Appendix

1. Neural Network Architectures

1.1. LeNet

The pre-trained LeNet architecture for the single instance attack on the MNIST dataset is summarized in Table 3. Before the poisoning attack, the network is trained with the Adam optimizer at the learning rate of 1×10^{-4} for 80 epochs on clean data and achieves a test accuracy of 99%.

Table 3. Network Architecture for the MNIST experiments

Layer		Number of Filter	Size	Kernel Size	Stride	Activation Function
Input	Image	-	32x32x1	-	-	-
1	Convolution	6	28x28	5x5	1	ReLU
2	Max Pooling	6	14x14	2x2	2	-
3	Convolution	16	10x10	5x5	1	ReLU
4	Max Pooling	16	5x5	2x2	2	-
5	FC	-	120	-	-	ReLU
6	FC	-	84	-	-	ReLU
Output	FC	-	10	-	-	Softmax

1.2. VggNet

We evaluate the effect of our proposed attacks on CIFAR-10 using the Vgg model. The detailed pre-trained Vgg network architecture is shown in Table 4. It achieves a test accuracy of 81.05% without poisoning. The model is trained with the Momentum optimizer with 0.9 momentum for 250 epochs. The learning rate starts at 0.01 and is scheduled with a decay rate of 0.5 every 25 epochs.

Table 4. Network Architecture for the CIFAR10 experiments

Layer		Number of Filter	Size	Kernel Size	Stride	Activation Function
Input	Image	-	32x32x3	-	-	-
1	Convolution	64	32x32	3x3	1	ReLU
2	Convolution	64	32x32	3x3	1	ReLU
3	Max Pooling	64	16x16	2x2	2	-
4	Convolution	128	16x16	3x3	1	ReLU
5	Covnvolution	128	16x16	3x3	1	ReLU
6	Max Pooling	128	8x8	2x2	2	-
7	Convolution	256	8x8	3x3	1	ReLU
8	Convolution	256	8x8	3x3	1	ReLU
9	Max Pooling	256	4x4	2x2	2	-
10	FC	-	1024	-	-	ReLU
11	FC	-	180	-	-	ReLU
Output	FC	-	10	-	-	Softmax

1.3. ResNet

For experiments on ImageNet, we use a pre-trained ResNet-50 model[1]. The benign model has a Top-1 accuracy of 74.87% and Top-5 accuracy of 92.02%, respectively. The training hyper-parameters and learning rate decay strategy follow the same settings as in the original paper [17].

2. Additional Experiments of COEG Attack

To have a comprehensive understanding of the effectiveness of COEG attack, we perform more in-depth experiments and analysis for both single instance attack and attack with a set of poisoned data.

2.1. Single instance attack

2.1.1 Different class pairs.

Recent research [36] found that the effect of the poisoning integrity attacks (a.k.a backdoor) is highly dependent on class pairs and the choice of seed images. For instance, the same seed image with different poisoned labels (seed-target pairs) will result in a dramatically inconsistent poisoning effect. Meanwhile, different seed images with the same poisoned label will also lead to varied performances. Note that a different poisoned label means a different supplanter class in the COEG attack. To study whether the issue also exists in the availability poisoning attack, we perform two experiments for the single instance attack: i) we pick a different seed image from the previous experiment on MNIST and assign the same poisoned labels; ii) we pick the same seed image from CIFAR-10 and assign different poisoned labels. We keep the same training setting and baseline attacks as the previous experiments. The results of the first experiment are presented in Table 5. We find a similar trend that different labels yield different poisoning effect, although the difference is not as large as in poisoning integrity attacks [36]. As shown, our proposed attack still significantly outperform the baseline attack in test error and CTT rate.

The results of the second experiment are presented in Figure 6 and Figure 7. Although some of the class pairs perform a bit worse than others (i.e., frog-cat for the baseline attack, frog-horse for our proposed attack), both the FL

[1] https://github.com/keras-team/keras

Table 5. Comparison of different seed images with the same poisoned label.

