Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Information Processing	
Series Title		
Chapter Title	Delving into Diversity in Substitute Ensembles and Transferability of Adversarial Examples	
Copyright Year	2018	
Copyright HolderName	Springer Nature Switzerland AG	
Author	Family Name	**Hang**
	Particle	
	Given Name	**Jie**
	Prefix	
	Suffix	
	Role	
	Division	School of Computer Science and Technology
	Organization	Nanjing University of Posts and Telecommunications
	Address	Nanjing, China
	Email	
Author	Family Name	**Han**
	Particle	
	Given Name	**KeJi**
	Prefix	
	Suffix	
	Role	
	Division	School of Computer Science and Technology
	Organization	Nanjing University of Posts and Telecommunications
	Address	Nanjing, China
	Email	
Corresponding Author	Family Name	**Li**
	Particle	
	Given Name	**Yun**
	Prefix	
	Suffix	
	Role	
	Division	School of Computer Science and Technology
	Organization	Nanjing University of Posts and Telecommunications
	Address	Nanjing, China
	Email	liyun@njupt.edu.cn

Abstract	Deep learning (DL) models, e.g., state-of-the-art convolutional neural networks (CNNs), have been widely applied into security-sensitivity tasks, such as facial recognition, automated driving, etc. Then their vulnerability analysis is an emergent topic, especially for black-box attacks, where adversaries do not know the model internal architectures or training parameters. In this paper, two types of ensemble-based black-box attack strategies, *iterative cascade ensemble strategy* and *stack parallel ensemble strategy*, are proposed to explore the vulnerability of DL system and potential factors that contribute to the high-

efficiency attacks are examined. Moreover, two pairwise and non-pairwise diversity measures are adopted to explore the relationship between the diversity in substitutes ensembles and transferability of crafted adversarial examples. Experimental results show that proposed ensemble adversarial attack strategies can successfully attack the DL system with ensemble adversarial training defense mechanism and the greater the diversity in substitute ensembles enables stronger transferability.

Keywords (separated by '-')	Black-box attack - Vulnerability - Ensemble adversarial attack - Diversity - Transferability

Delving into Diversity in Substitute Ensembles and Transferability of Adversarial Examples

Jie Hang, KeJi Han, and Yun Li(✉)

School of Computer Science and Technology,
Nanjing University of Posts and Telecommunications, Nanjing, China
liyun@njupt.edu.cn

Abstract. Deep learning (DL) models, e.g., state-of-the-art convolutional neural networks (CNNs), have been widely applied into security-sensitivity tasks, such as facial recognition, automated driving, etc. Then their vulnerability analysis is an emergent topic, especially for black-box attacks, where adversaries do not know the model internal architectures or training parameters. In this paper, two types of ensemble-based black-box attack strategies, *iterative cascade ensemble strategy* and *stack parallel ensemble strategy*, are proposed to explore the vulnerability of DL system and potential factors that contribute to the high-efficiency attacks are examined. Moreover, two pairwise and non-pairwise diversity measures are adopted to explore the relationship between the diversity in substitutes ensembles and transferability of crafted adversarial examples. Experimental results show that proposed ensemble adversarial attack strategies can successfully attack the DL system with ensemble adversarial training defense mechanism and the greater the diversity in substitute ensembles enables stronger transferability.

AQ1

Keywords: Black-box attack · Vulnerability
Ensemble adversarial attack · Diversity · Transferability

1 Introduction

Deep learning models are often vulnerable to adversarial examples: malicious inputs modified to yield erroneous model outputs, while appearing unmodified to human observers at inference phase [1–4]. Potential attacks include confusing vehicle behavior in automated driving or having malicious content like malware identified as legitimate. Yet, all existing adversarial example attacks require explicit knowledge of the model internals or its training data (white-box). However, to search for adversarial examples of a real world system, such knowledge may not be available. In this situation, the target model is a *black-box* to the attacker. Therefore, it is quite difficult to extract information about the decision boundary of target models, which is usually a pre-requisite to design input perturbations that result in erroneous predictions. However, previous works have shown that *transferability* exists between different models, i.e., the adversarial examples can transfer from one model to another [1, 5–8]. Such a property can be leveraged to perform black-box attacks. In other words, the attacker can query the target system, and establish a *substitute model* based on the query results [9]. Then the attacker can

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11303, pp. 1–13, 2018.
https://doi.org/10.1007/978-3-030-04182-3_16

generate the adversarial examples for the substitute model, and these adversarial examples may transfer to disorder the target system. For example, an adversary who seeks to penetrate a computer network rarely has access to the specifications of the deployed intrusion detection system, however they can observe its outputs for any chosen inputs [10]. These observed input-output pairs will be used to produce synthetic datasets, and to train a substitute model approximating the target system. Therefore, the adversarial examples generated by substitutes are more likely to transfer to confuse the target system.

