E-MART: An Improved Misclassification Aware Adversarial Training with Entropy-Based Uncertainty Measure

Songcao Hou[1,a,*], Tianying Cui[1,b]

[1]School of Modern Information Industry, Guangzhou College of Commerce, Guangzhou, China
[a]housc77@163.com, [b]now_work@126.com
**Corresponding author*

Keywords: Adversarial training, Misclassified example, Entropy-based uncertainty measure

Abstract: Recently, adversarial training (AT) has been demonstrated to be effective to improve deep neural network (DNN) robustness against adversarial examples. Among them, Misclassification Aware adveRsarial Training (MART) is the most promising one, which incorporates an explicit differentiation of misclassified examples as a regularizer. However, MART uses prediction error for identifying the misclassified examples, and yet it fails to achieve the greatest performance. This crux lies in the fact that the prediction error only focuses on the output probability regarding with the ground-truth label, neglecting the impact of the complement classes. In this paper, we offer a unique insight into the condition that emphasizes learning on misclassified examples, and propose an improved MART method with entropy-based uncertainty measure (termed as E-MART). Specifically, we consider the impact of the outputs from all classes and develop an entropy-based uncertainty measure (EUM) to provide reliable evidence that indicates the impact of the misclassified and correctly classified examples. Moreover, based on EUM, we conduct a soft decision scheme to optimize the loss function of AT, which help to make the efficient training for the model's final robustness. We have carried out experiments on CIFAR-10 dataset, and the experimental results demonstrate the effectiveness of our method.

1. Introduction

Although deep neural networks (DNNs) have shown impressive performance in numerous classification tasks, e.g., computer vision[1], speech recognition [2] and natural language processing [3, 4], they still suffer from some fatal threats such as the attack of adversarial example (AE)[5]. AE is carefully crafted by adding small adversarial perturbation to the natural instance [6, 7, 8, 9], leading to the false predictions by the target DNN model. Thus improving the model's robustness against AEs has played an important role in the practical applications of DNNs.

Over the past decade, a train of technologies have been proposed to improve the robustness of DNNs. The representative methods include adversarial training(AT)[10, 11, 12], input denoising [13, 14, 15], feature squeezing [16], adversarial detection [17, 18], defensive distillation [19, 20], gradient regularization [21, 22], among which AT has been shown to be the most effective. AT can be regarded as a data augmentation technique that trains DNNs under AEs by solving a min-max optimization problem, where the inner maximum loop aims to derive the adversarial perturbation to fool the classifier while the outer minimization loop is used to train robust DNNs[12, 23]. Note that the AEs can be obtained by adding targeted or untargeted adversarial perturbations to the natural data in the inner maximization [15, 19, 24].

Compared with the conventional training method under the natural examples, AT is particularly difficult. To this end, many works are developed for the improvement of AT. For instance, to reduce the difference of the predictions between the natural examples and AEs, Zhang et.al. proposed TRADE [25] method, wherein robust error is decomposed into natural error and boundary error. Zhang et al. proposed a friendly adversarial training method (FAT) [26] method, where friendly adversarial data is used to minimize the loss for finding the misclassified adversarial data. These methods overlook a fact that the formal definition of an AE is conditioned on it being correctly classified and the AEs generated from misclassified examples are "undefined". Hence, they treat all training examples equally in both the maximization and the minimization processes, regardless of whether or not they are correctly classified. In this regard, it is reasonable to pay much emphasis on the distinction of correctly classified and misclassified examples when adversarially training the robust DNNs. The pioneering work for this intuition is Max-Margin Adversarial (MMA) presented by Ding et.al. [27], they proposed to use maximum margin optimization for correctly classified examples while keeping the optimization for misclassified examples unchanged. Yet, MMA did not pay sufficient attention to misclassified examples and the improvements are limited. The most influential work of AT with misclassified examples is Misclassification Aware adveRsarial Training (MART) proposed by Wang et.al. [28]. Inspired by the observation that the misclassified examples have a significant impact on the final robustness, this method explicitly differentiates the misclassified and correctly classified examples during the training. That is, the AEs generated by the misclassified examples are regarded as more critical data and are shared with greater weight in AT. MART and its variant could significantly improve the state-of-the-art adversarial robustness.

