DUAL STUDENT NETWORKS FOR DATA-FREE MODEL STEALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-free model stealing aims to replicate a target model without direct access to either the training data or the target model. To accomplish this, existing methods use a generator to produce samples in order to train a student model to match the target model outputs. To this end, the two main challenges are estimating gradients of the target model without access to its parameters, and generating a diverse set of images that thoroughly explores the input space. We propose a Dual Student method where two students are symmetrically trained in order to provide the generator a criterion to generate samples that the two students disagree on. On one hand, disagreement on a sample implies at least one student has classified the sample incorrectly when compared with the target model. This push towards disagreeing samples implicitly encourages exploring a more diverse region of input space. On the other hand, our method utilizes gradients of student models to indirectly estimate gradients of the target model. We show that this novel training objective for the generator network is equivalent to optimizing a lower bound on the generator's loss if we had access to the target model gradients. In other words, our method alters the standard data-free model stealing paradigm by substituting the target model with a separate student model, thereby creating a lower bound which can be directly optimized without additional target model queries or separate synthetic datasets. We show that our new optimization framework provides more accurate gradient estimation of the target model and better accuracies on benchmark classification datasets. Additionally, our approach balances improved query efficiency with training computation cost. Finally, we demonstrate that our method serves as a better proxy model for transfer-based adversarial attacks than existing data-free model stealing methods.

1 INTRODUCTION

Model stealing has been shown to be a serious vulnerability in current machine learning models. Machine learning models are increasingly being deployed in products where the model's output is accessible through APIs, also known as Machine Learning as a Service. Companies put a large amount of effort into training these models through the collection and annotation of large amounts of data. However, recent work has shown that the ability to query a model and get its output – *without access to the target model's weights* – enables adversaries to utilize different model stealing approaches, where the attacker can train a student model to have similar functionality to the target model (Kesarwani et al., 2018; Yu et al., 2020; Yuan et al., 2022; Truong et al., 2021; Sanyal et al., 2022). Two major motivations for stealing a private model are utilizing the stolen model for downstream adversarial attacks as well as monetary gains (Tramèr et al., 2016; Zhang et al., 2022).

Data-free model stealing is a generalization of black-box model stealing which is an extension of knowledge distillation. In all three areas, the aim is to obtain a student model which imitates the target model. However, while knowledge distillation environments typically retain full knowledge of the target model training data and weights, black-box model stealing eliminates the need to have access to the target model weights and training data. Going one step further, black-box model stealing typically uses real-world data samples to train the student network, whereas data-free model stealing removes that requirement by leveraging a generator. The removal of explicit substitute training data in data-free model stealing provides more generalizable results, as the relation between the substitute dataset and true dataset are typically unknown in practice.

Existing methods for data-free model stealing involve either using a min-max adversarial approach inspired by Generative Adversarial Networks (GANs) (Truong et al., 2021; Kariyappa et al., 2021; Zhang et al., 2022), or a generator-discriminator setup incorporating GANs (Sanyal et al., 2022). DFME (Truong et al., 2021) and DFMS-HL (Sanyal et al., 2022) are the two most recent and highest performing methods for these approaches, respectively. In both approaches, a student is optimized to minimize the distance between the student and target model outputs. However, in the former, a generator is optimized to maximize the distance between student and target model outputs, while in the latter, the generator-discriminator is optimized to generate images similar to a synthesized dataset which have balanced student classifications. Even though utilizing a synthesized dataset relaxes the truly data-free requirement, the current state of the art (SOTA) is provided by the DFMS-HL method which does utilize synthetic data, while DFME method provided the previous SOTA and does not require synthetic data.

Our proposed Dual Student method alters the generator-student min-max framework by training two student models to match the target model instead of one, allowing the generator to have a much more effective criterion. The two students are symmetrically trained to match the target model, while employing the new generator objective to generate samples on which the two students disagree. Disagreement between the two students implies disagreement between at least one of the students and the target model, resulting in the generated sample being a hard sample (c.f. hard sample mining (Shrivastava et al., 2016)) to at least one student. This hard sample provides the driving force to explore controversial regions in the input space on which the students and teacher disagree on. Our new generator optimization not only results in samples more beneficial for efficiently training the student models, it also provides a more accurate gradient estimation of the target model by creating a substitute lower bound for the overall min-max portion of the generator-student framework. In other words, we propose a novel approach to estimate gradients of the target model without directly querying this model. This important change not only eliminates the need of some existing approaches (Truong et al., 2021; Kariyappa et al., 2021) to estimate the gradient through the target model which reduces the number of queries, it also allows these soft-label methods to be easily extended to the more difficult hard-label setting. Finally, we also explore a more difficult data-free setting where the number of classes is not know in advance. To the best of our knowledge, we are the first to address this setting and provide a solution to this new challenging data-free setup.

