Data from Model: Extracting Data from Non-robust and Robust Models

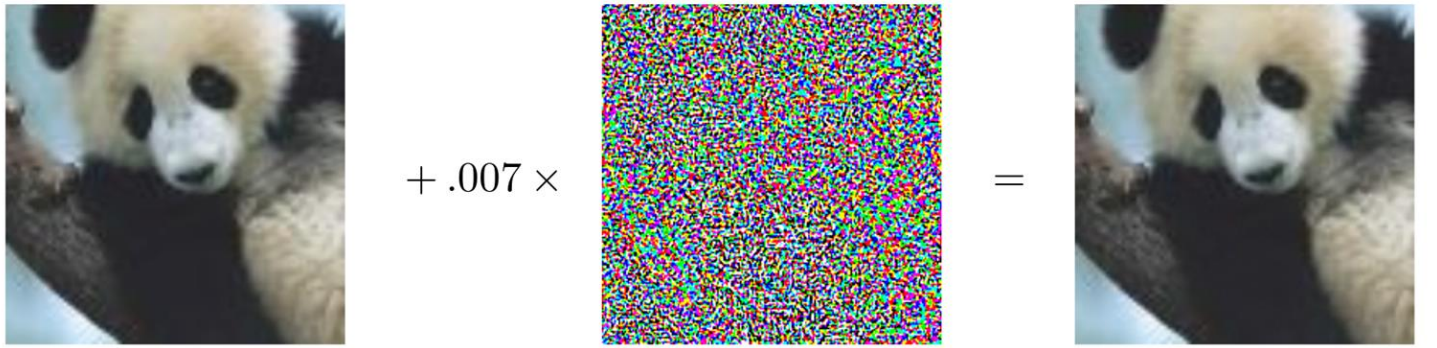
Philipp Benz*, Chaoning Zhang*, Tooba Imtiaz, In-So Kweon
* indicates equal contribution

Korea Advanced Institute of Science and Technology (KAIST)

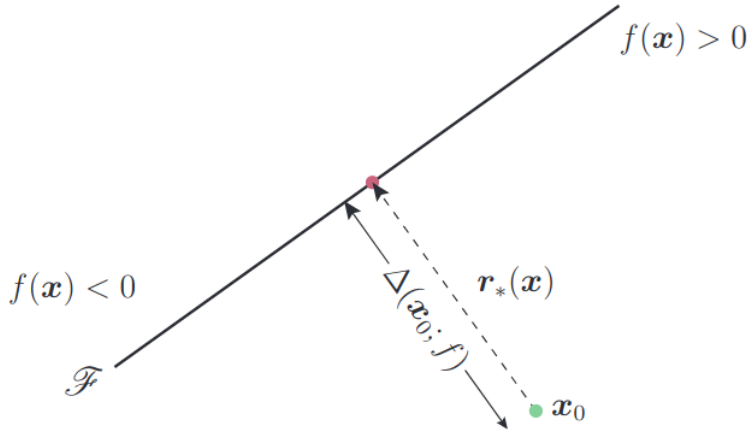
Adversarial Examples

Deep Neural Networks are sensitive to small perturbations in the image, which can lead to misclassifications. These changes are mostly imperceptible for human observers.

Image-dependant Adversarial Perturbations [1,2]



DeepFool [3]



Adversarial examples for a linear binary classifier

[1] Intriguing properties of neural networks; Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus; ArXiv 2013
[2] Explaining and Harnessing Adversarial Examples; Goodfellow, Shlens, Szegedy; ICLR 2015
[3] DeepFool: a simple and accurate method to fool deep neural networks; Moosavi-Dezfooli, Fawzi, Frossard; CVPR 2016

Adversarial Examples

Adversarial Examples can be directly attributed to the presence of non-robust features [1]

