ADVERSARIAL DEFENSE VIA LOCAL FLATNESS REGULARIZATION

Jia Xu*  Yiming Li*  Yang Bai  Yong Jiang  Shu-Tao Xia

Tsinghua Shenzhen International Graduate School, Tsinghua University, China
{xujia19, li-ym18, y-bai17}@mails.tsinghua.edu.cn; {jiangy, xiast}@sz.tsinghua.edu.cn

ABSTRACT

Adversarial defense is a popular and important research area. Due to its intrinsic mechanism, one of the most straightforward and effective ways is to analyze the property of loss surface in the input space. In this paper, we define the local flatness of the loss surface as the maximum value of the chosen norm of the gradient regarding to the input within a neighborhood centered at the sample, and discuss its relationship with adversarial vulnerability. Based on the analysis, we propose a new defense approach via regularizing the local flatness (LFR). We demonstrate the effectiveness of the proposed method also from other perspectives, such as human visual mechanism, and analyze the relationship between LFR and related methods theoretically. Experiments are conducted to verify our theory and demonstrate the superiority of the proposed method.

Index Terms—adversarial defense, loss surface geometry, gradient-based regularization.

1. INTRODUCTION

Deep neural networks (DNNs) have been widely and successfully used in many computer vision area, such as pose estimation [1, 2, 3], object detection [4, 5, 6] and super-resolution [7, 8, 9].

Despite their excellent performance under the standard setting, recently, researchers found that DNNs are vulnerable to some well-designed pixel-wise perturbations. Those perturbations are invisible to human, whereas is possible to fool network with high probability. For example, some attack methods, such as Fast Gradient Sign Method (FGSM) [10] and Deepfool [11], are proposed, which can reduce the accuracy of the network to be close to 0% on CIFAR-10 dataset.

To reduce the adversarial vulnerability, some adversarial defense methods are proposed, which can be roughly divided into four categories including adversarial training based defense [12, 13], detection based defense [14, 15], and reconstruction based defense [16]. Particularly, one of the most straightforward ways is to analyze the loss surface in the input space, since the adversarial attack is to find the worst-case perturbation in the input space. Specifically, in [17], the relationship between the average value of the chosen norm of the gradient and the adversarial vulnerability was discussed, based on which they proposed a defense method by regularizing that value. Recently, Qin [18] pointed out that the previous method did not consider the local characteristics, and it may have a relatively large error when the loss surface is not linear enough in the local area, Accordingly, they proposed a new defense method to achieve the local linearity based on regularizing the difference between the loss and its first-order Taylor expansion.

Is the local linearity really necessary for the adversarial defense? In this paper, we make a more systematical discussion on the relationship between adversarial robustness and the local flatness of loss surface. Specifically, the local flatness can be measured by the maximum value of the chosen norm of the gradient with respect to the input within a neighborhood centered at the sample. We prove that whether a sample is easy to be attacked is related to the flatness of loss surface around that sample. Based on this discussion, we propose a new defense method via regularizing the local flatness (LFR) to enhance the adversarial robustness. Besides, we compare our method with the human visual mechanism and the local Lipschitz property to further verify the validity of the method. Discussion on the relationship between our LFR and other previous related defense methods is theoretically conducted, which demonstrates that most of them are special cases of LFR under certain conditions.

The main contributions of this paper can be summarized as follows:

- We give a systematic discussion on the relationship between loss surface flatness and the adversarial robustness. Based on the analysis, we propose a new regularization, the LFR, for the adversarial defense.

- We theoretically discuss the relationship between LFR and previous related defense methods.

- Experiments and comparison with the human visual mechanism and the local Lipschitz property are included, which further verify the validity of the proposed method.
2. LOCAL FLATNESS REGULARIZATION

2.1. Preliminaries

Suppose \( L(\cdot) \) is the loss function (such as cross-entropy or the K-L divergence), and \( B_p(x, \epsilon) \) is the \( \epsilon \)-ball centering around \( x \) under \( \ell^p \) norm, i.e., \( B_p(x, \epsilon) = \{ x' \mid ||x' - x||_p \leq \epsilon \} \).

