Safeguarding Attention With Diffusion Denoised Smoothing

E. Nechanicky[1*], M. Morgan[1], G. Lesher[1]

August 20, 2023

Abstract

This paper explores the enhancement of attention-based image classification model robustness by incorporating diffusion denoised smoothing techniques. We build upon the methodology outlined in the seminal ”Certified! Adversarial Robustness for Free!”Carlini et al. (2022) paper, assessing the impact of these smoothing techniques on the robustness and performance of alternative attention models. Moreover, we directly compare the effect of diffusion denoising on the robustness of ResNet-50 with and without convolutional block attention modules. Our small-scale results indicate diffusion denoising is effective for both traditional and attention-based image classifiers.

1 Introduction

1.1 Background

Image classification is a common task in the field of computer vision that aims to categorize images into predefined classes. This task is often accomplished through supervised learning techniques using neural networks, often Convolutional Neural Networks (CNNs), which aim to identify patterns and features in images. On the other hand, attention mechanisms, largely used within the context of natural language processing, help models to identify and focus on the most relevant parts of input data while making predictions. In recent years, these mechanisms are seeing increased usage outside of natural language processing, including in image classification. In the field of adversarial robustness, diffusion denoised smoothing is a method used to enhance a model's defense against adversarial attacks. Adversarial attacks are intentional minor changes to input data that aim to mislead machine learning models. Diffusion denoised smoothing implements both diffusion and randomized smoothing concepts alongside classification models to reduce model sensitivity to these perturbations.

1.2 Research Objective

Acknowledging the increased presence of attention mechanisms within image classification, the primary objective of this research is to investigate the performance of diffusion denoised smoothing when applied to attention-based image classification models. The research seeks to answer the question: ”Does the application of diffusion denoised smoothing enhance the adversarial robustness of attention-based image classifiers?” More specifically, the study aims to discern whether diffusion denoised smoothing has a notable benefit for attention-based image classifiers compared to their conventional counterparts. The goal is to establish whether diffusion denoised smoothing can be reliably used as image classifiers continue to evolve, while maintaining or improving classifier performance.

2 Related Work

2.1 Image Classification and Attention Mechanisms

While there are many attention-based image classification models nowadays, we selected two distinct models for two different datasets which are popular among prior research. For the CIFAR-10 dataset, we chose to train our own ResNet-50 models, an architecture first conceived by He et al. (2015). In years following, the model was adapted with convolutional block attention modules (CBAM) developed by Woo et al. (2018). Recently, research on the adversarial robustness of these two models increased. Agrawal et al. (2021) studied the impact of attention on adversarial robustness by comparing vanilla ResNet-50 with CBAM+ResNet-50. Their results indicated that the robustness could be dependent on the the number of classes in the dataset used, as CBAM+ResNet-50 was more robust for the CIFAR-100 dataset but less robust on CIFAR-10 and Fashion MNIST than vanilla ResNet. In this paper, we similarly compare vanilla ResNet-50 and CBAM+ResNet-50, but focus on the effectiveness of diffusion denoising as a defense between the two models.

For the ImageNet-1k dataset, we opted for a pretrained attention-based model due to the extensive time needed to train on the large dataset. Specifically, we opted for *coatnet_rmlp_2_rw_384.sw_in12k_ft_in1k* Dai et al. (2021) which is pretrained on ImageNet-12k and fine-tuned on ImageNet-1k. Plus, it is publicly available by HuggingFace Wightman (2019). This CoAtNet model combines MBConv (depthwise-separable) convolutional blocks near the front with self-attention tranformer blocks towards the end of the model architecture, and is one of the top three models in terms of top-1 accuracy on the ImageNet-1k datasets as of 2023.

2.2 Diffusion Denoised Smoothing

Diffusion denoised smoothing is recent technique in machine learning that aims to increase the robustness of models against adversarial attacks. Diffusion denoised smoothing addresses vulnerability of generic classification models by creating a ’smoothed’ version of the input, which leaves the model less susceptible to minor alterations in the input data. This technique uses diffusion models in connection with randomized denoising concepts to generate perturbed samples around a given input, after which, a denoising step is performed to create a smoothed input used to infer class labels. The resultant prediction is less sensitive to adversarial noise.

