Black-Box Dissector: Towards Erasing-based Hard-Label Model Stealing Attack (Supplementary Material)

Yixu Wang[1], Jie Li[1], Hong Liu[2], Yan Wang[3], Yongjian Wu[4], Feiyue Huang[4], and Rongrong Ji[1,5]✉

[1] Media Analytics and Computing Lab, School of Informatics, Xiamen University
[2] National Institute of Informatics
[3] Pinterest
[4] Youtu Lab, Tencent Technology (Shanghai) Co.,Ltd
[5] Institute of Artificial Intelligence, Xiamen University
yxwang79@gmail.com, lijie.32@outlook.com, hliu@nii.ac.jp, yanw@pinterest.com, {littlekenwu,garyhuang}@tencent.com, rrji@xmu.edu.cn

A Appendix

Table I. The performance (Agreement and test accuracy) of previous methods under the soft-label and the hard-label settings. And the average performance reduction of each dataset under hard label is reported in the last row.

Method		CIFAR10		SVHN		Caltech256		CUBS200	
		Agreement	Acc	Agreement	Acc	Agreement	Acc	Agreement	Acc
KnockoffNets	soft-label	81.59%	80.03%	93.17%	92.14%	76.42%	74.42%	65.48%	59.15%
	hard-label	75.32%-6.27%	74.44%-5.59%	85.00%-8.17%	84.50%-7.64%	57.64%-18.78%	55.28%-19.14%	30.01%-35.47%	28.03%-31.12%
ActiveThief(Entropy)	soft-label	81.61%	79.85%	92.79%	91.95%	77.38%	70.91%	68.12%	60.39%
	hard-label	75.26%-6.35%	74.21%-5.64%	90.47%-2.32%	89.85%-2.10%	56.28%-21.10%	54.14%-16.77%	32.05%-36.07%	29.43%-30.96%
ActiveThief(k-Center)	soft-label	82.98%	81.42%	94.45%	93.62%	78.66%	72.20%	73.71%	65.34%
	hard-label	75.71%-7.27%	74.24%-7.18%	81.45%-13.00%	80.79%-12.83%	61.19%-17.47%	58.84%-13.36%	37.68%-36.03%	34.64%-30.70%
ActiveThief(DFAL)	soft-label	80.42%	78.88%	91.41%	90.57%	64.56%	59.81%	53.24%	47.65%
	hard-label	76.72%-3.70%	75.62%-3.26%	84.79%-6.62%	84.17%-6.40%	46.92%-17.64%	44.91%-14.90%	20.31%-32.93%	18.69%-28.96%
ActiveThief(DFAL+k-Center)	soft-label	82.05%	80.86%	93.03%	92.08%	67.27%	62.67%	61.39%	55.18%
	hard-label	74.97%-7.08%	73.98%-6.88%	81.40%-11.63%	80.86%-11.22%	55.70%-11.57%	53.69%-8.98%	26.60%-34.79%	24.42%-30.76%
Average difference		-6.13%	-5.71%	-8.35%	-8.04%	-17.31%	-14.63%	-35.06%	-30.50%

A.1 Gap between hard-label and soft-label setting

Here, we report the numerical results of previous methods under both the soft-label setting and the hard-label setting as a supplementary to the Fig.1. To be consistent with the experiment section, the victim models we use are trained using a ResNet-34 [2] architecture on four datasets: CIFAR10 [4], SVHN [5], Caltech256 [1], and CUBS200 [8]. And their test accuracy are 91.56%, 96.45%, 78.40%, and 77.10% respectively. We use the 1.2M images without labels presented in the ILSVRC-2012 challenge [6] as the attack dataset. We also adopt official source codes from the authors for a fair comparison. As in the Tab. I, the performance of all previous methods has a significant degradation on the four

Table II. Test accuracy of our method and previous methods with different architectures on CIFAR10 dataset. The smaller the standard deviation (Std), the more stable the method.