In summary, our contributions in this paper are as follows: (1) We propose the Dual Student method which provides a tractable approach for the data-free model extraction problem with minimal added components. (2) The Dual Student method can be incorporated into existing soft-label approaches to extend them to hard-label settings, and has better classification accuracy than existing methods. (3) We show the effectiveness of utilizing the Dual Student setup to fine-tune fully trained models by directly improving the existing SOTA from DFMS-HL and evaluate the effectiveness of the top data-free model stealing approaches for transfer-based adversarial attacks against the target model.

2 Related Work

In this section, we review the foundational papers our method is built upon, as well as other recent works in this area. With the overarching problem of trying to distill the knowledge out of a target into a student, knowledge distillation is one of the key areas for this problem (Buciluǎ et al., 2006; Romero et al., 2014; Ba & Caruana, 2014; Hinton et al., 2015). Shifting the perspective of knowledge distillation towards attacking, model *extraction* explores obtaining a model without full access to the target method (Bastani et al., 2017b;a; Wang et al., 2021; Orekondy et al., 2019; Barbalau et al., 2020; Yuan et al., 2022). And in-line with that objective, the data-free model *stealing* removes the data-requirement that most black-box model extraction techniques rely upon (Truong et al., 2021; Kariyappa et al., 2021; Sanyal et al., 2022). One of the primary motivations for data-free model stealing is its use in transfer-based adversarial attacks, which attempts to create adversarial examples on a substitute model that remain effective when transferred to the target model (Wang et al., 2021; Zhang et al., 2022).

2.1 Knowledge Distillation

Knowledge distillation is the process of using transfer learning to train a different, typically smaller, model. From the area of model compression (Buciluǎ et al., 2006; Romero et al., 2014), knowledge

distillation was initially used determine the necessary depth of deep learning models in Ba & Caruana (2014). The objective was then expanded towards the general goal of creating smaller and more efficient models (Hinton et al., 2015; Urban et al., 2016; Yin et al., 2020). Our approach to data-free model extraction builds off of knowledge distillation techniques for training a student model from a target model, otherwise referred to as the "teacher" or "victim" model in other methods.

Model extraction, also called model stealing, is used to obtain a new model which performs similarly to the target model without knowledge of the target model's gradients. The key difference between model extraction and knowledge distillation is the black-box access to the target model as compared with typical white-box access provided in knowledge distillation. First uses of model extraction were in extracting decision trees to help better explain deep learning models (Bastani et al., 2017b;a), however, recent applications have focused on obtaining a substitute model for the purpose of transfer-based attacks on the target model (Wang et al., 2021).

Early knowledge distillation works had some reliance on access to the original training data. To address this, data-free knowledge distillation methods utilize a generator to remove this dependence (Fang et al., 2019; Binici et al., 2022). Recent successful works utilize activation regularization between the target and student models, in addition to interpolation between generated examples to improve training (Qu et al., 2021).

2.2 Data-Free Model Extraction

Extending this problem, a relatively new area of model extraction is data-free model extraction which aims to train a student model without being given a dataset prior. Most existing model extraction techniques rely on either some small set of labeled samples, or an alternative training dataset (Orekondy et al., 2019; Barbalau et al., 2020; Wang, 2021). The first solutions to this data-free model extraction problem are called DFME and MAZE, and use an adversarial setup between a generator and a student (Truong et al., 2021; Kariyappa et al., 2021). The student aims to match the target model and the generator aims to find examples which result in different classifications between the student and target models. These works use zeroth-order gradient estimation in the form of the forward differences method to estimate the gradient through the target for training the generator. This direct gradient estimation is one aspect that our Dual Students method seeks to improve on. In particular, our generator attempts to generate samples that the two students are disagreeing upon. We show this criterion is equivalent to optimizing a lower bound on the generator loss if the generator had full access to the target model.

