Machine Learning (2024) 113:2301–2326
https://doi.org/10.1007/s10994-023-06328-7

Imbalanced gradients: a subtle cause of overestimated adversarial robustness

Xingjun Ma[1] · Linxi Jiang[1] · Hanxun Huang[2] · Zejia Weng[1] · James Bailey[2] · Yu-Gang Jiang[1]

Received: 14 November 2021 / Revised: 19 December 2022 / Accepted: 10 March 2023 /
Published online: 30 March 2023

Abstract

Evaluating the robustness of a defense model is a challenging task in adversarial robustness research. Obfuscated gradients have previously been found to exist in many defense methods and cause a false signal of robustness. In this paper, we identify a more subtle situation called *Imbalanced Gradients* that can also cause overestimated adversarial robustness. The phenomenon of imbalanced gradients occurs when the gradient of one term of the margin loss dominates and pushes the attack towards to a suboptimal direction. To exploit imbalanced gradients, we formulate a *margin decomposition (MD)* attack that decomposes a margin loss into individual terms and then explores the attackability of these terms separately via a two-stage process. We also propose a multi-targeted and ensemble version of our MD attack. By investigating 24 defense models proposed since 2018, we find that 11 models are susceptible to a certain degree of imbalanced gradients and our MD attack can decrease their robustness evaluated by the best standalone baseline attack by more than 1%. We also provide an in-depth investigation on the likely causes of imbalanced gradients and effective countermeasures.

Keywords Adversarial attack · Robustness evaluation · Deep neural networks

Editors: Dana Drachsler Cohen, Javier Garcia, Mohammad Ghavamzadeh, Marek Petrik, Philip S. Thomas.

Xingjun Ma and Linxi Jiang have contributed equally to this work.

✉ Hanxun Huang
hanxunh@student.unimelb.edu.au

✉ Yu-Gang Jiang
ygj@fudan.edu.cn

Extended author information available on the last page of the article

1 Introduction

Deep neural networks (DNNs) are vulnerable to adversarial examples, which are input instances crafted by adding small adversarial perturbations to natural examples. Adversarial examples can fool DNNs into making false predictions with high confidence, and transfer across different models (Szegedy et al., 2014; Goodfellow et al., 2015). A number of defenses have been proposed to overcome this vulnerability. However, a concerning fact is that many defenses have been quickly shown to have undergone incorrect or incomplete evaluation (Athalye et al., 2018; Engstrom et al., 2018; Carlini et al., 2019; Tramer et al., 2020; Croce & Hein, 2020b). One common pitfall in adversarial robustness evaluation is the phenomenon of gradient masking (Papernot et al., 2017; Tramèr et al., 2018) or obfuscated gradients (Athalye et al., 2018), leading to weak or unsuccessful attacks and false signals of robustness. To demonstrate “real" robustness, newly proposed defenses claim robustness based on results of white-box attacks such as Projected Gradient Decent (PGD) attack (Madry et al., 2018) and AutoAttack (Croce & Hein, 2020b; Croce et al., 2020), and demonstrate that they are not a result of obfuscated gradients. In this work, we show that the robustness may still be overestimated even when there are no obfuscated gradients. Specifically, we identify a subtle situation called *Imbalanced Gradients* that exists in several recent defense models and can cause highly overestimated robustness.

Imbalanced gradients is a new type of gradient masking effect where the gradient of one loss term dominates that of other terms. This causes the attack to move toward a suboptimal direction. Different from obfuscated gradients, imbalanced gradients are more subtle and are not detectable by the detection methods used for obfuscated gradients. To exploit imbalanced gradients, we propose a novel attack named *margin decomposition (MD)* attack that decomposes the margin loss into two separate terms and then exploits the attackability of these terms via a two-stage attacking process. We also derive the MultiTargeted (Gowal et al., 2019) and ensemble variants of MD attack, following AutoAttack. By examining the robustness of 24 adversarial training based defense models proposed since 2018. We find that 11 of them are susceptible to imbalanced gradients to a certain extent, and their robustness evaluated by the best baseline standalone attack drops by more than 1% against our MD attack. Our key contributions are:

- We identify a new type of effect called *Imbalanced Gradients*, which can cause overestimated adversarial robustness and cannot be detected by detection methods for obfuscated gradients. Especially, we highlight that label smoothing is one of the major causes of imbalanced gradients.
- We propose *margin decomposition (MD)* attacks to exploit imbalanced gradients. MD leverages the attackability of the individual terms in the margin loss in a two-stage attacking process. We also introduce the MultiTargeted and ensemble variants of MD.
- We conduct extensive evaluations on 24 state-of-the-art defense models and find that 11 of them suffer from imbalanced gradients to some extent and their robustness evaluated by the best standalone attack drops by more than 1% against our MD attack. Our MD ensemble (MDE) attack exceeds state-of-the-art attack AutoAt-

tack on 16/20 defense models on CIFAR-10. Our MD attack alone can outperform AutoAttack in evaluating the adversarial robustness of vision transformer and ResNet-50 models trained on ImageNet.

2 Related work

We denote a clean sample by $\boldsymbol{x}$, its class by $y \in \{1, \cdots, C\}$ with C the number of classes, and a DNN classifier by f. The probability of $\boldsymbol{x}$ being in the i-th class is computed as $\boldsymbol{p}_i(\boldsymbol{x}) = e^{z_i} / \sum_{j=1}^{C} e^{z_j}$, where z_i is the logits for the i-th class. The goal of an adversarial attack is to find an adversarial example $\boldsymbol{x}_{adv}$ that can fool the model into making a false prediction (e.g., $f(\boldsymbol{x}_{adv}) \neq y$), and is typically restricted to be within a small ϵ-ball around the original sample $\boldsymbol{x}$ (e.g., $\|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_\infty \leq \epsilon$).

Adversarial attack Adversarial examples can be crafted by maximizing a classification loss ℓ by one or multiple steps of adversarial perturbations. For example, the one-step Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) and the iterative FGSM (I-FGSM) attack (Kurakin et al., 2017). Projected Gradient Descent (PGD) (Madry et al., 2018) attack is another iterative method that projects the perturbation back onto the ϵ-ball centered at $\boldsymbol{x}$ when it goes beyond. Carlini and Wagner (CW) (Carlini & Wagner, 2017) attack generates adversarial examples via an optimization framework. There also exist other attacks such as Frank-Wolfe attack (Chen et al., 2018a), distributionally adversarial attack (Zheng et al., 2019) and elastic-net attacks (Chen et al., 2018b). In earlier literature, the most commonly used attacks for robustness evaluations are FGSM, PGD, and CW attacks.

Several recent attacks have been proposed to produce more accurate robustness evaluations than PGD. This includes fast adaptive boundary attack (FAB) (Croce & Hein, 2019), MultiTargeted (MT) attack (Gowal et al., 2019), Output Diversified Initialization (ODI) (Tashiro et al., 2020), and AutoAttack (AA) (Croce & Hein, 2020b). FAB finds the minimal perturbation necessary to change the class of a given input. MT (Gowal et al., 2019) is a PGD-based attack with multiple restarts and picks a new target class at each restart. ODI provides a more effective initialization strategy with diversified logits. AA attack is a parameter-free ensemble of four attacks: FAB, two Auto-PGD attacks, and the black-box Square Attack (Andriushchenko et al., 2019). AA has demonstrated to be one of the state-of-the-art attacks to date, according to the RobustBench (Croce et al., 2020).

Adversarial loss Many attacks use cross entropy (CE) as the adversarial loss: $\ell_{ce}(\boldsymbol{x}, y) = -\log \boldsymbol{p}_y$. The other commonly used adversarial loss is the margin loss (Carlini & Wagner, 2017): $\ell_{margin}(\boldsymbol{x}, y) = z_{max} - z_y$, with $z_{max} = \max_{i \neq y} z_i$. Shown in Gowal et al. (2019), CE can be written in a margin form (e.g., $\ell_{ce}(\boldsymbol{x}, y) = \log(\sum_{i=1}^{C} e^{z_i}) - z_y$), and in most cases, they are both effective. While FGSM and PGD attacks use the CE loss, CW and several recent attacks such as MT and ODI adopt the margin loss. AA has one PGD variant using the CE loss and the other PGD variant using the Difference of Logits Ratio (DLR) loss. DLR can be regarded as a "relative margin" loss. In this paper, we identify a new effect that causes overestimated adversarial robustness from the margin loss perspective and propose new attacks by decomposing the margin loss.

Adversarial defense In response to the threat of adversarial attacks, many defenses have been proposed such as defensive distillation (Papernot et al., 2016), feature/subspace analysis (Xu et al., 2017; Ma et al., 2018), denoising techniques (Guo et al., 2018; Liao et al., 2018; Samangouei et al., 2018), robust regularization (Gu & Rigazio, 2014; Tramèr et al., 2018; Ross & Doshi-Velez, 2018), model compression (Liu et al., 2018; Das et al., 2018; Rakin et al., 2018) and adversarial training (Goodfellow et al., 2015; Madry et al., 2018). Among them, adversarial training via robust min-max optimization has been found to be the most effective approach (Athalye et al., 2018). The standard adversarial training (SAT) (Madry et al., 2018) trains models on adversarial examples generated via the PGD attack. Dynamic adversarial training (Dynamic) (Wang et al., 2019) trains on adversarial examples with gradually increased convergence quality. Max-Margin Adversarial training (MMA) (Ding et al., 2018) trains on adversarial examples with gradually increased margin (e.g., the perturbation bound ϵ). Jacobian adversarially regularized networks (JARN) adversarially regularize the Jacobian matrices, and can be combined with 1-step adversarial training (JARN-AT1) to gain additional robustness (Chan et al., 2020). Sensible adversarial training (Sense) (Kim & Wang, 2020) trains on loss-sensible adversarial examples (perturbation stops when loss exceeds certain threshold). Adversarial training with pre-training (AT-PT) (Hendrycks et al., 2019) uses pre-training to improve model robustness. Adversarial training with early stopping (AT-ES) (Rice et al., 2020) suggests the use of early stopping to avoid the robust overfitting of adversarial training. Bilateral adversarial training (Bilateral) (Wang & Zhang, 2019) trains on PGD adversarial examples with adversarially perturbed labels. Adversarial interpolation (Adv-Interp) training (Zhang & Xu, 2020) trains on adversarial examples generated under an adversarial interpolation scheme with adversarial labels. Feature scattering-based (FeaScatter) adversarial training (Zhang & Wang, 2019) crafts adversarial examples using latent space feature scattering, then trains on these examples with label smoothing. TRADES (Zhang et al., 2019) replaces the CE loss of SAT by the KL divergence for a better trade-off between robustness and natural accuracy. Based on TRADES, RST (Carmon et al., 2019) and UAT (Alayrac et al., 2019) improve robustness by training with 10× more unlabeled data. Misclassification Aware adveRsarial Training (MART) (Wang et al., 2020) further improves the above three methods with a misclassification aware loss function. Adversarial weight perturbation (AWP) (Wu et al., 2020) perturbs inputs and model weights alternatively during adversarial training to improve robust generalization. Channel-wise activation suppressing (CAS) robustifies the intermediate layers of DNNs via an auxiliary channel suppressing module (Bai et al., 2020). There are also recent works on robust neural architectures (Shao et al., 2021; Du et al., 2021; Tang et al., 2021; Huang et al., 2021) and adversarial robustness distillation (Goldblum et al., 2020; Zhu et al., 2021; Zi et al., 2021). We will discuss and evaluate a set of the above adversarial training-based defenses in Sect. 5.

