Efficient Model Stealing Defense with Noise Transition Matrix

Appendix

7. Proof of Theorem 1.

In the label space $\mathbf{Y}$, let $\tilde{y}$ denote the noisy label, $\hat{P}(\cdot|\boldsymbol{x})$ and $P(\cdot|\boldsymbol{x})$ denote the true posteriors and predicted posteriors, respectively. $P(\tilde{y}|y)$ means the probability of y flipping to $\tilde{y}$ as defined in the noise transition matrix. Here we regard $f'(\boldsymbol{x})$ and $h(\boldsymbol{x})$ as vectors, representing the noise-perturbed posteriors of the victim model and the attack's predictive posteriors, respectively. Let $\mathcal{L}(a, b)$ and $\mathcal{I}(a, c)$ denote the same classification loss function over the label b and the vector c respectively, with a as the prediction vector. Note that $\mathcal{L}(\cdot, \cdot)$ and $\mathcal{I}(\cdot, \cdot)$ can be loss function, such as cross-entropy loss, mean absolute error, or other loss functions.

The classification risk of learning with noisy label and learning with query-response pairs with perturbed posteriors equals to Eq. (14) and Eq. (15), respectively,

$$\mathfrak{R}(h) = \mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\mathcal{L}(h(\boldsymbol{x}), \tilde{y})], \tag{14}$$

$$\mathfrak{R}_T(h) = \mathbb{E}_{(\boldsymbol{x},\tilde{y})}[q(\boldsymbol{x}, \tilde{y}) \cdot \mathcal{L}(h(\boldsymbol{x}), \tilde{y})], \tag{15}$$

where $q(\boldsymbol{x}, \tilde{y}) = P(\tilde{y}|\boldsymbol{x})/\hat{P}(\tilde{y}|\boldsymbol{x})$.

To prove this, we first rewrite the classification risk $\mathfrak{R}(h)$ as follows:

$$\begin{aligned} \mathfrak{R}(h) &= \mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\mathcal{L}(h(\boldsymbol{x}), \tilde{y})] \\ &= \sum_{\tilde{y}\in\mathbf{Y}} P(\tilde{y}) \mathbb{E}_{(\boldsymbol{x}|\tilde{y})}[\mathcal{L}(h(\boldsymbol{x}), \tilde{y})], \end{aligned} \tag{16}$$

where $P(\tilde{y})$ can be further calculated by

$$P(\tilde{y}) = \mathbb{E}_{(\boldsymbol{x}|\tilde{y})}[\sum_{y\in Y} \hat{P}(y|\boldsymbol{x}) P(\tilde{y}|y)]. \tag{17}$$

By substituting Eq. (17) into Eq. (16), we obtain

$$\mathfrak{R}(h) = \sum_{\tilde{y}\in\mathbf{Y}} \mathbb{E}_{(\boldsymbol{x}|\tilde{y})}[\sum_{y\in Y} \hat{P}(y|\boldsymbol{x}) P(\tilde{y}|y) \cdot \mathcal{L}(h(\boldsymbol{x}), \tilde{y})]. \tag{18}$$

Correspondingly, the classification risk $\mathcal{R}_T(h)$ can be expressed as:

$$\begin{aligned} \mathfrak{R}_T(h) &= \mathbb{E}_{(\boldsymbol{x})}[\mathcal{I}(h(\boldsymbol{x}), f'(\boldsymbol{x}))] \\ &= \mathbb{E}_{(\boldsymbol{x})}[\sum_{\tilde{y}\in\mathbf{Y}} P(\tilde{y}|\boldsymbol{x}) \cdot \mathcal{L}(h(\boldsymbol{x}), \tilde{y})] \\ &= \sum_{\tilde{y}\in\mathbf{Y}} \mathbb{E}_{(\boldsymbol{x}|\tilde{y})}[P(\tilde{y}|\boldsymbol{x}) \cdot \mathcal{L}(h(\boldsymbol{x}), \tilde{y})] \\ &= \sum_{\tilde{y}\in\mathbf{Y}} \mathbb{E}_{(\boldsymbol{x}|\tilde{y})}[\frac{P(\tilde{y}|\boldsymbol{x}) \cdot \sum_{y\in Y} \hat{P}(y|\boldsymbol{x}) P(\tilde{y}|y)}{\sum_{y\in Y} \hat{P}(y|\boldsymbol{x}) P(\tilde{y}|y)} \cdot \mathcal{L}(h(\boldsymbol{x}), \tilde{y})] \\ &= \mathbb{E}_{(\boldsymbol{x},\tilde{y})}[q(\boldsymbol{x}, \tilde{y}) \cdot \mathcal{L}(h(\boldsymbol{x}), \tilde{y})], \end{aligned}$$

where $q(\boldsymbol{x}, \tilde{y}) = \frac{\sum_{y\in Y} P(y|\boldsymbol{x}) P(\tilde{y}|y)}{\sum_{y\in Y} \hat{P}(y|\boldsymbol{x}) P(\tilde{y}|y)} = \frac{P(\tilde{y}|x)}{\hat{P}(\tilde{y}|x)}$. Here $\hat{P}(\tilde{y}|\boldsymbol{x})$ and $P(\tilde{y}|\boldsymbol{x})$ represent the perturbed confidences of the true and predicted posteriors on the noise label $\tilde{y}$. The above derivation from the penultimate step to the final step utilizes the previously established equivalence between Equation (18) and Equation (16). The term $q(\boldsymbol{x}, \tilde{y})$ can also be viewed as an importance sampling weight between two noise distributions - one generated from the true posteriors with noise perturbation, and the other from the victim model posteriors with noise perturbation.

