ROBUST LOCAL FEATURES FOR IMPROVING THE GENERALIZATION OF ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training has been demonstrated as one of the most effective methods for training robust models so as to defend against adversarial examples. However, adversarially trained models often lack adversarially robust generalization on unseen testing data. Recent works show that adversarially trained models are more biased towards global structure features. Instead, in this work, we would like to investigate the relationship between the generalization of adversarial training and the robust local features, as the robust local features generalize well for unseen shape variation. To learn the robust local features, we develop a Random Block Shuffle (RBS) transformation to break up the global structure features on normal adversarial examples. We continue to propose a new approach called Robust Local Features for Adversarial Training (RLFAT), which first learns the robust local features by adversarial training on the RBS-transformed adversarial examples, and then transfers the robust local features into the training of normal adversarial examples. To demonstrate the generality of our argument, we implement RLFAT in two currently state-of-the-art adversarial training frameworks. Extensive experiments on STL-10, CIFAR-10, CIFAR-100 datasets show that RLFAT significantly improves both the adversarially robust generalization and the standard generalization of adversarial training. Additionally, we demonstrate that our models capture more local features of the object on the images, aligning better with human perception.

1 INTRODUCTION

Deep learning has achieved a remarkable performance breakthrough on various challenging benchmarks in machine learning fields, such as image classification (Krizhevsky et al., 2012) and speech recognition (Hinton et al., 2012). However, recent studies (Szegedy et al., 2014; Goodfellow et al., 2015) have revealed that deep neural network models are strikingly susceptible to *adversarial examples*, in which small perturbations around the input are sufficient to mislead the predictions of the target model. Moreover, such perturbations are almost imperceptible to humans and often transfer across diverse models to achieve black-box attacks (Papernot et al., 2017; Liu et al., 2017).

Though the emergence of adversarial examples has received significant attention and led to various defend approaches for developing robust models (Madry et al., 2018; Dhillon et al., 2018; Wang & Yu, 2019; Song et al., 2019; Zhang et al., 2019a), many proposed defense methods provide few benefits for the true robustness but mask the gradients on which most attacks rely (Carlini & Wagner, 2017a; Athalye et al., 2018; Uesato et al., 2018; Li et al., 2019). Currently, one of the best techniques to defend against adversarial attacks (Athalye et al., 2018; Li et al., 2019) is *adversarial training* (Madry et al., 2018; Zhang et al., 2019a), which improves the adversarial robustness by injecting adversarial examples into the training data.

Among substantial works of adversarial training, there still remains a big robust generalization gap between the training data and the testing data (Schmidt et al., 2018; Zhang et al., 2019b; Ding et al., 2019). The robustness of adversarial training fails to generalize on unseen testing data. Recent works (Geirhos et al., 2019; Zhang & Zhu, 2019) further show that adversarially trained models capture more on global structure features but normally trained models are more biased towards local features. In intuition, global structure features tend to be robust against adversarial perturbations but hard to generalize for unseen shape variations, instead, local features generalize well for un-

seen shape variations but are hard to generalize on adversarial perturbation. It naturally raises an intriguing question for adversarial training:

For adversarial training, is it possible to learn the robust local features , which have better adversarially robust generalization and better standard generalization?

To address this question, we investigate the relationship between the generalization of adversarial training and the robust local features, and advocate for learning robust local features for adversarial training. Our main contributions are as follows:

- To our knowledge, this is the first work that sheds light on the relationship between adversarial training and robust local features. Specifically, we develop a Random Block Shuffle (RBS) transformation to study such relationship by breaking up the global structure features on normal adversarial examples.
- We propose a novel method called Robust Local Features for Adversarial Training (RLFAT), which first learns the robust local features, and then transfers the information of robust local features into the training on normal adversarial examples.
- To demonstrate the generality of our argument, we implement RLFAT in two currently state-of-the-art adversarial training frameworks, PGD Adversarial Training (PGDAT) (Madry et al., 2018) and TRADES (Zhang et al., 2019a). Empirical results show consistent and substantial improvements for both adversarial robustness and standard accuracy on several standard datasets. Moreover, the salience maps of our models on images tend to align better with human perception.

2 Preliminaries

In this section, we introduce some notations and provide a brief description on current advanced methods for adversarial attacks and adversarial training.

2.1 Notation

Let $F(x)$ be a probabilistic classifier based on a neural network with the logits function $f(x)$ and the probability distribution $p_F(\cdot|x)$. Let $\mathcal{L}(F; x, y)$ be the cross entropy loss for image classification. The goal of the adversaries is to find an adversarial example $x' \in \mathcal{B}_\epsilon^p(x) := \{x' : \|x' - x\|_p \leq \epsilon\}$ in the ℓ_p norm bounded perturbations, where ϵ denotes the magnitude of the perturbations. In this paper, we focus on $p = \infty$ to align with previous works.