Despite its great success, MART is still fails to achieve the greatest performance in the practical application. It is due to the fact that MART only uses the prediction error regarding with the ground truth label to identify the misclassified and correctly classified examples, which neglects the impacts from the other classes. To address this issue, we give an exceptional perception regarding with learning on misclassified examples and proposed an improved misclassification aware adversarial training with entropy-based uncertainty measure termed as E- MART. E-MART uses entropy-based uncertainty measure (EUM) to evaluate the misclassified examples, and then applies a soft decision schema to weight the impacts of the misclassified and correctly classified examples during AT. Compared to MART, our method achieves a more reliable distinguishment between the misclassified and correctly classified examples and is able to further improve the model's final robustness.

2. Related Work

Adversarial Attack Methods. Fast gradient sign method (FGSM) is a single-step white-box attack which is proposed by Goodfellow et.al. [11]. It generates the AE by adding a small perturbation in original example in the direction of the sign of the loss gradient. As an extension of FGSM, Kurakin et al. [29] proposed the Basic Iterative Method(BIM). This method lies in the fact that applying FGSM iteratively over the examples by using smaller step sizes would result in stronger AEs. At each iteration, the output is clipped to ensure that the AE lies within the ϵ-neighborhood of the original input. A further improvement of BIM is MI-FGSM [30], which is motivated by the conception of momentum gradient dependence. C&W attack [24] is an optimization-based attack method, which formultae an objective function to craft AEs to fool the models. Recently, the iterative Projected Gradient Descent (PGD) [11] has been empirically determined to be the most effective method for performing norm constrained attacks, which reasonably approximates the optimal attack.

Robustness Boosting Method. The most effective approach to enhance the robustness of the DNN is AT [7]. AT modeling the training process by formulating a mini-max game between attacker and defender, where the goal of the attacker is to generate powerful AE in the inner maximization procedure, while the defender aims to minimize the impact of the AE in the outer minimization loop. More recently, a body of work has emerged that also improves robustness with adversarial training examples. Zhang et al. [25] suggested that the robust error can be decomposed into the sum of natural (classification) error and boundary error to describe the trade-off between accuracy and robustness of the classification problem and proposed a new defense method, TRADES, as optimizing a regularized surrogate loss. Ding et al.[27] studied the adversarial robustness of neural networks from the perspective of margin maximization (margin is defined as the distance from the input to the classifier's decision boundary) and proposed Max-Margin Adversarial (MMA) training to directly maximize the margin to achieve adversarial robustness. Zhang et al. proposed a friendly adversarial training method (FAT) [26], which used friendly adversarial data to optimize the loss based on the misclassified adversarial data. Wang et al. [28] found that the misclassified examples have a significant impact on the final robustness and proposed a AT approach, MART, which explicitly differentiates the misclassified and correctly classified examples during the training. Carmon et al. [31] suggested using semi-supervised learning with unlabeled data to further improve robustness. Wong et al. [32] offered a way to train a robust model with a lower computational cost with weak adversaries.

3. Preliminary

For a *K* (*K*≥2) class classification problem, given a dataset of natural examples $S=\{x_i\}_{i=1}^{N}$, $x_i \epsilon \mathbb{R}^d$, along with labels $\{y_i\}_{i=1}^{N}$, let $h_\theta : \mathbb{R}^d \rightarrow \{1...K\}$ be a DNN classifier with parameter θ that is used to do classification on *S*. $z_k(x_i,\theta)$ is the predicted vector of x_i respect to class *K*, which detailed explanation as follows:

$$z_k(x_i,\theta)=[p_1(x_i,\theta),\ p_2(x_i,\theta),..,p_k(x_i,\theta)] \tag{1}$$

where $p_k(x_i,\theta)$ is the predicted probability of xi belonging to class k, which is a scalar. $z_k(x_i,\theta)$ contains *K* prediction probabilities, representing the prediction results of the model for *K* classes. Note that the sum of all scalars in [•] is 1.

