Safeguarding Attention With Diffusion Denoised Smoothing

Final Presentation

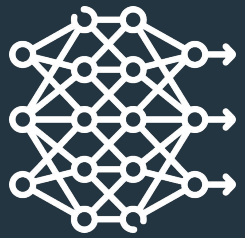
Matthew Morgan, Gavin Lesher, Ethan Nechanicky

Research Problem

- ***Is diffusion denoising a viable defense to adversarial attacks on attention-based image classifiers?***

Experiment Design

PGD Samples

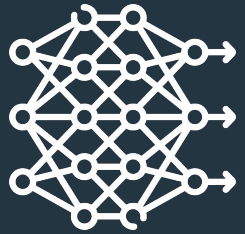


Attention Network

Measure Accuracy

VS.

Denoised Smoothing



Attention Network

Measure Accuracy

Threat Model

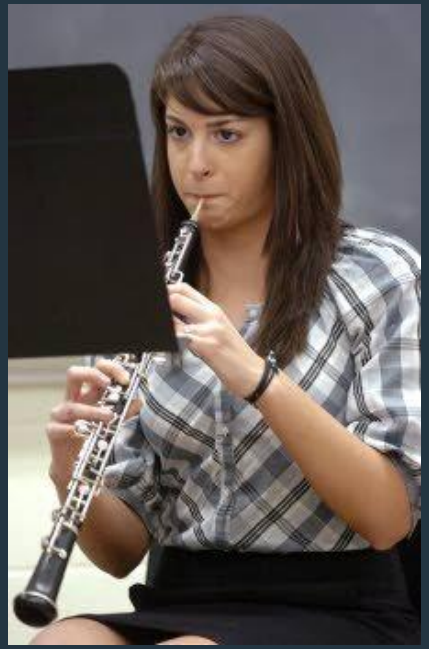
- **Mostly White-Box**
  - We assume attacker has white-box knowledge of classifier, but not diffusion denoiser.
    - Diffusion denoiser: “(Certified!!) Adversarial Robustness for Free!” [1]
  - Our experiment is designed such that a diffuser can be “added” independently from the classifier.

Experimental Design

- **ImageNet Model: coatnet_rmlp_2_rw_384.sw_in12k_ft_in1k from Hugging Face [3]**
- Pre-trained on ImageNet-12k (a 11821 class subset of full ImageNet-22k) and fine-tuned on ImageNet-1k by Ross Wightman.
- ImageNet-1k Results:
    - 87.39% Top-1, 98.31% Top-5

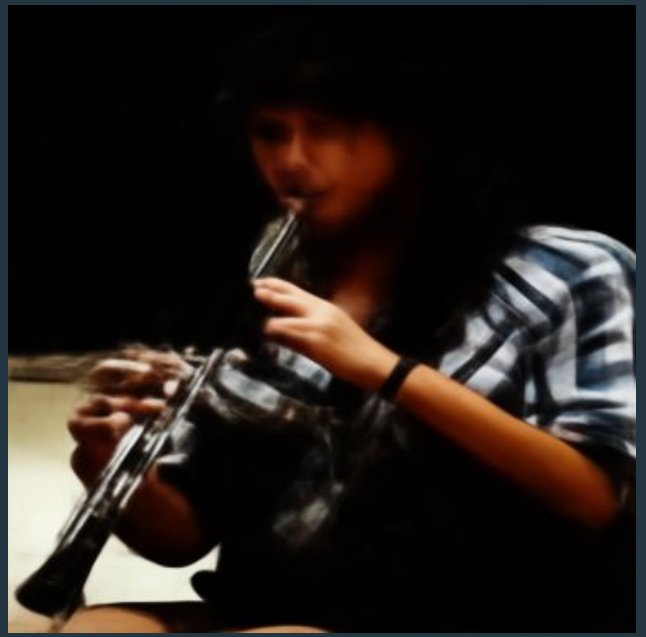
- **CIFAR-10 Model: ResNet50 & ResNet50+CBAM [2]**
- Trained for 20 epochs
- Batch size = 10, lr = 1e-4, decay rate = 0.98, batch size = 10

