FEATURE PRIORITIZATION AND REGULARIZATION IMPROVE STANDARD ACCURACY AND ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training has been successfully applied to build robust models at a certain cost. While the robustness of a model increases, the standard classification accuracy declines. This phenomenon is suggested to be an inherent trade-off between standard accuracy and robustness. We propose a model that employs feature prioritization by a nonlinear attention module and L_2 regularization as implicit denoising to improve the adversarial robustness and the standard accuracy relative to adversarial training. Focusing sharply on the regions of interest, the attention maps encourage the model to rely heavily on features extracted from the most relevant areas while suppressing the unrelated background. Penalized by a regularizer, the model extracts similar features for the natural and adversarial images, effectively ignoring the added perturbation. In addition to qualitative evaluation, we also propose a novel experimental strategy that quantitatively demonstrates that our model is almost ideally aligned with salient data characteristics. Additional experimental results illustrate the power of our model relative to the state of the art methods.

1 INTRODUCTION

Deep learning models have demonstrated impressive performance in a wide variety of applications (Goodfellow et al., 2016; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; Mnih et al., 2015). However, recent works have shown that these models are susceptible to adversarial attacks: imperceptible but carefully chosen perturbation added to the input can cause the model to make highly confident but incorrect predictions (Szegedy et al., 2013; Goodfellow et al., 2015; Kurakin et al., 2016).

Exploring the adversarial robustness of neural networks has recently gained significant attention and there is a rapidly growing body of work related to this topic (Kurakin et al., 2016; Fawzi et al., 2018; Athalye & Sutskever, 2017; Carlini & Wagner, 2017; Kolter & Wong, 2017; Wong & Kolter, 2018; Madry et al., 2017). There are a wide variety of methods proposed to defend a model against adversarial attacks. Preprocessing the images with denoising methods constitutes a possible natural approach in an effort to remove the added adversarial noise from the input before making a final classification. There are several works that use various denoising methods as a preprocessing step, which we call explicit denoising. Explicit denoising can be effective against some weak attacks, but most such techniques are subsequently shown to be ineffective against particularly designed attack methods such that they could be easily countered by white box attacks with optimization based adversaries (Athalye et al., 2018; Athalye & Carlini, 2018). Another type of methods are based on adversarial training, which uses adversarial samples in addition to clean images during the training process. This approach has been shown to be able to build relatively robust neural networks (Madry et al., 2017; Athalye et al., 2018; Dvijotham et al., 2018). With strong adversaries such as the Projected Gradient Descent (PGD) (Madry et al., 2017) or the Iterative Fast Gradient Sign Method (I-FGSM) (Kurakin et al., 2016) adversarially trained models are able to achieve state-of-the-art performance against a wide range of attacks.

Recent advances in the understanding of adversarial robustness of a model show that there is an inherent trade-off between classification accuracy and adversarial robustness (Tsipras et al., 2018;

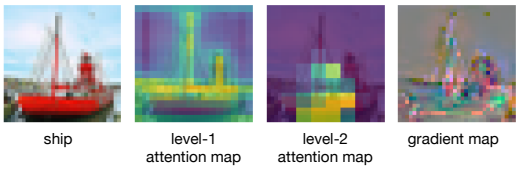

Figure 1: An example of the interpretability of the proposed model. The first image is an image in the CIFAR-10 dataset, the second and third images are the learned attention maps at two different layers of our model, and the last image is the gradient of the cross-entropy loss with respect to the input image.

Tanay et al., 2018). It has also been demonstrated that there are robust and non-robust features of an image depending on their correlation with the correct classification and their vulnerability to adversarial perturbation. Both robust and non-robust features contribute to standard classification accuracy, but adversarial attacks exploit those non-robust features to break the model and cause it to make erroneous predictions. Specifically, if a model relies on both robust and non-robust features when making a prediction, it is able to achieve a higher standard classification accuracy but is more subjective to attacks, thus less adversarially robust; while a robust model relies solely on robust features with a cost of standard classification accuracy. Therefore, a challenging problem is to design a model that inherently favors robust features in order to achieve better robustness and accuracy. However, it remains unclear of what the characteristics of robust features are and how to build such a model.