3.1. Adversarial Training

AT is a effective training method to improve the robustness and generalization ability of the model. AT includes making AEs and using them in the training process, so that the neural network model gradually adapts to this change and has certain robustness to the generated AEs. Specifically, the optimization objective of AT is:

$$\min_{\theta}\frac{1}{n}\sum_{i=1}^{n}\max_{\|x_i'-x_i\|_p\leqslant\epsilon}\mathcal{L}(h_\theta(x_i'),y_i) \tag{2}$$

where x_i' is the AE within the ϵ-ball(bounded by an L_p-norm) centered at natural example x_i, $\mathcal{L}(\cdot)$ is the standard classification loss (e.g., the cross-entropy(CE) loss), n is the number of training examples.

The internal maximization is used to generate the most powerful AEs, which is usually performed through an iterative gradient-based optimization, such as Projection gradient descent [11] (PGD), Fast Gradient Sign Method [7](FGSM). The external minimization is used to train the robust DNN model.

3.2. Misclassification Aware adversarial Training

Wang et al. claimed that the misclassified examples have a significant impact on the final robustness, so that the adversarial data generated by the misclassified examples should be regarded as more critical data. They divided the natural training examples into two subsets with respect to h_θ, i.e., correctly classified examples ($S^{+}_{h_\theta}$) and misclassified examples ($S^{-}_{h_\theta}$) :

$$\begin{aligned} S^{+}_{h_\theta} &= \{i:i\epsilon[n],h_\theta(x_i) = y_i\}; \\ S^{-}_{h_\theta} &= \{i:i\epsilon[n],h_\theta(x_i) \neq y_i\} \end{aligned} \tag{3}$$

Based on this intuition, they proposed a new AT approach, called Misclassification Aware adveRsarial Training (MART), incorporating an explicit differentiation of misclassified examples as a regularizer. The adversarial risk of MART is :

$$\begin{aligned} \mathcal{R}^{MART}(\theta) &= \frac{1}{n}\Big(\sum\nolimits_{i\epsilon S^{+}_{h_\theta}} R^{+}(h_\theta,x_i) + \sum\nolimits_{i\epsilon S^{-}_{h_\theta}} R^{-}(h_\theta,x_i)\Big) \\ &= \frac{1}{n}\sum\nolimits_{i=1}^{n}\{\mathbb{I}(h_\theta(x_i') \neq y_i) + \mathbb{I}(h_\theta(x_i) \neq h_\theta(x_i'))\cdot\mathbb{I}(h_\theta(x_i) \neq y_i)\} \end{aligned} \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathcal{R}^{+}(h_\theta,x_i)$ and $\mathcal{R}^{-}(h_\theta,x_i)$ are the adversarial risks for correctly classified examples and misclassified examples, respectively. In Equation 4, they combined two adversarial risks in an adversarial training framework, and show that the network can be trained by minimizing the risk $\mathcal{R}^{MATR}$.

MART uses the indicator function $\mathbb{I}(h_\theta(x_i) \neq y_i)$ as a condition that emphasizes learning on misclassified examples. However, the condition cannot be directly optimized during the training process. Hence, they proposed to use the prediction error of of the ground truth label, i.e., $1 - p_{y_i}(x_i,\theta)$ to replace $\mathbb{I}(h_\theta(x_i) \neq y_i)$, where $p_{y_i}(x_i,\theta)$ represents the predicted probability of the ground-truth class for the i-th example. This error will be large for misclassified examples and small for correctly classified examples.