Evaluation of Adversarial Robustness Adversarial robustness requires careful and rigorous evaluation. Many defenses that perform incomplete evaluation are quickly broken by new attacks. Several evaluation pitfalls have been identified as needing to be avoided for reliable robustness evaluation (Carlini et al., 2019). Although several general principles have been suggested around the regular attacks such as PGD (Carlini

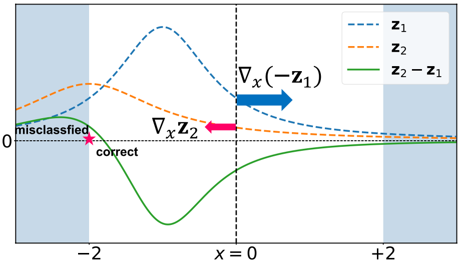


Fig. 1 A toy illustration of *Imbalanced Gradients* at $x = 0$: the gradient of margin loss $(z_2 - z_1)$ is dominated by its $-z_1$ term (i.e., $\nabla_x z_2$ will be canceled out by $\nabla_x(-z_1)$), pointing to a suboptimal attack direction towards +2, where x is still correctly classified

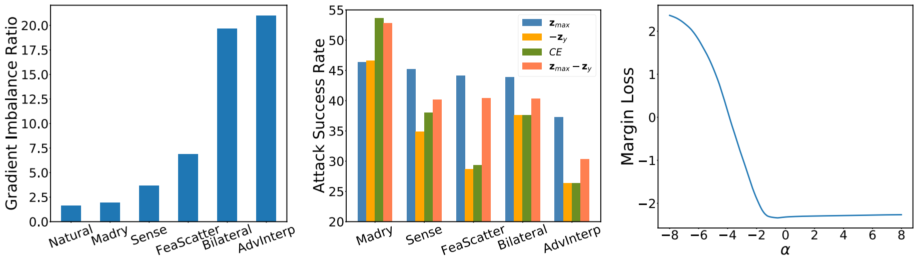
et al., 2019), there are scenarios where these attacks may give unreliable robustness evaluation. Gradient masking (Tramèr et al., 2018; Papernot et al., 2017) is a common effect that blocks the attack by hiding useful gradient information. Obfuscated gradients (Athalye et al., 2018), a type of gradient masking, has been exploited (unintentionally) by many defense methods to cause an overly optimistic evaluation of robustness. Obfuscated gradients exist in different forms such as non-differentiable gradients, stochastic gradients, or vanishing/exploding gradients. Such defenses have all been successfully circumvented by adaptive attacks in (Athalye et al., 2018; Carlini et al., 2019; Tramer et al., 2020). Gradient-free attacks such as SPSA (Spall et al., 1992) have also been used to identify obscured models (Uesato et al., 2018). In this paper, we identify a more subtle situation called *Imbalanced Gradients*, which also causes overestimated robustness, but is different from obfuscated gradients.

3 Imbalanced gradients and robustness evaluation

We first give a toy example of imbalanced gradients and show how regular attacks can fail in such a situation. We then empirically verify their existence in deep neural networks, particularly for some adversarially-trained models. Finally, we propose the margin decomposition attacks to exploit the imbalanced gradients. Since CE and margin loss are the two commonly used loss functions for adversarial attack and CE can be written in a margin form (Gowal et al., 2019), here we focus on the margin loss to present the phenomenon of imbalanced gradients.

Imbalanced gradients The gradient of the margin loss (e.g., $\ell_{margin}(\boldsymbol{x}, y) = z_{max} - z_y$) is the combination of the gradients of its two individual terms (e.g., $\nabla_{\boldsymbol{x}}(z_{max} - z_y) = \nabla_{\boldsymbol{x}} z_{max} + \nabla_{\boldsymbol{x}}(-z_y)$). *Imbalanced Gradients* is the situation where the gradient of one loss term dominates that of other term(s), pushing the attack towards a suboptimal direction.

Toy example Consider a one-dimensional classification task and a binary classifier with two outputs z_1 and z_2 (like logits of a DNN), Fig. 1 illustrates the distributions of z_1, z_2 and $z_2 - z_1$ around $x = 0$. The classifier predicts class 1 when $z_1 \geq z_2$, otherwise class 2. We consider an input at $x = 0$ with correct prediction $y = 1$, and a maximum perturbation constraint $\epsilon = 2$ (e.g., perturbation $\delta \in [-2, +2]$). The attack is successful if and only if $z_2 > z_1$. In this example, imbalanced gradients occurs at $x = 0$, where the gradients of the two terms $\nabla_x z_2$ and $\nabla_x(-z_1)$ have opposite directions, and the attack is dominated by the z_1 term as $\nabla_x(-z_1)$ is significantly larger than $\nabla_x z_2$. Thus, attacking x with the margin loss will converge to +2, where the sample is still correctly classified. However, for a successful



(a) (b) (c)

Fig. 2 **a** Gradient imbalance ratio of 5 models. **b** Attack success rate of PGD-20 with different losses. **c** The margin loss of the advinterp defense model on points $\boldsymbol{x}^* = \boldsymbol{x} + \alpha \cdot \mathrm{sign}\,(\nabla_{\boldsymbol{x}}(-z_y))$, where $\boldsymbol{x}$ is a natural sample and $\mathrm{sign}\,(\nabla_{\boldsymbol{x}}(-z_y))$ is the signed gradient of loss term $-z_y$. All these experiments are conducted on test images of CIFAR-10

attack, x should be perturbed towards -2. In this particular scenario, the gradient $\nabla_x z_2 < 0$ alone can provide the most effective attack direction. Note that this toy example was motivated by the loss landscape of DNNs when imbalanced gradient occurs.

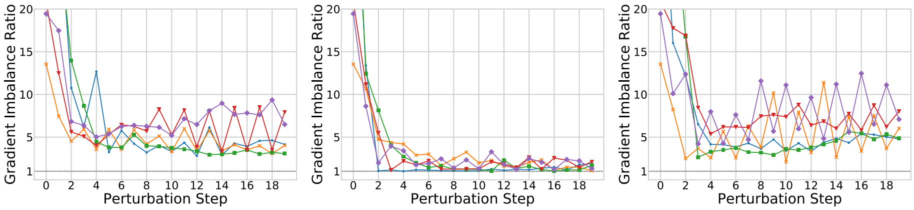
3.1 Imbalanced gradients in DNNs

The situation can be extremely complex for DNNs with high-dimensional inputs, as imbalanced gradients can occur at each input dimension. It thus requires a metric to quantitatively measure the degree of gradient imbalance. Here, we propose such a metric named *gradient imbalance ratio* (GIR) to measure the imbalance ratio for a single input $\boldsymbol{x}$, which can then be averaged over multiple inputs to produce the imbalance ratio for the entire model.

Definition of GIR To measure the imbalance ratio, we focus on the input dimensions that are dominated by one loss term. An input dimension x_i is dominated by a loss term (e.g., z_{max}) means that 1) the gradients of loss terms at x_i have different directions $(\nabla_{x_i} z_{max} \cdot \nabla_{x_i}(-z_y) < 0)$, and 2) the gradient of the dominant term is larger (e.g., $|\nabla_{x_i} z_{max}| > |\nabla_{x_i}(-z_y)|$). According to the dominant term, we can split these dimensions into two subsets $\boldsymbol{x}_{s_1}$ and $\boldsymbol{x}_{s_2}$ where $\boldsymbol{x}_{s_1}$ are dominated by the z_{max} term, while $\boldsymbol{x}_{s_2}$ are dominated by the $-z_y$ term. The overall dominance effect of each loss term can be formulated as $r_1 = \left\|\nabla_{\boldsymbol{x}_{s_1}}(z_{max} - z_y)\right\|_1$ and $r_2 = \left\|\nabla_{\boldsymbol{x}_{s_2}}(z_{max} - z_y)\right\|_1$. Here, we use the L_1-norms instead of L_0 -norms (i.e., the number of dominated dimensions) to also take into consideration the gradient magnitude. To keep the ratio larger than 1, GIR is computed as:

$$GIR = \max\{\frac{r_1}{r_2}, \frac{r_2}{r_1}\}. \tag{1}$$

Note that the GIR metric is not a general measure of imbalance. Rather, it is designed only for assessing gradient imbalance for *adversarial robustness evaluation*. GIR focuses specifically on the imbalanced input dimensions and uses the L_1 norm to take into account the influence of these dimensions to model output. The ratio reflects how far away the imbalance towards one direction than the other.



