Improving Adversarial Transferability with Gradient Refining

Guoqiu Wang[*], Huanqian Yan[*], Ying Guo[*], Xingxing Wei[†]

Beijing Key Laboratory of Digital Media, Beihang University, Beijing, China

{wangguoqiu, yanhq, yingguo, xxwei}@buaa.edu.cn

Abstract

Deep neural networks are vulnerable to adversarial examples, which are crafted by adding imperceptible perturbations to original images. Most existing adversarial attack methods achieve nearly 100% attack success rates under the white-box setting, but only achieve relatively low attack success rates under the black-box setting. To improve the transferability of adversarial examples for the black-box setting, several methods have been proposed, e.g., input diversity, translation-invariant attack, and momentum-based attack. In this paper, we propose a method named Gradient Refining, which can further improve the adversarial transferability by correcting negative gradients introduced by input diversity through multiple transformations. Our method is generally applicable to many gradient-based attack methods combined with input diversity. Extensive experiments are conducted on the ImageNet dataset and our method can achieve an average transfer success rate of 82.07% for three different models under single-model setting, which outperforms the other state-of-the-art methods by a large margin of 6.0% averagely. And we have applied the proposed method to the competition CVPR 2021 Unrestricted Adversarial Attacks on ImageNet and won the second place in attack success rates among 1558 teams.

1. Introduction

Though deep neural networks (DNNs) have achieved state-of-the-art performance on various vision tasks, including image classification [11, 20], object detection [6, 19, 18] and semantic segmentation [2, 15], they are vulnerable to adversarial examples [7, 22] which are crafted by adding imperceptible perturbations to original images, making models output wrong predictions expected by attackers. The existence of adversarial examples has raised concerns in security-sensitive applications, e.g., self-driving cars [13] and face recognition [8]. And two related research directions, one is trying to improve the attack ability of adversarial
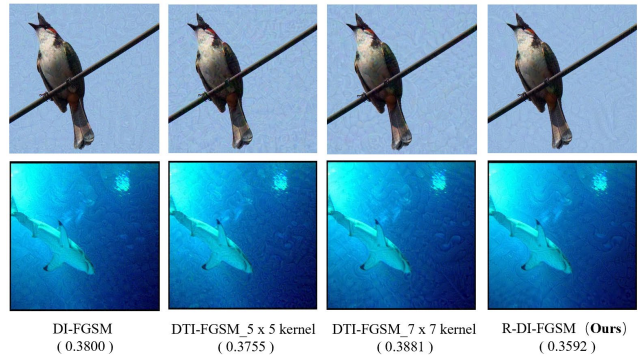

Figure 1. Several adversarial examples of different attacks. The text denotes the corresponding attack method and LPIPS [1] is shown in brackets. The transfer attack success rates are 67.10%, 69.88%, 69.92% and 70.02% respectively. The adversarial perturbation generated by our method in the fourth column shows better imperceptibility and transferability.

examples, the other is to improve the robustness of DNNs against the adversarial examples, promote each other and develop together.

The generation of adversarial examples can serve as an important surrogate to evaluate the robustness of different models, and many methods, such as FGSM [7], PGD [16], and C&W [1], have been proposed, which use the gradient information of the given model and are known as white-box attacks. Moreover, some works show that the adversarial examples have transferability [14, 17], which means the adversarial examples crafted for one model can attack a different model successfully. The existence of transferability makes black-box attacks, where the attackers have no knowledge of the model structure and parameters, practical to the real-world applications and induces serious security issues.

Although above white-box attacks can achieve high success rates, they have low transferability. Recently, many methods have been proposed to improve the transferability, such as multi-model ensemble transferable attack [14], input diversity [23], translation-invariant attack [5], and momentum-based attack [4]. By attacking an ensemble of

---

[*]Equal contribution.
[†]Corresponding author.

[1]LPIPS [24] is the perceptual similarity with original image (the lower the better).

models, an example remains adversarial for multiple models, and is more likely to transfer to other black-box models. Input diversity applies random and differentiable transformations to the input images to create hard and diverse input patterns. Translation-invariant attack smooths the gradient with a pre-defined kernel, making adversarial examples less sensitive to the discriminative region of the white-box model being attacked. Momentum-based iterative attack integrates the momentum term into the iterative process, which can stabilize the update directions and escape from poor local maxima during the iterations, leading to more transferable adversarial examples.

Fundamentally, the over-fitting of the source model is one of the key reasons for the low transferability of white-box attacks [23]. Input diversity [23] and translation-invariant attack [5] mentioned above decrease the over-fitting by random input transformations or smoothing the gradient with a kernel to improve the transferability. However, the randomness introduced by image transformations in input diversity does not always play a positive role in improving the transferability and sometimes brings the random **negative gradient information**, which has a negative effect in improving the transferability. We propose a method to neutralize the negative gradient information, which considers multiple perturbations from multiple random transformations of the same image comprehensively, retains the positive gradient information, and counteracts the negative gradient information to refine the final gradient. Combining our method with existing methods, such as translation-invariant attack and momentum-based attack, the transferability can be further improved. Extensive experiments demonstrate the effectiveness of our method and show that the generated perturbations have better imperceptibility (see Figure 1).