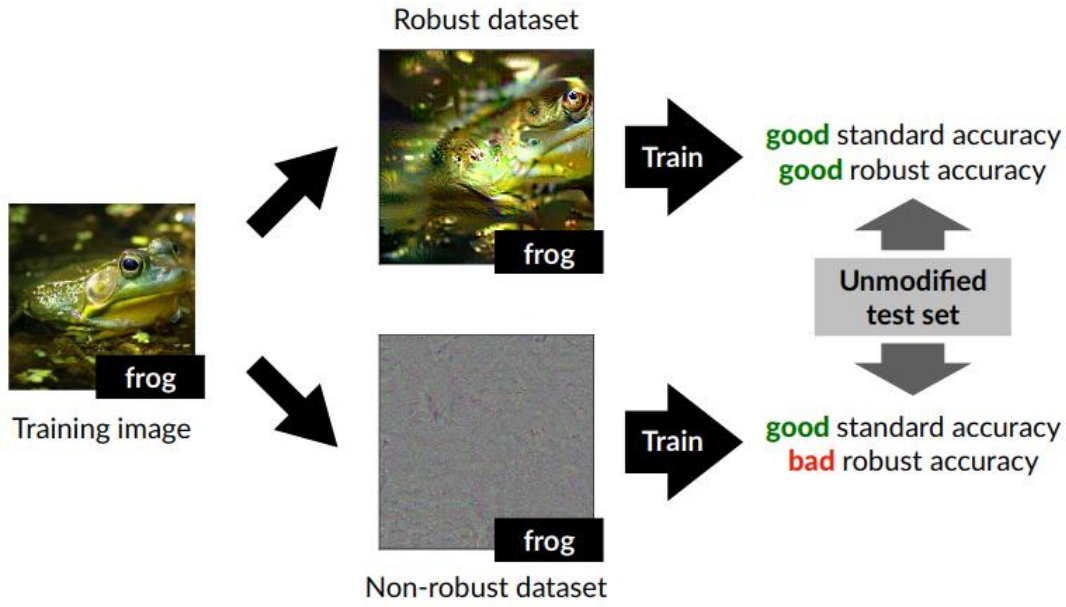


Figure 1: Disentanglement of features into combinations of robust and non-robust features [1]

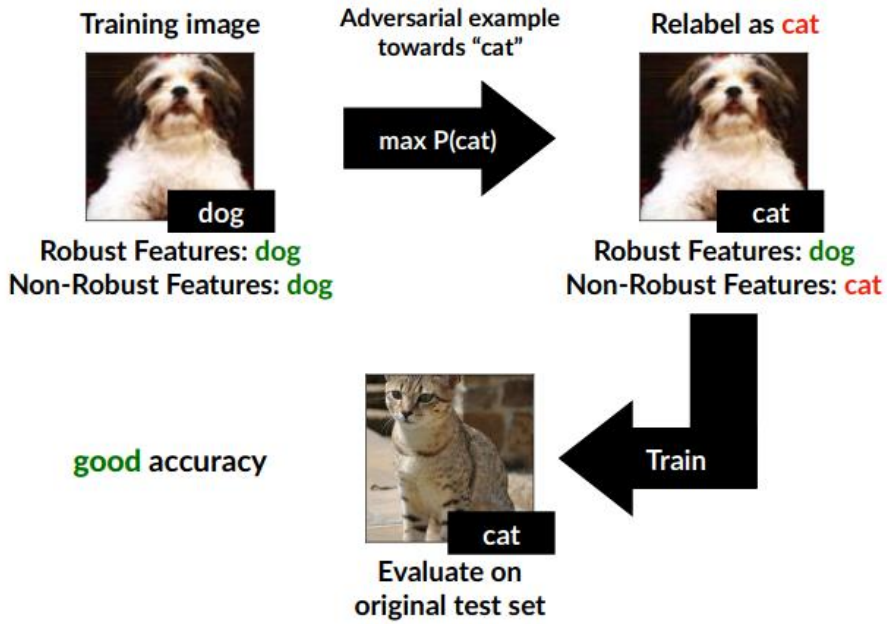
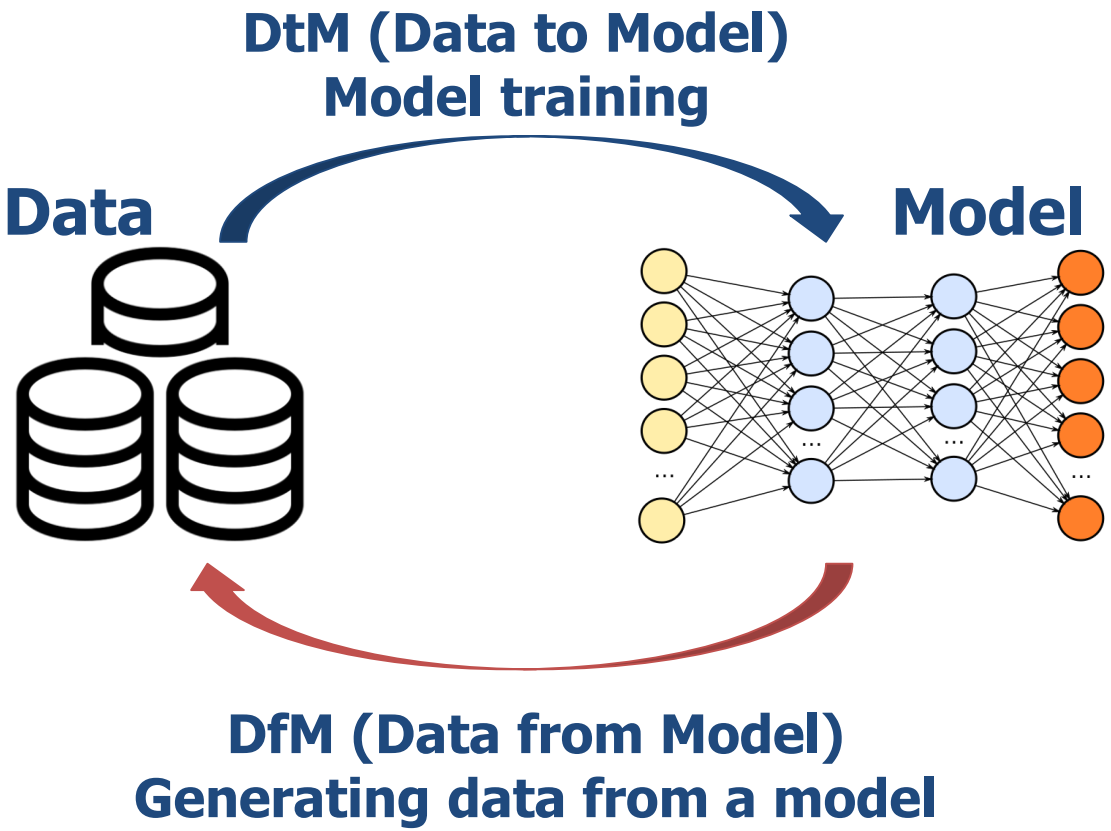


