EasyChair Preprint

№ 11020

CMMR: a Composite Multidimensional Models Robustness Evaluation Framework for Deep Learning

Liu Wanyi, Zhang Shigeng, Wang Weiping, Zhang Jian and Liu Xuan

October 4, 2023

CMMR: A Composite Multidimensional Models Robustness Evaluation Framework for Deep Learning

Wanyi Liu[1], Shigeng Zhang(✉)[1], Weiping Wang[1], Jian Zhang[1], and Xuan Liu[2]

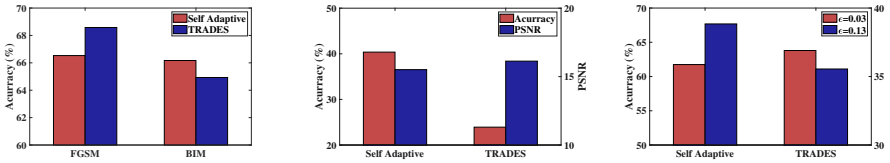
[1] School of Computer Science, Central South University, Changsha 410083, China
[2] School of Computer Science, HuNan University, Changsha 410082, China
{wyliu, sgzhang, wpwang, jianzhang}@csu.edu.cn
xuan_liu@hnu.edu.cn

Abstract. Accurately evaluating the defense models against adversarial examples has been proven to be a challenging task. We have recognized the limitations of mainstream evaluation standards, which fail to account for the discrepancies in evaluation results arising from different adversarial attack methods, experimental setups, and metrics sets. To address these disparities, we propose the Composite Multidimensional Model Robustness (CMMR) evaluation framework, which integrates three evaluation dimensions: attack methods, experimental settings, and metrics sets. By comprehensively evaluating the model's robustness across these dimensions, we aim to effectively mitigate the aforementioned variations. Furthermore, the CMMR framework allows evaluators to flexibly define their own options for each evaluation dimension to meet their specific requirements. We provide practical examples to demonstrate how the CMMR framework can be utilized to assess the performance of models in enhancing robustness through various approaches. The reliability of our methodology is assessed through both practical examinations and theoretical validations. The experimental results demonstrate the excellent reliability of the CMMR framework and its significant reduction of variations encountered in evaluating model robustness in practical scenarios.

Keywords: Robustness evaluation · Adversarial attacks · Adversarial machine learning.

1 Introduction

In recent years, with the deepening of deep learning models in research and practical applications, there has been rapid development in the field of deep learning models. Although deep learning models have been shown to be vulnerable to adversarial attacks [1], defense methods against them have also emerged [2–4].



(a) The accuracy of Self Adaptive and TRADES under FGSM and BIM with $\epsilon = 0.03$

(b) The accuracy and PSNR of Self Adaptive and TRADES under FGSM with $\epsilon = 0.12$

(c) The accuracy of Self Adaptive and TRADES under FGSM and BIM with $\epsilon = 0.03$ and $\epsilon = 0.13$

Fig. 1: Examples of evaluation under different adversarial attack methods, experimental settings, and metrics sets.

However, many proposed defense methods are quickly proven to have implemented incorrect or incomplete evaluations after their publication [5–7]. Therefore, we need a more comprehensive and accurate approach to evaluating the robustness of models.

In the current research on evaluating the robustness of deep learning models, several issues have been identified, including the following: inconsistent evaluation results due to the use of different attack methods, inconsistent evaluation results due to variations in the parameters set for the attack experiments, and inconsistent evaluation results due to the use of different evaluation metrics. We will now elaborate on each of these issues.

Attack Methods: It is common for researchers to propose new attack methods to circumvent newly developed defense methods, and subsequently, new defense methods are proposed to counter the previous attack methods, thus creating a continuous cycle. Therefore, evaluating the robustness of a deep learning model cannot rely solely on a single attack method. For instance, defense methods designed based on gradients can be easily defeated by attack methods that do not rely on gradient descent [8–10]. As shown in Fig. 1(a), models TRADES [4] and Self Adaptive [11] with the same perturbation budgets, the Self Adaptive achieves higher accuracy than the Trades model under the BIM, but its accuracy is lower than that of the TRADES model under the FGSM.

Experimental Parameters: When evaluating a model, it is generally assumed that the optimal parameters achieve the maximum attack success rate with the minimum attack cost, to observe the lower limit of model performance. However, for another attack method, the optimal experimental parameters may differ. Specifically, as shown in Fig. 1(c), at $\epsilon = 0.03$, TRADES exhibits higher classification accuracy than Self Adaptive under the FGSM attack method, but when $\epsilon = 0.13$, its performance is inferior to adaptive.

