Towards Data-Free Model Stealing in a Hard Label Setting

Sunandini Sanyal **Sravanti Addepalli** **R. Venkatesh Babu**
Video Analytics Lab, Department of Computational and Data Sciences
Indian Institute of Science, Bangalore

Abstract

Machine learning models as a service(MLaaS) are often susceptible to model stealing attacks. While existing works demonstrate near-perfect performance using softmax predictions of the classification network, most of the APIs allow access to only the top-1 labels. In this work, we show that it is indeed possible to steal Machine Learning models by accessing only top-1 predictions (Hard Label setting), without access to model gradients (Black-Box setting) and even the training dataset (Data-Free setting) within a low query budget. We propose a novel GAN-based framework[1] *that trains the student and generator in tandem to steal the model effectively while utilizing gradients of the clone network as a proxy to the victim's gradients. We propose to overcome the large query costs associated with a typical Data-Free setting by utilizing publicly available (potentially unrelated) datasets as a weak image prior. We additionally show that even in the absence of such data, it is possible to achieve state-of-the-art results within a low query budget using synthetically crafted samples. We are the first to show the scalability of Model Stealing on a 100 class dataset.*

1. Introduction

Deep learning based systems have progressed leaps and bounds over the past few years. Organizations often provide pretrained machine learning models as a service (MLaaS) where the end user is allowed to query the model and get access to its predictions via APIs for use in various applications. However, exposing the predictions of the models through queries makes the model susceptible to model stealing attacks, which attempt to clone the victim model without access to its gradients, in a black-box setting. Protecting the privacy of an ML model is of paramount importance as organizations invest significant resources on cutting edge research and also on gathering and labelling large amounts of training data [6]. In addition, recent works [13,15,17,22] have shown that an adversary could train a substitute model via model stealing and use it further for crafting adversarial examples [5] in a black-box setting, which poses a serious threat when the model is deployed in security critical applications. A stolen model could also compromise the privacy of users by leaking confidential data through a membership inference attack [14] or via model inversion [20,21]. Fig.1 showcases some of the possible malicious outcomes of Model Stealing. In order to prevent model stealing attacks, some defenses attempt to perturb the softmax predictions of the model, while preserving the top-1 prediction [9]. In this work we consider the problem of model stealing in a more practical and challenging hard label setting, where only the top-1 prediction of the model is accessible, and is thus effective even in the presence of such defenses. In a practical scenario, the adversary would not have access to the training data, and hence we consider the problem of Data-Free Model Stealing (DFMS) in this work. In such a data-free scenario, the attacker could use publicly available related datasets [12, 13], or synthetically generated samples [16] to query the model. While the use of publicly available datasets assumes access to related data, the data-free generative approach suffers from a large query budget, as the synthetic data can be far from the true training data distribution. In this work we overcome both challenges by utilizing the available data that may be potentially unrelated to the original training dataset, as a weak image prior. This enables the generation of representative samples under a low query budget.

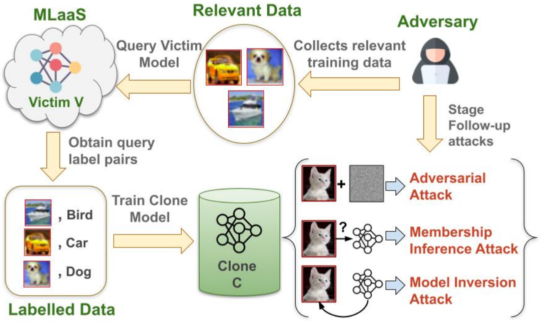

Figure 1. **Model Stealing Attack and its vulnerabilities**

While Data-Free Knowledge Distillation works [1, 3, 10,

[1] Project Page: https://sites.google.com/view/dfms-hl

11, 19] achieve near perfect accuracy, the additional challenges in a Model Stealing framework stem from the restriction of access to gradients and a hard label setting. Therefore, we consider the problem of data-free hard label model stealing and overcome the challenges by utilizing the clone model's gradients as a proxy to the gradients of the Victim model. This allows us to train the generator alternately with the clone model by enforcing the generation of a class-balanced dataset that is also more aligned with the distribution of the training dataset. We also utilize an adversarial loss in a GAN framework [4], by using a small amount of publicly available data, which we refer to as proxy data [1]. While this could be completely unrelated to the original training dataset, it still helps in enforcing a weak image prior in the generated data. This in turn reduces the number of Victim model queries needed to perform Model Stealing. In fact, we show that it is possible to even use synthetic samples, such as multiple overlapping shapes with a planar background, to steal a model in a completely data-free setting. Our method achieves a significant improvement over ZSDB3KD [18], a zero-shot data-free method in a similar hard label setting using only synthetic samples.

