*mathematics*

MDPI

*Article*

# Robustness Learning via Inference-Softmax Cross Entropy in Misaligned Distribution of Image

**Bingbing Song [1,†], Ruxin Wang [2,3], Wei He [2,3] and Wei Zhou [2,3,*,†]**

[1] School of Information Science and Engineering, Yunnan University, Kunming 650091, China
[2] Piolet School of Software, Yunnan University, Kunming 650091, China
[3] Engineering Research Center of Cyberspace, Yunnan University, Kunming 650091, China
* Correspondence: zwei@ynu.edu.cn
† These authors contributed equally to this work.

**Abstract:** Adversarial examples easily mislead vision systems based on deep neural networks (DNNs) trained with softmax cross entropy (SCE) loss. The vulnerability of DNN comes from the fact that SCE drives DNNs to fit on the training examples, whereas the resultant feature distributions between the training and adversarial examples are unfortunately misaligned. Several state-of-the-art methods start from improving the inter-class separability of training examples by modifying loss functions, where we argue that the adversarial examples are ignored, thus resulting in a limited robustness to adversarial attacks. In this paper, we exploited the inference region, which inspired us to apply margin-like inference information to SCE, resulting in a novel inference-softmax cross entropy (I-SCE) loss, which is intuitively appealing and interpretable. The inference information guarantees that it is difficult for neural networks to cross the decision boundary under an adversarial attack, and guarantees both the inter-class separability and the improved generalization to adversarial examples, which was further demonstrated and proved under the min-max framework. Extensive experiments show that the DNN models trained with the proposed I-SCE loss achieve a superior performance and robustness over the state-of-the-arts under different prevalent adversarial attacks; for example, the accuracy of I-SCE is 63% higher than SCE under the $PGD_{50}^{un}$ attack on the MNIST dataset. These experiments also show that the inference region can effectively solve the misaligned distribution.

**Keywords:** neural networks; robustness learning; loss function; adversarial examples

**MSC:** 68T07



**Citation:** Song, B.; Wang, R.; He, W.; Zhou, W. Robustness Learning via Inference-Softmax Cross Entropy in Misaligned Distribution of Image. *Mathematics* **2022**, *10*, 3716. https://doi.org/10.3390/math10193716

Academic Editor: Jakub Nalepa

Received: 12 September 2022
Accepted: 5 October 2022
Published: 10 October 2022



## 1. Introduction

Although deep neural networks have achieved a state-of-the-art performance in various tasks, such as image recognition [1] and natural language processing [2], it has been recently shown that adversarial examples by adding imperceptible disturbances do not find it difficult to fool *well-trained* neural networks [3,4], leading to malfunctions in intelligent systems [5–7]. The vulnerability to adversarial attacks indicates that the neural networks do not convey proper feature representations and may *overfit* on the training examples [8]. Schmidt et al. [9] show that the sample complexity of robust learning can be significantly larger than that of standard learning. Given the difficulty of training robust classifiers in practice, they further postulate that the difficulty could stem from the insufficiency of training examples in the commonly used datasets, e.g., CIFAR-10. Recent work intends to solve this problem by utilizing extra unlabeled data [10,11]. However, extra data are usually not available. Therefore, optimization strategies and training loss functions are very important for robust learning.

Take the softmax cross entropy (SCE) loss as an example, which is widely adopted in regressing probabilities and is a core building block for a high performance. The neural

networks trained with SCE are shown to be limited in their robustness to input perturbation, and hence are suboptimal in real applications where adversarial attacks exist [12].

The above issue brings many attempts that optimize SCE to enhance the robustness and anti-attack properties of neural networks. Several variants have been proposed to promote the effectiveness of the softmax loss, such as comparing loss [13], triplet loss [14], center loss [15], large-margin Gaussian mixture (L-GM) loss [16], and max-Mahalanobis center (MMC) loss [17]. These methods are led by the same principle in that they minimize the losses to maximally fit the training examples.

However, the adversarial examples have a misaligned distribution with the training data, meaning that the fitted models in training could be repellent to the adversarial data [8]. In fact, given a well-trained model, the distribution difference between the training and adversarial data is a blind region to the model, which we regard as the inference region. The examples in this region are expected to be generalizable by the well-trained model, which is not the case in existing methods, resulting in the vulnerability of neural networks [4]. The reason for why this region exists, according to our analyses, is that the model overfits on the training data even when large amounts of data are accessible in training and the adversarial data are clearly absent. Hence, how to generalize the examples in this region still remains unresolved. Unfortunately, the above methods fail to take this fact into consideration. Although the existing examples in the distribution space are unchangeable, adversarial training [18] adds adversarial examples as extra data to existing examples to change the distribution of training examples and expects to solve the issue of the misaligned distribution between the training examples and the adversarial examples. However, adversarial examples cannot be completely found and adversarial training only relieves the misalignment. A robust learning strategy urgently needs to solve the misalignment problem.

Considering the misalignment, we exploited the inference region between the distributions of training data and adversarial examples. This region guides us to develop an inference schema that imposes a margin-like inference information on the predicted logit of the network. Based on this, this study propose an inference-softmax cross entropy (I-SCE) loss as a substitute for the SCE loss. In this loss, the inference information is intuitively regarded as an additive term imposed on the prediction, which is extremely easy to implement and appealing. This study further shows the robustness of I-SCE under the min-max framework. Under severe adversarial attacks, I-SCE still maintains a high accuracy and robustness, and has a better performance. The experiments on MNIST and CIFAR10 demonstrate that the proposed loss produces an improved effectiveness and robustness compared with the state-of-the-art methods.

The main contributions are as follows:

(1) Considering the misalignment of related work, this study exploited the inference region to deal with the misalignment of adversarial examples and clean examples. In addition, this paper also discusses, in detail, the advantages of introducing an inference region.
(2) Based on this, this study proposes an inference-softmax cross entropy (I-SCE) loss as a substitute for the SCE loss without complicated implementation.
(3) Proved by rigorous theory and extensive experiments, I-SCE is more effective and robust compared with the state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we present a brief review of related works. In Section 3, we describe the methodology of the article (inference region, inference-softmax cross entropy) and a robustness analysis of I-SCE in detail (expected interval of correct class, proof of the property on the expected interval of correct class, and min-max framework). In Section 4, the experimental results are displayed and analyzed. Section 5 discusses practical and theoretical implications. Finally, Section 6 summarizes our work and looks forward to future research.

## 2. Related Work

Adversarial attacks exist widely in the open environment, imposing a critical robustness demand of neural networks regarding the security quality and overall performance of systems. Therefore, how to design an anti-attack and robust neural network has attracted the interest of many researchers, who are briefly reviewed in this section.

### *2.1. Adversarial Attack*

Szegedy et al. [4] first proposed the concept of adversarial examples and employed the L-BFGS method, as the solver of the disturbed problem, to mislead neural networks. Goodfellow et al. [3] proposed the fast gradient symbol method (FGSM) to generate adversarial examples with a single gradient step. Before backpropagation, FGSM was used to perturb the input of the model, which was an early form of adversarial training. Moosavi-Dezfooli et al. [19] proposed the DeepFool, which calculated the minimal necessary disturbance and applied it to construct adversarial examples. By imposing the $\ell_2$ regularization to limit the disturbance scale, DeepFool achieved a good performance. After this, Madry et al. [18] proposed the projected gradient descent (PGD) attack, which had a strong attack strength, and was used in adversarial training to improve robustness. Recently, Guo et al. [20] developed a local searching-based technique to construct a numerical approximation of the gradient, which was then used to perturb a small part of the input image.

