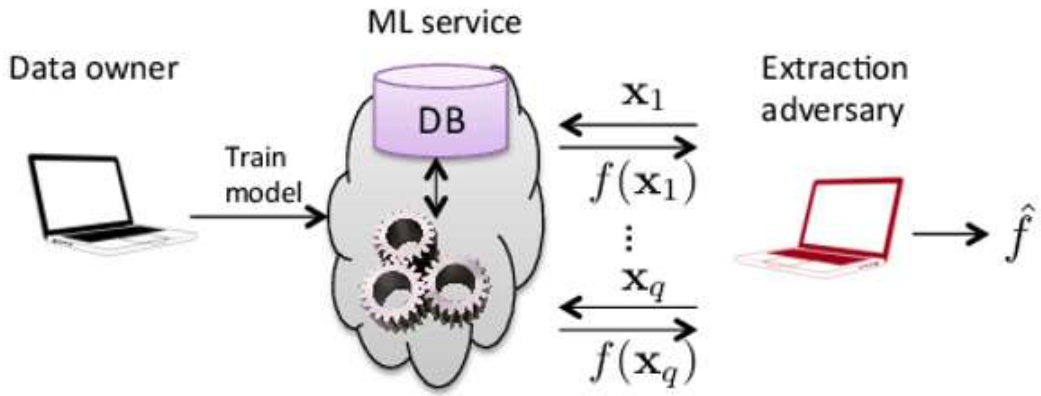
Dual Student Networks for Data-Free Model Stealing

James Beetham[1*], **Navid Kardan**[1*], **Ajmal Mian**[2], **Mubarak Shah**[1]
[1]Center for Research in Computer Vision
University of Central Florida
Orlando, Florida 32816, USA
[2]Department of Computer Science
University of Western Australia
Crawley WA 6009, Australia
{james.beetham, kardan}@knights.ucf.edu
ajmal.mian@uwa.edu.au
shah@crcv.ucf.edu

