Signal Processing: Image Communication 107 (2022) 116747

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Revisiting ensemble adversarial attack

Ziwen He [a,b], Wei Wang [a,*], Jing Dong [a], Tieniu Tan [a]

[a] Center for Research on Intelligent Perception and Computing, NLPR, CASIA, Beijing 100190, China
[b] School of Artificial Intelligence, University of Chinese Academy of Science (CAS), Beijing 100190, China

ARTICLE INFO

Keywords:
Adversarial attack
Ensemble strategies
Gradient-based methods
Deep neural networks
Image classification

ABSTRACT

Deep neural networks have shown vulnerability to adversarial attacks. Adversarial examples generated with an ensemble of source models can effectively attack unseen target models, posing a security threat to practical applications. In this paper, we investigate the manner of ensemble adversarial attacks from the viewpoint of network gradients with respect to inputs. We observe that most ensemble adversarial attacks simply average gradients of the source models, ignoring their different contributions in the ensemble. To remedy this problem, we propose two novel ensemble strategies, the Magnitude-Agnostic Bagging Ensemble (MABE) strategy and Gradient-Grouped Bagging And Stacking Ensemble (G^2BASE) strategy. The former builds on a bagging ensemble and leverages a gradient normalization module to rebalance the ensemble weights. The latter divides diverse models into different groups according to the gradient magnitudes and combines an intragroup bagging ensemble with an intergroup stacking ensemble. Experimental results show that the proposed methods enhance the success rate in white-box attacks and further boost the transferability in black-box attacks.

1. Introduction

In the past few years, deep learning has been widely used in many computer vision tasks, such as image classification [1], object detection [2] and action recognition [3]. However, recent research has shown that deep learning models can easily be fooled by adversarial examples that are crafted by maliciously adding designed perturbations to the inputs [4,5]. These examples are imperceptible to human eyes but lead to incorrect outputs, posing potential threats to face recognition [6], autonomous driving [7] and other real-world applications [8,9].

Crafting adversarial examples, i.e., adversarial attack, has drawn enormous attention since it can evaluate the adversarial robustness. There are two main typical adversarial attack protocols, i.e., white-box attacks [4,5] and black-box attacks [10,11]. In white-box attacks, attackers have full knowledge of the target model, including architectures, parameters and potential defense modules [12]. In contrast, in black-box attacks, attackers cannot access any information of the target model. Nonetheless, in such a scenario, it is probable to utilize transferability, which means adversarial examples crafted on one white-box surrogate model can be misclassified by unseen target models [13–15].

A series of methods have been proposed to enhance the attack ability of adversarial examples in both white-box and black-box settings, including (1) single-model attack [4,11,16–18] and (2) ensemble attack [11,13]. Many attempts have been made on the former, by improving attacking algorithms on a single model. For example, different tricks, such as momentum and input transformations, are applied to avoid adversarial examples trapping in local optima. The latter attacks an ensemble of deep learning models rather than a single model. In white-box attack, it can generate a cross-model adversarial example to fool all models in the ensemble. In black-box attack, if an example is adversarial for multiple models in the ensemble, it is hypothesized that this example remains adversarial for other unseen models. Almost all top-ranked solutions in prior competitions of adversarial attacks are based on ensemble-based methods [19]. However, existing ensemble attacks simply fuse the outputs of multiple models evenly. Ensemble strategies has not been thoroughly investigated yet.

In this paper, we observe that existing ensemble strategies [20] obtain a loss in attack capability when multiple models have noticeably different gradient magnitudes. The reason is shown as a toy example in Fig. 1. Considering Liu et al. [13] have demonstrated that the gradient directions of different source models are orthogonal to each other when conducting ensemble adversarial attacks, the gradient magnitude plays a vital role in the final search direction of adversarial examples. In the baseline, the model with the largest gradient magnitude dominates the ensemble direction, ignoring the effects of other models. Such a conflict decreases the attack success rate on the other two models and therefore hinders model diversity for boosting transferability.

To overcome this drawback, we propose two novel ensemble strategies, dubbed magnitude-agnostic bagging ensemble (MABE) and gradient-grouped bagging and stacking ensemble (G^2BASE). In MABE, the input gradients are normalized according to the gradient amplitude

* Corresponding author.
E-mail address: wwang@nlpr.ia.ac.cn (W. Wang).

https://doi.org/10.1016/j.image.2022.116747
Received 22 April 2021; Received in revised form 28 February 2022; Accepted 27 May 2022
Available online 4 June 2022

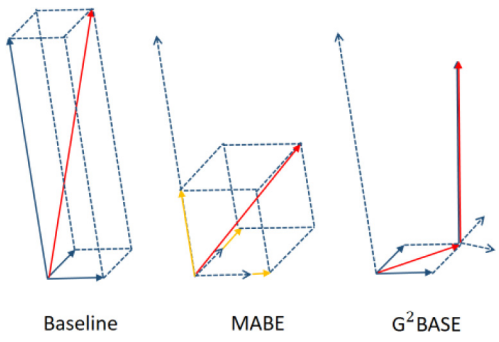

Fig. 1. Illustration of different ensemble strategies. The baseline method ensembles three orthogonal gradient vectors, shown as the blue solid lines. The final gradient direction, shown as the red line, is dominated by the gradient with the largest magnitude. The magnitude-agnostic bagging ensemble (MABE) applies gradient normalization on three models and obtains normalized gradient vectors, shown as the yellow line. The gradient-grouped bagging and stacking ensemble (G^{2}BASE) contains multiple steps and only fuses gradient vectors with similar magnitudes in each step. The end point in one step is taken as the start point in the next step.

so that each model contributes to the final gradient direction and therefore boosts diversity. In G^{2}BASE, according to gradient amplitudes, diverse models are divided into different groups. For intra-group ensemble, the gradient magnitudes of the models in each group are similar to each other. For inter-group ensemble, the gradient magnitudes between groups are significantly different. The bagging strategy is used in intra-group, and the stacking method is applied in inter-group. This group strategy avoids domination of a specific model and greatly facilitates attack ability. To conclude, MABE is based on the framework of bagging ensembles and efficiently achieves improvement through introducing a normalization module. G^{2}BASE, by contrast, is more effective and flexible, but a bit time-consuming due to the group selection module and the stacking strategy. Experimental results show that both proposed methods boost attack performance.

In summary, the major contributions of this work are:

• We find a drawback of the existing ensemble strategies from the viewpoint of gradient, i.e., they simply average the gradients and therefore ignore the different contributions of models in the ensemble.

