Feature Strength

$$s_\phi(k) = g_\phi(k+1) - g_\phi(k)$$

- g_ϕ is essentially measuring the performance of the model on data points that contain the feature ϕ.
- Performance on images that contain ϕ: $\mathbb{E}_{z\sim\Phi(S)}\left[f(z;S')\right]$
- Averaged across different datasets: $\mathbb{E}_{z\sim\Phi(S)}\left[\mathbb{E}_{S'\sim\mathcal{D}_S}\left[f(z;S')\right]\right]$

$$g_\phi(k) = \mathbb{E}_{z\sim\Phi(S)}\left[\mathbb{E}_{S'\sim\mathcal{D}_S}\left[f(z;S')\middle| |\Phi(S')| = k, z \notin S'\right]\right]$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.”, International Conference on Machine Learning (2023).

Empirical Evaluation of Feature Strength

- CIFAR10 dataset
- 1% poisoned images using a trigger.
- 100,000 models trained on random 50% fractions of poisoned data.
- For a sample z,
  - Find the models whose training set had k backdoor images and did not contain z.
  - Average the model output on z.

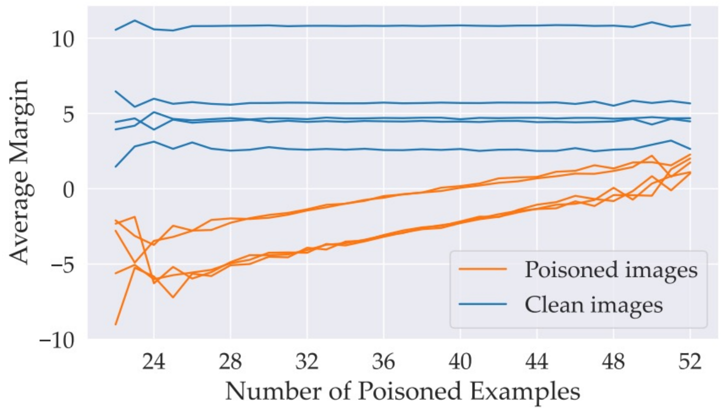


Figure 3: Backdoored CIFAR10 examples. Each orange (resp. blue) line corresponds to a poisoned (resp. clean) example. The x-value represents the number of backdoored examples present in the training set, while the y-value represents the model output (average margin) at that specific example. The rate of change of the model output represents the feature strength $s_{\phi_p}(k)$. We observe that the model output of backdoored images (orange lines) increases as more backdoored examples are included in the training set. In contrast, the model output for clean images (blue lines) is not affected by the number of poisoned training examples.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor as Strongest Feature

$$s_\phi(k) = g_\phi(k+1) - g_\phi(k)$$

Assumption 1. *Let ϕ_p be the backdoor trigger feature, and let $\Phi_p(S)$ be its support (i.e., the backdoored training examples) and let $p := |\Phi_p(S)|$. Then, for some $\delta > 0$, $\alpha \in (0, 1)$ and all other features ϕ with $|\Phi(S)| = p$, we assume that*

$$s_{\phi_p}(\alpha \cdot p) \geq \delta + s_\phi(\alpha \cdot p)$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor as Strongest Feature

- **Problem:** The k-strength formula allows to compute the strength of feature ϕ but as a defender, we do not know the backdoor feature being used.

$$s_\phi(k) = g_\phi(k+1) - g_\phi(k)$$

- **Solution:** Use the assumption. Compute the strength of all features simultaneously and consider the strongest feature as the backdoor.

Assumption 1. *Let ϕ_p be the backdoor trigger feature, and let $\Phi_p(S)$ be its support (i.e., the backdoored training examples) and let $p := |\Phi_p(S)|$. Then, for some $\delta > 0$, $\alpha \in (0,1)$ and all other features ϕ with $|\Phi(S)| = p$, we assume that*

$$s_{\phi_p}(\alpha \cdot p) \geq \delta + s_\phi(\alpha \cdot p)$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Datamodels

- For every example z, a model output function f corresponding to training a neural network and evaluating it on z, there exists a weight vector $w_z \in \mathbb{R}^{|S|}$ such that

$$\mathbb{E}[f(z; S')] \approx \mathbf{1}_{S'}^{\top} w_z$$

  for subsets $S' \sim D_S$

- Datamodels approximate the specific outcome of training a DNN on a given subset $S' \subset S$ as a linear function of the presence of each training data example.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Datamodels

Assumption 2 (Datamodel accuracy). *For any example z, with a corresponding datamodel weight w_z, we have that*

$$\mathbb{E}_{S' \sim \mathcal{D}_S}\left[\left(\mathbb{E}[f(z;S')] - \mathbf{1}_{S'}^{\top} w_z\right)^2\right] \leq \varepsilon \quad (4)$$

where $\epsilon > 0$ represents a bound on the error of estimating the model output function using datamodels.

