ATLPA: ADVERSARIAL TOLERANT LOGIT PAIRING WITH ATTENTION FOR CONVOLUTIONAL NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Though deep neural networks have achieved the state of the art performance in visual classification, recent studies have shown that they are all vulnerable to the attack of adversarial examples. To solve the problem, some regularization adversarial training methods, constraining the output label or logit, have been studied. In this paper, we propose a novel regularized adversarial training framework ATLPA,namely Adversarial Tolerant Logit Pairing with Attention. Instead of constraining a hard distribution (e.g., one-hot vectors or logit) in adversarial training, ATLPA uses Tolerant Logit which consists of confidence distribution on top-k classes and captures inter-class similarities at the image level. Specifically, in addition to minimizing the empirical loss, ATLPA encourages attention map for pairs of examples to be similar. When applied to clean examples and their adversarial counterparts, ATLPA improves accuracy on adversarial examples over adversarial training. We evaluate ATLPA with the state of the art algorithms, the experiment results show that our method outperforms these baselines with higher accuracy. Compared with previous work, our work is evaluated under highly challenging PGD attack: the maximum perturbation ϵ is 64 and 128 with 10 to 200 attack iterations.

1 INTRODUCTION

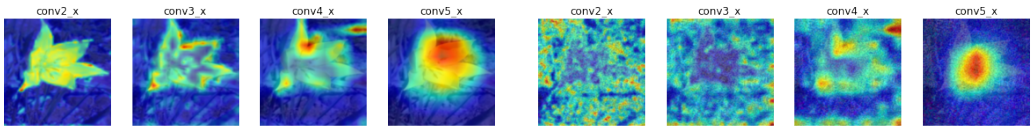
In recent years, deep neural networks have been extensively deployed for computer vision tasks, particularly visual classification problems, where new algorithms reported to achieve or even surpass the human performance(Krizhevsky et al., 2012; He et al., 2015; Li et al., 2019a). Success of deep neural networks has led to an explosion in demand. Recent studies(Szegedy et al., 2013; Goodfellow et al., 2014; Carlini & Wagner, 2016; Moosavi-Dezfooli et al., 2016; Bose & Aarabi, 2018) have shown that they are all vulnerable to the attack of adversarial examples. Small and often imperceptible perturbations to the input images are sufficient to fool the most powerful deep neural networks.

In order to solve this problem, many defence methods have been proposed, among which adversarial training is considered to be the most effective one(Athalye et al., 2018).Adversarial training(Goodfellow et al., 2014; Madry et al., 2017; Kannan et al., 2018; Tramèr et al., 2017; Pang et al., 2019) defends against adversarial perturbations by training networks on adversarial images that are generated on-the-fly during training. Although aforementioned methods demonstrated the power of adversarial training in defence, we argue that we need to perform research on at least the following two aspects in order to further improve current defence methods.

***Strictness** vs. **Tolerant**.* Most existing defence methods only fit the outputs of adversarial examples to the one-hot vectors of clean examples counterparts. Kannan et al. (2018) also fit confidence distribution on the all logits of clean examples counterparts, they call it as **Logits Pair**. Despite its effectiveness, this is not necessarily the optimal target to fit, because except for maximizing the confidence score of the primary class (i.e., the ground-truth), allowing for some secondary classes (i.e., those visually similar ones to the ground-truth) to be preserved may help to alleviate the risk of over-fitting(Yang et al., 2018). We fit **Tolerant Logit** which consists of confidence distribution on top-k classes and captures inter-class similarities at the image level. We believe that limited attention

should be devoted to top-k classes of the confidence score, rather than strictly fitting the confidence distribution of all classes. **A More Tolerant Teacher Educates Better Students.**

***Process** vs. **Result**.* In Fig. 1, we visualize the spatial attention map of a flower and its corresponding adversarial image on ResNet-101(He et al., 2015) pretrained on ImageNet(Russakovsky et al., 2015). The figure suggests that adversarial perturbations, while small in the pixel space, lead to very substantial noise in the attention map of the network. Whereas the features for the clean image appear to focus primarily on semantically informative content in the image, the attention map for the adversarial image are activated across semantically irrelevant regions as well. The state of the art adversarial training methods only encourage hard distribution of deep neural networks output (e.g., one-hot vectors(Madry et al., 2017; Tramèr et al., 2017) or logit(Kannan et al., 2018)) for pairs of clean examples and adversarial counterparts to be similar. In our opinion, it is not enough to align the difference between the clean examples and adversarial counterparts only at the output layer of the network, and we need to align the attention maps of middle layers of the whole network, e.g.,o uter layer outputs of $conv2.x$, $conv3.x$, $conv4.x$, $conv5.x$ in ResNet-101. **We can't just focus on the result, but also on the process.**



(a) original (b) adversarial

Figure 1: Spatial attention map of ResNet-101 pretrained on ImageNet(Russakovsky et al., 2015).(a) is original image and (b) is corresponding adversarial image.For ResNet-101, which we use exclusively in this paper, we grouped filters into stages as described in (He et al., 2015). These stages are $conv2.x$, $conv3.x$, $conv4.x$, $conv5.x$.