**Definition 1.** The local flatness of the loss surface (generated by classifier \( C \)) around \( B_p(x, \epsilon) \) is defined as

\[
\gamma_C(x, \epsilon) = \max_{x' \in B_p(x, \epsilon)} ||\partial_x L(x')||_q, \quad \left( \frac{1}{p} + \frac{1}{q} = 1 \right). \tag{1}
\]

Note that \( ||\cdot||_q \) is the dual norm of \( ||\cdot||_p \) when \( \frac{1}{p} + \frac{1}{q} = 1 \), which is introduced for the convex optimization in the proof. For example, \( \ell^1 \) is the dual norm of \( \ell^\infty \).

The reason why \( \gamma_C(x, \epsilon) \) measures the local flatness can be easily explained. By the definition, gradient (at point \( x' \)) indicates the direction in which loss is changed at the highest rate. As such, \( ||\partial_x L(x')|| \) can be regarded as the fluctuation at \( x' \), and therefore \( \max_{x' \in B_p(x, \epsilon)} ||\partial_x L(x')|| \) measures the flatness of local \( B_p(x, \epsilon) \).

Now we use the defense of \( \ell^\infty \) attack as an example to discuss the relation between local flatness and adversarial robustness systematically.

**Theorem 1.** \( \forall x' \in B_\infty(x, \epsilon), L(x') \leq L(x) + \epsilon \cdot \gamma(x, \epsilon), \) where \( \gamma(x, \epsilon) = \max_{x' \in B_\infty(x, \epsilon)} ||\partial_x L(x')||_1 \).

**Proof.** \( \forall x' \in B_\infty(x, \epsilon), \forall t \in [0, 1], \) we have

\[
tx' + (1-t)x \in B_\infty(x, \epsilon).
\]

Therefore

\[
L(x) = L(x) + \int_0^1 \nabla L(tx' + (1-t)x) (x' - x) dt \\
\leq L(x) + \epsilon \cdot \max_{x' \in B_\infty(x, \epsilon)} ||\partial_x L(x')||_1.
\]

\( \square \)

**Theorem 1** indicates that we can defend attack under \( \ell^\infty \) norm by regularizing the maximum of the dual norm of the local gradient.

Except from the previous perspective, the effectiveness of this regularization can also be verified through the following aspects:

**Verification from the aspect of human visual mechanism**

It is widely accepted that the human visual system relies mainly on key components rather than all pixels of the whole image. For example, when categorizing a picture as a cat, the eye only pays attention to the pixels of the cat and ignores background (such as the grass or the house).

Suppose \( x = (x_{ij})_{m \times n} \). Let

\[
\partial_x L(x) = \left( \frac{\partial L}{\partial x_{1,1}} \cdots \frac{\partial L}{\partial x_{1,n}} \cdots \frac{\partial L}{\partial x_{m,1}} \cdots \frac{\partial L}{\partial x_{m,n}} \right)
\]

indicates the matrix of the gradient of the loss function with respect to each pixel of the image \( x \). The (absolute value of) gradient of the loss function at the pixel measures how important the pixel is to the prediction.

Although there is no well-developed metric to identify which pixels are important to human vision system, using only key pixels at least means that the matrix \( \partial_x L(x) \) should be sparse, which can be constrained by the \( \ell^1 \) norm of the matrix. This is also consistent with the phenomenon that the gradients are significantly more human-aligned for adversarially trained networks, as observed in [19]. In addition, since the human visual system is not sensitive to small pixel-wise changes, i.e., human visual system is robust under certain pixel-wise perturbation, taking the local property of the gradients into consideration is rational.