Figure 2: A dataset appearing mislabeled to humans (via adversarial examples) can results in good accuracy on the original test set [1]

[1] Adversarial Examples Are Not Bugs, They Are Features; Ilyas, Santurkar, Tsipras, Engstrom, Tran, Madry; NeurIPS 2019

Data to Model & Data from Model



Differentiation from [1]:

- **Sequential iteration** of the DtM and DfM process and exploration of the non-robust and robust behavior

$$D_0 \xrightarrow{M_0} D_1 \xrightarrow{M_1} D_2 \xrightarrow{M_2} \cdots \cdots \xrightarrow{M_N} D_N$$

- **Restricted access** to either the data or the model generated in the previous step
- In the DfM process, we leverage a s**ubstitute dataset** and **virtual logits**
- In the subsequent DtM step we train the new model with the previously obtained images and logits utilizing the **MSE loss** instead of cross entropy

[1] Adversarial Examples Are Not Bugs, They Are Features; Ilyas, Santurkar, Tsipras, Engstrom, Tran, Madry; NeurIPS 2019

DtM & DfM chain

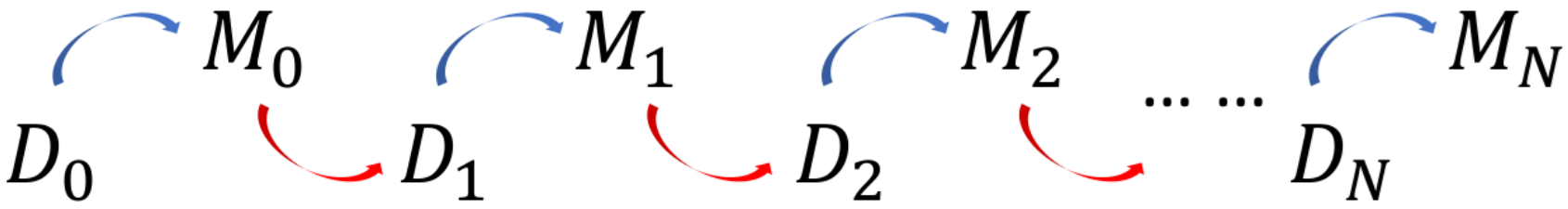


Figure 1: The chain of performing DtM and DfM repetitively. The blue arrows indicate the DtM process and the red arrows indicate the DfM process.

