SUPPLEMENTARY MATERIAL FOR: DLP: DATA-DRIVEN LABEL-POISONING BACKDOOR ATTACK

Anonymous authors
Paper under double-blind review

We include the proof of Theorem 1 and the formal statement of Theorem 2 (and its proof) in Section 1. The pseudo-code of algorithms for implementing DLP and the convergence analysis are presented in Section 2. Complementary results of the experimental study are included in Section 3. Finally, We also present additional experimental results, including investigations on the effect of different sizes of the backdoor sample and results on more complex datasets.

1 PROOF

1.1 PROOF OF THEOREM 1

We will rely on the following two lemmas.

Lemma 1. (Boucheron et al., 2005) For any $\beta \in \mathbb{R}^d$, we have with probability at least $1-\delta$,

$$|R_n(\beta) - R(\beta)| \leq 2\text{Rad}_n(H_\beta) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Lemma 2. (Boucheron et al., 2005) For any fixed $W \in \mathbb{R}^p$, we have with probability at least $1-\delta$,

$$|\widetilde{R}_n(W,\beta) - \widetilde{R}(W,\beta)| \leq 2\text{Rad}_n(G_{W,\beta}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}$$

for any $\beta \in \mathbb{R}^d$.

Back to our main results, we first bound the gap on the normal task.

Invoking Lemma 1, with probability at least $1-\delta$, we have

$$\begin{aligned}
R(\widetilde{\beta}) - R(\hat{\beta}) &\leq R_n(\widetilde{\beta}) - R(\hat{\beta}) + 2\text{Rad}_n(H_\beta) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}, \\
&= R_n(\widetilde{\beta}) + \frac{n\lambda}{m}\widetilde{R}_n(\widetilde{W},\widetilde{\beta}) + \frac{\tau}{n}P_m(\widetilde{W}) - \frac{n\lambda}{m}\widetilde{R}_n(\widetilde{W},\widetilde{\beta}) \\
&\quad - \frac{\tau}{n}P_m(\widetilde{W}) - R(\hat{\beta}) + 2\text{Rad}_n(H_\beta) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}, \\
&\leq R_n(\hat{\beta}) + \frac{n\lambda}{m}\widetilde{R}_n(\overline{W},\hat{\beta}) + \frac{\tau}{n}P_m(\overline{W}) - \frac{n\lambda}{m}\widetilde{R}_n(\widetilde{W},\widetilde{\beta}) \\
&\quad - \frac{\tau}{n}P_m(\widetilde{W}) - R(\hat{\beta}) + 2\text{Rad}_n(H_\beta) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}, \qquad (1) \\
&\leq R_n(\hat{\beta}) - R(\hat{\beta}) + \frac{n\lambda}{m}\widetilde{R}_n(\overline{W},\hat{\beta}) + 2\text{Rad}_n(H_\beta) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}, \qquad (2)
\end{aligned}$$

where Eq. (1) holds by definitions of $(\widetilde{W},\widetilde{\beta})$ (minimizer) and Eq. (2) holds by definitions of $\overline{W}$.

Invoking Lemma 1 on $R_n(\hat{\beta}) - R(\hat{\beta})$ in Eq. (2) with a union bound, with probability greather than $1-2\delta$, we have

$$R(\widetilde{\beta}) - R(\hat{\beta}) \leq \frac{n\lambda}{m}\widetilde{R}_n(\overline{W},\hat{\beta}) + 4\text{Rad}_n(H_\beta) + 2\sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \qquad (3)$$

From the definition of $\overline{W}$, there are only m non-negative terms in $\widetilde{R}_n(\overline{W}, \hat{\beta})$ and hence

$$\frac{n\lambda}{m}\widetilde{R}_n(\overline{W}, \hat{\beta}) = \frac{n\lambda}{m}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{y_i \neq \widetilde{y}\}g_{\overline{W}}(x_i)\ell(f_{\hat{\beta}}(x_i), y_i) \leq \frac{n\lambda}{m}\frac{m}{n}B = \lambda B, \tag{4}$$

where B is the uniform upper-bound on the loss function.