### *2.2. Adversarial Defense*

The features of adversarial examples could follow a different distribution from the clean training data, making the defense progress very difficult. Some works detect adversarial examples and remove adversarial noise. Metzen et al. [21] introduced a novel model to detect adversarial examples. Xie et al. [22] proposed the use of random resizing and random padding on images for defense. In addition, the regularization and penalty term can also make the model more robust. Ross et al. [23] and Yan et al. [24] proposed regularizing the gradients during training to improve the model robustness. Farnia et al. [25] used a spectral regularization as the gradient penalty, which was combined with adversarial training to alleviate vulnerability. In addition, data augmentation [26,27] was a typical option used to enhance the generalization ability of neural networks and to reduce the risk of overfitting on training data. However, this option could not completely solve the problem of an adversarial attack which always generated new kinds of adversarial examples. As a top performer, the adversarial training (AT), which can be seen as data augmentation, achieved advanced robustness in different adversarial attack environments [18,28–30]. By using extra adversarial examples, it enabled the model to learn more generalizable feature representations. The AT mechanism accepted various losses and regularizers and was a powerful tool used to resist attacks. Despite this, AT might sacrifice the performance in the clean input and is computationally expensive [31]. Schmidt et al. [9] showed that the example complexity of robust learning might be much larger than standard learning.

### *2.3. Robust Loss Functions*

Many studies have been conducted to improve the widely used SCE loss function, most of which focused on encouraging a higher intra-class compactness and greater separation between classes. The comparing loss [13] and the triplet loss [14] were first proposed to improve the internal compactness of each class, which, however, suffered from the slowed training process and the unstable convergence. Center loss [15] was proposed to avoid the problem of a slow convergence and instability by minimizing the Euclidean distance between features and the corresponding class centers. However, the center loss was used together with the SCE loss to balance the inter-class dispersion and the intra-class compactness, which made it unable to obtain reliable robustness. After this, Liu et al. [32] converted the softmax loss to the cosine space, and proposed that the angular distance margin favoured high intra-class compactness and inter-class separability. Then,

Wan et al. [16] proposed the large-margin Gaussian mixture (L-GM) loss, which used the Gaussian mixture distribution to fit the training data and increased the distance between feature distributions of different classes. Recently, Pang et al. [17] proposed the max-Mahalanobis center (MMC) loss to induce dense feature regions, which encouraged the model to concentrate on learning ordered and compacted representations.

Different from the previous works, which improve the loss function to better fit the data distribution, the proposed method (i.e., I-SCE) is a much simple and interpretable way to enable the neural networks to learn freely. Moreover, we advocate that I-SCE encourages the models to be more generalizable with respect to the adversarial data instead of overfitting on the training data.

## 3. Methods

In this section, we introduce the inference-softmax cross entropy loss by first presenting the definition of an inference region, which motivated us to develop an inference schema.

Current neural networks tend to overfit on the by-hand clean training data, which, however, cannot work out a robust model and, instead, makes them vulnerable to adversarial attacks. We advocate that this scenario is caused by the misaligned distribution between the clean training data and the adversarial data, and overfitting prevents the model from being tolerant to input perturbations. The distribution difference is termed as the *inference region*, which characterizes why adversarial examples are outliers to the neural networks trained on clean data. Figure 1 is given as an illustration, where the grey circle region contains the features of the clean data $x$ and the orange circle region contains the features of the adversarial data $x + \delta$, and $\delta$ is the adversarial perturbation. When using SCE, the optimized decision boundary is located closely to the clean data area as shown in subfigure (a), whereas the expected boundary is around the adversarial data area as shown in subfigure (b). Considering the isotropic expansion of the space caused by adversarial perturbation, the inference region is then induced from the annular area. The features of adversarial examples reside outside the feature space of training examples, whereas the decision boundary specified by the well-trained model closely fits the training data area. Considering that the type of adversarial attack incrementally appears in real scenarios, the decision boundary in Figure 1a is not good enough to give the right prediction, even if several kinds of input perturbations are involved in training. Instead, adversarial attacks are assumed to result in an isotropic expansion of the feature space, where the expanded region is the inference region as shown in Figure 1b. By expanding the inference region, it is more difficult for adversarial attacks to cross the decision boundary. Then, the task is to encourage the model to be generalizable to this region in this study.

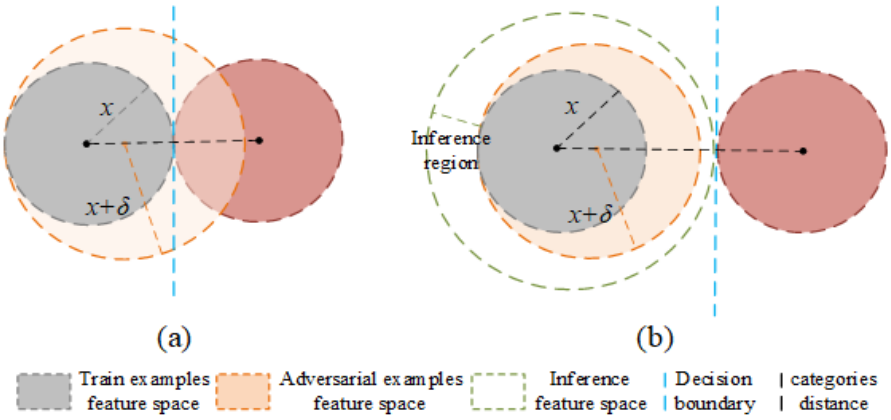


**Figure 1.** Illustration of the inference region. (**a**) is the decision boundary of SCE; (**b**) is the decision boundary using the inference region.

The softmax cross entropy (SCE) loss is a typical loss function used in training deep models, which imposes a hard constraint on the label of the input, i.e., regressing the

probability of 1 on the correct label and the probability of 0 on the incorrect labels (generally in the case of one-hot label representation). Unfortunately, the hard constraint, on the one hand, causes a difficult regression process in training. On the other hand, it makes the resultant model over-confident on the predictions, hence bringing the issue of vulnerability. This has already been mentioned in the literature of label smoothing [33–35], which solved the problem by designing a soft label or a soft output distribution, i.e., regressing the probabilities of $1-\epsilon$ and $\epsilon/K$ on the correct and incorrect labels, respectively, where $K$ is the number of task labels. As shown in Figure 2, (a) is the softmax cross entropy case, which regresses the probabilities from one side; (b) is the label smooth case, which regresses the soft labels from both sides; (c) is the ArcFace case, which regresses the targets on a circle axis in feature space, i.e., encouraging the circular margins between different classes; (d) is the inference softmax cross entropy case, which regresses the targets from all directions, i.e., encouraging the isotropic margins between different classes.

We give an intuitive explanation of the above discussion in Figure 2a,b. As shown, the SCE encourages the label regression from one side along the 0-1 axis, whereas the label smoothing drives the regression from both sides around the target probabilities.

In addition, we also identify that the margin-based idea in SCE is similar to the label smoothing. Specifically, the soft label implies a margin between the true distribution and the soft output distribution. Considering that softmax is a monotonically increasing function, a margin between the label distributions can induce a margin between features in the logit layer of a neural network, as in the ArcFace loss [36]. From Figure 2c, we can see that ArcFace pushes regression towards the target angles from both sides in a circle axis.

While the above analyses show that the regression is performed from either one side or both sides, here, we proposed an alternative definition of a soft label that could be regressed from arbitrary directions in feature space. Specifically, we freed the circle constraint in ArcFace and imposed the additive margin to the features only normalized by L2-norm. In this way, the resultant features are not necessarily located on a circle or a sphere, and, on the contrary, the margin is isotropically posed around each example in the feature space, as shown in Figure 2d. We will empirically demonstrate the effectiveness of this operation over the ArcFace. ArcFace loss normalizes the features and the weights such that the resultant features are located on a hyper-sphere, and the training process regresses the class targets along the surface of the hyper-sphere (as shown in Figure 2c). Similarly, Pang et al. [37] discuss the robust benefit of feature normalization and weight in a hyper-sphere. Different from it, I-SCE could be regressed from arbitrary directions in feature space (rather than a hyper-sphere), and normalize features using the L2-norm to ensure the effectiveness of added inference information (the max value of logit layer varies).