However, conventional attack strategies notoriously only consider to train a single substitute to craft adversarial examples with a weak transfer capability in black-box attack scenario, which is easily defended by existed defense mechanism [11–13]. Papernot et al. [14] have proposed ensemble adversarial training technique, which is an extension of adversarial training [1, 15], to increase robustness of DL models against black-box attacks. Thus, new attack strategies should be designed to explore the vulnerability of DL models with ensemble adversarial training.

In this paper, we propose two types of ensemble-based black-box attack strategies, *iterative cascade ensemble strategy* and *stack parallel ensemble strategy*, to implement more powerful black-box attacks against DL models and demonstrate that the ensemble adversarial training does not significantly increase the robustness and security of DL models. Besides, potential factors that contribute to the effective attacks against DL models are examined from three perspectives: the transferability of substitutes, the diversity of substitutes, and the number of substitutes. Ensemble adversarial black-box attack strategies and strategy analysis will be emphatically introduced in Sect. 2. The comparison experiment results on real world data sets and feasibility exploration are reported in Sect. 3 and paper concludes in Sect. 4.

2 Ensemble-Based Black-Box Attack Strategy

Before introducing the attack strategies, we will briefly introduce the architecture of substitutes and transferable adversarial examples generation algorithms used in this paper. For the input $x \in R^D$, the composition of functions modeled by the substitute can be formalized as [16]:

$$F(x) = softmax(f_n(\theta_n, f_{n-1}(\theta_{n-1}, \ldots f_2(\theta_2, f_1(\theta_1, x))))) \quad (1)$$

where each function f_i for $i \in 1 \ldots n$ is modeled by a layer of neurons, each layer is parameterized by a weight vector θ_i impacting each neuron's activation. The output of the last layer is computed by using the softmax function, which ensures that the output vector $F(x)$ satisfies $0 \leq F(x)_i \leq 1$, and $F(x)_1 + \ldots + F(x)_c = 1$, where c is the number of classes.

Transferable adversarial examples are generated by substitute through carefully introducing human indistinguishable perturbations to the original examples, then these generated adversarial examples $x^* \in R^D$ can transfer to confuse target model O, i.e., $O(x^*) \neq O(x)$. Currently proposed adversarial examples generation algorithms mainly

include gradient-based (e.g., FGSM [1], I-FGSM [2], R + FGSM [14], etc.) and optimization-based (e.g., Carlini L_∞ Attack [17]), and specific details are described below:

Fast Gradient Sign Method (FGSM) is a single-substitute attack method. It finds the adversarial perturbation that yields the highest increase of the loss function under L_∞-norm. The update equation is

$$x^* = x + \alpha \cdot sign\left(\nabla_x loss\left(1_y, F(x)\right)\right) \tag{2}$$

where α controls the magnitude of adversarial perturbation, 1_y is the one-hot encoding of the ground truth label of y. I-FGSM is a straightforward way to extend the FGSM by using a better iterative optimization strategy and R + FGSM significantly increases the power of the FGSM by adding gaussian noise to inputs before computing the gradient.

Carlini L_∞ Attack is a stronger single-substitute attack method proposed recently. It finds the adversarial perturbation r by using an auxiliary ω as

$$r = \frac{1}{2}(\tanh(\omega) + 1) - x \tag{3}$$

Then the loss function optimizes the auxiliary variable ω_n

$$\min_\omega \left\| \frac{1}{2}(\tanh(\omega) + 1) \right\| + c \cdot f\left(\frac{1}{2}(\tanh(\omega) + 1)\right) \tag{4}$$

The function $f(\cdot)$ is defined as

$$f(x) = \max\left(Z(x)_{1_y} - \max\left\{Z(x)_i : i \neq 1_y\right\}, -\kappa\right) \tag{5}$$

where $Z(x)_i$ is the logits output for class i, and κ controls the confidence gap between the adversarial class and true class.

Yet, these single-substitute attack algorithms achieve unsatisfactory attack performance in black-box attack scenario. Then, we attempt to ensemble multiple pre-trained substitutes to produce adversarial examples with more powerful transferability in the form of iterative cascade ensemble and stack parallel ensemble, as illustrated in Fig. 1.