In this paper, we propose an approach that enhances adversarial training with a feature prioritizing mechanism and an implicit denoising mechanism to improve the robustness of a model. Specifically, for feature prioritizing, we propose an attention module that uses the learned global features of the whole image as a way to assign weights to the lower level learned local features by a non-linear compatibility function followed by using the weighted combination of the local features to predict class labels. For implicit denoising, we add a L_2 regularization term to the training objective that penalizes the distance between the attention features of a clean sample x and that of its perturbed adversarial counterpart x'. By optimizing this regularizer, we are pushing the model to map the two examples to nearby points in the high dimensional manifold learned by the model and hence extract very similar features, thus ignoring the added perturbation. We call our method implicit because there is no denoising preprocessing model involved and cannot be used by the attacker to break it. In the proposed model, the attention module focuses on the area of an image that contains the actual object and helps the classifier to only rely on features extracted from those areas. We employ the attention module at multiple layers which allows focuses at different but complementary areas. The background clutter and irrelevant features which could be misleading are suppressed. The implicit denoising further encourages the model to extract robust features that are not perturbed by the adversarial attack. The resulting model has a highly interpretable gradient map that aligns perfectly with salient data characteristics (Figure 1).

The main contributions of this paper include:

- A method based on feature prioritization and regularization, which significantly outperforms adversarial training. Our model is evaluated on the MNIST and CIFAR-10 datasets, and demonstrates superior performance relative to both standard classification accuracy and adversarial robustness.

- Through qualitative evaluation, the attention maps generated by our non-linear attention estimator focus sharply on the regions of interest while suppressing irrelevant background clutter and misleading perturbations.

- In addition to qualitative evaluation of the gradient maps, we propose a novel experimental strategy that quantitatively demonstrates the alignment of the gradient maps generated by our model with salient data characteristics.

2 RELATED WORK

Due to the extensive amount of literature in this area and the limited length of this paper, we only review some of the most related works in this section. For a more comprehensive survey please refer to Akhtar & Mian (2018).

Adversarial training Kurakin et al. (2016). This approach involves adversarial training as a form of data augmentation where it injects adversarial examples during training. In every training mini batch, a mixture of clean images and adversarial images generated by one step Fast Gradient Sign Method (FGSM) are used to update the network's parameters. It is then improved by Na et al. (2017) by adding adversarial examples generated by iterative methods. In Madry et al. (2017) it is proposed to replace all clean images with adversarial images which is a direct result of optimizing a saddle point (min-max) formulation. By studying the loss landscape of the problem they suggest that PGD is a universal first-order adversary which is then used in their adversary generating process.

Denoising preprocessing. Various proposed defense approaches fall under the category of denoising preprocessing (Prakash et al., 2018; Liao et al., 2018; Song et al., 2017; Samangouei et al., 2018). Their effectiveness mostly comes from obfuscated gradients such that an optimization based attacker does not have usable gradients, and the lack of knowledge about their actual processing steps. These defenses are proved to be easily broken by designing attacks that could overcome the imposed obfuscated gradients and the corresponding preprocessing steps (Athalye et al., 2018; Athalye & Carlini, 2018). In our model, since we utilize implicit denoising by extracting similar features between natural and adversarial images through optimizing a L_2 regularizer, there is no dependency on a specific preprocessing model which could possibly be countered.

Attention Models. Attention in CNN is most commonly deployed for query-based tasks (Seo et al., 2016; Xu et al., 2015; Jetley et al., 2018). In Jetley et al. (2018) a method is presented to use a learned representation of the global image as a query to leverage multiple attention maps at different scales, which allows the expression of a complementary focus on different parts of the image. The application of attention to the adversarial robustness aspects has not been seriously explored. To the best of our knowledge, we are the first to employ an attention mechanism in training a robust deep neural network. In our application, we use a ReLU activated neural network instead of the linear-based method as the attention estimator. It allows highly non-linear compatibility between the learned global features and the lower-level local features.

