Feature Prioritization and Regularization Improve Standard Accuracy and Adversarial Robustness

Chihuang Liu[*] and **Joseph JaJa**

Department of Electrical and Computer Engineering
University of Maryland
College Park, MD 20742, USA

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. We propose a model that employs feature prioritization by a nonlinear attention module and L_2 feature regularization to improve the adversarial robustness and the standard accuracy relative to adversarial training. The attention module encourages the model to rely heavily on robust features by assigning larger weights to them while suppressing non-robust features. The regularizer encourages the model to extract similar features for the natural and adversarial images, effectively ignoring the added perturbation. In addition to evaluating the robustness of our model, we provide justification for the attention module and 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; Simonyan and Zisserman, 2014]. 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; Tramèr *et al.*, 2017; Fawzi *et al.*, 2018; Athalye and Sutskever, 2017; Carlini and Wagner, 2017; Kolter and Wong, 2017; Wong and Kolter, 2018; Madry *et al.*, 2017]. A wide variety of methods are proposed to defend a model against adversarial attacks [Prakash *et al.*, 2018; Liao *et al.*, 2017; Song *et al.*, 2017; Samangouei *et al.*, 2018]. Despite these advances, many techniques are subsequently shown to be ineffective [Athalye *et al.*, 2018; Athalye and Carlini, 2018], and adversarial training which uses adversarial samples in addition to clean images during the training process 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 training provide insights of its effectiveness. It is shown that standard and robust models depend on very different sets of features [Tsipras *et al.*, 2018; Tanay *et al.*, 2018]. While standard models utilize features including non-robust ones that are weakly correlated with class labels and easily manipulated by small input perturbations, robust models only use robust features that are highly correlated with class labels and invariant to those perturbations. Although adversarial training learns robust features, there are no explicit design components to encourage a model to depend solely on robust features. Therefore, to further improve the robustness of a model, we propose feature regularization and prioritization schemes.

We first propose to use an attention mechanism that introduces to the model the flexibility of prioritizing features and bias a model towards robust features. We call the learned features at the final layer of a network the global features, and the ones at lower level layers the local features. In our attention module, we use the global features as a way to assign weights to the local features by a non-linear compatibility function. Since global features are directly used to produce class label prediction, we are effectively assigning weights to local features depending on their correlation with the labels. Robust features have higher correlation and therefore will be assigned larger weights which in turn contribute to the model's robustness.

Next, we propose to use feature regularization to learn robust features that are invariant to input perturbations. We add an L_2 regularization term that penalizes the distance between

[*] Contact Author

the learned features of a clean sample x and that of its perturbed adversarial counterpart x' to the training objective. By optimizing this regularizer, we are pushing the model to extract very similar features from the original image and the adversarial image, and thus only features that are invariant to the perturbations are learned and the added noise is effectively ignored. From another point of view, a model with small L_2 feature distance maps the two nearby points in the image space to nearby points in the learned high dimensional manifold, which is a desirable behavior.

In this paper, we propose an approach that enhances adversarial training with feature prioritization and regularization to improve the robustness of a model. We use extensive experiments to demonstrate that the attention module focuses on the area of an image which contains the actual object and helps the classifier to only rely on features extracted from those areas. The background clutter and irrelevant features which could be misleading are effectively suppressed. The feature regularization further encourages the model to extract robust features that are not manipulated by the adversarial perturbations. The resulting model has a highly interpretable gradient map that aligns very well with salient data characteristics.

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, CIFAR-10, and CIFAR-100 datasets, and demonstrates superior performance relative to both standard classification accuracy and adversarial robustness.
- We provide empirical justification for attention and show that it helps the model to rely on robust features by assigning larger weights to them. Through qualitative inspection, we show that the attention maps generated by our non-linear attention estimator focus sharply on the regions of interest while suppressing irrelevant background clutter.
- In addition to qualitative evaluation of the gradient maps, we propose a novel experimental strategy that quantitatively demonstrates better 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, we only review some of the most related works in this section. For a more comprehensive survey, please refer to Akhtar and Mian [2018].

Adversarial training. Kurakin et al. [2016] use 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 was later improved by Na et al. [2017] by adding adversarial examples generated by iterative methods. Madry et al. [2017] proposed to replace all clean images with adversarial images which is a direct result of optimizing a saddle point (min-max) formulation. They suggest that PGD is a universal first-order adversary which is then used in their adversary generating process.