Attack (seed-label pairs)	Test Error	CTT Rate
FL attack (seed:1-label:4)	27.59%	27.33%
Proposed attack (1-4)	71.11%	79.90%
FL attack (seed:6-label:4)	19.00%	32.85%
Proposed attack (6-4)	85.50%	85.85%

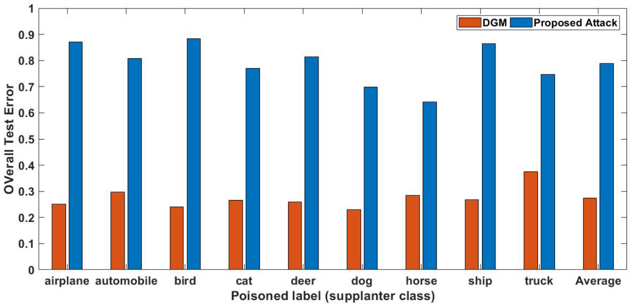

Figure 6. Test error with different seed-target pairs on CIFAR-10 for the single instance attack.

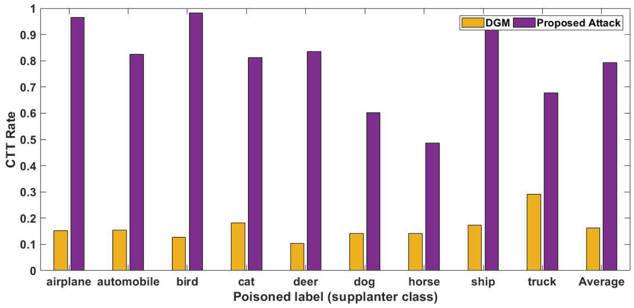

Figure 7. CTT rate with different seed-target pairs on CIFAR-10 for the single instance attack.

attack and our proposed attack show a relatively consistent performance with different poisoned labels. The proposed COEG attack achieves an test error of 78.87% and CTT rate of 79.20% on average, which significantly outperforms the baseline attack with an average test error of 27.44% and CTT rate of 16.30%. Note that the CTT of our approach is 65.9% higher than that of the baseline attack, which exceeds the performance difference of test error. This fact further validates that our proposed attack is particularly effective as class-oriented poisoning attacks.

2.1.2 Attack with more images.

While lower learning rates can reduce the effect of poisoning attacks, the limited number of poisoned data may also be a factor. To better understand the limitation of the single instance attack, we keep the lower learning rate and increase the number of poisoned data by duplicating the single poisoned sample. The results are summarized in Table 6. With the increase of poisoned data, the overall test error and CTT rate of the baseline attack remain nearly the same while ours even slightly drop by 13% and 7%, respectively. These results reveals that simply increasing the number of poisoned data by duplication will not improve the poisoning effect. Thus, the diversity of the poisoned dataset is crucial.

Table 6. Comparison of poisoning effect with different numbers of poisoned data.

Number of	Test Error		CTT Rate	
poisoned data	DGM	Ours	DGM	Ours
1	23.4%	54.8%	11.52%	23.32%
100	22.9%	41.1%	10.7%	16.8%
500	22.6%	41.1%	10.6%	16.5%

2.2. Attack with a set of poisoned data

2.2.1 Different target supplanter classes.

We study the impact of different target supplanter classes for the general poisoning attack, where a set of poisoned data is injected. This setting is slightly different from the single instance attack as the poisoned seed images are arbitrarily selected from all classes. We assign different targeted labels (supplanter classes) to the poisoned data and compare the performance between the FL attack and our proposed attack. The test error and CTT rate are presented in Figure 8 and Figure 9. We use $\alpha = 0.5$ and a learning rate of 1×10^{-5}. Similar to the single instance attack, our proposed attack is resilient to the variation of poisoned labels and achieves the class-oriented adversarial goal.