2.3 The ”Certified! Adversarial Robustness for Free!” Approach

In the paper ”Certified! Adversarial Robustness for Free!”Carlini et al. (2022), the authors explored the use of diffusion models as a method for randomized denoising. The methodology they used involved implementing these diffusion models on a specific set of models and datasets to test their efficacy. They discovered that their implementation resulted in both higher clean accuracies and higher robust accuracies compared to all previous smoothing techniques. A significant aspect of their method was that it utilized existing diffusion and classification models, eliminating the need to retrain the image classifier. This is a substantial advancement as it simplifies the process and makes it more accessible for use in practical applications.

3 Methodology

3.1 Attention Visualizations

The CBAM+ResNet-50 model allows us to visualize the final layer of the model’s feature maps as can be seen in Figures 1, 2, 3, and 4. We implement a PGD attack for our adversarial perturbations as per Madry et al. (2019), and variate the ε value to achieve different strengths of attack. In this case, the clean image is correctly predicted as automobile, but after a PGD attack of $\varepsilon = 0.005$, the classifier predicts the perturbed image as ”Truck”. We can see as the ε value increases, the blue heatmap values shift from left to right, while the dark red area on the back-left tire increases in size.

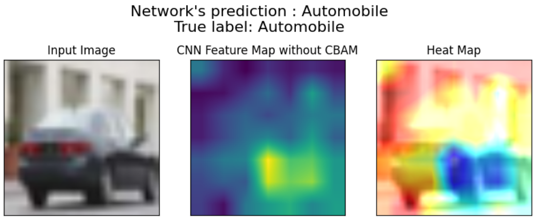
Network's prediction : Automobile
True label: Automobile
Input Image
CNN Feature Map without CBAM
Heat Map

Figure 1: Clean Image

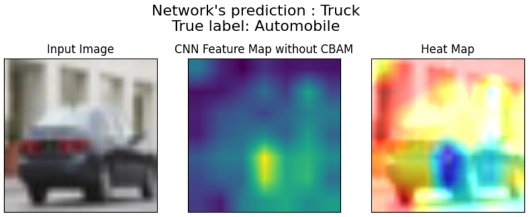

Figure 2: $\varepsilon = 0.005$

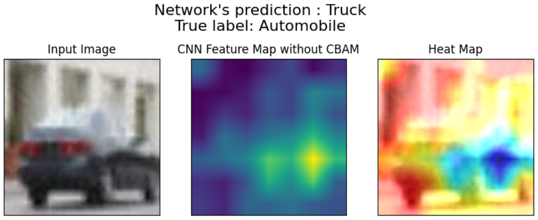

Figure 3: $\varepsilon = 0.05$

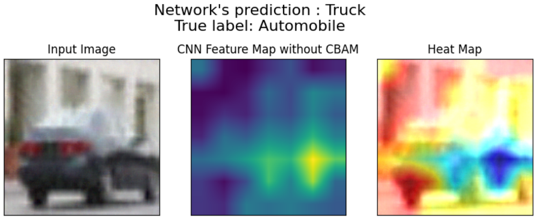

Figure 4: $\varepsilon = 0.1$

3.2 Implementation of Diffusion Denoised Smoothing

For our diffusion denoiser, we use two pretrained off-the-shelf models provided by OpenAI, which are the same models used in Carlini et al. (2022)'s work. For the ImageNet diffusion denoiser (Nichol and Dhariwal (2021a)), we integrate it directly in front of the ImageNet classifier on the forward pass on inference. Likewise, for the CIFAR-10 diffusion denoiser (Nichol and Dhariwal (2021b)), images for inference are resized to 224x224 using "bilinear" method, then run through the diffusion denoiser and then directly into the corresponding ResNet-50 or CBAM+ResNet-50 model.

3.3 Evaluation Criteria

To evaluate the effectiveness of diffusion denoising on an attention-based model we conduct two distinct trials, one control and one experimental. Our control experiments evaluate the attention models both using clean and adversarially generated test-data to identify baseline accuracy. We then compare these accuracies to our second trial's experimental results. The second trial consists of the same process as the control trial, but instead, we pass our model's input through our diffusion denoiser first, then pipe the denoised image into the attention model to be classified. In the paper we based these experiments off, the accuracies reported are their Certified counterparts following Cohen et al. (2019). We do not report certified accuracies in this paper due to time constraints.

