Can Attention Masks Improve Adversarial Robustness?

Pratik Vaishnavi[1], **Tianji Cong**[2], **Kevin Eykholt**[2], **Atul Prakash**[2], **Amir Rahmati**[1]
[1]Stony Brook University [2]University of Michigan
pvaishnavi@cs.stonybrook.edu,{congtj,keykholt,aprakash}@umich.edu, amir@rahmati.com

Abstract

Deep Neural Networks (DNNs) are known to be susceptible to adversarial examples. Adversarial examples are maliciously crafted inputs that are designed to fool a model, but appear normal to human beings. Recent work has shown that pixel discretization can be used to make classifiers for MNIST highly robust to adversarial examples. However, pixel discretization fails to provide significant protection on more complex datasets. In this paper, we take the first step towards reconciling these contrary findings. Focusing on the observation that discrete pixelization in MNIST makes the background completely black and foreground completely white, we hypothesize that the important property for increasing robustness is the elimination of image background using attention masks before classifying an object. To examine this hypothesis, we create foreground attention masks for two different datasets, GTSRB and MS-COCO. Our initial results suggest that using attention mask leads to improved robustness. On the adversarially trained classifiers, we see an adversarial robustness increase of over 20% on MS-COCO.

Introduction

Deep Neural Networks are employed in a wide range of applications ranging from autonomous systems to trading and healthcare. This has resulted in an increased attention to their security. One of the primary focuses of these efforts has been defense against adversarial examples. Adversarial examples (Szegedy et al. 2014) can be generated by adding carefully-crafted imperceptible noise to a normal input example. Such an adversarial example can be used to trigger a misclassification on a target model for image classification tasks (*e.g.,* a road-sign classifier in a self-driving car). Many techniques have been developed to tackle this problem (Xie et al. 2019; Madry et al. 2018; Lamb et al. 2018), one of the popular one being adversarial training.

In analyzing an adversarially trained DNN on MNIST, Madry *et al.* (Madry et al. 2018) found that the first layer filters turned out to be thresholding-filters that were acting as a de-noiser for the grayscale MNIST images. Follow up experiments by Schott *et al.* (Schott et al. 2018) showed that training a DNN with binarized MNIST images (where each pixel was discretized to be either made completely black or completely white using a static threshold) resulted in significantly improved adversarial accuracy without any adversarial training or negative impact on normal performace. In other words, a pipeline that first thresholds each pixel in an MNIST image to 1 or 0 and then classifies the resulting image with a naturally trained model has a very high degree of adversarial robustness, without requiring any adversarial training. Subsequent works, however, found that a simple binarization was not effective for more complex datasets such as CIFAR-10 (Chen et al. 2018).

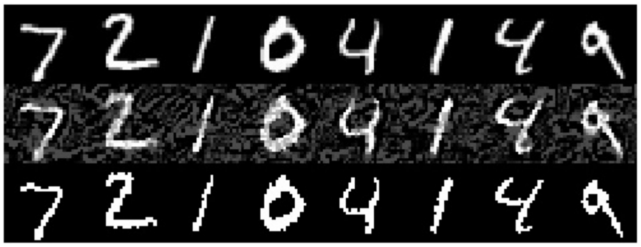

Figure 1: Visualizing images from the MNIST dataset. The first row contains natural images, the second row contains corresponding adversarial images and the third row contains binarized adversarial images (threshold = 0.5). Binarization removes almost all the adversarial noise.

In contrast to previous work, we observe that binarization on MNIST acts as an approximation for the process of *foreground-background* separation. Figure 1 presents a sample image from the MNIST dataset. MNIST dataset consist of 28×28 pixel images of handwritten white digits on a black background and are among the simplest datasets in ML. Adversarial attacks on the MNIST typically do not distinguish between the digits and the background and may similarly manipulate both to achieve their goal. When examining the adversarial noise in MNIST adversarial examples for a naturally-trained model, we see that the adversarial noise is spread throughout the image, including in background regions of the image that we might not consider to carry any predictive signals. Binarization on MNIST as a pre-processing step snaps most of the background pixels back to the original value on which the model was trained (*i.e.,* 0), hence reducing the attack surface.

