Enhancing the Transferability of Adversarial Attacks with Input Transformation

Bo Yang*, Hengwei Zhang*† and Kaiyong Xu

PLA SSF Information Engineering University, Zhengzhou 450001, China

*These authors contributed equally to this work
†Corresponding Author: Hengwei Zhang. Email: zhw11qd@aliyun.com

Abstract. Deep neural networks (DNNs) are challenged by their vulnerability to adversarial examples, which are crafted by adding small, human-imperceptible perturbations to the original images, but make the model output inaccurate predictions. Adversarial attacks can thus be an important tool to evaluate and select robust models before they are deployed. However, under the challenging black-box setting, most of existing adversarial attacks can only fool a model with a low success rate. Based on image augmentation methods, we found that random transformation of image size can eliminate overfitting in the generation of adversarial examples and improve their transferability. To this end, we propose an adversarial example generation method based on this phenomenon, which can be integrated with Fast Gradient Sign Method (FGSM)-related methods to build a stronger gradient-based attack and generate adversarial examples with better transferability. Extensive experiments on the ImageNet dataset demonstrate that our methods attack both normally trained models and adversarially trained models with higher attack success rates than existing baseline attacks. We hope that our proposed attack method can serve as a benchmark for evaluating the robustness of networks to adversaries and the effectiveness of different defense methods.

1. Introduction

In image recognition, some experiments on standard test sets have proven that deep neural networks (DNNs) are able to classify images with an accuracy approaching that of humans [1-4]. However, DNNs have been shown to be extremely vulnerable to attacks from adversarial examples [5,6], because to add perturbations to an original input image that are imperceptible to humans can result in failure predictions of the models. In addition, adversarial examples have an intriguing property of transferability, i.e., those generated for one model may also be adversarial to another, which enables black-box attacks [7]. Learning how to generate adversarial examples can help us understand the robustness of different models and investigate the insufficiency of current training algorithms.

Although adversarial examples are generally transferable, the optimal approach to improving their transferability needs to be examined further. Several gradient-based attacks have been proposed to generate adversarial examples, such as single-step [6] and iterative [8,9] methods. These show powerful attack capabilities in the white-box setting, however, they often exhibit low success rates under the black-box setting, which we attribute to overfitting of adversarial examples, i.e., the difference in attack performances for adversarial examples under white-box and black-box settings are similar to the neural network performance on training and test sets. Therefore, we can apply methods that improve the performance of DNN to the generation of adversarial examples to eliminate
overfitting and improve their transferability. Data augmentation [1,2] has been shown to be an effective way to prevent models from overfitting during the training process.

To this end, we optimize the process of adversarial example generation based on data augmentation [1,2] and propose the Random Transformation of Image Size Attack Method (RSM) to improve their transferability.

- Inspired by the fact that data augmentation [1,2] can prevent overfitting during training and improve the generalization ability of models, we adapt the random transformation of image size into the iterative gradient-based attack, so as to effectively eliminate overfitting in the generation of adversarial examples and improve their transferability.
- Our method is readily combined with gradient-based attack methods (e.g., iterative gradient-based [8] and momentum iterative gradient-based method [9]) to further boost the success rate of adversarial examples for black-box attacks.
- Our algorithm can be associated with the gradient-based adversarial attack methods by adjusting its parameter settings. It reflects the convenience and advantages of our method.

Extensive experiments on the ImageNet dataset [13] have demonstrated that, compared to existing baseline attacks [8,9], our method, RS-MI-FGSM (Random Transformation of Image Size Momentum Iterative Fast Gradient Sign Method), has a higher success rate for black-box attacks in normally and adversarially trained models. The method of attacking ensemble models simultaneously is used to further improve the transferability of adversarial examples [7]. Under the ensemble attack, RS-DIM reaches an average success rate of 45.1% for black-box attacks on adversarially trained networks, which outperforms MI-FGSM by a large margin of 12.8%. We hope that the proposed attack method can help evaluate the robustness of models and effectiveness of defense methods. We hope that the proposed attack method can serve as a benchmark for evaluating the robustness of models and effectiveness of defense methods.