(a) PGD-20 with loss $z_{max} - z_y$ (b) PGD-20 with loss z_{max} (c) PGD-20 with loss $-z_y$

Fig. 3 Changes in gradient imbalance ratio when apply PGD-20 ($\epsilon = 8/255$) attack with **a** the margin loss, **b)** only the z_{max} term, or **c** only the $-z_y$ term, on the AdvInterp model for 5 CIFAR-10 test images. The imbalance ratio is effectively reduced by attacking a single z_{max} term

GIR of both naturally- and adversarial-trained DNNs With the GIR metric, we next investigate 6 DNN models including a naturally-trained (Natural) model and 5 adversarially-trained models using standard adversarial training (Madry et al., 2018) (SAT), sensible adversarial training (Kim & Wang, 2020) (Sense), feature scattering-based adversarial training (Zhang & Wang, 2019) (FeaScatter), bilateral adversarial training (Wang & Zhang, 2019) (Bilateral), and adversarial interpolation training (Zhang & Xu, 2020) (AdvInterp). We present these defense models here because they (except SAT) represent different levels of gradient imbalance (more results are provided in Fig. 5). Natural, SAT and Sense are WideResNet-34–10 models, while others are WideResNet-28–10 models. We train Natural and SAT following typical settings in Madry et al. (2018), while others use their officially released models. We compute the GIR scores of the 6 models based on 1000 randomly selected test samples and show them in Fig. 2a. One quick observation is that some defense models have a much higher imbalance ratio than either naturally-trained or SAT models. This confirms that gradient imbalance does exist in DNNs, and some defenses tend to train the model to have highly imbalanced gradients. We will show in Sect. 5 that this situation of imbalanced gradients may cause overestimated robustness when evaluated by the commonly used PGD attack.

3.2 Imbalanced gradients reduce attack effectiveness

When there are imbalanced gradients, the attack can be pushed by the dominant term to produce weak attacks, and the non-dominant term alone can lead to more successful attacks. To illustrate this, in Fig. 2b, we show the success rates of PGD attack on the above 5 defense models (Natural has zero robustness against PGD) with different losses: CE loss, margin loss, and the two individual margin terms. We consider 20-step PGD (PGD-20) attacks with step size $\epsilon/4$ and $\epsilon = 8/255$ on all CIFAR-10 test images. One may expect the two margin terms to produce less effective attacks, as they only provide partial information about the margin loss. This is indeed the case for the low gradient imbalance model SAT. However, for highly imbalanced models Sense, FeaScatter, Bilateral and AdvInterp, attacking the z_{max} term produces even more powerful attacks than attacking the margin loss. This implies that the gradient of the margin loss is shifted by the dominant term (e.g., $-z_y$ in this case) towards a less optimal direction, which inevitably causes less powerful attacks. Compared between attacking CE loss and attacking $-z_y$, they achieve a very close performance on imbalanced models. This shows a stronger dominant effect of $-z_y$ in CE loss ($\ell_{ce}(\boldsymbol{x}, y) = \log(\sum_{i=1}^{C} e^{z_i}) - z_y$). It is worth mentioning that, while both GIR and this individual term-based test can be used to check whether there are significantly imbalanced gradients in a defense model, GIR alone cannot predict the amount of overestimated robustness. Figure 2c shows an example of how the $-z_y$ term leads the attack to a suboptimal direction: the margin loss is flat at the $\nabla_{\boldsymbol{x}}(-z_y)$ direction, yet increases drastically at an opposite direction. In this example, the attack can actually succeed if it increases (rather than decreases) z_y.

Gradients can be balanced by attacking individual loss terms Here, we show that, interestingly, imbalanced gradients can be balanced by attacking the non-dominant term. Consider the AdvInterp model tested above as an example, the dominant term is $-z_y$. Figure 3 illustrates the GIR values of 5 randomly selected CIFAR-10 test images by attacking them using PGD-20 with different margin terms or the full margin loss. As can be observed, for all three losses, the GIRs are effectively reduced after the first few steps. However, only the non-dominant term z_{max} manages to steadily reduce the imbalance ratio towards 1. This indicates that optimizing the individual terms separately can help avoid the situation of imbalanced gradients and the attack can indeed benefit from more balanced gradients (see the higher success rate of z_{max} in Fig. 2b).

4 Proposed margin decomposition attacks

Margin decomposition attack The above observations motivate us to exploit the individual terms in the margin loss so that the imbalanced gradients situation can be circumvented. Specifically, we propose margin decomposition (MD) attack that decomposes the attacking process with a margin loss into two stages: (1) attacking the two individual terms (e.g., z_{max} or $-z_y$) alternately with restarts; then (2) attacking the full margin loss. Formally, our MD attack and its loss functions used in each stage are defined as follows:

$$\boldsymbol{x}_{k+1} = \Pi_{\epsilon}\big(\boldsymbol{x}_k + \alpha \cdot \text{sign}\,(\nabla_{\boldsymbol{x}} \ell_k^r(\boldsymbol{x}_k, y))\big),$$

$$\ell_k^r(\boldsymbol{x}_k, y) = \begin{cases} z_{max} & \text{if } k < K' \text{ and } r \bmod 2 = 0 \\ -z_y & \text{if } k < K' \text{ and } r \bmod 2 = 1 \\ z_{max} - z_y & \text{if } K' \leq k \leq K, \end{cases} \tag{2}$$

where, Π is the projection operation that projects the perturbation back onto the ϵ-ball around $\boldsymbol{x}$ if it goes beyond, $k \in \{1, \dots, K\}$ is the perturbation step, $K' \in [1, K]$ is the allocated step for the first stage (i.e., $\forall k \in [1, K')$), $r \in \{1, \dots, n\}$ is the r-th restart, mod is the modulo operation for alternating optimization, and ℓ_k^r defines the loss function used at the k-th step and r-th restart. The loss function switches from the individual terms back to the full margin loss at step K'. The first stage exploits the two margin terms to rebalance the gradients, while the second stage ensures that the final objective (i.e., maximizing the classification error) is achieved. The complete algorithm of MD can be found in Algorithm 1.

Note that, not all defense models have the imbalanced gradients problem. A model is susceptible to imbalanced gradients if there is a substantial difference between robustness evaluated by PGD attack and that by our MD attack. In addition, to help escape the flat loss landscape observed in Fig. 2c, we randomly initialize the perturbation at different restarts (line 6 in Algorithm 1), and use large initial perturbation size $\alpha = 2\epsilon$ with cosine annealing for both stages (lines 8–12 in Algorithm 1).

Algorithm 1 Margin Decomposition Attack

```
1:  Input: clean sample $\boldsymbol{x}$, label $y$, model $f$, stage 1 steps $K'$, total steps $K$
2:  Output: adversarial example $\boldsymbol{x}_{adv}$
3:  Parameters: Maximum perturbation $\epsilon$, step size $\alpha$, number of restarts $n$, first
    stage steps $K'$, total steps $K$
4:  $\boldsymbol{x}_{adv} \leftarrow \boldsymbol{x}$
5:  for $r \in \{1, ..., n\}$ do
6:      $\boldsymbol{x}_0 \leftarrow \boldsymbol{x} + uniform(-\epsilon, \epsilon)$          ▷ uniform noise initialization
7:      for $k \in \{1, ..., K\}$ do
8:          if $k < K'$ then
9:              $\alpha \leftarrow \epsilon \cdot \big(1 + \cos(\frac{k-1}{K'}\pi)\big)$
10:         else if $k \geq K'$ then
11:             $\alpha \leftarrow \epsilon \cdot \big(1 + \cos(\frac{k-K'}{K-K'}\pi)\big)$
12:         end if
13:         $\boldsymbol{x}_k \leftarrow \Pi_\epsilon\big(\boldsymbol{x}_{k-1} + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} \ell_k^r(\boldsymbol{x}_{k-1}, y))\big)$    ▷ update $\boldsymbol{x}_k$ by Eqn. (2)
14:         if $\ell(\boldsymbol{x}_{adv}) < \ell(\boldsymbol{x}_k)$ then
15:             $\boldsymbol{x}_{adv} \leftarrow \boldsymbol{x}_k$
16:         end if
17:     end for
18: end for
19: $\boldsymbol{x}_{adv} = \Pi_{[0,1]}\big(\boldsymbol{x}_{adv}\big)$          ▷ final clipping
20: return $\boldsymbol{x}_{adv}$
```

Multi-targeted MD attack We also propose a multi-targeted version of our MD attack and call it MD-MT. The loss terms used by MD-MT at different attacking stages are defined as follows:

$$\boldsymbol{x}_{k+1} = \Pi_\epsilon\big(\boldsymbol{x}_k + \alpha \cdot \ \mathrm{sign}\,(\nabla_{\boldsymbol{x}} \ell_k^r(\boldsymbol{x}_k, t))\big),$$
$$\ell_k^r(\boldsymbol{x}_k, y) = \begin{cases} z_t & \text{if } k < K' \text{ and } r \bmod 2 = 0 \\ -z_y & \text{if } k < K' \text{ and } r \bmod 2 = 1 \\ z_t - z_y & \text{if } K' \leq k \leq K, \end{cases} \tag{3}$$

where, z_t is the logits of a target class $t \neq y$. Other parameters are the same as in Eq. (2). Like the MT attack, MD-MT will attack each possible target class one at a time, then select the strongest adversarial example at the end. That is, the target class $t \neq y$ will be switched to a different target class at each restart. The complete algorithm MD-MT can be found in Algorithm 2.

Algorithm 2 MultiTargeted Margin Decomposition Attack

```
Input: clean sample $\boldsymbol{x}$, class label $y$, class set $\mathcal{T}$, model $f$
Output: adversarial example $\boldsymbol{x}_{adv}$
Parameters: Maximum perturbation $\epsilon$, step size $\alpha$, number of restarts $n$,
    first stage steps $K'$, total steps $K$
$n_{\mathrm{r}} \leftarrow \lfloor n/|\mathcal{T}| \rfloor$, $\boldsymbol{x}_{adv} \leftarrow \boldsymbol{x}$
for $r \in \{1, ..., n_{\mathrm{r}}\}$ do
    for $t \in \mathcal{T}$ do
        $\boldsymbol{x}_0 \leftarrow \boldsymbol{x} + uniform(-\epsilon, \epsilon)$    ▷ uniform noise initialization
        for $k \in \{1, ..., K\}$ do
            if $k < K'$ then
                $\alpha \leftarrow \epsilon \cdot \big(1 + \cos(\frac{k-1}{K'}\pi)\big)$
            else if $k \geq K'$ then
                $\alpha \leftarrow \epsilon \cdot \big(1 + \cos(\frac{k-K'}{K-K'}\pi)\big)$
            end if
            Update $\boldsymbol{x}_k$ by Eqn. (3)
            if $\ell(\boldsymbol{x}_{adv}) < \ell(\boldsymbol{x}_k)$ then
                $\boldsymbol{x}_{adv} \leftarrow \boldsymbol{x}_k$
            end if
        end for
    end for
end for
$\boldsymbol{x}_{adv} = \Pi_{[0,1]}\big(\boldsymbol{x}_{adv}\big)$    ▷ final clipping
return $\boldsymbol{x}_{adv}$
```

MD ensemble attack Following AutoAttack (Croce & Hein, 2020b), here we also propose an ensemble attack to fully exploit the strengths of both existing attacks and the gradient exploration of our MD attacks. The AutoAttak ensemble consists of 4 attacks: (1) APGD$_{CE}$, which is the Auto-PGD with the cross entropy loss; (2) DLR, which is the Auto-PGD with the Difference of Logits Ratio (DLR) loss; (3) Fast Adaptive Boundary Attack (FAB) (Croce & Hein, 2019); and (4) the black-box Square Attack (Andriushchenko et al., 2019). The MultiTargeted version of both APGD$_{CE}$ and DLR are used in the latest version of AutoAttack. We first replace the MultiTargeted DLR (DLR-MT), the strongest attack in the AutoAttack ensemble by our MD-MT attack. We then replace the Square attack by our MD attack as we focus on white-box robustness and gradient issues. This gives us the MD Ensemble (MDE) of 4 attacks including (1) MD, (2) MD-MT, (3) APGD$_{CE}$ and (4) FAB. We did not replace the APGD$_{CE}$ attack as we find it is better to have a cross entropy loss based attack in the ensemble.