Table 1: Evaluation of the models obtained through the sequential DfM process on the original validation dataset

	LeNet (MNIST)		VGG8 (CIFAR10)	
	non-robust $\mathcal{M}_0$	robust $\mathcal{M}_0$	non-robust $\mathcal{M}_0$	robust $\mathcal{M}_0$
$\mathcal{M}_0$	99.5	98.7	92.2	87.0
$\mathcal{M}_1$	98.5	97.7	89.4	88.1
$\mathcal{M}_2$	96.6	96.1	80.1	82.5
$\mathcal{M}_3$	91.5	95.2	66.8	71.7
$\mathcal{M}_4$	87.4	94.2	52.5	58.7
$\mathcal{M}_5$	76.5	93.7	27.5	44.8

Table 2: Robustness evaluation of the models obtained during the chaining process for a non-robust (left) and robust (right) origin model. The results are reported for the LeNet architecture on MNIST.

	non-robust $\mathcal{M}_0$					robust $\mathcal{M}_0$				
ϵ	0	1	2	3	4	0	1	2	3	4
$\mathcal{M}_0$	99.5	74.2	4.5	0.1	0.1	98.7	91.0	58.9	10.6	0.7
$\mathcal{M}_1$	98.5	61.3	1.4	0.0	0.0	97.7	81.1	25.4	1.1	0.0
$\mathcal{M}_2$	96.6	31.2	0.2	0.0	0.0	96.1	71.2	12.4	0.1	0.0
$\mathcal{M}_3$	91.5	6.5	0.0	0.0	0.0	95.2	54.2	3.24	0.2	0.1
$\mathcal{M}_4$	87.4	1.1	0.0	0.0	0.0	94.2	31.8	1.0	0.1	0.3
$\mathcal{M}_5$	76.5	0.0	0.0	0.0	0.0	93.7	13.3	0.6	0.2	0.2