2. Related Work

2.1. Adversarial Attacks

Biggio et al. [15] first performed evasion attacks on traditional machine learning models based on MNIST, leading to misclassification of the models. Szegedy et al. [5] reported the intriguing property that DNNs are vulnerable to adversarial examples and proposed the L-BFGS method to generate them. Goodfellow et al. [6] demonstrated the fast gradient sign method that can generate adversarial examples with one gradient step. It reduces the computation needed to generate adversarial examples and forms the basis of subsequent FGSM-related methods. Alexey et al. [8] proposed the Iterative Fast Gradient Sign Method (I-FGSM), which greatly improved the success rate for white-box attacks and proved that adversarial examples also exist in the physical world. Dong et al. [9] proposed momentum-based iterative FGSM, improving the transferability of adversarial examples. Xie et al. [14] optimize the adversarial perturbations over the diverse transformation of the input image at each iteration. The transformations include the random resizing and the random padding. Instead of optimizing the adversarial perturbations on a single image, Dong et al. [16] use a set of translated images to optimize the adversarial perturbations. They further develop an efficient algorithm to calculate the gradients by convolving the gradient at untranslated images with a kernel matrix. In addition, the fact that adversarial examples may exist in the physical world brings security threats to the practical application of DNNs [8,17].

2.2. Adversarial Defenses

Many defense methods have been proposed to protect deep neural networks from the threat of adversarial examples [18-21]. Among these, adversarial training is the most effective way to improve model robustness [6,22,23]. In this process, adversarial examples are generated and added to the training set and participate in the model training procedure. Xie et al. [20] found that the effectiveness of adversarial examples can be reduced through random transformation. Guo et al. [21] found a range
of image transformations with the potential to remove adversarial perturbations while preserving the key visual information of an image. Tramèr et al. [23] proposed ensemble adversarial training, utilizing adversarial examples generated for other models to increase training data and further improve the robustness of models.

3. Methodology
Let $x$ be the original input image, $y$ the corresponding true label, and $\theta$ the parameter of the model. $J(\theta, x, y)$ is the loss function of the neural network, which is usually cross-entropy loss. We aim to generate an adversarial example $x_{\text{adv}}$ that is visually indistinguishable from $x$ by maximizing $J(\theta, x, y)$ to fool the model; i.e., the model misclassifies the adversarial example $x_{\text{adv}}$. In this paper, we used an $L_\infty$ norm bound to limit adversarial perturbations, such that $\|x^* - x\|_\infty \leq \epsilon$, where $\epsilon$ is the size of the perturbation. Hence adversarial example generation can therefore be converted into the following optimization problem:

$$\arg\max_{x_{\text{adv}}} J(\theta, x_{\text{adv}}, y), \quad \text{s.t. } \|x_{\text{adv}} - x\|_\infty \leq \epsilon.$$ (1)

### 3.1. Gradient-based adversarial attack methods

Here we provide a brief introduction to several methods of adversarial example generation. **Fast Gradient Sign Method (FGSM).** FGSM [6] is one of the simplest methods, which seeks adversarial examples in the direction of the loss gradient $\nabla J(\theta, x, y)$ with respect to the input and imposes infinity norm restrictions on adversarial perturbations. The updated equation is

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla J(\theta, x, y)).$$ (2)

**Iterative Fast Gradient Sign Method (I-FGSM).** Kurakin et al. [8] proposed an iterative version of FGSM. It divides the gradient operation in FGSM into multiple iterations that can be expressed as follows:

$$x_{\text{adv}}^0 = x, \quad x_{\text{adv}}^{t+1} = \text{Clip}_s\\{x_{\text{adv}}^{t+1} + \alpha \cdot \text{sign}(\nabla J(\theta, x_{\text{adv}}^{t+1}, y))\},$$ (3)

where $\alpha$ is the step size of each iteration and $\alpha = \epsilon / T$, in which $T$ is the number of iterations. The Clip function restricts the adversarial example to be within the $\epsilon$-ball of the original image $x$ to meet the infinity norm constraint. I-FGSM is more effective than FGSM in white-box environments, but less effective in black-box environments. In other words, I-FGSM exhibits poor transferability.