2.2 Adversarial Attacks

Projected Gradient Descent. Projected Gradient Descent (PGD) (Madry et al., 2018) is a stronger iterative variant of Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015), which iteratively solves the optimization problem $\max_{x':\|x'-x\|_\infty<\epsilon} \mathcal{L}(F; x', y)$ with a step size α:

$$\begin{aligned} x^0 &\sim \mathcal{U}\left(\mathcal{B}_\epsilon^\infty(x)\right), \\ x^{t+1} &= \Pi_{\mathcal{B}_\epsilon^\infty(x)}\left(x^t - \alpha\, \mathrm{sign}\left(\nabla_x \mathcal{L}(F; x, y)|_{x^t}\right)\right), \end{aligned} \tag{1}$$

where $\mathcal{U}$ denotes the uniform distribution, and $\Pi_{\mathcal{B}_\epsilon^\infty(x)}$ indicates the projection of the set $\mathcal{B}_\epsilon^\infty(x)$.

Carlini-Wagner attack. Carlini-Wagner attack (CW) (2017b) is a sophisticated method to directly solve for the adversarial example x^{adv} by using an auxiliary variable w:

$$x^{adv} = 0.5 \cdot (\tanh(w) + 1). \tag{2}$$

The objective function to optimize the auxiliary variable w is defined as:

$$\min_w \left\|x^{adv} - x\right\| + c \cdot \mathcal{F}\left(x^{adv}\right), \tag{3}$$

where $\mathcal{F}(x^{adv}) = \max\left(f_{y^{\text{true}}}(x^{adv}) - \max\left\{f_i(x^{adv}) : i \neq y^{\text{true}}\right\}, -k\right)$. The constant k controls the confidence gap between the adversarial class and the true class.

$\mathcal{N}$attack. $\mathcal{N}$attack (Li et al., 2019) is a derivative-free black-box adversarial attack and it breaks many of the defense methods based on gradient masking. The basic idea is to learn a probability density distribution over a small region centered around the clean input, such that a sample drawn from this distribution is likely to be an adversarial example.

2.3 ADVERSARIAL TRAINING

Despite a wide range of defense methods, Athalye et al. (2018) and Li et al. (2019) have broken most previous defense methods (Dhillon et al., 2018; Buckman et al., 2018; Wang & Yu, 2019; Zhang et al., 2019a), and revealed that adversarial training remains one of the best defense method. The basic idea of adversarial training is to solve the min-max optimization problem, as shown in Eq. (4):

$$\min_{F} \max_{x':\|x'-x\|_\infty<\epsilon} \mathcal{L}\left(F; x', y\right). \tag{4}$$

Here we introduce two currently state-of-the-art adversarial training frameworks.

PGD adversarial training. PGD Adversarial Training (PGDAT) (Madry et al., 2018) leverages the PGD attack to generate adversarial examples, and trains only with the adversarial examples. The objective function is formalized as follows:

$$\mathcal{L}_{\mathrm{PGD}}(F; x, y) = \mathcal{L}(F; x'_{\mathrm{PGD}}, y)\ , \tag{5}$$

where x'_{PGD} is obtained via the PGD attack on the cross entropy $\mathcal{L}(F; x, y)$.

TRADES. Zhang et al. (2019a) propose TRADES to specifically maximize the trade-off of adversarial training between adversarial robustness and standard accuracy by optimizing the following regularized surrogate loss:

$$\mathcal{L}_{\mathrm{TRADES}}(F; x, y) = \mathcal{L}(F; x, y) + \lambda D_{\mathrm{KL}}\left(\ p_F(\cdot|x)\ \|\ p_F\left(\cdot|x'_{\mathrm{PGD}}[x]\right)\ \right)\ , \tag{6}$$

where $x'_{\mathrm{PGD}}[x]$ is obtained via the PGD attack on the KL-divergence $D_{\mathrm{KL}}\left(\ p_F(\cdot|x)\ \|\ p_F\left(\cdot|x'\right)\ \right)$, and λ is a hyper-parameter to control the trade-off between adversarial robustness and standard accuracy.

3 ROBUST LOCAL FEATURES FOR ADVERSARIAL TRAINING

Unlike adversarially trained models, normally trained models are more biased towards the local features but vulnerable to adversarial examples (Geirhos et al., 2019). It indicates that, in contrast to global structural features, local features seems be more well-generalized but less robust against adversarial perturbation. Based on the basic observation, in this work, we focus on the learning of robust local features on adversarial training, and propose a novel form of adversarial training called RLFAT that learns the robust local features and transfers the robust local features into the training of normal adversarial examples. In this way, our adversarially trained models not only yield strong robustness against adversarial examples but also show great generalization on unseen testing data.