Evaluation Metrics: When assessing whether a model meets the user's requirements, it is necessary to select the metrics to be observed by the evaluated model and then assess the performance of these metrics. Currently, most evaluation metrics for model robustness only consider model classification accuracy, which is inadequate for defense methods. For instance, if a human observer can

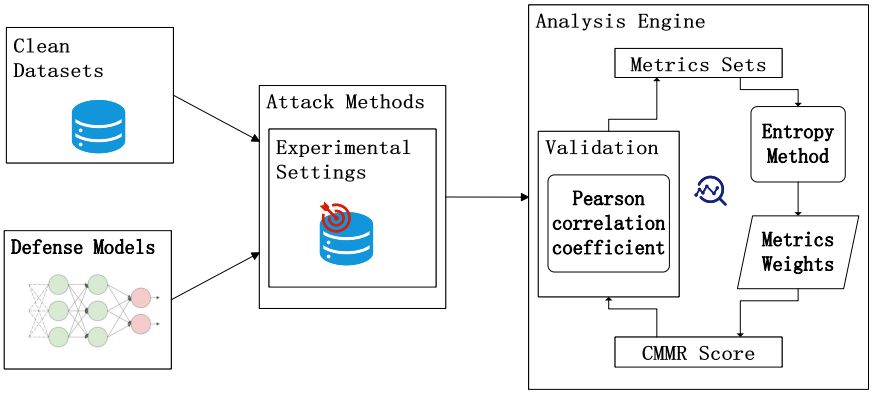


Fig. 2: Overview of CMMR. We confront clean data with defense models facing different types of adversarial attacks under different experimental parameter settings and finally obtain a multidimensional set of metric results. These metric result sets are fed into Analysis Engine to obtain our CMMR score.

directly identify the differences between an adversarial sample and a normal sample, it can be avoided from being input into the model. Moreover, the choice of different evaluation metrics leads to different evaluation results. As shown in Fig. 1(b), The values of Acc and PSNR for both Self Adaptive and TRADES under FGSM attack with $\epsilon = 0.12$. It can be seen that Self Adaptive's accuracy is higher than TRADES' accuracy, but Self Adaptive's PSNR is lower than TRADES' PSNR.

Due to the diverse range of evaluation dimensions mentioned above, it is challenging to establish a definitive criterion for assessing model robustness. Without a unified standard, fair comparisons cannot be made. DEEPSEC, proposed by Ling et al.[12], evaluates the robustness of each defense as an average, rather than based on the most effective attack against that defense [13]. Dong et al.[7] introduced robustness curves, but they did not demonstrate the model's performance under multiple attacks. Wu et al.[14] proposed the PSC Framework to address the issue of result discrepancies caused by different experimental settings in model robustness evaluation. However, they also did not consider the scenario where the model is subjected to multiple adversarial sample attacks. To address these issues, we propose a multidimensional comprehensive robustness evaluation method to accurately, comprehensively, and holistically assess the robustness of models. Specifically, our method is shown in Fig. 2. The evaluation process includes the following four steps: In the first step, select the counterattack method, set the parameters of the attack method, and select the set of metrics for evaluation. The second step inputs the evaluated model. In the third step, the entropy weight method is designed to obtain the weight of each metric. The new metrics M are calculated based on the metric weights and

metric values. In the fourth step, the metrics M under each attack method are calculated and equally weighted to obtain the final composite multidimensional model robustness evaluation score(CMMR). The CMMR is finally obtained to measure the robustness of the model under different attack methods, metrics, and experimental parameters.

Our contributions can be summarized as follows: We conducted extensive experiments to demonstrate the differences in model robustness under different attack methods, metrics, and experimental parameters. We analyze the factors that contribute to the discrepancies in model robustness evaluation and propose a Composite Multidimensional framework for evaluating Model Robustness(CMMR) in order to reduce the robustness evaluation discrepancies and provide a comprehensive and accurate assessment of model robustness. We selected two sets of adversarial training models with different perturbation budgets and observed and analyzed their robustness using our proposed method.

2 Attacks, Defense, and Metrics

In this section, we summarize typical adversarial attack methods, defense methods, and commonly used evaluation metrics.

2.1 Attack Methods

White-box Attacks: White-box attack means that the attacker knows the parameters and structure of the models. Most white-box attack methods craft adversarial examples based on input gradients. The Fast Gradient Sign Method (FGSM) [15] is a classical single-step attack algorithm that calculates the perturbation value for adversarial attacks solely based on the sign of the gradient. The Basic Iterative Method (BIM) [16] is an iterative version built upon FGSM, also known as the Iterative Fast Gradient Sign Method (I-FGSM). In BIM, the approach involves taking multiple small steps instead of a single large step as in FGSM. Projected Gradient Descent (PGD) [17] is another extension of FGSM that replaces the single large step with multiple small steps. Carlini and Wagner [18] proposed a group of optimization-based adversarial attacks, known as C&W attacks, which can generate adversarial samples CW_0, CW_2, and CW_∞ under L_0, L_2, and L_∞ norm constraints, respectively. Deep Fool [19] is an attack method based on hyperplane classification, aiming to find the minimum perturbation that leads to misclassification. Momentum Iterative Attack (MIA) [20] integrates momentum into the BIM attack and derives a new attack iteration algorithm. Its essence lies in the fact that the current perturbation is not only dependent on the current gradient direction but also on previous gradient directions.

