Supplementary material for *Decoupling Direction and Norm for Efficient Gradient-Based L_2 Adversarial Attacks and Defenses*

1. Model architectures

Table 1 lists the architectures of the CNNs used in the Attack Evaluation - we used the same architecture as in [1] for a fair comparison against the C&W and DeepFool attacks. Table 2 lists the architecture used in the robust model (defense) trained on CIFAR-10. We used a Wide ResNet with 28 layers and widening factor of 10 (WRN-28-10). The residual blocks used are the "basic block" [2, 4], with stride 1 for the first group and stride 2 for the second an third groups. This architecture is slightly different from the one used by Madry *et al*. [3], where they use a modified version of Wide ResNet with 5 residual blocks instead of 4 in each group, and without convolutions in the residual connections (when the shape of the output changes, e.g. with stride=2).

Layer Type	Size
Convolution	$3 \times 3 \times 16$
Residual Block	$\begin{bmatrix} 3 \times 3, 160 \\ 3 \times 3, 160 \end{bmatrix} \times 4$
Residual Block	$\begin{bmatrix} 3 \times 3, 320 \\ 3 \times 3, 320 \end{bmatrix} \times 4$
Residual Block	$\begin{bmatrix} 3 \times 3, 640 \\ 3 \times 3, 640 \end{bmatrix} \times 4$
Batch Normalization + ReLU	-
Average Pooling	8×8
Fully Connected + Softmax	10

Table 2: CIFAR-10 architecture used for the Defense evaluation

2. Hyperparameters selected for the C&W attack

We considered a scenario of running the C&W attack with 100 steps and a fixed C (1×100), and a scenario of running 4 search steps on C, of 25 iterations each (4×25). Since the hyperparameters proposed in [1] were tuned for a larger number of iterations and search steps, we performed a grid search for each dataset, using learning rates in the range [0.01, 0.05, 0.1, 0.5, 1], and C in the range [0.001, 0.01, 0.1, 1, 10, 100, 1 000]. We selected the hyperparameters that resulted in targeted attacks with lowest Median L_2 for each dataset. Table 3 lists the hyperparameters found through this search procedure.

Dataset	# Iterations	Parameters
MNIST	1×100	$\alpha = 0.1, C = 1$
MNIST	4×25	$\alpha = 0.5, C = 1$
CIFAR-10	1×100	$\alpha = 0.01, C = 0.1$
CIFAR-10	4×25	$\alpha = 0.01, C = 0.1$
ImageNet	1×100	$\alpha = 0.01, C = 1$
ImageNet	4×25	$\alpha = 0.01, C = 10$

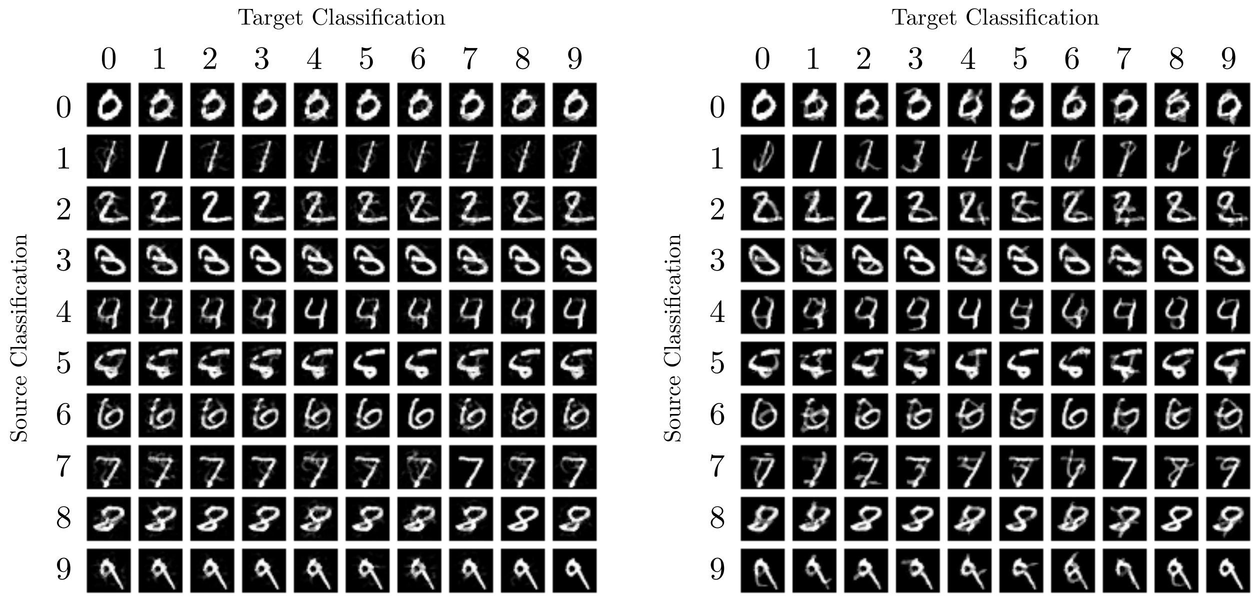
Table 3: Hyperparameters used for the C&W attack when restricted to 100 iterations.