Therefore, it may be more accurate to characterize binarization on MNIST as foreground-background separation rather than simple pixel discretization for more complex image datasets. If the hypothesis is true, then we should see improved robustness on other datasets by simply separating the background from the foreground and masking the background prior to training and classification, i.e., applying a foreground-attention mask on the dataset.

Towards validating the above hypothesis, given a classifier and a dataset, we introduce an additional pre-processing step where a foreground attention mask is applied to the model's input before classification. A challenge in testing our hypothesis is determining the foreground attention mask. Unfortunately, most image datasets on which adversarial testing is done (e.g., CIFAR-10, ImageNet) lack sufficient ground truth data for foreground attention masks. To address the challenge, we generate two datasets with foreground attention masks from existing datasets: The German Traffic Sign Recognition (GTSRB) (Stallkamp et al. 2012) and MS-COCO (Lin et al. 2014). For the GTSRB dataset, we took advantage of the typical color distribution in images and that a road sign often lies in the center of the image, and used them to design a custom attention mask generator by doing random sampling of pixels near the center of the image along with a min cut-max flow algorithm to create the foreground attention mask. For the MS-COCO dataset, we use the segmentation masks included with the dataset to create a cropped image of the object of interest and its foreground mask.

Our preliminary results suggest that a classification pipeline that utilizes foreground attention masks experiences improved adversarial robustness with, at worst, no impact on natural accuracy. On the naturally trained classifiers, the adversarial accuracy improves by 0.73% on MS-COCO and around 19.69% on GTSRB. Robustness improvements were also found on combining usage of attention masks with adversarial training. Thus, this paper makes the following contributions:

- We create two datasets based on the GTSRB and MS-COCO that allow exploration of attention mask effects on adversarial examples. These datasets are available at [link].
- We take the first step toward exploring the effect of attention masks on improving model robustness in image classification tasks. We show that attention masks have certain effect on improving adversarial performance against PGD adversary.

Background

Discrete Pixelization Defenses. Developing adversarial defenses towards robust classification has received significant attention in recent years (Madry et al. 2018; Liao et al. 2018). Among these, defense methods that pre-process inputs to improve robustness are potentially attractive because the pre-processed input can be passed to existing classifiers for improved robustness. Unfortunately, some of these methods were vulnerable to stronger adaptive adversarial attacks (Athalye, Carlini, and Wagner 2018), raising doubts on the effectiveness of pre-processing strategies.

One pre-processing strategy that has stood attacks well against stronger adaptive adversarial attacks is that of binarization for the MNIST dataset. Unfortunately, binarization, which converts each pixel value to fewer bits did not provide a significant benefit on more complex datasets and, in some cases, negatively impacted test accuracy (Chen et al. 2018). Chen *et al.* provide theoretical insights into the phenomenon, concluding that discrete pixelization is unlikely to provide significant benefit.

Semantic Segmentation. Semantic segmentation of images has applications in diverse fields like biomedical imaging (Ronneberger, Fischer, and Brox 2015). (Wu et al. 2019; Chen et al. 2017) describe semantic segmentation techniques for complex real-word datasets like MSCOCO and Cityscapes. However, (Arnab, Miksik, and Torr 2018) have shown that DNN-based segmentation pipelines can be susceptible to adversarial attacks, though other work has shown that such attacks may be successfully detected (Xiao et al. 2018), potentially providing a defense. We note that, unlike our work, prior work on robustness of semantic segmentation looked at robustness of segmentation model itself and not the impact of foreground-background separation on robustness of classification of a foreground object.

