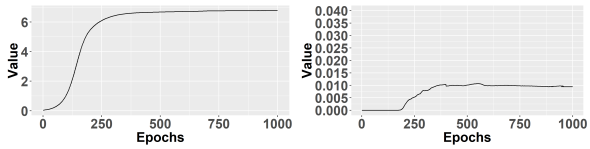
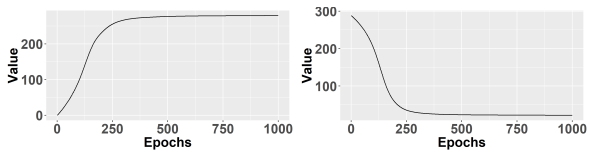


(a) Regarding the first term: $||x - b||_2^2$

(b) Regarding the second term: $\alpha \cdot ||f(x) - 100\%||^2$



(c) Regarding the third term: $\alpha \cdot ||f^{'}(x) - f^{'}(b)||_2^2$

(d) Regarding the fourth term: $\alpha \cdot ||f^{'}(x) - f^{'}(t)||_2^2$

Fig. 1: The value changing pattern of the four terms in Eq. (1).

Appendix Item 1: Empirical Observation on Optimizing Eq. (1) using Algorithm 1 (Sec.3)

As to be seen in Sec. 4, the effectiveness of applying Algorithm 1 to optimize Eq. (1) has been proved by extensive evaluation results. To help understand how applying Algorithm 1 optimizes the four terms included in Eq. (1), we now show a sample experiment to illustrate the typical value changing pattern of each term in Eq. (1) due to this optimization.

In this experiment, we randomly chose a sample from the mnist dataset as the base instance **b**, and use a blackcard as the target instance **t**. We then perform optimization using Algorithm 1 with a α value of set to be 3.125. We recorded the changing values of the four terms in Eq. 1 during the optimization process.

Fig. 1 shows the typical value changing pattern of the four terms in Eq. (1) when performing the optimization (note that we observe the similar changing pattern of these terms in other experiments using various input data). As seen in Fig. 1, as expected, the third and fourth terms in Eq. (1) gradually increase and decrease, respectively. Specifically, the third term increases from 0 to 279.32, ensuring that the feature space representation of **b** and **x** will not be collided under model **P**. The fourth term decreases from 296 to 11, representing that the feature spaces of **b** and **t** stay close under **P**. For the second term, its value actually increases a little from 0 to 0.0096 due to exploring optimization tradeoff among multiple terms by Eq. (1). Nonetheless, this small value of merely 0.0096 still implies that **x** would be classified as **b** with very high confidence (although not 100%) under model **A** after optimization. Also note that the value of the first term increases a little from 0 to 6.67, which is to ensure the poison instance **x** to appear like the base class instance **b** to a human labeler. Due to perturbation added to the input image, the first term will have to increase. Nonetheless, according to [4] and our

evaluation results, such increase is sufficiently small which guarantees that the generated poisoned data is visually-indistinguishable to human labelers.

Appendix Item 2: Detailed information about dataset, complexity, and model architecture of each task

Task	Dataset	Labels	Input Size	# of Training Images	Architecture of **P**	Architecture of **T**
Hand-written Digit Recognition	MNIST	10	28*28*1	60000	4 Conv + 1 Dense	LeNet-5 [5]
Fashion Item recognition	Fashion-MNIST	10	28*28*1	60000	4 Conv + 1 Dense	State-of-the-art [1]
Object Recognition in Images	CIFAR-10	10	32*32*3	50000	DenseNet121	RESNetv2.56
Traffic Sign Recognition	GTSRB	43	32*32*3	35288	6 Conv + 2 Dense	State-of-the-art [2]
Face Recognition using VGG Face	VGG Face dataset	2622	224*224*3	2622000	12 Conv+3 Dense	VGGNet16 [6]
Face Recognition using Asian Face	$CIASIA_V5Dataset$	500	224*224*3	2500	4 Conv + 2 Dense	State-of-the-art [3]

Table 1: Detailed information about the dataset, complexity, and model architecture used in the evaluation.

References

1. https://www.kaggle.com/anebzt/fashion-mnist-in-keras
2. https://github.com/bolunwang/backdoor
3. https://github.com/kongzelun/AsianFace.git
4. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE (2017)
5. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)