Key Contributions:

- We propose DFMS-HL, a data-free model stealing (DFMS) attack in a hard-label (HL) setting to train a clone model with the help of unrelated proxy data. We show that DFMS-HL outperforms the existing baseline ZSDB3KD [18] and results in a significant reduction of around 500× in the number of queries.
- We demonstrate state-of-the-art results on CIFAR-10 using unrelated proxy samples, such as 40 or 10 classes from CIFAR-100, or a synthetic dataset.
- We are the first to show noteworthy results of data-free model stealing on a dataset with a larger number of classes such as CIFAR-100.
- The soft-label variant (DFMS-SL) achieves a significant boost of 3% over the state-of-the-art model stealing attacks MAZE [7] and DFME [16].

2. Proposed Approach

We propose a data-free model stealing approach **DFMS-HL** that requires only hard-labels. At first, we train a DC-GAN by imposing an image prior using synthetic or unrelated proxy data. This gives a good initialization for the generator $\mathcal{G}$. We also train an initial clone model with the proxy images. Following this, we then begin our procedure of alternately training the clone model and the generator. The data flow is shown in Fig. 2 wherein the generator $\mathcal{G}$ generates data $x = \mathcal{G}(z)$ from a random normal vector z. The victim model takes input x and generates input, label pairs $(x, \hat{y}(x))$ for each instance in x. Since, the victim model is black-box, we do not backpropagate the gradients through it. The labelled input pairs are used to train the clone model with the cross-entropy loss as follows:

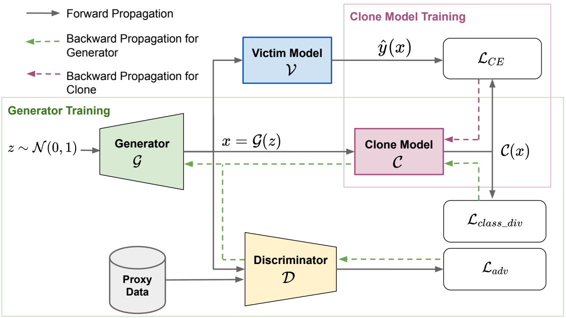

Figure 2. **Architecture of DFMS-HL**: Generator $\mathcal{G}$ generates data x with a proxy image prior. The clone model $\mathcal{C}$ is trained using the labels from the victim model $\mathcal{V}$ with cross-entropy loss $\mathcal{L}_{CE}$. The generator $\mathcal{G}$ is trained with the adversarial loss $\mathcal{L}_{adv}$ along with the class-diversity loss $\mathcal{L}_{class\_div}$.The generator and clone model are trained alternately in every iteration.

$$\mathcal{L}_C = \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,I)} \left[\mathcal{L}_{CE}(\hat{y}(x), \mathcal{C}(x))\right], \; x = \mathcal{G}(z) \tag{1}$$

where $\hat{y}(x) = \underset{i}{\operatorname{argmax}}\, \mathcal{V}_i(x)$ is the class label for the maximum probability class and $\mathcal{C}(x)$ is the output logits from the clone model. The generator is trained with the adversarial loss [4] and a unique diversity loss as shown below:

$$\mathcal{L}_{adv,real} = \mathop{\mathbb{E}}_{x\sim p_{data}(x)} \left[log\mathcal{D}(x)\right], \tag{2}$$

$$\mathcal{L}_{adv,fake} = \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,I)} \left[log(1 - \mathcal{D}(\mathcal{G}(z))\right] \tag{3}$$

Across a batch of N samples, we take the expected confidence value over the batch as α_j for every class j and obtain the entropy over K classes. Hence, the generator model learns to generate samples from different classes by minimizing the diversity loss formulation as below,

$$\mathcal{L}_{class\_div} = \sum_{j=0}^{K} \alpha_j \log \alpha_j, \;\; \alpha_j = \frac{1}{N}\sum_{i=1}^{N} \text{softmax}(\mathcal{C}(x_i))_j \tag{4}$$

The equations below describe the generator and discriminator losses, that are minimized alternately for training.

$$\mathcal{L}_G = \mathcal{L}_{adv,fake} + \lambda_{div}\mathcal{L}_{class\_div} \tag{5}$$

$$\mathcal{L}_D = \mathcal{L}_{adv,real} + \mathcal{L}_{adv,fake} \tag{6}$$

3. Experiments

Comparison with Knowledge distillation methods: We perform experiments on CIFAR-10 as the True dataset as shown in Table 1 for comparing with existing KD methods. DeGAN [1] and ZSKD [11] are data-free knowledge distillation methods with white-box teacher access. KnockoffNets [12] and Black-Box Ripper [2] are data-free KD methods in a black-box setting. Similar to the experimental setting of prior works [1, 2], we use 40 unrelated classes from CIFAR-100 dataset