$\sigma = 1.5 \quad \varepsilon = 0.01$



Clean
True 683
Oboe

PGD
Pred 683
Oboe



Denoised
Pred 683
Oboe

σ =0.5 ε = 0.05

Clean
True 459
brassiere

PGD
Predicted 638
maillot

Denoised
Predicted 459
brassiere

σ =0.25 ε = 0.01

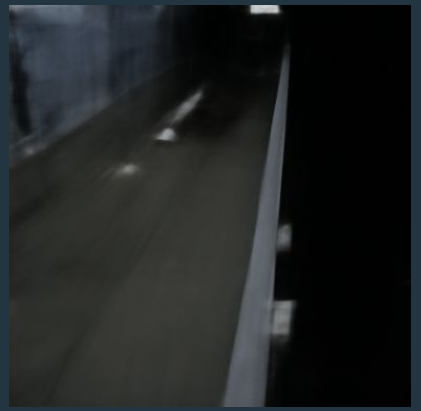
Clean
True 239
Bernese_mountain
_dog

PGD
Predicted 238
Greater-Swiss_Mountain_dog

Denoised
Predicted 239
Bernese_mountain
_dog

Clean:

σ = 0



σ = 0.25

σ = 0.5

σ = 1.0

σ = 1.5

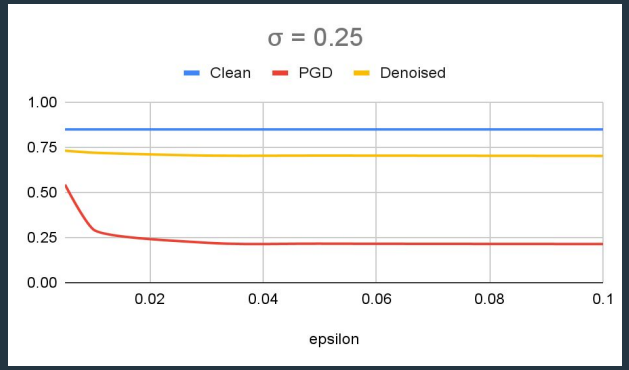
All attacked with ε = 0.1

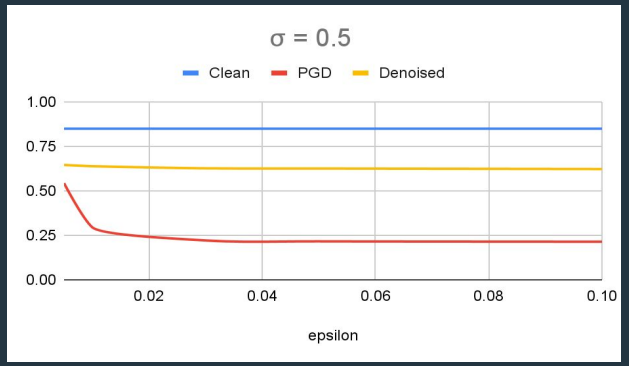
Results

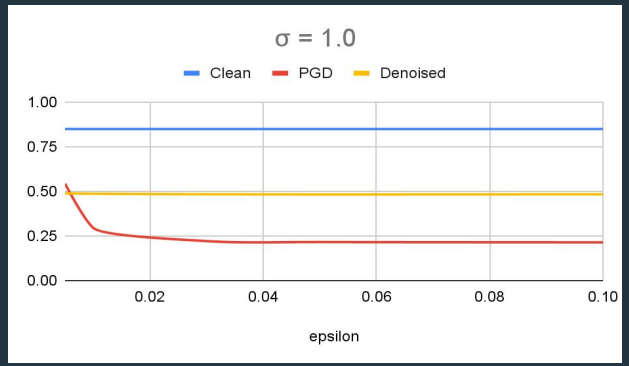
- ***1000 validation images***

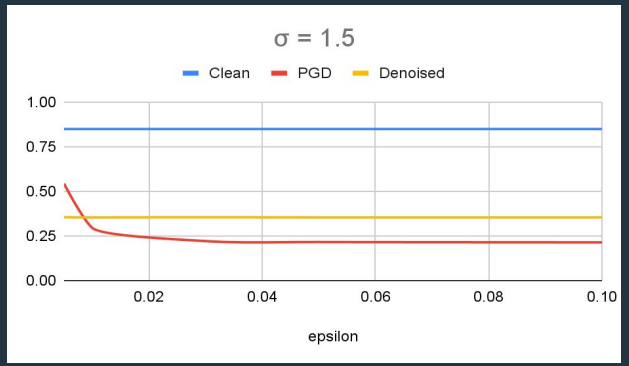
Sigma	Epsilon	Clean	PGD	Denoised
0.25	0.005	0.851	0.545	0.733
	0.01	0.851	0.296	0.722
	0.03	0.851	0.222	0.706
	0.05	0.851	0.217	0.706
	0.1	0.851	0.215	0.704
	0.5	0.851	0.167	0.702
0.5	0.005	0.851	0.545	0.647
	0.01	0.851	0.296	0.64
	0.03	0.851	0.222	0.628
	0.05	0.851	0.217	0.627
	0.1	0.851	0.215	0.624
	0.5	0.851	0.167	0.62
1	0.005	0.851	0.545	0.49
	0.01	0.851	0.296	0.488
	0.03	0.851	0.222	0.485
	0.05	0.851	0.217	0.484
	0.1	0.851	0.215	0.485
	0.5	0.851	0.167	0.481
1.5	0.005	0.851	0.545	0.356
	0.01	0.851	0.296	0.355
	0.03	0.851	0.222	0.356
	0.05	0.851	0.217	0.355
	0.1	0.851	0.215	0.355
	0.5	0.851	0.167	0.354