4 Results and Discussion

4.1 Performance and Robustness Assessment

We compared the ImageNet model top-1 classification accuracy on a sample of 1000 unseen images from the ImageNet-1k validation set. On the clean dataset, the model scored an accuracy of 0.851. We then performed an adversarial attack using PGD with epsilon values of 0.005, 0.01, 0.03, 0.05, 0.1, and 0.5. PGD's step-size and number of iterations were fixed at $2/255$ and 5 respectively, and a random initialization strategy was not used. The accuracy across these ε values is graphed in Figure 5 as a red curve. Last, we denoised these adversarial images using the diffusion denoiser at the sigma values: 0.25, 0.5, 1.0, and 1.5, plotted as the yellow curve. As expected, the diffusion denoiser improved accuracy on perturbed examples from less than 0.25 up to just under 0.75 for $\sigma = 0.25$. However, the diffusion denoiser did not restore the original clean accuracy of 0.851, still leaving a around a 0.15 accuracy decrease available to an attacker. We found that a σ value of 0.25 was most effective against adversaries across all ε values. For detailed results, including for $\sigma = 1.0$ and $\sigma = 1.5$, see Table 1 in the Appendix.

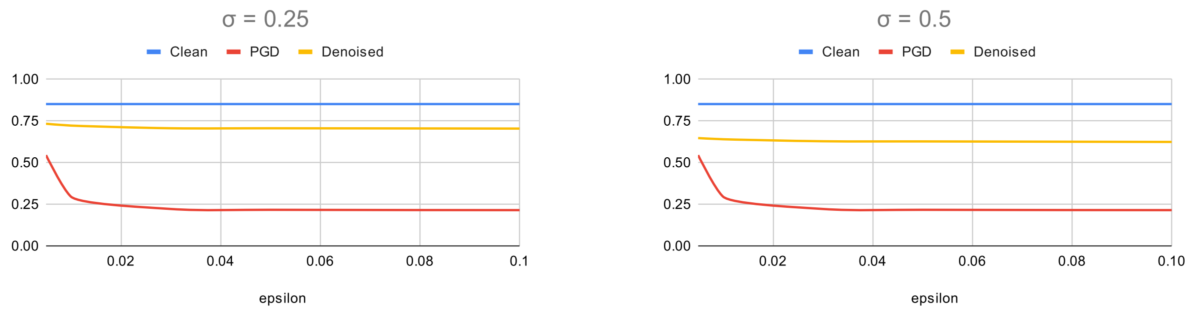

Figure 5: Evaluation of accuracy compared to the adversarial images' ε value on the ImageNet dataset.

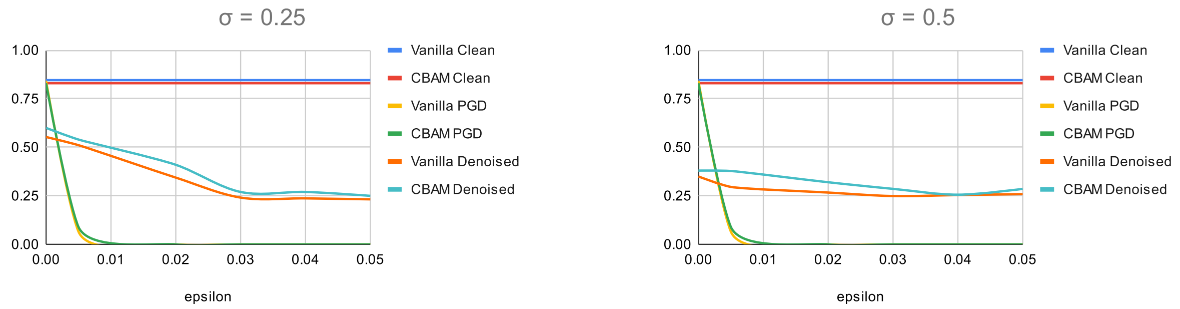

Figure 6: Evaluation of CBAM and non-CBAM (vanilla) model accuracy on the CIFAR dataset compared to ε

While the ImageNet experiment allowed us to evaluate the effectiveness of diffusion denoising, we then performed a comparison of similar attacks on two CIFAR-10 dataset classifiers: Vanilla ResNet-50 and CBAM+ResNet-50. The two classifiers scored near identical accuracy on a test set of 1000 unseen images, with an accuracy of 0.846 for Vanilla ResNet-50 versus 0.83 for CBAM+ResNet-50. After PGD attacks at 0.005, 0.02, 0.04, 0.05, and 0.1, both models' accuracy severely dropped, with not a single correct prediction for every epsilon value above 0.01 (see figure 6). Even though PGD dropped adversarial accuracy down to 0, the diffusion denoiser was able to increase accuracy up above 0.23 across all epsilon values. Just like the ImageNet model, $\sigma = 0.25$ was the optimal parameter for the diffusion denoiser, and for epsilon values less than 0.01, was able to increase accuracy above 0.5. Note that our PGD hyperparameters were a niter of 5 and stepsize of 2/255.