Other data-free model stealing works seek to remove this gradient estimation step by altering the generator objective. A slightly different training scheme which does this alters the generator objective to maximize the confidence of the student output (Hong et al., 2022), though the approach suffers from training stability issues. The current SOTA for data-free model stealing is DFMS-HL by Sanyal et al. (2022), which not only sets the bar for student model classification accuracy, it also extends the problem to the hard-label setting. Their approach utilizes a full generator-discriminator GAN setup to generate samples similar to a synthetic dataset while retaining balanced class classifications of the images by the student model. These two aspects – keeping samples similar to the synthetic dataset, and balancing the generated class distribution – remove the reliance on gradient estimation techniques that methods such as DFME require. We note that DFMS-HL utilizes a synthetic dataset of random shapes generated on colored backgrounds, which breaks the truly data-free setting.

2.3 Transfer-based Attacks

Transfer-based adversarial attacks Akhtar et al. (2021) use a substitute model's gradients to generate an adversarial image to fool the target model. Initially explored using models trained in a similar setting, recent work has moved towards a more realistic data-free setting (Zhang et al., 2021). A prominent use-case for model stealing attacks is to use the student model as a proxy for transfer-based adversarial attacks. Several works exploring the area of data-free model extraction include experiments on the effectiveness of transfer-attacks on the student models (Hong et al., 2022; Li & Merkel, 2021; Yue et al., 2021). The method proposed in Zhang et al. (2022) leverages data-free model stealing work to use orders of magnitude fewer queries than existing data-free model stealing approaches to fully train for transfer-based adversarial attacks. However, we do not include

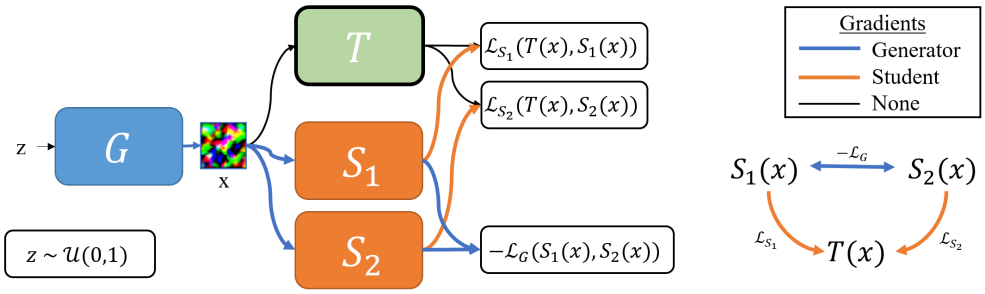


Figure 1: Our network consists of a generator, two students, and a target model. The generator generates images which the two students are trained to classify. The student losses in the Dual Student setup aim to reduce the distance between the target and student model outputs, and only require the soft-label or hard-label outputs of the target model. The generator loss competes against the two student losses by maximizing the distance between the target and student model outputs.

this work in our transfer-based attack comparison, as that work is targeted towards transfer-based adversarial attacks as opposed to our primary objective of extracting classification performance.

3 Dual Students Method

This section outlines the framework of the Dual Students method, an explanation for why our proposed optimization is better, and an empirical analysis of our algorithm. The Dual Students method consists of a generator, two students, and a target model with limited black-box access. Similar to the adversarial objective in GANs where the generator tries to fool the discriminator, the generator in the Dual Students method generates images on which the two students should have disagreeing outputs. A mismatch between outputs of the two students on a generated image implies that at least one of the student outputs does not match the teacher output for this image. This is beneficial for optimizing the students to match the target model, especially in the hard-label setting where hard examples are crucial for correcting the student models. These differing objectives – generator promoting disagreement between the students and the students aiming for agreement with the teacher – form the min-max framework of our setup, which facilitates the exact computation of the gradients for all losses.

Architecture of the proposed method is show in Figure 1, where a key contribution is that the new generator optimization no longer requires the estimated gradient of the target model. The student model gradients are not a direct proxy for the target model gradients. However, when the gradient of the generated image is computed from the generator loss, Dual Students results in the generator gradient being more aligned to the generator gradient computed in a standard white-box setting than when the generator gradient is computed using single-step forward differences method.