Initialization perspective interpretation of MD attacks Previous works have shown that random or logits diversified initialization are crucial for generating strong adversarial attacks (Madry et al., 2018; Tashiro et al., 2020). Compared to random or logits diversified initialization, our MD attacks can be interpreted as a type of *adversarial initialization*, i.e., initialization at the adversarial sub-directions defined by the two margin terms. Rather than a single step of initialization, our MD attacks iteratively explore the optimal starting point during the first attacking stage.

Extending MD to complex losses with more than two terms. Our MD strategy is not restricted to the margin loss, it is also suitable for other margin-based losses like CW and DLR attacks. For more complex loss functions with more than two terms, one can group the individual terms into two conflicting groups that could produce opposite gradient directions or investigate the two most conflicting terms (in the form of 'A - B'). In this way, a complex loss could be reformulated in a margin form where our margin decomposition strategy can be easily applied.

5 Experiments

Defense models We apply our MD attacks to evaluate the robustness of 24 defense models proposed since 2018. Here, we focus on adversarial training models, which are arguably the strongest defense approach to date (Athalye et al., 2018; Croce & Hein, 2020b). The selected defense models are as follows.

The standard adversarial training (SAT) (Madry et al., 2018) trains models on adversarial examples generated by PGD attack. Dynamic adversarial training (dynamic) (Wang et al., 2019) trains on adversarial examples with gradually increased convergence quality. Max-Margin Adversarial training (MMA) (Ding et al., 2018) trains on adversarial examples with increasing margins (e.g., the perturbation bound ϵ). For MMA, we evaluate the released "MMA-32" model. Jacobian adversarially regularized networks (JARN) adversarially regularize the Jacobian matrices and can be combined with 1-step adversarial training (JARN-AT1) to gain additional robustness (Chan et al., 2020). For JARN, we only evaluate the JARN-AT1 as its none adversarial training version has been completely broken in (Croce & Hein, 2020b). We implement JARN-AT1 on the basis of their released implementation of JARN. Sensible adversarial training (Sense) (Kim & Wang, 2020) trains on loss-sensible adversarial examples (perturbation stops when the loss exceeds a certain threshold). Bilateral adversarial training (Bilateral) (Wang & Zhang, 2019) trains on PGD adversarial examples with adversarially perturbed labels. For Bilateral, we mainly evaluate its released strongest model "R-MOSA-LA-8". Adversarial Interpolation (Adv-Interp) training (Zhang & Xu, 2020) trains on adversarial examples generated under an adversarial interpolation scheme with adversarial labels. Feature Scattering-based (FeaScatter) adversarial training (Zhang & Wang, 2019) crafts adversarial examples using latent space feature scattering, then trains on these examples with label smoothing. Adversarial Training with Hypersphere Embedding (AT-HE) (Pang et al., 2020) regularizing the features onto compact manifolds. Adversarial Training with Pre-Training (AT-PT) (Hendrycks et al., 2019) uses pre-training to improve model robustness. TRADES (Zhang et al., 2019) replaces the CE loss of SAT with the KL divergence for a better trade-off between robustness and natural accuracy. Based on TRADES, RST (Carmon et al., 2019) and UAT (Alayrac et al., 2019) improve robustness by training with 10× more unlabeled data. Misclassification Aware adveRsarial Training (MART) (Wang et al., 2020) further improves the above three methods with a misclassification aware loss function. Adversarial Training with Early Stopping (AT-ES) (Rice et al., 2020) improves SAT by using early stopping to avoid robust overfitting. Adversarial weight perturbation (AWP) (Wu et al., 2020) proposes a double perturbation mechanism to explicitly regularize the flatness of the weight loss landscape. Robust WideResNet (R-WRN) (Huang et al., 2021) explores the most robust configurations of WideResNet (Zagoruyko & Komodakis, 2016) and trains the model following the same procedure as RST. This robust configuration can bring additional stability to the model and improve robustness. The 4 defenses we consider for the ImageNet dataset are as follows. (1) Engstrom et al. (2019) showed that robust representations obtained via adversarial training on ImageNet are approximately invertible. (2) Fast adversarial training (FastAT) (Wong et al., 2020) has also been found to be crucial for efficient adversarial training on ImageNet. (3) Salman et al. (2020) trained adversarially robust models on ImageNet to demonstrate their transferability to other tasks. (4) A recent work by Debenedetti et al. (2022) explored the adversarial training hyperparameters for vision transformers and trained robust XCiT models on ImageNet.

The architectures of the defense from CIFAR-10 models are all WideResNet variants (Zagoruyko & Komodakis, 2016). We also evaluated other architectures, including VGG-19 (Simonyan & Zisserman, 2014), DenseNet-121 (Huang et al., 2017), and DARST (Liu et al., 2019) trained with SAT (Madry et al., 2018). The configuration for these models follows the same setting as in Huang et al. (2021). For each defense model, we either download their shared models or retrain the models using the official implementations, unless explicitly stated. For ImageNet, we consider the adversarially pre-trained ResNet-50 and vision transformer XCiT-S models. The defense models were trained against maximum perturbation $\epsilon = 8/255$ on CIFAR-10 and $\epsilon = 4/255$ on ImageNet. We apply the current state-of-the-art and our MD attacks to evaluate the robustness of these models in a white-box setting.

Baseline attacks and settings Following the current literature, we consider 4 existing untargeted attacks: (1) the L_∞ version of CW attack (Madry et al., 2018; Wang et al., 2019), (2) projected gradient descent (PGD) (Madry et al., 2018), (3) output diversified initialization (ODI) (Tashiro et al., 2020), (4) fast adaptive boundary attack (FAB) (Croce & Hein, 2020a), (5) auto-PGD with difference of logits ratio (DLR) (Croce & Hein, 2020b). We consider 3 multi-target attacks, (1) multi-targeted (MT) attack (Gowal et al., 2019), (2) multi-targeted FAB (FAB-MT), (3) multi-targeted DLR (DLR-MT). The evaluation was conducted under the same maximum perturbation $\epsilon = 8/255$ and $\epsilon = 4/255$ for CIFAR-10 and ImageNet respectively. For all untargeted attacks, we use the same total perturbation steps $K = 100$. For DLR and FAB, we use the official implementation and parameter setting. For PGD and ODI, we use 5 restarts and step size set to $\alpha = 0.8/255$. For both stages of our MD attack, we use a large initial step size $\alpha = 2\epsilon$, then gradually decrease it to 0 via cosine annealing. The number of steps is set to $K' = 20$ for the first stage (i.e., $K - K' = 80$ for the second stage). For all multi-target attacks, we use the same total perturbation steps $K = 100$ for each class. For CIFAR-10, this means a total of 900 steps for each image (as there are only 9 possible target classes). For DLR-MT and FAB-MT, we use the official implementation and parameter setting. For MT attack, we use 5 restarts and step size set to $\alpha = 0.8/255$. For our MD-MT attack, we use the same parameter setting as the untargeted MD when attacking each target class. For fair comparison, the total number of perturbation steps are set to be the same for all attacks. We also compare our MD Ensemble (MDE) with the AutoAttack ensemble. Adversarial robustness is measured by the model's accuracy on adversarial examples crafted by the attack on CIFAR-10 and ImageNet test images. We excluded FAB from the untargeted evaluation on ImageNet and all ImageNet models in targeted and ensemble evaluations due to the attack's efficiency.

5.1 Robustness evaluation results

Table 1 reports the untargeted evaluation result, where R-WRN, AWP, and RST are the top 3 best defenses. The 'Diff' column shows that there are 11 defense models demonstrating more than 1% robustness drop against our MD attack compared to the best baseline attack. This implies that these models are susceptible to imbalanced gradients. Out of the 11 models, 4 of them are ranked at the very bottom (worst robustness on CIFAR-10) of the list with a robustness that is much lower than SAT. In most cases, our MD attack is able to decrease the PGD or DLR robustness by more than 2% even on the top 5 defense models.