3.1 ROBUST LOCAL FEATURE LEARNING

It's known that adversarial training tends to capture global structure features so as to increase invariance against adversarial perturbations (Zhang & Zhu, 2019; Ilyas et al., 2019). To advocate for the learning of robust local features during adversarial training, we propose a simple and straightforward image transformation called Random Block Shuffle (RBS) to break up the global structure features of the images, at the same time retaining the local features. Specifically, for an input image, we randomly split the target image into k blocks horizontally and randomly shuffle the blocks, and then we perform the same split-shuffle operation vertically on the resulting image. As illustrated in Figure 1, RBS transformation can destroy the global structure features of the images to some extent and retain the local features of the images.

Then we apply the RBS transformation on adversarial training. Different from normal adversarial training, we use the RBS-transformed adversarial examples rather than normal adversarial examples as the adversarial information to encourage the models to learn robust local features. Note that we only use the RBS transformation as a tool to learn the robust local features during adversarial training and will not use RBS transformation in the inference phase. we refer to the form of adversarial training as *RBS Adversarial Training* (RBSAT).

To demonstrate the generality of our argument, we consider two currently state-of-the-art adversarial training frameworks, PGD Adversarial Training (PGDAT) (Madry et al., 2018) and TRADES (Zhang et al., 2019a), to demonstrate the effectiveness of the robust local features.

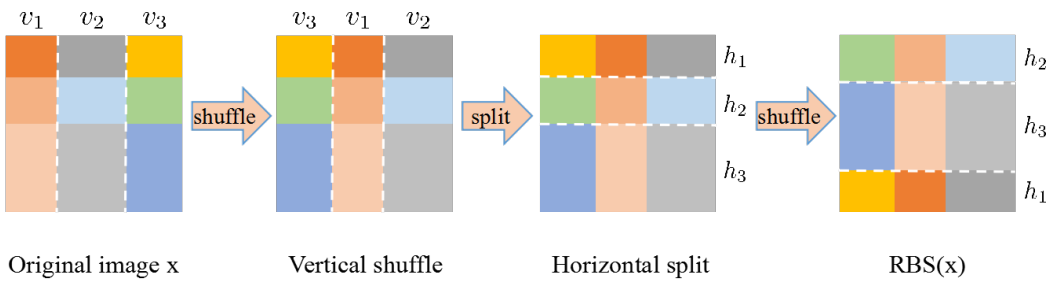

Figure 1: Illustration of the RBS transformation for $k = 3$. For a better understanding on the RBS transformation, we paint the split image blocks with different colors.

We use the following loss function as the alternative to the objective function of PGDAT:

$$\mathcal{L}^{\mathrm{RLFL}}_{\mathrm{PGDAT}}(F; x, y) = \mathcal{L}(F; \mathrm{RBS}(x'_{\mathrm{PGD}}), y) \ , \tag{7}$$

where $\mathrm{RBS}(\cdot)$ denotes the RBS transformation; x'_{PGD} is obtained via the PGD attack on the cross entropy $\mathcal{L}(F; x, y)$.

Similarly, we use the following loss function as the alternative to the objective function of TRADES:

$$\mathcal{L}^{\mathrm{RLFL}}_{\mathrm{TRADES}}(F; x, y) = \mathcal{L}(F; x, y) + \lambda D_{\mathrm{KL}} \left[\, p_F(\cdot|x) \parallel p_F \left(\cdot|\mathrm{RBS}\left(x'_{\mathrm{PGD}}[x]\right)\right) \, \right] \ , \tag{8}$$

where $x'_{\mathrm{PGD}}[x]$ is obtained via the PGD attack on the KL-divergence $D_{\mathrm{KL}} \left(\, p_F(\cdot|x) \parallel p_F \left(\cdot|x'\right) \, \right)$.

3.2 Robust Local Feature Transfer

Since the type of input images in the training phase and the inference phase is different (RBS transformed images for training, versus original images for inference), we consider to transfer the knowledge of the robust local features learned by RBSAT to the normal adversarial examples. Specifically, we present a knowledge transfer scheme, called *Robust Local Feature Transfer (RLFT)*. The goal of RLFT is to learn the representation that minimizes the feature shift between the normal adversarial examples and the RBS-transformed adversarial examples.