**Verification from the aspect of local Lipschitz property**

**Lemma 1.** Let \( L(\cdot) \) be a Lipschitz continuous function, then

\[
\max_{x' \in B_\infty(x, \epsilon)} ||\partial_x L(x')|| \leq Lip(L), \text{ where } Lip(L) \text{ is the Lipschitz constant of } L \text{ in the local } B_\infty(x, \epsilon). \tag{20}
\]

**Lemma 1** indicates that regularizing \( \gamma(x, \epsilon) \) has a direct connection with regularizing \( Lip(L) \), which is effective for the defense since \( \forall x' \in B_\infty(x, \epsilon), \)

\[
||L(x') - L(x)|| \leq ||x' - x|| \cdot Lip(L) \leq \epsilon \cdot Lip(L),
\]

according to the definition of Lipschitz property.

2.2. Proposed method

Following the analysis above, we now propose the following objective for the adversarial defense

\[
L(D) = E_{(x,y) \in D} \left\{ L_{normal}(x,y) + \lambda \cdot \max_{x' \in B_\infty(x, \epsilon)} ||\partial_x L(x')||_q \right\},
\]

where \( \frac{1}{p} + \frac{1}{q} = 1 \), \( L_{normal}(\cdot) \) is the normal training loss (such as cross-entropy or K-L divergence), and \( \lambda \) is a hyper-parameter balancing the standard accuracy and the adversarial accuracy. The training process is shown in Algorithm 1.

3. COMPARISON WITH RELATED METHODS

In this section, we compare our method with some previous related defense methods. Specifically, we prove that both adversarial training and first-order based adversarial defense is a special case of our method under certain conditions.
Algorithm 1 The training process of LFR

1: **Input:** Training set \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \), and hyperparameter \( \lambda \).
2: Generate the fluctuated sample \( x'_i \) by maximizing the LFR around image \( x_i \) through PGD, \( (i = 1, \ldots, N) \).
3: Learn model parameter \( \theta \) by minimizing

\[
L_{normal}(x) + \lambda \max_{x' \in B_p(x, \epsilon)} ||\partial_x L(x')||_q, \left( \frac{1}{p} + \frac{1}{q} = 1 \right),
\]

as shown in equation (2).
4: **Return:** Optimized model parameter \( \theta^* \).

### 3.1. Link to the first-order based defense

**Definition 2.** To defend the attack under \( \ell^p \) norm, first-order based adversarial defense is to minimize \( L(x) + \lambda ||\partial_x L(x)||_q, \left( \frac{1}{p} + \frac{1}{q} = 1 \right) \).

**Theorem 2.** First-order based adversarial defense is a special case of LFR.

Proof. \( ||\partial_x L(x)||_q = \max_{x' \in B_p(x,0)} ||\partial_x L(x')||_q \). \( \square \)

### 3.2. Link to the adversarial training

**Lemma 2.** \( \max_{||\alpha||_p \leq 1} \alpha \beta = ||\beta||_q, \left( \frac{1}{p} + \frac{1}{q} = 1 \right) \).

**Theorem 3.** Adversarial training using scaled FGSM attack under \( \ell^p \) norm is equivalent to minimizing \( L(x) + \epsilon \max_{\alpha' \in B_p(x,0)} ||\partial_x L(x')||_q, \left( \frac{1}{p} + \frac{1}{q} = 1 \right) \).

Proof. According to the first-order Taylor expansion,

\[
L(x + \alpha) = L(x) + \alpha \cdot \partial_x L(x) + O(\alpha^2).
\]

Therefore, the optimal adversarial training using scaled FGSM attack can be solved approximately up to terms of order \( \epsilon^2 \) by

\[
\max_{||\alpha||_p \leq \epsilon} L(x) + \alpha \cdot \partial_x L(x) = L(x) + \max_{||\alpha'||_p \leq 1} \epsilon \cdot \partial_x L(x). \tag{4}
\]

According to Lemma 2,

\[
L(x) + \epsilon \cdot ||\partial_x L(x)||_q. \tag{5}
\]

In other words, adversarial training is a special case of first-order defense to a certain extent. By Theorem 2, the statement is proved. \( \square \)

### 4. EXPERIMENTS

#### 4.1. Setup

**Baselines Selection** We select trade-off inspired adversarial defense via surrogate-loss minimization (TRADES) [13], local linearization Regularization (LLR) [18] and PGD-based adversarial training (AT) [12] as the baseline methods in the following experiments, since they are the representative of the state-of-the-art defense methods, most advanced loss surface geometry based defenses, and the most classical defense methods, respectively. Besides, we also train a model with a standard training process as an important baseline in this paper.

