An Adversarial Example Attack Method Based on Ensemble Learning

Zebin Zhuang[1], **Chenliang Luo**[2], **Kaiwei Lan**[3]

[1]30920241154588, [2]30920241154554, [3]30920241154551

Abstract

This paper aims to enhance the success rate of adversarial example attacks. The reason for choosing this issue lies in the challenges that deep neural networks (DNNs) face due to adversarial example attacks. Such attacks generate models by adding imperceptible noise to legitimate samples, thereby resulting in inaccurate prediction results desired by attackers. The research on adversarial samples is of great significance in safety-critical fields such as autonomous driving.However, although the attack success rates in white-box environments are relatively high, they cannot accurately reflect real-world conditions. Moreover, the attack success rates in black-box scenarios remain unsatisfactory. The current state-of-the-art methods include integrating iterative strategies based on the Fast Gradient Sign Method (FGSM) and introducing the concept of momentum to improve the ability to escape from poor local optima, among others. Based on previous research, this paper first replicates existing attack methods under the environment with limited computing power. Subsequently, it investigates the effects of various attack algorithms in the environment with limited computing power. Finally, we will adopt simple ensemble learning to improve the attack success rate.

1.Introduction

1.1What is Adversarial examples?

Adversarial examples are a significant concept in the field of machine learning. They are crafted by adding imperceptible perturbations to legitimate input samples. The purpose of creating adversarial examples is to mislead deep neural networks (DNNs) into making incorrect predictions. These adversarial perturbations are designed in such a way that they are barely noticeable to human observers but can cause significant changes in the output of the neural network. Adversarial examples play a crucial role in evaluating the robustness and security of machine learning models. They help identify vulnerabilities in the models and can be used for adversarial training to enhance the resilience of the models against malicious attacks. Moreover, they contribute to a better understanding of the inner workings of DNNs and promote the development of more interpretable artificial intelligence. In critical applications such as autonomous driving and healthcare, the study of adversarial examples is essential to ensure the safety and reliability of machine learning systems.

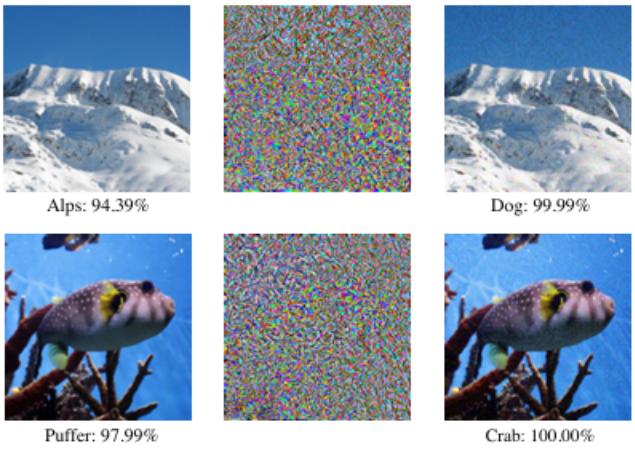

Figure 1: Introduction to adversarial examples

1.2Why are adversarial examples important?

Adversarial examples are crucial in machine learning for assessing model robustness and security. They serve as essential tools for identifying vulnerabilities, facilitating adversarial training to enhance a model's resilience against malicious inputs. Additionally, they contribute to the diversification of training data, improving generalization capabilities. Research on adversarial examples also aids in understanding the internal mechanisms of deep learning models, advancing the field of interpretable artificial intelligence. Ultimately, their implications extend to critical applications across various industries, underscoring the necessity for robust and reliable systems.

1.3What challenges does the field of adversarial examples face?

The field of adversarial example attacks faces many challenges. On the one hand, it is necessary to balance the effectiveness and concealment of attacks. Ensure high success rate while maintaining similarity with the original sample

so as not to be detected. On the other hand, the diversity and complexity of models increase the difficulty of attacks. Different types and models with complex structures require specific attack methods, and attackers find it difficult to understand their internal mechanisms. In black-box attacks, information acquisition is difficult and generalization ability is limited. When extending attacks from the digital world to the physical world, environmental factors affect the attack effect, and practical physical limitations and feasibility also need to be considered. In summary, the field of adversarial examples is facing the following challenges:

- Although the attack success rate in white-box environments is high, it does not accurately reflect real-world conditions.
- The attack success rate in black-box scenarios remains insufficient.
- Effectively enhancing the transferability of adversarial examples continues to be a significant challenge.