3 APPROACH

We now present our model that combines the attention module and implicit denoising, and show how it can be applied to enhance the adversarial training to improve the adversarial robustness of a model and its accuracy. Figure 2 provides an overview of our method. We adopt the terminology in Jetley et al. (2018), in particular that local features and global features refer to features extracted at certain layers of a network whose effective receptive fields are subsets of the image and the entire image. We start by forwarding each of the clean and adversarial images and computing the attention weights by a non-linear estimator. Then the individual attention feature is defined to be the weighted combination of the corresponding local features. Next, we define an L_2 regularization loss as the Euclidean distance between the two sets of learned attention features. The attention features of the adversarial image are then used to produce the logits, which is followed by softmax layer to produce the cross-entropy loss. The final loss function of our model is a combination of cross-entropy loss and the regularization loss.

3.1 ADVERSARIAL TRAINING

We adopt the adversarial training described in Madry et al. (2017) as the basic training approach. It replaces natural training examples by PGD examples, which is suggested to represent a universal first-order adversary. So far PGD has been shown to represent the strongest attack model (Athalye et al., 2018; Athalye & Carlini, 2018). A model that is trained with PGD adversaries is also robust against a wide range of other attacks and not yet outperformed by any other approach. The adversarial training has a saddle point formulation:

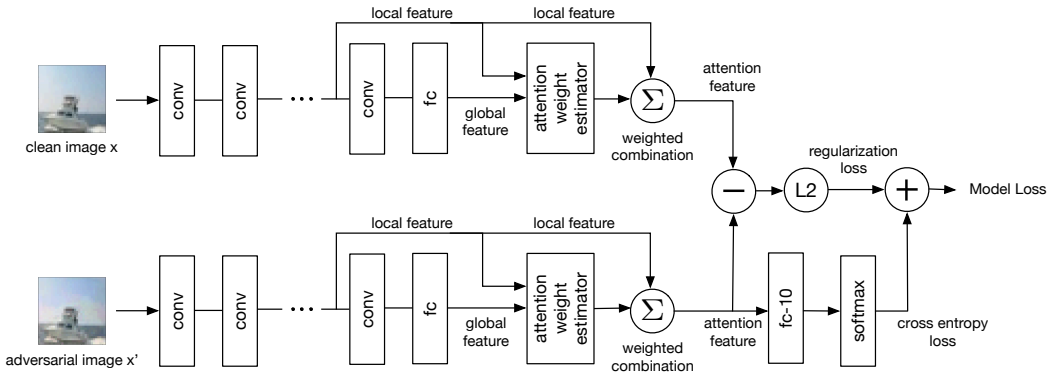

Figure 2: Overview of the proposed model. The top and bottom networks are the same copy that share all network parameters. Both the clean and adversarial images are forwarded through the network to produce the corresponding attention features. The L_2 regularization loss is defined as the Euclidean distance between the two sets of attention features. The final model loss is a combination of the L_2 regularization loss and the cross-entropy loss for only the adversarial input.

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\max_{\boldsymbol{\delta}\in S} L(\theta, \boldsymbol{x}+\boldsymbol{\delta}, y)] \tag{1}$$

where $\mathcal{D}$ is the distribution of data $\boldsymbol{x}$ and class labels y, L is the cross-entropy loss function for a model with parameters θ, $\boldsymbol{\delta}$ is the additive adversarial perturbation with bound S. In this paper we consider l_∞ bound as in Madry et al. (2017). Our adversarial samples $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$ are created by PGD:

$$\boldsymbol{x}^{t+1} = \Pi_{\boldsymbol{x}+S}\left(\boldsymbol{x}^t + \alpha\,\mathrm{sgn}(\nabla_{\boldsymbol{x}} L(\theta, \boldsymbol{x}, y))\right) \tag{2}$$

PGD adversaries are computed at each iteration as an approximated optimum of the inner maximization in equation 1 and an update of the parameters θ is made according to the outer minimization formulation.