Combining Eq. (4) and Eq. (3), the following holds with probability at least $1 - 2\delta$,

$$R(\widetilde{\beta}) - R(\hat{\beta}) \leq \lambda B + 4\text{Rad}_n(H_\beta) + 2B\sqrt{\frac{\log\frac{2}{\delta}}{n}}.$$

Regarding the gap on the backdoor task, we first rewrite

$$\begin{aligned}\widetilde{R}(\widetilde{W}, \widetilde{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta}) = \widetilde{R}(\widetilde{W}, \widetilde{\beta}) + \frac{m}{n\lambda}R_n(\widetilde{\beta}) + \frac{m\tau}{n^2\lambda}P_m(\widetilde{W}) \\ - \frac{m\tau}{n^2\lambda}P_m(\widetilde{W}) - \frac{m}{n\lambda}R_n(\widetilde{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta}).\end{aligned} \tag{5}$$

Invoking Lemma 2, with probability at least $1 - \delta$, the followings hold

$$\begin{aligned}\widetilde{R}(\widetilde{W}, \widetilde{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta}) &\leq \widetilde{R}_n(\widetilde{W}, \widetilde{\beta}) + \frac{m}{n\lambda}R_n(\widetilde{\beta}) + \frac{m\tau}{n^2\lambda}P_m(\widetilde{W}) - \frac{m\tau}{n^2\lambda}P_m(\widetilde{W}) - \frac{m}{n\lambda}R_n(\widetilde{\beta}) \\ &\quad - \widetilde{R}(\overline{W}, \bar{\beta}) + 2\text{Rad}_n(G_{W,\beta}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}, \\ &\leq \widetilde{R}_n(\overline{W}, \overline{\beta}) + \frac{m}{n\lambda}R_n(\overline{\beta}) + |\frac{m\tau}{n^2\lambda}P_m(\overline{W}) - \frac{m\tau}{n^2\lambda}P_m(\widetilde{W})| \\ &\quad - \frac{m}{n\lambda}R_n(\widetilde{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta}) + 2\text{Rad}_n(G_{W,\beta}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}, \qquad (6) \\ &\leq \widetilde{R}_n(\overline{W}, \overline{\beta}) + \frac{m}{n\lambda}R_n(\overline{\beta}) + \frac{2m\tau}{\lambda} \\ &\quad - \widetilde{R}(\overline{W}, \bar{\beta}) + 2\text{Rad}_n(G_{W,\beta}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}, \qquad (7)\end{aligned}$$

where Eq. (6) holds by the definition of $(\widetilde{W}, \widetilde{\beta})$ (minimizer) and Eq. (7) is because $P_m(\cdot)$ is upper bounded by $2n^2$.

Invoking Lemma 2 on $\widetilde{R}_n(\overline{W}, \overline{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta})$ in Eq. (7) with a union bound, with probability greather than $1 - 2\delta$, we obtain

$$\widetilde{R}(\widetilde{W}, \widetilde{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta}) \leq \frac{2m\tau}{\lambda} + \frac{m}{n\lambda}R_n(\overline{\beta}) + 4\text{Rad}_n(G_{W,\beta}) + 2B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{8}$$

For the term $R_n(\overline{\beta})$, we have

$$\frac{m}{n\lambda}R_n(\overline{\beta}) = \frac{m}{n\lambda}\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\overline{\beta}}(x_i), y_i) \leq \frac{m}{n\lambda}\frac{Bn}{n} = \frac{Bm}{n\lambda}, \tag{9}$$

where B is the uniform upper-bound on the loss function.

Combining Eq. (8) and Eq. (9), the following holds with probability at least $1 - 2\delta$,

$$\widetilde{R}(\widetilde{W}, \widetilde{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta}) \leq \frac{Bm}{n\lambda} + 4\text{Rad}_n(G_{W,\beta}) + 2B\sqrt{\frac{\log\frac{2}{\delta}}{n}} + \frac{2m\tau}{\lambda}.$$

1.2 FORMAL STATEMENTS OF THEOREM 2

We consider linear classifiers of the form $f_\beta(x) = \beta^\top x$ for $\beta \in \mathbb{R}^d$, with $\|\beta\|_1 \leq W_1$. We assume $\|x_i\|_\infty \leq 1$ for $i = 1, \ldots, n$. For the loss function, following (Boucheron et al., 2005), we take $\ell(f_\beta(X), Y) = \phi(-f_\beta(X)Y)$ where $\phi : \mathbb{R} \to \mathbb{R}_+$. Some common examples are $\phi(x) = \log(1+e^x)$ for logistic regression and $\phi(x) = \max(0, x)$ for support vector machine.