Black-box Attacks: Transfer-based black-box attack: transfer-based attack craft adversarial examples against a substitute model against another unknown model with different parameters. The basic idea of SVRG [21] is to reduce

the intrinsic variance of Stochastic Gradient Descent (SGD) using prediction variance reduction, while reducing the intrinsic gradient variance of multiple models. Object-based diverse input (ODI) method [22] is proposed, which expands objects to draw counter images on 3D objects and classifies rendered images as target classes. **Decision-based black-box attack**: in this setting, only the probabilities or logits of the model are provided. Boundary [23] is a method for decision-based black-box attacks that simulates local geometric shapes to search for directions, effectively reducing the dimensionality of the search space. **Score-based black-box attack**: refers to situations where the attacker has access to the predicted probabilities from the model's final layer. CG-attack [24], whose main idea is to develop a new adversarial transferable mechanism that is robust to agent bias. The "$\mathcal{N}$ attack" [25] method focuses on finding the probability density distribution within a small region centered around the input, allowing the sampling from this distribution to potentially yield adversarial examples without accessing the internal layers or weights of the DNN.

2.2 Defense Methods

The field of adversarial attacks and defenses can be seen as a game, where the continuous emergence of new attack methods leads to the development of corresponding defense techniques. In this section, we classify the defense techniques into two categories, including model and data perspectives. We will provide an overview of defense strategies from these two angles: the model and the data.

Model: The methods [26, 27]were initially proposed to enhance the model's generalization ability and render it highly resilient to adversarial examples. It involves the utilization of defensive distillation to smooth the trained model during the training process. However, in 2017, Carlini and Wagner [18] declared the ineffectiveness of this method. Adversarial training [28, 29]is another defense method in which noise is introduced and parameters are regularized to alter the model's parameters, thereby improving its robustness. However, Shafahi et al. [30] demonstrated that no matter how much adversarial training is performed, there will always exist adversarial examples capable of deceiving neural networks. APMSA [31], AID [32] as a model-assisted classifier that does not change the original model structure and assists in defending against adversarial attacks. In addition, there are adversarial examples detection methods [33] for detecting adversarial samples to avoid input models.

Data: Luo et al. [34]proposes a defense mechanism based on the foveation mechanism, which can defend against adversarial perturbations generated by L-BFGS and FGSM methods. The assumption behind this defense is that a CNN classifier trained on a large dataset is robust to image scaling and transformation variations. Xie et al. [35] discovered that introducing random resizing to training images can weaken the strength of adversarial attacks. Other methods include random padding and image augmentation during the training process.

2.3 Comparison Metrics

To ascertain whether a model satisfies the criteria set by evaluators, it is imperative to carefully select the metrics that will be monitored to evaluate the performance of the model under consideration. In the case of deep learning models, classification accuracy is undoubtedly a fundamental metric used for evaluating model performance. However, in practical applications, if an image undergoes substantial perturbations that render it easily identifiable to the human eye, subsequent evaluations become inconsequential. Therefore, supplementary metrics are employed to gauge the quality of images before and after they are subjected to adversarial attacks. This section presents an introduction to the chosen metrics.

Accuracy If $\mathcal{A}_{\epsilon,p}$represents an attack setting for generating adversarial examples with perturbation size ϵ under the ℓ_p, and $x^{adv} = \mathcal{A}_{\epsilon,p}(x)$ denotes the adversarial example generated from a clean sample x under this attack setting, C represents a model classifier with a defense method. Then, the classification accuracy of the model classifier C under adversarial attacks can be expressed as

$$ACC\left(C, \mathcal{A}_{,p}\right) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(C\left(\mathcal{A}_{,p}\left(\mathbf{x}_i\right)\right) = y_i\right), \tag{1}$$

where $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$ is test set, $\mathbf{1}(\cdot)$ is the indicator function.

Average Structural Similarity(ASS) The change of ASS [36] before and after an image against attack is expressed as

$$ASS(\mathbf{x}, \mathcal{A}_{\epsilon,p}) = [l(\mathbf{x}, \mathbf{x^{adv}})]^{\alpha}[c(\mathbf{x}, \mathbf{x^{adv}}])]^{\beta}[s(\mathbf{x}, \mathbf{x^{adv}}])]^{\gamma}, \tag{2}$$

where $l()$ means luminance, $c()$ means contrast, and $s()$ means structure.

Mean Squared Error (MSE) x denotes the original image with dimensions $m * n$, and x^{adv}denotes the image obtained by subjecting it to an adversarial attack method. The average mean squared error (MSE) of a dataset can be mathematically expressed as

$$MSE\left(x, \mathcal{A}_{\epsilon,p}\right) = \frac{1}{N} \sum_{i=1}^{N} (\frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [x(i,j) - x^{adv}(i,j)]^2). \tag{3}$$

Average L_2 Distortion(ALD_2) is used to measure the similarity of two images.The ALD_2 of the original dataset after the adversarial attack is expressed as

$$ALD_2(\mathbf{x}, \mathcal{A}_{\epsilon,p}) = \frac{1}{N} \sum_{i=1}^{N} (\|x_i - x_i^{adv}\|_2). \tag{4}$$

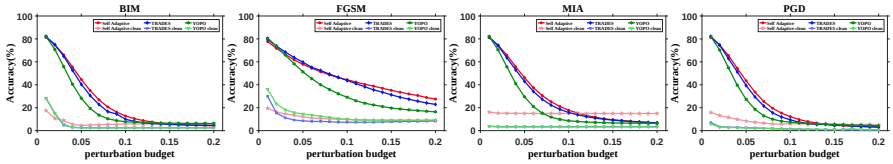


Fig. 3: The accuracy vs. perturbation budget curves of the 6 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