The contributions of this paper are the following:

- We propose a novel regularized adversarial training framework **ATLPA** : a method that uses **Tolerant Logit** and encourages attention map for pairs of examples to be similar. When applied to clean examples and their adversarial counterparts, **ATLPA** improves accuracy on adversarial examples over adversarial training. Instead of constraining a hard distribution in adversarial training, **Tolerant Logit** consists of confidence distribution on top-k classes and captures inter-class similarities at the image level.
- We explain the reason why our **ATLPA** can improve the robustness of the model from three dimensions: **average activations on discriminate parts**, **the diversity among learned features of different classes** and **trends of loss landscapes**.
- We show that our **ATLPA** achieves **the state of the art** defense on a wide range of datasets against strong **PGD** gray-box and black-box attacks. Compared with previous work, our work is evaluated under highly challenging PGD attack: the maximum perturbation $\epsilon \in \{0.25, 0.5\}$ i.e. $L_\infty \in \{0.25, 0.5\}$ with 10 to 200 attack iterations. To our knowledge, such a strong attack has not been previously explored on a wide range of datasets.

The rest of the paper is organized as follows: in Section 2 related works are summarized, in Section 3 definitions and threat models are introduced, in Section 4 our **ATLPA** is introduced, in Section 5 experimental results are presented and discussed, and finally in Section 6 the paper is concluded.

2 Related Work

Athalye et al. (2018) evaluate the robustness of nine papers(Buckman et al., 2018; Ma et al., 2018; Guo et al., 2017; Dhillon et al., 2018; Xie et al., 2017; Song et al., 2017; Samangouei et al., 2018; Madry et al., 2017; Na et al., 2017) accepted to ICLR 2018 as non-certified white-box-secure defenses to adversarial examples. They find that seven of the nine defenses use obfuscated gradients, a kind of gradient masking, as a phenomenon that leads to a false sense of security in defenses

against adversarial examples.Obfuscated gradients provide a limited increase in robustness and can be broken by improved attack techniques they develop. The only defense they observe that significantly increases robustness to adversarial examples within the threat model proposed is **adversarial training**(Madry et al., 2017).

Adversarial training(Goodfellow et al., 2014; Madry et al., 2017; Kannan et al., 2018; Tramèr et al., 2017; Pang et al., 2019) defends against adversarial perturbations by training networks on adversarial images that are generated on-the-fly during training. For adversarial training, the most relevant work to our study is (Kannan et al., 2018), which introduce a technique they call **Adversarial Logit Pairing(ALP)**, a method that encourages logits for pairs of examples to be similar. (Engstrom et al., 2018; Mosbach et al., 2018) also put forward different opinions on the robustness of ALP. Our **ATLPA** encourages attention map for pairs of examples to be similar. When applied to clean examples and their adversarial counterparts, **ATLPA** improves accuracy on adversarial examples over adversarial training.(Araujo et al., 2019) adds random noise at training and inference time,(Xie et al., 2018) adds denoising blocks to the model to increase adversarial robustness, neither of the above approaches focuses on the attention map. Following (Pang et al., 2018; Yang et al., 2018; Pang et al., 2019), we propose **Tolerant Logit** which consists of confidence distribution on top-k classes and captures inter-class similarities at the image level.

In terms of methodologies, our work is also related to deep transfer learning and knowledge distillation problems, the most relevant work to our study are (Zagoruyko & Komodakis, 2016; Li et al., 2019b), which constrain the L_2-norm of the difference between their behaviors (i.e., the feature maps of outer layer outputs in the source/target networks). Our **ATLPA** constrains attention map for pairs of clean examples and their adversarial counterparts to be similar.

3 Definitions and Threat Models

In this paper, we always assume the attacker is capable of forming untargeted attacks that consist of perturbations of limited L_∞-norm. This is a simplified task chosen because it is more amenable to benchmark evaluations. We consider two different threat models characterizing amounts of information the adversary can have:

- **Gray-box Attack** We focus on defense against gray-box attacks in this paper. In a gray-back attack, the attacker knows both the original network and the defense algorithm. Only the parameters of the defense model are hidden from the attacker. This is also a standard setting assumed in many security systems and applications(Pfleeger & Pfleeger, 2004).
- **Black-box Attack** The attacker has no information about the models architecture or parameters, and no ability to send queries to the model to gather more information.

4 Methods

4.1 Architecture

Fig.2 represents architecture of **ATLPA** : a baseline model is adversarial trained so as, not only to make similar the output labels, but to also have similar **Tolerant Logits** and spatial attention maps to those of original images and adversarial images.