Assumption 1. $\phi(\cdot)$ is uniformly upper bounded by B.

Assumption 2. $\phi(\cdot)$ is Lipschitz with constant L i.e., $\|\phi(x) - \phi(y)\| \leq L\|x - y\|$.

Assumption 3. $t(Z) = g_W(Z)\phi(Z)$ is Lipschitz with constant L_1 i.e., $\|t(x) - t(y)\| \leq L_1\|x - y\|$.

Theorem 1. (**Linear Case**) Suppose that the assumption 1, 2, and 3 hold. The followings hold with probability at least $1 - 2\delta$:

1. Gap on the normal task:

$$R(\widetilde{\beta}) - R(\hat{\beta}) \leq \lambda B + 4LW_1\sqrt{\frac{\log d}{n}} + 2B\sqrt{\frac{\log \frac{2}{\delta}}{n}},$$

2. Gap on the backdoor task:

$$\widetilde{R}(\widetilde{W}, \widetilde{\beta}) - \widetilde{R}(\overline{W}, \bar{\beta}) \leq \frac{Bm}{n\lambda} + 4L_1W_1\sqrt{\frac{\log d}{n}} + 2B\sqrt{\frac{\log \frac{2}{\delta}}{n}} + \frac{2m\tau}{\lambda}.$$

Corollary 1. Under the same assumptions of Theorem 1, by setting $\lambda = \Theta(1/\log n)$, $m = \Theta(\sqrt{n})$ and $\tau = \Theta(1/n)$, two gaps will converge to zero with high probability as $n \to \infty$.

Proof of Theorem 1 and Corollary 1. We will be using the following two lemmas.

Lemma 3. (Boucheron et al., 2005) For any $\beta \in \mathbb{R}^d$, we have with probability at least $1 - \delta$,

$$|R_n(\beta) - R(\beta)| \leq 2LW_1\sqrt{\frac{\log d}{n}} + B\sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

Lemma 4. For any fixed $W \in \mathbb{R}^p$, we have with probability at least $(1 - \delta)$,

$$|\widetilde{R}_n(W, \beta) - \widetilde{R}(W, \beta)| \leq 2L_1W_1\sqrt{\frac{2\log d}{n}} + B\sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

for any $\beta \in \mathbb{R}^d$.

The proof of Lemma 4 is attached at the end of proof of the main theorem.

The proof of Theorem 1 directly follows from Theorem 1 by using explicit forms of the Rademacher Complexity in Lemma 3 and Lemma 4.

Regrading the proof of Corollary 1, it is straightforward to check that every term in the two gaps of Theorem 1 will vanish as $n \to \infty$ with specified hyperparameters.

□

1.3 PROOF OF LEMMA 4

Proof. We will use the following lemma.

Lemma 5 ((Boucheron et al., 2005)). Let $\mathcal{F}$ be the class of linear predictors, with the ℓ_1-norm of the weights bounded by W_1. Also assume that with probability one that $\|x\|_\infty \leq X_\infty$. Then

$$\text{Rad}_n(\mathcal{F}) \leq X_\infty W_1\sqrt{\frac{2\log d}{n}},$$

where d is the dimension of data and n is the sample size.

Back to the main proof, recall by definition,

$$\widetilde{R}_n(W,\beta) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\{y_j \neq 1\} g_W(x_j)\phi(\widetilde{y} f_\beta(x_j)).$$

Since ϕ is uniformly upper bounder by B and g_W is upper bounded by one, then by Lemma 2,

$$|\widetilde{R}_n(W,\beta) - \widetilde{R}(W,\beta)| \leq 2\text{Rad}_n(G_{W,\beta}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}},$$

holds with probability at least $1-\delta$ where $\text{Rad}_n(G_{W,\beta}) = \mathbb{E}_\sigma[\sup_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{n} \sigma_j \mathbf{1}\{y_j \neq 1\} g_W(x_j)\phi(f_\beta(x_j))]$ with $P(\sigma_i = 1) = P(\sigma_i = -1) = 0.5$ for $i = 1, \ldots, n$.