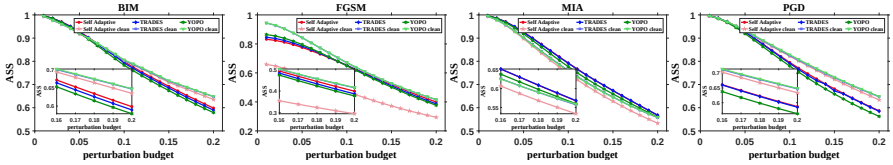


Fig. 4: The ASS vs. perturbation budget curves of the 6 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

Peak Signal-to-Noise Ratio (PSNR) is one of the standards used to measure image quality. For color channels, the MSE values of the three RGB channels are calculated separately and then averaged to obtain the PSNR value, and the formula for calculating PSNR is

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right), \tag{5}$$

where MAX_I represents the maximum pixel value of the image.

3 Composite Multidimensional Model Robust Evaluation Method

The extensive deployment of deep learning models in practical applications underscores the critical importance of accurately evaluating the effectiveness of a classification model in the face of the continuous and profound development of adversarial attack methods. In this section, we will analyze how to evaluate the robustness of a model in three dimensions and introduce our Composite Multidimensional Model Robustness Evaluation Framework (CMMR).

3.1 Motivation

Some defense methods are specifically designed to counter a particular type of adversarial attack, but their defensive capabilities are greatly diminished against other attack methods. For instance, the original Fast Gradient Sign Method (FGSM) [37] generates adversarial samples based on gradient information. However, this method becomes ineffective when confronted with gradient masking

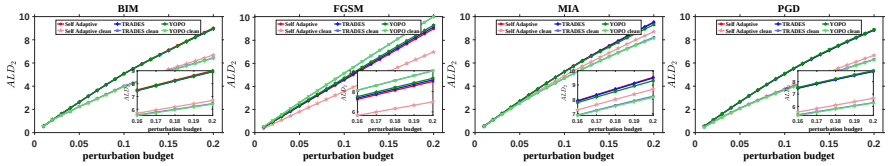


Fig. 5: The ALD_2 vs. perturbation budget curves of the 6 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

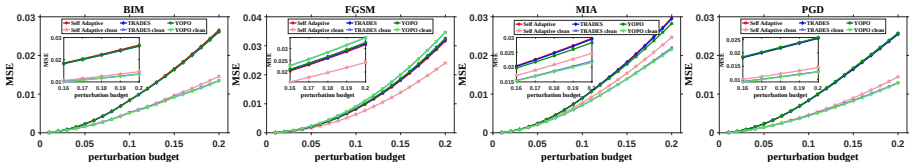


Fig. 6: The MSE vs. perturbation budget curves of the 6 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

caused by simple adversarial training [38]. As indicated in Fig. 3 under FGSM attack, in terms of the achieved accuracy (Acc) in the experimental results, the Self Adaptive Robust model performs significantly better than the other three models under FGSM attacks. This suggests that the defense method employed by Self Adaptive effectively evades FGSM attacks. During the training process, Self Adaptive dynamically corrects mislabeled samples based on the model's predictions. Since FGSM is a simple single-step gradient-based attack, the defense mechanism of Self Adaptive can successfully evade such attacks and achieve a higher accuracy rate. Therefore, in order to comprehensively evaluate the robustness of a model, it is advisable to consider employing multiple adversarial attack methods for model evaluation.

Considering the second scenario, the experimental results vary even for models under the same attack method with different adversarial perturbation size settings. Specifically, as shown in Fig. 3 under FGSM attack, the TRADES model consistently outperforms the Self Adaptive model when $\epsilon < 0.09$. However, when $\epsilon > 0.09$, the performance gap between TRADES and Self Adaptive models increases with the increase of perturbation budget, and the Self Adaptive model obtains higher accuracy (Acc). In order to comprehensively evaluate the performance of the model under different perturbation strengths, the second feature of this study's methodology is to evaluate the model in multiple experimental environments. More precisely, we compare different adversaries in a specific comparison range $\epsilon = [0.01, 0.2]$ with an interval of 0.1, under the same attack method, resulting in a total of 20 parameter settings.

Furthermore, the concept of "robustness" refers to the capacity of a system to maintain specific performance characteristics when subjected to perturbations in certain parameters (such as structure or magnitude) [39]. Robustness plays a vital role in ensuring the system's survival in abnormal and hazardous circum-

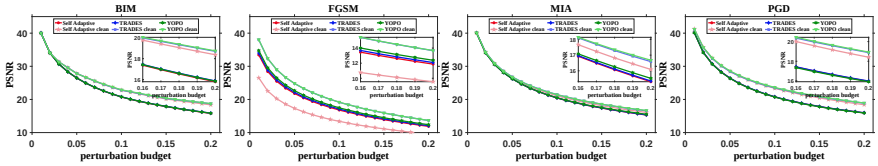


Fig. 7: The PSNR vs. perturbation budget curves of the 6 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

stances. When it comes to assessing the robustness of models, many existing evaluation methods are confined to measuring model classification accuracy [7, 40], which is evidently inadequate. Specifically, in Fig. 3and Fig. 4, under the MIA attack with $\epsilon = 0.05$, Self Adaptive outperforms the other models when classification accuracy is used as the evaluation criterion, while the TRADES model outperforms the other models when ASS is used as the evaluation criterion. Therefore, in order to comprehensively evaluate the robustness of the model, we chose multiple evaluation metrics to assess the performance of the model under different attack methods.