Table 1. Comparison of DFMS-HL with state-of-the-art KD methods(Top) and ZSDB3KD (Bottom)

Method	Hard Label	Black Box	Data Free	Victim Acc	Data Free	CIFAR-100 (40C)	CIFAR-100 (10C)
Victim Accuracy = 82.5%							
ZSKD	×	×	✓	82.50	**69.50**	-	-
DeGAN	×	×	✓	82.50	-	76.30	72.60
KnockoffNets	×	✓	✓	82.50	-	65.70	46.60
Black-Box Ripper	×	✓	✓	82.50	-	**76.50**	**77.90**
DFMS-HL (Ours)	✓	✓	✓	82.52	65.70	76.02	71.36
Victim Accuracy $\sim$ 80%							
ZSDB3KD	✓	✓	✓	79.30	59.46	-	-
DFMS-HL (Ours)	✓	✓	✓	80.18	**67.03**	**74.27**	**70.57**

Table 2. Performance of DFMS-HL on CIFAR-100

Method	Proxy Data	Victim Acc	Clone Acc
DeGAN	CIFAR-10	78.52	75.62
DFMS-HL	CIFAR-10	78.52	72.83
DFMS-HL	Synthetic	78.52	43.56

as the proxy dataset for CIFAR-10 model stealing. We also show results on 10 random classes from these 40 classes. We achieve comparable results with the data-free KD methods despite having more restrictions on access to the victim model. Apart from proxy data, we also show results on synthetically crafted data. We generate a synthetic dataset (shown in Fig 3) of 50k samples using skimage python library[2] by drawing shapes of triangle, rectangle, ellipse and circles at random locations on top of a clear background. These manually crafted images are converted to grey-scale and then used as proxy data. From Table 1, it can be observed that our approach not only outperforms ZSDB3KD by a large margin, but also achieves a comparable accuracy with respect to the DeGAN and Black-Box Ripper for the CIFAR-100 40 classes proxy data. We also use a significantly lower query budget of 8M as compared to ZSDB3KD which requires 4000M queries. We also perform experiments on CIFAR-100 (Table 2) with CIFAR-10 [1, 2] and synthetic data as proxy datasets. DFMS-HL reaches a comparably close accuracy of 72.83% using CIFAR-10 as the proxy without any access to the victim model's gradients and only using hard labels

Comparison with Model Stealing methods. We compare our approach with the state-of-the-art data-free Model Stealing approaches [8, 16] in Table 3. We obtain an accuracy of 84.51% by merely using synthetic samples in a completely data-free hard-label setting. We use a lower query budget of 8M, as compared to that of DFME and MAZE that require 20M queries for CIFAR-10. We further extend our attack to the soft-label black-box scenario (denoted as DFMS-SL in Table 3) where the softmax predictions of the victim model are available. We get a boost of almost 3% using synthetic data and CIFAR-100 10 classes with the same query budget of 20M.

[2] https://scikit-image.org/docs/stable/auto_examples/edges/plot_random_shapes.html

Table 3. Comparison of DFMS-HL with data-free model stealing methods MAZE and DFME (Top) and with ZSDB3KD (Bottom)

Method	Hard Label	Black Box	Data Free	Victim Acc	Data Free	CIFAR-100 (40C)	CIFAR-100 (10C)
Victim Accuracy $\sim$ 95.5%							
MAZE	×	✓	✓	95.50	45.60	-	-
DFME	×	✓	✓	95.50	88.10	-	-
DFMS-HL (Ours)	✓	✓	✓	95.59	84.51	**92.06**	**85.53**
DFMS-SL (Ours)	×	✓	✓	95.59	**91.24**	**93.96**	**90.88**
Victim Accuracy $\sim$ 93.7%							
ZSDB3KD	✓	✓	✓	93.65	50.18	-	-
DFMS-HL (Ours)	✓	✓	✓	93.83	**85.92**	**90.51**	**83.37**

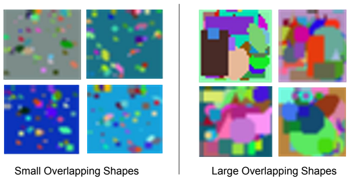

Figure 3. **Synthetic images:** Equal share of large(right) and small(left) overlapping shapes on planar background used to train the clone model.