4.2 Adversarial training

We use adversarial training with **Projected Gradient Descent(PGD)**(Madry et al., 2017) as the underlying basis for our methods:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\in\hat{p}_{\text{data}}} \left(\max_{\delta\in S} L(\theta, x+\delta, y) \right) \tag{1}$$

where $\hat{p}_{\text{data}}$ is the underlying training data distribution, $L(\theta, x+\delta, y)$ is a loss function at data point x which has true class y for a model with parameters θ, and the maximization with respect to δ is approximated using PGD. In this paper, the loss is defined as:

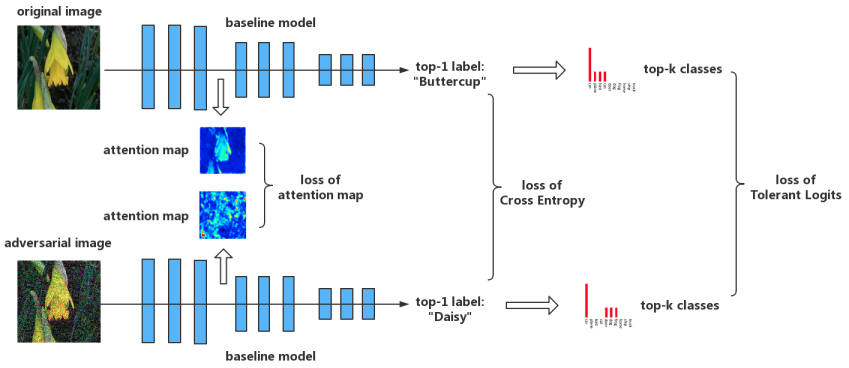


Figure 2: Schematic representation of **ATLPA**: a baseline model is adversarial trained so as, not only to make similar the output labels, but to also have similar **Tolerant Logits** and spatial attention maps to those of original images and adversarial images.

$$L = L_{CE} + \alpha L_{TL} + \beta L_{AT} \tag{2}$$

Where L_{CE} is cross entropy,α and β are hyper-parameters which balance **Tolerant Logit Loss** L_{TL} and **Attention Map Loss** L_{AT}. When β=0, we call it **ATLPA**(w/o ATT), i.e., **ATLPA** without attention.

4.3 Tolerant Logit Loss

Instead of computing an extra loss over all classes just like **ALP**(Kannan et al., 2018), we pick up a few classes which have been assigned with the highest confidence scores, and assume that these classes are more likely to be semantically similar to the input image. We use top-k classes of confidence distribution which capture inter-class similarities at the image level. The logit of model is $Z(x)$,f_{a_k} is short for the k-th largest element of $Z(x)$. Then we can define the following loss:

$$L_{TL} = max(f_{a_1} - \sum_{k=2}^{K} w_k \cdot f_{a_k}, 0) \tag{3}$$

where w_k is non-negative weight, used to adjust the influence of the k-th largest element of $Z(x)$. In the experiments we use $K = 5$.

4.4 Attention Map Loss

We use **Attention Map Loss** to encourage the attention map from clean examples and their adversarial counterparts to be similar to each other. Let also I denote the indices of all activation layer pairs for which we want to pay attention. Then we can define the following total loss:

$$L_{AT} = \sum_{j \in \mathcal{I}} \left\| \frac{Q^j_{ADV}}{\left\| Q^j_{ADV} \right\|_2} - \frac{Q^j_O}{\left\| Q^j_O \right\|_2} \right\|_p \tag{4}$$

Let O, ADV denote clean examples and their adversarial counterparts.where $Q^j_O = \text{vec}\left(F\left(A^j_O\right)\right)$ and $Q^j_{ADV} = \text{vec}\left(F\left(A^j_{ADV}\right)\right)$ are respectively the j-th pair of clean examples and their adversarial counterparts attention maps in vectorized form, and p refers to norm type (in the experiments we use $p = 2$). F sums absolute values of attention maps raised to the power of p.

5 EXPERIMENTS

5.1 GRAY AND BLACK-BOX SETTINGS

To evaluate the effectiveness of our defense strategy, we performed a series of image-classification experiments on **17 Flower Category Database**(Nilsback & Zisserman, 2006) and **BMW-10 Database**. Following (Athalye et al., 2018; Xie et al., 2018; Dubey et al., 2019), we assume an adversary that uses the state of the art PGD adversarial attack method.

We consider *untargeted* attacks when evaluating under the gray and black-box settings; *untargeted* attacks are also used in our adversarial training. We evaluate top-1 accuracy on validation images that are adversarially perturbed by the attacker. In this paper, adversarial perturbation is considered under L_∞ norm (i.e., maximum perturbation for each pixel), with an allowed maximum value of ϵ. The value of ϵ is relative to the pixel intensity scale of 256, we use $\epsilon = 64/256 = 0.25$ and $\epsilon = 128/256 = 0.5$. PGD attacker with 10 to 200 attack iterations and step size $\alpha = 1.0/256 = 0.0039$. Our baselines are ResNet-101/152. There are four groups of convolutional structures in the baseline model, which are described as $conv2\_x$, $conv3\_x$,$conv4\_x$ and $conv5\_x$ in (He et al., 2015)

5.2 IMAGE DATABASE

We performed a series of image-classification experiments on a wide range of datasets.Compared with data sets with very small image size e.g., MNIST is $28 * 28$,CIFAR-10 is $32 * 32$, the image size of our data sets is closer to the actual situation. All the images are resized to $256 * 256$ and normalized to zero mean for each channel, following with data augmentation operations of random mirror and random crop to $224 * 224$.