In this study, we assume that all adversaries possess the capability to apply the maximum perturbation to attack the model, in order to observe the model's performance under the strongest attack method.

3.2 CMMR Framework

Getting metrics weights by Entropy Weight Method. For the selected multiple metrics, different evaluators find it challenging to make consistent judgments regarding the relative importance of each metric. Therefore, we employ the entropy weighting method to objectively assign weights to the selected metrics. The entropy weighting method was initially proposed within Shannon's formula [41]. In the field of statistics, it is widely acknowledged that as data becomes more dispersed, the entropy value decreases, indicating greater importance of the corresponding metric. This concept is also applicable in the context of adversarial attacks. For example, when evaluating models, if the variation of metric A is relatively small among all models under the same attack and at the same attack intensity, selecting it as part of the evaluation result would have minimal impact on the final evaluation outcome. Conversely, if metric B exhibits a significant variation, it would have a substantial influence on the final evaluation result. The following are the detailed steps of the improved entropy weighting method used in this approach.

Step 1: Define the evaluation target and establish the evaluation metric system. Construct the preference matrix R', where the horizontal vectors of R' represent the set of evaluation metrics. R'_1 to R'_5 represent Acc, ASS, MSE, ALD_2, $PSNR$, respectively. Additionally, an extra metric R'_6 is included to balance the importance of selecting metric weights. The column vector represents the values

of these metrics under the setting $\epsilon = [0.01, 0.2]$. For example, R'_{11}represents the value of Acc under attack method A with $\epsilon = 0.01$.

Step 2: Normalize the preference matrix to obtain R.

Step 3: Calculate the entropy value for each matrix using the formula

$$H_j = -k \sum_{i=1}^{m} f_{ij} \cdot \ln f_{ij}, \tag{6}$$

where $f_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}}, k = \frac{1}{\ln m}$, f_{ij}represents the weight of the i parameter setting's metric value under the j metric.

Step 4: Determine the weight of each metric as

$$\lambda_j = \frac{\lambda'_j w_j}{\sum_{j=1}^{n} \lambda'_j w_j}, \tag{7}$$

where $w_j = \frac{1-H_j}{\sum_{j=1}^{n}(1-H_j)}$ denotes the entropy weight of the j metric.

Step 5: Based on the weights assigned to each indicator, a new evaluation indicator M,indicator m is calculated as

$$M = \sum_{j=1}^{n} R \cdot w_j. \tag{8}$$

Finally, employ Pearson's coefficients to analyze the reliability and consistency between M and R_i. Adjust the weights iteratively until the three coefficients reach their maximum values, indicating that the final evaluation metric is reliable.

Synthesizing adversarial attacks by Equivalent Weighting. Assuming a lack of prior knowledge regarding the evaluated model's defense mechanisms against specific types of attacks, the model's susceptibility to any attack is considered equally probable. Prior to this, we computed the value of M for each adversarial attack method using Eq. 8. Following that, equal weights are assigned to the integrated metrics, M, for each adversarial attack method. The resulting values are then used to calculate the Comprehensive Multidimensional Model Robustness (CMMR) Score, which serves as a combined assessment score for evaluating the robustness of the model. The formula for calculating the CMMR Score is as

$$CMMR = \sum_{j=1}^{N} M \cdot w_j, w_j = \frac{1}{N}, \tag{9}$$

where N represents the total number of adversarial attack methods employed against the model.

Table 1: We show the structure of the defense models and their clean models that were incorporated into our adversarial robustness evaluation framework. We also show the original threat models (i.e., the threat models in the original paper where the defense system was trained to be robust or evaluated;), and the accuracy (%) of each method on clean data. The accuracies are recalculated by ourselves.

Defense Method	Model	Intended Threat	Clean Acc.
TRADES	WRN-34-10	L(=0.031)	84.59
YOPO	WRN-34-10	L(=0.03)	86.8
Self-adaptive	WRN-34-10	L(=0.031)	83.48
TRADES Natural	WRN-34-10	-	94.93
YOPO Natural	WRN-34-10	-	95.05
Self-adaptive Natural	WRN-34-10	-	66.33
FAT_MART_62	WRN-28-10	L(=16/255)	80.64
FAT_TRADES_62	WRN-34-10	L(=16/255)	82.41
FAT_MART_031	WRN-28-10	L(=8/255)	90.56
FAT_TRADES_031	WRN-34-10	L(=8/255)	89.44

Table 2: The Summary of Notations

Attack Method	Description	Utility Metrics	Description
FGSM	Fast Gradient Sign Method	ASS	Average Structural Similarity
BIM	Basic Iterative Method	PSNR	Peak Signal to Noise Ratio
PGD	Projected L Gradient Descent attack	MSE	Mean Square Error
MIA	Momentum Iterative Attack	ALDp	Average Lp Distortion

Curving CMMR Scores. If the model consistently exhibits superior performance compared to the adversary across the entire spectrum of perturbations, the problem can be considered straightforward. However, evaluations frequently exhibit intersections at specific points. To achieve a more comprehensive assessment of the model's robustness performance, we utilize the dimensionality reduction technique mentioned earlier to reduce the evaluation results from three dimensions to a single dimension. Subsequently, we plot the aggregated scores of diverse models at varying levels of perturbation intensity to gain insights into a more comprehensive evaluation outcome.