4. Methodology

MART only uses the prediction error regarding with the ground truth label to identify the misclassified and correctly classified examples, neglecting the impacts from the other classes. To tackle this problem, we proposed an improved MART method termed as E-MART. In our method, we develops an entropy-based uncertainty measure (EUM) to provide reliable evidence for indicating the impact of the misclassified and correctly classified examples, and then apply a soft decision approach to evaluate the weight of the impact of the misclassified examples.

Inspired by the concept of information entropy [33], we offer a unique insight into the case that distinguishes misclassified examples. We observe that the uncertainty of the prediction results can be used as a liable measure to indicate the learning effect of a model. That is, smaller uncertainty of a prediction implies superior learning effect a model, and further contends the lower probability of an example to be misclassified; and vice versa. Since the prediction uncertainty can be expressed as the function of entropy, our proposed EUM can be represented as

$$EUM(x_i,\theta) = -\sum_{k=1}^{K} p_k(x_i,\theta)\log\left(p_k(x_i,\theta)\right) \tag{5}$$

Since EUM considers all the components of the model output, rather than the prediction probability of the ground-truth class, It provides more reliable evidence to indicate the impact of the misclassified and correctly classified examples and fills the gap that complement classes have not been explicitly optimized in MART. Furthermore, Based on EUM, we propose an improved MART AT framework termed as E-MART. E-MART use EUM to evaluate the misclassified examples in the inner maximizing loop of AT.

The adversarial risk of Equation 4 is give in the form of 0-1 loss. However, optimization over 0-1 loss is intractable in practice. for our E-MART framework, we replace the 0-1 losses with proper surrogate loss functions which are both physical meaningful and computationally tractable. As shown in Equation 4, the loss required to be substituted is composed of three indicator functions: (1) $\mathbb{I}(h_\theta(x_i') \neq y_i)$, (2) $\mathbb{I}(h_\theta(x_i) \neq h_\theta(x_i'))$, (3) $\mathbb{I}(h_\theta(x_i) \neq y_i)$.

For the first indication function, instead of the commonly used CE loss in previous works[35, 36], we used a boosted cross entropy (BCE) loss as the surrogate loss as MART does. The BCE loss can be represented as

$$BCE(p(x_i',\theta),y_i) = -\log(p_{y_i}(x_i',\theta)) - \log\left(1 - \max_{k \neq y_i} p_k(x_i',\theta)\right) \tag{6}$$

For the second indication function, drawing on MART, we used KL divergence as the surrogate loss function [35, 36] to measure the different output distributions between AEs and natural examples. Then, we have

$$KL(p(x_i,\theta) \parallel p(x_i',\theta)) = \sum_{k=1}^{K} p_k(x_i,\theta)\log\frac{p_k(x_i,\theta)}{p_k(x_i',\theta)} \tag{7}$$

For the third indicator function, instead of using the prediction error of ground-truth in MART to identify the misclassified examples, we employ the proposed EUM to substitute this indicator function of 0-1 loss.

Eventually, the final objective function for the proposed E-MART can be representted as

$$\mathcal{L}^{E-MART}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\{BCE(p(x_i',\theta),y_i) + \lambda \cdot KL(p(x_i,\theta) \parallel p(x_i',\theta)) \cdot EUM(x_i,\theta)\right\} \tag{8}$$

where λ is a tunable scaling parameter that balances the supervised and unsupervised loss, and is fixed for all training examples.

Note that the EUM in Equation 8 is served as a soft decision schema, which weights the impacts of the misclassified and correctly classified examples during AT. Specifically, larger/smaller EUM implies greater/smaller probability of an example to be misclassified.

5. Experiments

In this section, we conducted experiments under various white box attacks on the CIFAR-10 [37] dataset to evaluate our method. The architectures of the base classifiers are ResNet (ResNet-18) [1] and wide ResNet (WRN-34-10) [38].