ICLR 2023

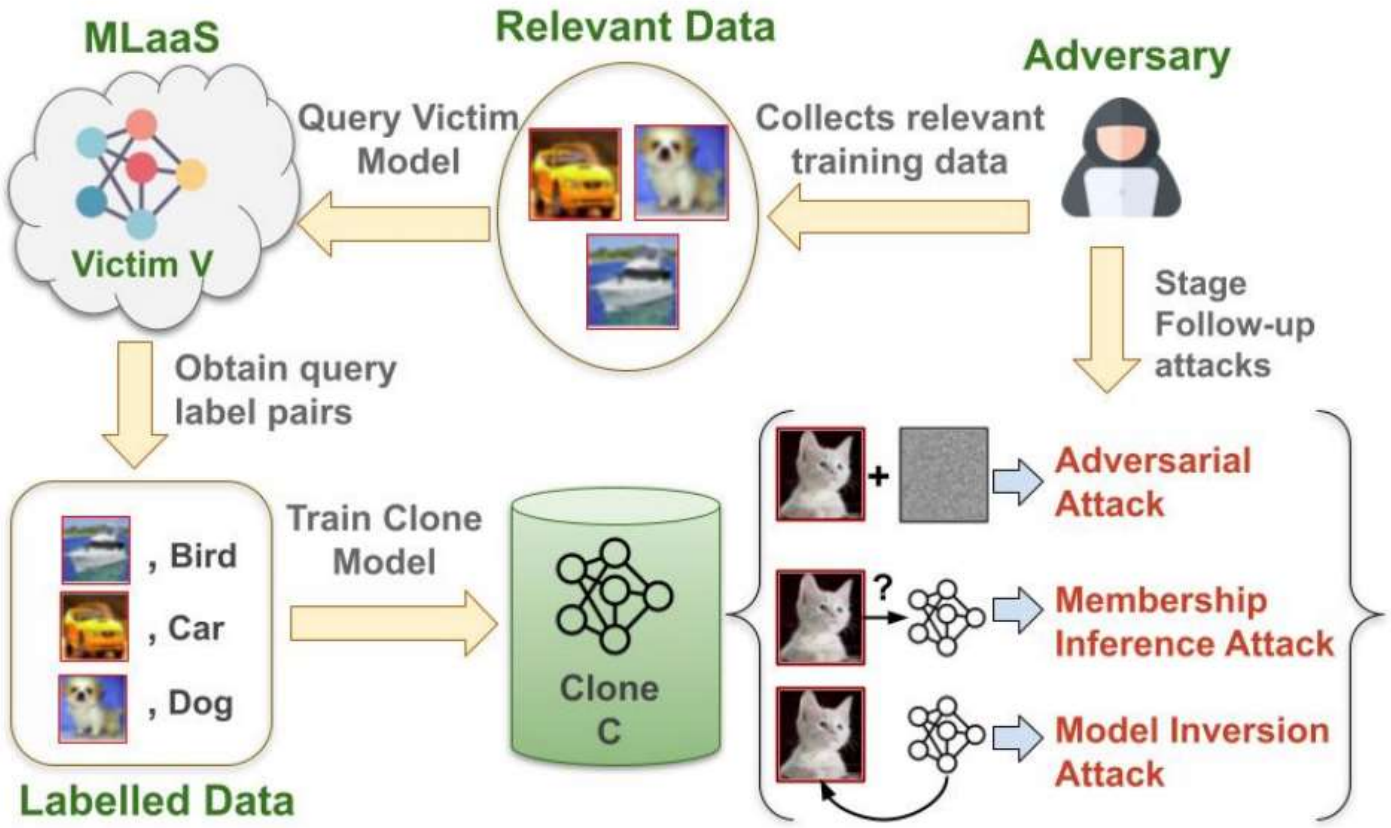
Model Stealing



Motivation:

1. Downstream adversarial attacks
2. Monetary gains

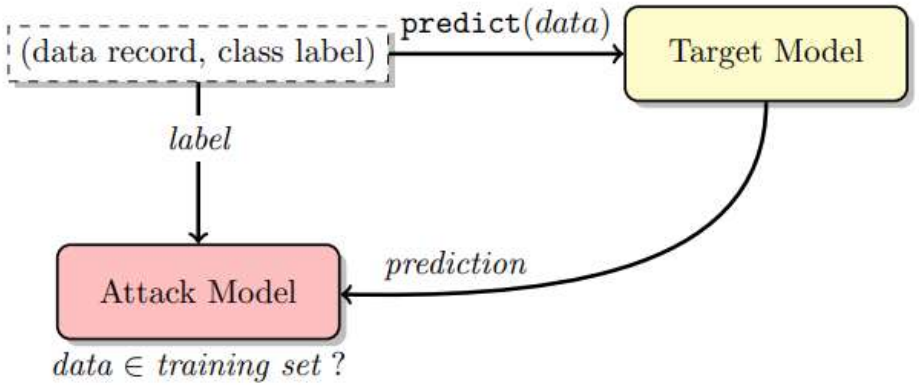
Downstream adversarial attacks



Attacks

Membership Inference Attack:

Refers to the black box access permission of a given data record and a model to determine whether the record is in the training data set of the model.



Model Inversion Attack:

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Data-free Model Stealing

knowledge distillation vs **black-box model stealing** vs **Data-free model stealing**

- **Same**: obtaining a student model which imitates the target model

- **Differences**:

1. Knowledge distillation environments typically retain full knowledge of the target model training data and weights;

2. Black-box model stealing eliminates the need to have access to the target model weights and training data (black-box model stealing typically uses real-world data samples to train the student network);

3. Data-free model stealing leveraging a generator to produce data samples.

Previous Data-free Model Stealing

- DFME Data-free model extraction (CVPR 2021)

The generator is optimized to maximize the distance between student and target model outputs.

- DFMS-HL Towards Data-Free Model Stealing in a Hard Label Setting (CVPR 2022)

The generator-discriminator is optimized to generate samples similar to a synthesized dataset which have balanced student classifications.

In both approaches, a student is optimized to minimize the distance between the student and target model outputs.

DUAL STUDENTS METHOD

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

$$\min_{S} \max_{G} ||T(G(z)) - S(G(z))||_p.$$

- For student S:

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

- For generator G:

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

- Due to the limited black-box access to the target model, we can't update G directly.
- Generator loss doesn't directly promote a diversity of classes within the generated images.

APPROXIMATING GRADIENT WITH DUAL STUDENTS

Towards making objective differentiable

- we propose adding an additional student S_2 to solve the following optimization problem:

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

- For generator G :

$$\max_{G} ||S_1(G(z)) - S_2(G(z))||_p.$$

This optimization for Generator G removes the Target model T and makes the problem directly differentiable.