- **17 Flower Category Database**(Nilsback & Zisserman, 2006) contains images of flowers belonging to 17 different categories. The images were acquired by searching the web and taking pictures. There are 80 images for each category. We use only classification labels during training. While part location annotations are used in a quantitative evaluation of show cases, to explain the effect of our algorithm.
- **BMW-10 dataset**(Krause et al., 2013) contains 512 images of 10 BMW sedans. The data is split into 360 training images and 152 testing images, where each class has been split roughly in a 70-30 split.

5.3 EXPERIMENTAL SETUP

To perform image classification, we use ResNet-101/152 that were trained on our data sets. We consider two different attack settings: (1) a gray-box attack setting in which the model used to generate the adversarial images is the same as the image-classification model, viz. the ResNet-101; and (2) a black-box attack setting in which the adversarial images are generated using the ResNet-152 model; The backend prediction model of gray-box and black-box is ResNet-101 with different implementations of the state of the art defense methods,such as IGR(Ross & Doshi-Velez, 2017), PAT(Madry et al., 2017), RAT(Araujo et al., 2019), Randomization(Xie et al., 2017), ALP(Kannan et al., 2018), and FD(Xie et al., 2018). ALL the defence methods are all trained under the same adversarial training parameters: batch size is 16, iteration number is 6000, learning rate is 0.01, the ratio of original images and adversarial images is 1:1, under 2-iteration PGD attack, step size is 0.125.

Ensemble learning among different algorithms and models(Tramèr et al., 2017; Pang et al., 2019; Raff et al., 2019) is good idea,but here we only consider the use of one single algorithm and one single model. The hyper-parameters settings of the above algorithms use the default values provided in their papers.We will open source our code implementation if this paper is accepted.

5.4 RESULTS AND DISCUSSION

5.4.1 MAIN RESULTS

Here, we first present results with **ATLPA** on **17 Flower Category Database**. Compared with previous work, (Kannan et al., 2018) was evaluated under 10-iteration PGD attack and $\epsilon = 0.0625$,our

work are evaluated under highly challenging PGD attack: the maximum perturbation $\epsilon \in \{0.25, 0.5\}$ i.e. $L_\infty \in \{0.25, 0.5\}$ with 10 to 200 attack iterations.The bigger the value of ϵ, the bigger the disturbance, the more significant the adversarial image effect is. To our knowledge, such a strong attack has not been previously explored on a wide range of datasets.As shown in Fig.3 that **our ATLPA outperform the state-of-the-art in adversarial robustness against highly challenging gray-box and black-box PGD attacks**.

Table 1 shows **Main Result** of our work:under strong 200-iteration PGD gray-box and black-box attacks,**our ATLPA outperform the state-of-the-art in adversarial robustness on all these databases**. For example, under strong 200-iteration **PGD** gray-box and black-box attacks on **BMW-10 Database** where prior art has 35% and 36% accuracy, our method achieves **61%** and **62%**.