2.Related Work

By understanding the structure and parameters of a given model, several methods can successfully generate adversarial examples in a white-box setting, such as L-BFGS[1], Fast Gradient Sign Method[2], and iterative variants based on gradient methods[5]. However, in black-box scenarios, a significant issue is their poor transferability[1, 3, 4]; that is, adversarial examples designed for one model may not retain their adversarial properties when applied to other models, which undermines the practicality of black-box attacks and raises genuine security concerns. The phenomenon of transferability arises from the fact that different machine learning models learn similar decision boundaries around data points, allowing adversarial examples crafted for one model to be effective against others as well.

In the realm of black-box attacks, numerous scholars are currently conducting research. For instance, ensemble adversarial training[6] has significantly enhanced the robustness of deep neural networks, rendering most existing methods unable to successfully attack them in a black-box manner. This phenomenon can largely be attributed to the trade-off between attack capability and transferability. Papernot et al. [7] employed adaptive queries to train surrogate models that sufficiently capture the behavior of the target model, thereby transforming black-box attacks into white-box attacks. However, this approach necessitates complete prediction confidence provided by the target model and a substantial number of queries, particularly for large-scale datasets such as ImageNet[8]. Such requirements are impractical in real-world applications.

2.1FGSM

The Fast Gradient Sign Method (FGSM) is a well - known adversarial attack technique. It quickly generates adversarial examples by adding a small perturbation to the input data in the direction of the gradient sign of the loss function. The magnitude of the perturbation is controlled by a parameter. It's a simple yet effective way to expose the vulnerability of machine - learning models.

2.2I-FGSM

Iterative Fast Gradient Sign Method (I-FGSM) is an advanced adversarial attack method. I-FGSM conducts iterative attacks by repeatedly applying small perturbations in the direction of the gradient sign. At each iteration, it takes the previously perturbed sample as the starting point and computes the gradient with respect to the current sample to determine the direction of the perturbation. This iterative process allows for more targeted and effective attacks compared to FGSM. It can generate adversarial examples that are more likely to mislead the target model. Moreover, by controlling the step size and the number of iterations, one can fine-tune the strength and effectiveness of the attack. I-FGSM has shown significant effectiveness in attacking various deep learning models and has raised concerns about the robustness and security of these models.

2.3MI-FGSM

Momentum Iterative Fast Gradient Sign Method (MI - FGSM) is an advanced adversarial attack method. It's based on I - FGSM and uses a momentum term. The momentum helps accumulate gradient directions over iterations, allowing for more stable and effective attacks. It overcomes some limitations of I - FGSM, like getting stuck in local optima, and can better deceive machine - learning models.

3.Proposed Solution

Based on previous research, this paper first replicates the existing attack methods. Subsequently, it investigates the effects of various attack algorithms in the environment with limited computing power. Although white-box attacks demonstrate a relatively high success rate, this is often not the case in real-world environments. Black-box attacks are more in line with practical scenarios, and many scholars have already explored this field. Finally, we will adopt simple ensemble learning to improve the attack success rate. The plan of this paper is as follows:

1. Verify whether the advanced attack methods can work under the condition of low computing power.
2. Compare the replicated results with the advanced results in the paper.
3. Employ simple ensemble learning to improve the accuracy rate of adversarial samples in black-box attacks.

4.Experiments

In this section, experiments were conducted on the ImageNet dataset to verify the effectiveness of the proposed method. Firstly, the experimental setup will be elaborated in detail in Section 4.1. Subsequently, we will report the results of attacks carried out by various algorithms against a single model in Section 4.2, and present the results of attacks targeting model ensembles in Section 4.3.

4.1.Setup

At present, the research on the security of deep learning models has attracted significant attention. We focus on four typical models for in-depth exploration. Among them,

	VGG	ALEXNET	SQUEEZENET
FGSM	327	364	346
I-FGSM	345	388	365
MI-FGSM	456	480	471

Table 1: Attacking a single model

the three conventionally trained models, namely VGG11, ALEXNET, and SQUEEZENET,each have their own highlights. VGG11 is adept at precisely extracting image features by relying on its regular convolutional layer stacking. ALEXNET, as a pioneer, has revolutionized learning efficiency with large convolutional kernels and the ReLU function. SQUEEZENET focuses on being lightweight and can guarantee classification performance under limited computing power. There is also a model constructed through ensemble technology, which integrates the advantages of multiple sub-models and exhibits better stability and stronger anti-interference ability in complex image environments. To ensure the effectiveness of experimental data, models must first be able to correctly classify the original images. Otherwise, it is meaningless to study the attack success rate. For this purpose, we target the ILSVRC 2012 validation set and use a random sampling procedure to accurately select 1,000 images from its vast collection covering 1,000 categories. After strict verification, these images can all be correctly classified by the above four models.