APPROXIMATING GRADIENT WITH DUAL STUDENTS

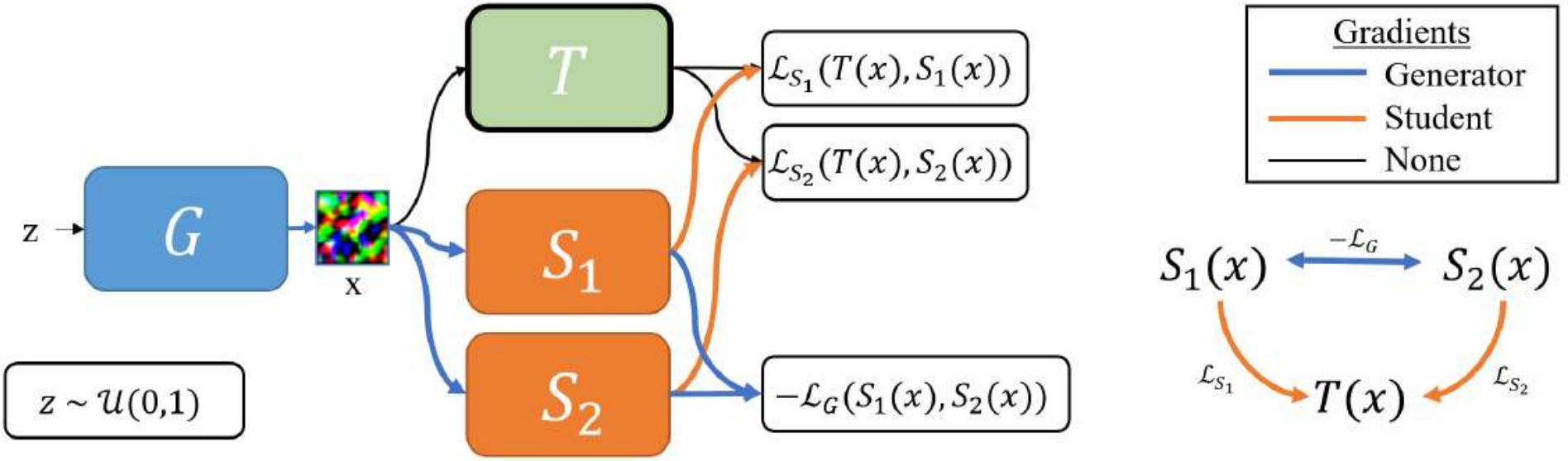
$$\max_G ||S_1(G(z)) - S_2(G(z))||_p. \qquad \text{Why?}$$

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

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

$$\max_x ||S_1(x) - S_2(x)||_p \leq \max_x ||S_1(x) - T(x)||_p + ||S_2(x) - T(x)||_p.$$