5.1. Experimental Setup

Training Parameters. For the sake of fair comparison, we adopt the training parameters as MART used. That is, the classifiers are trained using SGD with momentum 0.9, weight decay 3.5×10^{-3}, batch size 128, and an initial learning rate of 0.01. For generating the adversarial examples, we use the PGD-10 attack method with random start and step size $\epsilon/4$, the perturbation is subjected to L_∞ constraint of $\epsilon = 8/255$. For the trade-off of parameter λ, it is set to 6(for ResNet-18) / 7 (for WRN-34-10).

Baselines. The comparing method include standard AT(Standard) and its variants.

- Standard[11]: standard adversarial training (PGD), which is the most effective method for performing norm constrained attacks.
- MMA[27] : max-margin adversarial training, which proposed to use maximal margin optimization for correctly classified examples while keeping the optimization on misclassified examples unchanged.
- Dynamic[34] : the adversarial training with a criterion that dynamically controls the convergence quality of the inner maximization.
- TRADES[35] : the robust error can be decomposed into the sum of natural (classification) error and boundary error to describe the trade-off between accuracy and robustness of the classification problem.
- MART[28] : misclassification aware adversarial training, which explicitly differentiates the misclassified and correctly classified examples during the training.

Robustness Evaluation. We evaluated our methods and baselines using the standard test accuracy on natural data (Natural) and the adversarial robustness based on several attack methods, including the PGD method with 20 iterations [11], FGSM attack [7], CW attack [24]. All these methods have full access to the model parameters (i.e., white-box attacks) and are constrained by the same perturbation limit as above.

5.2. Performance Evaluation

In this section, to verify the effectiveness of our proposal, we compare our method with its competitors on WideResNet-34-10 and ResNet-18 under the CIFAR-10 dataset. Each method is executed six repeated trials with different random seeds, and the medians and standard deviations of the experimental results are recorded. We compare different methods on the **best** (the “best” refers to the highest robustness that ever achieved at different checkpoints.) checkpoint model (suggested by [39]) and the **last** checkpoint model (used by [11]), respectively. We evaluate the robust models based on the five evaluation metrics, i.e., standard test accuracy on natural data (Natural), robust test accuracy on adversarial data generated by FGSM [7], PGD-20 [11] and CW [24].

Table 1 shows the experimental results of the comparing methods on WideResNet-34-10. To clearly exhibit the variation tendency of the model robustness, we plot the robustness curves at each epoch in training for MART and our method in Figure 1. From Table 1, we can clearly observe that compared to MART, our method significantly enhances both the best checkpoint model and the last checkpoint model in terms of adversarial robustness, without declining the prediction accuracy of the natural examples. In addition, we can see from Figure 1 that the adversarial robustness of our method is higher than that of MART under variant attack (e.g.,FGSM, CW, PGD-20) in the late training period (90-120 epoch), implying the proposed method is more stable than MART.

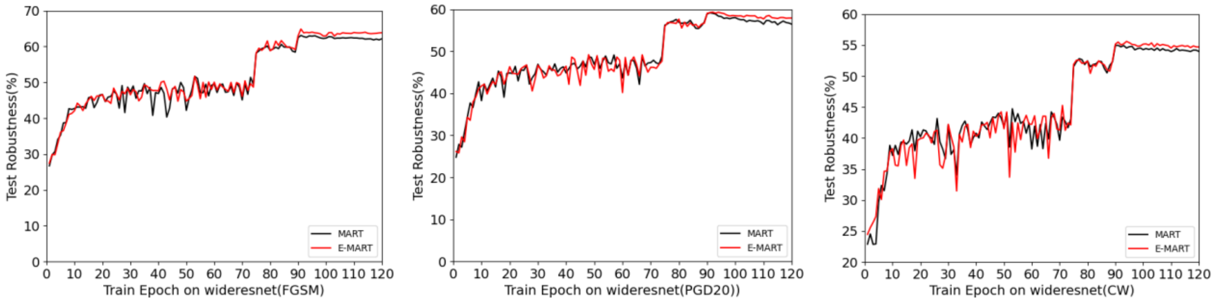
Table 2 shows the adversarial robustness of the testing methods on ResNet-18. Figure 2 depicts the variation tendency of the model robustness obtained by our method and MART. From Table 2 and Figure 2, we can observe that our method superior to the competitors in terms of the model best and last robustness in the absent of losing the prediction accuracy of natural instances. Meanwhile, the proposed method is more stable than MART.