Most importantly, the experiment allowed us to answer the question: how does diffusion denoising compare as a defense for attention-based image classifiers versus traditional CNNs? Our results indicate CBAM+ResNet-50 was slightly more resilient to adversarial attacks than Vanilla ResNet-50, with only about a 0.05 difference at most after denoising with $\sigma = 0.25$. At higher σ values, the vanilla model became narrowly more resilient than the attention-based model. Overall, the two models did not deviate much from each other. The full results are available in Table 2 and 3.

4.2 Comparison with Existing Literature

In comparison to the work that inspired our paper, Carlini et al. (2022), our ImageNet experiments yielded comparable results. Carlini et al. (2022) achieved 82.8% clean accuracy and 71.1% certified accuracy at $\varepsilon = 0.5$. Our results are not certified, as they consist of an accuracy measurement on a small 1,000 sample set. Yet, we similarly scored an 85.1% clean accuracy and 70.2% with PGD at an $\varepsilon = 0.5$ and denoising at $\sigma = 0.25$. On the other hand, Carlini et al. (2022) more thoroughly explored diffusion denoising and found better results than we achieved at higher σ values, in some cases differing by 50% accuracy.

On CIFAR-10, Carlini et al. (2022) use a pretrained Wide-ResNet model which scored 95.2% clean accuracy, meanwhile we train our own ResNet-50 and CBAM+ResNet-50 models to 84.6% and 83.0% respectively. We hope that our comparison between the same mode with and without attention preview that diffusion denoising is worthy of use for traditional and attention-based models, with Carlini et al. (2022)'s work as a backbone to ours.

5 Conclusion and Future Work

5.1 Summary of Findings

In our experiments we found that, in general, while a diffusion denoiser recovers a significant amount of the accuracy lost from adversarially images, using an attention based model does not provide much benefit or cost over other model types. Our experiments show a small increase in recovered accuracy (at most 5%) when comparing our attention model to a vanilla model.

It is worth noting that in our experiments with σ greater than 1.0, the vanilla denoised accuracies are higher than the CBAM accuracies (see Appendix B and C). The difference between the vanilla and CBAM accuracies at these σ values gradually match the differences observed in the clean samples fed to the model (without the denoiser) as the σ increases. This, along with the overall accuracy degrading towards the base 10% accuracy, suggests that σ values within this range destroy the image within the denoising step, and as such are not viable for defense.

5.2 Implications and Future Directions

The results we see imply that the application of a diffusion denoiser on an attention based model has little to no benefit over a non-attention CNN. Our results also imply that even for high ε values, a denoiser tuned to small values of noise (small σ parameter) will preform better at defending against adversarial attacks. This intuition makes sense because adversarial attacks generally preform minute perturbations which can be abstracted to ”noise”.

Our experiments are by no means comprehensive, and we have identified multiple avenues for future study. One potential avenue is attacking the denoiser and model as a whole, rather than treating the denoiser as an unknown addition. It is reasonable that an attacker would have knowledge of a denoising defense and so crafting an attack with this knowledge could be valuable to explore. The models used in our experiments were pre-trained and as such had no ”knowledge” of the denoiser pre-filter. Training the attention models on the diffused images would likely increase classification accuracy to rates similar to non-adversarial accuracies, however proper research must be done to verify this claim.

Another potential avenue to explore is how spatial-attention feature maps are affected by the diffusion denoising. Zoran et al. (2020) found that sequential attention models are more vulnerable to global, salient, and spatially coherent adversarial examples. We generated sample heatmaps (Figure 1) of various PGD attack strengths, but would be keen to explore and quantify how spatially an attention-based model's robustness may differ from a traditional CNN when combined with diffusion denoising.