DUAL STUDENTS METHOD



Experiment

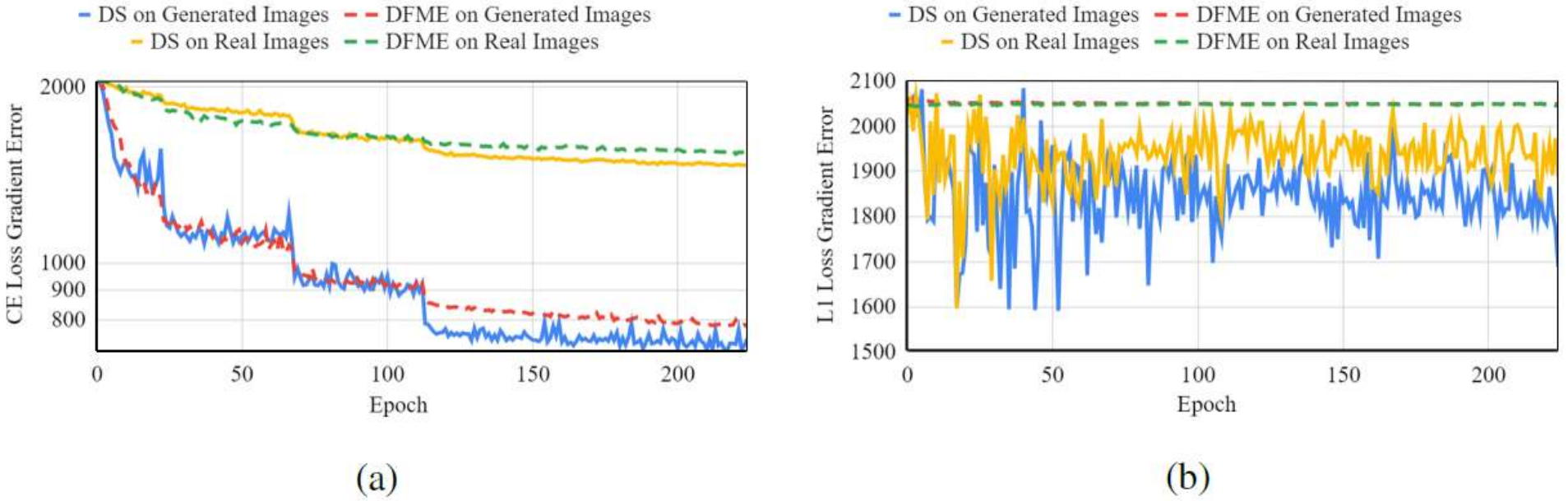


Figure 2: The distance between the true and estimated target model gradients of either Generated images or Real images. The gradient is computed w.r.t. input image x, and a formal description of the distance used is provided in Eq. 6. (a) Cross-Entropy loss is used to compute the gradient in the hard-label setting, and (b) ℓ_1 loss is used in the soft-label setting. Real images are from the test split of the CIFAR10 dataset.

Experiment

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	88.51	79.61
	95.5	DFMS-SL/HL + DS	89.38	**85.06**
CIFAR100	77.99	DFME	26.46	6.91
	77.99	DS	45.32	9.77
	77.99	DFMS-SL/HL	44.86	35.78
	77.99	DFMS-SL/HL + DS	**50.98**	**36.38**

Experiment

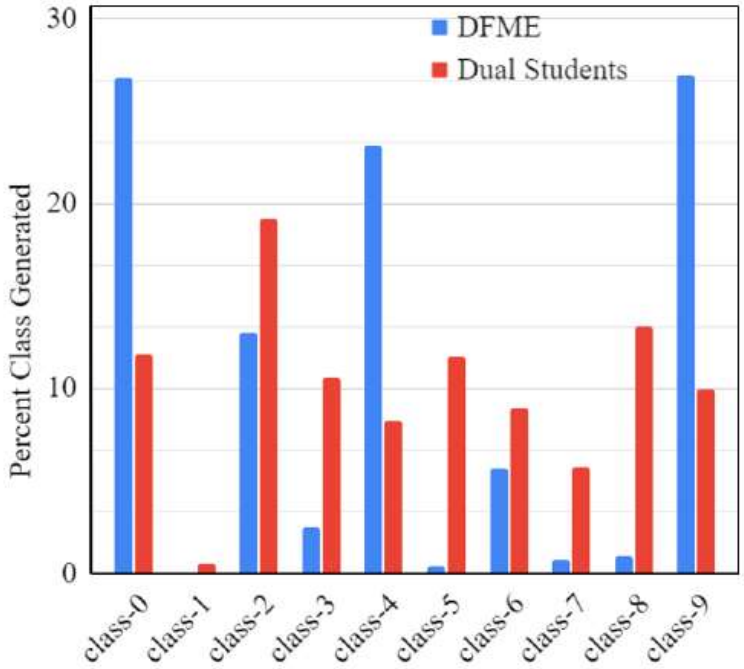
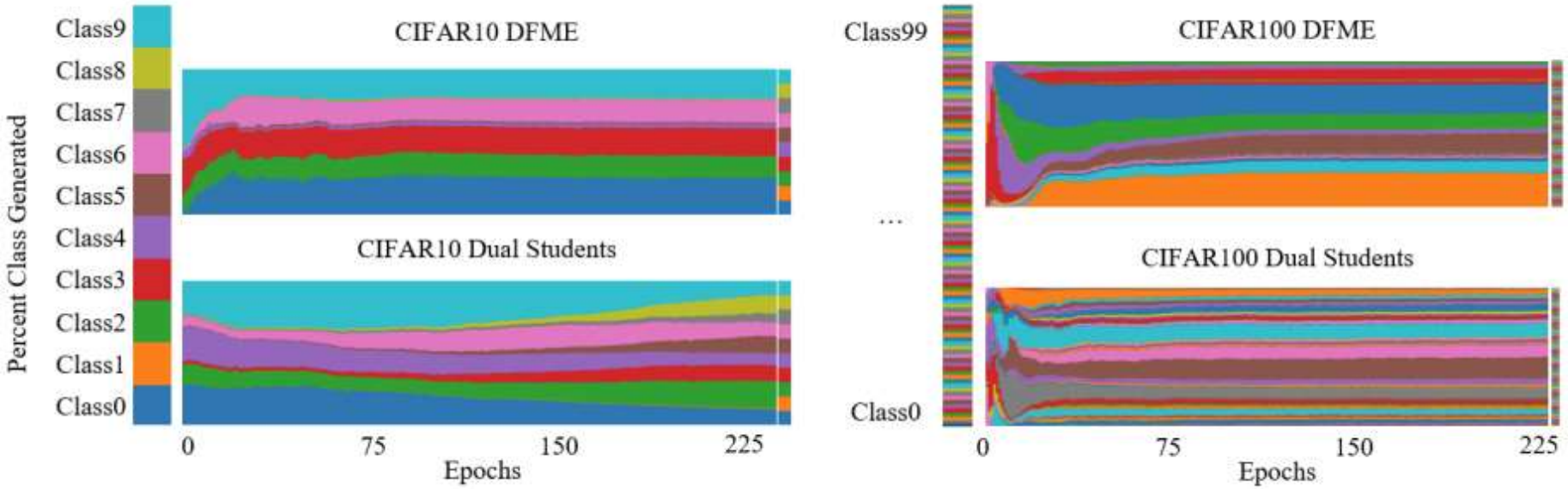