L_2 regularization. A similar idea with feature regularization is proposed in Kannan et al. [2018] which they call adversarial logit pairing (ALP), to prevent a model from being over-confident when making predictions. Compared with ALP, feature regularization is more intuitive as it motivates a model to learn very robust features that are invariant to input perturbations, which leads to a robust model. In addition, we propose to also use attention module to further encourage the model to favor robust features which will improve the robustness.

Attention Models. Attention in CNN is most commonly deployed for query-based tasks [Seo *et al.*, 2016; Jetley *et al.*, 2018]. Jetley et al. [2018] presented a method 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. However, 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 L_2 feature regularization, and show how it can be applied to enhance the adversarial training to improve the adversarial robustness of a model and its accuracy. Figure 1 provides an overview of our method. 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 to be 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 method [Athalye *et al.*, 2018; Athalye and 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:

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

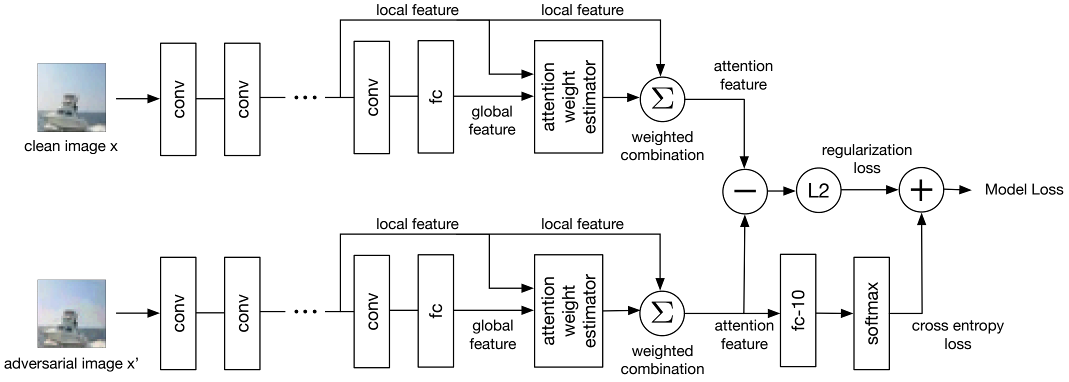

Figure 1: Overview of the proposed model. The top and bottom networks are the same copy that share all network parameters. Both the clean and adversarial images are forwarded through the network to produce the corresponding attention features. The L_2 regularization loss is defined as the Euclidean distance 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.

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:

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

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, our goal of attention model is to favor robust features in making predictions. 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 one-hidden-layer ReLU 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 use the outputs of the last residual block as the local feature for computing attention, and replace the global feature $\boldsymbol{g}$ with the corresponding attention descriptor $\boldsymbol{h}^i$ for final classification.

By using a 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 Feature Regularization

In addition to the attention mechanism, we also propose to use an L_2 regularization term to encourage 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, which are robust features since they are invariant to the adversarial perturbation. 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 a 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:

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

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, CIFAR-10 and CIFAR-100 datasets, and present empirical justification to attention module and some quantitative and qualitative results.

4.1 Robustness on MNIST

We use a CNN with two convolutional layers with 32 and 64 filters respectively, followed by two fully connected layers of size 1024 and 10. The network is trained with 40-step PGD adversary 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, but only study the effectiveness of the proposed feature regularization method.

Method	**Natural**	**White box**	**Transfer**
Madry et al.	98.72%	92.86%	95.97%
AT-reg	**98.97%**	**95.95%**	**96.90%**

Table 1: Performance comparison of the adversarial training and our proposed adversarial training with feature regularization (AT-reg) on MNIST against PGD 5 adversaries. Transfer attack accuracies are evaluated against adversaries generated from an independently trained copy of the same method with identical configurations.

The evaluation results are presented in Table 1. Regarding the value of weight λ of the L_2 regularizer, we find that roughly any $\lambda \in [0.001, 0.1]$ works well. The reported results are obtained with $\lambda = 0.1$. Table 1 shows that a model trained with the proposed feature regularization method is significantly more robust against PGD adversary than the baseline model with adversarial training alone. The improvement is more than 3% for white box attack and 1% for transfer attack.