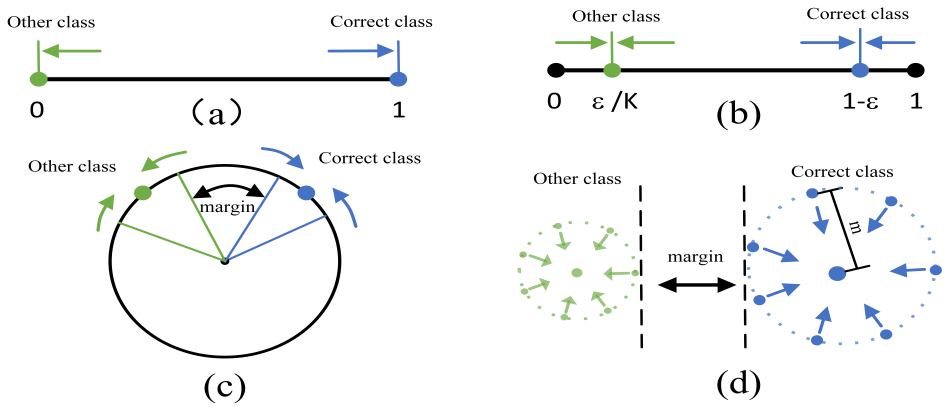


**Figure 2.** Intuitive explanation of label regression. (**a**) is the softmax cross entropy case; (**b**) is the label smooth case; (**c**) is the ArcFace case; (**d**) is the inference softmax cross entropy case.

### *3.1. Inference Region*

While the proposed method (I-SCE) could be viewed as a margin-based loss, the difference to the ArcFace loss is how the margin (or inference information) is applied to the logits. Instead, the proposed method only normalizes the features without normalizing the weights in order to locate the features in a free space, in which case, the regression process can be performed in any direction (as shown in Figure 2d). We advocate that freeing the sphere constraint will bring a performance improvement in adversarial defencing, which was demonstrated in the experiments. The reason for the effectiveness may be that the adversarial perturbation causes a large variation in the feature space. Constraining the features on a hyper-sphere would bring a large feature shift if the normalization direction (to the sphere) is undesirable. By contrast, the proposed method prefers isotropic tolerance to feature perturbations, and is hence better. By implementing this margin idea, the inference information is then contained in the margin, which could help (1) to avoid overfitting and (2) to improve the generalization ability of the feature representation, driving the decision boundary towards the boundary of the inference region in feature space. Hence, it is not easy for a small perturbation of an adversarial example to cross the decision boundary, greatly alleviating the problem of vulnerability. This schema is simple, interpretable, and effective, as demonstrated in the experiments. In the following, we present the inference-softmax cross entropy in detail.

### *3.2. Inference-Softmax Cross Entropy*

To derive a robust loss for neural network training, in this Section, we apply the inference-schema on SCE and propose an inference-softmax cross entropy (I-SCE) loss, which could encourage the tolerance of the model to adversarial perturbations, thus avoiding overfitting.

Given a $k$-class classification task, the posterior probability predicted by the deep model using softmax is

$$P(y' = i \| x) = \frac{e^{f_i(x)}}{\sum_j e^{f_j(x)}}, \tag{1}$$

where $i \in [1, k]$ is the label candidate and $f_i$ is the prediction function for the $i$-th class, which specifies both the backbone and the softmax layer in a typical classification network.

To improve the vulnerability of SCE, we imposed the inference information to the logits produced by the neural networks and proposed an inference softmax as

$$P_I(y' = i \| x) = \frac{e^{sf_i(x)+m}}{e^{sf_i(x)+m} + \sum_{j \neq i} e^{f_j(x)}}, \tag{2}$$

When $y' \neq i$,

$$P_I(y' = k \| x) = \frac{e^{f_k(x)}}{e^{sf_i(x)+m} + \sum_{j \neq i} e^{f_j(x)}}, \tag{3}$$

which then induces the inference-softmax cross entropy loss as

$$\text{I-SCE} = -\sum_{i=1}^{k} y_i \ln \frac{e^{y_i(sf_i(x)+m)+(1-y_i)f_i(x)}}{e^{y_i(sf_i(x)+m)+(1-y_i)f_i(x)} + \sum_{j \neq i} e^{f_j(x)}}, \tag{4}$$

where $y_i = 1$ if the ground truth label of $x$ is $i$ and otherwise 0, and $s \geq 1$ is used to scale the prediction $f_i(x)$ and control the gradient update rate on the right class. Note that we used $y_i$ as an indicator of the inference information, that is, $s$ and $m$ were only imposed on the right class instead of all classes. $s$ is a scaling factor that controls the gradient of the predicted class and $m$ is the inference region, which is explained in Section 3.1. As shown, the implementation of this loss is very easy: simply adding a scalar and a constant on the prediction of the right class, which is unaggressive to the original training code of neural networks.

In the implementation of I-SCE, we found that the case of $f_i \gg f_j, j \neq i$ possibly occurs, which reduces the effect of $m$. To address this issue, we normalized $f(x)$ by $L_2$ to increase the numerical stabilization. During the inference process, Equation (2) was calculated by firstly finding the index $i$ of the maximal value $f_i(x)$ among $i \in [1, k]$ and then applying $s$ and $m$ on the $i$-th class according to this equation. This operation does not change the class decision since $s \geq 1$ and $m > 0$.

### *3.3. Robustness Analyses of I-SCE*

#### 3.3.1. Expected Interval of Correct Class

To demonstrate the robustness of I-SCE, we analyzed the expected interval of the correct class predicted by both I-SCE and SCE. The bigger the expected interval is , the more adversarial perturbations are added to mislead neural networks. Here, assume the minimum perturbation $\delta$, which makes the model just misclassified. The probability that the SCE model recognizes the adversarial example $x + \delta$ as the correct label $i$ is

$$P(i|x+\delta) = \frac{e^{f_i(x+\delta)}}{\sum_j e^{f_j(x+\delta)}}. \tag{5}$$

Regarding the I-SCE model, the probability is then

$$P_I(i|x+\delta) = \frac{e^{sf_i(x+\delta)+m}}{e^{sf_i(x+\delta)+m} + \sum_{j\neq i} e^{f_j(x+\delta)}}. \tag{6}$$

The expected intervals of the correct class by using SCE are defined as

$$L = P(i|x) - P(i|x+\delta) = \frac{e^{f_i(x)}}{\sum_j e^{f_j(x)}} - \frac{e^{f_i(x+\delta)}}{\sum_j e^{f_j(x+\delta)}}, \tag{7}$$

The expected intervals of the correct class by using I-SCE are defined as

$$L_I = P_I(i|x) - P_I(i|x+\delta) = \frac{e^{sf_i(x)+m}}{e^{sf_i(x)+m} + \sum_{j\neq i} e^{f_j(x)}} - \frac{e^{sf_i(x+\delta)+m}}{e^{sf_i(x+\delta)+m} + \sum_{j\neq i} e^{f_j(x+\delta)}}. \tag{8}$$

The vulnerability of SCE to adversarial attacks states that $f(x+\delta) < f(x)$. Considering that the perturbation $\delta$ is a just value that misleads the SCE model, the expected interval measures the maximal level of perturbation that the model is robust on. The larger the interval, the more robust the model. Starting from this point, we show the following property of I-SCE:

*When* $s \geq 1$, $m > 0$, *and* $\frac{se^{sf_i(x)+m}}{(e^{sf_i(x)+m}+\sum_{j\neq i} e^{f_j(x)})^2} - \frac{e^{f_i(x)}}{(\sum_j e^{f_j(x)})^2} > 0$, $L_I$ *is larger than* $L$.

The condition in the above property is both theoretically demonstrated and empirically validated in Section 3.3.2. This states that the robustness of I-SCE is improved compared with SCE.