3.2 Attention model

As we discussed in Section 1 a possible goal is to build a model that relies solely on robust features in making predictions. While adversarially trained models exhibit such behavior to some extent in that they utilize robust features that align relatively well with salient data characteristics (Tsipras et al., 2018), there is no design component to proactively select different features. We propose a non-linear attention model that acts as a feature prioritizing scheme, which is able to put more weight on robust features and less weight on non-robust features to increase the robustness of a classifier.

Let $\boldsymbol{l}_n^i$ be the learned feature vector at layer $i \in \{1, 2, ..., I\}$ of a neural network at spatial location $n \in \{1, 2, ..., N\}$, and let $\boldsymbol{g}$ be the feature vector of the layer just before the final fully connected layer which produces the class label prediction scores (logits). We use a small ReLU activated neural network to generate compatibility scores between the global feature $\boldsymbol{g}$ and local features $\boldsymbol{l}_n^i$:

$$c_n^i = f(\boldsymbol{l}_n^i, \boldsymbol{g}) \tag{3}$$

where f is the neural network and the concatenation of $\boldsymbol{g}$ and $\boldsymbol{l}_n^i$ is fed to the network to produce the compatibility scores c_n^i. We then normalize the scores with a softmax operation to get the attention weights:

$$w_n^i = \frac{\exp c_n^i}{\sum_m \exp c_m^i} \tag{4}$$

Next, we compute the weighted sum of local feature vectors which is the attention feature vector at layer i:

$$\boldsymbol{h}^i = \sum_n w_n^i \boldsymbol{l}_n^i \tag{5}$$

We may generate attention features at multiple layers in order to focus on different but complementary parts of an image. In this case, we concatenate $\boldsymbol{h}^i$ for all layers i considered and replace the global feature $\boldsymbol{g}$ with the new descriptor for final classification.

By using a small neural network instead of the linear alignment models as in Jetley et al. (2018), we are able to capture non-linear compatibility between the local and global features when producing the attention weights, which is beneficial considering the multiple non-linear function activated layers between the local and global features.

3.3 Implicit denoising

In addition to the attention mechanism, we also propose a method to implicitly denoise the adversarial samples via a L_2 regularization term that encourages the model to extract similar features for the clean image and the corresponding adversarial image. Denote by $\mathcal{G}_\theta$ the deep neural network, $\boldsymbol{x}$ and $\boldsymbol{x}'$ the natural image and adversarial image. Denote by $\mathcal{G}_\theta(\boldsymbol{x})$, $\mathcal{G}_\theta(\boldsymbol{x}')$ the learned features of the layer just before the final fully connected layer (in our case this is the attention weighted global descriptor) which produces the class label prediction scores. The L_2 regularizer has the following form:

$$L_r(\boldsymbol{x}, \boldsymbol{x}') = \|\mathcal{G}_\theta(\boldsymbol{x}) - \mathcal{G}_\theta(\boldsymbol{x}')\|_2 \tag{6}$$

By minimizing the regularization function, the model effectively learns very similar features for the clean sample and the adversarial sample, therefore ignoring the added adversarial noise. From another perspective, the learned features of a neural network lie on a high dimensional manifold that is linearly separable for different classes because the classification layer is a linear classifier followed by softmax function. With adversarial training alone, a model only tries to map $\boldsymbol{x}$ and $\boldsymbol{x}'$ to the same side of the decision boundary, while with the additional regularization, they are not only on the same side but also mapped to nearby points in the space. This mapping is a desired behavior considering that, in the original image space, they are very close points representing essentially the same image.