Furthermore, it can be deduced that the more closely the victim model's predicted posteriors match the true posteriors, the more this training process resembles learning with noisy labels.

8. Tailored query sets description

In the distribution-aware attack scenario, we use tailored query sets ImageNet-C10, ImageNet-C100 and ImageNet-CUB200 that match the evaluation datasets. Specifically, these tailored sets are constructed from ImageNet-1K by manually selecting overlapping classes. ImageNet-C10, ImageNet-C100 and ImageNet-CUB200 contain 183,763, 161,653 and 30,000 examples respectively. Notably, ImageNet-C10 and ImageNet-C100 query sets exhibit a long-tailed class distribution, as illustrated in the Figure 8.

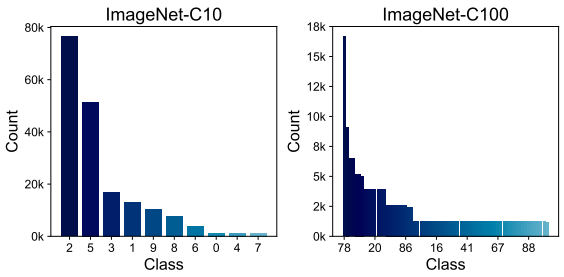

Figure 7. Class distribution of ImageNet-C10 and ImageNet-C100.

Since CUB200 is a fine-grained dataset on bird species, ImageNet-CUB200 only contains a bird class from ImageNet-1K.

9. Visualization of comparison results

For most defense budgets, our proposed method induces lower classification accuracy in the attacker compared to the best baseline methods, achieving the minimal trade-off between attacker performance and defender utility.

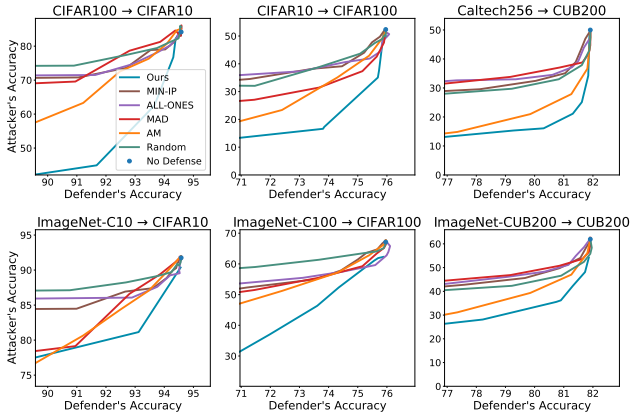

Figure 8. Visualization results of the baseline method at different defense budgets across six attack scenarios.

10. Robustness across architectures

An important consideration for our defense is their robustness when the attacker uses a different network than the surrogate model. In Figure 9, we keep the ResNet34 surrogate and posterior perturbations but use DenseNet121 for the attacker model. We find EMMA remains a strong defense in this setting, with minimal decrease in performance slope, indicating perturbations crafted on the surrogate transfer effectively to the attacker network.

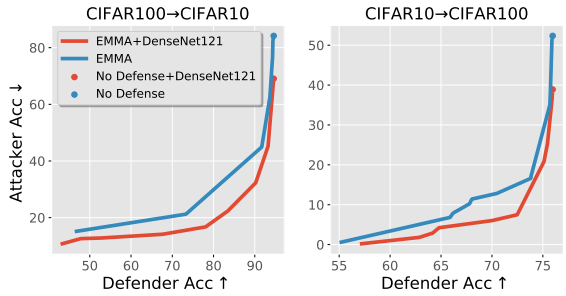

Figure 9. Robustness across architecture.

11. Data-free model stealing defense

In practical scenarios, it is common for companies to train their models using proprietary datasets, which are not readily available to potential attackers. In such cases, the attackers may resort to data-free model stealing attacks, where they attempt to extract the model's knowledge without accessing the actual dataset. Therefore, we also conduct experiments to evaluate the effectiveness of our method EMMA in defending against such attacks. In Table 4, the performance of EMMA was evaluated against two data-free model stealing attacks: DFMS [44] and MAZE [22] (with ΔAcc = 2). The results demonstrate that EMMA can still effectively defend against these data-free model stealing attacks.

Eval Data	Victim Acc	DFMS	DFMS+EMMA	MAZE	MAZE+EMMA
CIFAR10	95.33	90.56	**65.18**	90.23	**73.18**
CIFAR100	75.96	67.83	**39.53**	69.37	**48.09**
CUB200	81.44	64.15	**35.70**	68.92	**39.91**

Table 4. Attacker accuracy (%) with no defense and with EMMA defense.

12. Extension to foundation model defense

To extend our method to other foundation models like image encoder, we first replace the to-be-optimized matrix with a trainable adapter network (MLP), that transforms the original outputting feature embedding to a perturbed one, then modify the attacker and defender objective functions based on similarity loss, and finally perform alternate optimization to obtain the optimized MLP. To validate our extended method's effectiveness, we took CSTEAL [45] as a baseline and followed its experimental setup and pipeline. The results in Table 5 show EMMA still works and effectively defends against stealing attacks by slightly sacrificing victim model performance.

	Victim	CSTEAL	Victim+EMMA	CSTEAL+EMMA
F-MNIST	83.81±0.12	68.02±0.20	**81.26±0.31**	**56.14±0.52**
CIFAR10	88.47±0.19	84.29±0.13	**85.24±0.70**	**71.81±0.35**
STL10	77.61±0.39	72.10±0.14	**74.09±1.21**	**60.03±0.82**

Table 5. Accuracy (%) of victim and attacker on downstream classification.