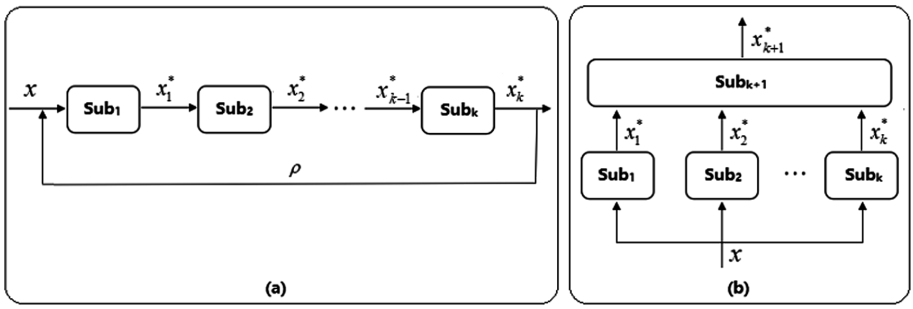

Fig. 1. Illustration of Iterative Cascade Ensemble Strategy (a) and Stack Parallel Ensemble Strategy (b).

2.1 Iterative Cascade Ensemble Strategy

Iterative cascade ensemble strategy employs a cascade structure, as shown in Fig. 1(a), where each substitute of cascade will receive adversarial examples $x_j^*(j \in [0,k])$ generated by its preceding substitute, and output its counterparts to the next substitute. During each iteration, the output of the k-th substitute x_k^* will be used as the input to the first substitute. Output results obtained from the k-th substitute after ρ iterations are final adversarial examples. Before implementing the iterative cascade ensemble strategy, the adversary first requires to train k heterogeneous substitute models with various synthetic datasets, which are constructed by observed input-output pairs and their augmentation with Jacobian-based technique [9]. In order to obtain more effective adversarial examples, each substitute is trained based on various architectures of deep neural networks. Afterwards, FGSM or Carlini L_∞ Attack is adopted as a classic attack algorithm for each substitute to craft adversarial examples. Finally, the adversaries can cascade multiple pre-trained substitutes and iteratively maximize each loss of substitute to obtain the final adversarial examples. The iterative cascade attack procedure is outlined in Algorithm 1.

Algorithm 1: Iterative Cascade Ensemble Strategy for Generating Transferable Adversarial Examples —— Iter_Casc

Input: normal example x, the ground true label y, a substitute F, the number of substitutes k, perturbation amplitude α, gaussian noise amplitude ε , the maximum iterative epochs ρ

Output: Transferable adversarial example x_k^*

Initialize the value of ε, α , k , ρ
$x_0^* = x + \varepsilon \cdot sign(N(0^D, 1^D))$
while $\rho > 0$ **do**
  for j = *1 to* k **do**
    $loss\left(1_y, F_j\left(x_{j-1}^*\right)\right) = -\sum_{t=1}^{c}\left(1_{y_t} \cdot \log F_j\left(x_{j-1}^*\right)_t\right)$
    $grads = \nabla_{x_{j-1}^*} loss(1_y, F_j(x_{j-1}^*))$
    $x_j^* = x_{j-1}^* + \alpha \cdot sign(grads)$ //$L_\infty(L_1\ or\ L_2)$ distance metric
  end for
  $x_0^* = x_k^*$
  $\rho = \rho - 1$
end while
return x_k^*

The algorithm first requires to initialize the value of all input variables ε, α, k, ρ (where $\varepsilon = \alpha/2$ and $k = \rho$), and add gaussian noise to original normal examples. For each substitute, the standard cross entropy loss function [18] should be constructed to compute gradient to maximize the loss function optimized for the L_∞ distance metric. The gradient of loss function determines the direction which feature should be changed. During each iteration, generated adversarial example x_k^* will be assigned to x_0^* as input

of the first substitute. Until the loop iteration ends, the final transferable adversarial examples are obtained from the output of k-th substitute.

2.2 Stack Parallel Ensemble Strategy

Stack parallel ensemble strategy employs a parallel structure, as shown in the Fig. 1(b), where each substitute of parallel will receive the original legitimate example x, and output result $x_j^*(j \in [1,k])$ will be combined with a linear way as new input of the $k + 1$ substitute. Output results obtained from the $k + 1$ substitute are final adversarial examples. Before implementing the parallel ensemble strategy, the adversary first requires to train $k + 1$ heterogeneous substitute models with various synthetic datasets, which are constructed by observed input-output pairs and their augmentation with Jacobian-based technique [9]. In order to achieve more effective adversarial examples, each substitute is still trained based on various architectures of deep neural networks. Afterwards, FGSM or Carlini L_∞ Attack is adopted as a classic attack algorithm for each substitute to craft adversarial examples. Finally, the adversary can parallel multiple pre-trained substitutes and maximize each loss of substitute to obtain adversarial examples. The stack parallel attack procedure is outlined in Algorithm 2.