Attention masks. According to work by Xie *et al.* on feature denoising, they discovered that adversarial noise causes machine learning models to focus on semantically uninformative information in the input, whereas the opposite is true for natural clean inputs. Thus, rather than relying on the model to identify relevant input features, we explore if we can force the network to focus on important portions of the image (*e.g.,* the foreground object). Harley *et al.* (Harley, Derpanis, and Kokkinos 2017) proposed the idea of segmentation-aware convolution networks that rely on local attention masks, which are generated based on color-distance among neighboring pixels, and found that it can improve semantic segmentation and optical flow estimation. Our work aims to understand if attention masks can also be useful for improvement of robustness of image classification.

Our Approach

In this work, our goal is to examine if isolating predictive signals in the form of foreground features has benefits in terms of adversarial robustness while having minimal impact on model performance on natural inputs. We examine, using two datasets, that training a model on foreground pixels helps it perform well not only on natural images, but makes it robust against adversarial images as well.

Let $\mathbf{X}$ be a set of images drawn from a distribution. Let's consider the task of image classification defined on $\mathbf{X}$. Traditionally, in image classification, we restrict each image to contain only one object. An image $x^{(i)} \in \mathbf{X}$ can be divided into foreground and background pixels. By definition, foreground pixels are pixels that are a part of the object and every other pixel can be considered as a part of the background. In this paper we make the assumption that the foreground pixels, on their own, carry sufficient predictive power for the task of image classification. Additionally, removing background pixels restricts the input space that the adversary can attack, inhibiting its ability to trigger misclassification in the target model.

For an image $x^{(i)} \in \mathbf{X}$ of resolution $m \times n$, let's define a foreground image $x_{FG}^{(i)} = F(x^{(i)})$, where

$$F(x^{(i)}) = \begin{cases} x_{jk}^{(i)} & \text{if } x_{jk}^{(i)} \in S_{FG}^{(i)} \\ 0 & \text{else} \end{cases} \quad j=1...m, k=1...n$$

where, $S_{FG}^{(i)}$ is the set of foreground pixels for image $x^{(i)}$. We generate $\mathbf{X}_{FG}$ containing foreground images $x_{FG}^{(i)}$ $\forall x^{(i)} \in \mathbf{X}$.

Based on the above, we evaluate two class of models: (1) model trained on $\mathbf{X}$, (2) model trained on $\mathbf{X}_{FG}$. For both, we perform Natural (N) and Adversarial (A) training.

We assume access to foreground masks in our experiments. Thus, our work provides an upper bound on the potential benefit that can be provided by foreground attention masks on model robustness, assuming foreground attention masks can be robustly found. There is some potential hope since recent work has shown that adversarial attacks may be successfully detected on segmentation algorithms (Xiao et al. 2018) using statistical tests, potentially providing a basis for a defense.

Dataset Creation with Foreground Masks

MS-COCO. We pre-process the MS-COCO dataset to make it compatible with the task of image-classification. Particularly, we use the following pre-processing steps:

- We make use of the semantic segmentation masks and object bounding box annotations to generate image, mask, label pairs such that each image contains object(s) of one label only, or in other words, the mask corresponding to an image contains annotations for one object only.
- To deal with objects having overlapping bounding box regions, we explicitly black-out pixels corresponding to the extra objects.
- We adjust the crop dimensions in order to extract square image patches, and resize all the extracted image patches to 32×32.
- Due to high class-imbalance in the resultant dataset, we ignore the person class (over frequent) and short-list top 10 classes from the remaining classes based on the frequency.

We call this modified MS-COCO dataset as **MS-COCO-IC**. Table 1 shows the statistics for this dataset. The images in this dataset contain $\approx 56\%$ foreground pixels. Figure 3 gives an example of an original image, the cropped image, and the final image that is used for training a classifier.