References

Agrawal P, Punn NS, Sonbhadra SK, and Agarwal S (2021) Impact of attention on adversarial robustness of image classification models

Carlini N, Tramer F, Dvijotham KD, Rice L, Sun M, and Zico Kolter J (2022) (Certified!!) Adversarial Robustness for Free!. *arXiv e-prints* page arXiv:2206.10550. provided by the SAO/NASA Astrophysics Data System

Cohen JM, Rosenfeld E, and Kolter JZ (2019) Certified Adversarial Robustness via Randomized Smoothing. *arXiv e-prints* page arXiv:1902.02918

Dai Z, Liu H, Le QV, and Tan M (2021) Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:210604803*

He K, Zhang X, Ren S, and Sun J (2015) Deep residual learning for image recognition

Madry A, Makelov A, Schmidt L, Tsipras D, and Vladu A (2019) Towards deep learning models resistant to adversarial attacks

Nichol A and Dhariwal P (2021a) Guided diffusion. `https://github.com/openai/guided-diffusion`

Nichol A and Dhariwal P (2021b) Improved diffusion. `https://github.com/openai/improved-diffusion`

Wightman R (2019) Pytorch image models. `https://github.com/huggingface/pytorch-image-models`

Woo S, Park J, Lee JY, and Kweon IS (2018) Cbam: Convolutional block attention module

Zoran D, Chrzanowski M, Huang PS, Gowal S, Mott A, and Kohli P (2020) Towards robust image classification using sequential attention models. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

Appendices

A CBAM vs. Vanilla results

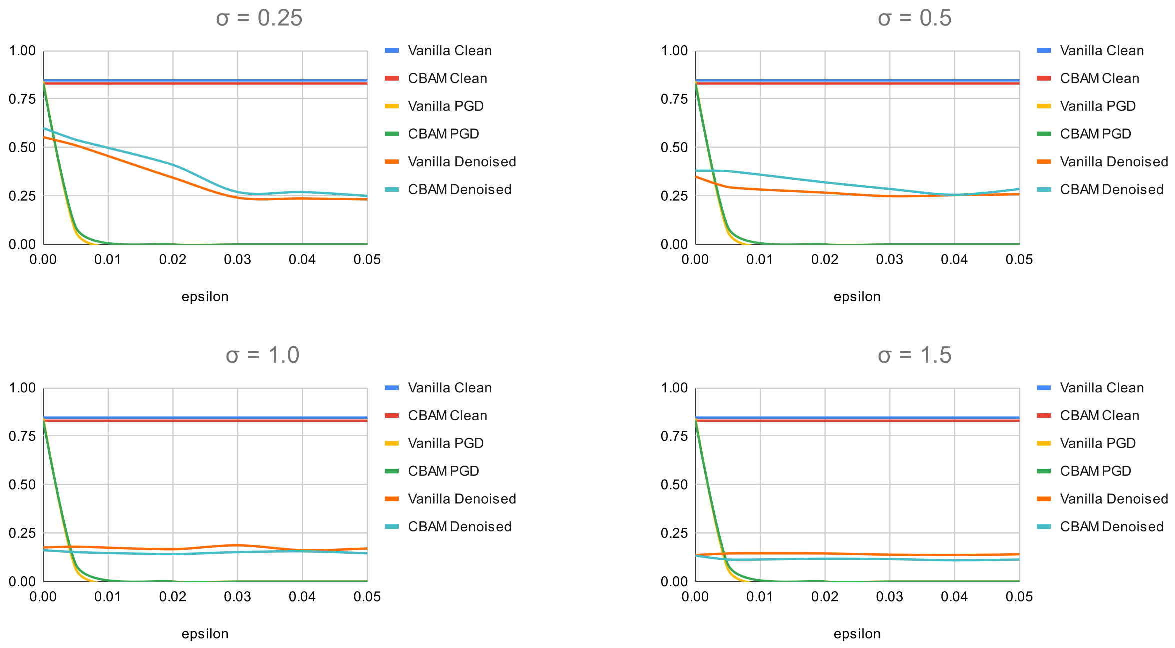

Figure 7: Evaluation of CBAM and non-CBAM (vanilla) model accuracy on the CIFAR dataset across ε values