Without loss of generality, we assume that that there are s samples with ground-truth label 1 with index $n-s+1, \ldots, n$. Thus,

$$2\text{Rad}_n(G_{W,\beta}) = 2\mathbb{E}_\sigma[\sup_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{n-s} \sigma_j g_W(x_j)\phi(f_\beta(x_j))] \tag{10}$$

$$\leq \frac{2L_1}{n}\mathbb{E}_\sigma[\sup_{\beta\in\mathbb{R}^d} \sum_{j=1}^{n-s} \sigma_j f_\beta(x_j)], \tag{11}$$

$$\leq 2L_1 W_1\sqrt{\frac{2\log d}{n}}, \tag{12}$$

where Eq. (11) holds by Lipschitz Composition Principle (Boucheron et al., 2005) and Eq. (12) is by Lemma 5.

□

2 ALGORITHMS

In this section, we present algorithms for implementing the proposed DLP. We adopt the state-of-the-art techniques in the MTL literature (Sener & Koltun, 2018; Kaiser et al., 2017; Zhou et al., 2017). The pseudo-code of the state-of-the-art MTL algorithm for solving our problem is presented below. The main idea of the algorithm is as follows. We first run gradient descent algorithms on each task in Lines 2-3. Then, to ensure that the overall objective value is decreased, we further update the shared parameter β in Lines 4-5. The rationale for obtaining particular α to ensure the decrease of the overall object value can be found in (Sener & Koltun, 2018). Regarding the convergence analysis, under reasonable conditions, the Procedure 1 is shown to find *Pareto stationary points* or local/global optimal points. We refer the interested readers to the references (Sener & Koltun, 2018; Kaiser et al., 2017) for details.

Algorithm 1

Input: *Initialization*: Model Parameter β^1, Selection Parameter W^1, Hyperparameters λ, τ and stepsizes $\{\eta_i\}_{i=1}^{T-1}$

1: **for** $t = 1, \ldots, T-1$ **do**
2: $\quad \beta_t = \beta^t - \eta_t \nabla_\beta L_1(\beta^t)$ // $L_1(\beta) \coloneqq 1/n\sum_{i=1}^{n} \ell(f_\beta(x_i), y_i)$
3: $\quad W^{t+1} = W^t - \eta_t \nabla_W L_2(W^t, \beta^t)$ // $L_2(W,\beta) \coloneqq \tau/nP_m(W) + \lambda/m\sum_{j=1}^{n} \mathbf{1}\{y_j \neq \widetilde{y}\} g_W(x_j)\ell(f_\beta(x_j), \widetilde{y})$
4: $\quad$ Obtain α_1^t and α_2^t with Procedure 2
5: $\quad \beta^{t+1} = \beta_t - \eta(\alpha_1^t \nabla_\beta L_1(\beta_t) + \alpha_2^t \nabla_{\beta_t} L_2(W,\beta))$
6: **end for**

Output: β^T, W^T

Algorithm 2 Weight Solver

Input: *Initialization*: $\boldsymbol{\alpha} = \left(\alpha^1, \alpha^2\right) = \left(\frac{1}{2}, \frac{1}{2}\right)$, W and β

1: Compute $\mathbf{M}$ st. $\mathbf{M}_{1,2} = (\nabla_\beta L_1(\beta))^\top (\nabla_\beta L_2(W, \beta))$, $\mathbf{M}_{2,1} = (\nabla_\beta L_2(W, \beta))^\top (\nabla_\beta L_1(\beta))$
2: Compute $\hat{t} = \arg\min_r \sum_t \alpha^t \mathbf{M}_{rt}$
3: Compute $\hat{\gamma} = \arg\min_\gamma \left((1-\gamma)\boldsymbol{\alpha} + \gamma \boldsymbol{e}_{\hat{t}}\right)^\top \mathbf{M} \left((1-\gamma)\boldsymbol{\alpha} + \gamma \boldsymbol{e}_{\hat{t}}\right)$
4: $\boldsymbol{\alpha}^* = (1-\hat{\gamma})\boldsymbol{\alpha} + \hat{\gamma}\boldsymbol{e}_{\hat{t}}$