3.1 Data-Free Model Stealing Framework

For target T, generator G, student S, and noise z, our broad objective for solving data-free model stealing is to optimize:

$$\min_S \max_G ||T(G(z)) - S(G(z))||_p. \tag{1}$$

The original Data-Free Model Extraction framework aimed to extract the target model by utilizing a generator G with parameters θ_G to generate images x from noise $z \sim \mathcal{U}([0,1]^d)$, where d is the dimension of the latent space, in which the student S with parameters θ_S is trained to match the output of target model T. The loss function for the student is:

$$\min_{\theta_S} \mathcal{L}_S(S(G(z;\theta_G);\theta_S), T(G(z;\theta_G))),$$

where student loss $\mathcal{L}_S$ is typically $\mathcal{L}_S = D_{KL}$ or $\mathcal{L}_S = \ell_1$. We choose to use the latter when applicable, as it has been shown to perform better than D_{KL} (Truong et al., 2021). The generator,

on the other hand, is optimized to maximize the difference between student and teacher. To this end, the negative of the student loss is typically used:

$$\min_{\theta_G} -\mathcal{L}_G(S(G(z;\theta_G);\theta_S), T(G(z;\theta_G))).$$

However, whereas the student loss only requires the gradient through the student model, the generator loss requires the gradient through the student, target, and generator. Due to the limited black-box access to the target model, existing approaches utilize the forward differences method to estimate the gradient through the target model. An additional problem is that the generator loss does not directly promote a diversity of classes within the generated images. Though this has not been a problem for 10-class datasets like SVHN and CIFAR10, we note in Section 4.2 that diversity of classes becomes a larger challenge when using datasets with more classes like CIFAR100.

3.2 Approximating Gradient with Dual Students

Towards making the broad DFMS objective outlined in Eq. 1 differentiable, we propose adding an additional student S_2 to solve the following optimization problem:

$$\min_{S_1,S_2} \max_{G} ||S_1(x) - T(x)||_p + ||S_2(x) - T(x)||_p, \tag{2}$$

for any ℓ_p norm loss. Though neither the original objective in Eq. 1 or this new objective are differentiable due to requiring the gradients through the Target model T, adding the second student S_2 allows us to make a substitution to use the generator optimization:

$$\max_{G} ||S_1(G(x)) - S_2(G(x))||_p. \tag{3}$$

This optimization for Generator G removes the Target model T and makes the problem directly differentiable. This new generator optimization is related to our new objective in Eq. 2 through the triangle inequality,

$$||S_1(x) - S_2(x)||_p \leq ||S_1(x) - T(x)||_p + ||S_2(x) - T(x)||_p. \tag{4}$$

The LHS of inequality 4 corresponds to the generator's loss in Dual Student with $x = G(z)$ in Eq. 3, and the RHS is the minimization desired in Eq. 1 with an additional student. In other words, when the LHS is maximized during generator optimization, the lower bound is increasing:

$$\max_{x} ||S_1(x) - S_2(x)||_p \leq \max_{x} ||S_1(x) - T(x)||_p + ||S_2(x) - T(x)||_p.$$

This ensures that either $S_1(x)$ or $S_2(x)$ is getting farther from $T(x)$ during the generator optimization. To optimize minmax optimization in Eq. 3, we iteratively update students and generator networks. In the first step, Student S_i with parameters θ_{S_i} is optimized to match the output of the target model for a particular query:

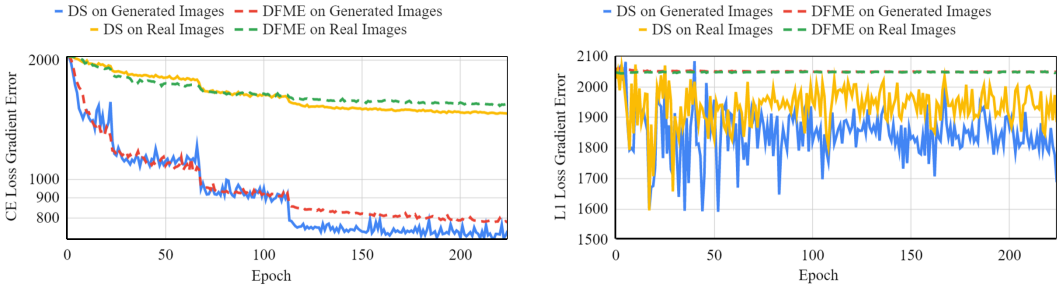
$$\min_{\theta_{S_i}} \mathcal{L}_S(S_i(G(z;\theta_G);\theta_{S_i}), T(G(z;\theta_G))). \tag{5}$$

Next, the generator is updated according to:

$$\max_{\theta_G} \mathcal{L}_G(S_1(G(z;\theta_G);\theta_{S_1}), S_2(G(z;\theta_G);\theta_{S_2})). \tag{6}$$

The student loss used for soft-labels is $\mathcal{L}_S = \ell_1$, whereas the student loss used for hard-labels is cross entropy, i.e. $\mathcal{L}_S = ce$, as this is standard for classification tasks. The generator optimization maximizes the difference between the outputs of the two students, and $\mathcal{L}_G = \ell_1$ is used in the generator loss when training on both soft-labels and hard-labels. We provide an algorithm for our method in the supplemental materials.