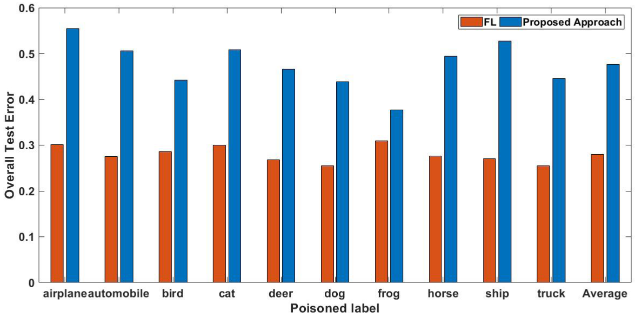

Figure 8. Test error with different target labels on CIFAR10 for the attack with a set of poisoned data.

2.2.2 Different dataset sizes.

Another interesting finding in paper [36] is that the effect of poisoning attacks is related to the size of the retraining dataset. For instance, with a fixed percentage of poisoned data, some attacks achieve better poisoning effect on a larger retraining dataset, while others perform worse. The impact

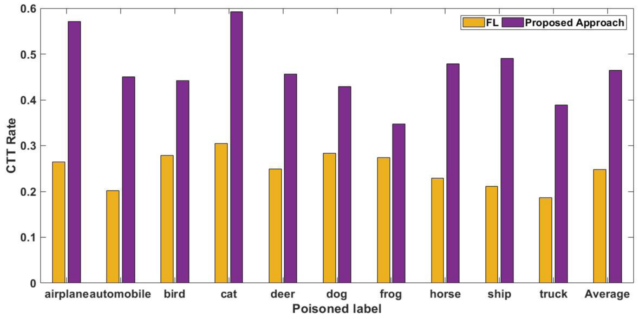

Figure 9. CTT rate with different target labels on CIFAR10 for the attack with a set of poisoned data.

of dataset size is particularly worth studying in the scenario of end-to-end fine-tuning on a pre-trained model, where retraining data are scarce. We test three different sizes with the same percentage of poisoned data for each size. We use the same training setting as above. The results are presented in Tables 7 and 8.

It can be seen that a larger dataset size always yields better poisoning performance at all percentage levels for the baseline FL attack. However, our proposed attack has the same trend when the percentage of poisoned data is small and shows superior performance on the dataset size of 500. When the dataset size is small, the FL attack is barely effective with any percentage of poisoned data, while the proposed attack achieves improved test error and CTT rate. Moreover, the performance improvement of FL is much smaller than the proposed approach with the increase of dataset size. These observations provide sufficient evidence that our proposed attack is much more effective than the baseline attack and more elastic to dataset size change.

Table 7. Comparison of test error with different dataset sizes (lr = 1×10^{-5}).

Test Error	Training dataset size = 100		Training dataset size = 500		Training dataset size = 1000	
	FL	**Ours**	FL	**Ours**	FL	**Ours**
$\alpha = 0.1$	18.95%	**18.97%**	19.21%	**23.84%**	19.65%	**36.38%**
$\alpha = 0.2$	18.92%	**19.29%**	19.54%	**37.74%**	20.84%	**47.00%**
$\alpha = 0.3$	18.79%	**19.64%**	20.37%	**54.49%**	23.05%	**49.53%**
$\alpha = 0.4$	18.92%	**20.70%**	21.20%	**64.23%**	26.10%	**51.84%**
$\alpha = 0.5$	18.99%	**22.27%**	22.41%	**70.15%**	30.14%	**55.42%**

Table 8. Comparison of CTT rate with different dataset sizes (lr = 1×10^{-5}).

CTT Rate	Training dataset size = 100		Training dataset size = 500		Training dataset size = 1000	
	FL	**Ours**	FL	**Ours**	FL	**Ours**
$\alpha = 0.1$	10.35%	**10.76%**	11.22%	**16.00%**	12.11%	**31.25%**
$\alpha = 0.2$	10.50%	**11.46%**	12.20%	**33.46%**	14.46%	**46.70%**
$\alpha = 0.3$	10.06%	**12.25%**	13.64%	**56.83%**	17.48%	**49.75%**
$\alpha = 0.4$	10.80%	**13.28%**	14.83%	**69.46%**	21.4%	**52.43%**
$\alpha = 0.5$	10.98%	**14.88%**	16.57%	**76.75%**	26.45%	**57.11%**

3. Additional Experiments of COES Attack

As discussed in the main manuscript, the adversarial goal for the COES attack is more challenging than COEG. We also extensively study the effectiveness of our proposed approach on the COES attack with more experiments.