3.3.2. Proof of the Property on Expected Interval of Correct Class

According to the definition $L$ and $L_I$ in Equations (7) and (8), we can derive $L_I - L$; if $L_I - L > 0$, the expected intervals of the correct class of I-SCE are larger than SCE, and therefore I-SCE cannot break down easily under an adversarial attack.

$$\begin{aligned} L_I - L = & \frac{e^{sf_i(x)+m}}{e^{sf_i(x)+m} + \sum_{j\neq i} e^{f_j(x)}} - \frac{e^{sf_i(x+\delta)+m}}{e^{sf_i(x+\delta)+m} + \sum_{j\neq i} e^{f_j(x+\delta)}} \\ & - \frac{e^{f_i(x)}}{\sum_j e^{f_j(x)}} + \frac{e^{f_i(x+\delta)}}{\sum_j e^{f_j(x+\delta)}}. \end{aligned} \tag{9}$$

By defining $h(f(x)) = P_I(i|x) - P(i|x)$,

$$\begin{aligned} h(f(x)) &= P_I(i|x) - P(i|x) \\ &= \frac{e^{sf_i(x)+m}}{e^{sf_i(x)+m} + \sum_{j\neq i} e^{f_j(x)}} - \frac{e^{f_i(x)}}{\sum_j e^{f_j(x)}} \\ &= \frac{\sum_{j\neq i} e^{f_j(x)} (e^{sf_i(x)+m} - e^{f_i(x)})}{(e^{sf_i(x)+m} + \sum_{j\neq i} e^{f_j(x)}) \sum_j e^{f_j(x)}}. \end{aligned} \tag{10}$$

The above equation shows that $h(f(x)) > 0 \iff e^{sf_i(x)+m} - e^{f_i(x)} > 0 \iff \frac{e^{sf_i(x)+m}}{e^{f_i(x)}} > 1 \iff e^{(s-1)f_i(x)+m} > 1 \iff (s-1)f_i(x) + m > 0 \iff s > \frac{f_i(x)-m}{f_i(x)}$. Hence, when the parameters $s$ and $m$ satisfy $s \geq 1 > \frac{f_i(x)-m}{f_i(x)}$ and $m > 0$, $h(f(x)) > 0$. When $s = 1$ and $m = 0$, $P_I(i|x)$ degenerates to $P(i|x)$.

Similarly, regarding $h(f(x+\delta))$, we have

$$\begin{aligned} h(f(x+\delta)) &= P_I(i|x+\delta) - P(i|x+\delta) \\ &= \frac{e^{sf_i(x+\delta)+m}}{e^{sf_i(x+\delta)+m} + \sum_{j\neq i} e^{f_j(x+\delta)}} - \frac{e^{f_i(x+\delta)}}{\sum_j e^{f_j(x+\delta)}} \\ &= \frac{\sum_{j\neq i} e^{f_j(x+\delta)} (e^{sf_i(x+\delta)+m} - e^{f_i(x+\delta)})}{(e^{sf_i(x+\delta)+m} + \sum_{j\neq i} e^{f_j(x+\delta)}) \sum_j e^{f_j(x+\delta)}}. \end{aligned} \tag{11}$$

When $s \geq 1 > \frac{f_i(x+\delta)-m}{f_i(x+\delta)}$ and $m > 0$, $h(f(x+\delta)) > 0$.

Based on the above derivations, we calculated

$$L_I - L = h(f(x)) - h(f(x+\delta)). \tag{12}$$

To analyze the sign of the above equation, we computed the derivative of $h(f(x))$ with respect to $f_i(x)$ as

$$\begin{aligned} \frac{\partial h(f(x))}{\partial f_i(x)} &= \frac{\partial (\frac{e^{sf_i(x)+m}}{e^{sf_i(x)+m} + \sum_{j\neq i} e^{f_j(x)}} - \frac{e^{f_i(x)}}{\sum_j e^{f_j(x)}})}{\partial f_i(x)} \\ &= \sum_{j\neq i} e^{f_j(x)} (\frac{s e^{sf_i(x)+m}}{(e^{sf_i(x)+m} + \sum_{j\neq i} e^{f_j(x)})^2} - \frac{e^{f_i(x)}}{(\sum_j e^{f_j(x)})^2}). \end{aligned} \tag{13}$$

Considering that the perturbation $\delta$ is a just value that misleads the network, $f(x) > f(x+\delta)$. When $s$ and $m$ satisfy $\frac{s e^{sf_i(x)+m}}{(e^{sf_i(x)+m} + \sum_{j\neq i} e^{f_j(x)})^2} - \frac{e^{f_i(x)}}{(\sum_j e^{f_j(x)})^2} > 0$, $h(f(x))$ is monotonically increasing. This guarantees that $L_I - L > 0$.

However, the condition $\frac{se^{sf_i(x)+m}}{(e^{sf_i(x)+m}+\sum_{j\neq i}e^{f_j(x)})^2} - \frac{e^{f_i(x)}}{(\sum_j e^{f_j(x)})^2} > 0$ is not easy to validate in the case of $s \geq 1$ and $m > 0$. Here, we conducted an experiment to empirically demonstrate its validation. Specifically, we computed the empirical values $\sum_{j\neq i} e^{f_j(x)} \approx 8$, $f_i(x) \approx 0.97$ by averaging the corresponding values of all examples in MNIST and CIFAR10. By employing these two values, we plotted the 3D surface of $z = \frac{se^{sf_i(x)+m}}{(e^{sf_i(x)+m}+\sum_{j\neq i}e^{f_j(x)})^2} - \frac{e^{f_i(x)}}{(\sum_j e^{f_j(x)})^2}$ with respect to $s$ and $m$, which is shown in Figure 3. The surface indicates that $z$ is always larger than 0 when $s > 1$ and $m > 0$. This empirically demonstrates the validity of the conditions in the property.

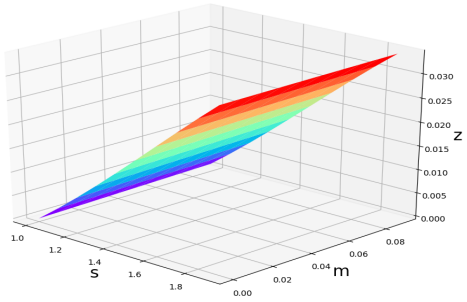


**Figure 3.** Parameter selection for $s$ and $m$.

#### 3.3.3. Min-Max Framework

The above robustness conclusion is also applicable to the min-max framework [18], which is a typical framework of adversarial attack and defense. The min-max framework is formulated as

$$\min_{\theta} \rho(\theta), \quad where \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim D}\left[\max_{\delta\in S} \ell(\theta, x+\delta, y)\right], \tag{14}$$

where $\theta$ is the model parameter and $\delta$ is the input perturbation. The internal maximization is an attack process that finds the perturbation that maximally misleads the model $\theta$. The external minimization is a defense process that encourages the model to be tolerant to such an attack. We used $\rho_I$ and $\rho$ to represent the objective losses by using I-SCE and SCE, respectively. Given an input perturbation $\delta$ and a trained model $\{f_i\}$, we had, *when* $s \geq 1$ *and* $m > 0$, $P_I(i|x+\delta) > P(i|x+\delta)$, which is proven in Equation (10). This states that the $P_I$ results in a lower loss than $P$, i.e., $\rho_I < \rho$. Hence, the lower loss indicates the better defense performance on adversarial attacks, which demonstrates the improved robustness of I-SCE.