4 Experimental Analysis for CMMR

In section 3, we introduced the steps of the Composite Multidimensional Models Robustness Evaluation method(CMMR). In this section, we will employ the CMMR method to demonstrate the process of evaluating model performance from three dimensions to CMMR. Finally, we will show the validation of the CMMR method from practical and theoretical aspects respectively.

Dimension. Given that the evaluation of model robustness mentioned in the introduction requires multiple dimensions, including adversarial attack methods, evaluation metric sets, and experimental parameter settings, we will visually

demonstrate the process of evaluating model robustness in three dimensions, as well as the CMMR evaluation process of our method, in order to demonstrate the intuitiveness, wholeness, and correctness of our method.

Setting. We selected five groups of classification models under the CIFAR10 dataset, three of which were used as baseline comparisons, including the original model and the model with the defense method applied. The other two groups consist of models trained under different levels of adversarial perturbations using each of the two defense methods. Table 1 provides details of the defense models, the structure of the models, the budget of the adversarially trained perturbations, and the clean accuracy.

Validation. The CMMR validation consists of two steps: Practical validation and Theoretical validation. In the practical validation step, we compare the results of models with applied defense methods and models without defense methods, using CMMR, to determine if the outcomes align with real-world scenarios. In the theoretical validation part, we assess the reliability of the reduced data through Kendall's coefficient of concordance, as well as examine the correlation between the reduced data and the original set of metrics using Pearson's coefficient.

Acronyms and Notations. For convenient reference, we summarize the acronyms and notations in Table 2.

4.1 3-dimension Models Robustness Evaluation.

Fig. 3, 4, 5, 6,and 7 presents line graphs that demonstrate the relationship between the evaluation metric sets Acc, ASS, ALD_2, MSE, and $PSNR$ and the perturbation budget curves for three comparative models. The graphs are arranged from top to bottom and represent the performance of the models under non-targeted attacks such as FGSM, BIM, PGD, and MIA.

Acc. As the perturbation budget increases, both the natural models and the defense models experience a gradual decrease in classification accuracy. However, there is a difference in behavior. The classification accuracy of the model without defense methods significantly drops to its lowest point when the perturbation is small, and then only slightly decreases as the perturbation size increases. In contrast, robust models with the same structure maintain relatively high classification accuracy even with small perturbation budgets. Interestingly, the selected defense methods are all effective in defending against the FGSM attack. Even at the maximum perturbation budget, the model's classification accuracy remains better than the corresponding natural model. However, it is worth noting that the FGSM attack was not specifically designed to evaluate model robustness against strong attacks [42]. Another interesting observation is that although the Self Adaptive natural model exhibits a significant decrease in classification accuracy at small perturbation budgets, there is no significant change in accuracy as the perturbation budget increases. In fact, Self Adaptive clean model outperforms even most defense models for $\epsilon > 0.11$ under FGSM attacks. Additionally, we can observe that for the BIM attack method, the YOPO model demonstrates the highest robustness at $\epsilon > 0.13$. However, within this perturbation budget

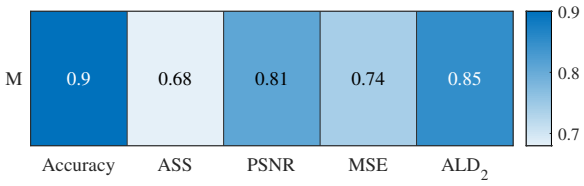


Fig. 8: The Pearson correlation coefficients of the metric M with the five metrics.

range, the YOPO model may not be the most robust against the other attack methods.

ASS. Among the different adversarial attack methods, the ranking of the three robust classification models in terms of $SSIM$ under the FGSM attack differs. When the perturbation size $\epsilon < 0.1$, the model rankings, from best to worst, are as follows: Self Adaptive, Trades, YOPO. At $\epsilon = 0.1$, the rankings are nearly the same, but as the perturbation size increases, the rankings change to YOPO, Trades, Self Adaptive, moving in the opposite direction. For the other adversarial attack methods, the model rankings do not change with the perturbation size.

PSNR. We are aware that a higher PSNR indicates better image quality, and it aligns with prior knowledge that the $PSNR$ of all models decreases as the perturbation budget increases. However, as the perturbation size increases, there is typically an increasing gap in $PSNR$ between natural models and robust models, with the curves of the robust models positioned below those of the natural models. Does this imply that natural models are less vulnerable to attacks compared to robust models? Not necessarily. This observation suggests that adversarially trained models often require more substantial changes in the image to induce misclassifications when confronted with adversarial attacks of the same perturbation size. Similar to the classification accuracy findings, the Self Adaptive clean model demonstrates distinct behavior compared to other robust models under the FGSM attack.