Table 1 Untargeted evaluation

Dataset	Defense (*ranked by MD robustness*)	Clean	CW	PGD	ODI	FAB	DLR	MD	Diff
CIFAR10	R-WRN (Huang et al., 2021)	91.23	64.28	64.98	64.76	64.49	64.38	**62.91**	**– 1.37**
	AWP (Wu et al., 2020)	88.25	63.64	63.45	62.14	60.74	60.73	**60.19**	**– 0.54**
	RST (Carmon et al., 2019)	89.69	62.25	61.99	61.89	60.92	60.88	**59.77**	**– 1.11**
	UAT (Alayrac et al., 2019)	86.46	61.19	60.96	60.17	60.06	62.97	**58.44**	**– 1.62**
	AT-PT (Hendrycks et al., 2019)	87.11	57.57	57.53	56.84	55.57	57.25	**55.22**	**– 0.35**
	AT-ES (Rice et al., 2020)	85.52	55.80	55.80	54.50	53.14	54.67	**52.94**	**– 0.20**
	TRADES (Zhang et al., 2019)	84.92	55.07	54.98	54.83	53.50	53.66	**52.74**	**– 0.76**
	AT-HE (Pang et al., 2020)	81.43	53.05	52.87	52.89	52.20	54.62	**51.85**	**– 0.35**
	MART (Wang et al., 2020)	83.62	56.09	55.84	53.43	51.80	52.90	**51.50**	**– 0.30**
	SAT-DenseNet121 (Huang et al., 2021)	86.07	52.75	52.70	53.01	51.39	53.06	**50.83**	**– 0.56**
	SAT-DARTS (Huang et al., 2021)	86.76	49.16	46.35	48.64	46.65	47.38	**45.78**	**– 0.57**
	Adv-Interp (Zhang & Xu, 2020)	90.25	73.12	73.05	50.85	**43.96**	52.87	45.07	1.11
	MMA (Ding et al., 2018)	84.36	51.34	51.40	46.90	48.97	51.20	**44.94**	**– 1.96**
	SAT (Madry et al., 2018)	87.25	51.49	45.32	47.23	45.94	46.85	**44.70**	**– 0.62**
	Dynamic (Wang et al., 2019)	84.36	51.34	51.40	44.30	46.90	51.20	**44.94**	**– 0.48**
	SAT-VGG19 (Huang et al., 2021)	77.06	45.51	46.06	44.86	43.88	46.08	**43.56**	**– 0.32**
	FeaScatter (Zhang & Wang, 2019)	89.98	69.56	69.35	44.89	43.78	51.82	**42.16**	**– 1.62**
	Sense (Kim & Wang, 2020)	91.51	60.68	60.61	46.32	42.24	50.54	**39.91**	**– 2.33**
	Bilateral (Wang & Zhang, 2019)	90.73	61.50	61.20	44.73	43.79	46.36	**39.39**	**– 4.11**
	JARN-AT1 (Chan et al., 2020)	83.86	50.15	19.36	19.73	20.14	20.57	**17.32**	**– 2.04**
ImageNet	XCiT-S (Debenedetti et al., 2022)	72.53	55.55	42.84	41.63	-	44.48	**40.26**	**– 1.37**
	ResNet-50 (Salman et al., 2020)	63.87	51.14	38.60	36.41	-	37.92	**35.31**	**– 1.09**
	ResNet-50 (Engstrom et al., 2019)	62.40	49.16	32.46	31.23	-	33.00	**29.76**	**– 1.47**
	FastAT (Wong et al., 2020)	53.82	44.55	27.28	26.31	-	27.48	**25.34**	**– 0.98**

Robustness (%) of 24 defense models on CIFAR-10 and 4 defense models on ImageNet, all evaluated by untargeted attacks. The 'diff' column shows the robustness decrease by our MD attack compared to the *best* baseline attack (i.e., best baseline - MD)

The best results are boldfaced

Compared to the best baseline attack, our MD attack can reduce its evaluated robustness by more than 2% on Sense, Bilateral, FeaScatter, and JARN-AT1. It also shows more than 1% improvement on top-ranking defenses, including R-WRN, RST, and UAT. These results demonstrate the importance of addressing the imbalanced gradients problem in robustness evaluation. Circumventing imbalanced gradients via margin decomposition and exploration makes our MD the best overall standalone attack for robustness evaluation. In terms of evaluating ImageNet defenses, our MD attack alone can outperform the AA ensemble attack on XCiT-S and FastAT, compared to the robustness evaluated and reported on RobustBench leaderboard (Croce et al., 2021). For instance, AA evaluates XCiT-S to be of 41.78% robustness, while our MD evaluates it to be of 40.26% robustness. This highlights the significance of the imbalanced gradients problem on the large-scale dataset and the advantage of our proposed technique in evaluating vision transformers.

The multi-targeted evaluation results are presented in Table 2, where it shows that our MD-MT attack is the strongest attack on average. By comparing Table 1 with Table 2, we find that the multi-targeted robustness is lower than the untargeted robustness for all defenses. This indicates that multi-targeted evaluation is more accurate than untargeted evaluation. Overall, our MD-MT attack demonstrates the most reliable robustness on 13/20 defense models, while DLR-MT is slightly better on the other 7 models. The improvement of DLR-MT over DLR indicates that exploring different targets can help circumvent imbalanced gradients to some extent. We will conduct a detailed analysis of different attack techniques against imbalanced gradients in Sect. 5.3.

We further compare our MDE attack with the AutoAttack in Table 3. Compared to multi-targeted evaluation, the use of ensemble attacks produces the most accurate evaluation for all defense models. This is because ensemble attacks combine the strengths of multiple attacks. It is worth mentioning that the current adversarial robustness leaderboard is created based on the AutoAttack ensemble (Croce et al., 2020). By replacing two of its attacks, our MDE is able to produce even better evaluation with lower robustness in most cases, except a tie on SAT and R-WRN, and 0.05% worse on UAT. The improvements are more pronounced on the imbalanced gradients models (e.g., the bottom 6 models). This result again verifies the importance of margin exploration in robustness evaluation.

Table 2 Multi-targeted evaluation.

Defense (*ranked by MD-MT robustness*)	Clean	MT	FAB-MT	DLR-MT	MD-MT	Diff
R-WRN (Huang et al., 2021)	91.23	62.64	63.18	**62.55**	62.57	0.02
AWP (Wu et al., 2020)	88.25	60.12	60.52	**60.05**	60.07	0.02
RST (Carmon et al., 2019)	89.69	59.74	60.20	59.58	**59.55**	**– 0.03**
UAT (Alayrac et al., 2019)	86.46	56.50	59.98	**56.16**	56.19	0.03
AT-PT (Hendrycks et al., 2019)	87.11	55.04	55.28	54.88	**54.87**	**– 0.01**
TRADES (Zhang et al., 2019)	84.92	52.61	53.06	52.53	**52.51**	**– 0.02**
AT-ES (Rice et al., 2020)	85.52	52.42	52.75	52.35	**52.31**	**– 0.04**
AT-HE (Pang et al., 2020)	81.43	51.17	51.53	51.10	**51.04**	**– 0.06**
MART (Wang et al., 2020)	83.62	51.00	51.39	50.95	**50.91**	**– 0.04**
SAT-DenseNet121 (Huang et al., 2021)	86.07	50.24	50.64	**50.12**	50.14	0.02
SAT-DARTS (Huang et al., 2021)	86.76	45.39	45.97	**45.09**	45.11	0.02
SAT (Madry et al., 2018)	87.25	44.67	45.29	44.52	**44.49**	**– 0.03**
Dynamic (Wang et al., 2019)	84.48	43.01	43.40	42.93	**42.91**	**– 0.02**
SAT-VGG19 (Huang et al., 2021)	77.06	42.67	43.30	42.62	**42.60**	**– 0.02**
MMA (Ding et al., 2018)	84.36	42.74	42.66	41.72	**41.45**	**– 0.27**
ADVInterp (Zhang & Xu, 2020)	90.25	66.30	39.10	**37.53**	37.54	0.01
Bilateral (Wang & Zhang, 2019)	90.73	57.51	38.36	38.55	**37.03**	**– 1.33**
FeaScatter (Zhang & Wang, 2019)	89.98	43.37	38.54	37.29	**36.72**	**– 0.57**
Sense (Kim & Wang, 2020)	91.51	48.40	35.50	35.94	**34.87**	**– 0.63**
JARN-AT1 (Chan et al., 2020)	83.86	17.10	17.49	**16.58**	16.63	0.05

Robustness (%) of 20 defense models on CIFAR-10 evaluated by different multi-targeted attacks. The 'diff' column shows the robustness decrease by our MD-MT attack compared to the *best* baseline attack (i.e., best baseline - MD-MT)

The best results are boldfaced

Table 3 Ensemble evaluation

Defense (ranked)	Clean	AutoAttack	MDE	Diff
R-WRN (Huang et al., 2021)	91.23	62.54	62.54	0.00
AWP (Wu et al., 2020)	88.25	60.05	**60.00**	**– 0.05**
RST (Carmon et al., 2019)	89.69	59.56	**59.53**	**– 0.03**
UAT (Alayrac et al., 2019)	86.46	**56.11**	56.16	0.05
AT-PT (Hendrycks et al., 2019)	87.11	54.91	**54.86**	**– 0.05**
TRADES (Zhang et al., 2019)	84.92	52.54	**52.51**	**– 0.03**
AT-ES (Rice et al., 2020)	85.52	52.34	**52.30**	**– 0.04**
AT-HE (Pang et al., 2020)	81.43	51.09	**51.06**	**– 0.03**
SAT-DenseNet121 (Huang et al., 2021)	86.07	50.11	50.09	**– 0.02**
SAT-DARTS (Huang et al., 2021)	86.76	**45.00**	45.01	0.01
MART (Wang et al., 2020)	83.62	50.94	**50.89**	**– 0.05**
SAT (Madry et al., 2018)	87.25	44.45	44.45	0.00
Dynamic (Wang et al., 2019)	84.48	42.90	**42.89**	**– 0.01**
SAT-VGG19 (Huang et al., 2021)	77.06	42.61	**42.57**	**– 0.04**
MMA (Ding et al., 2018)	84.36	41.51	**41.34**	**– 0.17**
ADVInterp (Zhang & Xu, 2020)	90.25	36.46	**36.55**	**– 0.09**
Bilateral (Wang & Zhang, 2019)	90.73	36.61	**36.48**	**– 0.13**
FeaScatter (Zhang & Wang, 2019)	89.98	36.65	**36.25**	**– 0.40**
Sense (Kim & Wang, 2020)	91.51	34.19	**33.84**	**– 0.35**
JARN-AT1 (Chan et al., 2020)	83.80	16.55	**16.51**	**– 0.04**

Robustness (%) of 20 defense models evaluated by ensemble attacks. The ‘diff’ column shows the robustness decrease by our MDE attack compared to the *best* baseline attack (i.e., best baseline - MDE)

The best results are boldfaced

5.2 Defenses that cause imbalanced gradients

Here, we explore common defense techniques that cause imbalanced gradients by focusing on the 6 defense models that are relatively more susceptible to imbalanced gradients: MMA, Bilateral, Adv-Interp, FeaScatter, Sense, and JARN-AT1. To avoid other factors introduced by the different attack techniques used by ODI and DLR attacks, here we directly compare the robustness evaluated by the classic PGD attack and that evaluated by our MD attack. Note that imbalanced gradients are not the only possible but rather a subtle cause of over-estimated robustness. That is, it is more subtle than obfuscated gradients and needs specific techniques to evade.