• We provide a new insight into ensemble adversarial attack, i.e., to alleviate the conflict between models with different gradient magnitudes, and propose two novel ensemble attack methods, namely, magnitude-agnostic bagging ensemble (MABE) and gradient-grouped bagging and stacking ensemble (G^{2}BASE), to boost the attack performance.

The remainder of the paper is organized as follows. Section 2 briefly discusses the related work. In Section 3, we elaborate the two proposed methods, GABE and G^{2}BASE. We then present the experimental results and analysis in Section 4. Finally, Section 5 concludes this paper.

2. Related work

Adversarial attack methods can be categorized into single-model attack and ensemble attack. We briefly introduce the two types of attack in this section.

2.1. Single-model adversarial attack

Goodfellow et al. [4] introduce the fast gradient sign method (FGSM) to craft white-box adversarial examples by a one-step gradient update along the direction of the sign of the gradient at each pixel. Madry et al. [21] propose the projected gradient descent (PGD) attack, starting from a random point within the regularized ball space of the input example and iteratively updating the adversarial example. Dong et al. [11] propose the momentum iterative method (MIM), which uses momentum in the optimization step to speed up the convergence rate and avoid getting trapped in a local optima. Xie et al. [16] optimize the adversarial perturbations over the diverse transformation of the input image at each iteration, namely, the diverse input method (DIM). To generate more transferable adversarial examples against defense models, Dong et al. [17] propose the translation invariant method (TIM), which uses a set of translated images to optimize adversarial perturbations. Recently, Lin et al. [18] propose two new attack methods, the Nesterov iterative fast gradient sign method (NI-FGSM) and the scale invariant attack method (SIM), to further improve the transferability of adversarial attacks.

These methods are all focused on improving attacking algorithms on a single model and the attack performance is closely related to the choice of this specific model. They can be naturally integrated into ensemble adversarial attacks to further promote the attack ability.

2.2. Ensemble adversarial attack

Ensemble methods have been widely adopted in previous studies to enhance the performance of neural networks [22–24]. For example, bagging [20] and stacking [25] can both improve the accuracy and robustness of neural networks. Recently, ensemble methods have been introduced into adversarial attacks.

Liu et al. [13] proposed an ensemble-based approach to generate adversarial examples, which prevents the noise from overfitting a single model architecture and thus bolsters the transferability. Dong et al. [11] further investigate three manners of organizing the base models and demonstrate that the ensemble of averaging logits outperforms the others for boosting the attack effectiveness. Hang et al. [26] propose two types of ensemble-based black-box attack strategies to produce adversarial examples with more powerful transferability. Li et al. [27] apply feature-level perturbations to an existing model to potentially create a huge set of diverse models and propose a longitudinal ensemble method specifically for their networks. Che et al. [28] divide a large number of pretrained source models into several batches and introduce long-term gradient memories in their new ensemble algorithm for specific networks or tasks (e.g., pix-to-pix image translation).

Despite achieving impressive results, none of the above methods consider the negative impact of gradient amplitudes. When attacking an ensemble model composed of diverse gradient amplitudes, the generated adversarial examples can fool only parts of models and therefore are at the risk of lowering transferability. This work solves this problem by introducing a gradient normalization module or model grouping to leverage the diversity of models with similar gradient magnitudes and alleviate the conflict between models with different gradients.

3. Method

In this section, after brief descriptions of preliminaries, we elaborate the two proposed strategies, MABE and G^{2}BASE.

3.1. Preliminaries

We describe the ensemble adversarial attack in this subsection. For clarity, we focus on a nontargeted attack, which crafts adversarial examples misclassified as wrong labels. Denote the data as $(x; y) \sim D$, attacking a single target model f solves the optimization problem with regard to δ

$$\max_{\delta} J\left(l\left(x+\delta\right), y\right),$$
$$subject\ to \left\|\delta\right\|_p < \epsilon, \tag{1}$$

where J is the classification loss, usually cross entropy (CE) loss with softmax as the activation. y is the ground-truth label, $l\left(x+\delta\right)$ represents the logits of the model f, and the input values are softmax. The perturbation is bounded by the ℓ_p norm with a small constant ϵ, and we focus on the ℓ_∞ norm in this paper.

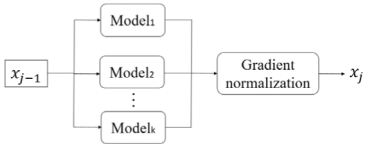

Fig. 2. Illustration of the magnitude-agnostic bagging ensemble.

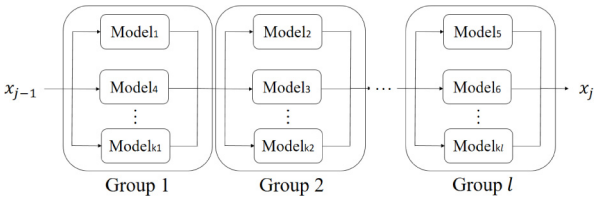

Fig. 3. Illustration of gradient-grouped bagging and stacking ensembles.

To generate more transferable adversarial examples, the ensemble adversarial attack utilizes an ensemble model as the substitute and performs white-box attack on it. Dong et al. [11] report three different ensemble methods, including ensembles in logits, ensembles in predictions and ensembles in loss. All methods simply add model outputs together and then average. The only difference is where to combine the outputs of multiple models.

Ensemble in logits modifies the objective in (1) as

$$\max_{\delta} J\left(\sum_{i=1}^{k}\alpha_i l_i\left(x+\delta\right), y\right), \tag{2}$$

where k is the number of models, $\sum_{i=1}^{k}\alpha_i l_i\left(x+\delta\right)$ is the ensemble model, and α_i is the ensemble weight of the ith model and satisfies $\sum_{i=1}^{k}\alpha_i = 1$.

Ensemble in loss directly averages in loss as

$$\operatorname*{Max}_{\delta}\sum_{i=1}^{k}\alpha_i J(l_i\left(x+\delta\right), y). \tag{3}$$

To boost the ensemble adversarial attack, we propose two novel ensemble strategies. We have elaborated on their drawbacks from the viewpoint of gradient information in Section 1. In the following subsections, we present our ensemble strategies, enabling us to efficiently craft adversarial examples in both white-box and black-box attack protocols.