A major benefit of our approach is the simplicity of requiring only an additional student to overcome the non-differentiability of the black-box teacher. Our method is fully differentiable, in contrast to methods like DFME which approximate teacher gradients using forward differences. Additionally, the Dual Students approach is compatible with both soft-label and hard-label learning through a simple change to the loss function. Existing gradient-estimation based methods like DFME can



(a) (b)

Figure 2: The distance between the true and estimated target model gradients of either Generated images or Real images. The gradient is computed w.r.t. input image x, and a formal description of the distance used is provided in Eq. 7. (a) Cross-Entropy loss is used to compute the gradient in the hard-label setting, and (b) ℓ_1 loss is used in the soft-label setting. Real images are from the test split of the CIFAR10 dataset.

only be used in the soft-label setting due the forward differences method needing the additional information provided by the soft-label to remain effective. For other fully trained DFMS, our method can be used to fine-tune the trained models as long as there is a student and generator. In the next subsection we evaluate the accuracy of the different gradient approximation methods compared to our Dual Students method.

3.3 Evaluating the Gradient Approximation

The aim for the generator optimization under the black-box scenario proposed in Eq. 6 is to get as close as possible to the generator optimization under the white-box scenario outlined in Eq. 1. One way to evaluate the quality of the estimated gradients between the forward differences method and the Dual Students method, is to compare the gradients computed from $\mathcal{L}_G$ w.r.t. the generated image x when using the black-box estimate versus the gradients computed when using the white-box Target model directly. By comparing difference between the normalized gradients, the difference can be shown as:

$$f(x) = \left|\left|\nabla_x \frac{\mathcal{L}_G(S(x;\theta_S), T(x;\theta_T))}{||\mathcal{L}_G(S(x;\theta_S), T(x;\theta_T))||} - \nabla_x \frac{\mathcal{L}_G(S_1(x,\theta_{S_1}), S_2(x;\theta_{S_2}))}{||\mathcal{L}_G(S_1(x;\theta_{S_1}), S_2(x;\theta_{S_2}))||}\right|\right|_2 . \tag{7}$$

The results of the forward differences method compared with our Dual Students using this setup is shown in Figure 2. The figure plots the difference between the normalized gradient of the generated image using the target model (white-box gradient) versus the estimate (forward differences or Dual Students). Figure 2a shows that our method provides a gradient estimate closer to the white-box setting than the forward differences method used in DFME in the hard-label setting for both generated and real images. Figure 2b shows error for the soft-label setting, where we see DFME has similar estimation error between real and generated images, while our Dual Students has a smaller error for both between real and generated images. We note that the performance on generated images are more important as generated images are the only images the models have access to during training.

Empirically, we see that this new generator loss results in more accurate gradient estimations for the target model than the forward differences method, as is shown in Figure 2. We note that the forward differences method utilizes soft-labels, and performance is much worse when the soft-label information is reduced to hard-labels.

4 Experiments

We evaluate the Dual Students method on various datasets for standard classification accuracy and transfer-based attack effectiveness. The datasets used are MNIST, FashionMNIST, GTSRB, SVHN, CIFAR10, and CIFAR100 (Xiao et al., 2017; Stallkamp et al., 2011; Netzer et al., 2011; Krizhevsky et al., 2009). The target model architecture is ResNet-34, while the students are ResNet-18 and

Table 1: Percent student accuracy for different methods with and without the Dual Student (DS) method.

Dataset	Target Accuracy	Method	Probabilities	Hard-Labels
MNIST	99.66	DFME	99.15	97.85
	99.66	DS	**99.36**	**99.25**
FashionMNIST	93.84	DFME	85.17	48.91
	93.84	DS	**91.17**	**80.53**
GTSRB	97.21	DFME	96.35	85.69
	97.21	DS	**96.40**	**93.20**
SVHN	96.20	DFME	95.33	93.87
	96.20	DS	**95.72**	**95.43**
CIFAR10	95.5	DFME	88.10	68.40
	95.5	DS	**91.34**	78.72
	95.5	DFMS-SL/HL	83.02	79.61
	95.5	DFMS-SL/HL + DS	89.38	**85.06**

are trained for 2 million queries for MNIST, FashionMNIST, GTSRB, and SVHN and 20 million queries for CIFAR10 and CIFAR100 (He et al., 2016). The students and generator are trained using SGD with weight decay and momentum. Moving average weights are used for the students and generators (Cai et al., 2021), and for the classification task an average of the two student's outputs are used in an ensemble. However, for transfer-based attacks, only the higher accuracy student is used as a proxy in order to help make the performance more comparable across methods.