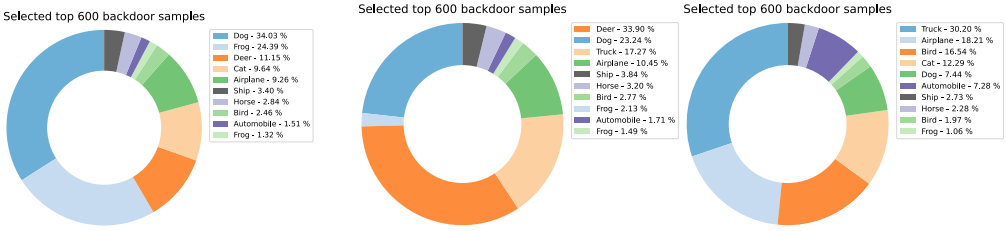
Output: $\boldsymbol{\alpha}^*$

3 Complementary Experimental Results

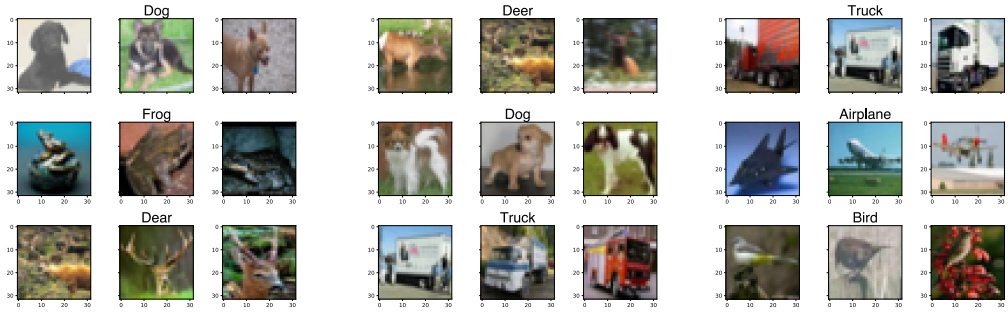
We provide detailed experimental results that are omitted in the main text due to the page limit in this section.

3.1 Selected Samples for CIFAR10

For task II, we demonstrate the categories of the top 1% selected backdoor training samples in Fig. 1a, and several images of the top-3-category selected training backdoor samples associated with the backdoor label 'Cat', 'Deer' and 'Truck' in Fig. 1b. Similar to the task I in the main text, the top-3-category backdoor samples are semantically consistent with their target labels. For example, the images of 'Deer' and 'Dog' resemble the images of 'Horse' most than any other categories in the Fashion-MNIST dataset. Such observations provide concrete evidence to support the conjecture that backdoor images should be similar to their backdoor label category (Bagdasaryan et al., 2020). The similarity is measured in terms of semantic meanings in our case.



(a) Pie-charts of selected backdoor samples corresponding to backdoor label Cat (left), Horse (middle) and Automobile (right).



(b) Snapshots of top-3-category selected backdoor samples corresponding to backdoor label Cat (left), Horse (middle) and Automobile (right).

Figure 1: Illustration of (a) selected categories of backdoor samples in pie chart and (b) selected examples for CIFAR10 dataset.

3.2 TEST PERFORMANCES ON GTSRB

In this section, we test the proposed method on the GTSRTB dataset. There are in total 43 types of traffic signs (labels) in GTSRB, and many of them are of the same type, e.g., "20 speed","30 speed". For the sake of clarity, we only select one of each types for testing.

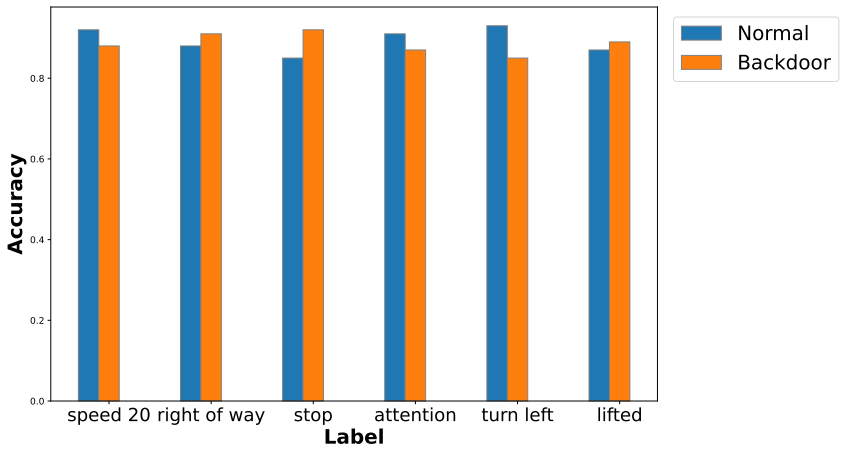


Figure 2: Accuracy of proposed method on GTSRB

3.3 COMPARISON WITH STATE-OF-THE-ART ON CIFAR10

For semantic backdoor attacks, following the framework in (Bagdasaryan et al., 2020), we relabel a whole category of the training data with the same semantic meaning. For example, we can relabel images of Top (with ground-truth label 1) with label 0. For edge-case attacks, we follow the ideas in (Wang et al., 2020) to first create a new training dataset by excluding the samples of one whole category. Then, we relabel the samples of the previously excluded sub-category and added them to the previous training data to form a new training data.