Table 1: White-box robustness (accuracy (%) on white-box test attacks) on CIFAR-10 dataset using the **WRN-34-10**.

Defence	Natural	FGSM		PGD20		CW$_\infty$	
		Best	Last	Best	Last	Best	Last
Standard	**87.41**	55.32	54.67	52.17	49.13	50.04	48.36
Dynamic	82.01	62.14	62.16	55.10	51.36	50.22	49.31
TRADES	83.11	62.81	62.67	56.01	52.61	51.19	50.23
MART	83.97	63.08	62.15	58.85	56.81	54.51	54.45
E-MART	83.87	**65.84**	**64.05**	**59.95**	**57.93**	**56.74**	**56.15**

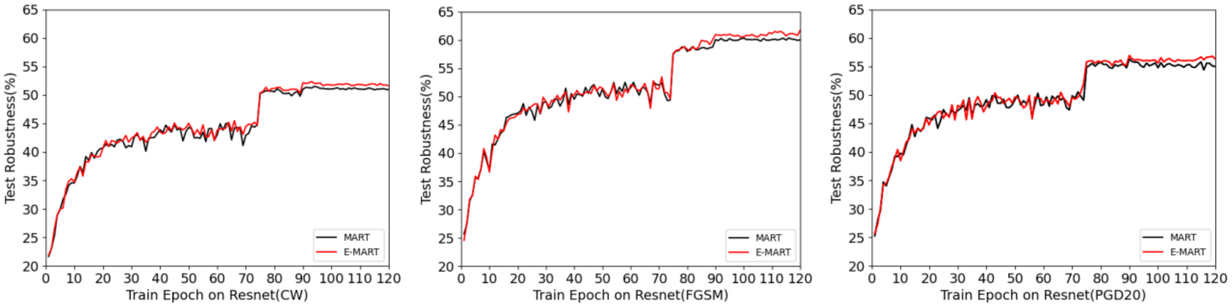
Table 2: White-box robustness (accuracy (%) on white-box test attacks) on CIFAR-10 dataset using the **ResNet-18**.

Defence	Natural	FGSM		PGD20		CW$_\infty$	
		Best	Last	Best	Last	Best	Last
Standard	83.38	60.01	59.76	47.31	45.81	45.33	44.99
MMA	**83.89**	61.14	59.77	48.30	47.91	44.78	43.09
Dynamic	82.12	60.83	58.77	48.44	46.56	46.08	44.31
TRADES	81.97	61.01	59.31	50.77	50.21	48.31	46.93
MART	82.83	**61.85**	60.00	55.61	54.17	51.49	50.93
E-MART	82.81	61.70	**61.94**	**57.60**	**56.53**	**53.51**	**52.84**



(a) (b) (c)

Figure 1: Comparison of MART and E-MART on CIFAR-10 dataset using the WideResNet-34-10.



(a) (b) (c)

Figure 2: Comparison of MART and E-MART on CIFAR-10 dataset using the ResNet-18 .