4.1 Classification Performance

The standard evaluation setting for data-free model extraction is to test the student model accuracy on the test set respective to the dataset the target model was trained on. Table 1 provides a comparison between DFME (Truong et al., 2021), DFMS-HL (Sanyal et al., 2022), our proposed Dual Students method, and when our Dual Students method is used with DFMS-HL. As can be seen, our method outperforms DFME and DFMS-HL on the different datasets using soft-labels (probabilities), with the most notable increase being on FashionMNIST. On hard-labels, our method outperforms DFMS-HL when used to fine-tune the DFMS-HL trained student.

Using Dual Students to fine-tune fully trained generator-student-teacher methods is an effective way to increase accuracy. This is especially beneficial for methods that have a different optimization scheme such as DFMS-HL. From a high level view, the DFMS-HL method optimizes the generator to generate images that are similar to the proxy data and that have a balanced distribution of classes when classified by the *student model*. Due to the student model being optimized to match the target model outputs, the generator is indirectly optimized to generate a balanced distributions of classes as classified by the *target model*. There is no loss in the setup which directly incentivizes samples on which the student and teacher disagree. Thus when our method is used to fine-tune the fully trained DFMS-HL models for 1 million queries (5% of total), we see an improvement of over 5% in Table 1. We note that the DFMS-SL/HL results shown were trained using our target model in order to better compare with other methods, and don't fully match the accuracies reported in Sanyal et al. (2022). They achieve 85.92% accuracy on hard-labels compared with the 79.61% accuracy we found. When their model is fine-tuned using Dual Students, we observe an increased accuracy of 88.46%.

4.2 Query Efficiency and Number of Classes

The Dual Student generator loss avoids the extra queries necessary for zeroth order gradient estimation when optimizing the Generator. Using zeroth order gradient estimation adds additional queries to the target model during the generator training. The DFME method has the same alternating generator-student training and uses the same parameters generator and student iterations. However, due to their use of the forward differences method when updating the generator, each generator iteration requires an additional 2 queries to the target model. This results in our method using $\frac{5}{7} \approx 71.42\%$ of the queries that DFME does.

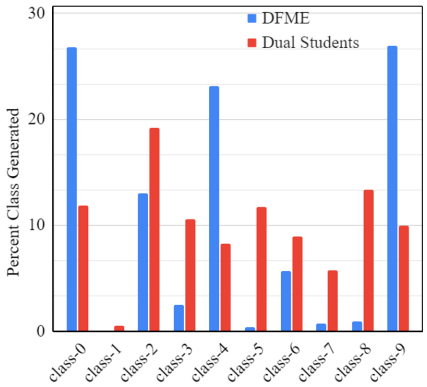


Figure 3: Class distribution of generated classes at the end of training for CIFAR10 dataset. Dual Student is able to produce more balanced samples.

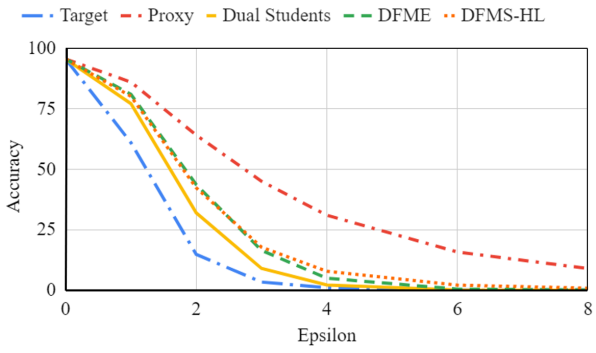


Figure 4: Accuracy on Target Model of perturbed images generated using PGD attack with varying epsilons ($\frac{\epsilon}{255}$) where different soft-label DFMS student models are used.

When compared with the DFMS-HL, our Dual Student method benefits from a simpler training setup and requires less compute. Though both methods have the same number of separate models, DFMS-HL requires the generator to be pretrained on the synthetic data, whereas our method can be trained from scratch using alternating generator-student training loops.