ALD_2. Compared to the aforementioned three metrics, its variation with respect to the perturbation budget is more consistent, meaning the curve of this metric is closer to a straight line.

MSE. The difference in MSE between the model groups increases with the perturbation size. However, under the FGSM attack, the MSE of the Self Adaptive clean model is lower than that of all models. For the other adversarial attack methods, the MSE curve of the Self Adaptive model lies between the curves of the robust models and the clean model.

The previous section showed the values of the five metric sets for the three comparison models under four attacks, and in this section, we show the process of evaluating the three comparison models using the CMMR method based on the above test results. Based on the test results, the weights of each metric are obtained by the entropy weighting method, where ω is the hyperparameter used to balance the importance of each weight. We set $\omega = 0.08$, and the calculated

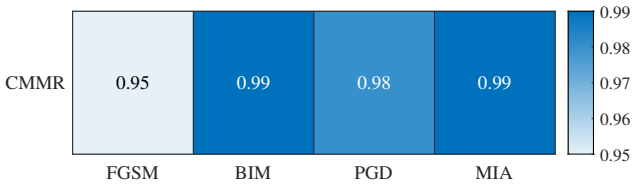


Fig. 9: The Pearson correlation coefficients of CMMR with each model indicator M under each attack method.

Table 3: The weights of each metric are obtained by the entropy weighting method

Metric	Weights Value
Accuracy	0.49
ASS	0.12
PSNR	0.12
MSE	0.12
ALD_2	0.07
w	0.08

weights of each index are shown in Table. 3 to obtain the composite metric M. Then we calculate the composite metric M under FGSM, BIM, PGD, and MIA respectively, and assign them with equal weights to obtain CMMR.

Fig. 10 shows the CMMR scores of TRADES, Self Adaptive, Yopo model, and their clean model. It is evident that the robust model consistently outperforms the clean model in terms of CMMR scores. Furthermore, as the perturbation magnitude increases, the disparity in CMMR scores between the clean and robust models diminishes, leading to score convergence. While the CMMR score of the Self Adaptive clean model surpasses that of the Self Adaptive robust model at $\epsilon = 0.13$, the difference is not statistically significant. Interestingly, empirical observations demonstrate that in the presence of attacks like BIM, PGD, and MIA, the accuracy of the Self Adaptive clean model at $\epsilon = 0.13$ exceeds that of the Self Adaptive robust model, thus validating the observations.

In a subsequent stage, we performed a theoretical validation. We used the Pearson correlation coefficient to test the correlation between the data since the metric M value and the five metrics two-by-two satisfy the following conditions:

- The relationship between the two variables is linear and both are continuous data.
- The overall distribution of the two variables is normal or near-normal with a single-peaked distribution.
- The observations of the two variables are paired, and each pair of observations is independent of each other.

Fig. 8 shows the Pearson correlation coefficients of the metric M with the five metrics, all of which are greater than 0.6, indicating that metric M is strongly

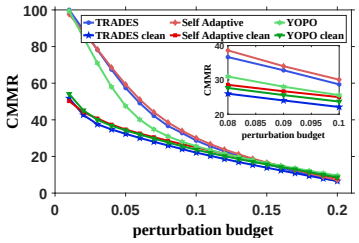


Fig. 10: The CMMR vs. perturbation budget curves of three sets of robust and clean comparison models.

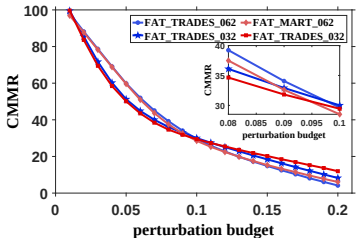


Fig. 11: The CMMR vs. perturbation budget curves of two sets of robust models trained by different perturbation budget.

correlated with the five metrics of the original measure. Similarly, Fig. 9 shows the Pearson correlation coefficients of CMMR with each model indicator M under each attack method. Since the assumption is that the model suffers from the same probability of each type of attack, the difference between the metric M and the final CMMR under each attack method is not significant and all are greater than 0.9, which is strongly correlated. Therefore, we can consider that in theory the CMMR represents the performance of the model exhibited by the combined selected metrics.

4.2 Evaluating the robustness of two sets of models by CMMR.

The robustness of the robust and clean models has been analyzed above. Now, let's examine the robustness performance of models trained with different perturbation magnitudes in adversarial training. We have selected two sets of models trained with different perturbation magnitudes for this study. The first set includes FAT for TRADES models trained with adversarial perturbations at $\epsilon = 8/255$ and $\epsilon = 16/255$, denoted as FAT_TRADES_031 and FAT_TRADES_062, respectively. The second set includes FAT for MART models trained with adversarial perturbations at $\epsilon = 8/255$ and $\epsilon = 16/255$, denoted as FAT_MART_031 and FAT_MART_062, respectively.