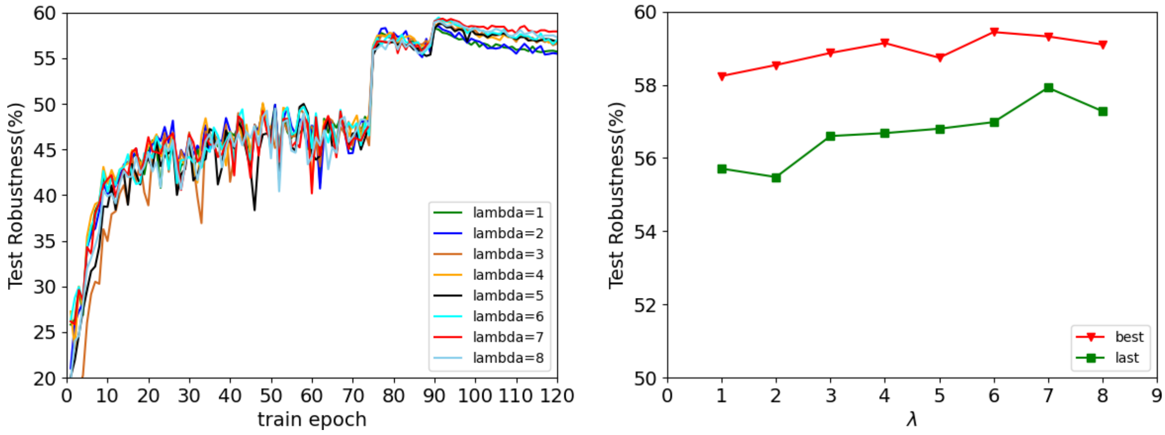
5.3. Ablation Experiment

In this section, we investigate the effect of the super-parameter λ, which controls the strength of the regularization, for the objective function of E-MART defined in Equation 8 .

Figure 3(a) shows the variation tendency of the model robustness obtained by our method under different $\lambda \in [1, 8]$. We run eight repeated experiments on the WRN-34-10 network with the same random seed and different λ values, and plotted the robustness curves in (a) of Figure 3. We can see that the different λ values make a significant difference in the robustness curves of the model, among which the red curve(λ=7) has favorable performance in the late training period.

Figure 3(b) shows the robustness of the best checkpoint model and the last checkpoint under different $\lambda \in [1, 8]$. In order to select the best λ more clearly, we take " the robustness of the best checkpoint model and the last checkpoint model " as two criteria, and the results are shown in (b) of Figure 3. We can see that the model has favorable performance in terms of test robustness under λ=7. Eventually, we set λ to 7 (for WRN-34-10) as a the tradeoff parameter.

For the ResNet-18 network, the tradeoff parameter is λ to 6 in our ablation experiment.



(a) (b)

Figure 3: The ablation experiments of E-MART.

6. Conclusion

This paper mainly focus on improving the DNN robustness for AT depending upon misclassified examples. Instead of using the prediction error of the ground truth label to identify the misclassified and correctly classified examples as MART does, we develop an entropy-based uncertainty measure to provide a more reliable evidence to indicate the impact of the misclassified and correctly classified examples. Moreover, we employ a soft decision scheme based on the entropy-based uncertainty measure to optimize the loss function of AT, which is contribute to making the efficient training for the model's final robustness. Experimental results demonstrated that our proposal is superior to MART in terms of the model robustness and prediction accuracy. Our future work will pay more attention to the exploration of the correlations between the misclassified and correctly classified examples and study more capable approach to enhance the roustness of DNN model.

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[2] Y. Wang, X. Deng, S. Pu, Z. Huang, Residual convolutional CTC networks for automatic speech recognition, arXiv preprint arXiv:1702.07793.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language under- standing, arXiv preprint arXiv:1810.04805.

[4] M. Zeng, Y. Wang, Y. Luo, Dirichlet latent variable hierarchical recurrent encoder-decoder in dialogue generation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 1267–1272.

[5] X. Wang, J. Li, X. Kuang, Y.-a. Tan, J. Li, The security of machine learning in an adversarial setting: A survey, Journal of Parallel and Distributed Computing 130 (2019) 12–23.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199.

[7] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572.

[8] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, X. Ma, Skip connections matter: On the transferability of adversarial examples generated with resnets, arXiv preprint arXiv:2002.05990.