In particular, we apply RLFT on the logit layer for high-level feature alignment. Formally, the objective functions of robust local feature transfer for PGDAT and TRADES are formalized as follows, respectively:

$$\begin{aligned} \mathcal{L}^{\mathrm{RLFT}}_{\mathrm{PGDAT}}(F; x, y) &= \| f(\mathrm{RBS}(x'_{\mathrm{PGD}})) - f(x'_{\mathrm{PGD}}) \|_2^2 \, , \\ \mathcal{L}^{\mathrm{RLFT}}_{\mathrm{TRADES}}(F; x, y) &= \| f(\mathrm{RBS}(x'_{\mathrm{PGD}}[x])) - f(x'_{\mathrm{PGD}}[x]) \|_2^2 \, , \end{aligned} \tag{9}$$

where $f(\cdot)$ denotes the mapping of the logit layer, and $\|\cdot\|_2^2$ denotes the squared Euclidean norm.

3.3 Overall Objective Function

Since the quality of robust local feature transfer depends on the quality of the robust local features learned by RBSAT, we integrate RBSAT and RLFT into an end-to-end training framework, which we refer to as *RLFAT* (Robust Local Features for Adversarial Training). The general training process of RLFAT is summarized in Algorithm 1. Note that the computational cost of RBS transformation (line 7) is negligible in the total computational cost.

We implement RLFAT in two currently state-of-the-art adversarial training frameworks, PGDAT and TRADES, and have new objective functions to learn the robust and well-generalized feature representations, which we call $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$:

$$\begin{aligned} \mathcal{L}_{\mathrm{RLFAT_P}}(F; x, y) &= \mathcal{L}^{\mathrm{RLFL}}_{\mathrm{PGDAT}}(F; x, y) + \eta \mathcal{L}^{\mathrm{RLFT}}_{\mathrm{PGDAT}}(F; x, y), \\ \mathcal{L}_{\mathrm{RLFAT_T}}(F; x, y) &= \mathcal{L}^{\mathrm{RLFL}}_{\mathrm{TRADES}}(F; x, y) + \eta \mathcal{L}^{\mathrm{RLFT}}_{\mathrm{TRADES}}(F; x, y), \end{aligned} \tag{10}$$

where η is a hyper-parameter to balance the two terms.

Algorithm 1 Robust Local Features for Adversarial Training (RLFAT).

Randomly initialize network $F(x)$;
Number of iterations $t \leftarrow 0$;
repeat
  $t \leftarrow t + 1$;
  Read a minibatch of data $\{x_1, ..., x_m\}$ from the training set;
  Generate the normal adversarial examples $\{x_1^{adv}, ..., x_m^{adv}\}$
  Obtain the RBS-transformed adversarial examples $\{\mathrm{RBS}(x_1^{adv}), ..., \mathrm{RBS}(x_m^{adv})\}$;
  Calculate the overall loss following Eq. (10).
  Update the parameters of network F through back propagation;
until the training converges.

4 EXPERIMENTS

In this section, to validate the effectiveness of RLFAT, we empirically evaluate our two implementations, denoted as $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$, and show that our models make significant improvement on both robust accuracy and standard accuracy on standard benchmark datasets, which provides strong support for our main hypothesis. Codes are available online[1].

4.1 EXPERIMENTAL SETUP

Baselines. Since most previous defense methods provide few benefit in true adversarially robustness (Athalye et al., 2018; Li et al., 2019), we compare the proposed methods with state-of-the-art adversarial training defenses, PGD Adversarial Training (PGDAT) (Madry et al., 2018) and TRADES (Zhang et al., 2019a).

Adversarial setting. We consider two attack settings with the bounded ℓ_∞ norm: the white-box attack setting and the black-box attack setting. For the white-box attack setting, we consider existing strongest white-box attacks: Projected Gradient Descent (PGD) (Madry et al., 2018) and Carlini-Wagner attack (CW) (Carlini & Wagner, 2017b). For the black-box attack setting, we perform the powerful black-box attack, $\mathcal{N}$attack (Li et al., 2019), on a sample of 1,500 test inputs as it is time-consuming.

Datasets. We compare the proposed methods with the baselines on widely used benchmark datasets, namely CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009). Since adversarially robust generalization becomes increasingly hard for high dimensional data and a little training data (Schmidt et al., 2018), we also consider one challenging dataset: STL-10 (Coates et al.), which contains $5,000$ training images, with 96×96 pixels per image.

Neural networks. For STL-10, the architecture we use is a wide ResNet 40-2 (Zagoruyko & Komodakis, 2016). For CIFAR-10 and CIFAR-100, we use a wide ResNet w32-10. For all datasets, we scale the input images to the range of $[0, 1]$.

Hyper-parameters. To avoid posting much concentrate on optimizing the hyper-parameters, for all datasets, we set the hyper-parameter λ in TRADES as 6, set the hyper-parameter η in $\mathrm{RLFAT_P}$ as 0.5, and set the hyper-parameter η in $\mathrm{RLFAT_T}$ as 1. For the training jobs of all our models, we set the hyper-parameters k of the RBS transformation as 2. More details about the hyper-parameters are provided in Appendix A.