4.2 Robustness on CIFAR-10

We use the same wide residual network as Madry et al. [2017] with [16, 16, 32, 64] filters and its 3-times wider variant with [16, 48, 96, 192] 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 ReLU neural network with 64 hidden units as the non-linear attention weight estimator. The model is trained with 5-step PGD adversary and a step size of 2 and l_∞ bound of $\epsilon = 8/255$. In order to isolate and analyze the effectiveness of attention module and feature regularization independently, we train three models with the following configurations: adversarial training with feature regularization (AT-reg), adversarial training with attention (AT-att), and adversarial training with both attention and feature regularization (AT-att-reg). We then evaluate the models using PGD and CW [Carlini and Wagner, 2017] adversaries optimized with various number of steps under white box and transfer attack settings. Transfer attack accuracies are evaluated against adversaries generated from an independently trained copy of the same method with identical configurations.

Method	**Madry et al.**	**AT-reg**	**AT-att**	**AT-att-reg**
Natural	80.79%	79.52%	**82.43%**	81.20%
White, PGD 5	49.89%	52.35%	51.22%	**53.38%**
White, PGD 20	39.72%	44.25%	41.40%	**45.20%**
White, PGD 100	38.76%	43.73%	40.85%	**44.60%**
White, PGD 200	38.64%	43.70%	40.72%	**44.54%**
White, CW 30	40.27%	42.96%	40.60%	**44.37%**
White, CW 100	39.98%	42.87%	40.31%	**44.26%**
Transfer, PGD 5	60.13%	61.82%	**63.26%**	62.32%
Transfer, PGD 20	56.60%	56.40%	**58.26%**	57.52%
Transfer, PGD 100	56.44%	56.28%	**58.08%**	57.53%
Transfer, PGD 200	56.49%	56.25%	**58.04%**	57.53%
Transfer, CW 30	57.11%	56.86%	**58.07%**	57.57%
Transfer, CW 100	57.10%	56.79%	**58.13%**	57.41%

Table 2: Performance comparison of the adversarial training, adversarial training with feature regularization (AT-reg), adversarial training with attention (AT-att), and adversarial training with both (AT-att-reg) on CIFAR-10 against PGD and CW adversaries.

Method	**Madry et al.**	**AT-reg**	**AT-att**	**AT-att-reg**
Natural	85.41%	84.65%	**86.48%**	85.98%
White, PGD 5	49.15%	52.21%	50.91%	**53.23%**
White, PGD 20	38.19%	41.00%	39.52%	**41.55%**
White, PGD 100	37.39%	40.28%	38.98%	**40.78%**
White, PGD 200	37.20%	40.24%	38.89%	**40.67%**
White, CW 30	38.92%	**42.20%**	40.75%	42.12%
White, CW 100	38.71%	41.88%	40.32%	**42.06%**
Transfer, PGD 5	68.03%	68.00%	**69.78%**	69.00%
Transfer, PGD 20	62.77%	63.14%	**64.70%**	64.01%
Transfer, PGD 100	62.65%	63.12%	**64.78%**	64.15%
Transfer, PGD 200	62.73%	63.12%	**64.69%**	64.11%
Transfer, CW 30	63.89%	64.23%	**65.65%**	64.50%
Transfer, CW 100	63.77%	64.04%	**65.43%**	64.35%

Table 3: Performance comparison of the adversarial training, adversarial training with feature regularization (AT-reg), adversarial training with attention (AT-att), and adversarial training with both (AT-att-reg) on CIFAR-10 using the 3-times wide ResNet network against PGD and CW adversaries.

The evaluation results of aforementioned narrow and 3-times wide models on CIFAR10 are presented in Table 2 and

Tabel 3. We find that roughly any $\lambda \in [0.01, 1]$ works well for feature regularization. The reported results are obtained with $\lambda = 1$ for AT-reg and $\lambda = 0.1$ for AT-att-reg. 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. Note that the white box accuracy of the 3-times wide models are lower than the narrow models, which is due to overfitting of adversarial training [Song *et al.*, 2018], and could be solved by early stopping.

We assess the effectiveness of feature regularization by comparing baseline to AT-reg in both Tables. The robustness is significantly improved especially under stronger attack with more steps, more than 5% in some cases.

Next, by comparing the results of models with and without the attention module, we can see that attention contributes to both standard and adversarial accuracy such that the robustness improves by 1.5% and at the same time standard accuracy by a similar margin. 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 background clutter and misleading perturbations in irrelevant areas, the model with attention module more precisely learns the underlying distribution of the data which leads to better accuracy.