[9] H. Chen, K. Lu, X. Wang, J. Li, Generating transferable adversarial examples based on perceptually-aligned perturbation, International Journal of Machine Learning and Cybernetics 12 (11) (2021) 3295–3307.

[10] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial machine learning at scale, arXiv preprint arXiv:1611.01236.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083.

[12] F. Tram`er, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, P. McDaniel, Ensemble adversarial training: Attacks and defenses, arXiv preprint arXiv:1705.07204.

[13] C. Guo, M. Rana, M. Cisse, L. Van Der Maaten, Countering adversarial images using input transformations, arXiv preprint arXiv:1711.00117.

[14] Y. Bai, Y. Feng, Y. Wang, T. Dai, S.-T. Xia, Y. Jiang, Hilbert-based generative defense for adversarial examples, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4784–4793.

[15] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, J. Zhu, Defense against adversarial attacks using high-level representation guided denoiser, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1778–1787.

[16] W. Xu, D. Evans, Y. Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, arXiv preprint arXiv:1704.01155.

[17] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, J. Bailey, Characterizing adversarial subspaces using local intrinsic dimensionality, arXiv preprint arXiv:1801.02613.

[18] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Advances in neural information processing systems, Vol. 31, 2018.

[19] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: 2016 IEEE European symposium on security and privacy (EuroS&P), IEEE, 2016, pp. 372–387.

[20] N. Papernot, P. McDaniel, Extending defensive distillation, arXiv preprint arXiv:1705.05264.

[21] A. Ross, F. Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[22] Q. Liu, T. Liu, Z. Liu, Y. Wang, Y. Jin, W. Wen, Security analysis and enhancement of model compressed deep learning systems under adversarial attacks, in: 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE, 2018, pp. 721–726.

[23] X. Wang, J. Li, Q. Liu, W. Zhao, Z. Li, W. Wang, Generative adversarial training for supervised and semi-supervised learning, Frontiers in Neurorobotics 16.

[24] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 ieee symposium on security and privacy (sp), Ieee, 2017, pp. 39–57.

[25] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, M. Jordan, Theoretically principled trade-off between robustness and accuracy, in: International conference on machine learning, PMLR, 2019, pp. 7472–7482.

[26] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, M. Kankanhalli, Attacks which do not kill training make adversarial learning stronger, in: International conference on machine learning, PMLR, 2020, pp. 11278–11287.

[27] G. W. Ding, Y. Sharma, K. Y. C. Lui, R. Huang, Mma training: Direct input space margin maximization through adversarial training, arXiv preprint arXiv:1812.02637.

[28] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, Q. Gu, Improving adversarial robustness requires revisiting misclassified examples, in: International Conference on Learning Representations, 2019.

[29] A. Kurakin, I. J. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: Artificial intelligence safety and security, Chapman and Hall/CRC, 2018, pp. 99–112.

[30] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.

[31] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, P. S. Liang, Unlabeled data improves adversarial robustness, Advances in Neural Information Processing Systems 32.

[32] E. Wong, L. Rice, J. Z. Kolter, Fast is better than free: Revisiting adversarial training, arXiv preprint arXiv:2001.03994.

[33] C. E. Shannon, A mathematical theory of communication, ACM SIGMOBILE mobile computing and communications review5 (1) (2001) 3–55.

[34] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, Q. Gu, On the convergence and robustness of adversarial training, arXiv preprint arXiv:2112.08304.

[35] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, M. Jordan, Theoretically principled trade-off between robustness and accuracy, in: International conference on machine learning, PMLR, 2019, pp. 7472–7482.

[36] S. Zheng, Y. Song, T. Leung, I. Goodfellow, Improving the robustness of deep neural networks via stability training, in: Proceedings of the ieee conference on computer vision and pattern recognition, 2016, pp. 4480–4488.

[37] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images , cs.toronto.edu, 2009.

[38] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146.

[39] L. Rice, E. Wong, Z. Kolter, Overfitting in adversarially robust deep learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 8093–8104.