Class	**Number of Images**	
	Train	**Test**
Chair	21674	11077
Car	18498	9594
Book	12094	6188
Bottle	10923	5735
Dinning table	10291	5274
Umbrella	6510	3309
Boat	5531	2797
Motorcycle	5340	2703
Sheep	4748	2432
Cow	4391	2162
Total	100000	51271

Table 1: Number of images per class in the train and test set of our MSCOCO-IC dataset.

GTSRB. The German Traffic Sign Recognition Benchmark (GTSRB) is a dataset containing 50,000 color images of 43 different classes of road signs, with high class imbalance. Images in GTSRB model various viewpoints and light conditions. We use a customized segmenter based on the graph cut algorithm to obtain foreground masks. One favorable aspect of GTSRB is that majority of the traffic signs are centrally located in the image, have regular shapes, and usually possess a sharp color in contrast to the background. These features match the assumptions of our customized segmenter. The images in this dataset contain $\approx 25\%$ foreground pixels. We associate the low percentage of foreground pixels to the imperfections of our ad hoc segmenter. Figure 2 gives an example of an image, computed mask by our segmenter, and the final image that is used for training a classifier.

Table 2: Visualizing examples in the GTSRB-IC dataset. We display the natural image, foreground mask and the foreground masked image from left to right.

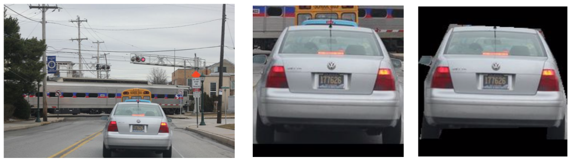

Table 3: Visualizing examples in the MS-COCO-IC dataset. We display the original image from the MS-COCO dataset, cropped image of object of interest, and the foreground masked image from left to right.

Results

For our experiments, we train two set of models: (1) on natural images; (2) on foreground masked image, both naturally and adversarially. For both, we use the VGG-19 classifier. We evaluate these models against a 10-step $\mathbf{L}_\infty$ bounded PGD adversary with step size of $\frac{2}{255}$ and $\epsilon = \frac{8}{255}$. Treating the performance of models trained on $\mathbf{X}$ as a baseline, we calculate potential gains in robustness in the models trained on $\mathbf{X}_{FG}$. We repeat the above experiments for both GTSRB-IC and MS-COCO-IC datasets. Note that in the case of $\mathbf{X}_{FG}$ models, the adversary is only given access to the foreground pixels. We summarize our results in Table 4.

MS-COCO-IC. We can observe from the results that both the naturally trained classifiers exhibit comparable vulnerability to PGD adversary. However, in case of the adversarially trained classifiers, natural and adversarial accuracy increases by 11.41% and 22.82% respectively, on using foreground attention masks. This validates our hypothesis that foreground attention has a positive effect on a model's classification performance and robustness.

GTSRB-IC. Similar to the previous set of results, we see that the model adversarially trained using $\mathbf{X}_{FG}$ is more robust to a PGD adversary than a model adversarially trained using $\mathbf{X}$

Data	Training	MSCOCO-IC		GTSRB-IC	
		Natural	PGD	Natural	PGD
$\mathbf{X}$	N	79.46%	2.28%	98.04%	18.69%
$\mathbf{X}_{FG}$	N	**81.64%**	**3.01%**	98.04%	**38.38%**
$\mathbf{X}$	A	61.51%	30.80%	89.54%	55.25%
$\mathbf{X}_{FG}$	A	**72.92%**	**53.62%**	**91.20%**	**64.57%**

Table 4: Comparing adversarial robustness of models trained on natural images versus foreground masked images. For both datasets. we observe increased robustness against a PGD adversary when the model and the adversary have access to foreground features only.

(+9.32%). Additionally, we see that the model naturally trained on $\mathbf{X}_{FG}$ exhibits considerable improvement in robustness as compared to the model naturally trained on $\mathbf{X}$ (+19.69%).