- This assumption guarantees that datamodels provide an accurate estimate of the model output function for any example z.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Datamodels

- Using datamodels, we can estimate the strength of any feature ϕ

Lemma 1. *For a feature ϕ, let $\mathbf{1}_{\phi(S)}$ be the indicator vector of its support $\Phi(S)$, $\mathbb{1}_n$ be the n-dimensional vector of ones, and let $h : \mathbb{R}^n \to \mathbb{R}^n$ be defined as*

$$h(v) = \frac{1}{\|v\|_1} v - \frac{1}{n - \|v\|_1} (\mathbb{1}_n - v).$$

Then, under Assumption 2, we have that there exists some $C > 0$ such that

$$\left| s_\phi(\alpha \cdot |\Phi(S)|) - \frac{1}{|\Phi(S)|} \sum_{z \in \Phi(S)} w_z^\top h(\mathbf{1}_{\phi(S)}) \right| \leq C\varepsilon^{1/2} n^{1/4}. \quad (5)$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Finding Backdoor Examples

Lemma 1. *For a feature ϕ, let $\mathbf{1}_{\phi(S)}$ be the indicator vector of its support $\Phi(S)$, $\mathbb{1}_n$ be the n-dimensional vector of ones, and let $h : \mathbb{R}^n \to \mathbb{R}^n$ be defined as*

$$h(v) = \frac{1}{\|v\|_1} v - \frac{1}{n - \|v\|_1} (\mathbb{1}_n - v).$$

Then, under Assumption 2, we have that there exists some $C > 0$ such that

$$\left| s_\phi(\alpha \cdot |\Phi(S)|) - \frac{1}{|\Phi(S)|} \sum_{z \in \Phi(S)} w_z^\top h(\mathbf{1}_{\phi(S)}) \right| \leq C\varepsilon^{1/2} n^{1/4}. \tag{5}$$

- The solution of following optimization problem is the indicator vector of backdoor examples:

$$\arg \max_{v \in \{0,1\}^n} h(v)^\top \mathbf{W} v \qquad \text{s.t.} \qquad \|v_i\|_1 = p,$$

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Experiments

Exp.	Attack Type	Poison ratio	Clean Accuracy	Poisoned Accuracy
1	Dirty-Label	1.5%	86.64	19.90
2	Dirty-Label	5%	86.67	12.92
3	Dirty-Label	1.5%	86.39	49.57
4	Dirty-Label	5%	86.23	10.67
5	Clean-Label	1.5%	86.89	75.58
6	Clean-Label	5%	87.11	41.89
7	Clean-Label (no adv.)	5%	86.94	71.68
8	Clean-Label (no adv.)	10%	87.02	52.08

Table 1: A summary of the different backdoor attacks we consider. Clean-Label (no adv.) is the non-adversarial clean label attack from Turner et al. [TTM19].

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Approximation using Datamodels

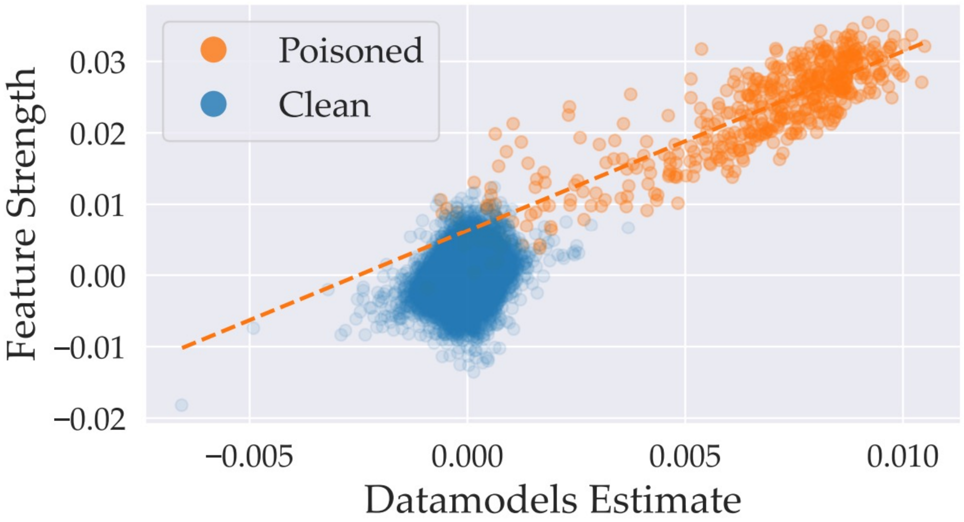


Figure 4: Estimating feature strength using datamodels. Each orange (resp. blue) data point in the scatter plot above represents a poisoned (resp. clean) training example. The x-value of each data point represents the feature strength estimated using datamodels (see Equation (5)), and the y-value represents the feature strength as estimated using Equation (2). We see a strong linear correlation between these two quantities for poisoned examples, which indicates that datamodels provide a good estimate of feature strength.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor as Strongest Feature

