Research on Adversarial Attack Algorithm Based on AI Recognition

Si Zhang[1*]

[1] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, 518000, China

Corresponding Email: zhangsi@szu.edu.cn

https://doi.org/10.70695/prtypb17

Abstract

Deep learning-based artificial intelligence algorithms are widely used in critical areas such as autonomous driving and medical diagnosis. However, the lack of interpretability of deep neural networks results in unpredictable prediction outcomes, posing significant security threats to AI applications and deployments. Adversarial examples, specially designed to introduce imperceptible perturbations, cause neural network models to produce confused and erroneous predictions. Therefore, it is crucial to explore both adversarial example generation and attack algorithms to understand the security of deep neural networks and enhance the interpretability of deep learning models. Existing adversarial example generation algorithms for image recognition still face issues such as low generation efficiency, poor sample quality, and unstable transferability during adversarial attacks. This study proposes HDSAttack, a transferable adversarial attack algorithm that maps low-dimensional dense information to high-dimensional sparse information, thereby enhancing transferability. To address the problem of unstable transferability in existing adversarial attacks, this paper suggests mapping samples from a low-dimensional dense input space to a high-dimensional latent space to expand the search space and obtain more effective information. Additionally, KL divergence is used to enforce sparsity constraints throughout the training process, yielding linearly separable high-dimensional sparse information for efficient information search. Further, an ensemble attack on multiple target networks is conducted to enable the search network to learn more about neural network structures, improving the transferability of adversarial examples. Experimental results show that, compared to traditional hourglass autoencoder structures, the proposed search network structure enhances the transfer attack success rate by 10.39%.

Keywords Deep learning; artificial intelligence; secure image recognition; adversarial samples

1 Introduction

As a crucial component of new infrastructure, Advanced AI computing capabilities, represented by AI computing centers, have become a pivotal foundation for the development of the digital economy [1]. Throughout the entire lifecycle of a model, different stages present corresponding security risks. Although various measures can partially mitigate these risks, internal algorithmic imperfections and unresolved issues, particularly the lack of interpretability in deep learning algorithms, continue to hinder further development [2].

In recent years, adversarial examples have become a hot research topic, with some progress made in adversarial attack research focused on image AI recognition. Szegedy et al. experimentally discovered that the units in the final layer of neural networks form a solid basis for extracting semantic information [3]. Goodfellow et al. further investigated the nature and properties of adversarial examples, attributing the vulnerability of deep neural networks to the model's local linear characteristics, and found that adversarial examples exhibit unstable transferability across different target classifiers [4].

However, adversarial example generation methods that involve iteratively modifying inputs to achieve maximum classification loss require extensive iterative updates to the samples themselves, resulting in each adversarial example needing a considerable amount of generation time [5]. When faced with the need to produce adversarial examples in bulk, these iterative methods can lead to significant time consumption. Due to their flexibility and mappability, Generative Adversarial Networks (GANs) have been effectively applied to rapidly generate adversarial examples. Xiao et al. proposed using AdvGAN to generate adversarial examples, where the generator maps the original input samples to adversarial

perturbations, and the discriminator determines whether the image with the added adversarial perturbations is an adversarial example [6].

This study proposes HDSAttack, a transferable adversarial attack algorithm based on high-dimensional sparse mapping. By optimizing the search network structure, this method achieves adversarial examples with better transferability, thus enhancing the transferability of adversarial attacks. Unlike traditional hourglass-shaped (i.e., dimensionality reduction followed by dimensionality expansion) autoencoder structures, this chapter introduces a high-dimensional mapping search network that first maps input information to a high-dimensional space. By utilizing KL divergence to optimize the loss function, this model performs constrained updates during the gradient descent process based on neuronal relationships, enhancing search efficiency and ultimately yielding sparse information with more logical relationships. The adversarial examples generated through this process exhibit stronger transferability, offering potential for further research in black-box adversarial attacks.

2 Methodology

The study proposes to abandon the traditional downsampling design in the feature extraction process, instead directly expanding the dimensionality of input images within shallow neural networks. This approach enlarges the search space for adversarial perturbation information, allowing for the acquisition of more effective information and thereby enhancing the transferability of adversarial examples, which facilitates the implementation of black-box attacks. To further improve transferability, a composite loss function is designed, consisting of the loss functions required for each attack process against multiple target networks. Consequently, the search network updates its parameters based on feedback from multiple networks, enabling the search for adversarial perturbation distributions that incorporate information from multiple attacked networks.