Algorithm 2: Stack Parallel Ensemble Strategy for Generating Transferable Adversarial Examples —— Stack_Paral

Input: normal example x, the ground true label y, a substitute F, the number of substitutes $k+1$, perturbation amplitude α, gaussian noise amplitude ε
Output: Transferable adversarial example x_{k+1}^*

Initialize the value of ε, α, k and $x_{mid}^* = 0^D$
$x_0^* = x + \varepsilon \cdot sign(N(0^D, 1^D))$
for $j = 1$ *to* k **do**
  $loss(1_y, F_j(x_0^*)) = -\sum_{t=1}^{c}\left(1_{y_t} \cdot \log F_j(x_0^*)_t\right)$
  $grads = \nabla_{x_0^*} loss(1_y \cdot F_j(x_0^*))$
  $x_j^* = x_0^* + \alpha \cdot sign(grads)$ //$L_\infty(L_1 \text{ or } L_2)$ distance metric
  $x_{mid}^* = x_{mid}^* + x_j^*$ // sum and save output results of each substitute
end for
$x_{mid}^* = x_{mid}^* / k$ // linear combination
$loss(1_y, F_{k+1}(x_{mid}^*)) = -\sum_{t=1}^{c}\left(1_{y_t} \cdot \log F_{k+1}(x_{mid}^*)_t\right)$
$grads = \nabla_{x_{mid}^*} loss(1_y, F_{k+1}(x_{mid}^*))$
$x_{k+1}^* = x_{mid}^* + \alpha \cdot sign(grads)$
return x_{k+1}^*

The algorithm still requires to initialize the value of all input variables ε, α, k, x_{mid}^* (where $\varepsilon = \alpha/2, x_{mid}^* = 0^D$), add gaussian noise to original legitimate examples and compute gradient of constructed loss function. The gradient of loss function determines

the direction which feature should be changed. For the top-k substitutes, generated adversarial example $x_j^*(j \in [1,k])$ will be combined with a linear way and save to x_{mid}^* as new input of the $k + 1$ substitute. The final transferable adversarial examples are achieved from the output of $k + 1$ substitute.

2.3 Strategy Analysis

Empirical evidence has shown that adversarial examples appear in wide regions, spanning a contiguous subspace of high dimensionality and a large portion of this space is shared between different models, thus enabling transferability [1, 7, 19]. Ian Goodfellow et al. first proposed Gradient Aligned Adversarial Subspace (GAAS) [7] method to find multiple independent orthogonal adversarial directions to directly evaluate the dimensionality of the adversarial subspace. The dimensionality of adversarial subspaces is relevant to the transferability problem: the higher the dimensionality, the more likely the subspaces of substitute and target model will intersect significantly. As proposed in [7], the decision boundaries learned by both the substitute and target model must be extremely close to each another in adversarial direction. Adversarial direction is defined by x and $x^* : d_{adv} = (x^* - x)/x^* - x_2$, where adversarial example x^* (blue dot) is generated from test example (brown dot) x to be misclassified by substitute $F(x)$: $\text{argmin}_{\varepsilon > 0} F(x^* : x + \varepsilon \cdot d_{adv}) \neq F(x)$, as shown in Fig. 2(a). That is, the cross-boundary distance (the red double-ended arrows) in adversarial direction between the decision boundaries of substitute and target model must be very short. In other words, the shorter the distance, the stronger transferability.

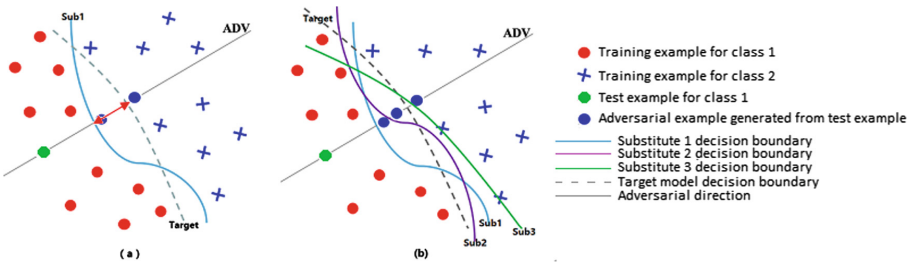

Fig. 2. Illustration of a binary misclassification procedure in the adversarial direction over a 2D input domain. (Color figure online)

Actually, it is difficult to guarantee the trained substitute accurately approximating the target black-box model and the adversarial direction is also not unique, which lead to the weak transferability of crafted adversarial examples. However, if adversarial examples remain adversarial for multiple substitutes, it is more likely to transfer to disorder the target model, as shown in Fig. 2(b). From the Fig. 2(b), we can observe that an adversarial example (blue dot) generated by our proposed ensemble-based black-box attack strategies crossing the decision boundaries of k (e.g. $k = 3$) substitutes, has a greater probability to cross the decision boundary of target model. This fully illustrates the ensemble-based black-box attack strategies effectively shorten the cross-boundary distance and improve the transferability of generated adversarial examples.