Layer Type	MNIST Model	CIFAR-10 Model
Convolution + ReLU	$3 \times 3 \times 32$	$3 \times 3 \times 64$
Convolution + ReLU	$3 \times 3 \times 32$	$3 \times 3 \times 64$
Max Pooling	2×2	2×2
Convolution + ReLU	$3 \times 3 \times 64$	$3 \times 3 \times 128$
Convolution + ReLU	$3 \times 3 \times 64$	$3 \times 3 \times 128$
Max Pooling	2×2	2×2
Fully Connected + ReLU	200	256
Fully Connected + ReLU	200	256
Fully Connected + Softmax	10	10

Table 1: CNN architectures used for the Attack Evaluation

3. Examples of adversarial images

Fig. 1 plots a grid of attacks (obtained with the C&W attack) against the first 10 examples in the MNIST dataset. The rows indicate the source classification (label), and the columns indicate the target class used to generate the attack (images on the diagonal are the original samples). We can see that in the adversarially trained model, the attacks need to introduce much larger changes to the samples in order to make them adversarial, and some of the adversarial samples visually resemble another class.



(a) Baseline (without adversarial training)

(b) Adversarially trained

Figure 1: Adversarial examples obtained using the C&W L_2 attack on two models: (a) Baseline, (b) model adversarially trained with our attack.

Fig. 2 shows randomly-selected adversarial examples for the CIFAR-10 dataset, comparing the baseline model (WRN 28-10), the Madry defense and our proposed defense. For each image and model, we ran three attacks (DDN 1 000, C&W 9×10 000, DeepFool 100), and present the adversarial example with minimum L_2 perturbation among them. Fig. 3 shows cherry-picked adversarial examples on CIFAR-10, that visually resemble another class, when attacking the proposed defense. We see that on the average case (randomly-selected), adversarial examples against the defenses still require low amounts of noise (perceptually) to induce misclassification. On the other hand, we notice that on adversarially trained models, some examples do require a much larger change on the image, making it effectively resemble another class.

4. Attack performance curves

Fig. 4 reports curves of the perturbation size against accuracy of the models for three attacks: Carlini 9×10 000, DeepFool 100 and DDN 300.

References

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*, 2018.

[4] S. Zagoruyko and N. Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, pages 87.1–87.12, 2016.