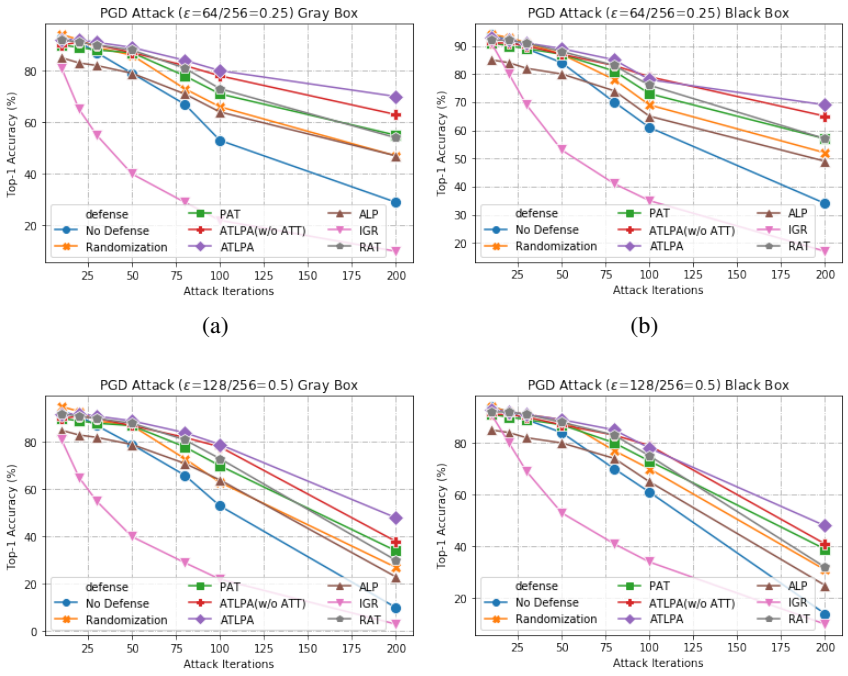


Figure 3: **Defense against gray-box and black-box attacks on 17 Flower Category Database**.(a)(c) shows results against a gray-box PGD attacker with 10 to 200 attack iterations.(b)(d) shows results against a black-box PGD attacker with 10 to 200 attack iterations.The maximum perturbation is $\epsilon \in \{0.25, 0.5\}$. Our **ATLPA**(purple line) **outperform the state-of-the-art in adversarial robustness** against highly challenging gray-box and black-box PGD attacks. Even our **ATLPA**(w/o ATT) does well, which is red line. **ATLPA**(w/o ATT): **ATLPA** without Attention.

5.4.2 THE AVERAGE ACTIVATIONS ON DISCRIMINATE PARTS

We visualized activation attention maps for defense against PGD attacks. Baseline model is ResNet-101(He et al., 2015), which is pre-trained on **ImageNet**(Russakovsky et al., 2015) and fine-tuned on **17 Flower Category Database**. We found from APPENDIX(Fig. 5) that has a higher level of activation on the whole flower,compared with other defence methods.

To further understand the effect, we compared average activations on discriminate parts of **17 Flower Category Database** for different defense methods. **17 Flower Category Database** defined discriminative parts of flowers. So for each image, we got several key regions which are very important to discriminate its category.Using all testing examples of **17 Flower Category Database**, we calculated normalized activations on these key regions of these different defense methods. As shown in Table 2, **ATLPA** got the highest average activations on those key regions, demonstrating that **ATLPA**

Table 1: **Defense against gray-box and black-box attacks** . The adversarial perturbation were produced using PGD with step size α=1.0/256=0.0039 and 200 attack iterations. As shown in this table, **ATLPA got the highest Top-1 Accuracy on all these database**. **ATLPA**(w/o ATT): **ATLPA** without Attention.

17 Flower Category Database	Gray-Box		Black-Box	
	L_∞=0.25	L_∞=0.5	L_∞=0.25	L_∞=0.5
No Defence	0	0	15	10
IGR(Ross & Doshi-Velez, 2017)	10	3	17	10
PAT(Madry et al., 2017)	55	34	57	39
RAT(Araujo et al., 2019)	54	30	57	32
Randomization(Xie et al., 2017)	12	6	27	16
ALP(Kannan et al., 2018)	47	23	49	25
FD(Xie et al., 2018)	33	10	33	10
Our **ATLPA**(w/o ATT)	63	38	65	41
Our **ATLPA**	**69**	**48**	**70**	**48**

BMW-10 Database	Gray-Box		Black-Box	
	L_∞=0.25	L_∞=0.5	L_∞=0.25	L_∞=0.5
No Defence	6	6	18	18
IGR(Ross & Doshi-Velez, 2017)	12	12	16	16
PAT(Madry et al., 2017)	36	36	35	35
RAT(Araujo et al., 2019)	17	16	20	20
Randomization(Xie et al., 2017)	11	10	15	14
ALP(Kannan et al., 2018)	16	16	20	20
FD(Xie et al., 2018)	20	20	22	22
Our **ATLPA**(w/o ATT)	58	58	55	55
Our **ATLPA**	**61**	**60**	**62**	**62**

focused on more discriminate features for flowers recognition. In addition, the score of **ATLPA** is more bigger than **ATLPA**(w/o ATT), so it can be seen that the main factor is our **Attention**.

Table 2: Comparing **average activations** on discriminate parts of **17 Flower Category Database** for different defense methods. **ATLPA** got the highest average activations on those key regions, demonstrating that **ATLPA** focused on more discriminate features for flowers recognition.

Defense	Gray-Box		Black-Box	
	L_∞=0.25	L_∞=0.5	L_∞=0.25	L_∞=0.5
No Defense	0.10	0.10	0.15	0.15
ALP(Kannan et al., 2018)	0.15	0.15	0.14	0.14
IGR(Ross & Doshi-Velez, 2017)	0.14	0.14	0.13	0.13
PAT(Madry et al., 2017)	0.17	0.17	0.15	0.15
RAT(Araujo et al., 2019)	0.22	0.22	0.21	0.21
Our **ATLPA**(w/o ATT)	0.31	0.31	0.27	0.27
Our **ATLPA**	**0.39**	**0.39**	**0.36**	**0.36**

5.4.3 THE DIVERSITY AMONG LEARNED FEATURES OF DIFFERENT CLASSES

Previous work has shown that for a single network, promoting the diversity among learned features of different classes can improve adversarial robustness(Pang et al., 2018; 2019). As shown in APPENDIX(Fig. 7),the **ATLPA** and **ATLPA**(w/o ATT) training procedure conceals normal examples on low-dimensional manifolds in the final-layer hidden space. Then the detector allowable regions can also be set low-dimensional as long as the regions contain all normal examples. Therefore the white-box adversaries who intend to fool our detector have to generate adversarial examples with preciser calculations and larger noises.