In brief, the main contributions of this paper can be summarized as follows:

- We propose an effective method to improve the adversarial transferability, which counteracts the negative gradients generated during the input diversity process by comprehensively considering the gradients of multiple transformed images.

- Our method has good scalability. It can be well integrated with existing attack strategies (e.g., I-FGSM [12], MI-FGSM [4], TI-FGSM [5]) to further improve the adversarial transferability.

- Extensive experimental results show that the proposed method can get a higher transfer success rate of 82.07% on average, which outperforms the other mainstream methods by a large margin of 6.0%. By applying the proposed method to the competition CVPR 2021 Unrestricted Adversarial Attacks on ImageNet, we ranked fourth overall, and won the second place in attack success rates among 1558 teams.

2. Related Work

In this section, we provide the background knowledge of adversarial attack as well as several classic methods which are designed to improve the attack ability and transferability of adversarial examples.

Given a classification network f_θ parameterized by θ, let (x, y) denote the original image and its corresponding ground-truth label, the goal of adversarial attacks is to find an example x_{adv} which is in the vicinity of x but misclassified by the network. In most cases, we use the L_p norm to limit the adversarial perturbations below a threshold ϵ, where p could be $0, 2, \infty$. This can be expressed as:

$$f_\theta(x_{adv}) \neq y, \ s.t. \ \|x_{adv} - x\|_p \leq \epsilon \qquad (1)$$

Input diversity iterative (DI) attack [23] applies random transformations to the input images at each iteration to create hard and diverse input patterns, which brings randomness to the adversarial perturbations and improves the generalization of adversarial examples efficiently. This method can be formalized as:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(\bigtriangledown_x L(T(x_t^{adv}, p), y)) \qquad (2)$$

where $\bigtriangledown_x L(T(x_t^{adv}, p), y)$ denotes the gradient of the loss function $L(\cdot, \cdot)$ w.r.t. the transformed image of adversarial example x_t^{adv} generated in step t. $sign(\cdot)$ is the sign function and α denotes the step size. $T(\cdot, p)$ denotes the stochastic transformation function with probability p. If $p = 1.0$, then only transformed inputs are used for the attack. Since the original inputs are not seen by the attackers, the generated adversarial examples tend to have much higher attack success rates on black-box models but lower success rates on white-box models.

Translation-invariant iterative (TI) attack [5] improves attack ability by smoothing the gradient of the image with a pre-defined kernel, making the adversarial example less sensitive to the discriminative region of the white-box model being attacked. This method can be formulated as:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(\boldsymbol{W} * \bigtriangledown_x L(x_t^{adv}, y)) \qquad (3)$$

where $\boldsymbol{W}$ denotes the pre-defined kernel. The size of $\boldsymbol{W}$ plays a key role for improving the transferability, and with the size increasing, the adversarial perturbations will be more perceptible (see Figure 1).

Momentum-based iterative (MI) method [4] boosts adversarial attacks by integrating the momentum term into the iterative process to stabilize the update directions. The momentum-based iterative attack of FGSM (MI-FGSM) can be expressed as:

$$g_{t+1} = \mu \cdot g_t + \frac{\bigtriangledown_x L(x_t^{adv}, y)}{\| \bigtriangledown_x L(x_t^{adv}, y)\|_1} \qquad (4)$$
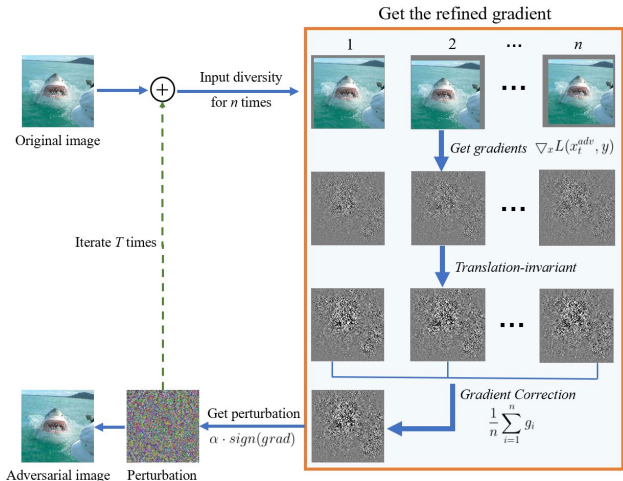
Figure 2. The framework of R-DTI-FGSM algorithm.