**Momentum Iterative Fast Gradient Sign Method (MI-FGSM).** MI-FGSM [9] is the first method to apply momentum to adversarial example generation, which can stabilize gradient update directions, improve convergence, and greatly increase the attack success rate. It can be expressed as:

$$x_{\text{adv}}^0 = x, \quad g_0 = 0, \quad g_{t+1} = \mu \cdot g_t + \frac{\nabla J(\theta, x_{\text{adv}}^t, y)}{\|\nabla J(\theta, x_{\text{adv}}^t, y)\|},$$ (4)

$$x_{\text{adv}}^{t+1} = \text{Clip}_s\\{x_{\text{adv}}^{t+1} + \alpha \cdot \text{sign}(g_{t+1})\},$$ (5)

where $\mu$ is the decay factor of the momentum term, and $g_t$ is the gradient weighted accumulation of the previous $t$ iterations.

**Projected Gradient Descent (PGD).** PGD [18] is a strong iterative version of FGSM, which improves the attack success rate of adversarial examples. It consists of a random start within the allowed norm ball and then follows by running several iterations of I-FGSM to generate adversarial examples.

### 3.2. Random transformation of image size attack method

Data augmentation [1,2] has been shown to be an effective way to prevent networks from overfitting during the training process. Based on this, we propose the Random Transformation of Image Size Attack Method (RSM), which randomly transforms the size of the original input image with probability $p$ in each iteration to alleviate overfitting. It optimizes the adversarial perturbations of the image with randomly transformed size:
arg max \( x^{\text{adv}} \in \mathbb{R}^d \) \( J(\theta, T(x^{\text{adv}}_T), y) \), \( \text{s.t.} \| x^{\text{adv}} - x \|_\infty \leq \varepsilon \) \hspace{1cm} (6)

\[
T(X^{\text{adv}}, p) = \begin{cases} 
T(X^{\text{adv}}_T), & \text{with probability } p \\
X^{\text{adv}}, & \text{with probability } 1 - p
\end{cases} \hspace{1cm} (7)
\]

For the transformation function \( T() \), the input image is randomly padded to a \( \text{rnd} \times \text{rnd} \times 3 \) image, with \( \text{rnd} \in [299, 330] \), and then resized to the size \( 299 \times 299 \times 3 \). For intuitive understanding, Figure 1 shows some images after random transformation. The transformation probability \( p \) controls the balance between the original input image and the transformed image. With this method, we can achieve effective attacks on the model through data augmentation, avoid overfitting attacks of white-box models, and improve the transferability of adversarial examples.

![Figure 1](image.png)

Figure 1. Images from first to third line are randomly picked original images, randomly transformed images, and their corresponding adversarial images, respectively. The adversarial examples are crafted on Inc-v3 by the RS-MI-FGSM method. It can be seen that these generated adversarial perturbations are human imperceptible.

### 3.3. Attack algorithms

For the gradient processing of crafting adversarial examples, RSM introduces data augmentation to alleviate overfitting. RSM is easily combined with MI-FGSM to form a stronger attack, which we refer to as RS-MI-FGSM (Random transformation of image Size Momentum Iterative Fast Gradient Sign Method). Our algorithm can be associated with the family of FGSM via different parameter settings. For example, RS-MI-FGSM degrades to MI-FGSM if \( p = 0 \), i.e., we can remove step 4 of algorithm 1 to realize MI-FGSM. The detail of RT-MI-FGSM attack method is shown in Algorithm 1.
Algorithm 1 RS-MI-FGSM

Input: A clean example \(x\) with ground-truth label \(y\); a classifier \(f\) with loss function \(J\);

Input: Perturbation size \(\varepsilon\); maximum iterations \(T\) and decay factor \(\mu\).