Results - ImageNet

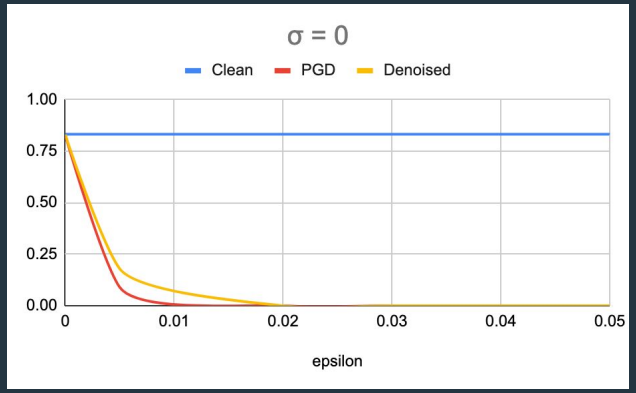


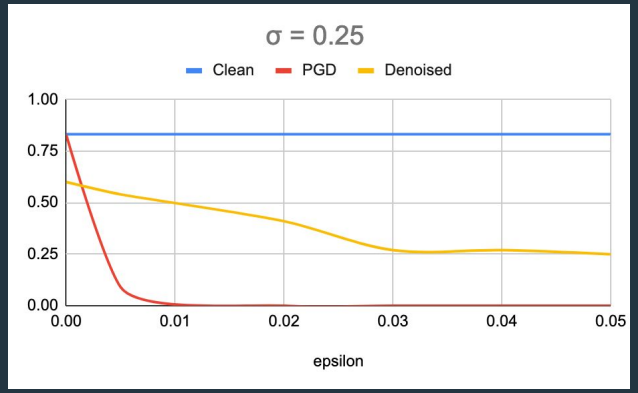


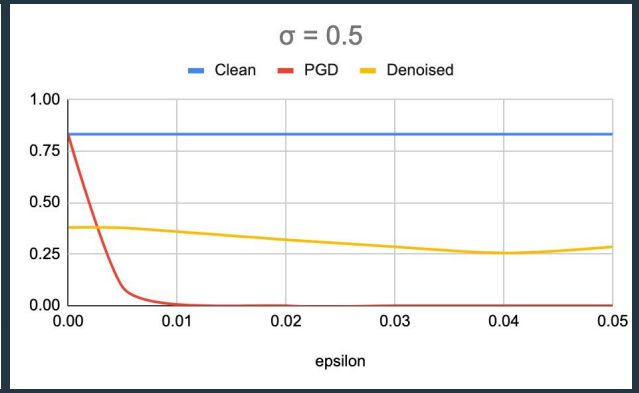


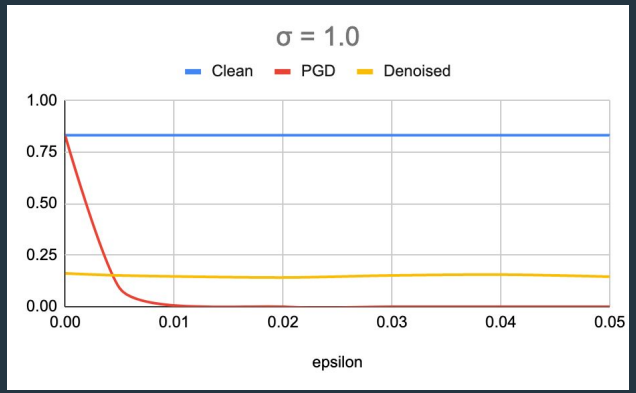


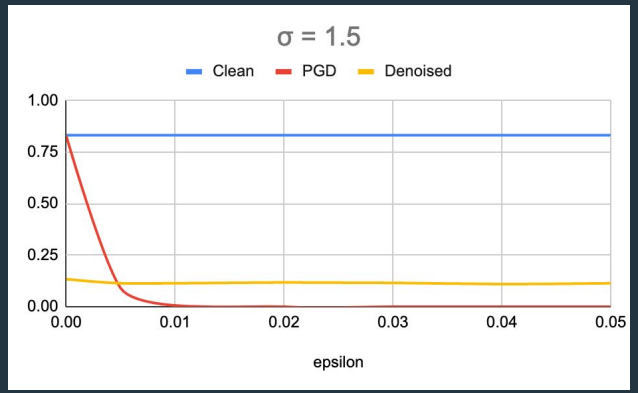
Results - CIFAR-10 with ResNet50+CBAM

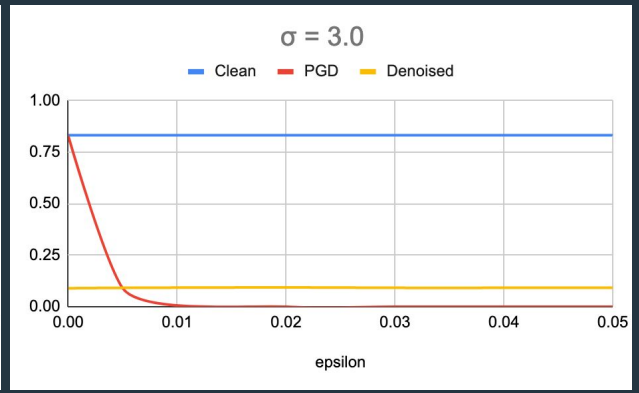












Takeaways

Yes, diffusion denoising a viable defense to adversarial attacks on attention-based image classifiers.

There's lots more to study!

- We were not able to get to how spatial-attention is affected by PGD and denoising.
  - May be able to compare ResNet50 and ResNet50+CBAM
- While the diffuser in our threat model is not considered when we are attacking, we are interested on how adversarial attacks would perform given the classifier and diffuser are white box.
  - Unsure on how to attack said model

Questions / Comments?

Thank you

References

[1] N. Carlini, F. Tramèr, K. Dvijotham, L. Rice, M. Sun, and Z. Kolter, “(Certified!!) Adversarial Robustness for Free!,” Int. Conf. Learn. Represent. ICLR, 2023.

[2] P. Agrawal, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, “Impact of Attention on Adversarial Robustness of Image Classification Models,” CoRR, vol. abs/2109.00936, 2021, [Online]. Available: https://arxiv.org/abs/2109.00936

[3] R. Wightman, “PyTorch Image Models,” GitHub repository. GitHub, 2019. doi: 10.5281/zenodo.4414861.