3.2. Magnitude-agnostic bagging ensemble

The magnitude-agnostic bagging ensemble (MABE) employs a bagging structure of ensemble learning, as shown in Fig. 2. This method is based on the framework of ensembles in loss. The main difference is that MABE leverages the gradient normalization module to balance the contributions of each substitute model to the final adversarial gradient direction.

For one-step methods, such as FGSM, and the first step of iteration-based methods, the x_{j-1} in Fig. 2 is the clean sample x_0. Otherwise, it represents the adversarial example at the $j-1$ iteration. MABE inputs x_{j-1} into each substitute model and computes the loss $J_i = J(l_i\left(x+\delta\right), y)$.

Afterwards, the gradient alignment module takes all losses as inputs and reconstructs the total loss by the following equation:

$$J = \sum_{i=1}^{k}\alpha_i\frac{J_i}{\left\|\nabla_\delta J_i\right\|_2}. \tag{4}$$

Here, we analyze the relationship between the ensemble in loss and MABE. As both methods are based on the update of gradients, we compare their ensemble gradients with respect to the perturbation δ. For ensemble in loss, its final gradient is

$$\nabla_\delta J = \sum_{i=1}^{k}\alpha_i\nabla_\delta J(l_i\left(x+\delta\right), y). \tag{5}$$

Each term $\nabla_\delta J(l_i\left(x+\delta\right), y)$ is only related to one single model f_i. When one model has a small gradient magnitude, it has mild effects on the total gradient.

For our MABE, the gradient is

$$\nabla_\delta J = \sum_{i=1}^{k}\alpha_i\frac{\nabla_\delta J(l_i\left(x+\delta\right), y)}{\left\|\nabla_\delta J(l_i\left(x+\delta\right), y)\right\|_2}. \tag{6}$$

Each term can be seen as a unit vector to ensemble the final gradient direction. In this way, we directly align the gradients of source models, ensuring the contributions of models with small gradient magnitudes.

In other words, when there is a large difference between the magnitudes of ensemble models, MABE effectively utilizes the diversity of each model rather than focusing on one principle model. Without the normalization module in MABE, the adversarial direction in the generation process may only cross the decision boundary of the model with the largest gradient magnitude. In this scenario, this ensemble strategy is highly likely to degrade to a single-substitute method. Instead, the gradient direction in MABE is more likely to cross the decision boundaries of each ensemble model. Recall the key idea of ensemble adversarial attacks: if an adversarial example can fool multiple substitute models, it is more likely to disorder the target model. Therefore, MABE has a greater possibility to cross the decision boundary of the target model.

3.3. Gradient-grouped bagging and stacking ensemble

The gradient-grouped bagging and stacking ensemble (G^2BASE) combines the bagging and stacking structure of ensemble learning, as shown in Fig. 3. G^2BASE first divides source models into different groups and then utilizes bagging and stacking ensemble strategies for intro-groups and inter-groups, respectively. In this case, the models are divided into l groups with the proposed Algorithm 1. For example, k_l models including the 1st and 4th models are selected in group 1 and ensemble with the bagging way. Afterwards, different groups ensemble with the stacking method, and the order of groups in stacking is fixed in each iteration. We observe that for different batches of inputs x_{j-1}, the gradient magnitudes of a fixed classifier have low variance. Thus, the grouping process is unnecessary to perform in each iteration. Instead, when a new batch of clean examples is fed in, the grouping of source models has been finished. Therefore, G^2BASE is computationally inexpensive compared with traditional ensemble methods.

In G^2BASE, all substitute models are divided into several groups according to their input gradient magnitudes. The main intuition is to ensure that the input gradient magnitudes of intra-groups are close while inter-groups are far enough. Denote the magnitude of the ith model's gradient w.r.t. the perturbation as g_i. To measure the distance between gradient magnitudes of two models, we use the following equation:

$$d\left(g_i, g_j\right) = \log\left(\max\left\{g_i, g_j\right\}\right) - \log\left(\min\left\{g_i, g_j\right\}\right). \tag{7}$$

If this distance is lower than the predefined threshold h, these two models are divided into the same group. The grouping process can be seen as a clustering problem and solved with Algorithm 1. Afterwards, we run the optimization process by utilizing different ensemble intragroup and intergroup strategies.

First, the models in the same group compose an ensemble model via the bagging strategy. Different from MABE, which relies on the framework of ensembles in loss, G^2BASE is more flexible to combine with other bagging strategies. Each intragroup can be ensembled via either an ensemble in logits method or the proposed MABE. Thus, the intragroup obtains a lower variance by averaging the outputs of independent models.

In another aspect, different groups compose an ensemble model in the stacking ensemble strategy. The traditional stacking method collects the outputs of each base model to create a new dataset, and then applies this dataset to train a meta model at a high level. Similarly, in our method, the outputs of the last bagging ensemble model are fed into the next model as inputs. For example, one group G_i takes x^* as its input and optimizes the formulation

$$\max_{\delta_i} J(G_i, x^* + \delta_i, y), \tag{8}$$

obtaining the adversarial example $x_i^* = x^* + \delta_i$. Iteratively, the next group G_{i+1} will take x_i^* as input and optimize the formulation

$$\max_{\delta_{i+1}} J(G_{i+1}, x_i^* + \delta_{i+1}, y). \tag{9}$$

We categorize ensembles in loss and ensembles in logits as bagging ensemble methods. Here, we analyze the relationship between the bagging ensemble and G^{2}BASE.

Algorithm 1 Group selection in G^{2}BASE.

Input: k classifiers $\{f_0, f_1, \dots, f_{k-1}\}$ and their corresponding outputs $\{l_0, l_1, \dots, l_{k-1}\}$; a batch of real examples x and their ground-truth labels y; a threshold h.
Output: Several groups of source models G.

$G_0 = \{\{f_0\}\}$, $g_0 = \|\nabla_\delta J(l_0(x+\delta), y)\|_2$;
for $i = 1$ to $k-1$ **do**
  $g_i = \|\nabla_\delta J(l_i(x+\delta), y)\|_2$.
  for $j = 0$ to $i-1$ **do**
    $d_j(g_i, g_j) = \log(\max\{g_i, g_j\}) - \log(\min\{g_i, g_j\})$.
  end for
  if Exist a group satisfying $d_j \le h$ for all models f_j in the group **then**
    Update G by merging f_i into this group.
  else
    Update G by dividing f_i into a new group.
  end if
end for
return Groups G.