3 Experiments

All experiments[1] use Tensorflow[2] framework and cleverhans library[3]. To demonstrate the effectiveness and feasibility of the proposed ensemble-based black-box attack strategy, we empirically compare the conventional single-substitute attack algorithms described previously, e.g., FGSM, I-FGSM, R + FGSM and Carlini L_∞ attack, and expose the potential factors that contribute to the high-efficiency attacks.

3.1 Setup

Four benchmark datasets for two tasks, i.e., digit recognition and traffic sign recognition, are used in experiments. Details about datasets are listed in Table 1. The target classifier as black-box model in this work are trained with training data of each dataset. For each dataset, few unused test examples, as query inputs, are used to query target classifier and produce synthetic datasets augmented by observed input-output pairs. Then, diverse convolutional neural network architectures, as shown in Table 2, are selected to train substitutes with various synthetic datasets for ensemble to implement black-box attack tasks.

Table 1. Summary of 4 benchmark datasets

Name	Training data	Test data	Features	Labels	Task
MNIST	50000	10000	$28 \times 28 \times 1$	10	Digit recognition
USPS	7291	2007	$16 \times 16 \times 1$	10	Digit recognition
GTSRB	39209	12630	$32 \times 32 \times 3$	43	Traffic sign recognize
BelgiumTSC	4575	2534	$32 \times 32 \times 3$	62	Traffic sign recognize

Table 2. Neural network architectures used in this work for substitute and target model training. Conv: convolution layer, FC: fully connected layer, Relu: activation function

Target Model	Substitute1	Substitute2	Substitute3	Substitute4	Substitute5
Conv(128,3,3)+Relu Conv(64,3,3)+Relu Dropout(0.25) FC(128)+Relu Dropout(0.5) FC+Softmax	Conv(64,8,8)+Relu Conv(128,6,6)+Relu Conv(128,5,5)+Relu Dropout(0.5) FC+Softmax	FC(300)+Relu Dropout(0.5) FC(300)+Relu Dropout(0.5) FC(300)+Relu Dropout(0.5) FC(300)+Relu Dropout(0.5) FC+Softmax	Conv(32,3,3)+Relu Conv(32,3,3)+Relu Conv(64,3,3)+Relu Conv(64,3,3)+Relu FC(200)+Relu Dropout(0.5) FC+Softmax	Conv(32,3,3)+Relu Conv(32,3,3)+Relu Conv(64,3,3)+Relu Conv(64,3,3)+Relu Conv(128,3,3)+Relu Conv(128,3,3)+Relu Drop(0.2) FC(512)+Relu Dropout(0.5) FC+Softmax	Conv(64,3,3)+Relu Conv(64,3,3)+Relu Conv(128,3,3)+Relu Conv(128,3,3)+Relu FC(256)+Relu FC(256)+Relu FC+Softmax

[1] Codes is available at https://github.com/HangJie720/Ensemble_Adversarial_Attack.

[2] https://www.tensorflow.org/?hl=zh-cn.

[3] https://github.com/tensorflow/cleverhans.

Two diverse measurements, *Success rate* and *Transfer rate*, are redefined to evaluate the vulnerability of DL models according to Eqs. 6. and 7.

$$\frac{1}{nk}\sum_{j=1}^{k}\sum_{i=1}^{n}\mathbb{I}\left(F_j\left(x_i^*\right) \neq F_j(x_i)\right) \tag{6}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(O\left(x_i^*\right) \neq O(x_i)\right) \tag{7}$$

where $\mathbb{I}(\cdot) = 1$ represents generated adversarial example is misclassified, and 0, otherwise. These two metrics are used to measure the error rate of substitute and target model respectively.

3.2 Results

This section first quantitatively analyzes the vulnerability of DL models under success rate and transfer rate measurement. Afterwards, we empirically compare the conventional single-substitute attack algorithms based on FGSM and Carlini L_∞ attack for different datasets. Finally, possible factors that contribute to the higher transfer rate are explored from two aspects, the diversity of substitutes and the number of substitutes k.