## 4. Experiments

In this Section, we conducted a series of experiments on MNIST [38] and CIFAR-10 [39] to demonstrate the effectiveness of the proposed I-SCE. The MNIST and CIFAR-10 are classic data sets in image classification task. The MNIST is a handwritten digital dataset of 0–9, and the size of each picture is $28 \times 28 \times 1$. The CIFAR-10 has 10 classes, and the size of each picture is $32 \times 32 \times 3$. The backbone used in our implementation was ResNet-32, with five stages [40], which was optimized by using the Adam algorithm [41]. We employed the white-box attack and the black-box attack, including the targeted and untargeted PGD [18], Deepfool [19], and SimBA [20]. The white-box attack means that the attacker can obtain the network structure and weight of the model, whereas the black-box attack means that the attacker does not know the network structure and weight of the model. The untargeted attack means that the attacker only makes the neural network misclassify, whereas the targeted attack makes the neural network with a specific wrong class. In this paper, the specific wrong class of the targeted attack is the class with the lowest prediction probability of the model. We selected the state-of-the-art models as competitors, such as the center

loss [15], the large-margin Gaussian mixture (L-GM) loss [16], ArcFace loss [36], the max-Mahalanobis center (MMC) loss [17], the random method [22], label smoothing [33], and the adversarial training (AT) method [18]. Extensive experiments show that the model trained by I-SCE is more robust compared with competitors.

### *4.1. Parameter Setting*

I-SCE was directly used as the loss function to train the neural network model. There were two hyper-parameters $s$ and $m$ in the proposed I-SCE, which affects the defense performance. We set the ranges as $s \in [1, 2]$ and $m \in (0, 0.1]$, and densely evaluated the performance of I-SCE under different settings and different attacks. Figure 4 illustrates the results, from which, we can see that the performance is highly correlated with the settings, the attack types, and the datasets. Therefore, to obtain better robustness, the parameters needed to be reset in different tasks by using a small validation set. In the following experiments, to make a fair comparison, we set $s = 1$ and $m = 0.1$.

Through a detailed analysis, using the inference information can significantly improve the robustness of neural networks. Intuitively, the bigger the inference interval, the better the robustness. However, the inference interval also influences the accuracy, and the inference interval needs to balance robustness and accuracy. Through our research, if hyper-parameter $s > 1$, it will increase the risk of gradient exposure and will lead to a model susceptible to adversarial attacks, specially based on a gradient attack, such as a PGD attack. Some works have reduced the amplitude of the gradient to improve robustness, such as the constraint of the Lipschitz constant and distillation temperature. If $s < 1$, it may cause a reduction in the inference interval. Therefore, we set $s = 1$.

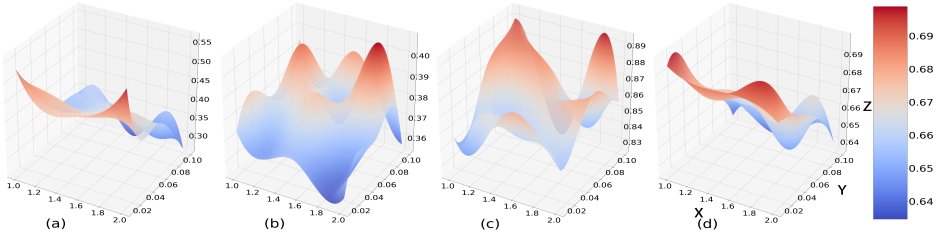


**Figure 4.** Performance of I-SCE under different parameter settings. The x-axis is $s$, the y-axis is $m$, and the z-axis is the accuracy. (**a**) Deepfool attack on MNIST. (**b**) Deepfool attack on CIFAR10. (**c**) Untargeted PGD attack on MNIST. (**d**) Untargeted PGD attack on CIFAR10.

### *4.2. Comparison with State-of-the-Arts*

**PGD attack:**

The PGD attack is a strong white-box untargeted and targeted attack. We used $L_2$ constrained untargeted and targeted PGD attacks for comparison. The results are listed in Figures 5 and 6. The Acc is the accuracy on clean examples, $\epsilon$ is the perturbation level, and $\text{PGD}^{tar,un}_{10,50}$ represents the targeted or untargeted attacks with 10 or 50 iterations. The results indicate that I-SCE produces a better performance than the others in most cases. In particular, under a $PGD^{un}_{50}$ attack, the accuracy of I-SCE is 63.13% higher than SCE on the MNIST dataset. The accuracy of I-SCE is 60.31% higher than center loss. The accuracy of I-SCE is 61.06% higher than L-GM loss. The accuracy of I-SCE is 11.35% higher than center loss. The accuracy of I-SCE is 17.41% higher than MMC loss. While AT sometimes achieves a good performance, it has a noticeable sacrifice of accuracy on clean examples, e.g., on CIFAR10, and it has a weaker defense against strong PGD attacks than I-SCE. In constrast, I-SCE preserves a high performance on clean data. Under several attack cases, e.g., $\epsilon = 0.04$, I-SCE performs better than the others and is comparable with MMC and AT. In particular, considering the accuracy of AT in clean examples, we reduced the adversarial disturbance of AT to improve the accuracy; for example, as shown in Tables 1 and 2, the adversarial disturbance of AT is $\epsilon = 0.02$ with 10 iterations. Although AT is an effective method used to improve robustness, it can only learn the existing distribution of adversarial examples.

When the adversarial disturbance of a PGD attack exceeds the adversarial disturbance of AT, we can find that AT does not have a good defensive effect on the PGD attack and AT does not learn the distribution of higher adversarial disturbances with higher iterations. In particular, we find that increasing the adversarial disturbance does not always improve the robustness during adversarial training. As shown in Figure 7, the X axis represents the perturbation level during adversarial training and the Y axis represents the accuracy of classification. The red curve is the classification accuracy of clean examples. The blue curve is the PGD attack with 0.03/10 (0.03 is the perturbation level of the adversarial attack, 10 is the iterations of the adversarial attack). Similarly, the green curve, sky-blue curve, and yellow curve represent 0.03/30, 0.06/10, and 0.06/30. Two point coordinates represents the maximum values of the green curve and yellow curve. When the robustness reaches the peak, the robustness of adversarial training decreased, with an increase in the perturbation level during adversarial training with ResNet-32. It shows that adversarial training cannot increase robustness without limitation via sacrificing accuracy. In addition, the two point coordinates of the curve's maximum values indicate that the model cannot achieve the best robustness under different attacks. In general, adversarial training has an upper bound for robustness.

