Appendix A

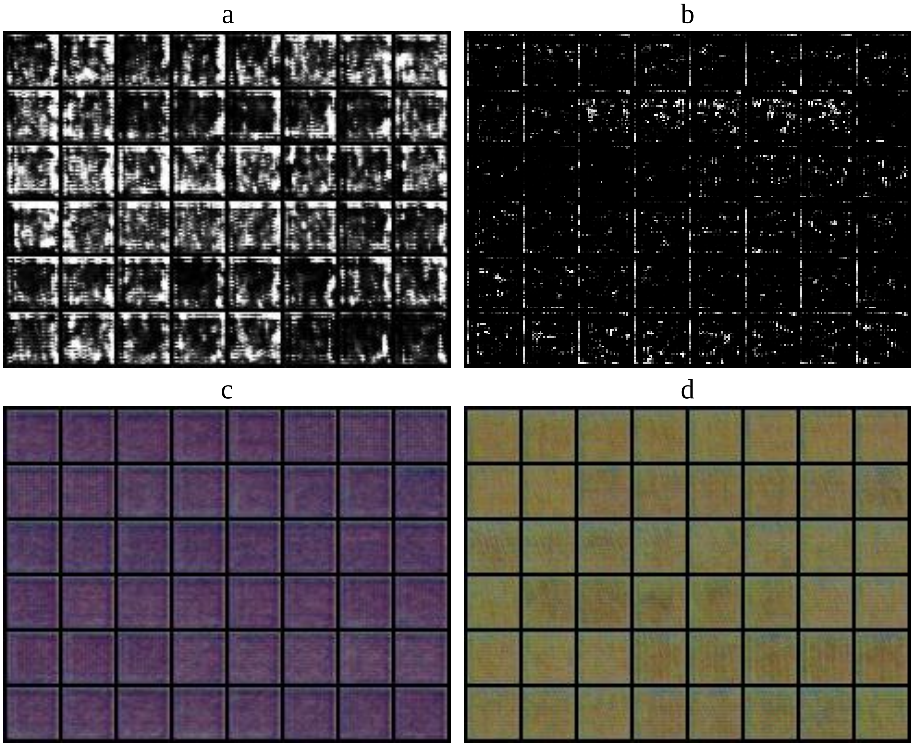

Figure 1: Visualization of the synthetic samples generated by the generator in DaST. These samples are produced by the GANs which are converged in training. We generate 48 samples for each scenario. a and b: samples generated by DaST in label-only and probability-only attack scenario on MNIST dataset (the medium network is the target model), respectively. c and d: samples generated by DaST in label-only and probability-only attack scenario on CIFAR-10 dataset (VGG-16 is the target model), respectively.

Appendix B

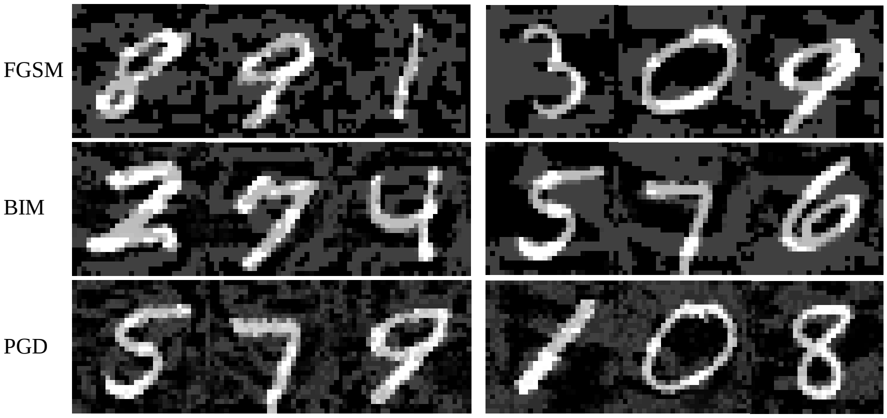

Figure 2: Visualization of the adversarial examples generated by DaST for attacking the medium model on MNIST. We only show the adversarial examples which can fool the model successfully. Left part: examples generated by DaST-P. Right part: examples generated by DaST-L.), respectively.

Appendix C

Table 1: Network architectures for MNIST. Convolutional kernel $(A \times B, C)$ denotes the kernel size and channel number, respectively.

ConvBlock	Small net	Medium net	Large net
ConvLayer $(A \times B, C)$	ConvBlock $(5 \times 5, 20)$	ConvBlock $(5 \times 5, 20)$	ConvBlock $(5 \times 5, 20)$
ReLU	ConvBlock $(5 \times 5, 50)$	ConvBlock $(5 \times 5, 50)$	ConvBlock $(5 \times 5, 50)$
MaxPooling (2×2)	DenseLayer	ConvBlock $(3 \times 3, 50)$	ConvBlock $(3 \times 3, 50)$
	ReLU	DenseLayer	ConvBlock $(3 \times 3, 50)$
	DenseLayer	ReLU	DenseLayer
	Sigmoid	DenseLayer	ReLU
		Sigmoid	DenseLayer
			Sigmoid