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

Anonymous Author(s)
Affiliation
Address
`email`

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%

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**

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 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.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

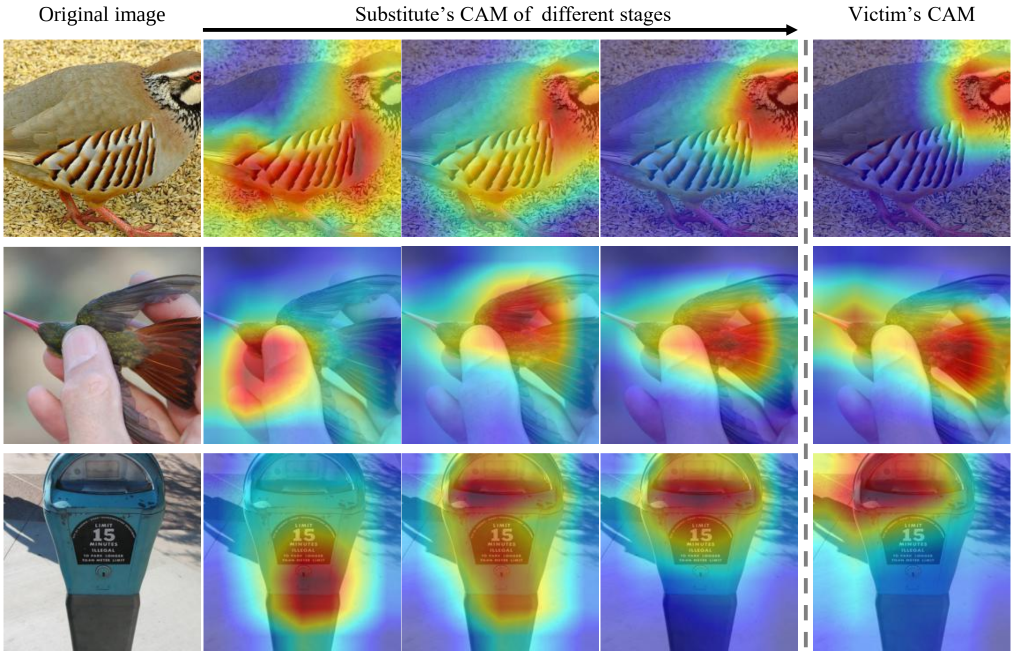

Figure I: The additional visualized attention maps of the victim model and different stages substitute models using the Grad-CAM. Along with the training stages, the attention map of the substitute model tends to fit the victim model's.

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.

References

[1] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[5] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[8] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[9] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.