3.4 Model loss

Equipped with the presented methods, the total loss of our model is:

$$\text{Loss} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[L(\theta, \boldsymbol{x}', y) + \lambda\|\mathcal{G}_\theta(\boldsymbol{x}) - \mathcal{G}_\theta(\boldsymbol{x}')\|_2] \tag{7}$$

where λ is a hyperparameter that controls the relative weight of the L_2 regularization loss.

4 Experiments and Results

In this section, we evaluate our model on the MNIST and CIFAR-10 datasets, and present both quantitative and qualitative results.

4.1 Results on MNIST

First we present the results on the MNIST dataset. We use a CNN with two convolutional layers with 32 and 64 filters respectively, each followed by 2×2 max-pooling, and a fully connected layer of size 1024. For the PGD adversary, we run 40 iterations with a step size of 0.01 and l_∞ bound of $\epsilon = 0.3$. The settings are the same as in Madry et al. (2017). Since MNIST is a very small scale dataset and the model is very robust with just adversarial training, we do not employ the attention mechanism on MNIST. We only study the effectiveness of the proposed implicit denoising method.

The evaluation results on MNIST are presented in Table 1. Regarding the value of weight λ of the L_2 regularizer, we find that roughly any $\lambda \in [0.001, 0.05]$ works well. The reported results are obtained with $\lambda = 0.01$. From Table 1, we can see that a model trained with the proposed implicit denoising method is significantly more robust against PGD adversary than the baseline model that uses adversarial training alone. The improvement is nearly 3% for white box attack and 1% for black box attack.

On the other hand, we also note that there is a slight decline in the accuracy of natural examples even though the difference is very small relative to the gain in adversarial accuracy. This is caused

Table 1: Performance comparison of the adversarial training method (Madry et al., 2017) and our proposed adversarial training with implicit denoising (AT-id) method on MNIST. Black box accuracies are evaluated against adversaries generated from an independently trained copy of the same method with identical configurations.

Method	**Natural**	**White box**	**Black box**
Madry et al. (2017)	**98.72%**	92.86%	95.97%
AT-id	98.66%	**95.69%**	**96.90%**

by the inherent trade-off of classification accuracy between natural and adversarial examples as we discussed in Section 1. Implicit denoising only modifies the training objective function and the structure of a model is the same, therefore the trade-off applies in this case. In the next section we will show that by using attention module and implicit denoising, a model can improve both the standard classification accuracy on natural examples and its adversarial robustness.

4.2 Results on CIFAR-10

4.2.1 Classification Accuracy

Here we present our results on the CIFAR-10 dataset. For the base network, we use the original Resnet model that has four residual blocks with [16, 16, 32, 64] filters respectively. For our attention model, we modify the Resnet by replacing the spatial global average pooling layer after the residual block 4 with a convolutional layer sandwiched between two max-pooling layers to obtain the global feature $\boldsymbol{g}$. We use a one-hidden-layer neural network with 64 hidden units and ReLU activation function as the non-linear attention weight estimator. For the PGD adversary, we run 5 iterations with a step size of 2 and l_∞ bound of $\epsilon = 8/255$. In order to isolate and analyze the effectiveness of attention module and implicit denoising independently, we train three models with the following configurations: AT-id is an adversarial trained model with implicit denoising, AT-att is an adversarial trained model with attention module, and AT-att-id is the model with both attention and implicit denoising.

Table 2: Performance comparison of the adversarial training (Madry et al., 2017), adversarial training with implicit denoising (AT-id), adversarial training with attention model (AT-att), and adversarial training with both (AT-att-id) on CIFAR-10. Black box accuracies are evaluated against adversaries generated from an independently trained copy of the same method with identical configurations.