To further understand the effect,we compute **silhouette score**(Rousseeuw, 1999) of the final hidden features of different defense after t-SNE(Laurens & Hinton, 2008). The range of silhouette score is $[-1, 1]$. The closer the samples of the same category are, the farther the samples of different categories are, the higher the score is.We compute the silhouette score to quantify the quality of

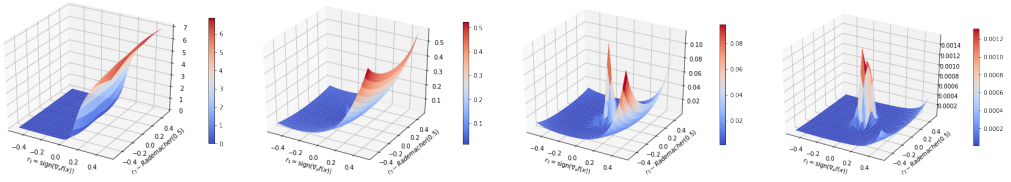
diversity among learned features of different classes.As shown in Table 3, **ATLPA** got the highest silhouette score, demonstrating that **ATLPA** promotes the diversity among learned features of different classes.In addition, the scores of **ATLPA** and **ATLPA**(w/o ATT) are very close, so it can be seen that the main factor is our **Tolerant Logit**.

Table 3: Comparing **silhouette score** on adversarial images of **17 Flower Category Database** testing set for different defense methods.

Defense	Gray-Box		Black-Box	
	L_∞=0.25	L_∞=0.5	L_∞=0.25	L_∞=0.5
No Defense	0.09	0.09	0.04	0.04
ALP(Kannan et al., 2018)	0.05	0.05	0.04	0.04
IGR(Ross & Doshi-Velez, 2017)	0.00	0.00	-0.03	-0.03
PAT(Madry et al., 2017)	0.13	0.13	0.12	0.13
RAT(Araujo et al., 2019)	0.13	0.13	0.12	0.12
Our **ATLPA**(w/o ATT)	0.16	0.16	0.15	0.15
Our **ATLPA**	**0.17**	**0.17**	**0.16**	**0.16**

5.4.4 THE TRENDS OF LOSS LANDSCAPES

We generate loss plots by varying the input to the models, starting from an original input image chosen from the testing set of **17 Flower Category Database**. The z axis represents the loss. If x is the original input, then we plot the loss varying along the space determined by two vectors: $r1 = sign(\bigtriangledown_x f(x))$ and $r2 \sim Rademacher(0.5)$. We thus plot the following function: $z = loss(x \cdot r1 + y \cdot r2)$. As shown in Fig. 4,the input varies in the same range and the landscape of our **ATLPA** varies in the smallest range, our **ATLPA** has better robustness.



(a) No defence (b) ALP (c) **ATLPA**(w/o ATT) (d) **ATLPA**

Figure 4: **Comparison of loss landscapes**. Loss plots are generated by varying the input to the models, starting from an original input image chosen from the testing set of **17 Flower Category Database**. **ATLPA**(w/o ATT): **ATLPA** without Attentions.

6 CONCLUSION

In this paper, we propose a novel regularized adversarial training framework **ATLPA** a method that uses **Tolerant Logit** which consists of confidence distribution on top-k classes and captures inter-class similarities at the image level, and encourages attention map for pairs of examples to be similar. We show that our **ATLPA** achieves **the state of the art** defense on a wide range of datasets against strong **PGD** gray-box and black-box attacks. We explain the reason why our **ATLPA** can improve the robustness of the model from three dimensions: **average activations on discriminate parts**, **the diversity among learned features of different classes** and **trends of loss landscapes**. The results of visualization and quantitative calculation show that our method is helpful to improve the robustness of the model.

ACKNOWLEDGMENTS

We would like to thank Anish Athalye for his helpful feedback on this paper. Furthermore, we thank the reviewers for their valuable comments.

REFERENCES

Alexandre Araujo, Rafael Pinot, Benjamin Negrevergne, Laurent Meunier, and Jamal Atif. Robust neural networks using randomized adversarial training. 2019.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018. URL https://arxiv.org/abs/1802.00420.

Avishek Joey Bose and Parham Aarabi. Adversarial attacks on face detectors using neural net based constrained optimization. 2018.

Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S18Su--CW.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. 2016.

Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

Abhimanyu Dubey, Van Der Maaten Laurens, Zeki Yalniz, Yixuan Li, and Dhruv Mahajan. Defense against adversarial images using web-scale nearest-neighbor search. 2019.

Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Computer Science*, 2014.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018. URL http://arxiv.org/abs/1803.06373.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, 2012.