Label smoothing causes imbalanced gradients According to Table 1, the PGD robustness of Adv-Interp, FeaScatter, and Bilateral decreases the most (i.e., 20–27%) against our MD attack. This indicates that these defenses have caused the imbalanced gradients problem, as also confirmed by their high GIR values in Fig. 2a. All three defenses use label smoothing as part of their training scheme to improve adversarial training, which we suspect is one major cause of imbalanced gradients. Given a sample $\boldsymbol{x}$ with label y, label smoothing encourages the model to learn a uniform logit or probability distribution over classes $j \neq y$. This tends to smooth out the input gradients of $\boldsymbol{x}$ with respect to these classes, resulting in smaller gradients. In order to confirm label smoothing indeed causes imbalanced gradients, we train a WideResNet-34–10 model using natural training (‘Natural’)

Table 4 Robustness (%) of WideResNet-34–10 (may be different to those evaluated in Table 1) models trained with/without label smoothing

Defense	FGSM	CW	PGD	ODI	DLR	MD	z_{max}	$-z_y$	$z_{max} - z_y$
Natural	31.96	37.73	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.02	**0.00**
+ Label Smoothing	58.28	49.27	7.87	0.28	0.14	**0.06**	0.11	0.29	0.13
SAT	65.92	50.56	47.25	48.86	48.87	**46.31**	61.10	55.07	48.47
+ Label Smoothing	66.57	51.15	48.67	48.25	50.35	**46.60**	52.19	48.82	48.37

The lowest robustness results are boldfaced. z_{max}, $-z_y$, and $z_{max} - z_y$ denote MD attack with the single loss terms, respectively

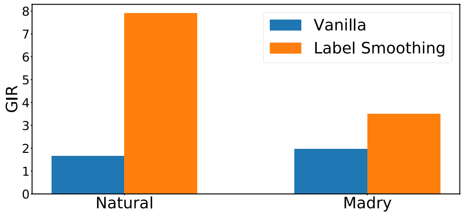


Fig. 4 Gradient imbalance ratio (GIR) of models trained with/without label smoothing

and SAT with or without label smoothing (smoothing parameter 0.5). We report their robustness in Table 4, and show their gradient imbalance ratios (GIRs) in Fig. 4. According to the GIRs, adding label smoothing into the training process immediately increases the imbalance ratio, especially in natural training. The PGD robustness of the naturally-trained model also “increases" to 7.87%, which is almost zero (0.06%) according to our MD attack. Using smoothed labels in SAT defense also “increases" PGD robustness by more than 1%, which in fact is only 0.3% according to our MD. Other attacks including CW and DLR are also affected by the label-smoothing effect. This is because CW and DLR are both logit-based attacks, which are more prone to imbalanced gradients. ODI is less sensitive to label smoothing, yet is still not as effective as our MD attack. These evidences confirm that label smoothing indeed causes imbalanced gradients, leading to overestimated robustness if evaluated by regular attacks like PGD or DLR. Figure 4 demonstrates that adversarial training can inhibit imbalanced gradients caused by label smoothing to large extent. This is because the adversarial examples used for adversarial training are specifically perturbed to the $j \neq y$ classes, thus helping avoid uniform logits over classes $j \neq y$ to some extent.

The last three columns in Table 4 show the robustness results when different loss terms are used in MD, in the presence of label smoothing and thus imbalanced gradients. They are consistent with our illustration in Fig. 2b, i.e., $-z_y$ works similarly as $z_{max} - -z_y$, while z_{max} is the worst. This means that the non-dominant term is more effective when there are imbalanced gradients.

Other defense techniques that cause imbalanced gradients The other 3 imbalanced defenses MMA, Sense and JARN-AT1 adopt different defense techniques to “improve" robustness. MMA is a margin-based defense that maximizes the shortest successful perturbation for each data point. MMA only perturbs correctly-classified examples, and the perturbation stops immediately at misclassification (into a $j \neq y$ class). In other words, MMA focuses on examples that are around the decision boundary (i.e., $z_{max} = z_y$) between class y and all other classes $j \neq y$. During training, the decision boundary margin is maximized by pulling the boundary away from these examples. This process maximizes the distance to the closest decision boundary (e.g., towards the weakest class) and results in equal distances to all other classes. This tends to generate a uniform prediction over classes $j \neq y$, a similar effect of label smoothing, and causes imbalanced gradients.

Similar to MMA, Sense perturbs training examples until a certain loss threshold is satisfied. While in MMA the threshold is misclassification, in Sense, it is the loss value with respect to probability (e.g., $\boldsymbol{p}_y = 0.7$). This type of training procedures with a specific logits or probability distribution regularization has caused the imbalanced gradients problem for both MMA and Sense. Note that, Sense causes more imbalanced gradients than MMA. We conjecture it is because optimizing over a probability threshold is much easier than moving the decision boundary.

JARN-AT1 is also a regularization-based adversarial training method. Different from MMA or Sense, it regularizes the model's Jacobian (e.g., input gradients) to resemble natural training images. Such explicit input gradient regularizations reduce the input gradients to a much smaller magnitude and only keep the salient part of input gradients. The input gradients associated with other $j \neq y$ classes will be minimized to cause an imbalance to that associated with class y.

The above analysis indicates that future defense should avoid using label smoothing, margin smoothing or input gradient regularization techniques, or should be carefully evaluated against our MD attack to check for imbalanced gradients.

Correlation between GIR and robustness According to the GIR scores shown in Figs. 2a and 5, models exhibit high GIR scores (e.g., Adv-Interp, FeaScatter and Bilateral) are generally more prone to imbalanced gradients and potentially more vulnerable to our MD attack. However, GIR is not a measure for robustness nor should be used as an exact metric to determine whether one defense is more robust than the other. For example, MART demonstrates a higher GIR score than Sense or JARN-AT1, however, according to our MD attack, it is 11% and 34% more robust than Sense and JARN-AT1, respectively. This is because the GIR score of a model only measures the gradient situation of the model at its current state, which could decrease during the attack as shown in Figs. 3 and 6. Our MD attack iteratively exploits and circumvents imbalanced gradients during the first attacking stage, thus can produce reliable robustness evaluation at the end.

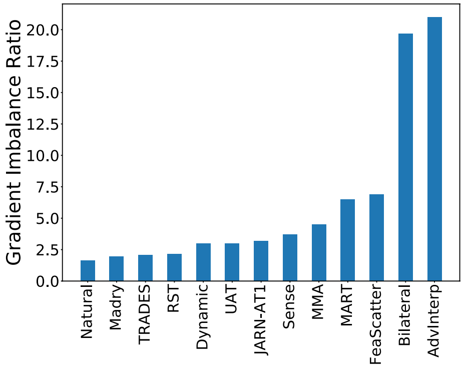


Fig. 5 Gradient imbalance ratios (GIRs) of 12 defense models and a naturally trained model ("Natural"). All models are trained on CIFAR-10 dataset. Those defense models that are not susceptible to imbalanced gradients are omitted here

5.3 An attack view of imbalanced gradients

As shown in Table 1, advanced attacks ODI and DLR are more effective than traditional attacks PGD and CW against imbalanced gradients. Here, we provide more insights into what attack techniques are effective against imbalanced gradients. Before that, we first show imbalanced gradients are more subtle than obfuscated gradients and cannot be easily circumvented by common techniques like random restarts or momentum.

5.3.1 Imbalanced gradients are different from obfuscated gradients

Imbalanced gradients occur when one loss term dominates the attack towards a suboptimal gradient direction, which does not necessarily block gradient descent like obfuscated gradients. Therefore, it does not have the characteristics of obfuscated gradients, and cannot be detected by the five checking rules for obfuscated gradients (Athalye et al., 2018). Here, we test all five rules on the four defense models that exhibit significantly imbalanced gradients: Adv-Interp, FeaScatter, Bilateral, and Sense. Note that all these models were trained and tested on the CIFAR-10 dataset.

One-step attacks outperform iterative attacks. When gradients are obfuscated, iterative attacks are more likely to get stuck in a local minimum. To test this, we compare the success rate of one-step attack FGSM and iterative attack PGD in Table 5. We see that PGD outperforms FGSM consistently on all four defense models with no obvious sign of obfuscated gradients.

Unbounded attacks do not reach 100% success. Increasing the distortion bound does not increase success. Larger perturbation bound gives the attacker more ability to attack. So, if gradients are not obfuscated, an unbounded attack should reach 100% success rate. To test this, we run an “unbounded" PGD attack with $\epsilon = 255/255$. As shown in Table 5, all models are completely broken by this unbounded attack, i.e., the over-estimated robustness is caused by a more subtle effect than obfuscated gradients.

Black-box attacks are better than white-box attacks If a model is obfuscating gradients, it should fail to provide useful gradients in a small neighborhood. Therefore, using a substitute model should be able to evade the defense, as the substitute model was not trained to be robust to small perturbations. To test this, we run a black-box transfer PGD attack on naturally trained substitute models. We find that all four defenses are robust to transferred attacks (“Transfer" in Table 5). We also attack the four defense models using gradient-free attack SPSA (Uesato et al., 2018). For SPSA, we use a batch size of 8192 with 100 iterations and run on 1000 randomly selected CIFAR-10 test images. We confirm that SPSA cannot degrade its performance. None of these results indicate obfuscated gradients.

Table 5 Test of obfuscated gradients for four defense models that have significant imbalanced gradients following (Athalye et al., 2018): attack success rate (%) of different attacks

Defense	FGSM	PGD	Unbounded	Transfer	SPSA	Random
Adv-Interp (Zhang & Xu, 2020)	23.06	27.52	100.00	10.89	24.80	0.00
FeaScatter (Zhang & Wang, 2019)	22.60	31.36	100.00	11.11	28.20	0.00
Bilateral (Wang & Zhang, 2019)	28.90	39.05	100.00	9.23	36.00	0.00
Sense (Kim & Wang, 2020)	27.29	40.14	100.00	9.90	37.90	0.00

The results indicate no sign of obfuscated gradients

Random sampling finds adversarial examples Brute force random search within some ϵ -neighbourhood should not find adversarial examples when gradient-based attacks do not. Following (Athalye et al., 2018), we choose 1000 test images on which PGD fails. We then randomly sample 10^5 points for each image from its $\epsilon = 8/255$-ball region and check if any of them are adversarial. The results (i.e., "Random" in Table 5) show that random sampling cannot find adversarial examples when PGD does not.