**MNIST Setup** We adopt a simple CNN architecture, which consists of four convolutional layers followed by three fully-connected layers. We apply both Fast Gradient Sign Method (FGSM) and Project Gradient Descent (PGD) attacks to evaluate the adversarial robustness of different methods. Specifically, we set the perturbation \( \epsilon = 0.3 \), the perturbation step size \( \eta = 0.01 \), number of iterations \( K = 40 \) in inner maximization problem, and run 100 epochs with learning rate \( \alpha = 0.01 \) and batch size \( m = 128 \). These settings are learned from [13]. For the hyperparameters of different methods, we set \( 1/\lambda = 1 \) for TRADES, \( \lambda = 4 \) and \( \mu = 3 \) for LLR according to the setting suggested in their paper.

#### 4.2. Adversarial defense under white-box attacks

In this section, we evaluate the adversarial robustness of different methods. We define robust accuracy \( A_{rob}(f) \) as the test accuracy on perturbed images while natural accuracy \( A_{nat}(f) \) refers to the result on the clean testing set.

As shown in Table 1, LFR performs best under both FGSM and PGD attacks with accuracy of 98.14% and 96.99% respectively. For FGSM, LFR achieves improvements of 0.73%, 0.45%, 0.64% over AT, TRADES and LLR respectively. Under stronger attack i.e. the PGD 40, LFR outperforms the SOTA baseline, the LLR, by a even larger margin.

| Defense      | Attack Type | \( A_{nat}(f) \) | \( A_{rob}(f) \) |
|--------------|-------------|-----------------|-----------------|
| No Defense   | FGSM        | 99.29%          | 56.13%          |
| AT           | FGSM        | 99.45%          | 97.41%          |
| TRADES       | FGSM        | 99.49%          | 97.69%          |
| LLR          | FGSM        | 99.55%          | 97.50%          |
| LFR          | FGSM        | 99.40%          | 98.14%          |
| No Defense   | PGD 40      | 99.29%          | 0.16%           |
| AT           | PGD 40      | 99.45%          | 95.14%          |
| TRADES       | PGD 40      | 99.49%          | 95.52%          |
| LLR          | PGD 40      | 99.55%          | 95.59%          |
| LFR          | PGD 40      | 99.40%          | 96.99%          |
4.3. Visualization of loss surfaces

In this section, we analyze the effectiveness of LFR from the aspect of loss surface geometrically. In particular, we visualise the modified loss surfaces of chosen models using the method proposed in [22]. This new loss \( \hat{L}(x) > 0 \) is defined by the difference between the logits corresponding to the ground-truth label and the remaining maximal one. Hence, it can evaluate the decision confidence where \( \hat{L}(x) > 0 \) indicates that the prediction of \( x \) is correct and vice versa.

Specifically, We visualize the modified loss surface of the LFR model with \( \lambda = 6500 \) and the natural model with standard training (\( \lambda = 0 \)) in the input space, as shown in Fig. 1.

As shown in Fig. 1, the loss surfaces between these two models behave quite differently. Compared with LFR, the loss surfaces of the natural model have sharper peaks and larger slopes, implying that the decision is vulnerable to small perturbation. In other words, the decision confidence could quickly drop to negative areas where the model is fooled after attacked by a small pixel-wise adversarial perturbation. In contrast, the loss surfaces of LFR are rather flat and located on a plateau with positive decision confidence in the vicinity around the sample. As such, the output of LFR still lies in the correct classification regions after being attacked.