Output: An adversarial example \(x^{adv}\)

1: \(\alpha = \varepsilon / T\)
2: \(x_0^{adv} = x ; g_0 = 0\)
3: for \(t = 0\) to \(T - 1\) do
4:   Get \(x_t^{adv}\) by \(x_t^{adv} = T(x_t^{adv}; p) \) > apply random transformation of the inputs size with the probability \(p\)
5:   Get the gradients by \(\nabla_x J(\theta, x_t^{adv}, y)\)
6:   Update \(g_t+1\) by \(g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(\theta, x_t^{adv}, y)}{\|\nabla_x J(\theta, x_t^{adv}, y)\|}\)
7:   Update \(x_t^{adv}\) by Eq. (5)
8: return \(x_T^{adv}\)

4. Experiments

Extensive validation experiments were conducted to demonstrate the effectiveness of the proposed methodology. Below, we specify the experimental settings, show the results of attacking a single network, and validate our method on ensemble models.

4.1. Experimental setup

Dataset. It is less meaningful to attack the images that are already classified wrongly. Therefore, we randomly selected 1000 images belonging to 1000 categories (i.e., one image per category) from the ImageNet verification set, which were correctly classified by our testing networks. All these images were adjusted to \(299 \times 299 \times 3\) beforehand.

Models. We consider four normally trained networks and three adversarially trained networks. The four networks are Inception-v3 (Inc-v3) [24], Inception-v4 (Inc-v4) [25], Inception-Resnet-v2 (IncRes-v2) [25], and Resnet-v2-101 (Res-101) [26]; the three adversarially trained networks [23] are ens3-adv-Inception-v3 (Inc-v3ens3), ens4-adv-Inception-v3 (Inc-v3ens4), and ens-adv-Inception-ResNet-v2 (IncRes-v2ens).

Baselines. We integrated our method with I-FGS M [8] and MI-FGS M [9] to evaluate the improvement of RSM over these baseline methods.

Implementation details. For the hyper-parameters, we follow the default settings in [9] with the maximum perturbation \(\varepsilon = 16\), number of iterations \(T = 10\), and step size \(\alpha = 1.6\). For MI-FGS M, the decay factor is defaulted to \(\mu = 1.0\). For our methods, \(p\) is set to 0.5 for the transformation function.

4.2. Attacking a single network

We first perform adversarial attacks on a single network. We use I-FGS M, RS-I-FGS M, MI-FGS M and RS-MI-FGS M to generate adversarial examples only on the normally trained networks and tested them on all seven networks. The results are shown in Table 1, where the success rate is the model classification error rate with adversarial examples as input. We show six adversarial images in Fig. 1 generated for Inc-v3.

The results in Table 1 show that the attack success rates of RT-MI-FGS M under mostly black-box settings are much higher than those of other baseline attacks, and maintains high success rates on all white-box models. For example, when generating adversarial examples on the Inc-v3 network to attack the Inc-v4 network, the success rate for black-box attacks of RT-MI-FGS M reaches 68.5%, the highest among these methods. RS-MI-FGS M also performs better on the adversarially trained networks. Compared to the other three attack methods, our method greatly improves the success rates...
for black-box attacks. For example, when generating adversarial examples on the IncRes-v2 network to attack the adversarially trained networks, the average attack success rates of RT-MI-FGSM and I-FGSM are 29.1% and 19.8%, respectively. This convincingly demonstrates the effectiveness of the combination of input Transformation and momentum for improving the transferability of adversarial examples.

Table 1. The success rates (%) of adversarial attacks against seven models under single model setting.
Adversarial examples are crafted on Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively, using I-FGSM, MI-FGSM, and TI-MI-FGSM. * indicates white-box attacks.