All the above test results lead to one conclusion that the robustness of the four defenses is not a result of obfuscated gradients. This indicates that imbalanced gradients does not share the characteristics of obfuscated gradients, and thus cannot be detected following the five test principles for obfuscated gradients. Therefore, imbalanced gradients should be addressed independently for more reliable robustness evaluation.

5.3.2 Momentum, random restart cannot circumvent imbalanced gradients

As we discussed in Sect. 3, random restarts can potentially increase the probability of finding an adversarial example. Momentum method is another way to help escape overfitting to local gradients (Sutskever et al., 2013). Here, we test their effectiveness in circumventing imbalanced gradients. For random restart, we run a 400-step PGD attack with 100 restarts ($\text{PGD}^{100\times400}$). For momentum, we use momentum iterative FGSM (MI-FGSM) (Dong et al., 2018) with 40 steps, 2 restarts, and momentum 1.0. For both attacks, we set $\epsilon = 8/255$ and step size $\alpha = 2/255$. Note that, we removed samples that were successfully perturbed from the batch, thus only restarting samples that are not successful. This slightly improves the performance and shows it is stronger than PGD without restarts. We apply the two attacks on 1000 randomly chosen CIFAR-10 test images, and report the robustness in Table 6 for the four defense models checked in Sect. 5.3.1. Compared to traditional PGD with 40 steps, the robustness can indeed be decreased by $\text{PGD}^{100\times400}$ except on Bilateral, a consistent observation with our analysis in Sect. 3 that more restarts can lower model accuracy. However, the robustness is still highly overestimated compared to that of our MD attack. This indicates that imbalanced gradients can exist in wide-spanned input regions, resulting in a low probability of random restart to find successful attacks. To our surprise, MI-FGSM performs even worse than traditional PGD. On three defense models (i.e., Adv-Interp, FeaScatter, and Sense), it produces even higher robustness than PGD. This implies that accumulating velocity in the gradient direction can make the overfitting even worse when there are imbalanced gradients. This again confirms that the imbalanced gradients problem should be explicitly addressed.

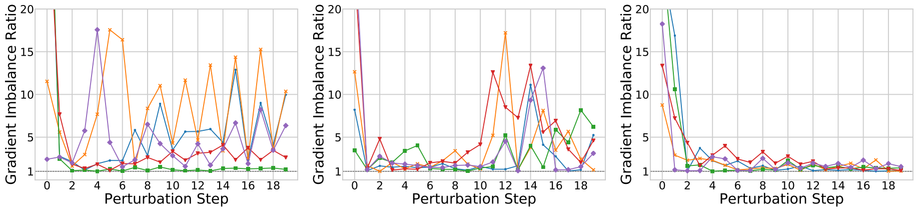
5.3.3 What can help circumvent imbalanced gradients?

Logits diversified initialization ODI randomly initializes the perturbation by adding random weights to logits at its first 2 steps. The random weights change the gradient

Table 6 Robustness (%) of four defense models that have significant imbalance gradients against PGD, $\text{PGD}^{100\times400}$, MI-FGSM and our MD attacks

Defense	PGD	MD (ours)	$\text{PGD}^{100\times400}$	MI-FGSM
Adv-Interp	72.48	**45.07**	71.64	73.25
FeaScatter	68.64	**42.16**	67.00	70.79
Sense	59.86	**39.91**	58.51	62.41
Bilateral	60.95	**39.39**	59.61	51.52

The best results are boldfaced



(a) ODI (b) FAB (c) MD (Ours)

Fig. 6 Gradient imbalance ratio at the first 20 steps of ODI (**a**), FAB (**b**) and our MD (**c**) attacks on the AdvInterp model for 5 randomly selected CIFAR-10 test images

magnitude, thus can also mitigate imbalanced gradients, as shown in Fig. 6a. However, the initialization can only help with the first two attack steps, and the imbalance ratio fluctuates drastically in the following steps. Our MD attack provides a more direct and efficient exploration of imbalance gradients, thus can maintain a low imbalance ratio even after the first few steps (see Fig. 6c). In Table 1, our MD attack was also found to be more effective than ODI.

Exploration beyond the ϵ-ball By inspecting the individual attacks in AutoAttack, we found that the exploration technique used by the FAB attack is also effective against imbalanced gradients to some extent. FAB first finds a successful attack using unbounded perturbation size, then minimizes the perturbation to be within the ϵ-ball. As shown in Fig. 6b, the first few steps of exploration outside the ϵ-ball can effectively avoid imbalanced gradients. This is also why our MD attacks use a large step size in the first stage. However, the imbalance ratio tends to increase when FAB attempts to minimize the perturbation (steps 10–16). We believe FAB can be further improved following a similar strategy to our margin decomposition.

Circumventing imbalanced gradients improves black-box attacks Here we show gradient estimation based black-box attacks can also benefit from our MD method when there are imbalanced gradients. We take SPSA as an example and use the two-stage losses of our MD attack for SPSA. This version of SPSA is denoted as SPSA+MD. For both SPSA and SPSA+MD, we use the same batch size of 8192 with 100 iterations. Since black-box attack is quite time-consuming, we only run on 1000 randomly selected CIFAR-10 test images. The attack success rates on Adv-Interp, FeaScatter, and Sense models are reported in Table 7. Compared to SPSA, SPSA+MD can lower the robustness by at least 10.9%. This indicates that imbalanced gradient also has a negative impact on back-box attacks, and our method can be easily applied to produce more query-efficient and successful black-box attacks.

Table 7 Attack success rate (ASR, %) of the SPSA attack with or without our MD losses on three defense models. '↑' marks the ASR boost by MD

Attack	Adv-Interp	FeaScatter	Sense
SPSA	24.80	28.20	37.90
SPSA+MD	**40.30** (↑15.5)	**45.60** (↑17.4)	**48.80** (↑10.9)

The best results are boldfaced

5.4 Ablation and parameter analysis of MD attack

In this section, we provide a more detailed analysis of our proposed MD attack via an ablation study and a parameter analysis. The ablation study focuses on the two attacking stages of MD, while the parameter analysis focuses on the perturbation-related parameters including the number of steps and initial step size.

5.4.1 Ablation study

Here, we investigate two influential factors to our MD attack: 1) the second attacking stage, and 3) the stage order. We use AdvInterp as our target model, and conduct the following attack experiments on CIFAR-10 test data.

The second attacking stage We further investigate the importance of the second stage of attacking with the full margin loss in our MD and MD-MT attacks. The attack success rates with or without the second stage are also reported in Table 8. It shows that attacking the full margin loss via the second stage can increase the success rate, for both MD and MD-MT. This verifies the importance of the second stage for generating the strongest attacks.

Ordering of the two stages To verify that the ordering of the two stages is suitable for MD attacks, we evaluate a new version of our MD attacks with the two stages switched: the first stage optimizes the full margin loss and the second stage explores the individual loss terms. The results are also reported in Table 8 (the last two columns). As can be observed, MD attacks become less effective when the two stages are switched, even compared to that without the second stage. This indicates that the imbalanced gradients should be addressed first before producing a reliable robustness evaluation.

5.4.2 Parameter analysis

We further investigate the sensitivity of our MD attack to two parameters: 1) the number of perturbation steps, and 2) the initial step size. Here, we focus on the first attacking stage as the second stage is similar to the PGD attack, which has been investigated in (Wang et al., 2019).

Number of steps for the first stage Here, we fix the total number of steps for the two stages to $K = 100$ and vary the steps allocated for the first stage. Note that MD attack will reduce to the regular PGD attack if the step of its first stage is set to 0. Here, we vary the steps for the first stage from 5 to 50 in granularity of 5. The initial step size is fixed to $\alpha = 2\epsilon$ and gradually decreased to 0 via cosine annealing. The robustness of four defense models including Bilateral, Sense, Adv-Interp, and FeaScatter are illustrated in Fig. 7a. As

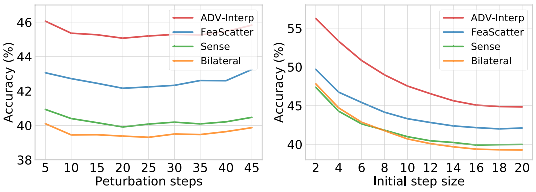
Table 8 Attack success rates (%) of our MD and MD-MT attacks 1) with or without the second stage, and 2) with or without the two stages being switched

Attacks	Second Stage		Switching Stage	
	Without	With	Yes	No
MD	44.53	**45.18**	43.17	**45.18**
MD-MT	52.56	**52.71**	51.47	**52.71**

Experiments are conducted on the defense model AdvInterp and CIFAR-10 dataset

The best results are boldfaced

Fig. 7 Parameter analysis: the accuracies (robustness) of four defense models under our MD attack with a different **a** number of steps, or **b** initial step sizes for the first stage



(a) Perturbation Steps

(b) Initial Step Size

can be observed, the effectiveness of our MD attack tends to drop at both ends, and the best performance (lowest evaluated robustness) is achieved at 20, except for Bilateral (which is 25). Therefore, we suggest using 20 steps as the optimal choice for the first stage.

Initial step size for the first stage We vary the initial step size used in the first stage from 2/255 to 20/255 in a granularity of 2/255. The initial step size will decrease to 0 with the cosine annealing. Following the above experiment, here we fix the number of steps for the first stage to 20. The evaluated robustness (or model accuracy on the generated attacks) of the four defense models are illustrated in Fig. 7b. A clear improvement of using a large initial step size can be observed. Based on the trend, we suggest using a relatively large ($\alpha \in [2\epsilon, 3\epsilon)$) initial step size to help circumvent imbalanced gradients during the first stage.

6 Conclusion

In this paper, we identified a subtle situation called *Imbalanced Gradients*, where existing attacks may fail to produce the most accurate adversarial robustness evaluation. We proposed a new metric to investigate the imbalanced gradients problem in current defense models. We also proposed a new attack called margin decomposition (MD) attack to leverage imbalanced gradients via a two-stage attacking process. The multi-targeted and ensemble version of MD attacks were also introduced to generate the strongest attacks. By re-evaluating 24 defense models proposed since 2018, we found that 11 of them are susceptible to imbalanced gradients to some extent and their robustness evaluated by the best standalone attack can be further reduced for more than 1% by our MD attack. We identified a set of possible causes of imbalanced gradients, and effective countermeasures. Our results indicate that future defenses should avoid causing imbalanced gradients to achieve more reliable adversarial robustness.