Figure 3 demonstrates that deep learning models are extremely susceptible to adversarial examples generated by proposed ensemble-based black-box attack strategies under different perturbation amplitude α.

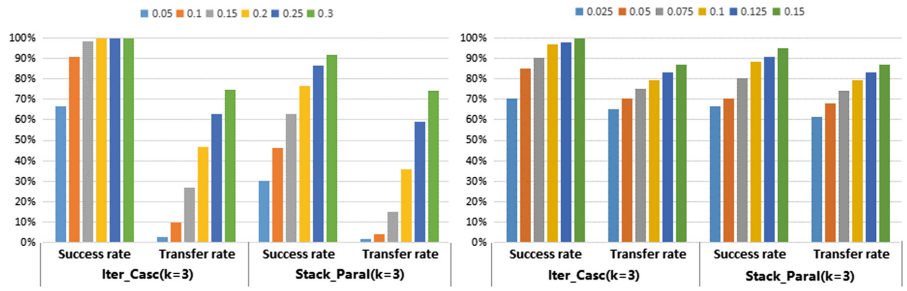

Fig. 3. Success rate and Transfer rate of adversarial examples generated by ensemble-based black-box attack strategies under different perturbation amplitude on MNIST and GTSRB.

The transferability of adversarial examples generated by each substitute and cascading or paralleling any k substitutes (e.g. $k = 3$, 5) are illustrated in Table 3 and Fig. 4. Experiments demonstrate that the adversarial examples crafted by *iterative cascade ensemble strategy* achieve higher transfer rate than *stack parallel ensemble strategy* dramatically. Both obtain superior attack performance to other single-substitute attack algorithms. We also can observe that optimization-based algorithm (e.g. Carlini L_∞ attack) provided for each substitute to iterative cascade ensemble obtain greater transferability than gradient-based algorithm (e.g. FGSM). Figure 5

demonstrates that our proposed ensemble-based black-box attack strategies are still aggressive to target classifier trained with ensemble adversarial training defense mechanism.

Table 3. Transfer rate of adversarial examples generated by single-substitute, iterative cascade ensemble strategy and stack parallel ensemble strategy based on **FGSM** and **Carlini L_∞ attack** for different datasets.

FGSM (α= 0.3)	MNIST	USPS	GTSRB	BelgiumTSC
Sub1	32.47%	30.44%	59.32%	58.22%
Sub2	37.00%	38.20%	55.27%	49.64%
Sub3	18.57%	25.42%	50.23%	50.12%
Sub4	19.04%	27.62%	45.55%	43.65%
Sub5	16.61%	25.29%	40.29%	49.23%
Iter_Casc (k = 3)	**58.01%**	**53.23%**	**65.89%**	**64.68%**
Stack_Paral (k = 3)	**50.00%**	**48.27%**	**61.36%**	**60.00%**
Carlini L_∞ attack ($\kappa = 0\kappa = 0$)	MNIST	USPS	GTSRB	BelgiumTSC
Sub1	12.50%	10.50%	12.50%	28.20%
Sub2	12.50%	12.00%	20.50%	19.65%
Sub3	1.50%	2.50%	9.50%	2.10%
Sub4	0.50%	1.50%	5.55%	3.55%
Sub5	1.00%	0.50%	8.50%	1.20%
Iter_Casc (k = 3)	**94.50%**	**90.00%**	**100.00%**	**100.00%**
Stack_Paral (k = 3)	**17.50%**	**20.00%**	**30.50%**	**35.00%**

Moreover, possible factors that contribute to the higher transfer rate are explored from two perspectives: the diversity of substitutes and the number of substitutes k.

(1) The number of substitutes k. The experiment results are shown in Fig. 6. for our proposed ensemble-based black-box attack strategies, which indicates that the larger the value of k, the higher transfer rate of generated adversarial examples.

(2) The diversity of substitutes. Two averaged pairwise measures [20] (the Q statistics, the correlation coefficient ρ) and two non-pairwise measures [20] (The entropy measure E, the Kohavi-Wolpert variance KW) are selected to analyze the relationship between the diversity of substitute and transferability of generated adversarial examples. Experimental results are listed in Table 4, where I, II, III and IV represent the four strategies to generate the substitutes, such as, the substitutes are same, trained with different training sets, trained with different architectures and trained with different training sets and architectures respectively. Comparative experimental results demonstrate that the greater the diversity of substitutes, the stronger the transferability of adversarial examples. Thus, all same substitutes used in I-FGSM obtain the lowest transfer rate, as shown in Fig. 4.