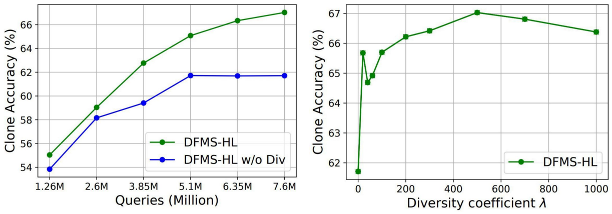

Figure 4. **Query Ablation (Left):** Sensitivity Plot of Clone model accuracy to number of queries. A significant boost of 6% in the clone model accuracy is evidenced after using class-diversity loss. **Class-diversity Loss (Right):** Clone model accuracy increases with increase in diversity coefficient λ_{div}.

4. Ablation Experiments

Effect of Query Budget: We analyse the impact of the query budget on the clone model accuracy. Our approach achieves a good accuracy with a query budget of 7.6 million on synthetic data for AlexNet as victim model and AlexNet-half as clone model. From Fig.4, we observe that even with a small query budget of 1.26M, our method performs well and it almost saturates within 8M. We report the saturating accuracies in Table 1 and 3. We use a query budget of 10M for the CIFAR-100 experiments (Table 2) and 8M for CIFAR-10 experiments (Tables1 and 3). The class-diversity loss has a huge impact with a significant boost of 6% in the clone accuracy for synthetic data using 7.6M queries.

Effect of Class Diversity Loss: We gradually increase the loss coefficient from 0 to 1000 for synthetic data as proxy with CIFAR-10 as the true dataset as shown in Fig. 4 and measure the clone model accuracy. We run the ablations till 7.6M queries and observe that increasing the coefficient λ_{div} of class-diversity loss improves the clone model accuracy. We reported our final results with a λ_{div} value of 500 for CIFAR-10 experiments in Table 1 and 3 and set λ_{div} as 100 for CIFAR-100 experiments in Table 2.

5. Conclusions

In this paper, we propose an effective model stealing attack in a practical setting of having access to only hard-labels of a black-box victim model. Extensive experiments show that our method DFMS-HL performs better than the state-of-the art model stealing method at a 500x lower query budget. We further show that our attack is effective in a completely data-free setting using a synthetic dataset. We demonstrate the scalability of the proposed model stealing attack to CIFAR-100 as well with a low query budget, which has not been attempted in prior works

Acknowledgements This work was supported by a project grant from MeitY (No.4(16) /2019-ITEA), Govt. of India and Uchhatar Avishkar Yojana (UAY, IISC_010), MHRD, Govt. of India, Prime Minister's Research Fellowship (Sunandini), and Google PhD Fellowship (Sravanti).

References

[1] Sravanti Addepalli, Gaurav Kumar Nayak, Anirban Chakraborty, and Venkatesh Babu Radhakrishnan. DeGAN: Data-enriching gan for retrieving representative samples from a trained classifier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 1, 2, 3

[2] Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, and Marius Popescu. Black-Box Ripper: Copying black-box models using generative evolutionary algorithms. *arXiv preprint arXiv:2010.11158*, 2020. 2, 3

[3] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019. 1

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[6] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. 1

[7] Sanjay Kariyappa, Atul Prakash, and Moinuddin K Qureshi. Protecting dnns from theft using an ensemble of diverse models. In *International Conference on Learning Representations*, 2020. 2

[8] Sanjay Kariyappa, Atul Prakash, and Moinuddin K Qureshi. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[9] Taesung Lee, Benjamin Edwards, Ian Molloy, and Dong Su. Defending against model stealing attacks using deceptive perturbations. *arXiv preprint arXiv:1806.00054*, 2018. 1

[10] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017. 1

[11] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, pages 4743–4751. PMLR, 2019. 1, 2

[12] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[13] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 1

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. 1

[15] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017. 1

[16] Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4771–4780, 2021. 1, 2, 3

[17] Wenxuan Wang, Bangjie Yin, Taiping Yao, Li Zhang, Yanwei Fu, Shouhong Ding, Jilin Li, Feiyue Huang, and Xiangyang Xue. Delving into data: Effectively substitute training for black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

[18] Zi Wang. Zero-shot knowledge distillation from a decision-based black-box model. *arXiv preprint arXiv:2106.03310*, 2021. 2

[19] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 1

[20] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[21] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Y Lim. Exploiting explanations for model inversion attacks. *arXiv preprint arXiv:2104.12669*, 2021. 1

[22] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. Dast: Data-free substitute training for adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1