During the experimental stage, we mainly compare the momentum-based method, the single-step gradient-based method, and the iterative method. The momentum-based method introduces the concept of physical momentum and can reduce the interference of local gradients during optimization, thus steadily seeking the optimal solution. The single-step gradient method is simple and straightforward, quickly adjusting parameters according to the current gradient, but it is prone to getting trapped in local optima. The iterative method refines the results through multiple iterations. As for the optimization-based method, since it is difficult to control the distance between adversarial examples and real examples, it cannot be directly compared with our method. Therefore, it is not included in this core comparison. We are fully committed to digging deep into the differences and advantages among the former three methods to promote the advancement of model attack and defense technologies.

4.2.Attacking a single model

We report in Table 1 the success rates of attacks against the models under our study. Adversarial examples are generated for VGG11, ALEXNET, and SQUEEZENET respectively by employing the Fast Gradient Sign Method (FGSM), Iterative Fast Gradient Sign Method (I-FGSM), and Momentum Iterative Fast Gradient Sign Method (MI-FGSM). The success rate refers to the misclassification rate of the corresponding models when taking adversarial images as inputs. In all experiments, the maximum perturbation is set to 16, with the pixel value ranging within [0, 255]. The number of iterations for both the Iterative Fast Gradient Sign Method (I-FGSM) and the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) is 10. From the table, we can observe that the Momentum Iterative Fast Gradient Sign Method (MI-FGSM), like the Iterative Fast Gradient Sign Method (I-FGSM), remains a powerful white-box adversary since it can attack white-box models with a success rate close to 100On the other hand, it can be seen that by integrating momentum, our proposed Momentum Iterative Fast Gradient Sign Method (MI-FGSM) significantly outperforms both the Fast Gradient Sign Method (FGSM) and the Iterative Fast Gradient Sign Method (I-FGSM) in black-box attacks, which demonstrates the effectiveness of the proposed algorithm.

4.3. White-box attack

In the research field of adversarial examples, white-box attacks are a crucial and thorny type of attack, attracting significant attention both in academic discussions and practical application scenarios. What distinguishes them from other attack methods is that attackers possess all the detailed information about the target machine learning model, including the overall architecture design, specific internal parameter settings, the algorithm logic used in model training, and how input data flows and output results are generated. With such comprehensive information at their disposal, attackers are empowered with extremely powerful attacking capabilities, posing a huge threat to model security. When launching white-box attacks, attackers will fully utilize the known model information and carry out attacking actions in an orderly manner. Due to their familiarity with the internal structure and parameters of the model, attackers can accurately locate the weak links of the model. For example, in the task of image classification, they know which combinations of neurons are responsible for recognizing specific categories and which layers play a key role in feature extraction, and thus can skillfully target these critical parts. During the attack process, gradient information plays a pivotal role. Although specific formulas are not delved into here, the principle is that, relying on their understanding of the model, attackers can efficiently calculate the key gradient changes

	VGG	ALEXNET	SQUEEZENET
FGSM	357	394	376
I-FGSM	379	423	396
MI-FGSM	493	520	509

Table 2: Attacking an ensemble of models

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	FGSM	72.3*	28.2	26.2	25.3	11.3	10.9	4.8
	I-FGSM	**100.0***	22.8	19.9	16.2	7.5	6.4	4.1
	MI-FGSM	**100.0***	**48.8**	**48.0**	**35.6**	**15.1**	**15.2**	**7.8**
Inc-v4	FGSM	32.7	61.0*	26.6	27.2	13.7	11.9	6.2
	I-FGSM	35.8	**99.9***	24.7	19.3	7.8	6.8	4.9
	MI-FGSM	**65.6**	**99.9***	**54.9**	**46.3**	**19.8**	**17.4**	**9.6**
IncRes-v2	FGSM	32.6	28.1	55.3*	25.8	13.1	12.1	7.5
	I-FGSM	37.8	20.8	**99.6***	22.8	8.9	7.8	5.8
	MI-FGSM	**69.8**	**62.1**	99.5*	**50.6**	**26.1**	**20.9**	**15.7**
Res-152	FGSM	35.0	28.2	27.5	72.9*	14.6	13.2	7.5
	I-FGSM	26.7	22.7	21.2	**98.6***	9.3	8.9	6.2
	MI-FGSM	**53.6**	**48.9**	**44.7**	98.5*	**22.1**	**21.7**	**12.9**