4.2 EVALUATION RESULTS

We first validate our main hypothesis: for adversarial training, is it possible to learn the robust local features that have better adversarially robust generalization and better standard generalization?

In Table 1, we compare the accuracy of $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$ with the competing baselines on three standard datasets. The proposed methods lead to consistent and significant improvements on adversarial robustness as well as standard accuracy over the baseline models on all datasets.

[1] https://drive.google.com/drive/folders/183Sb5q_RQbzeZkw-uQbpBd7S1yTnn-Au?usp=sharing

With the robust local features, $\mathrm{RLFAT_T}$ achieves better adversarially robust generalization and better standard generalization than TRADES. $\mathrm{RLFAT_P}$ also works similarly, showing a significant improvement on the robustness against all attacks and standard accuracy than PGDAT.

Table 1: The classification accuracy (%) of defense methods under white-box and black-box attacks on STL-10, CIFAR-10 and CIFAR-100.

(a) **STL-10.** The magnitude of perturbation is 0.03 in ℓ_∞ norm.

Defense	No attack	PGD	CW	$\mathcal{N}$attack
PGDAT	67.05	30.00	31.97	34.80
TRADES	65.24	38.99	38.35	42.07
$\mathrm{RLFAT_P}$	71.47	38.42	38.42	44.80
$\mathrm{RLFAT_T}$	**72.38**	**43.36**	**39.31**	**48.13**

(b) **CIFAR-10.** The magnitude of perturbation is 0.03 in ℓ_∞ norm.

Defense	No attack	PGD	CW	$\mathcal{N}$attack
PGDAT	82.96	46.19	46.41	46.67
TRADES	80.35	50.95	49.80	52.47
$\mathrm{RLFAT_P}$	**84.77**	53.97	**52.40**	**54.60**
$\mathrm{RLFAT_T}$	82.72	**58.75**	51.94	**54.60**

(c) **CIFAR-100.** The magnitude of perturbation is 0.03 in ℓ_∞ norm.

Defense	No attack	PGD	CW	$\mathcal{N}$attack
PGDAT	55.86	23.32	22.87	22.47
TRADES	52.13	27.26	24.66	25.13
$\mathrm{RLFAT_P}$	56.70	**31.99**	**29.04**	**32.53**
$\mathrm{RLFAT_T}$	**58.96**	31.63	27.54	30.86

The results demonstrate that, the robust local features can significantly improve both the adversarially robust generalization and the standard generalization over the state-of-the-art adversarial training frameworks, and strongly support our hypothesis. That is, for adversarial training, it is possible to learn the robust local features, which have better robust and standard generalization.

4.3 Loss Sensitivity under Distribution Shift

Motivation. Ding et al. (2019) and Zhang et al. (2019b) found that the effectiveness of adversarial training is highly sensitive to the "semantic-loss" shift of the test data distribution, such as gamma mapping. To further investigate the performance of the proposed methods, we quantify the smoothness of the models under the distribution shifts of brightness perturbation and gamma mapping.

Loss sensitivity on brightness perturbation. To quantify the smoothness of models on the shift of the brightness perturbation, we propose to estimate the Lipschitz continuity constant $\ell_{\mathcal{F}}$ by using the gradients of the loss function with respect to the brightness perturbation of the testing data. We adjust the brightness factor of images in the HSV (hue, saturation, value) color space, which we refer to as $x^b = \mathcal{V}(x, \alpha)$, where α denotes the magnitude of the brightness adjustment. The lower the value of $\ell^b_{\mathcal{F}}(\alpha)$ is, the smoother the loss function of the model is:

$$\ell^b_{\mathcal{F}}(\alpha) = \frac{1}{m} \sum_{i=1}^{m} \left\| \nabla_x \mathcal{L}(F; \mathcal{V}(x_i, \alpha), y_{true}) \right\|_2 \tag{11}$$

Loss sensitivity on gamma mapping. Gamma mapping (Szeliski, 2011) is a nonlinear element-wise operation used to adjust the exposure of images by applying $\tilde{x}^{(\gamma)} = x^\gamma$ on the original image x. Similarly, we approximate the loss sensitivity under gamma mapping, by using the gradients of the loss function with respect to the gamma mapping of the testing data. A smaller value indicates a smoother loss function.

$$\ell^g_{\mathcal{F}}(\gamma) = \frac{1}{m} \sum_{i=1}^{m} \left\| \nabla_x \mathcal{L}(F; x_i^\gamma, y_{true}) \right\|_2 \tag{12}$$

Sensitivity analysis. The results for the loss sensitivity of the adversarially trained models under brightness perturbation are reported in Table 2a, where we adopt various magnitude of brightness adjustment on each testing data. In Table 2b, we report the loss sensitivity of adversarially trained models under various gamma mappings. We observe that $\mathrm{RLFAT_T}$ provides the smoothest model under the distribution shifts on all the three datasets. The results suggest that, as compared to PGDAT and TRADES, both $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$ show lower gradients of the models on different data distributions, which we can directly attribute to the robust local features.