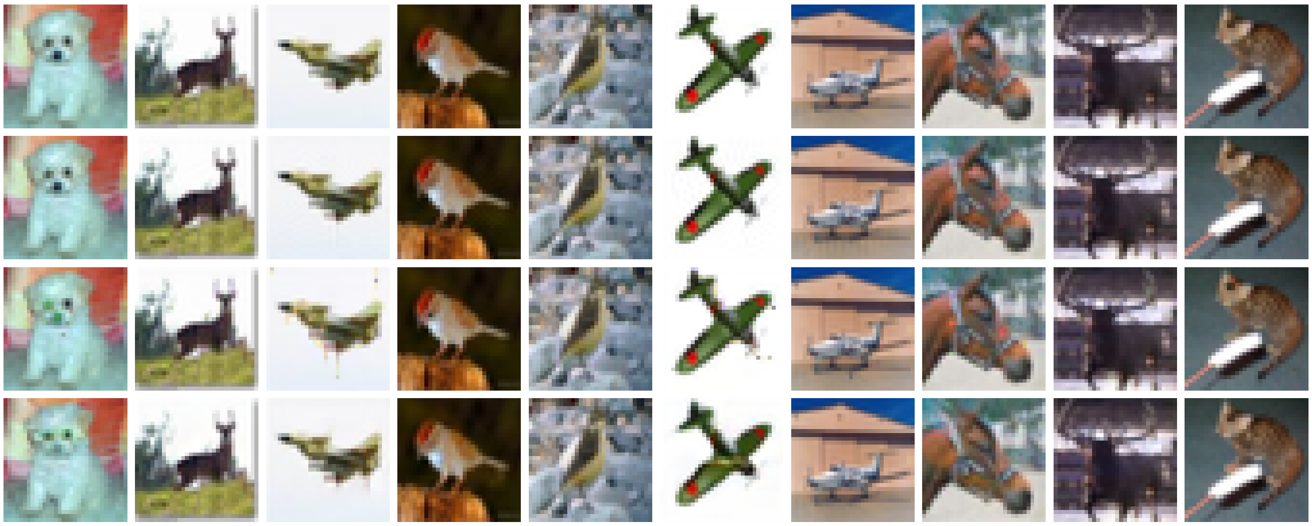


Figure 2: Randomly chosen adversarial examples on CIFAR-10 for three models. **Top row**: original images; **second row**: attacks against the baseline; **third row**: attacks against the Madry defense.

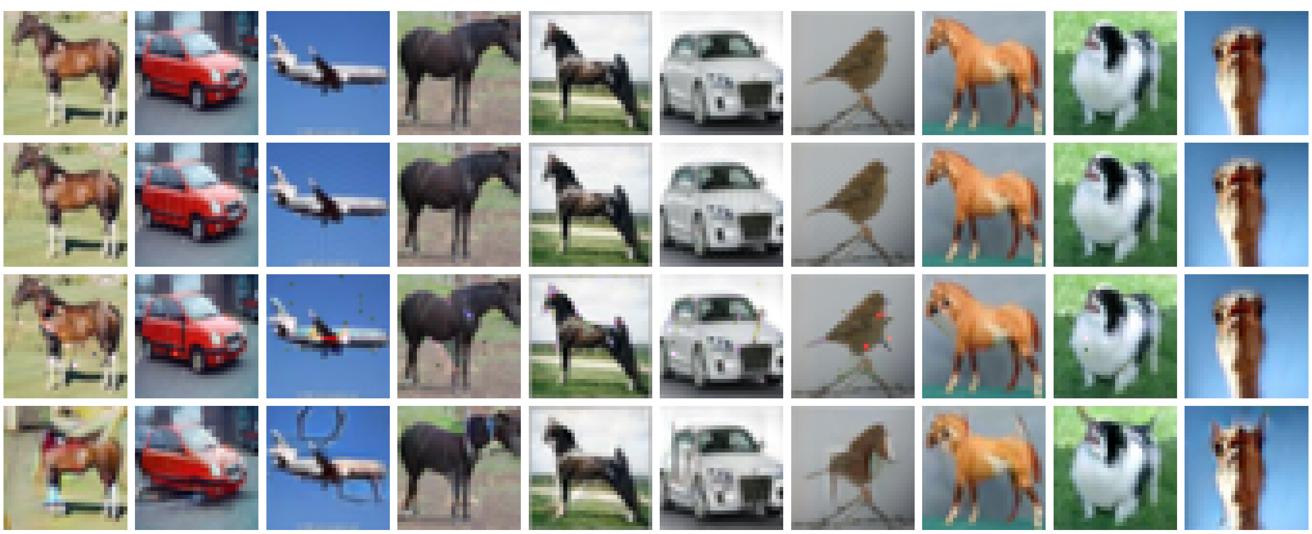
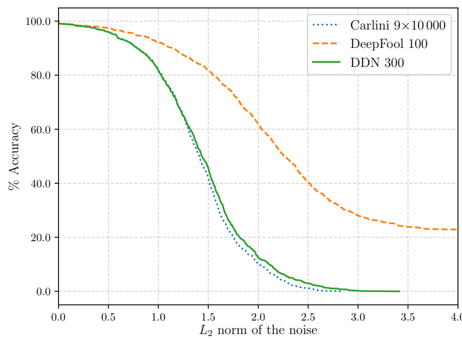
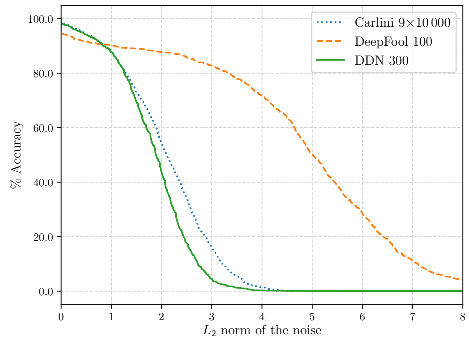