One challenge that comes when scaling the DFMS problem relates to the number of classes in the dataset. Intuitively, datasets with more classes tend to be more difficult than datasets with fewer classes, and this is what we found when extending our method to CIFAR100. With no changes to our setup, DFME achieves around 26.46% accuracy on CIFAR100 while our Dual Students method achieves around 45.32%. A reason for this large reduction in accuracy when going from CIFAR10 to CIFAR100 is the lack of explicit diversity within the generator loss optimization L_G that we define. Though the generator is optimized to find confusing images for the two students, there is no guarantee that those confusing images are evenly distributed across classes. Despite this, Figure 3 shows that the generator for the Dual Students method generates a more balanced set of classes than the DFME generator on CIFAR10, where perfectly balanced classes would be 10% prevalence for all classes. We provide additional figures and analysis of generated image class distributions in our supplementary material.

A natural extension to the data-free black-box hard-label environment that DFMS works in is to remove the knowledge about the number of classes in the target dataset. This raises the challenge of how to learn about the number of classes, and how to alter the classifier to grow as the number of classes grows. To extend our method and address these challenges, the seen classes are mapped in the order they are observed, and student classifier training is limited to the number of seen classes. As additional classes are discovered naturally by the generator traversing the image-space, the final classification layer of the student model expands to include these additional classes as separate outputs. Training on CIFAR10 using hard-labels without the number of classes known, our method achieves 74.95% accuracy, which is comparable to the hard-label accuracy when the number of classes are known.

4.3 Black-box Attacks

The effectiveness of using the trained student model as a substitute for the target model in a black-box transfer-based adversarial attack can be seen in Table 2. The DaST method (Zhou et al., 2020) and the work done by Zhang et al. (2022) also have a similar generator-student setup, however the focus of those works are primarily on being used as a transfer-based proxy model, whereas this work focuses on classification accuracy of CIFAR10. In the fooling rates in Table 2, we can see that our method has a much higher fooling rate for the different attacks. We chose to use a small epsilon of $\epsilon = \frac{3}{255}$ instead of the standard $\epsilon = \frac{8}{255}$ to highlight the difference in effectiveness. For higher epsilons the attack success rate becomes 100%, as can be seen in Figure 4.

Table 2: Attack success percentage of different DFMS methods on the Target Model trained on CIFAR10 when attack $\epsilon = \frac{3}{255}$. All attacks are evaluated on the Target Model. The Target Model row is a white-box attack, Proxy Model is a transfer-based black-box attack where the proxy is trained using the same data as the target model. The other DFMS methods provide trained student models which act as the proxy in transfer-based black-box attacks.

Attack	**Method**	**Untargeted Attacks**		**Targeted Attacks**	
		Probabilities	**Hard-Labels**	**Probabilities**	**Hard-Labels**
FGSM	*Target Model*	*45.00*		*19.64*	
	Proxy Model	*33.12*		*14.38*	
	DFME	56.84	39.22	21.07	17.15
	Dual Students	**62.35**	44.58	**21.58**	21.04
	DFMS-HL/SL	54.88	48.89	19.74	21.85
	DFMS-HL/SL + DS	54.99	**50.41**	20.53	**23.59**
BIM	*Target Model*	*96.74*		*76.90*	
	Proxy Model	*57.50*		*29.74*	
	DFME	84.38	55.52	53.39	31.95
	Dual Students	**91.56**	62.76	**63.39**	34.85
	DFMS-HL/SL	83.25	73.19	52.99	41.00
	DFMS-HL/SL + DS	82.23	**74.60**	51.89	**43.47**
PGD	*Target Model*	*96.78*		*76.32*	
	Proxy Model	*55.01*		*28.33*	
	DFME	83.59	54.71	51.97	31.49
	Dual Students	**91.04**	62.21	**61.96**	33.39
	DFMS-HL/SL	81.97	72.40	52.64	39.95
	DFMS-HL/SL + DS	81.04	**73.08**	51.38	**42.53**

Two observations of note in Table 2 are that the data-free methods have a higher attack success rate than the proxy model trained on the real dataset, and that the data-free methods outperform the white-box attack when using FGSM. For the former, it might seem counter intuitive that the student models significantly outperform the regularly trained proxy model. However, this result is expected, as the millions of queries to the target model used to train the student model result in the students having gradients more similar to the target model. For the latter, the DFMS methods having a higher fooling rate than the white-box setting is surprising. One explanation for this is that the FGSM attack is more susceptible to noise within the attacked-model's gradients due to only taking a single step in the direction of the signed gradient. The white-box model is a larger model trained on the real dataset, and thus has a higher tendency for noise within its gradients as compared with the student models. The student models, on the other hand, are trained on images designed to be near the class boundaries, as, if there is noise within the student model gradients, the generator will leverage that noise to cause disagreement between the students. PGD and BIM both take multiple smaller steps and are thus less susceptible to the noise which explains why the white-box setting is more optimal for those attacks.