The group selection algorithm is essential in G^{2}BASE. If the threshold h is too large, then all models are divided into the same group, and G^{2}BASE equals the bagging ensemble. When a suitable h is selected, all models in a single group have similar gradient magnitudes, which is similar to indirectly normalizing the gradients and the effects of all models are considered. Moreover, the stacking strategy also boosts the diversity of ensembles. The outputs of different groups are updated in different iterative steps, so the final adversarial example considers many model combinations in different steps and probably crosses the decision boundaries of all ensembled models.

4. Experiments

In this section, we introduce the setup for experiments first. Then, we report the results against diverse undefended and defended models and make comparisons with state-of-the-art benchmarks. Finally, we make some analysis on the proposed methods.

4.1. Setup

Dataset. We conduct experiments on CIFAR-10 [29] and ImageNet [30]. Due to limitation of compute resources, for ImageNet, we evaluate on 2000 randomly chosen images, of which each category contains 2 images. For both datasets, our evaluation is repeated 5 times with different random seeds, and the experimental results are averaged.

Threat Model. We evaluate with both black-box and white-box threat models. In the black-box threat model, the adversarial examples are generated by attacking a source model and then fed into the target model to test the success rate. The white-box threat model can be treated as directly using the target model as the source model. For all methods, the perturbation δ is bounded with the ℓ_∞ norm, and its maximum budget is set to $\varepsilon = 8$ for CIFAR-10 and $\varepsilon = 16$ for ImageNet, with pixel values in [0, 255].

Table 1
Natural accuracy of models on CIFAR-10 and ImageNet.

CIFAR-10	Accuracy (%)	ImageNet	Accuracy (%)
IncV3	93.27	RN50	74.93
DN169	92.84	DN121	74.97
RN18	92.59	VGG19bn	72.89
VGG11bn	91.93	IncV3	76.41
AdvRN20	86.32	Mnas	71.90
AdvRN56	85.46	WRN101	78.37
AWP	85.36	AdvInc	76.80
FS	90.00	AdvIR	79.61
HE	85.14	AdvEnsInc	74.31
AdvWR	85.36	AdvEnsIR	78.85
		AdvRN	55.01
		AdvDe	64.68

Source and target models. We ensemble models with diverse architectures and parameters as the substitute source model to generate adversarial examples. For CIFAR-10, source models contain InceptionV3 (IncV3), DenseNet121 (DN121) [31] and our defense models with adversarial training, AdvResNet20 (AdvRN20) [32] and AdvResNet56 (AdvRN56) [32]. Target models include some pretrained models, such as ResNet18 (RN18) [32] and VGG11 with batch normalization (VGG11bn) [33], and some defense models, such as an adversarial trained WideResNet model (AdvWR) [34]. We also test with some strong defense models including adversarial weight perturbation (AWP) [35], feature scattering (FS) [36] and hypersphere embedding (HE) [37]. For ImageNet, the model pool also includes both normally trained models and defense models. Pretrained models consist of DenseNet121 (DN121) [31], ResNet50 (RN50) [32], InceptionV3 (IncV3) [38] and VGG19 with batch normalization (VGG19bn) [33] and MnasNet (Mnas) [39] and WideResNet101 (WRN101) [34]. Defense models include some adversarially trained models, AdvResnext101Denoise (AdvDe) [40], AdvResNet121 (AdvRN) [40], AdvInceptionV3 (AdvInc) [41] and AdvInceptionResNetV2 (AdvIR) [41]. Moreover, we evaluate with strong defense models through ensemble adversarial training (AdvEnsInc [41] and AdvEnsIR [41]). The natural accuracy of these models is reported in Table 1.

Baselines. We compare the proposed ensemble attack methods with ensembles in loss and ensembles in logits. All these ensemble strategies need to be combined with adversarial attack methods that perform on a single model, including FGSM, PGD and MIM in our experimental setting. For example, when using FGSM as the base method, we attack the ensemble model for only one step. When using MIM, instead, we iteratively attack the ensemble model and compute the gradient with momentum. For G^{2}BASE, the optimization toward constructing the adversary conducted per group is ensembled in logits.

Metric. Following the evaluation method in previous works [11, 13], we compare different attack methods via the nontargeted attack success rate, i.e., the proportion of constructed adversarial examples misclassified by the target model. It is formulated as

$$\frac{1}{n}\sum_{i=1}^{n} \hbar(f_t(x_i) \neq f_t(x_i^*)), \tag{10}$$

where n denotes the number of test data and f_t represents the target model. x_i is the ith test sample and x_i^* is the corresponding adversarial example. $\hbar(\cdot) = 1$ means the adversarial example is misclassified by the target model, and otherwise $\hbar(\cdot) = 0$. In the nontargeted protocol, the attack succeeds if the adversarial examples are misclassified by the target model as any wrong labels. Accordingly, a higher success rate of adversarial examples on the target models represents better attack performance.

Table 2
White-box attack success rate (%±std) on **CIFAR-10**. The ensemble model includes IncV3, DN169, AdvRN20 and AdvRN56.

Base	Ensemble method	IncV3	DN169	AdvRN20	AdvRN56
FGSM	Ensemble in loss	54.48 ± 0.78	48.04 ± 0.34	29.84 ± 0.46	28.81 ± 0.68
	Ensemble in logits	51.27 ± 0.21	44.16 ± 1.15	22.60 ± 0.49	21.71 ± 1.18
	MABE	57.53 ± 0.51	56.88 ± 0.71	62.59 ± 0.61	60.81 ± 0.11
	G^2BASE	**81.16 ± 0.90**	**69.52 ± 0.31**	**70.21 ± 0.29**	**73.22 ± 0.35**
PGD	Ensemble in loss	90.75 ± 0.44	71.19 ± 0.17	20.15 ± 1.05	19.51 ± 0.89
	Ensemble in logits	91.48 ± 0.15	94.22 ± 0.74	19.56 ± 1.08	18.87 ± 0.40
	MABE	91.34 ± 0.90	**94.71 ± 0.62**	79.38 ± 0.10	78.08 ± 0.33
	G^2BASE	**92.21 ± 0.20**	92.50 ± 0.32	**81.83 ± 0.63**	**82.20 ± 0.60**
MIM	Ensemble in loss	90.55 ± 0.23	84.95 ± 0.08	26.87 ± 0.25	25.90 ± 0.16
	Ensemble in logits	89.16 ± 0.09	91.17 ± 0.43	24.60 ± 0.27	23.60 ± 0.34
	MABE	88.61 ± 0.73	90.63 ± 0.52	78.77 ± 0.36	77.55 ± 0.44
	G^2BASE	**94.42 ± 0.25**	**98.16 ± 0.07**	**85.74 ± 0.41**	**85.20 ± 0.63**

Parameters setting. The number of iterations is set as 20, which is widely used in prior adversarial attack experiments. For all experiments, the ensemble weight α_i is set to $\alpha_i = 1/k$, $i = 0, 1, \ldots, k-1$, for k source models. The threshold in Algorithm 1 is set to $h = 0.5$.