Van Der Maaten Laurens and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2605):2579–2605, 2008.

Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5060–5069, 2019a.

Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. *arXiv preprint arXiv:1901.09229*, 2019b.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

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

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Marius Mosbach, Maksym Andriushchenko, Thomas Trost, Matthias Hein, and Dietrich Klakow. Logit pairing methods can fool gradient-based attacks. 2018.

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

M-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1447–1454, 2006.

Tianyu Pang, Du Chao, Yinpeng Dong, and Jun Zhu. Towards robust detection of adversarial examples. 2018.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. 2019.

Charles P. Pfleeger and Shari Lawrence Pfleeger. *Security in Computing*. 2004.

Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. 2017.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational & Applied Mathematics*, 20(20):53–65, 1999.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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

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.

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.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. 2017.

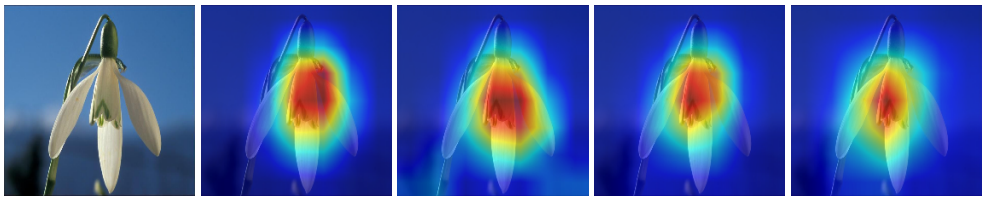
Cihang Xie, Yuxin Wu, Laurens Van Der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. 2018.

Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan Yuille. Knowledge distillation in generations: More tolerant teachers educate better students. *arXiv preprint arXiv:1805.05551*, 2018.

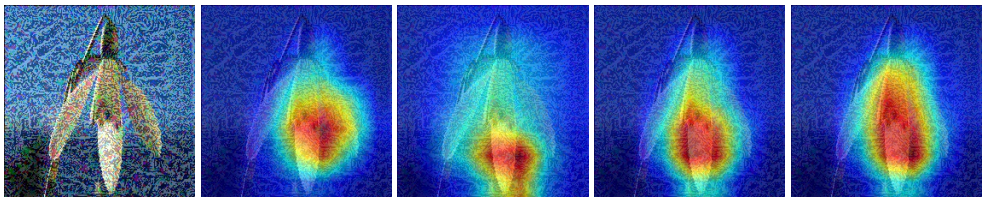
Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. 2016.

A APPENDIX

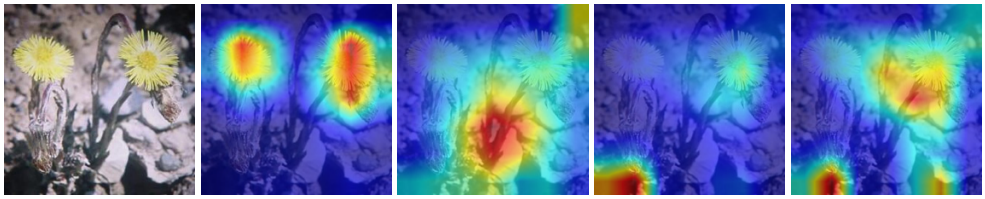
17 Flower Category Database



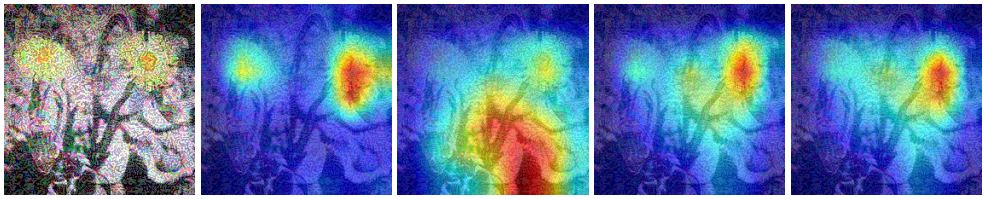
(a) Original (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**



(a) Adversarial (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**



(a) Original (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**



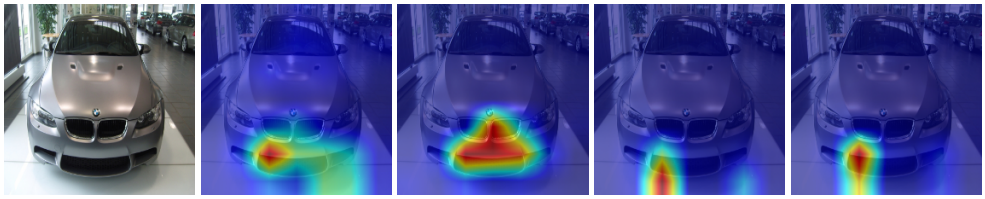
(a) Adversarial (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**

Figure 5: Illustration of the effect of the attention mechanism of **ATLPA**. Using attention, however, changes the activation map significantly. **ATLPA** show obviously improved concentration. With attention (the right-most column), we observed a large set of pixels that have high activation at important regions e.g. the whole petal. **ATLPA** w/o ATT: **ATLPA** without Attentions.