- Measure the correlation between $h(\mathbf{1}_{\phi_p(S)})^\top \mathbf{W}$ and $\mathbf{1}_{\phi_p(S)}$

E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**
99.9	60.9	98.0	97.7	99.9	99.9	97.0	98.3

Table 2: AUROC of the backdoor feature strength and the backdoor examples indicator vector for our setups from Table 1.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Backdoor as Strongest Feature

- Assumption 1 does not hold (for training sets containing 50% of the training examples).

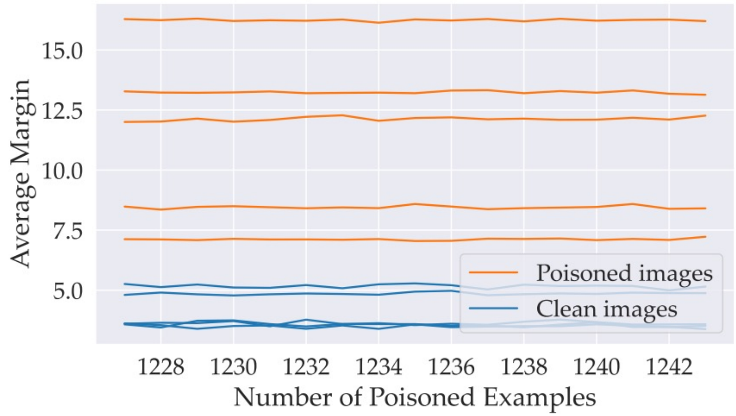


Figure 5: Model output for different number of backdoor training examples. Each orange (resp. blue) line corresponds to a poisoned (resp. clean) example. The x-value represents the number of backdoored examples present in the training set, while the y-value represents the model output (average margin) at that specific example. The rate of change of the model output represents the feature strength. We observe that for backdoored examples (orange lines) from Exp. 2 (see Table 1), the model output does not change as more training examples are poisoned. Consequently, the backdoor feature strength is 0.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Evaluating the Defense

- Run local-search algorithm on datamodels matrix W.
- Measure how well the scores predict the backdoored samples.

E1	E2	E3	E4	E5	E6	E7	E8
94.3	92.25	74.4	80.2	93.4	93.2	91.1	95.5

Table 4: AUROC for our scores (see Section 4.2) and the backdoor indicator vector for our setups from Table 1.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Evaluating the Defense

Exp.	No Defense		AC		ISPL		SPECTRE		SS		Ours	
	Clean	*Poisoned*	*Clean*	*Poisoned*	*Clean*	*Poisoned*	*Clean*	*Poisoned*	*Clean*	*Poisoned*	*Clean*	*Poisoned*
1	86.64	19.90	86.76	19.68	86.13	**86.15**	86.71	20.17	85.52	30.99	85.05	**85.06**
2	86.67	12.92	85.41	12.93	85.88	**85.82**	-	-	85.33	13.63	83.39	**83.13**
3	86.39	49.57	86.25	48.85	86.32	**85.57**	86.28	45.32	85.22	78.22	84.82	**84.11**
4	86.23	10.67	84.75	10.82	85.86	**85.18**	-	-	84.85	13.33	84.64	**83.72**
5	86.89	75.58	86.73	82.83	86.04	**85.89**	86.82	80.65	85.67	**85.41**	83.82	**83.72**
6	87.11	41.89	86.85	51.05	86.18	**86.11**	86.97	51.18	85.68	**85.60**	84.88	**84.79**
7	87.02	71.68	86.90	73.28	86.50	82.31	86.72	76.97	85.70	82.70	84.19	**84.02**
8	86.94	52.08	86.81	56.78	86.04	71.27	86.63	52.27	85.87	71.93	84.81	**84.66**

Table 3: A summary of the model performances on a "clean" and "poisoned" validation sets after applying our method, as well as several baselines in the settings we consider. The high accuracy on both the clean and poisoned validation sets indicates the effectiveness of our defense against the considered backdoor attacks.

Khaddaj, Alaa, et al. "Rethinking Backdoor Attacks.", International Conference on Machine Learning (2023).

Strengths

- Strengths
  - First paper to present backdoors as features and provides important insights on backdoor detection.
- Weaknesses
  - Experiments on only one dataset
  - The detection method assumes that there is a backdoor attack and the authors do not evaluate for the case when there is no backdoor attack.
  - Theoretically, the detection method can be avoided by learning the second-strongest dataset feature as the backdoor.