Fig. 11 shows the relationship between the CMMR scores of the two groups of models and the perturbation budget. It is evident that the robustness relationships of the FAT for TRADES group and the FAT for MART group are not consistently aligned. The red curve corresponds to the FAT for MART group, which exhibits two inflection points in its robustness relationship. For $\epsilon < 0.01$, the CMMR of FAT_TRADES_031 is higher than that of FAT_TRADES_062. However, as epsilon increases, FAT_TRADES_062 surpasses FAT_TRADES_031 until reaching $\epsilon = 0.95$. Subsequently, for epsilon values greater than 0.06, FAT_TRADES_031 outperforms FAT_TRADES_062 until $\epsilon = 0.2$. On the other hand, the blue curve represents the FAT for TRADES group, which also displays two inflection points in its robustness relationship. For $\epsilon < 0.01$, the CMMR of FAT_MART_031 is higher than that of FAT_MART_062. Sim-

ilar to the previous group, as epsilon increases, FAT_MART_062 surpasses FAT_MART_031 until reaching $\epsilon = 0.1$. Once again, for $\epsilon > 0.098$, FAT_MART_031 outperforms FAT_MART_062 until $\epsilon = 0.2$, and this pattern persists. Therefore, we can draw the conclusion that for models trained through adversarial training to enhance robustness, the robustness of the models is not directly correlated with the perturbation magnitude used during adversarial training. It is not necessarily the case that larger perturbations lead to better robustness.

5 Conclusion

In this stduy, we propose a Composite Multi-dimensional Robustness (CMMR) score to evaluate the robustness of models from multiple dimensions, including adversarial attack methods, selected metrics, and experimental settings. Currently, there is no unified framework in the literature for comprehensive multi-dimensional evaluation. The computation of the CMMR score involves three main steps, all of which can be standardized. We provide examples of how to assess the performance of models that enhance robustness in different ways. To ensure its reliability, we employ three coefficients that measure the consistency and reliability of the evaluation data. Evaluators also have the flexibility to define their own options for each evaluation dimension to meet their specific requirements. We believe that this approach will standardize and expedite the equitable comparison of model robustness.

References

1. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
2. Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13398–13408, 2022.
3. Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15273–15283, 2022.
4. Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.
5. Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.
6. Cory Cornelius, Nilaksh Das, Shang-Tse Chen, Li Chen, Michael E Kounavis, and Duen Horng Chau. The efficacy of shield under different threat models. *arXiv preprint arXiv:1902.00541*, 2019.

7. Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 321–331, 2020.
8. Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018.
9. Hesamodin Mohammadian, Ali A Ghorbani, and Arash Habibi Lashkari. A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems. *Applied Soft Computing*, 137:110173, 2023.
10. Yangguang Zhang, Can Wang, Qihao Shi, Yan Feng, and Chun Chen. Adversarial gradient-based meta learning with metric-based test. *Knowledge-Based Systems*, page 110312, 2023.
11. Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *Advances in neural information processing systems*, 33:19365–19376, 2020.
12. Xiang Ling, Shouling Ji, Jiaxu Zou, Jiannan Wang, Chunming Wu, Bo Li, and Ting Wang. Deepsec: A uniform platform for security analysis of deep learning model. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 673–690. IEEE, 2019.
13. Nicholas Carlini. A critique of the deepsec platform for security analysis of deep learning models. *arXiv preprint arXiv:1905.07112*, 2019.
14. Jing Wu, Mingyi Zhou, Ce Zhu, Yipeng Liu, Mehrtash Harandi, and Li Li. Performance evaluation of adversarial attacks: Discrepancies and solutions. *arXiv preprint arXiv:2104.11103*, 2021.
15. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
16. Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.
17. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
18. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.
19. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.
20. Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.
21. Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14983–14992, 2022.
22. Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through

object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2022.
23. Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.
24. Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shu-Tao Xia. Boosting black-box attack with partially transferred conditional adversarial distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15095–15104, 2022.
25. Aritran Piplai, Sai Sree Laya Chukkapalli, and Anupam Joshi. Nattack! adversarial attacks to bypass a gan based classifier trained to detect network intrusion. In *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 49–54. IEEE, 2020.
26. Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014.
27. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
28. Jianan Zhou, Jianing Zhu, Jingfeng Zhang, Tongliang Liu, Gang Niu, Bo Han, and Masashi Sugiyama. Adversarial training with complementary labels: On the benefit of gradually informative attacks. *arXiv preprint arXiv:2211.00269*, 2022.
29. Zhong-Han Niu and Yu-Bin Yang. Defense against adversarial attacks with efficient frequency-adaptive compression and reconstruction. *Pattern Recognition*, 138:109382, 2023.
30. Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5636–5643, 2020.
31. Muhammad Asghar Khan, Hosam Alhakami, Insaf Ullah, Wajdi Alhakami, Syed Agha Hassnain Mohsan, Usman Tariq, and Nisreen Innab. A resource-friendly certificateless proxy signcryption scheme for drones in networks beyond 5g. *Drones*, 7(5):321, 2023.
32. Duhun Hwang, Eunjung Lee, and Wonjong Rhee. Aid-purifier: A light auxiliary network for boosting adversarial defense. *Neurocomputing*, 541:126251, 2023.
33. Jiefei Wei, Luyan Yao, and Qinggang Meng. Self-adaptive logit balancing for deep neural network robustness: Defence and detection of adversarial attacks. *Neurocomputing*, 531:180–194, 2023.
34. Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015.
35. Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017.
36. Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
37. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
38. Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

39. Peter J Rousseeuw, Frank R Hampel, Elvezio M Ronchetti, and Werner A Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.
40. Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15105–15114, 2022.
41. Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.
42. Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.