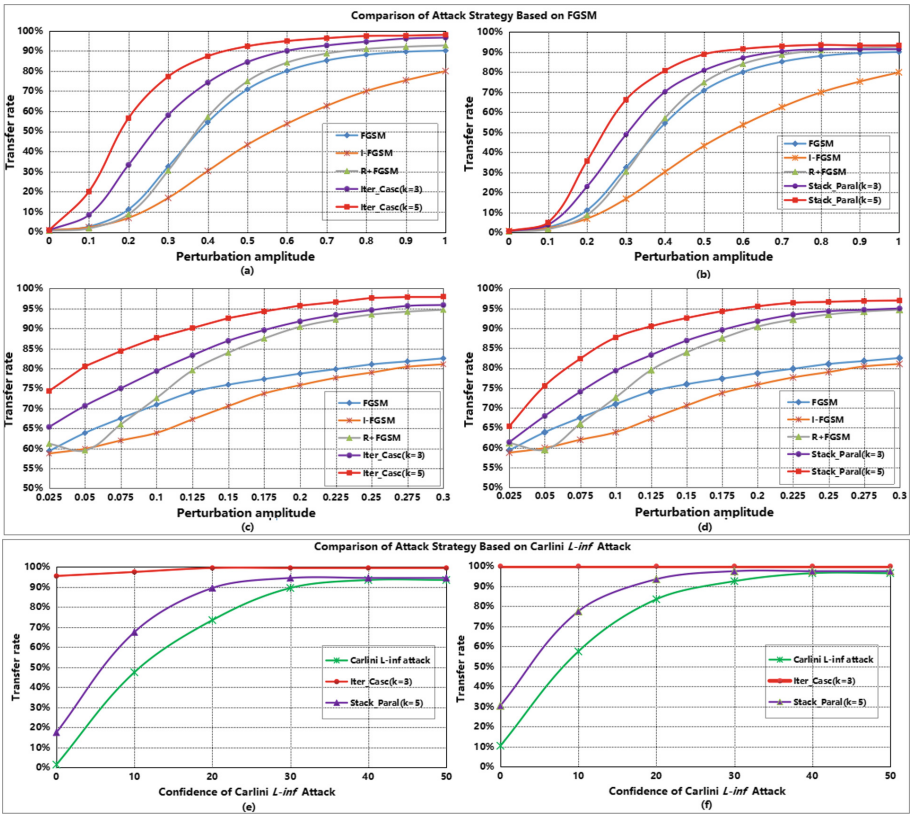

Fig. 4. Transfer rate of adversarial examples crafted by disparate attack strategies on two major classification tasks. Ensemble strategies compared with single-substitute attack algorithms based on FGSM under differ perturbation amplitude α are shown in Fig. (a)–(d). Ensemble strategies compared with single-substitute attack algorithms based on Carlini L_∞ Attack under different confidence κ are shown in Fig. (e) and (f).

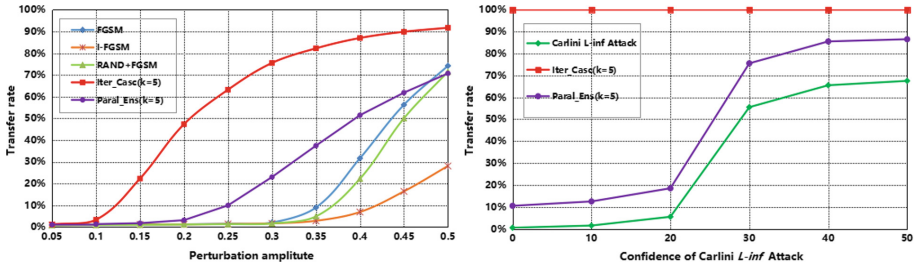

Fig. 5. Weakly defense performance of target classifier trained with ensemble adversarial training defense mechanism.

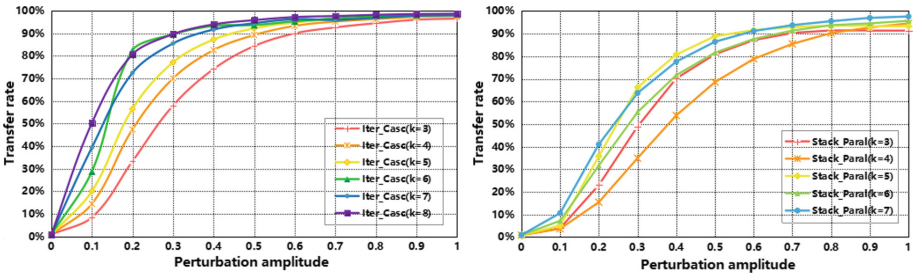

Fig. 6. Transfer rate of adversarial examples crafted by iterative cascade ensemble strategy and stack parallel ensemble strategy with different number of substitutes k.