Table 2: The loss sensitivity of defense methods under different testing data distributions.

(a) Loss sensitivity on brightness perturbation for the adversarially trained models.

Dataset	$\ell^b_{\mathcal{F}}(-0.15)$ / $\ell^b_{\mathcal{F}}(0.15)$			
	PGDAT	TRADES	$\mathrm{RLFAT_P}$	$\mathrm{RLFAT_T}$
STL-10	0.85 / 0.82	0.46 / 0.48	0.32 / 0.34	**0.22 / 0.23**
CIFAR-10	1.41 / 1.36	0.88 / 0.90	0.70 / 0.71	**0.54 / 0.56**
CIFAR-100	3.93 / 3.66	2.31 / 2.12	1.25 / 1.31	**1.00 / 0.98**

(b) Loss sensitivity on gamma mapping for the adversarially trained models.

Dataset	$\ell^g_{\mathcal{F}}(0.8)$ / $\ell^g_{\mathcal{F}}(1.2)$			
	PGDAT	TRADES	$\mathrm{RLFAT_P}$	$\mathrm{RLFAT_T}$
STL-10	0.77 / 0.79	0.44 / 0.42	0.30 / 0.29	**0.21 / 0.19**
CIFAR-10	1.27 / 1.20	0.84 / 0.76	0.69 / 0.62	**0.54 / 0.48**
CIFAR-100	2.82 / 2.80	1.78 / 1.76	1.09 / 1.01	**0.95 / 0.88**

4.4 Ablation Studies

To further gain insights on the performance obtained by the robust local features, we perform ablation studies to dissect the impact of various components (robust local feature learning and robust local feature transfer). As shown in Figure 2, we conduct additional experiments for the ablation studies of $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$ on STL-10, CIFAR-10 and CIFAR-100, where we report the standard accuracy over the clean data and the *average* robust accuracy over all the attacks for each model.

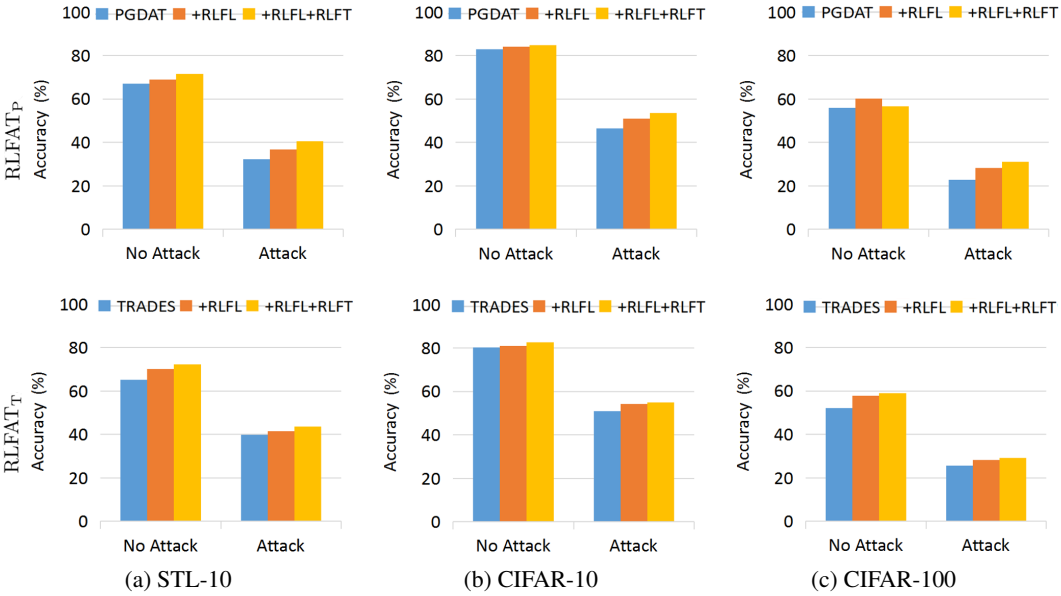

Figure 2: Ablation studies for $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$ to investigate the impact of Robust Local Feature Learning (RLFL) and Robust Local Feature Transfer (RLFT).

Does robust local feature learning help? We first analyze that as compared to adversarial training on normal adversarial examples, whether adversarial training on RBS-transformed adversarial examples produces better generalization and more robust features. As shown in Figure 2, we observe that Robust Local Features Learning (RLFL) exhibits stable improvements on both standard accuracy and robust accuracy on all datasets for $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$, providing strong support for our hypothesis.