KL divergence is a commonly used method for measuring the difference between two probability distributions. For a given dataset X that follows two probability distributions, P and

$$\text{Q: } KL(P \,||\, Q) = \sum_{i \in X} P(i) \bullet \log(\frac{P(i)}{Q(i)}) \tag{1}$$

Adding KL divergence as a penalty term to the loss function can achieve sparsity limitation. Given an input x, where $a_j(x)$ represents the activation value of hidden layer neuron j, the formula for the average activation value of j is:

$$\hat{\rho}_j = \frac{1}{M} \sum_{m=1}^{M} a_j(x^{(m)}) \tag{2}$$

Where M represents the number of samples in the dataset, and m represents the index number of each sample. The sparsity parameter ρ is a custom parameter, which is a value close to 0 (taken as 0.05 in the implementation process of this project). In loss function, the average activation value $\hat{\rho}_j$ is expected to approach ρ to achieve sparsity limitation. Therefore, a penalty term based on KL divergence is added to the loss function to penalize cases where the difference between $\hat{\rho}_j$ and ρ is too large, ensuring that the average activation value of hidden neurons is limited to a smaller range. KL divergence is used to measure the difference between $\hat{\rho}_j$ and ρ :

$$KL(\rho \,||\, \hat{\rho}_j) = \sum_{i \in X} \rho \bullet \log(\frac{\rho}{\hat{\rho}_j}) \tag{3}$$

When $\hat{\rho}_j = \rho$, divergence value of KL ($\hat{\rho}_j \,||\, \rho$) is 0, gradually increase as difference between ρ and $\hat{\rho}_j$ increases. Therefore, minimizing the penalty term KL can produce the effect of bringing $\hat{\rho}_j$ closer to ρ, in order to achieve sparsity limitation on neural networks. Sparsity limitation is beneficial for the sparse representation of input information in high-dimensional space. More specifically, it can activate effective neurons and avoid the impact of redundant information on gradient feedback. As shown in

Figure 1, the bold red circles and lines represent the redundant neurons that are suppressed under the sparsity constraint, while the remaining sparse effective neurons work normally, which is beneficial for more effective model training or search processes.

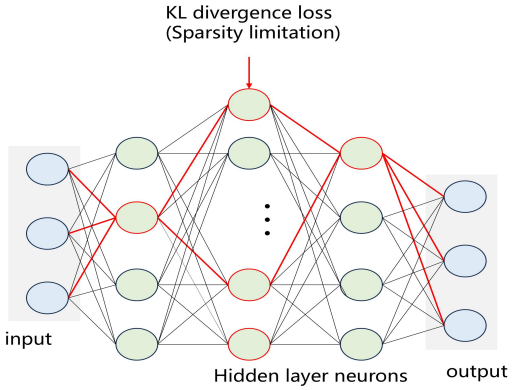

Fig. 1. KL divergence loss achieves sparsity limitation

Two transferability exploration experiments were conducted, comparing the adversarial samples generated by the CCL-Adv method before high-dimensional sparse mapping with those generated by the HDSAttack algorithm after high-dimensional sparse mapping, in terms of their transferability during adversarial attacks. In the comparison of model settings, the search network structure before high-dimensional sparse mapping utilized an implicit space feature search network, whereas the search network structure after high-dimensional sparse mapping employed a high-dimensional mapping search network.

The transferability of the universal adversarial samples was tested on the four target classifiers and extended to other classifiers, including AlexNet, DenseNet, ResNet-152, and ResNet-34. Additionally, the transferability of the generated universal adversarial samples was compared with that of FGSM and PGD. In the black-box attack comparison experiments, FGSM and PGD adversarial samples were first generated through white-box attacks on ResNet-50 and then used for subsequent black-box attacks.

The experimental environment of this paper is shown in Table 1.

Table 1. HDSAttack algorithm experimental environment

Environmental type	Parameters
Hardware environment	GPU: NVIDIA GeForce RTX 2080Ti CPU: Intel(R) Xeon(R) Bronze 3104 CPU @ 1.70GHz
Software environment	Pytorch, Opencv, PIL, Numpy
Language	Python 3.6
Objection	ResNet-50, VGG-16, GoogleNet, MobileNet-v2
Comparison algorithm	FGSM, PGD
Dataset	ImageNet-1000

3 Results

Tables 2 and 3 respectively present the transferability test results of adversarial samples generated by the CCL-Adv method before and by the HDSAttack algorithm after high-dimensional sparse mapping. The phrase "does not significantly demonstrate" is awkward; consider revising to: "A comparison between the two tables reveals that the high-dimensional sparse mapping structure does not significantly enhance white-box attack performance on the diagonal, with the average white-box attack success rate across the four target networks increasing from 96.36% to 97.99%. However, the black-box attack success rates in the off-diagonal regions do reflect the advantages of the high-dimensional sparse mapping structure proposed by the HDSAttack algorithm, with the overall average black-box attack success rate improving from 53.96% to 64.35%. This indicates better transferability of adversarial attacks. Although the high-dimensional sparse mapping structure enhances adversarial attack transferability to some extent, the overall transfer attack success rate still exhibits instability related to the similarity

between the source model and target model structures. Therefore, based on the research work in this chapter, an ensemble attack across multiple target models is introduced to generate universal adversarial samples, further improving the transferability of adversarial attack algorithms.