4.2. Experimental results

In this subsection, we study the performance of our attack in both white-box and black-box protocols. Considering that the target model may contain a defended model, attackers will naturally ensemble adversarially trained models in the local substitute. To simulate this scenario, we employ diverse models including pretrained models and defense models as our source models.

White-box results. We present the white-box results on two datasets in Tables 2 and 3, respectively. We observe our methods achieve better performance than baselines on all models. More importantly, the proposed methods apparently increase the success rate on defended models. For instance, in Table 3, when using MIM as the base method, the attack success rates of ensembles in logits on two defended ImageNet models are only 69.67% and 65.85%, respectively, while G^2BASE achieves 91.49% and 90.75%. The huge gain comes from the full use of the gradient information of adversarially trained models. The baselines ignore the contribution of gradient information from defended models due to their small magnitude. In contrast, by directly (MABE) or implicitly (G^2BASE) redistributing the contributions of different models to the final ensemble gradient, our methods outperform in all cases.

Black-box results. To simulate the black-box protocol, we first perform attacks on the ensemble source model to craft adversarial examples. The resultant examples are then fed into unseen target models with different architectures and parameters. As shown in Tables 4 and 5, our methods consistently improve the transferability to a great extent and defeat all the other benchmark methods with a significant margin. In the CIFAR-10 experiments, MABE and G^2BASE achieve similar results, largely surpassing baselines. Even against defense models, the attack transferability achieves an improvement of around 5%. In the ImageNet experiments, our G^2BASE always ranks first among the four methods, demonstrating its effectiveness in boosting transferability. For example, when using FGSM as the base method, G^2BASE increase the attack success rate of ensembles in logits on AdvEnsIR by 17.96%. Moreover, with PGD as the base method, the proposed G^2BASE obtains the highest success rate of 73.92% on AdvEnsIR, surpassing the result of the ensemble in logits, 30.37%.

Visualization results. Fig. 4 displays one original image in ImageNet and the corresponding adversarial examples generated by running the proposed MABE and G^2BASE on the ensemble model. We show that the manipulations to the clean image are hardly visible. Therefore, our ensemble methods can obtain high attack performance while ensuring imperceptibility.

Fig. 5 displays four original images and their corresponding adversarial perturbations generated with ensembles in logits and the proposed G^2BASE on the ensemble model. For G^2BASE, we observe that the resulting perturbations are perceptually similar to the original images, i.e., they have salient semantic features of ground-truth images. On the other hand, for ensembles in logits, the perturbations often look more similar to random noises. A previous study found that adversarial examples generated by attacking adversarial robust models possess similar high-level features with corresponding natural images. Thus, our method possesses the potential to fully leverage the gradient information of adversarially trained models and extract transferable features in the attacking process.

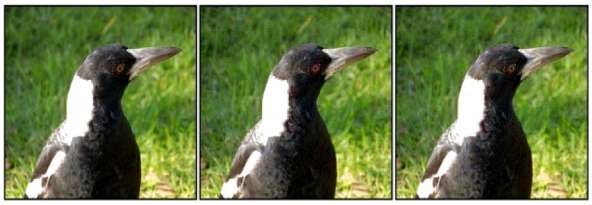

Fig. 4. One original image (left) and its corresponding adversarial examples generated by running the proposed MABE (middle) and G^2BASE (right) on the ensemble model composed of RN50, DN121, VGG19bn and IncV3.

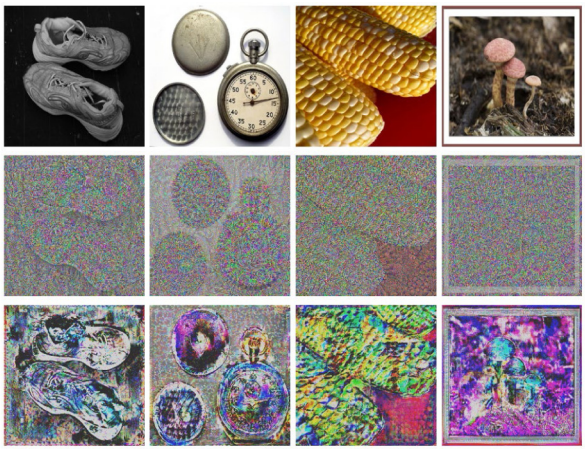

Fig. 5. Original images (top) and corresponding adversarial perturbations (×5 for better view, actually they are imperceptible), generated by running ensemble in logits (middle) and the proposed G^2BASE (bottom) on the ensemble model composed of RN50, DN121, AdvRN and AdvDe. Both ensemble methods are based on MIM.

4.3. Analysis of the proposed methods

We have shown the effectiveness of our methods to achieve high attack success rate while maintaining imperceptibility and the superior performance compared with baseline methods. In this subsection, we

Table 3
White-box attack success rate (%±std) on **ImageNet**. The ensemble model includes RN50, DN121, AdvRN and AdvDe.

Base	Ensemble method	RN50	DN121	AdvRN	AdvDe
FGSM	Ensemble in loss	90.01 ± 0.96	88.37 ± 0.73	64.80 ± 1.05	52.23 ± 1.27
	Ensemble in logits	91.50 ± 1.25	91.97 ± 0.38	56.18 ± 0.92	48.49 ± 1.66
	MABE	90.96 ± 0.69	90.09 ± 0.71	**77.52 ± 1.48**	**75.41 ± 0.59**
	G^2BASE	**93.35 ± 0.90**	**93.46 ± 0.49**	72.25 ± 0.80	68.70 ± 1.35
PGD	Ensemble in loss	99.89 ± 0.09	99.67 ± 0.13	61.04 ± 1.46	52.45 ± 1.00
	Ensemble in logits	**100.00 ± 0.00**	**100.00 ± 0.00**	61.54 ± 0.84	55.17 ± 1.58
	MABE	**100.00 ± 0.00**	**100.00 ± 0.00**	90.02 ± 0.88	**89.97 ± 1.08**
	G^2BASE	**100.00 ± 0.00**	**100.00 ± 0.00**	**90.53 ± 0.72**	89.87 ± 0.88
MIM	Ensemble in loss	99.97 ± 0.03	99.90 ± 0.05	67.17 ± 0.98	57.43 ± 1.47
	Ensemble in logits	**100.00 ± 0.00**	**100.00 ± 0.00**	69.67 ± 1.33	65.85 ± 1.44
	MABE	**100.00 ± 0.00**	**100.00 ± 0.00**	89.63 ± 0.82	89.29 ± 1.06
	G^2BASE	**100.00 ± 0.00**	**100.00 ± 0.00**	**91.49 ± 0.51**	**90.75 ± 0.85**

Table 4
Black-box attack success rate (%±std) on **CIFAR-10**. The source model ensembles IncV3, DN169, AdvRN20 and AdvRN56.