3.1. Different values of α.

We first study the impact of various α values on the COES attack. We keep the same training setting as in previous experiments. The results are presented in Table 9. As expected, we have a similar finding as the COEG attack that the poisoning effect is proportional to the number of injected poisoned data. On the other hand, we find that for all different α, FL-1 always performs badly in reducing the CFT rates of non-victim classes, while FL-2 reduces the CFT rates of non-victim classes at the cost of lowering the CFT rate of the victim class at the same time. Our proposed approach addresses for the shortcomings of both FL-1 and FL-2, achieving a high CFT rate for the victim class while keeping low CFT rates for non-victim classes.

3.2. Different learning rates.

We present the results of different learning rates of the COES attack in Table 10. We keep the dataset size at 1000 and set the α value to 0.5. Our proposed approach consistently outperforms the baseline attacks. With the increase of the learning rate, the CFT rates of the victim class increase for all three attacks. Interestingly, our proposed attack achieves the best CFT rate of the victim class at a learning rate of 4×10^{-5} other than a higher learning rate. In comparison, with the decrease of the learning rate, FL-1 and FL-2 suffer drastic performance drops. Thus, it can be concluded that the proposed approach is more resilient to the variation of the learning rate for the COES attack.

3.3. Different training dataset sizes.

Lastly, we evaluate the impact of different training dataset sizes. We keep the learning rate at 1×10^{-5} and α value at 0.5 for the experiment and present the results in Table 11. Unlike the COEG attack that achieves a better CTT performance with a dataset size of 500, a larger dataset size in COES attack always yields a better CFT rate of the victim class without significantly affecting the CFT rates of non-victim classes. Note that the poisoning effect is greatly reduced by scaling down the training dataset size. This further shows the challenging adversarial goal of the COES attack as we need to manipulate the performance of all classes with limited poisoned data. However, even training on an extremely small dataset, we can still achieve nearly 5 times better performance than the FL-2 attack, which is a clear showcase of the effectiveness of our proposed attacks.

Table 9. CFT rate of each CIFAR-10 class by poisoning with different α at learning rate 1×10^{-5}.

Fraction of poisoned data	Attacks	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck (victim)
$\alpha = 0.1$	FL-1	-5.57%	-0.99%	-0.27%	-0.65%	5.53%	2.21%	3.53%	-0.69%	1.23%	2.75%
	FL-2	-1.62%	-0.99%	-0.68%	-1.47%	2.19%	0.55%	0.35%	-0.69%	-0.78%	1.83%
	Ours	-3.48%	-1.99%	-0.82%	-3.11%	3.47%	1.11%	-0.35%	-0.23%	-0.67%	7.89%
$\alpha = 0.2$	FL-1	-9.27%	-0.44%	2.73%	0.49%	11.05%	4.01%	6.35%	-0.34%	4.37%	3.66%
	FL-2	-2.20%	-0.88%	-1.50%	-1.47%	1.67%	1.24%	-0.12%	-0.69%	-1.23%	3.20%
	Ours	-4.17%	-1.10%	-0.82%	-3.11%	3.37%	2.21%	-0.12%	0.23%	-0.56%	15.79%
$\alpha = 0.3$	FL-1	-11.24%	0.11%	5.05%	3.27%	17.10%	7.88%	9.75%	1.49%	7.05%	8.01%
	FL-2	-2.67%	-1.10%	-1.91%	-1.47%	2.57%	1.66%	-0.24%	-0.57%	-1.01%	5.03%
	Ours	-5.01%	-1.66%	-1.23%	-2.13%	5.91%	3.04%	0.24%	0.91%	-0.67%	28.60%
$\alpha = 0.4$	FL-1	-12.98%	0.55%	10.52%	6.71%	24.29%	11.20%	17.16%	4.46%	11.09%	12.01%
	FL-2	-3.13%	-1.10%	-1.50%	-2.29%	2.44%	2.77%	0.00%	-0.67%	-1.01%	5.84%
	Ours	-6.37%	-0.33%	-0.27%	-0.49%	6.43%	3.32%	0.47%	0.69%	-0.67%	38.10%
$\alpha = 0.5$	FL-1	-13.09%	3.87%	16.80%	16.20%	32.65%	19.09%	22.44%	8.34%	15.45%	18.42%
	FL-2	-3.01%	-1.44%	-2.05%	-2.29%	2.57%	2.67%	-0.59%	-0.11%	-0.78%	8.01%
	Ours	-7.07%	0.88%	-0.41%	2.29%	5.27%	3.32%	-0.47%	0.11%	0.22%	51.14%