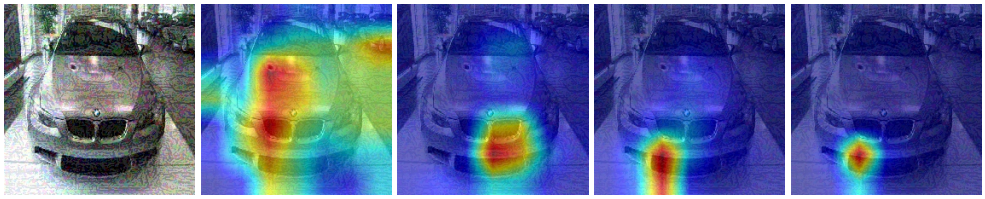
BMW-10 Database

(a) Original (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**

(a) Adversarial (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**



(a) Original (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**



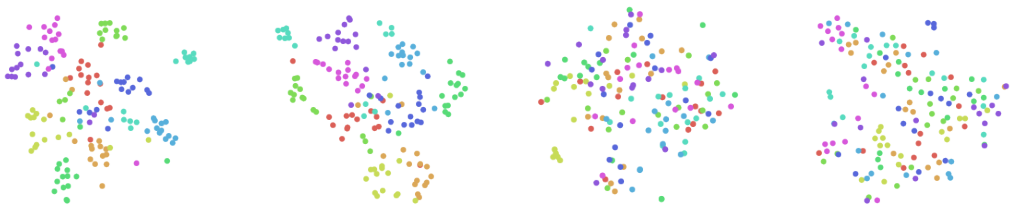
(a) Adversarial (b) No defence (c) ALP (d) **ATLPA** w/o ATT (e) **ATLPA**

Figure 6: Illustration of the effect of the attention mechanism of **ATLPA**.Using attention, however, changes the activation map significantly. **ATLPA** show obviously improved concentration. With attention (the right-most column), we observed a large set of pixels that have high activation at important regions e.g. front bumper of car .**ATLPA** w/o ATT: **ATLPA** without Attentions.

17 Flower Category Database

(a) **ATLPA**(w/o ATT) (b) **ATLPA** (c) ALP (d) No defence

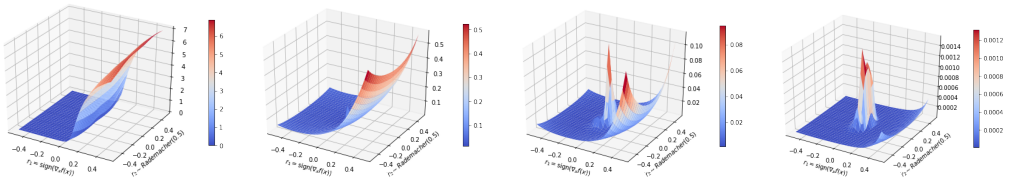
BMW-10 Database



(a) **ATLPA**(w/o ATT) (b) **ATLPA** (c) ALP (d) No defence

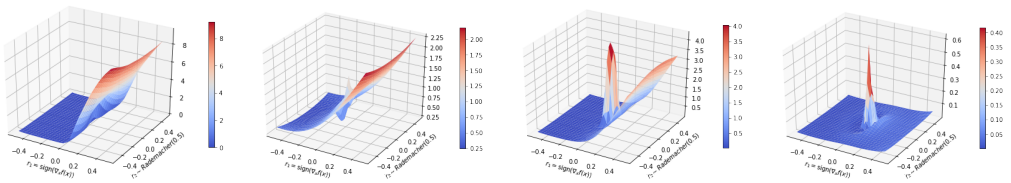
Figure 7: t-SNE visualization results on the final hidden features of different defense methods. The inputs are adversarial images of **17 Flower Category Database** and **BMW-10 Database** testing set.**ATLPA**(w/o ATT): **ATLPA** without Attention.

17 Flower Category Database #0083



(a) No defence (b) ALP (c) **ATLPA**(w/o ATT) (d) **ATLPA**

17 Flower Category Database #1163



(a) No defence (b) ALP (c) **ATLPA**(w/o ATT) (d) **ATLPA**

Figure 8: **Comparison of loss landscapes**. Loss plots are generated by varying the input to the models, starting from an original input image chosen from the testing set of **17 Flower Category Database**. The z axis represents the loss. If x is the original input, then we plot the loss varying along the space determined by two vectors: $r1 = sign(\bigtriangledown_x f(x))$ and $r2 \sim Rademacher(0.5)$. We thus plot the following function: $z = loss(x \cdot r1 + y \cdot r2)$. We see that the input varies in the same range and the landscape of our **ATLPA** varies in the smallest range,our **ATLPA** has better robustness. **ATLPA**(w/o ATT): **ATLPA** without Attentions.