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1}) \quad (5)$$

where g_t is the moment gradient and μ is the decay factor which is often set to 1.0. MI-FGSM updates the moment gradient g_{t+1} by Eq. (4) and then updates x_{t+1}^{adv} by Eq. (5).

3. Methodology

Input diversity creates diverse input patterns and the image will be transformed once randomly with a certain probability during each iteration in the process of attacking an image, thus enhancing the transferability. But this inspires us a question: *The randomness brought by input diversity does always positive? If not, how can we counteract negative randomness?*

Input diversity method brings random gradients and translation-invariant method magnifies the randomness. Though the random gradients reduce the over-fitting and make the adversarial examples more generalized and transferable, there exists random negative information of gradients, which inhibits the generalization and transferability of adversarial examples to some extent.

To reduce the effects of random negative gradients introduced by input diversity, we design a method named Gradient Refining, which can enhance the transferability of adversarial examples efficiently. In the process of attacking an image, our method randomly transforms the image with several times during each iteration and calculates their respective gradients, then averages them to get a refined gradient, which preserves the gradients helpful for the transferability and counteracts the negative gradients for the attack.

Gradient Refining can be combined with DI, DI-TI or DI-TI-MI, which called R-DI-FGSM, R-DTI-FGSM and R-DTMI-FGSM respectively. R-DTI-FGSM method can be

Algorithm 1 R-DTMI-FGSM Algorithm
***
Input: A classifier $f(\cdot)$ with loss function L, an original image x and ground-truth label y, correction number n, decay factor μ, iterations T, the max perturbation ϵ.
Output: adversarial example x_{adv}
1: $\alpha = \epsilon / T$;
2: $g_r = 0, x_0^{adv} = x, g_0 = 0$;
3: **for** $t = 0$ to $T-1$ **do**
4:     Input x_t^{adv} to $f(\cdot)$ and obtain gradient g_r by
    $g_r = \frac{1}{n}\sum_{i=1}^{n}(W * \bigtriangledown_x L(T_i(x_t^{adv}, p), y))$;
5:     Update g_{t+1} by $g_{t+1} = \mu \cdot g_t + g_r$;
6:     Update x_{t+1}^{adv} by $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1})$
7:     $g_r = 0$;
8: **end for**
9: **return** $x_{adv} = x_T^{adv}$
***

formalized as:

$$g_r = \frac{1}{n}\sum_{i=1}^{n}(W * \bigtriangledown_x L(T_i(x_t^{adv}, p), y)) \quad (6)$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_r) \quad (7)$$

where n denotes the correction times of the gradient. With the increase of n, the attack ability will be improved, and when $n = 1$, it is DTI-FGSM.

The framework of R-DTI-FGSM algorithm is in Figure 2. For an original image, we use input diversity, which introduces randomness, for several times, to get multiple transformed images, and calculate their respective gradients. Translation-invariant method is used to amplify the randomness. Then we retain the positive information of gradients and counteract the negative gradients through average operation to get a refined gradient. We iterate the above process for T times and get an adversarial image with high transferability.

For R-DTMI-FGSM algorithm, we use refined g_r to update moment gradient g_{t+1} and then update x_{t+1}^{adv} by Eq. (5). The overall R-DTMI-FGSM algorithm is outlined in Algorithm 1. In this paper, we show the algorithm and its experimental results under the constraint of L_∞ norm. The method can also be used in L_2 norm.

4. Experiments

4.1. Setup

Dataset: We perform experiments on the dataset provided by alibaba group, which contains 5000 images selected from ImageNet [3], five images in each class, and all these images are resized to 224×224.

Networks: We study four models with different structures and parameters, where three of them are normally trained models and the last one is the adversarially trained

Table 1. The attack success rates (%) under single-model setting and ensemble-model setting. $*$ indicates the white-box model being attacked. The right columns of $*$ for both setting show the transfer attack success rates in three different models and the average results for different attack methods. The improvement of average ASR by our method is shown in the brackets.