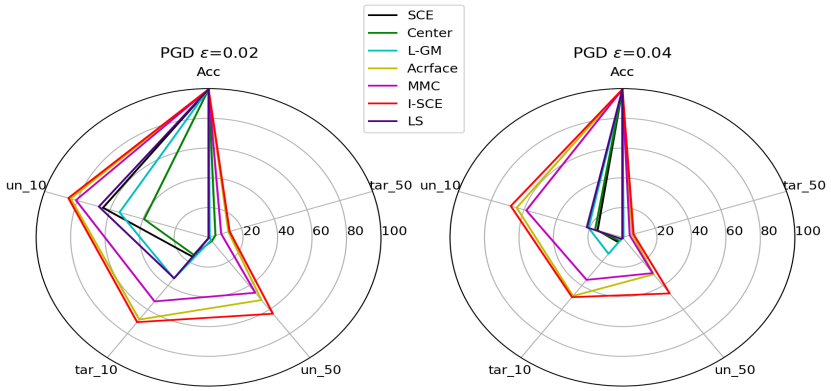


**Figure 5.** Classification accuracy (%) under PGD attack on MNIST.

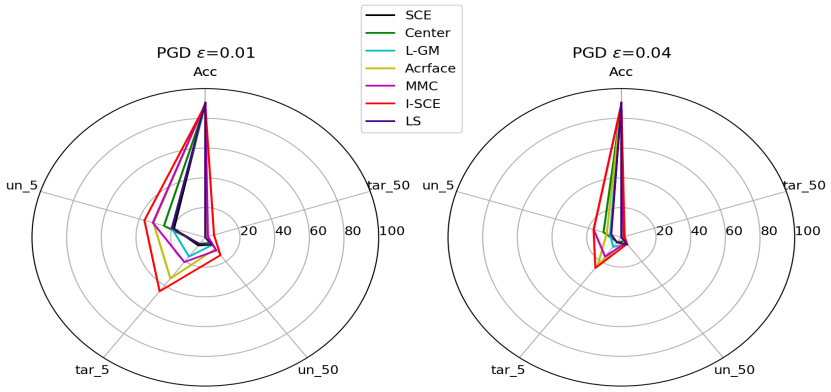


**Figure 6.** Classification accuracy (%) under PGD attack on CIFAR10.

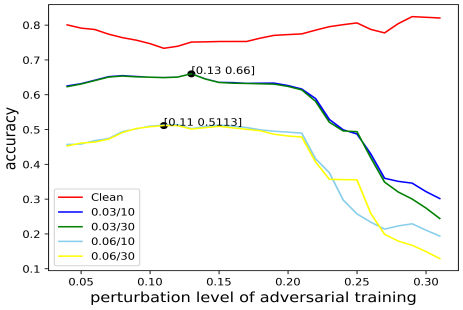


**Figure 7.** The robust evaluation of adversarial training.

**Table 1.** Adversarial training with PGD attack on MNIST.

| | | $\mathbf{PGD}_{50}^{un}$ | | $\mathbf{PGD}_{50}^{tar}$ | |
|---|---|---|---|---|---|
| **Method** | **Clean** | $\epsilon = 0.02$ | $\epsilon = 0.04$ | $\epsilon = 0.02$ | $\epsilon = 0.04$ |
| I-SCE | **99.57** | **63.47** | **46.41** | **12.74** | **6.78** |
| AT | 99.25 | 13.62 | 0.74 | 6.73 | 0.01 |

**Table 2.** Adversarial training with PGD attack on CIFAR10.

| | | $\mathbf{PGD}_{50}^{un}$ | | $\mathbf{PGD}_{50}^{tar}$ | |
|---|---|---|---|---|---|
| **Method** | **Clean** | $\epsilon = 0.01$ | $\epsilon = 0.04$ | $\epsilon = 0.01$ | $\epsilon = 0.04$ |
| I-SCE | **89.09** | **14.82** | 5.51 | **5.26** | **1.95** |
| AT | 83.48 | 7.87 | **7.15** | 0.08 | 0.07 |

**Deepfool attack:**

The Deepfool attack generates minimal input perturbations to mislead the neural Networks. Here, we used the $L_2$ constrained Deepfool attack on MNIST and CIFAR10. From the results in Figure 8, it is clearly observed that I-SCE produces a much higher performance than all competitors, which have a very limited defense ability against Deepfool. The improvement in I-SCE is above 50% in most cases, which is significant and exciting. In real applications, the minimal disturbance generated by Deepfool is more usual than the strong offensive disturbance generated by PGD. Therefore, the results indicate that I-SCE is more suitable and can achieve a better performance in real scenarios than the other methods.

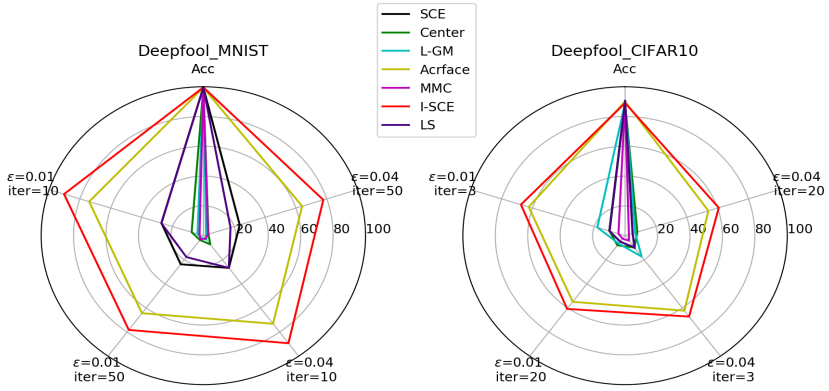


**Figure 8.** Classification accuracy (%) under Deepfool attack.

**Black-box attack:**

A robust performance is critical to claiming reliable robustness against the black-box attacks. SimBA [20] is a black-box query-based attack, which was employed here. We set the frequency of the query to 300 times per image on MNIST and 500 times per image on

CIFAR10. The results under different disturbance levels are shown in Table 3, from which, we can see that I-SCE has a higher accuracy and a lower sacrifice of accuracy compared with the others. This evidence indicates that I-SCE can induce a reliable robustness rather than the false one caused by, e.g., the gradient mask [42].

**Table 3.** Classification accuracy (%) under SimBA attack.

| | MNIST | | | | CIFAR10 | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Clean** | $\epsilon = 0.5$ | $\epsilon = 1$ | $\epsilon = 1.5$ | **Clean** | $\epsilon = 0.5$ | $\epsilon = 1$ | $\epsilon = 1.5$ |
| SCE | 99.33 | 97.00 | 91.23 | 88.43 | **90.17** | 77.60 | 72.77 | 70.17 |
| Center | 99.25 | 96.20 | 91.04 | 88.17 | 89.27 | 81.80 | 77.94 | 76.48 |
| MMC | 99.18 | 98.38 | 97.42 | 96.15 | 89.77 | 82.57 | 77.83 | 76.17 |
| I-SCE | **99.40** | **98.73** | **97.90** | **97.10** | 89.63 | **85.93** | **83.33** | **82.07** |
| Random | 98.90 | 94.84 | 92.32 | 86.12 | 89.48 | 76.54 | 72.74 | 71.30 |
| AT | 98.86 | 97.32 | 66.98 | 49.94 | 83.46 | 80.14 | 75.40 | 72.22 |

**Feature embedding:** To visually investigate the effect of I-SCE, we computed a three-dimensional (3D) representation of the input by adding a three-dimensional embedding layer before the output layer. The embedded points are plotted in Figure 9, where the examples are selected from the test set of MNIST and CIFAR10 without any perturbation. The embedded points are composed of 5000 examples. If the embedded points look fewer in Figure 9, they can indicate that embedding points are coincident and the class centers are more concentrated. As seen, the distribution examples of SCE are confused in the space, where little perturbations in the examples could change the category decision. The center loss adds a penalty term for the class center together with the SCE loss, and we can clearly see the center of the class. However, center loss still has many embedded points far from the center of the class. L-GM loss uses Gaussian mixture distribution to fit the data distribution, and embedded points also fit very well with a Gaussian distribution. MMC loss induces dense feature regions to improve robustness. L-GM loss and MMC loss still have some embedded points close to other centers of the class, which means that adversarial examples are more likely to cross the decision boundary. In contrast, I-SCE produces separable clusters for each class with large margins among them, and, hence, has a higher tolerance to the perturbations than the other competitors.

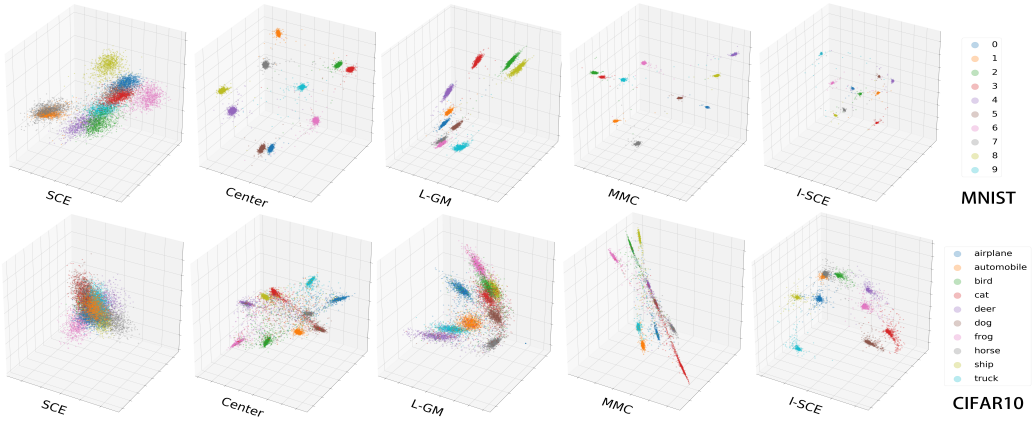


**Figure 9.** Illustration of three-dimensional feature embedding.

### *4.3. Experiments on Shallow Networks*

In real applications on mobiles, shallow networks are generally preferred because of low computational costs. Hence, in this section, we evaluated the robustness of the proposed I-SCE with shallow networks. Specifically, we followed the same settings of

competitors and attack methods as in Section 4. The backbone network was LeNet-5 [38] for MNIST and an eight-layer neural network for CIFAR10.