AT-att-reg takes advantage of both techniques and offers the most improvement in robustness and also better standard accuracy. We think attention and feature regularization complements each other. While regularization encourages the model to learn robust features from the input, attention assigns larger weights to robust features to make the model rely even more on them.

4.3 Empirical Justification for Attention

First of all, in order to show the advantage of our proposed non-linear attention model over the linear alignment model in Jetley et al. [2018], we train two narrow ResNet on natural CIFAR-10 samples with the two attention schemes respectively and the accuracy is 92.87% for our attention and 91.34% for the linear attention model.

Next, we empirically demonstrate that the attention module assigns larger weights to more robust features by examining the attention weights relative to the feature robustness. Figure 2 shows the relationship between the robustness of a feature and the magnitude of its assigned attention weight. The robustness measure we use is the L_2 distance between the learned features of a clean and an adversarial image, i.e. the smaller the distance between the features, the more invariant the feature is against input perturbations, therefore the feature is more robust. We rank the features accordingly and compute the average attention weights for each feature across all images. As shown in Figure 2, our proposed attention mechanism indeed assigns larger weights to more robust features and less weights to non-robust features, so the model is more invariant to adversarial perturbation.

Finally, we show the attention maps of our model in Figure 3 to visualize the attention weights. The attention maps focus sharply on the objects in the images, and the most relevant features like the head and legs of an animal and the wings of an airplane contribute more to the model's prediction and ignore the irrelevant background clutter.

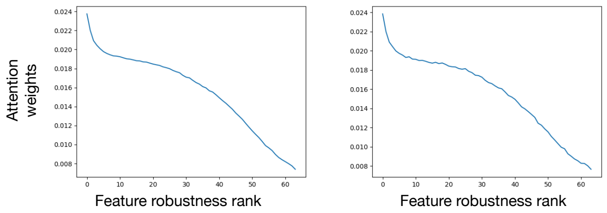

Figure 2: The relationship between attention weights and feature robustness. The horizontal axis is the robustness rank, with 0 being the most robust and 63 the least robust, and the vertical axis is the corresponding attention weights. Left plot shows the results for training set and right plot is for test set of CIFAR-10.

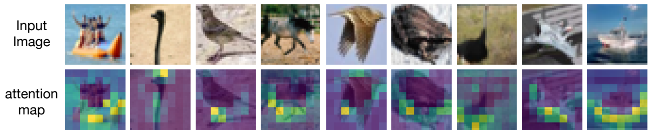

Figure 3: The learned attention maps of our model. The first row are the input images and the second are the attention maps learned at residual block 4.

4.4 Gradient Map

In this section, we study the gradient maps, which are the gradients of the cross-entropy loss with respect to input image pixels. The gradient maps directly indicate how the input features are utilized by a model for prediction and highlight the features which affect the loss most strongly. A large gradient on an input feature signifies a heavy dependence from the model. Human vision is robust against small input perturbations and the perception of an image reflects the reliance on input features. Therefore, models that depend on robust features will be better aligned with human vision, and the alignment can be used to evaluate the robustness of a model. Next, we show that the gradient maps generated from our model align better with the salient data characteristics.

First we present the qualitative result. Figure 4 shows the gradient maps from Madry et al. [2017] and our model. Overall, we note that gradient maps from both models are highly interpretable and align very well with the image features. Upon careful inspection, it is evident that the gradient maps generated from our model are better than the baseline. To point out a few, note that, in columns 2, 3, 6 and 8, 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 9 the edges of the faces and heads of the animals are depicted clearer in our model. Human inspection

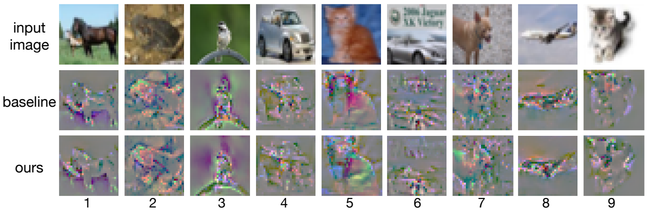

Figure 4: Original CIFAR-10 images (top row) and corresponding gradient maps from Madry et al. [2017] model (mid row) and our model (bottom row). The raw gradients are clipped to within ±3 standard deviation and rescaled to lie in the [0, 1] range for visualization.

which could be very subjective, therefore, we also 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, the more recognizable the gradient images are. A standard neural network extracts relevant features and make predictions based on them. When a gradient map is highly aligned with the original image, the neural net is able to identify more relevant features and thus the classification accuracy will be higher. Therefore, we can quantitatively compare the alignment by the classification accuracy of gradient maps. We use a pretrained network to classify the gradient maps for images in both the training set and the test set.