|	Single-model attack					Ensemble-model attack					
Attack methods	Res-101*	Den-161	Inc-v4	advIncRes-v2	Average	Res-101*	VGG-16*	Den-161	Inc-v4	advIncRes-v2	Average
DI-FGSM	99.96	72.58	62.28	37.28	57.38	99.76	99.90	86.52	74.74	40.04	67.10
R-DI-FGSM(**Ours**)	100.0	76.90	64.84	38.06	59.93 (↑2.55)	99.98	99.98	90.76	78.06	41.26	70.02 (↑2.92)
DTI-FGSM	99.96	74.36	65.50	40.76	60.20	99.82	99.92	88.48	77.60	43.58	69.88
R-DTI-FGSM(**Ours**)	100.0	85.06	76.26	46.44	69.25 (↑9.05)	99.98	99.98	95.34	85.28	49.26	76.62 (↑6.74)
DTMI-FGSM	99.90	90.50	82.64	55.30	76.14	99.82	99.90	94.84	88.34	59.16	80.78
R-DTMI-FGSM(**Ours**)	99.98	**94.90**	**88.42**	**62.90**	**82.07** (↑5.93)	99.92	99.96	**97.00**	**91.16**	**64.82**	**84.32** (↑3.54)
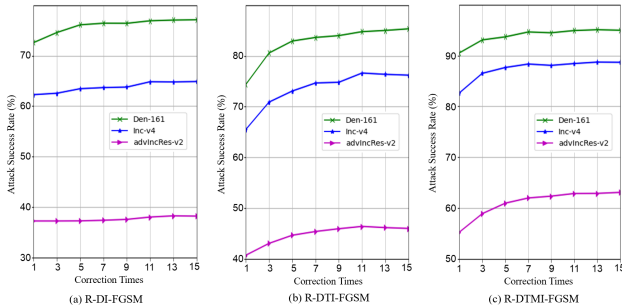
Figure 3. The transfer attack success rates (%) against Den-161, Inc-v4 and advIncRes-v2. The adversarial examples are generated on Res-101 model with the correction times ranging from 1 to 15.

model. They are Resnet 101 (Res-101) [9], Denset 161 (Den-161) [10], Inception v4 (Inc-v4) [21], and adv-Inception Resnet v2 (advIncRes-v2) [21].

Implementation details: We set the maximum perturbation of each pixel to be $\epsilon = 16 / 255$ and the total iteration number T is equal to 8. For the stochastic transformation function $T(x, p)$, the probability p is set to be 0.7. The size of gaussian kernel W is 5×5.

Metrics: We use Attack Success Rate (ASR), which refers to the percentage of all images that can be misclassified by the target model, to evaluate different attack methods.

4.2. Results

4.2.1 The Effect of Correction Times

The correction times n plays a key role in improving the transfer attack success rates. If n is equal to 1, our method will degenerate to their vanilla versions. Therefore, we conduct an ablation study to examine the effect of correction times. We attack the Res-101 model by R-DI-FGSM, R-DTI-FGSM, and R-DTMI-FGSM with different correction times, which range from 1 to 15 with a granularity 2, and the results are shown in Figure 3. The transfer attack success rates continuously increase at first, and turn to remain stable after the correction times exceed 11. Considering attack ability and computational complexity, the correction times are set to 11 in the following experiments.

4.2.2 Attack Success Rates on Different Settings

We perform adversarial attacks using DI-FGSM, DTI-FGSM and DTMI-FGSM, and their extensions by combining with Gradient Refining attack method as R-DI-FGSM, R-DTI-FGSM and R-DTMI-FGSM. We report the attack success rates of the Gradient Refining based attacks with three baseline methods under single-model setting and ensemble-model setting. The results are shown in Table 1.

Under single-model attacks, the transfer attack success rates are improved by a large margin when using the proposed Gradient Refining on the listed attack methods. On average, the transfer attack success rates of Gradient Refining based attacks outperform the baseline attacks by 6.0%. In particular, by using R-DTI-FGSM, the generated adversarial examples have 69.15% average transfer attack success rates, which is improved by 9.05% from the baseline method.

Under the ensemble-model setting, the results in table 1 show that the transferability can also be improved by using Gradient Refining. Here we use two models, Res-101 [9] and VGG-16 [20]. R-DTI-FGSM improves the transfer attack success rates from the baseline method largely, because translation-invariant method magnifies the randomness brought by input diversity, and our method can correct the negative randomness and get a refined gradient to generate stronger adversarial perturbations. Besides, we visualize some adversarial images generated by DI-FGSM, DTI-FGSM and R-DI-FGSM in Figure 1. These visualization results show that the adversarial perturbations generated by our method have better imperceptibility.

5. Conclusion

In this paper, we proposed Gradient Refining method to generate adversarial perturbations that have higher transferability and better imperceptibility. By combining with input diversity, our method can be integrated into many existing attack methods. We conducted extensive experiments to validate the effectiveness. By using R-DTMI-FGSM algorithm, we can fool three different state-of-the-art models with an average success rate of 82.07% under the single-model setting and higher success rates under the ensemble-model setting. Experimental results reveal the vulnerability of current models and inspire us to develop more robust models.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 1, 2

[5] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 1, 2

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[8] Ying Guo, Xingxing Wei, Guoqiu Wang, and Bo Zhang. Meaningful adversarial stickers for face recognition in physical world. *arXiv preprint arXiv:2104.06728*, 2021. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[12] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. 2

[13] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1028–1035, 2019. 1

[14] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 1

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[17] Nicolas Papernot, Patrick McDaniel, Ian 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*, pages 506–519, 2017. 1

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 4

[21] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 4

[22] 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. 1

[23] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 1, 2

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1