Method	**Natural**	**White box**	**Black box**
Madry et al. (2017)	80.79%	49.89%	60.13%
AT-id	79.52%	52.35%	61.82%
AT-att	**82.71%**	51.06%	**63.88%**
AT-att-id	81.86%	**52.50%**	63.66%

The evaluation results of aforementioned four models on CIFAR10 are presented in Table 2. Similar with MNIST, we find that roughly any $\lambda \in [0.01, 0.1]$ works well for implicit denoising. The reported results are obtained with $\lambda = 0.1$ for AT-id and $\lambda = 0.01$ for AT-att and AT-att-id. From the table, we see that all of the three proposed models have better adversarial robustness over the baseline model that only uses adversarial training, and both models with attention show improvement on the classification accuracy on natural examples as well.

By comparing AT-id with adversarial training (Madry et al., 2017), we note that the implicit denoising method provides a significant improvement on robustness over the baseline method, with a decline on the natural example accuracy. Same trade-off appears in the comparison of AT-att-id and AT-att. With the same structure, a model trained with implicit denoising utilizes more robust fea-

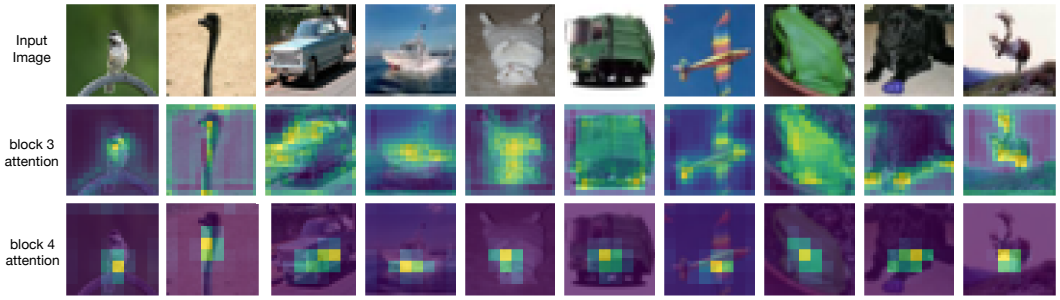

Figure 3: The learned attention maps of our model.The first row are the input images, and the second and third row are the attention maps learned at residual block 3 and 4 respectively.

tures and less non-robust features. Since standard accuracy benefits from both robust and non-robust features, it drops when such non-robust features are less used.

On the other hand, by comparing the results of models with and without the attention module, we can see that attention contributes to both standard and adversarial accuracy. The attention structure not only favors robust features, it also relies heavily on features extracted from the spatial area that contains the actual object of concern. By suppressing the features extracted from the background clutter and misleading perturbations in irrelevant areas, the model with attention module more precisely learns the underlying distribution of the data therefore the better accuracy. In the next section, we will demonstrate that the attention weights sharply focus on the object in the image and ignore the irrelevant areas.

4.2.2 Attention Map

We show the attention maps at residual block 3 and 4 of our model in Figure 3. The attention maps at both levels focus sharply on the objects in the images, and the two level attention maps captures different and complementary parts of the objects. The attention maps at block 3 appear to attend to most part of the objects, since at early stages, a CNN only learns basic features like curves and lines. At block 4, the attention maps are localized on the most relevant features like the face of a cat or a dog and the wings of an airplane because high-level abstract features are learned.

4.2.3 Gradient Map

In this section we study the gradient maps, which are the gradients of the cross-entropy loss with respect to the input image pixels, of both the baseline and our model. The gradient maps are a direct indicator on how the input features are utilized by a model to produce a final prediction. A large magnitude of the gradient on an input feature signifies a heavy dependence the model has on that feature. By studying the gradient map we can determine which part of the image the model relies on. If the gradient map aligns perfectly with the input image, the model depends solely on those robust features. Next, we show that the gradient maps generated from our model align better with the salient data characteristics by evaluating the gradient maps both qualitatively and quantitatively.