We pretrained the same ResNet model as in Section 4.2 with only natural training data of CIFAR-10. It achieves 88.79% accuracy on the test set. The classification results are presented in Table 4. 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 original images.

	With clipping		**Without clipping**	
Method	**Train data**	**Test data**	**Train data**	**Test data**
Madry et al.	27.10%	26.78%	28.60%	28.72%
Ours	**30.11%**	**30.32%**	**31.46%**	**31.59%**

Table 4: Classification accuracy on the gradient maps from baseline and our model on both the training and test set of CIFAR-10. We run the experiment on gradient maps both with and without clipping to avoid the influence of gradient clipping.

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 depends on robust features of the input images which explains the improved performance on both standard accuracy and adversarial robustness.

4.5 Results on CIFAR-100

Here we present our results on the CIFAR-100 dataset. The experiment setup is the same as CIFAR-10 in Section 4.2.

Method	**Madry et al.**	**AT-reg**	**AT-att**	**AT-att-reg**
Natural	52.70%	49.53%	**53.67%**	50.66%
White, PGD 5	25.14%	26.99%	26.33%	**27.76%**
White, PGD 20	19.65%	23.16%	20.82%	**23.80%**
White, PGD 100	19.47%	23.07%	20.59%	**23.62%**
White, PGD 200	19.41%	22.96%	20.53%	**23.62%**
White, CW 30	18.64%	20.78%	19.39%	**20.88%**
White, CW 100	18.61%	20.63%	19.26%	**20.76%**
Transfer, PGD 5	35.37%	35.10%	**35.95%**	35.17%
Transfer, PGD 20	31.99%	31.88%	**32.48%**	32.04%
Transfer, PGD 100	32.03%	31.84%	**32.38%**	32.00%
Transfer, PGD 200	32.00%	31.80%	**32.37%**	32.06%
Transfer, CW 30	32.50%	31.96%	**32.75%**	29.81%
Transfer, CW 100	32.46%	31.86%	**32.74%**	29.71%

Table 5: Performance comparison of the adversarial training, adversarial training with feature regularization (AT-reg), adversarial training with attention (AT-att), and adversarial training with both (AT-att-reg) on CIFAR-100 against PGD and CW adversaries.

From Table 5, similar with CIFAR-10, our model outperforms baseline by as much as 4% on CIFAR-100 dataset.

4.6 Comparison of Feature Regularization with Adversarial Logit Pairing

Though feature regularization is more intuitive than adversarial logit pairing (ALP) as explained in Section 2, their performance under PGD attack is similar. However, because ALP is based on logits, it is not as robust as feature regularization under CW attack [Carlini and Wagner, 2017], which is based on logit margin.

Method	**ALP**		**AT-reg**	
Adversary	CW 30	CW 100	CW 30	CW 100
Accuracy	39.17%	39.15%	**42.96%**	**42.87%**

Table 6: Performance comparison of ALP and feature regularization on CIFAR-10 against CW adversaries.

As shown in Table 6, AT-reg outperforms ALP by around 3.8% against CW attacks.

5 Conclusion

In this paper, we propose feature prioritization and regularization to enhance both standard classification accuracy and adversarial robustness of a model over the baseline adversarial training approach. We provide empirical justifications for attention to show that it effectively favors robust features and focuses sharply on the region of interest. We then conduct quantitative and qualitative evaluation on gradient maps to show that they align perfectly with salient data characteristics, which further proves that our model heavily relies on the robust features.

References

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

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

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

[Athalye *et al.*, 2018] 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.

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

[Dvijotham *et al.*, 2018] 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.

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

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

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

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

[Kannan *et al.*, 2018] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

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

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

[Liao *et al.*, 2017] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. *arXiv preprint arXiv:1712.02976*, 2017.

[Madry *et al.*, 2017] 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.

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

[Prakash *et al.*, 2018] 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*, pages 8571–8580, 2018.

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

[Seo *et al.*, 2016] 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.

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

[Song *et al.*, 2017] 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.

[Song *et al.*, 2018] Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Improving the generalization of adversarial training with domain adaptation. *arXiv preprint arXiv:1810.00740*, 2018.

[Szegedy *et al.*, 2013] 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.

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

[Tramèr *et al.*, 2017] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[Tsipras *et al.*, 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.

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