Table 10. CFT rate of each CIFAR-10 class by poisoning with different learning rates when $\alpha = 0.5$.

Learning Rate	Attacks	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck (victim)
1×10^{-4}	FL-1	-14.60%	26.85%	45.49%	37.15%	48.97%	46.47%	46.65%	36.80%	43.23%	53.66%
	FL-2	-6.84%	0.55%	0.82%	-6.55%	3.21%	-0.55%	-2.82%	-0.34%	0.90%	41.30%
	Ours	-1.85%	-1.22%	3.28%	-6.71%	2.70%	-0.69%	-1.06%	2.40%	-2.35%	65.90%
4×10^{-5}	FL-1	-14.83%	22.54%	39.21%	39.28%	54.24%	39.42%	49.24%	31.66%	40.76%	48.86%
	FL-2	-6.72%	0.33%	-1.23%	-3.93%	4.24%	3.04%	-0.65%	-0.34%	1.34%	25.51%
	Ours	-3.01%	0.88%	1.91%	-2.45%	3.47%	-0.28%	0.24%	2.29%	-2.02%	66.59%
5×10^{-6}	FL-1	-10.66%	-0.22%	4.37%	1.31%	15.55%	7.47%	8.70%	1.37%	5.38%	7.21%
	FL-2	-1.74%	-0.88%	-2.05%	-1.31%	1.67%	1.52%	0.12%	-0.46%	-0.78%	4.46%
	Ours	-5.10%	-1.10%	-2.19%	-0.16%	3.21%	3.04%	0.12%	0.80%	-0.90%	27.80%

Table 11. CFT rate of each CIFAR-10 class by poisoning with different training dataset sizes.

Dataset size	Attacks	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck (victim)
100	FL-1	-2.90%	-0.55%	-1.91%	-1.96%	2.96%	1.80%	1.76%	-0.69%	-0.67%	1.83%
	FL-2	-0.93%	-0.33%	-0.68%	-0.16%	0.51%	0.55%	0.47%	-0.34%	-0.34%	1.14%
	Ours	-1.97%	0.00%	-1.23%	-1.64%	1.41%	1.94%	-0.24%	-0.11%	-0.67%	4.92%
500	FL-1	-10.66%	-0.33%	3.69%	1.96%	14.65%	8.30%	8.81%	1.03%	5.28%	7.09%
	FL-2	-2.67%	-0.66%	-2.19%	-1.15%	1.54%	3.18%	0.59%	-0.69%	-0.78%	4.35%
	Ours	-4.75%	-0.88%	-1.50%	-3.11%	3.73%	3.32%	0.94%	0.69%	-0.67%	24.03%
1000	FL-1	-13.09%	3.87%	16.80%	16.20%	32.65%	19.09%	22.44%	8.34%	15.45%	18.42%
	FL-2	-3.01%	-1.44%	-2.05%	-2.29%	2.57%	2.67%	-0.59%	-0.11%	-0.78%	8.01%
	Ours	-7.07%	0.88%	-0.41%	2.29%	5.27%	3.32%	-0.47%	0.11%	0.22%	51.14%