Does robust local feature transfer help? We further add Robust Local Feature Transfer (RLFT), the second term in Eq. (10), to get the overall loss of RLFAT. The robust accuracy further increases on all datasets for $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$. The standard accuracy further increases also, except for $\mathrm{RLFAT_P}$ on CIFAR-100, but it is still clearly higher than the baseline model PGDAT. It indicates that transferring the robust local features into the training of normal adversarial examples does help promote the standard accuracy and robust accuracy in most cases.

4.5 Visualizing the Salience Maps

We would like to investigate the features of the input images that the models are mostly focused on. Following the work of Zhang & Zhu (2019), we generate the salience maps using *SmoothGrad* (Smilkov et al., 2017) on STL-10 dataset. The key idea of SmoothGrad is to average the gradients of class activation with respect to noisy copies of an input image. As illustrated in Figure 3, all the adversarially trained models basically capture the global structure features of the object on the images. As compared to PGDAT and TRADES, both $\mathrm{RLFAT_P}$ and $\mathrm{RLFAT_T}$ capture more local feature information of the object, aligning better with human perception. Note that the images are correctly classified by all these models. For more visualization results, see Appendix B.

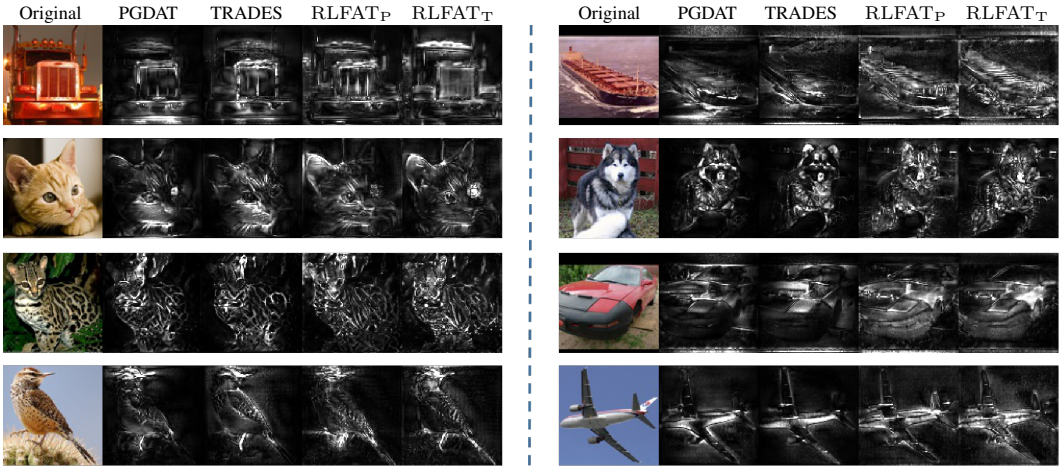

Figure 3: Salience maps of the four models on sampled images. For each group of images, we have the original image, and the salience maps of the four models sequentially.

5 Conclusion

Differs to existing adversarially trained models that are more biased towards the global structure features of the images, in this work, we hypothesize that robust local features can improve the generalization of adversarial training. To validate this hypothesis, we propose a new stream of adversarial training approach called Robust Local Features for Adversarial Training (RLFAT) and implement it in currently state-of-the-art adversarial training frameworks, PGDAT and TRADES. We provide strong empirical support for our hypothesis and show that the proposed methods based on RLFAT not only yield better standard generalization but also promote the adversarially robust generalization. Our findings open a new avenue for improving adversarial training. Furthermore, we show that the salience maps of our models on images tend to align better with human perception, uncovering certain unexpected benefit of the robust local features for adversarial training.

REFERENCES

Anish Athalye, Nicholas Carlini, and David A. 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*, pp. 274–283, 2018.

Jacob Buckman, Aurko Roy, Colin Raffel, and Ian J. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017*, pp. 3–14, 2017a.

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

Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011*.

Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang. On the sensitivity of adversarial robustness to input data distributions. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *CoRR*, abs/1905.02175, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, pp. 1106–1114, 2012.

Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 3866–3876, 2019.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017*, pp. 506–519, 2017.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pp. 5019–5031, 2018.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Richard Szeliski. *Computer Vision - Algorithms and Applications*. Texts in Computer Science. Springer, 2011.

Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aäron van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pp. 5032–5041, 2018.

Huaxia Wang and Chun-Nam Yu. A direct approach to robust deep learning using adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC*, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 7472–7482, 2019a.

Huan Zhang, Hongge Chen, Zhao Song, Duane S. Boning, Inderjit S. Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. In *7th International Conference on Learning Representations, ICLR 2019*, 2019b.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 7502–7511, 2019.