We follow the same setup for **Task II** in the main text. The two SOTA methods to be compared are listed below. For semantic attacks, the work of (Bagdasaryan et al., 2020) relabeled a whole category of the training data. For edge-case attacks, the authors in (Wang et al., 2020) first relabeled the images of Southwest Airplanes (NOT in CIFAR10) as Truck and then injected the relabeled sample-label pairs into the training data. For completeness, we also test for different backdoor labels as summarized in Table 1. The proposed DLP consistently outperforms the SOTA clean-sample attack for all backdoor target labels. Also, the proposed DLP is comparable with the more powerful edge-case attack in some cases, e.g., a backdoor label of 'Dog'.

Table 1: Test Accuracy (in %) on CIFAR10

Method	**DLP**		**Semantic**		**Edge-case**	
Accuracy	**AccN**	**AccB**	**AccN**	**AccB**	**AccN**	**AccB**
Label: Frog	84.2	85.7	83.2	81.6	85.1	84.8
Label: Truck	86.9	87.2	79.3	82.9	86.5	89.2
Label: Dog	83.5	89.3	82.1	87.3	88.5	90.9
Label: Horse	84.1	89.2	83.2	81.5	87.5	91.1

3.4 EXPERIMENTAL RESULTS UNDER DEFENSE MECHANISMS

In this section, we test the proposed method under certain defenses. As mentioned in the main text, because of the weak threat model of (centralized) clean-sample backdoor attacks, many existing defenses against perturbation-based attacks are inappropriate for defending against our attacks.

But defenses against label flipping attacks, e.g., label sanitization (Paudice et al., 2018) and label cerition (Rosenfeld et al., 2020) can be suitably tailored to defend against our method. The results are summarized in Fig. 3 and Fig. 4 respectively. It can be concluded from the figure that the proposed method can escape the two defenses.

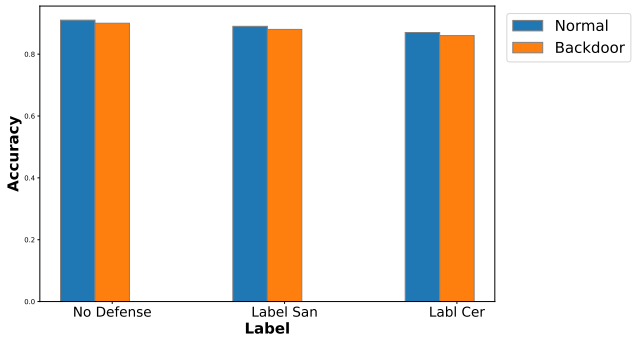


Figure 3: Test performance on Fashion-MNIST

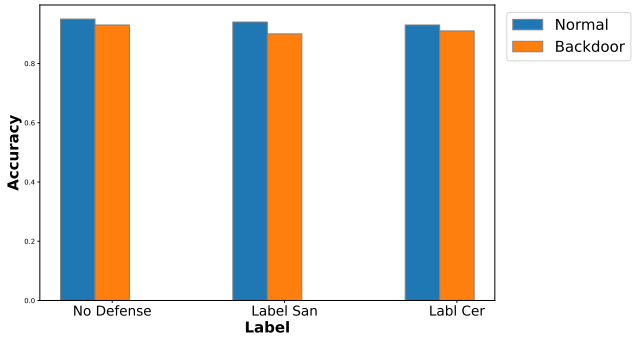


Figure 4: Test performance on CIFAR10

REFERENCES

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948. PMLR, 2020.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

Andrea Paudice, Luis Muñoz-González, and Emil C Lupu. Label sanitization against label flipping poisoning attacks. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 5–15. Springer, 2018.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pp. 8230–8241. PMLR, 2020.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *arXiv preprint arXiv:2007.05084*, 2020.

Di Zhou, Jun Wang, Bin Jiang, Hua Guo, and Yajun Li. Multi-task multi-view learning based on cooperative multi-objective optimization. *IEEE Access*, 6:19465–19477, 2017.