Conclusion

We study the use of foreground attention masks for improving the robustness of deep neural networks against L_∞-bounded PGD attack. We develop two new datasets based on MS-COCO and GTSRB, to examine these effects. Our preliminary results suggest positive effects in using foreground masks for improving adversarial robustness against PGD adversary. For an adversarially trained model on the MS-COCO-IC dataset, foreground attention masks improved adversarial accuracy by 21.8%. For the GTSRB-IC dataset, when adversarially training a model with foreground masked images, we observe a smaller improvement of 1.7% in adversarial accuracy. Initial results are promising however, further work must be done to verify the effect of foreground attention masks on adversarial robustness and to develop an reliable method to extract these foreground masks automatically.

Prior work by Simon-Gabriel *et al.* (Simon-Gabriel et al. 2018) suggest that our masking technique improves adversarial robustness due to a reduction in the number of input features. Against a first-order adversary (*e.g.,* PGD attack), they establish that the adversarial robustness scales as $1/\sqrt{d}$ where d is the number of input features. Therefore, masking some of the input features should improve adversarial robustness to some degree. For future work, we intend to investigate this relationship further as we believe that the relative importance of a feature for classification may suggest additional robustness. Also, this relationship might help better explain certain trends that we observe in the results. Such work will help better understand the usefulness of foreground masks in context of adversarial robustness against first-order adversaries.

References

[Arnab, Miksik, and Torr 2018] Arnab, A.; Miksik, O.; and Torr, P. H. 2018. On the robustness of semantic segmentation models to adversarial attacks. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

[Athalye, Carlini, and Wagner 2018] Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Intl. Conf. on Machine Learning*.

[Chen et al. 2017] Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

[Chen et al. 2018] Chen, J.; Wu, X.; Liang, Y.; and Jha, S. 2018. Improving adversarial robustness by data-specific discretization. *arXiv preprint arXiv:1805.07816*.

[Harley, Derpanis, and Kokkinos 2017] Harley, A. W.; Derpanis, K. G.; and Kokkinos, I. 2017. Segmentation-aware convolutional networks using local attention masks. In *Intl. Conf. on Computer Vision*.

[Lamb et al. 2018] Lamb, A.; Binas, J.; Goyal, A.; Serdyuk, D.; Subramanian, S.; Mitliagkas, I.; and Bengio, Y. 2018. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. *arXiv preprint arXiv:1804.02485*.

[Liao et al. 2018] Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Zhu, J.; and Hu, X. 2018. Defense against adversarial attacks using high-level representation guided denoiser. *IEEE Conf. on Computer Vision and Pattern Recognition*.

[Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

[Madry et al. 2018] Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *Intl. Conf. on Learning Representation*.

[Ronneberger, Fischer, and Brox 2015] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Intl. Conf. on Medical image computing and computer-assisted intervention*. Springer.

[Schott et al. 2018] Schott, L.; Rauber, J.; Bethge, M.; and Brendel, W. 2018. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*.

[Simon-Gabriel et al. 2018] Simon-Gabriel, C.-J.; Ollivier, Y.; Bottou, L.; Schölkopf, B.; and Lopez-Paz, D. 2018. Adversarial vulnerability of neural networks increases with input dimension. *arXiv preprint arXiv:1802.01421*.

[Stallkamp et al. 2012] Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*.

[Szegedy et al. 2014] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Intl. Conf. on Learning Representations*.

[Wu et al. 2019] Wu, H.; Zhang, J.; Huang, K.; Liang, K.; and Yu, Y. 2019. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*.

[Xiao et al. 2018] Xiao, C.; Deng, R.; Li, B.; Yu, F.; Liu, M.; and Song, D. 2018. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *European Conference on Computer Vision*.

[Xie et al. 2019] Xie, C.; Wu, Y.; Maaten, L. v. d.; Yuille, A. L.; and He, K. 2019. Feature denoising for improving adversarial robustness. In *IEEE Conf. on Computer Vision and Pattern Recognition*.