Method	Substitute's architecture					Std($\times 10^2$)↓
	ResNet-34	ResNet-18	ResNet-50	VGG-16	DenseNet	
KnockoffNets	74.44%	77.12%	66.78%	53.52%	78.50%	9.22
ActiveThief(k-Center)	74.24%	72.90%	71.25%	35.56%	74.48%	15.11
ActiveThief(Entropy)	74.21%	78.77%	73.52%	37.88%	**79.09%**	15.57
Ours	**80.47%**	**79.93%**	**80.34%**	**75.22%**	74.43%	**2.68**

datasets in this scenario, and the averages of the loss are in the last row of the Tab. I, which are 5.71%, 8.04%, 14.63%, and 30.50% respectively. The above results show that in the hard-label scenario, the previous model stealing methods are not effective enough.

A.2 The influence of model architectures

Instead of assuming that the substitute model and the victim one share the same architecture, we show the effect of different model architectures here on the CIFAR10 dataset. Keeping ResNet-34 as the victim model, we choose the structure of substitute model from ResNet-34, ResNet-18, ResNet-50 [2], VGG-16 [7], DenseNet [3], respectively. With the same architecture included, we use the standard deviation to evaluate the impact of architectures on different methods. As in Tab. II, the standard deviation of our method is about 1/6 to 1/3 of others, which means that our method is less susceptible to the influence of the model structure. In real situations, the structure of the victim model is often unknown. Since our method is less affected by the structure, our method performs better in real-world attacks.

A.3 The visualization of the attention alignment.

As we point out in the section 3.1, the novel CAM-driven erasing strategy we designed can not only dig out more class information, but also help the substitute model to align the victim model's attention. As shown in Fig. I, at the beginning time, the substitute model learns the wrong attention map. Along with the iterative training stages, the attention area of the substitute model tends to fit the victim model's, which conforms to our intention. As [9] stated, we transfer the victim's attention to the substitute model, which is one of the reasons why our method is effective enough.

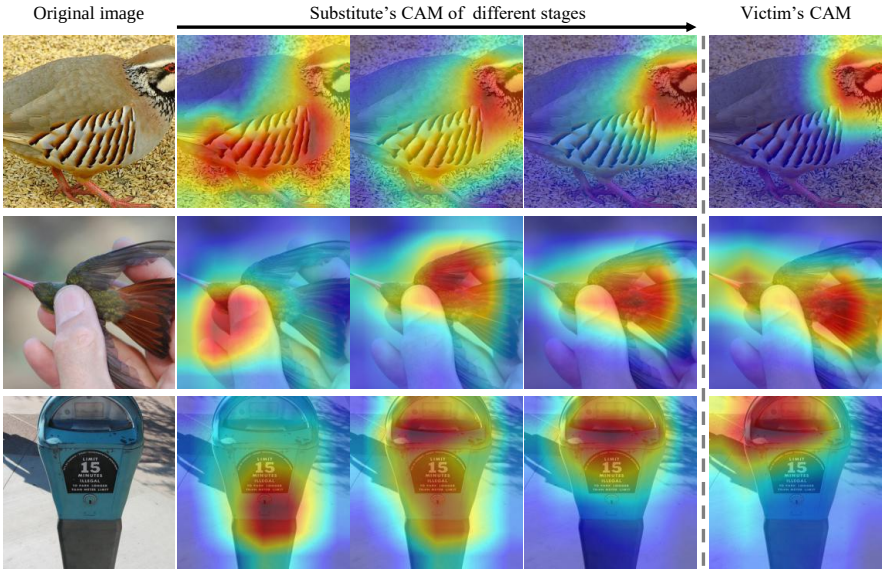

Fig. I. The additional visualized attention maps of the victim model and different stages substitute models using the Grad-CAM. Along with the training stages, the attention map of the substitute model tends to fit the victim model's.

References

1. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
3. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
4. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
5. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
8. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
9. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)