5 Conclusion

In summary, outlined our new Data-Free Model Stealing optimization and outlined the Dual Students method. We showed the Dual Students method is effective in stealing the functionality of the target model as well as providing a good proxy from which to generate transfer-based adversarial attacks. Towards improving model stealing methods, reducing the number of queries required and extending to more complex datasets and tasks are major challenges to be explored in future work.

Ethics Statement: Advancing the model stealing area has the potential to improve the effectiveness of attacks from adversaries with bad intentions. However, for the data-free area specifically, this is not an immediate concern due to current DFMS methods being very limited when it comes to scaling to more useful or valuable problems. And, as with many areas of machine learning research, it is not clear that work in this area will not bring about benefits towards knowledge distillation and machine learning as a whole which may outweigh potentially exacerbating the model stealing threat.

Reproducibility Statement: We outline the general idea of our method and provide the key hyperparameters used during training. In the supplemental material we provide the algorithm as well as additional hyperparameters necessary for replicating the results shown in the tables. Because the dual students method does not require pretraining, other synthetic datasets, or models that require separate training routines, the environment can be setup and trained from scratch without using proprietary datasets or environments.

REFERENCES

Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, and Marius Popescu. Black-box ripper: Copying black-box models using generative evolutionary algorithms. *Advances in Neural Information Processing Systems*, 33:20120–20129, 2020.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*, 2017a.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*, 2017b.

Kuluhan Binici, Shivam Aggarwal, Nam Trung Pham, Karianto Leman, and Tulika Mitra. Robust and resource-efficient data-free knowledge distillation by generative pseudo replay. *arXiv preprint arXiv:2201.03019*, 2022.

Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 194–203, 2021.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Chi Hong, Jiyue Huang, and Lydia Y Chen. Mega: Model stealing via collaborative generator-substitute networks. *arXiv preprint arXiv:2202.00008*, 2022.

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*, pp. 13814–13823, 2021.

Manish Kesarwani, Bhaskar Mukhoty, Vijay Arya, and Sameep Mehta. Model extraction warning in mlaas paradigm. In *Proceedings of the 34th Annual Computer Security Applications Conference*, pp. 371–380, 2018.

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

Tommy Li and Cory Merkel. Model extraction and adversarial attacks on neural networks using switching power information. In *International Conference on Artificial Neural Networks*, pp. 91–101. Springer, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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 (CVPR)*, June 2019.

Xiaoyang Qu, Jianzong Wang, and Jing Xiao. Enhancing data-free adversarial distillation with activation regularization and virtual interpolation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3340–3344. IEEE, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Sunandini Sanyal, Sravanti Addepalli, and R. Venkatesh Babu. Towards data-free model stealing in a hard label setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15284–15293, June 2022.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pp. 1453–1460. IEEE, 2011.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

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*, pp. 4771–4780, 2021.

Gregor Urban, Krzysztof J Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shengjie Wang, Rich Caruana, Abdelrahman Mohamed, Matthai Philipose, and Matt Richardson. Do deep convolutional nets really need to be deep and convolutional? *arXiv preprint arXiv:1603.05691*, 2016.

Yixu Wang, Jie Li, Hong Liu, Yan Wang, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Black-box dissector: Towards erasing-based hard-label model stealing attack. *arXiv preprint arXiv:2105.00623*, 2021.

Zi Wang. Zero-shot knowledge distillation from a decision-based black-box model. In *International Conference on Machine Learning*, pp. 10675–10685. PMLR, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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*, pp. 8715–8724, 2020.

Honggang Yu, Kaichen Yang, Teng Zhang, Yun-Yun Tsai, Tsung-Yi Ho, and Yier Jin. Cloudleak: Large-scale deep learning models stealing through adversarial examples. In *NDSS*, 2020.

Xiaoyong Yuan, Leah Ding, Lan Zhang, Xiaolin Li, and Dapeng Oliver Wu. Es attack: Model stealing against deep neural networks without data hurdles. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.

Zhenrui Yue, Zhankui He, Huimin Zeng, and Julian McAuley. Black-box attacks on sequential recommenders via data-free model extraction. In *Fifteenth ACM Conference on Recommender Systems*, pp. 44–54, 2021.

Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Data-free universal adversarial perturbation and black-box attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7868–7877, 2021.

Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15115–15125, June 2022.

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 (CVPR)*, June 2020.