Table 3: Attacking an ensemble of models

when the model processes inputs. They know which dimensions of the input to adjust and in which direction to make slight modifications to interfere with the normal judgment of the model. After obtaining the key information, attackers set about carefully constructing adversarial examples. Through a series of complex operations, they make subtle changes to the originally normal samples to generate adversarial examples. In the field of images, adversarial examples are almost indistinguishable from the original images at first glance, perhaps with only extremely subtle adjustments to a few pixels. However, it is precisely these slight differences that are enough to mislead the model. For instance, an image that was originally accurately identified as a "cat" will be firmly misclassified as a "dog" by the model after being processed into an adversarial example through a white-box attack, completely deviating from the correct result. White-box attacks have several advantages, making their destructive power not to be underestimated. In terms of efficiency, compared with black-box attacks that grope in the dark, repeatedly probing and querying a large number of model characteristics to find attack breakthroughs, white-box attacks are like explorers with a map, directly hitting the vital parts. Relying on the known model information, attackers can often quickly lock in an attack strategy and achieve the goal of making the model make mistakes with less time and computing power. In addition, accuracy is also a major strength of white-box attacks. Attackers can precisely identify the sensitive areas of the model and determine which details of the input, once changed, will seriously interfere with the model's output. Thus, they skillfully apply tiny perturbations to these critical dimensions, just like placing a feather on a precision balance, yet enough to disrupt the balance and greatly increase the probability of a successful attack. However, white-box attacks also bring unprecedented challenges to defense work. Traditional defense methods, such as restricting external access to model information and encrypting parts of the model structure, are ineffective in the face of white-box attacks. After all, attackers already have all the model information, and these simple protection measures simply cannot stop their offensive steps. Currently, defenders have to explore other paths and seek more sophisticated and elaborate defense technologies. For example, adversarial training involves integrating adversarial examples into the normal training process to expose the model to various attack means in advance and enhance its resistance. There is also the gradient masking technique, which hides the real gradient information to prevent attackers from taking advantage of it. However, the confrontation between offense and defense is endless. While defense technologies are constantly upgrading, new white-box attack methods are also emerging continuously, constantly testing and challenging the security line of models.As shown in Table 3, under white-box attacks, the success rate of the attacks is extremely high. For the MI-FGSM algorithm, it can even approach 100

4.4. Black-box attack

Under the black-box condition, attacks become more difficult. In the realm of adversarial examples research, the extreme difficulty of black-box attacks stems from multiple

factors. Firstly, attackers struggle to access crucial information of the target model, including structural details like the number of layers, neuron layouts per layer, and inter-layer connections, as well as parameter settings such as weight matrices and bias vectors. Even details about the training dataset and algorithm remain unknown. For instance, in the autonomous driving object recognition model, due to commercial secrecy and security concerns, its internal structure and training data are withheld. Secondly, existing defense mechanisms have upped the ante for black-box attacks. Many practical models employ complex defenses. Some preprocess inputs to filter out abnormal pixel patterns via statistical principles, while adversarial training incorporates adversarial examples during model training, endowing models with the ability to resist attacks. Without understanding these defenses, carefully crafted adversarial examples often get intercepted before reaching the model's core decision-making. Moreover, the limited information obtained from the model in black-box attacks curtails attack effectiveness. Unlike white-box attacks leveraging gradient information for precise adversarial example generation, black-box attacks only yield basic outputs like classification results and probability estimates. The process of creating adversarial examples demands a high level of precision and finesse. It requires delicate modifications to the original text, alterations that are meticulously calibrated to deceive the model without being overly conspicuous. But due to the attackers' lack of insight into the model's focus, they struggle to identify the optimal points of intervention within the text. They cannot ascertain which words or phrases, if tweaked ever so slightly, would trigger the model to misclassify the sentiment, making it an arduous task to engineer those subtly deceptive and misleading adversarial examples that could potentially undermine the integrity and reliability of the text sentiment classification model. Finally, significant differences between models and their inherent uncertainties impede black-box attacks. Varied model architectures and training paradigms lead to diverse decision boundaries and feature extraction logics. Even with prior attack experience, attackers often find past strategies ineffective against new, unknown models. Additionally, the influence of data noise and random initialization parameters renders model behavior unpredictable, further complicating black-box attacks.As shown in Table 3, under black-box attacks, the success rate of the attacks is extremely low, and there is a significant difference compared with the effect of white-box attacks.