Base	Ensemble method	AdvWR	AWP	FS	HE	VGG11bn	RN18
FGSM	Ensemble in loss	17.60 ± 0.45	17.26 ± 0.64	13.05 ± 0.35	16.30 ± 0.48	36.39 ± 0.74	40.05 ± 0.49
	Ensemble in logits	16.95 ± 0.43	16.73 ± 0.48	12.46 ± 1.07	15.52 ± 0.73	33.09 ± 0.99	36.26 ± 0.63
	MABE	21.00 ± 0.56	20.71 ± 0.44	17.77 ± 1.07	19.96 ± 0.64	46.15 ± 0.37	50.41 ± 0.91
	G^2BASE	**21.43 ± 0.16**	**21.09 ± 0.11**	**18.44 ± 1.31**	**20.30 ± 0.67**	**61.34 ± 0.49**	**65.78 ± 0.58**
PGD	Ensemble in loss	16.08 ± 0.13	15.59 ± 0.52	11.37 ± 0.41	14.90 ± 1.19	28.20 ± 1.02	42.16 ± 0.86
	Ensemble in logits	16.13 ± 0.92	15.64 ± 0.56	11.39 ± 0.94	14.91 ± 0.93	32.28 ± 0.71	50.72 ± 0.79
	MABE	**21.22 ± 0.16**	**20.88 ± 0.28**	**18.13 ± 0.43**	**20.10 ± 0.84**	58.75 ± 0.54	**71.27 ± 0.13**
	G^2BASE	21.08 ± 0.16	20.74 ± 0.14	18.05 ± 0.73	19.95 ± 0.64	**60.27 ± 0.42**	71.19 ± 0.14
MIM	Ensemble in loss	17.11 ± 0.36	16.66 ± 0.30	12.32 ± 0.55	15.83 ± 0.11	49.70 ± 0.93	65.72 ± 0.36
	Ensemble in logits	16.89 ± 0.46	16.44 ± 0.51	12.28 ± 0.18	15.62 ± 0.44	50.07 ± 0.75	66.89 ± 0.74
	MABE	**21.66 ± 0.23**	**21.44 ± 0.30**	**18.61 ± 0.53**	**20.48 ± 0.12**	62.59 ± 0.69	73.36 ± 0.54
	G^2BASE	21.37 ± 0.43	21.24 ± 0.51	**18.61 ± 0.61**	20.28 ± 0.53	**65.27 ± 0.57**	**76.70 ± 0.65**

Table 5
Black-box attack success rate (%±std) on **ImageNet**. The source model ensembles RN50, DN121, AdvRN and AdvDe.

Base	Ensemble method	AdvInc	AdvIR	AdvEnsInc	AdvEnsIR	VGG19bn	IncV3
FGSM	Ensemble in loss	48.36 ± 0.94	44.96 ± 0.69	40.16 ± 1.54	32.71 ± 0.68	74.83 ± 0.56	62.03 ± 0.43
	Ensemble in logits	45.17 ± 1.43	42.31 ± 0.84	38.18 ± 1.17	31.23 ± 0.77	74.84 ± 0.87	60.07 ± 0.68
	MABE	**64.84 ± 1.56**	**60.36 ± 1.44**	55.43 ± 1.07	46.81 ± 1.64	78.04 ± 0.97	**71.47 ± 0.94**
	G^2BASE	61.49 ± 1.16	57.54 ± 0.61	**56.19 ± 1.31**	**49.19 ± 1.56**	**78.29 ± 0.69**	70.80 ± 0.98
PGD	Ensemble in loss	36.39 ± 0.39	33.95 ± 0.30	36.56 ± 0.64	31.37 ± 0.58	86.66 ± 0.79	60.02 ± 0.94
	Ensemble in logits	38.65 ± 1.45	32.84 ± 0.36	35.22 ± 0.88	30.37 ± 0.72	85.88 ± 0.59	58.93 ± 0.57
	MABE	67.69 ± 1.11	61.65 ± 1.20	66.61 ± 0.84	60.21 ± 1.84	**88.59 ± 0.91**	77.19 ± 0.68
	G^2BASE	**80.54 ± 1.16**	**76.66 ± 1.04**	**78.94 ± 1.11**	**73.92 ± 1.53**	87.18 ± 1.01	**81.95 ± 0.82**
MIM	Ensemble in loss	47.86 ± 0.99	46.72 ± 0.98	43.46 ± 0.54	36.17 ± 0.88	92.38 ± 0.93	75.35 ± 0.36
	Ensemble in logits	51.99 ± 0.96	50.59 ± 1.51	46.49 ± 1.11	39.16 ± 0.64	**94.82 ± 0.85**	80.34 ± 0.76
	MABE	73.00 ± 1.70	69.10 ± 1.30	70.05 ± 1.55	63.12 ± 1.88	89.30 ± 0.64	82.51 ± 0.34
	G^2BASE	**77.29 ± 0.66**	**72.09 ± 1.01**	**72.37 ± 1.63**	**65.12 ± 2.03**	92.35 ± 0.57	**85.11 ± 0.65**

take a further step to analyze the mechanism of our methods. We first explore the reason of performance boosting with our methods and then study effects of key hyperparameters in our methods.

Why are our methods effective? In transfer-based black-box attack, the key assumption is that, if an example is adversarial for multiple source models, this example remains adversarial for other unseen target models. Therefore, the cornerstone of strong transferability is a high attack success rate on the source model. But we have shown that baseline ensemble methods cannot fool the defended source model with a satisfactory success rate. The reason is that the baseline methods does not utilize gradient information from defended models due to their small magnitude. Our methods remedy this problem and can therefore gain the improvement.