Table 2. Migration between models before high-dimensional sparse mapping (represented by attack success rate (%))

migrate	MobileNet-v2	ResNet-50	VGG-16	GoogleNet
MobileNet-v2	**97.22**	42.58	44.39	10.76
ResNet-50	59.08	**96.54**	81.25	80.96
VGG-16	62.33	45.72	**93.18**	24.89
GoogleNet	61.08	57.57	76.97	**98.52**

Table 3. Migration between models after high-dimensional sparse mapping (represented by attack success rate (%))

migrate	MobileNet-v2	ResNet-50	VGG-16	GoogleNet
MobileNet-v2	**99.44**	50.07	58.29	17.09
ResNet-50	80.53	**98.13**	93.57	82.05
VGG-16	77.45	49.00	**98.31**	33.62
GoogleNet	71.29	72.10	87.22	**96.06**

Further explore the attack capability and transfer performance of general adversarial samples obtained from integrated attacks on multiple target networks. Not only did we compare the performance of the general adversarial samples obtained from high-dimensional sparse mapping before and after ensemble attacks, but we also compared them with classical adversarial attack algorithms FGSM and PGD.

The results in Tables 4-5 indicate that the universal adversarial samples generated by the proposed algorithm exhibit higher and more stable transferability between models. Since universal adversarial samples are generated by attacking all four networks simultaneously, these samples contain generalized information relevant to the structures of the four models. As a result, the adversarial features are more likely to include generalized structural information applicable to other models and thus are more easily transferable to models sharing this common structural information. Experimental results show that while the transfer success rate of universal adversarial samples before high-dimensional sparse mapping is still lower than that of samples after high-dimensional sparse mapping, it still demonstrates some improvement in transferability compared to single-target attacks. Moreover, universal adversarial samples after high-dimensional sparse mapping achieve an attack success rate of no less than 95% on white-box attacks against the four target classifiers and a success rate of no less than 75% on black-box attacks against other models with similar structures.

This study also compares the proposed algorithm with classic white-box attack algorithms FGSM and PGD in terms of transferability. Table 4 presents the white-box attack success rates of FGSM and PGD on the four target networks for single-target attacks, while Table 5 shows the black-box attack success rates of these two algorithms on the same networks. The results indicate that PGD achieves the highest attack success rate for white-box attacks across the four target models. However, universal adversarial samples, which contain structural information from all four target networks, lead to reduced specificity for each network, resulting in relatively weaker white-box attack success rates compared to PGD. On the other hand, the black-box attack success rates presented in Table 5 demonstrate the superior transferability of universal adversarial samples. Thus, this chapter adopts the concept of ensemble attacks and shows that algorithms searching for networks trained on multiple target networks simultaneously can effectively enhance adversarial attack transferability. Further analysis of the transferability comparison between FGSM and PGD reveals that adversarial attack algorithms with stronger white-box attack success rates tend to exhibit weaker transferability. This observation is consistent with the phenomenon where universal adversarial samples show relatively weaker specificity for individual target models but stronger transferability.

Table 4. Generic adversarial sample transferability for white box attack (represented by attack success rate (%))

White box adversarial attack	MobileNet-v2	ResNet-50	VGG-16	GoogleNet
Before high-dimensional sparse mapping	94.29	96.01	96.97	94.98
After high-dimensional sparse mapping	96.22	96.66	97.33	95.44
FGSM (Single Target)	70.82	71.52	68.36	70.69
PGD (Single Target)	**98.64**	**99.27**	**99.8**	**98.12**

Table 5. Generic adversarial sample transferability for black box attack (represented by attack success rate (%))

Black box adversarial attack	AlexNet	DenseNet	ResNet-152	ResNet-34
Before high-dimensional sparse mapping	67.35	79.88	74.66	72.95
After high-dimensional sparse mapping	**75.74**	**87.48**	**80.90**	**89.31**
FGSM	22.56	35.29	20.20	43.86
PGD	15.32	21.67	12.04	25.54

4 Conclusion

The HDSAttack algorithm proposed here enhances the transferability of adversarial samples by mapping input data to a high-dimensional space via high-dimensional mapping search networks. This expansion of the adversarial distribution search space exposes more effective information, improving the transferability of adversarial samples during the attack process. To address the issue of reduced search efficiency due to the enlarged search space, the algorithm employs KL divergence as a medium to impose sparsity constraints during training, thereby obtaining high-dimensional sparse information and identifying effective connections within the vast neural network structure to enhance search efficiency. Furthermore, this chapter describes using the search network to simultaneously obtain feedback from multiple target models, guiding the entire search process and uncovering adversarial perturbation distributions enriched with structural information from various models. Adversarial samples generated from these universal adversarial perturbation distributions demonstrate improved and more stable transferability.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural Networks. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.
2. He K， Zhang X， Ren S， et al. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016. 770-778.
3. Szegedy C, Zaremba W， Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.
4. Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

5. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine, 2019, 380(14):1347-1358.
6. Inkawhich N , Liang K , Carin L, et al. Transferable perturbations of deep feature distributions. International Conference on Learning Representations. ICLR, 2020. 1-14.