4.5. Attacking an ensemble of models

In this section, a comprehensive comparison of the ensemble methods for attacks has been meticulously conducted. Our research scope encompasses three prominent models, namely VGG11, ALEXNET, and SQUEEZENET, which have been widely recognized and studied in the field of deep learning. In the course of our experiments, we carried out targeted attacks on the ensemble constructed by integrating these three models. To be specific, we employed three distinct yet influential attack methods: the Fast Gradient Sign Method (FGSM), the Iterative Fast Gradient Sign Method (I-FGSM), and the advanced Momentum Iterative Fast Gradient Sign Method (MI-FGSM), respectively. During the experimental setup, we meticulously calibrated the parameters to ensure the reliability and comparability of the results. The maximum perturbation was deliberately set to 16, which functions as a crucial constraint dictating the extent of alterations permissible on the input data. Additionally, for both the Iterative Fast Gradient Sign Method (I-FGSM) and the Momentum Iterative Fast Gradient Sign Method (MI-FGSM), the number of iterations was uniformly fixed at 10, aiming to standardize the iterative process and maintain consistency across different attack scenarios. Simultaneously, we adopted an equal weighting strategy for the ensemble models, granting each component model an equivalent influence within the integrated framework, thereby eliminating potential biases caused by uneven weight distribution. The outcomes of these elaborate experiments are methodically presented in Table 2, which serves as a crucial repository of our empirical findings. Upon close examination of Table 2, a remarkable observation comes to the fore: the adversarial examples generated by the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) exhibit a strikingly high transfer rate. This feature endows them with the remarkable ability to execute potent black-box attacks, effectively circumventing the target model's defenses without prior knowledge of its internal structure, thus posing a significant challenge to the security and robustness of deep learning models. Overall, these findings not only shed light on the performance disparities among different attack methods but also underline the potential threats that the MI-FGSM approach might pose to real-world applications reliant on deep learning systems, urging further investigations into enhanced defense mechanisms.

Conclusion

In this paper, we reproduce the momentum - based iterative methods to enhance adversarial attacks under the condition of weak computing power. These methods can effectively deceive white - box models and black - box models. Our methods always outperform the one - step gradient - based methods and vanilla iterative methods in the black - box attack mode. We have carried out a large number of experiments to verify the effectiveness of the proposed methods and explain why they are effective in practical applications. In order to further improve the transferability of the generated adversarial examples, we propose to attack a simple ensemble model. Specifically, from the perspective of surrogate models, if only a single surrogate model is considered, the adversarial examples generated with the VGG11 classifier as the surrogate model usually have poor transferability in attacks, while those generated with Alexnet have higher transferability. It is hypothesized that the decision boundaries of Alexnet and the defense models in Kaggle are more similar. If an ensemble learning model is used as the surrogate model, the transferability of the generated adversarial examples will be improved. From the perspective of attack algorithms, as the attack algorithms become more advanced, the update capabilities of the algorithms become more stable, the abilities to escape from poor local optima become stronger, and the attack success rates become higher.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus.Intriguing properties of neural networks.In ICLR, 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy.Explaining and harnessing adversarial examples.In ICLR, 2015.

[3] Y. Liu, X. Chen, C. Liu, and D. Song.Delving into transferable adversarial examples and black-box attacks.In ICLR, 2017.

[4] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard.Universal adversarial perturbations.In CVPR, 2017.

[5] A. Kurakin, I. Goodfellow, and S. Bengio.Adversarial examples in the physical world.arXiv preprint arXiv:1607.02533, 2016.

[6] A. Kurakin, I. Goodfellow, and S. Bengio.Adversarial examples in the physical world.arXiv preprint arXiv:1607.02533, 2016.F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel.Ensemble adversarial training: Attacks and defenses.In ICLR, 2018.

[7] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami.Practical black-box attacks against machine learning.In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al.