B Clean Image vs. Adversarial Image vs. Denoised Image

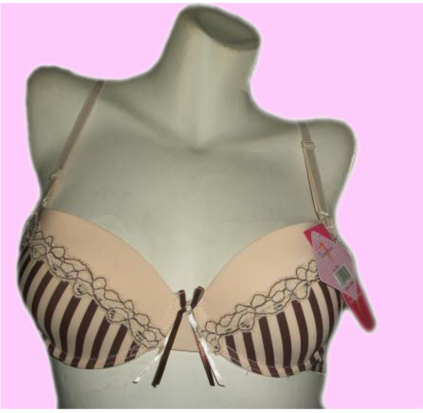

Figure 8: Clean Image

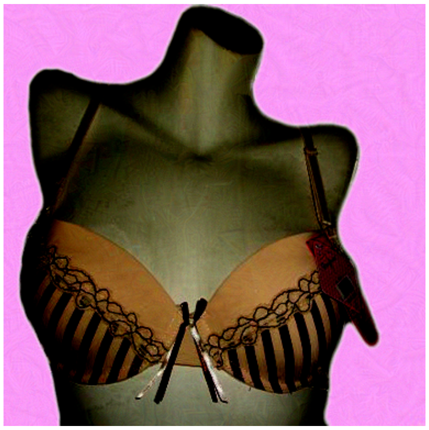

Figure 9: Adversarial Attack with $\varepsilon = 0.05$

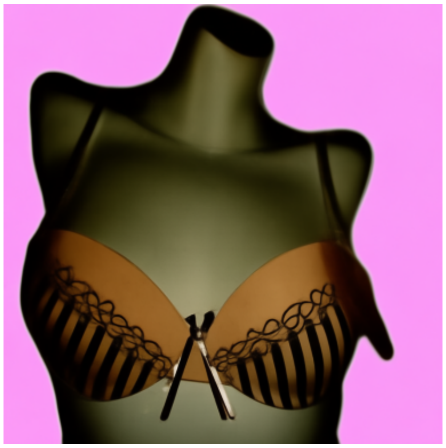

Figure 10: Denoised with $\sigma = 0.5$

C Expiremental Result Tables

Table 1: Clean, Adversarial, and Denoised Results on ImageNet

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

Table 2: Vanilla ResNet-50 Clean, Adversarial, and Denoised Results on CIFAR-10

Sigma	**Epsilon**	**Clean Acc.**	**PGD Acc.**	**Denoised Acc.**
0.25	0	0.846	0.843	0.553
0.25	0.005	0.846	0.066	0.511
0.25	0.02	0.846	0	0.344
0.25	0.04	0.846	0	0.241
0.25	0.05	0.846	0	0.237
0.25	0.1	0.846	0	0.232
0.5	0	0.846	0.843	0.35
0.5	0.005	0.846	0.066	0.296
0.5	0.02	0.846	0	0.267
0.5	0.04	0.846	0	0.249
0.5	0.05	0.846	0	0.254
0.5	0.1	0.846	0	0.259
1	0	0.846	0.843	0.176
1	0.005	0.846	0.066	0.18
1	0.02	0.846	0	0.167
1	0.04	0.846	0	0.187
1	0.05	0.846	0	0.162
1	0.1	0.846	0	0.171
1.5	0	0.846	0.843	0.137
1.5	0.005	0.846	0.066	0.145
1.5	0.02	0.846	0	0.145
1.5	0.04	0.846	0	0.139
1.5	0.05	0.846	0	0.137
1.5	0.1	0.846	0	0.141
3	0	0.846	0.843	0.107
3	0.005	0.846	0.066	0.113
3	0.02	0.846	0	0.108
3	0.04	0.846	0	0.108
3	0.05	0.846	0	0.111
3	0.1	0.846	0	0.107

Table 3: CBAM+ResNet-50 Clean, Adversarial, and Denoised Results on CIFAR-10

Sigma	**Epsilon**	**Clean Acc.**	**PGD Acc.**	**Denoised Acc.**
0.25	0	0.83	0.83	0.6
0.25	0.005	0.83	0.09	0.54
0.25	0.02	0.83	0	0.41
0.25	0.04	0.83	0	0.27
0.25	0.05	0.83	0	0.25
0.25	0.1	0.83	0	0.24
0.5	0	0.83	0.83	0.38
0.5	0.005	0.83	0.09	0.378
0.5	0.02	0.83	0	0.32
0.5	0.04	0.83	0	0.256
0.5	0.05	0.83	0	0.286
0.5	0.1	0.83	0	0.268
1	0	0.83	0.83	0.162
1	0.005	0.83	0.09	0.152
1	0.02	0.83	0	0.142
1	0.04	0.83	0	0.156
1	0.05	0.83	0	0.146
1	0.1	0.83	0	0.146
1.5	0	0.83	0.83	0.134
1.5	0.005	0.83	0.09	0.114
1.5	0.02	0.83	0	0.118
1.5	0.04	0.83	0	0.11
1.5	0.05	0.83	0	0.114
1.5	0.1	0.83	0	0.106
3	0	0.83	0.83	0.09
3	0.005	0.83	0.09	0.092
3	0.02	0.83	0	0.094
3	0.04	0.83	0	0.092
3	0.05	0.83	0	0.092
3	0.1	0.83	0	0.088