Figure 10 and Figure 11 illustrate the performance under the PGD attack on MNIST and CIFAR10, respectively. The results indicate that I-SCE performs surprisingly well in all cases of attack, while a slight sacrifice of accuracy on the clean data remains. Notably, the performance gaps between I-SCE and the others are above 50% in many cases, which validates the effectiveness of the proposed schema. More importantly, under severe attacks, I-SCE still shows strong robustness. Significantly, under the $PGD_{50}^{un}$ attack, the accuracy of I-SCE is 97% higher than SCE on the MNIST dataset and 79% higher than SCE on the CIFAR-10 dataset. The I-SCE is above a 48% accuracy compared with other methods on the MNIST dataset and a 55% accuracy compared with other methods on the CIFAR-10 dataset.

Figure 12 lists the results of all methods under the Deepfool attack. We find that the performance of I-SCE is comparable with the state-of-the-arts. MMC produces the best accuracy under attacks, but has a noticeable sacrifice in accuracy on clean data. By contrast, I-SCE shows a better trade-off between accuracy and robustness.

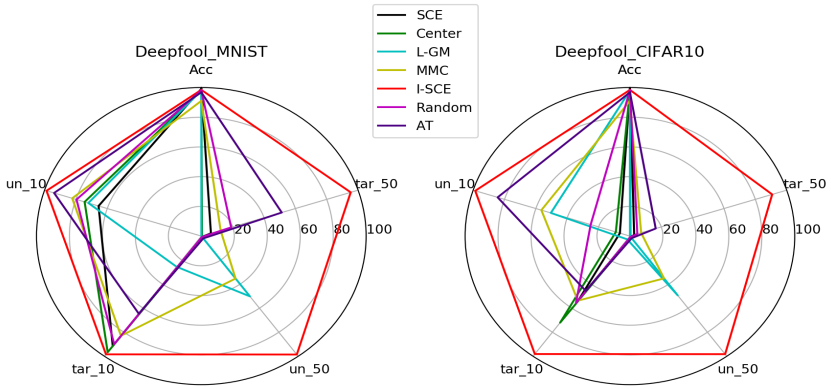


**Figure 10.** Performance (%) of shallow neural networks under PGD attack on MNIST.

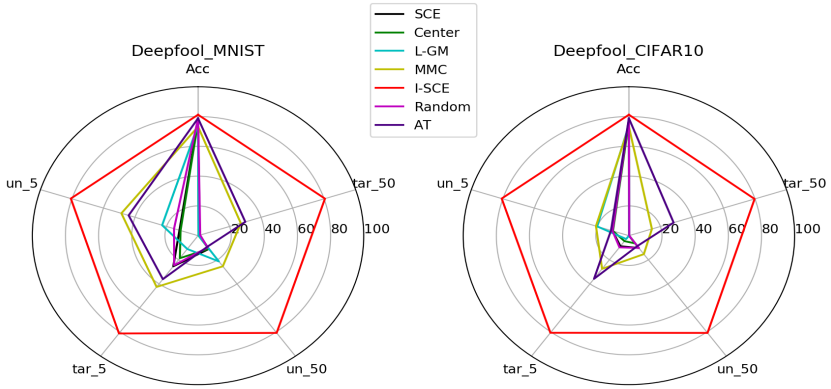


**Figure 11.** Performance (%) of shallow neural networks under PGD attack on CIFAR10.

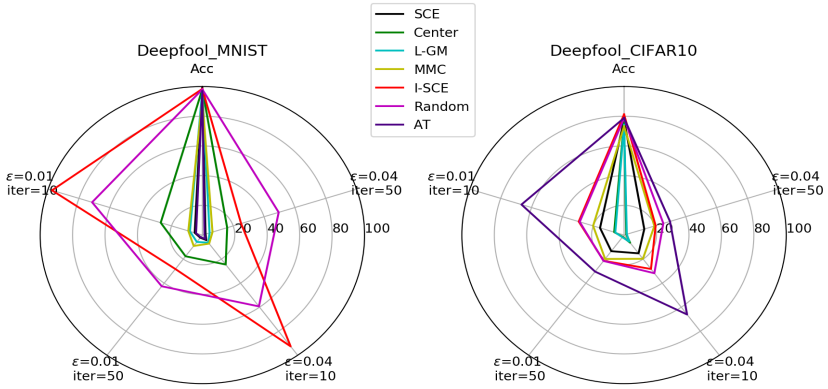


**Figure 12.** Performance (%) of shallow neural networks under Deepfool attack.

## 5. Discussion

In image classification tasks, because the loss function of DNN aims to overfit the training examples, DNN is vulnerable to out-of-distribution attacks (e.g., adversarial examples) far from the training data. Therefore, it is essential to improve the loss function to guarantee the neural network's generalization to out-of-distribution, and related work is lacking. In an attempt to fill this knowledge gap, we advocate that this scenario is caused by the misaligned distribution between the clean training data and the adversarial data. Overfitting training data prevents the model from being tolerant to input perturbations.

The misaligned distribution between the training and adversarial data is a blind region to the model, which we regard as the inference region. We exploit the inference region, which inspires us to involve margin-like inference information to SCE, resulting in a novel inference-softmax cross entropy (I-SCE) loss. The inference information guarantees the generalization of the neural network to out-of-distribution. Furthermore, I-SCE ensures inter-class separability and that it is difficult for the adversarial example to cross the decision boundary. The softmax cross entropy (SCE) loss is a typical loss function used in training deep models that generally imposes a hard constraint on the label of the input, regressing the probability of 1 on the correct label and the probability of 0 on the incorrect labels, where the hard constraint makes the resultant model over-confident on the predictions, bringing the issue of vulnerability. Several instances of literature have proposed promoting the effectiveness of the softmax loss, such as comparing loss [13], triplet loss [14], center loss [15], large-margin Gaussian mixture (L-GM) loss [16], and max-Mahalanobis center (MMC) loss [17]. These methods are still led by the same principle: minimizing the losses to fit the training examples maximally. These methods make it difficult for adversarial attacks to find the minimum perturbation, making it harder for adversarial examples to cross the decision boundary. These methods ignore the problem of distributional differences. Unlike other methods, our proposed ISCE loss function for distribution differences avoids model overfitting and enlarges the training set distribution. It also has good robustness under adversarial attacks.

In the Methods section, we detailed how adversarial attacks faced by neural network models cross the decision boundary, the design ideas of existing loss functions, and the loss ideas that we propose. In addition, we conducted a theoretical analysis of our method. I-SCE generalizes the clean example distribution to the adversarial example distribution, which solves the misaligned distribution to a certain extent. Compared with SCE, I-SCE has a higher correct expectation in adversarial attacks. We further demonstrated the robustness of I-SCE under the min-max framework.

In the experimental part, we used different adversarial attack methods to verify the effectiveness of our method, including a white-box attack and black-box attack; under severe adversarial attacks, I-SCE still maintains a high accuracy and robustness and performs better. On the MNIST dataset, the accuracy of I-SCE is 63% higher than SCE under the $PGD_{50}^{un}$ attack of ResNet-32. On the CIFAR-10 dataset, the accuracy of I-SCE is 38% higher

than SCE under the $PGD_5^{tar}$ attack. Significantly, under the $PGD_{50}^{un}$ attack of LeNet-5, the accuracy of I-SCE is 97% higher than SCE on the MNIST dataset and 79% higher than SCE on the CIFAR-10 dataset. The I-SCE is above a 48% accuracy compared with other methods on the MNIST dataset and 55% accuracy compared with other methods on the CIFAR-10 dataset. In addition, we also verified the feature space embedding of different loss functions. Figure X shows that our method has a better inter-class separability; it is more difficult for adversarial examples to cross the decision boundary.