It is worth noting that the LFR loss surfaces are flat rather than linear, while their corresponding models are still robust. Therefore, it is the local flatness rather than the local linearity that is critical to adversarial defense.

4.4. Analysis of hyperparameter \( \lambda \)

In this section, we further analyze how \( \lambda \) could affect the performance.

![Fig. 1. Modified loss surfaces of the natural and LFR models on two randomly selected samples.](image)

As shown in the Fig. 1, the improvement of adversarial robustness led by LFR is significant, especially when the \( \lambda \) is well-selected.

5. CONCLUSIONS

In this paper, we propose a new gradient-based regularization, the local flatness regularization (LFR), based on the relationship between the adversarial vulnerability and the local flatness of loss surface, which is defined as the maximum value of the chosen norm of the gradient regarding to the input within a neighborhood centered at the sample. We theoretically discuss the relationship between LFR with previous related defense methods, and further verify the effectiveness also from the aspect of human visual mechanism and local Lipschitz property. Verification experiments are conducted, which demonstrates the superiority of the proposed method.
6. REFERENCES

[1] Alexander Toshev and Christian Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in CVPR, 2014, pp. 1653–1660.

[2] Rıza Alp Guler, Natalia Neverova, and Iasonas Kokkinos, “Densepose: Dense human pose estimation in the wild,” in CVPR, 2018, pp. 7297–7306.

[3] Jun Liu, Henghui Ding, Amir Shahroudy, Ling-Yu Duan, Xudong Jiang, Gang Wang, and Alex Kot Chichung, “Feature boosting network for 3d pose estimation,” IEEE transactions on pattern analysis and machine intelligence, 2019.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, “You only look once: Unified, real-time object detection,” in CVPR, 2016, pp. 779–788.

[5] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei, “Relation networks for object detection,” in CVPR, 2018, pp. 3588–3597.

[6] Peiliang Li, Xiaozhi Chen, and Shaojie Shen, “Stereo r-cnn based 3d object detection for autonomous driving,” in CVPR, 2019, pp. 7644–7652.

[7] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in CVPR, 2016, pp. 1646–1654.

[8] Guimin Lin, Qingxiang Wu, Lida Qiu, and Xixian Huang, “Image super-resolution using a dilated convolutional neural network,” Neurocomputing, vol. 275, pp. 1219–1230, 2018.

[9] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos, “Generative adversarial networks and perceptual losses for video super-resolution,” IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3312–3327, 2019.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples,” in ICLR, 2015.

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in CVPR, 2016.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, “Towards deep learning models resistant to adversarial attacks,” in ICLR, 2018.

[13] Hongyang Zhang, Yaodong Yu, Jiatao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan, “Theoretically principled trade-off between robustness and accuracy,” in ICML, 2019.

[14] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff, “On detecting adversarial perturbations,” in ICLR, 2017.

[15] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu, “Detection based defense against adversarial examples from the steganalysis point of view,” in CVPR, 2019.

[16] Pouya Samangouei, Maya Kabkab, and Rama Chellappa, “Defense-gan: Protecting classifiers against adversarial attacks using generative models,” in ICLR, 2018.

[17] Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz, “First-order adversarial vulnerability of neural networks and input dimension,” in ICML, 2019, pp. 5809–5817.

[18] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Alhussein Fawzi, Soham De, Robert Stanforth, Pushmeet Kohli, et al., “Adversarial robustness through local linearization,” in NeurIPS, 2019.

[19] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, “Adversarial examples are not bugs, they are features,” in ICLR, 2019.

[20] Lawrence Craig Evans and Ronald F Gariepy, Measure Theory and Fine Properties of Functions, CRC Press, 2015.

[21] Stephen Boyd and Lieven Vandenberghe, Convex optimization, Cambridge university press, 2004.

[22] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, and Xiang Chen, “Interpreting and evaluating neural network robustness,” in IJCAI, 2019.