Decreasing Robustness

Qualitative Results

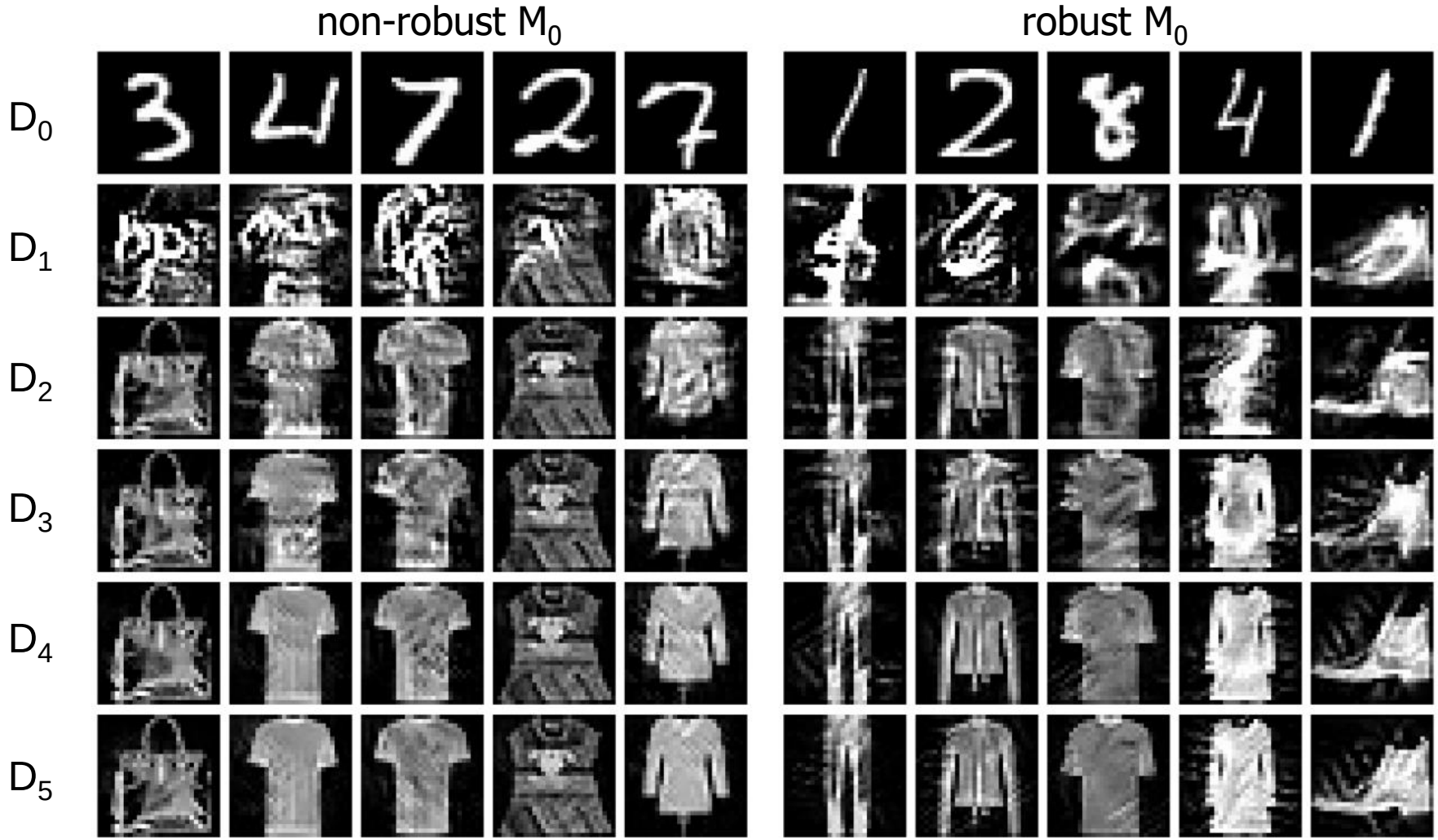


Figure 1: Qualitative results for the DfM process starting from a non-robust origin model (5 columns on the left) and a robust origin model (5 columns on the right). The results are shown for the LeNet architecture with Fashion MNIST as the background images.

Cross-Training

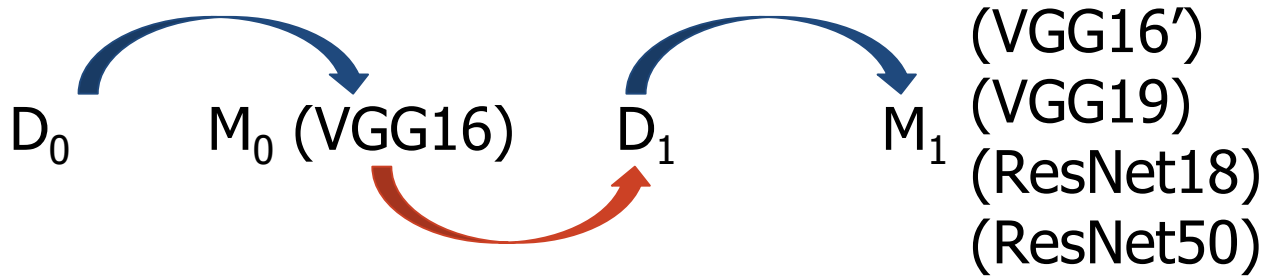


Figure 1: Schematic overview of the cross-training process

Table 1: Cross-training of the extracted datasets from non-robust (top) and robust (bottom) models. The models were originally trained on CIFAR10. The robust models were adversarially trained with the l2 variant of PGD. The rows indicate the model from which the data was extracted. The columns indicate the trained model. The values indicate the accuracy of the CIFAR-10 test dataset.

		VGG16	VGG19	ResNet18	ResNet50
non-rob.	VGG16 (93.8)	89.6	90.1	89.9	90.3
	VGG19 (93.6)	89.7	90.1	90.6	90.3
	ResNet18 (95.1)	87.9	88.0	89.7	89.6
robust	VGG16 (88.7)	90.3	90.5	90.3	90.5
	VGG19 (87.6)	87.9	88.0	88.0	88.1
	ResNet18 (90.2)	91.3	91.1	91.6	91.5

Cross-Evaluation

Figure 1: Schematic overview of the cross-evaluation process

Table 1: Cross-evaluation of datasets extracted from non-robust (top) and robust (bottom) models. The models were originally trained on CIFAR-10. The diagonal values were obtained for the same architecture but a different training run. The accuracy of the extracted data on the original model is 100%.

		VGG16	VGG19	ResNet18	ResNet50
non-robust	VGG16	43.2	40.4	36.5	30.5
	VGG19	50.1	48.7	45.4	37.8
	ResNet18	36.9	34.9	55.2	43.8
	ResNet50	41.3	40.6	60.0	62.6
robust	VGG16	48.5	43.6	45.4	44.0
	VGG19	38.3	38.9	36.9	36.0
	ResNet18	42.0	37.8	50.1	47.7
	ResNet50	35.4	31.9	39.7	35.7

Summary

- We introduced DfM (Data from Model) by extracting data back from the model using a substitute dataset and virtual logits

- Performance and robustness of models can be partially sustained over multiple iterations of DtM/DfM. Moreover, there is a trend for the robust features to shift towards non-robust features

- Different model architectures can be trained on the feature mappings from other architectures. Different models learn a portion of shared features, as well as different individual features

A better understanding of the relationship between data and model with features as the link.