We propose the inference region to solve the misaligned distribution problem for robust loss function. Furthermore, we recommend that data-driven neural networks avoid overfitting training examples. More optimization objectives (loss functions) need to be designed to learn from existing distributions.

## 6. Conclusions

The original SCE loss induces the model to fit the distribution of the clean data, which is shown as being vulnerable to adversarial attacks. We advocate that the vulnerability is caused by the unawareness of the inference region during learning. Targeting this issue, we proposed an I-SCE loss that avoids overfitting by imposing an additive inference information on the output of the neural network such that the sensitive class region of the model is expanded. In this way, the model has a higher generalization to the adversarial examples. Extensive experiments demonstrated the superiority of I-SCE compared with the state-of-the-arts. Especially in the case of strong attacks, I-SCE still remains highly robust. I-SCE is only currently suitable for supervision training on image classification, and other tasks need to be confirmed. Despite the limitations, these are valuable in light of image classification. We believe that the focus of future research on the robustness loss function is to avoid overfitting the training data and to learn the out-of-distributed examples for generalization. Our analyses in this paper also provide valuable insights for future work on designing new objectives beyond the SCE framework. It also provides design guidance for solving the issue of misaligned distribution in other fields.

**Author Contributions:** Conceptualization, B.S. and R.W.; methodology, B.S.; validation, B.S. and W.H.; formal analysis, W.Z.; investigation, B.S.; writing—original draft preparation, B.S.; writing—review and editing, R.W. and W.Z.; visualization, W.H.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported in part by the National Natural Science Foundation of China under Grant 62162067 and 62101480, in part by the Yunnan Province Science Foundation under Grant No.202005AC160007 and No.202001BB050076.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare that they have no known competing financial interests or personal relationships that could appear to influence the work reported in this paper.

## References

1. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18-22 June 2018; pp. 7132–7141.
2. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021, 2021; pp. 9992–10002.
3. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7-9 May 2015, Conference Track Proceedings.
4. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14-16 April 2014, Conference Track Proceedings.

5. Li, L.; Huang, Y.; Wu, J.; Gu, K.; Fang, Y. Predicting the Quality of View Synthesis With Color-Depth Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2509–2521.
6. Carlini, N.; Wagner, D.A. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In Proceedings of the 2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, 24 May 2018; pp. 1–7.
7. Liu, A.; Liu, X.; Fan, J.; Ma, Y.; Zhang, A.; Xie, H.; Tao, D. Perceptual-Sensitive GAN for Generating Adversarial Patches. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019; pp. 1028–1035.
8. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial Examples Are Not Bugs, They Are Features. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8-14 December 2019; pp. 125–136.
9. Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; Madry, A. Adversarially Robust Generalization Requires More Data. In Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada, 3-8 December 2018; pp. 5019–5031.
10. Alayrac, J.; Uesato, J.; Huang, P.; Fawzi, A.; Stanforth, R.; Kohli, P. Are Labels Required for Improving Adversarial Robustness? In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8-14 December 2019; pp. 12192–12202.
11. Carmon, Y.; Raghunathan, A.; Schmidt, L.; Duchi, J.C.; Liang, P. Unlabeled Data Improves Adversarial Robustness. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8-14 December 2019; pp. 11190–11201.
12. Ganeshan, A.; S., V.B.; Radhakrishnan, V.B. FDA: Feature Disruptive Attack. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, October 27 - November 2, 2019; pp. 8068–8078.
13. Sun, Y.; Chen, Y.; Wang, X.; Tang, X. Deep Learning Face Representation by Joint Identification-Verification. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, 8-13 December 2014; pp. 1988–1996.
14. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7-12 June 2015; pp. 815–823.
15. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In Proceedings of the Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, 11-14 October 2016; Proceedings, Part VII; pp. 499–515.
16. Wan, W.; Zhong, Y.; Li, T.; Chen, J. Rethinking Feature Distribution for Loss Functions in Image Classification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18-22 June 2018; pp. 9117–9126.
17. Pang, T.; Xu, K.; Dong, Y.; Du, C.; Chen, N.; Zhu, J. Rethinking Softmax Cross-Entropy Loss for Adversarial Robustness. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26-30 April 2020.
18. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018; Conference Track Proceedings.
19. Moosavi-Dezfooli, S.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27-30 June 2016; pp. 2574–2582.
20. Guo, C.; Gardner, J.R.; You, Y.; Wilson, A.G.; Weinberger, K.Q. Simple Black-box Adversarial Attacks. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, 9-15 June 2019; pp. 2484–2493.
21. Cohen, G.; Sapiro, G.; Giryes, R. Detecting Adversarial Samples Using Influence Functions and Nearest Neighbors. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13-19 June 2020; pp. 14441–14450.
22. Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A.L. Mitigating Adversarial Effects Through Randomization. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.
23. Ross, A.S.; Doshi-Velez, F. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, 2-7 February 2018; pp. 1660–1669.
24. Yan, Z.; Guo, Y.; Zhang, C. Deep Defense: Training DNNs with Improved Adversarial Robustness. In Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada, 3-8 December 2018; pp. 417–426.
25. Farnia, F.; Zhang, J.M.; Tse, D. Generalizable Adversarial Training via Spectral Normalization. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6-9 May 2019.

26. Hendrycks, D.; Mu, N.; Cubuk, E.D.; Zoph, B.; Gilmer, J.; Lakshminarayanan, B. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In Proceedings of the8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26-30 April 2020.
27. Zhang, H.; Cissé, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018; Conference Track Proceedings.
28. Miyato, T.; Dai, A.M.; Goodfellow, I.J. Adversarial Training Methods for Semi-Supervised Text Classification. In Proceedings of the5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24-26 April 2017, Conference Track Proceedings.
29. Shafahi, A.; Najibi, M.; Ghiasi, A.; Xu, Z.; Dickerson, J.P.; Studer, C.; Davis, L.S.; Taylor, G.; Goldstein, T. Adversarial training for free! In Proceedings of theAdvances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8-14 December 2019; pp. 3353–3364.
30. Najafi, A.; Maeda, S.; Koyama, M.; Miyato, T. Robustness to Adversarial Perturbations in Learning from Incomplete Data. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8-14 December 2019; pp. 5542–5552.
31. Xie, C.; Wu, Y.; van der Maaten, L.; Yuille, A.L.; He, K. Feature Denoising for Improving Adversarial Robustness. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16-20 June 2019; pp. 501–509.
32. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-Margin Softmax Loss for Convolutional Neural Networks. In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, 19-24 June 2016; pp. 507–516.
33. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8-14 December 2019; pp. 4696–4705.
34. Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, L.; Hinton, G.E. Regularizing Neural Networks by Penalizing Confident Output Distributions. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24-26 April 2017, Workshop Track Proceedings.
35. Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.V.K.V.; Wang, J. Confidence Regularized Self-Training. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, October 27 - November 2, 2019; pp. 5981–5990.
36. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16-20 June 2019; pp. 4690–4699.
37. Pang, T.; Yang, X.; Dong, Y.; Xu, T.; Zhu, J.; Su, H. Boosting Adversarial Training with Hypersphere Embedding. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, virtual, 6-12 December 2020.
38. Lecun, Y.; Bottou, L. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
39. Krizhevsky, A.; Hinton, G. Learning multiple layers of features from tiny images. Computer Science Department, University of Toronto, Tech. Rep 2009.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV; Volume 9908, pp. 630–645.
41. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7-9 May 2015, Conference Track Proceedings.
42. Athalye, A.; Carlini, N.; Wagner, D.A. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10-15 July 2018; Volume 80, pp. 274–283.