A HYPER-PARAMETER SETTING

Here we show the details of the training hyper-parameters and the attack hyper-parameters for the experiments.

Training Hyper-parameters. For all training jobs, we use the Adam optimizer with a learning rate of 0.001 and a batch size of 32. For CIFAR-10 and CIFAR-100, we run 79,800 steps for training. For STL-10, we run 29,700 steps for training. For STL-10 and CIFAR-100, the adversarial examples are generated with step size 0.0075, 7 iterations, and $\epsilon = 0.03$. For CIFAR-10, the adversarial examples are generated with step size 0.0075, 10 iterations, and $\epsilon = 0.03$.

Attack Hyper-parameters. For the PGD attack, we use the same attack parameters as those of the training process. For the CW attack, we use PGD to minimize its loss function with a high confidence parameter ($k = 50$) following the work of Madry et al. (2018). For the $\mathcal{N}$attack, we set the maximum number of optimization iterations to $T = 200$, $b = 300$ for the sample size, the variance of the isotropic Gaussian $\sigma^2 = 0.01$, and the learning rate $\eta = 0.008$.

B MORE FEATURE VISUALIZATION

We provide more salience maps of the adversarially trained models on sampled images in Figure 4.

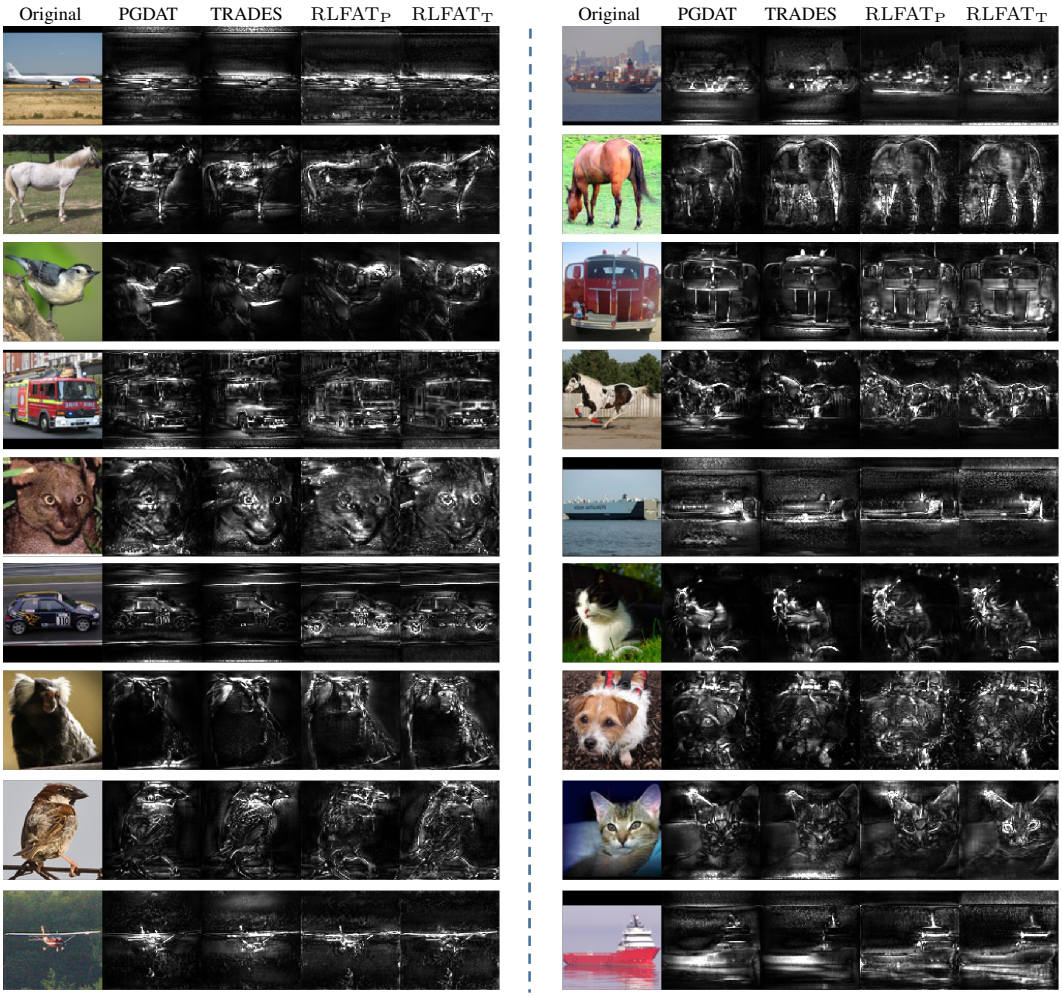

Figure 4: **More Salience maps of the four models.** For each group of images, we have the original image, and the salience maps of the four models sequentially.