Figure 3: Class distribution of generated classes at the end of training for CIFAR10 dataset. Dual Student is able to produce more balanced samples.

Experiment


Class9
Class8
Class7
Class6
Class5
Class4
Class3
Class2
Class1
Class0
Percent Class Generated
CIFAR10 DFME
CIFAR10 Dual Students
0
75
150
225
Epochs
Class99
...
Class0
CIFAR100 DFME
CIFAR100 Dual Students
0
75
150
225
Epochs

Experiment

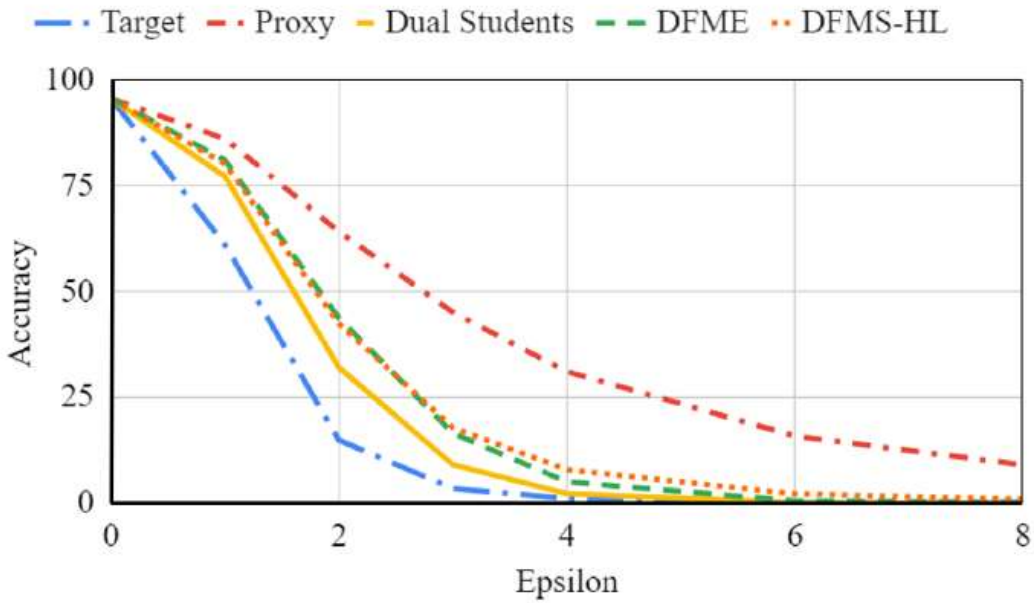


Figure 4: Accuracy on Target Model of perturbed images generated using PGD attack with varying epsilons ($\frac{\epsilon}{255}$) where different soft-label DFMS student models are used.

Experiment

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
	DS	**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**
PGD	*Target Model*	*96.78*		*76.32*	
	Proxy Model	*55.01*		*28.33*	
	DFME	83.59	54.71	51.97	31.49
	DS	**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**

Thanks