Table 4. The relationship of diversity in substitute cascade/parallel ensembles and transferability of generated adversarial examples. (↑) represents the measure value of diversity is increased, (↓) represents the measure value of diversity is decreased.

MNIST	Transfer Rate		Diversity Measure Value			
	Iter_Casc (k = 3)	Stack_Paral (k = 3)	$Q(\downarrow)$	$\rho(\downarrow)$	$Ent(\uparrow)$	$KW(\uparrow)$
I	16.89%	10.89%	1.0000	1.0000	0.0000	0.0000
II	20.35%	18.52%	0.8900	0.7343	0.4900	0.1089
III	40.23%	34.53%	0.6432	0.5321	0.6235	0.2336
IV	58.01%	50.00%	0.3411	0.2300	0.7800	0.3345
GTSRB	Iter_Casc (k = 3)	Stack_Paral (k = 3)	$Q(\downarrow)$	$\rho(\downarrow)$	$Ent(\uparrow)$	$KW(\uparrow)$
I	70.44%	66.24%	1.0000	1.0000	0.0000	0.0000
II	79.26%	72.81%	0.7100	0.6911	0.5300	0.2033
III	88.12%	80.36%	0.5302	0.3510	0.7122	0.3010
IV	95.89%	93.80%	0.2201	0.1800	0.8201	0.4700

4 Conclusion

In this paper, we propose two types of ensemble-based black-box attack strategies, *iterative cascade ensemble strategy* and *stack parallel ensemble strategy*, to explore the vulnerability of deep learning system. Experimental results show that our proposed ensemble adversarial attack strategies can successfully attack the deep learning system trained with ensemble adversarial training defense mechanism. The adversarial examples generated by *iterative cascade ensemble strategy* achieve better transferability than *stack parallel ensemble strategy* dramatically. Both obtain superior attack performance to other single-substitute attack algorithms. We also can observe that the diversity in substitute ensembles is an important factor to influence the transferability of generated adversarial examples.

Acknowledgement. This work was partially supported by Natural Science Foundation of China (No. 61603197, 61772284, 41571389).

References

1. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations (2015)
2. Kurakin, K., Goodfellow, J., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations (2017)
3. Biggio, B., et al.: Evasion attacks against machine learning at test time. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS (LNAI), vol. 8190, pp. 387–402. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40994-3_25
4. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Berkay Celik, Z., Swami, A.: The limitations of deep learning in adversarial settings. In: 1st IEEE European Symposium on Security and Privacy (EuroS&P), pp: 372–387. IEEE Press, New York (2016)
5. Szegedy, C., et al.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations (2014)
6. Papernot, N., McDaniel, P.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
7. Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: The space of transferable adversarial examples. arXiv preprint arXiv:1704.03453 (2017)
8. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. In: 5th International Conference on Learning Representations (2017)
9. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Berkay Celik, Z., Swami, A.: Practical black-box attacks against machine learning. In: ASIACCS, pp: 506–519. ACM, New York (2017)
10. Papernot, N., Mcdaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814 (2016)
11. Meng, D.Y., Chen, H.: MagNet: a two-pronged defense against adversarial examples. In: ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147. ACM, New York (2017)
12. Wang, Q.L., Guo, W.B., Zhang, K.X., Ororbia Ii, A. G., Xing, X., Liu, X.: Adversary resistant deep neural networks with an application to malware detection. In: 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 1145–1153. ACM, New York (2017)
13. Bhagoji, A, J., Cullina, D., Mittal, P.: Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers. arXiv preprint arXiv:1704.02654 (2017)
14. Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, P.: Ensemble adversarial training: attacks and defenses. In: 6th International Conference on Learning Representations (2018)
15. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. In: 5th International Conference on Learning Representations (2017)
16. Papernot, N., McDaniel, P., Wu, X.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP), vol. 00, pp. 582–597. IEEE, USA (2016)

17. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 38th IEEE Symposium on Security and Privacy, pp. 39–57. IEEE, USA (2017)
18. Boer, P.T.D., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**(1), 19–67 (2005)
19. Grosse, K., Papernot, N., Manoharan, P., Backes, M., McDaniel, P.: Adversarial perturbations against deep neural networks for malware classification. arXiv preprint arXiv:1606.04435 (2016)
20. Ludmila, I.K., Whiker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. **2**(51), 181–207 (2003)

Author Query Form

Book ID : **476537_1_En**

Chapter No : **16**

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

Dear Author,

During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	