Acknowledgements This work was supported in part by the National Key R&D Program of China (Grant No. 2021ZD0112804), National Natural Science Foundation of China (Grant No. 62276067 and 62032006), and Shanghai Science and Technology Committee (Grant No. 20511101000). This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

Author contributions All authors contributed to the study conception and design. Conceptualization, material preparation, data collection and analysis were performed by XM, LJ and HH. Part of the experiments were also performed by ZW. Writing review and editing were performed by JB and YJ. YJ also contributed to funding acquisition. The first draft of the manuscript was written by XM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported in part by the National Key R&D Program of China (Grant No. 2021ZD0112804), National Natural Science Foundation of China (Grant No. 62276067 and 62032006), and Shanghai Science and Technology Committee (Grant No. 20511101000). This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

Data availibility statement All datasets used in this work are public datasets. The examined defense models are also publicly available in GitHub. This work did not introduce new datasets or models.

Declarations

Conflict of interest Xingjun Ma is currently employed at Fudan University and is also affiliated with The University of Melbourne as an honorary fellow. Linxi Jiang and Zejia Weng are master students at Fudan University. Hanxun Huang is a Ph.D. candidate at The University of Melbourne. James Bailey is employed at The University of Melbourne. Yu-Gang Jiang is employed at Fudan University.

Editorial board members and editors The authors declare that none of them is an Editorial Board Member or Editor of Machine Learning journal.

Financial interests The authors declare they have no financial interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability The code used to produce the results in this work is publicly available at https://github.com/HanxunH/MDAttack.

References

Alayrac, J., Uesato, J., & Huang, P. et al. (2019). Are labels required for improving adversarial robustness?. In *Neural information processing systems*.

Andriushchenko, M., Croce, F., & Flammarion, N., et al. (2019). Square attack: a query-efficient black-box adversarial attack via random search. arXiv:1912.00049.

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*.

Bai, Y., Zeng, Y., Jiang, Y., et al. (2020). Improving adversarial robustness via channel-wise activation suppressing. In *International conference on learning representations*.

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In S &P.

Carlini, N., Athalye, A., Papernot, N., et al. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

Carmon, Y., Raghunathan, A., Schmidt, L. et al. (2019). Unlabeled data improves adversarial robustness. In *Neural information processing systems*.

Chan, A., Tay, Y., Ong, Y.S., et al. (2020). Jacobian adversarially regularized networks for robustness. In *International conference on learning representations*.

Chen, J., Zhou, D., & Yi, J., et al. (2018a). A frank-wolfe framework for efficient and effective adversarial attacks. arXiv preprint arXiv:1811.10828.

Chen, P.Y., Sharma, Y., & Zhang, H., et al. (2018b). Ead: elastic-net attacks to deep neural networks via adversarial examples. In *AAAI conference on artificial intelligence*.

Croce, F., & Hein, M. (2019). Minimally distorted adversarial examples with a fast adaptive boundary attack. arXiv:1907.02044.

Croce, F., & Hein, M. (2020a). Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International conference on machine learning*, PMLR.

Croce, F., & Hein, M. (2020b). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. arXiv preprint arXiv:2003.01690.

Croce, F., Andriushchenko, M., & Sehwag, V. et al. (2020). Robustbench: A standardized adversarial robustness benchmark. arXiv preprint arXiv:2010.09670.

Croce, F., Andriushchenko, M., & Sehwag, V., et al. (2021). Robustbench: A standardized adversarial robustness benchmark. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track*.

Das, N., Shanbhogue, M., & Chen, S.T., et al. (2018). Compression to the rescue: Defending from adversarial attacks across modalities. In *KDD*.

Debenedetti, E., Sehwag, V., & Mittal, P. (2022). A light recipe to train robust vision transformers. arXiv preprint arXiv:2209.07399.

Ding, G.W., Sharma, Y., & Lui, K.Y.C. et al. (2018). Max-margin adversarial (MMA) training: Direct input space margin maximization through adversarial training. arXiv preprint arXiv:1812.02637.

Dong, Y., Liao, F., & Pang, T., et al. (2018). Boosting adversarial attacks with momentum. In *IEEE/CVF international conference on computer vision*.

Du, X., Zhang, J., Han, B., et al. (2021). Learning diverse-structured networks for adversarial robustness.

Engstrom, L., Ilyas, A., & Athalye, A. (2018). Evaluating and understanding the robustness of adversarial logit pairing. arXiv preprint arXiv:1807.10272.

Engstrom, L., Ilyas, A., Santurkar, S., et al. (2019). Adversarial robustness as a prior for learned representations. arXiv preprint arXiv:1906.00945.

Goldblum, M., Fowl, L., & Feizi, S., et al. (2020). Adversarially robust distillation. In *AAAI conference on artificial intelligence*.

Goodfellow, I.J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International conference on learning representations*.

Gowal, S., Uesato, J., & Qin, C., et al. (2019). An alternative surrogate loss for pgd-based adversarial testing. arXiv preprint arXiv:1910.09338.

Gu, S., & Rigazio, L. (2014). Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068.

Guo, C., Rana, M., & Cisse, M., et al. (2018). Countering adversarial images using input transformations. In *International conference on learning representations*.

Hendrycks, D., Lee, K., & Mazeika, M. (2019). Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*.

Huang, G., Liu, Z., & Van Der Maaten, L., et al. (2017). Densely connected convolutional networks. In *IEEE/CVF international conference on computer vision*.

Huang, H., & Wang, Y., Erfani, S.M., et al. (2021). Exploring architectural ingredients of adversarially robust deep neural networks. In *Neural information processing systems*.

Kim, J., & Wang, X. (2020). Sensible adversarial learning. https://openreview.net/forum?id=rJlf_RVKwr.

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale.

Liao, F., Liang, M., & Dong, Y., et al. (2018). Defense against adversarial attacks using high-level representation guided denoiser. In *IEEE/CVF international conference on computer vision*.

Liu, H., Simonyan, K., & Yang, Y. (2019). DARTS: Differentiable architecture search. In *International conference on learning representations*.

Liu, Q., Liu, T., & Liu, Z., et al. (2018). Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In: *ASPDAC*.

Ma, X., Li, B., & Wang, Y., et al. (2018). Characterizing adversarial subspaces using local intrinsic dimensionality. In *International conference on learning representations*.

Madry, A., Makelov, A., & Schmidt, L., et al. (2018). Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations*.

Pang, T., Yang, X., & Dong, Y., et al. (2020). Boosting adversarial training with hypersphere embedding. In *Neural information processing systems*.

Papernot, N., McDaniel, P., Wu, X., et al. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In S &P.

Papernot, N., McDaniel, P., Goodfellow, I., et al. (2017). Practical black-box attacks against machine learning. In *Asia CCS*.

Rakin, A.S., Yi, J., & Gong, B., et al. (2018). Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions. arXiv preprint arXiv:1807.06714.

Rice, L., Wong, E., & Kolter, Z. (2020). Overfitting in adversarially robust deep learning. In *International conference on machine learning*, PMLR (pp. 8093–8104).

Ross, A.S., & Doshi-Velez, F. (2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI conference on artificial intelligence*.

Salman, H., Ilyas, A., & Engstrom, L., et al. (2020). Do adversarially robust imagenet models transfer better? *Neural Information Processing Systems*

Samangouei, P., Kabkab, M., & Chellappa, R. (2018). Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International conference on learning representations*.

Shao, R., Shi, Z., Yi, J., et al. (2021). On the adversarial robustness of visual transformers. arXiv preprint arXiv:2103.15670.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Spall, J.C., et al. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*

Sutskever, I., Martens, J., & Dahl, G., et al. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*.

Szegedy, C., Zaremba, W., & Sutskever, I., et al. (2014). Intriguing properties of neural networks. In *International conference on learning representations*.

Tang, S., Gong, R., & Wang, Y., et al. (2021). Robustart: Benchmarking robustness on architecture design and training techniques. arXiv preprint arXiv:2109.05211.

Tashiro, Y., Song, Y., & Ermon, S. (2020). Diversity can be transferred: Output diversification for white- and black-box attacks. In *Advances in neural information processing systems*.

Tramèr, F., Kurakin, A., Papernot, N., et al. (2018). Ensemble adversarial training: Attacks and defenses. In *International conference on learning representations*.

Tramer, F., Carlini, N., & Brendel, W., et al. (2020). On adaptive attacks to adversarial example defenses. In *Neural information processing systems*.

Uesato, J., O'Donoghue, B., & Kohli, P., et al. (2018). Adversarial risk and the dangers of evaluating against weak attacks. In *International conference on machine learning*.

Wang, J., & Zhang, H. (2019). Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *International conference on computer vision*.

Wang, Y., Ma, X., & Bailey, J., et al. (2019). On the convergence and robustness of adversarial training. In *International conference on machine learning*.

Wang, Y., Zou, D., & Yi, J., et al. (2020). Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*.

Wong, E., Rice, L., & Kolter, J.Z. (2020). Fast is better than free: Revisiting adversarial training. In *International conference on learning representations*.

Wu, D., Xia, S.T., & Wang, Y. (2020). Adversarial weight perturbation helps robust generalization. *Neural Information Processing Systems*, *33*.

Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155.

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. In *BMVC*.

Zhang, H., & Wang, J. (2019). Defense against adversarial attacks using feature scattering-based adversarial training. In *Neural information processing systems*.

Zhang, H., & Xu, W. (2020). Adversarial interpolation training: A simple approach for improving model robustness. https://openreview.net/forum?id=Syejj0NYvr.

Zhang, H., Yu, Y., & Jiao, J., et al. (2019). Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*.

Zheng, T., Chen, C., & Ren, K. (2019). Distributionally adversarial attack. In *AAAI conference on artificial intelligence*.

Zhu, J., Yao, J., & Han, B., et al. (2021). Reliable adversarial distillation with unreliable teachers. arXiv preprint arXiv:2106.04928.

Zi, B., Zhao, S., & Ma, X., et al. (2021). Revisiting adversarial robustness distillation: Robust soft labels make student better. In *International conference on computer vision*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Xingjun Ma[1] · Linxi Jiang[1] · Hanxun Huang[2] · Zejia Weng[1] · James Bailey[2] · Yu-Gang Jiang[1]

Xingjun Ma
xingjunma@fudan.edu.cn

Linxi Jiang
lxjiang18@fudan.edu.cn

Zejia Weng
zjweng20@fudan.edu.cn

James Bailey
baileyj@unimelb.edu.au

[1] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China

[2] School of Computing and Information Systems, The University of Melbourne, Parkville, Australia