For comparison, we then conduct an experiment by employing four undefended models as local source models. Different from previous setting, the four undefended models have similar gradient magnitudes. According to our analysis, due to the similarity of gradient magnitudes, each single model should contribute to the final ensemble result. Results in Tables 6 and 7 are in line with our expectations. We find our methods obtain similar results with baselines.

On the other hand, using undefended models as source models leads to low transferability on defended models. For example, all methods obtain a success rate lower than 40% on AdvEnsIR. This demonstrate the necessity of employing diverse models containing defense models as the source model. In this scenario, the baselines behave poor while our methods achieve a superior attack success rate.

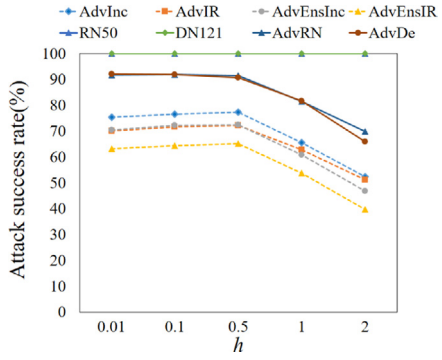

Fig. 6. The effect of hyperparameter h on attack success rates.

Table 6
White-box attack success rate (%±std). The source model ensembles **four undefended models**, RN50, DN121, VGG19bn and IncV3.

Base	Ensemble method	RN50	DN121	VGG19bn	IncV3
FGSM	Ensemble in loss	85.24 ± 0.81	84.18 ± 1.07	84.04 ± 0.51	71.06 ± 1.34
	Ensemble in logits	88.23 ± 0.72	89.03 ± 0.62	**93.20 ± 0.35**	**81.26 ± 0.49**
	MABE	**88.49 ± 0.86**	88.33 ± 0.87	91.88 ± 0.32	79.30 ± 0.45
	G^2BASE	88.31 ± 1.31	**89.20 ± 0.90**	93.18 ± 0.48	81.16 ± 0.39
PGD	Ensemble in loss	99.39 ± 0.26	99.07 ± 0.32	99.29 ± 0.24	96.98 ± 0.37
	Ensemble in logits	**100.00 ± 0.00**	99.99 ± 0.01	**99.99 ± 0.01**	99.87 ± 0.13
	MABE	**100.00 ± 0.00**	**100.00 ± 0.00**	**99.99 ± 0.01**	**99.99 ± 0.01**
	G^2BASE	**100.00 ± 0.00**	99.99 ± 0.01	99.94 ± 0.06	99.40 ± 0.15

Table 7
Black-box attack success rate (%±std). The source model ensembles **four undefended models**, RN50, DN121, VGG19bn and IncV3.

Base	Ensemble method	AdvInc	AdvIR	AdvEnsInc	AdvEnsIR	Mnas	WRN101
FGSM	Ensemble in loss	43.29 ± 1.46	41.73 ± 1.32	37.65 ± 1.25	30.84 ± 0.51	74.02 ± 0.63	65.23 ± 0.52
	Ensemble in logits	**44.47 ± 1.33**	**43.50 ± 1.30**	39.03 ± 1.32	31.79 ± 0.51	75.62 ± 0.18	67.58 ± 0.77
	MABE	44.41 ± 0.89	43.33 ± 1.37	38.40 ± 1.70	31.72 ± 0.37	75.54 ± 1.06	66.86 ± 0.99
	G^2BASE	44.38 ± 1.42	43.44 ± 1.21	**39.09 ± 1.26**	**31.85 ± 0.55**	**75.73 ± 0.33**	**67.70 ± 0.90**
PGD	Ensemble in loss	35.98 ± 0.87	35.16 ± 0.54	36.16 ± 1.04	30.10 ± 0.60	85.54 ± 0.54	84.40 ± 1.65
	Ensemble in logits	36.06 ± 1.19	32.25 ± 0.60	33.36 ± 1.54	28.09 ± 0.76	85.27 ± 0.83	85.22 ± 0.88
	MABE	36.03 ± 0.82	35.40 ± 0.40	36.54 ± 1.06	30.61 ± 1.04	84.10 ± 0.45	84.18 ± 1.17
	G^2BASE	**37.69 ± 0.46**	**37.44 ± 0.56**	**37.82 ± 0.93**	**31.86 ± 0.89**	**89.98 ± 0.68**	**91.27 ± 0.58**

Effects of parameters. G^2BASE is the most effective among the tested ensemble methods. The threshold h is the dominant hyperparameter in G^2BASE, and here, we explore its effect on the attack performance. Specifically, we vary h while fixing the other parameters when generating adversarial examples. Similar to previous experiments, we report the attack success rate of target models on crafted adversarial images to measure the attack effectiveness.

Fig. 6 illustrates the effect of h on attack success rates against diverse models, with adversarial examples generated on the ensemble model composed of RN50, DN121, AdvRN and AdvDe. We observe that when h changes from 0.01 to 0.5, the attack success rate shows a mild improvement. When h increases, the attack success rate drops dramatically. Note that when h is large, e.g., $h = 2$, all models are in the same group so that G^2BASE degrades as the bagging ensemble method.

5. Conclusions

In this work, we discover that gradient magnitude is of vital importance to ensemble adversarial attacks. Based on this discovery, we propose two novel ensemble strategies to balance the effects of different models in the ensemble process. Consequently, the proposed methods take full advantage of the gradient information of each model in the ensemble, resulting in a significant boosting of the attack success rate. We conduct extensive experiments to validate the effectiveness of our approach and confirm its superiority to state-of-the-art baselines. Therefore, our attack can serve as a strong benchmark of ensemble adversarial attacks to measure the robustness of defense models.

CRediT authorship contribution statement

Ziwen He: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Wei Wang:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. **Jing Dong:** Resources, Supervision, Writing – review & editing, Project administration. **Tieniu Tan:** Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61972395.