First we present the qualitative result. Figure 4 contains the gradient maps from Madry et al. (2017) and our model of the corresponding input images. These are raw gradients with only being clipped and rescaled for visualization. Overall, we note that both models generate highly interpretable gradient maps that align very well with the image features. However, upon carefully inspecting the images, it is evident that the gradient maps generated from our model are better than the baseline model. To point out a few: in columns 2, 3, 6 and 9, our gradient maps have cleaner backgrounds and the gradients only focus on the objects; especially in column 6, the baseline model has large gradients on the text field in the background which is irrelevant to the class label (automobile), while in our model gradients in that area are much more suppressed. In columns 1, 5, and 10 the faces and heads of the animals are depicted clearer in our model.

Next, we introduce a quantitative evaluation method for gradient maps. The problem we consider here is to decide how well the gradient maps align with the original images. The better they align,

Figure 4: Gradient maps generated by compute gradient of model's cross-entropy loss with respect to the input images. Top row are the original CIFAR-10 images, midrow are the gradient maps of Madry et al. (2017), and bottom row are the gradient maps of our model. The raw gradients are clipped to within ± 3 standard deviation and rescaled to lie in the [0, 1] range for visualization. No other preprocessing is applied.

the more recognizable the gradient images are. Therefore, instead of relying on human inspection to decide the quality of alignment, which is very subjective, we propose to use a pretrained neural network to classify the gradient maps for all images in the dataset, both the training set and the test set. If the classifier is able to predict the correct label from the gradient map image, it is verified to match perfectly with the original image. With the classification accuracy of gradient maps from both models of all images we are able to determine quantitatively the relative performance.

The pretrained model we use is the same Resnet model as in Section 4.2.1 trained with only natural training data of CIFAR-10. It achieves an accuracy of 88.79% on the test set. The classification results are presented in Table 3. To avoid the possible influence of gradient clipping we also run the evaluation on raw gradients. As demonstrated by the classification accuracy, the gradient maps from our model express significantly better alignment with the corresponding original images.

Table 3: Classification accuracy on the gradient maps from baseline and our methods on both the training dataset and test dataset of CIFAR-10. We run the experiment on gradient maps both with and without clipping to avoid the influence of gradient clipping.

	With clipping		**Without clipping**	
Method	**Train data**	**Test data**	**Train data**	**Test data**
Madry et al. (2017)	27.10%	26.78%	28.60%	28.72%
Ours	**30.13%**	**30.25%**	**31.26%**	**31.53%**

To summarize, both the qualitative and quantitative results show that the gradient maps from our model have better interpretability and alignment with the original images. It suggests that our model heavily depends on the features that are very closely correlated with the robust features of the input images. Background clutter and irrelevant features are suppressed by our model which explains the improved performance on both standard accuracy and adversarial robustness.

5 Conclusion

In this paper we propose feature prioritization and regularization to enhance the adversarial robustness. With the non-linear attention module and implicit denoising, a model is improved on both the standard classification accuracy and the adversarial robustness over the adversarial training approach. In addition to qualitative study of the attention maps and gradient maps, we propose and conduct a quantitative evaluation to demonstrate that the attention maps focus sharply on the region of interest and the gradient maps align perfectly with salient data characteristics, and therefore our model heavily relies on the robust features.

REFERENCES

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *arXiv preprint arXiv:1801.00553*, 2018.

Anish Athalye and Nicholas Carlini. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018.

Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015.

Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

J Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Jun Zhu, and Xiaolin Hu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Taesik Na, Jong Hwan Ko, and Saibal Mukhopadhyay. Cascade adversarial machine learning regularized with a unified embedding. *arXiv preprint arXiv:1708.02582*, 2017.

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8571–8580, 2018.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

Paul Hongsuck Seo, Zhe Lin, Scott Cohen, Xiaohui Shen, and Bohyung Han. Progressive attention networks for visual attribute prediction. *arXiv preprint arXiv:1606.02393*, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thomas Tanay, Jerone TA Andrews, and Lewis D Griffin. Built-in vulnerabilities to imperceptible adversarial perturbations. *arXiv preprint arXiv:1806.07409*, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5283–5292, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.