| Model       | Attack          | Inc-v3* | Inc-v4* | IncRes-v2* | Res-101* | Inc-v3ens3 | Inc-v3ens4 | IncRes-v2ens |
|-------------|-----------------|---------|---------|------------|----------|------------|------------|--------------|
| Inc-v3      | I-FGSM          | 99.9    | 22.7    | 20.0       | 18.3     | 7.3        | 7.6        | 4.0          |
|             | RS-I-FGSM(Ours) | 99.2    | 38.8    | 32         | 28       | 9.3        | 8.7        | 5.2          |
|             | MI-FGSM         | 99.9    | 48.3    | 46.9       | 39.8     | 15.0       | 14.4       | 7.1          |
|             | RS-MI-FGSM(Ours)| 99.2    | 68.5    | 63.3       | 55.7     | 19.6       | 20.0       | 9.0          |
| Inc-v4      | I-FGSM          | 37.8    | 99.9    | 26.3       | 21.8     | 8.8        | 8.1        | 5.1          |
|             | RS-I-FGSM(Ours) | 54.6    | 99.4    | 39.6       | 33.3     | 10.7       | 9.7        | 6.1          |
|             | MI-FGSM         | 63.7    | 99.9    | 53.6       | 47.6     | 19.8       | 16.8       | 9.5          |
|             | RS-MI-FGSM(Ours)| 77.0    | 99.2    | 68.8       | 62.4     | 26.7       | 23.7       | 13.4         |
| IncRes-v2   | I-FGSM          | 37.1    | 31.9    | 99.6       | 25.8     | 8.8        | 7.4        | 5.0          |
|             | RS-I-FGSM(Ours) | 58.0    | 51.0    | 98.9       | 41.1     | 15.2       | 12.6       | 8.5          |
|             | MI-FGSM         | 68.7    | 61.8    | 99.6       | 52.0     | 25.2       | 20.1       | 14.3         |
|             | RS-MI-FGSM(Ours)| 80.7    | 76.1    | 98.7       | 69.7     | 36.8       | 29.3       | 21.2         |
| Res-101     | I-FGSM          | 27.6    | 23.2    | 21.5       | 98.2     | 9.2        | 7.8        | 5.7          |
|             | RS-I-FGSM(Ours) | 42.0    | 35.6    | 34.1       | 96.9     | 13.4       | 12.1       | 6.9          |
|             | MI-FGSM         | 52.5    | 48.1    | 45.7       | 98.2     | 22.4       | 18.5       | 11.9         |
|             | RS-MI-FGSM(Ours)| 68.8    | 64.4    | 63.2       | 97.3     | 31.5       | 28.1       | 16.7         |

4.3. Attacking an ensemble of networks
Though the results in Table 1 show that RS-I-FGSM and RS-MI-FGSM can improve the transferability of adversarial examples on the black-box models, they are still relatively weak at attacking an adversarially trained network under the black-box setting. Therefore, we follow the strategy in [9] to attack multiple networks simultaneously in order to further boost the transferability. We consider all seven networks discussed above. Adversarial examples are crafted on an ensemble of four normally trained networks, and tested them on all seven networks, using I-FGSM, RS-I-FGSM, MI-FGSM, and RS-MI-FGSM, respectively. The number of iterations in the iterative method is $T = 10$, the perturbation size is $\epsilon = 16$, and the ensemble weights of networks are equal, i.e., $\omega_i = 1/4$.

The experimental results are summarized in Table 2, which shows that in the black-box settings, RS-MI-FGSM has higher attack success rates than the other methods. For example, the success rate of RS-MI-FGSM attacking Inc-v3ens3 is 53.5%, while those of I-FGSM, RT-I-FGSM, and MI-FGSM are 18.9%, 28.6%, and 38.4%, respectively. On challenging adversarially trained networks, it can be seen that the average success rate of RS-MI-FGSM for black-box attacks is 45.1%, which is 12.8% higher than that of MI-FGSM. These results show the effectiveness and advantages of our method.

Table 2. The success rates (%) of adversarial attacks against seven models under multi-model setting.
The adversarial examples are crafted for the ensemble of Inc-v3, Inc-v4, IncRes-v2, and Res-152. * indicates the white-box models being attacked.