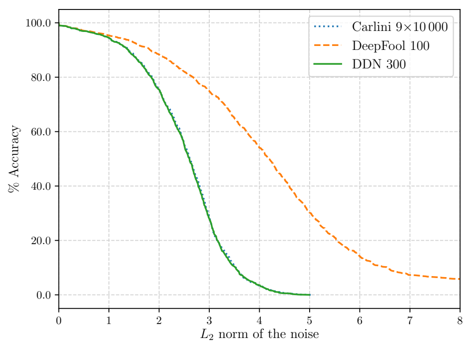
Figure 3: Cherry-picked adversarial examples on CIFAR-10 for three models. **Top row**: original images; **second row**: attacks against the baseline; **third row**: attacks against the Madry defense; **bottom row**: attacks against the proposed defense. Predicted labels for the last row are, from left to right: dog, ship, deer, dog, dog, truck, horse, dog, cat, cat.



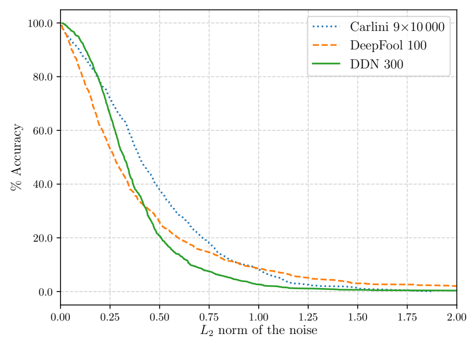
(a) MNIST / Baseline model.



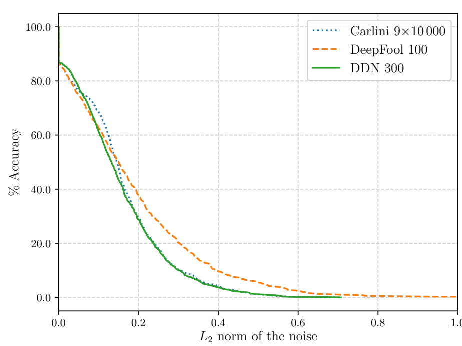
(b) MNIST / Madry defense.



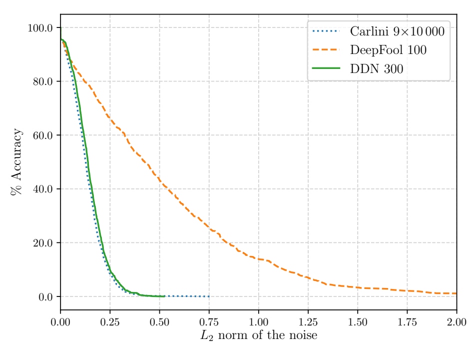
(c) MNIST / Our Defense



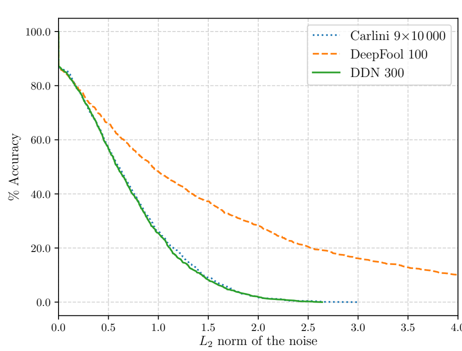
(d) ImageNet / Inception V3.



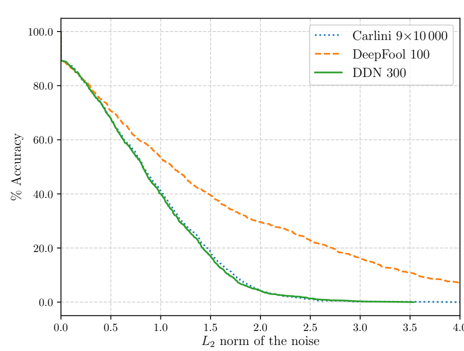
(e) CIFAR-10 / Baseline model.



(f) CIFAR-10 / Baseline WRN 28-10.



(g) CIFAR-10 / Madry defense.



(h) CIFAR-10 / Our Defense.

Figure 4: Attacks performances on different datasets and models.