References

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[2] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[3] Karen Simonyan, Andrew Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[4] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, Explaining and harnessing adversarial examples, in: Proceedings of the 3rd International Conference on Learning Representations, 2015.
[5] Christian Szegedy, et al., Intriguing properties of neural networks, in: Proceedings of the 2nd International Conference on Learning Representations, 2014.
[6] Yinpeng Dong, et al., Efficient decision-based black-box adversarial attacks on face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7706–7714.
[7] Yulong Cao, et al., Adversarial sensor attack on lidar-based perception in autonomous driving, in, in: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 2267–2281.
[8] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael backes and patrick mcdaniel, adversarial examples for malware detection, in: Proceedings of the European Symposium on Research in Computer Security, 2017, pp. 62–79.
[9] Kevin Eykholt, et al., Robust physical-world attacks on deep learning visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.
[10] Jianbo Chen, Michael I. Jordan, Martin J. Wainwright, HopSkipJumpAttack: A query-efficient decision-based attack, 2019, arXiv Prepr. arXiv1904.02144.
[11] Yinpeng Dong, et al., Boosting adversarial attacks with momentum, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185–9193.
[12] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad shahbaz khan and fatih porikli, a self-supervised approach for adversarial robustness, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 259–268.
[13] Yanpei Liu, Xinyun Chen, Chang Liu, Dawn Song, Delving into transferable adversarial examples and black-box attacks, in, in: Proceedings of the 5th International Conference on Learning Representations, 2017.
[14] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha ans Z. Berkay Celik, Ananthram Swami, Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, 2017, pp. 506–519.

[15] Nicolas Papernot, Patrick McDaniel and ian goodfellow, in: Transferability in Machine Learning: From Phenomena to Black-Box Attacks using Adversarial Samples, 2016, arXiv, abs/1605.0.
[16] Cihang Xie, et al., Improving transferability of adversarial examples with input diversity, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2730–2739.
[17] Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Evading defenses to transferable adversarial examples by translation-invariant attacks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4312–4321.
[18] Lin Jiadong, Song Chuanbiao, He Kun, Wang Liwei, John E. Hopcroft, Nesterov accelerated gradient and scale invariance for adversarial attacks, in: Proceedings of the 8th International Conference on Learning Representations, 2020.
[19] Alexey Kurakin, et al., Adversarial attacks and defences competition, 2018, pp. 195–231.
[20] Leo Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[21] Aleksander Madry, Aleksandar Makelov, Ludwigand Dimitris Tsipras Schmidt, Adrian Vladu, Towards deep learning models resistant to adversarial attacks, in: Proceedings of the 6th International Conference on Learning Representations, 2018.
[22] Anders Krogh, Jesper Vedelsby, Neural network ensembles, cross validation, and active learning, in: Proceedings of Advances in Neural Information Processing Systems, 1995, pp. 231–238.
[23] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, Alex Ksikes, Ensemble selection from libraries of models, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, p. 18.
[24] Lars Kai Hansen, Peter Salamon, Neural network ensembles, IEEE Trans. Pattern Anal. Mach. Intell. 12 (10) (1990) 993–1001.
[25] David H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.
[26] Jie Hang, Keji Han, Hui Chen, Yun Li, Ensemble adversarial black-box attacks against deep learning systems, Pattern Recognit. (2020).
[27] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, Alan L. Yuille, Learning transferable adversarial examples via ghost networks, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020, pp. 11458–11465.
[28] Zhaohui Che, Ali Borji, Guangtao Zhai, Suiyi Ling, Jing Li, Patrick Le Callet, A new ensemble adversarial attack powered by long-term gradient memories, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020, pp. 3405–3413.
[29] Alex Krizhevsky, Geoffrey Hinton, et al., Learning multiple layers of features from tiny images, 2009.
[30] Olga Russakovsky, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[31] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[33] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the 3rd International Conference on Learning Representations, 2015.
[34] Sergey Zagoruyko, Nikos Komodakis, Wide residual networks, in: Proceedings of British Machine Vision Conference, 2016, pp. 1–12.
[35] Dongxian Wu, Shu Tao Xia, Yisen Wang, Adversarial weight perturbation helps robust generalization, in: Advances in Neural Information Processing Systems, 2020.
[36] Haichao Zhang, Jianyu Wang, Defense against adversarial attacks using feature scattering-based adversarial training, Adv. Neural Inf. Process. Syst. (2019).
[37] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, Hang Su, Boosting adversarial training with hypersphere embedding, in: Advances in Neural Information Processing Systems, 2020.
[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[39] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710.
[40] Cihang Xie, Yuxin Wu, Laurens Van Der Maaten, Alan L. Yuille, Kaiming He, Feature denoising for improving adversarial robustness, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 501–509.
[41] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel, Ensemble adversarial training: Attacks and defenses, in: Proceedings of the 6th International Conference on Learning Representations, 2018.

Ziwen He received B.Eng. degree in Shanghai Jiao Tong University, China in 2018. He is a Ph.D. degree candidate in the Center for Research on Intelligent Perception and Computing (CRIPAC) at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China.

His current research focuses on adversarial examples and computer vision.

E-mail: ziwen.he@cripac.ia.ac.cn

Wei Wang received his B.E. in Computer Science and Technology from North China Electric Power University, China in 2007. Since July 2012, Prof. Wang has joined the National Laboratory of Pattern Recognition (NLPR). He is currently an associate professor.

His research interests are pattern recognition, image processing and digital image forensics, including watermarking, steganalysis and tampering detection.

E-mail: wwang@nlpr.ia.ac.cn

Jing Dong received her Ph.D. in Pattern Recognition from the Institute of Automation, Chinese Academy of Sciences, China in 2010. Then, she joined the National Laboratory of Pattern Recognition (NLPR), and she is currently an Associate Professor.

Her research interests are toward pattern recognition, image processing and digital image forensics, including digital watermarking, steganalysis and tampering detection. She is a senior member of IEEE. She also has served as the deputy general of the Chinese Association for Artificial Intelligence.

E-mail: jdong@nlpr.ia.ac.cn

Tieniu Tan received MS and Ph.D. degrees in electronic engineering from the Imperial College of Science, Technology and Medicine, London, United Kingdom, in 1986 and 1989, respectively. In January 1998, he returned to China to join the Institute of Automation of the Chinese Academy of Sciences (CAS). He is currently a professor and the director of the Center for Research on Intelligent Perception and Computing (CRIPAC). He is also a fellow of CAS, TWAS, IEEE and IAPR and an International Fellow of the UK Royal Academy of Engineering. He was founding chair of the IAPR Technical Committee on Biometrics, the IAPR-IEEE International Conference on Biometrics, the IEEE International Workshop on Visual Surveillance and Asian Conference on Pattern Recognition (ACPR). He has published more than 500 research papers in refereed journals and conferences in the areas of image processing, computer vision, and pattern recognition.

His research interests include biometrics, image and video understanding, and information forensics and security.

E-mail: tnt@nlpr.ia.ac.cn