| Attack               | Inc-v3* | Inc-v4* | IncRes-v2* | Res-101* | Inc-v3ens3 | Inc-v3ens4 | IncRes-v2ens |
|----------------------|---------|---------|------------|----------|------------|------------|--------------|
| I-FGSM               | 99.7    | 96.1    | 91.9       | 86.6     | 18.9       | 15.9       | 9.3          |
| RS-I-FGSM(Ours)      | 98.5    | 95.8    | 92.3       | 88.0     | 28.6       | 25.0       | 15.9         |
| MI-FGSM              | 99.8    | 97.7    | 95.1       | 91.1     | 38.4       | 36.1       | 22.4         |
| RS-MI-FGSM(Ours)     | 99.1    | 96.3    | 94.3       | 90.8     | 53.5       | 49.4       | 32.3         |
5. Conclusions
In this paper, we propose a new attack method, namely Random transformation of image Size Momentum Iterative Fast Gradient Sign Method (RS-MI-FGSM) to improve the transferability of adversarial examples. RS-MI-DFGSM aims to achieve data augmentation by randomly transforming the size of the input image at each iteration in the attack process. Compared with existing baseline attacks, the results on the ImageNet dataset demonstrate that our proposed attack method has much higher success rates for black-box models, and maintains similar success rates for white-box models. In addition, we use the method of attacking ensemble models simultaneously to further boost the transferability of adversarial examples. The results of this enhanced attack show that the average black-box attack success rate of RS-MI-FGSM on adversarially trained networks outperforms MI-FGSM by a large margin of 12.8%. Our work of RS-MI-FGSM suggests that other data augmentation methods may also be helpful to build strong attacks, which will be our future work, and the key is how to find effective data augmentation methods for iterative attacks. It is hoped that the proposed attack method can serve as a benchmark for evaluating the robustness of networks to adversaries and the effectiveness of different defense methods in future.

Acknowledgments
This work was supported by the National Key Research and Development Program of China under Grant no. 2017YFB0801900.

References
[1] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.
[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556.
[3] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 1–9.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.
[5] C. Szegedy et al., "Intriguing properties of neural networks," 2013, arXiv: 1312.6199. [Online]. Available: https://arxiv.org/abs/1312.6199.
[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, arXiv: 1412.6572. [Online]. Available: https://arxiv.org/abs/1412.6572.
[7] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016, arXiv: 1611.02770. [Online]. Available: https://arxiv.org/abs/1611.02770.
[8] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, arXiv: 1607.02533. [Online]. Available: https://arxiv.org/abs/1607.02533.
[9] Y. Dong et al., "Boosting adversarial attacks with momentum," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018, pp. 9185–9193.
[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv: 1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980.
[11] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in Proc. Int. Conf. Mach. Learn., Atlanta, GA, USA, 2013, pp. 1139–1147.
[12] T. Tieleman and G. Hinton, "Lecture 6.5 - RMSProp, coursera: Neural networks for machine learning," Tech. Rep., 2012.
[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge.International Journal of Computer Vision, 115(3):211–252,2015.
[14] C. Xie et al., "Improving transferability of adversarial examples with input diversity," in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 2725–2734.
[15] B. Biggio et al., "Evasion attacks against machine learning at test time," in Proc. Mach. Learn. Knowl. Discovery Databases, Berlin, Heidelberg, 2013, pp. 3876–402.
[16] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4312–4321, 2019.
[17] K. Eykholt et al., "Robust physical-world attacks on deep learning visual classification," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018, pp. 1625–1634.
[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladiu, "Towards deep learning models resistant to adversarial attacks," 2018, arXiv: 1706.06083. [Online]. Available: https://arxiv.org/abs/1706.06083.
[19] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in Proc. IEEE Symp. Secur. Privacy, San Jose, CA, USA, 2016, pp. 582–597.
[20] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In International Conference on Learning Representations, 2018.
[21] C. Guo, M. Rana, M. Cissé, and L. V. D. Maaten, "Countering adversarial images using input transformations," 2017, arXiv: 1711.00117. [Online]. Available: https://arxiv.org/abs/1711.00117.
[22] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, arXiv: 1611.01236. [Online]. Available: https://arxiv.org/abs/1611.01236.
[23] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2017, arXiv: 1705.07204. [Online]. Available: https://arxiv.org/abs/1705.07204.
[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 2818–2826.
[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," presented at the AAAI Conf. Artif. Intell., San Francisco, California, USA, 2017.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in Proc. Comput. Vision – ECCV 2016, Cham, 2016, pp. 630–645.