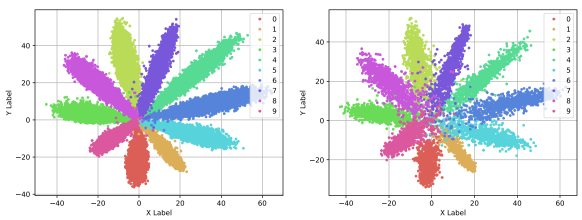
Boosting Robustness of Neural Networks via Angular Boundary-oriented Cosine Loss Framework

Hongtian Zhao, Hua Yang *Member, IEEE*, Hang Su *Member, IEEE*, Shibao Zheng

APPENDIX

APPENDIX A

To better understand the reasons why the norm of y is supposed to be fixed as a constant, we make the following analysis. First, in $L_{Cos-SCE}$, if s is replaced with $\|y\|$, and the θ_{d_i} and $\theta_{\hat{d}_i}$ are both constants, the loss is a function only concerning $\|y\|$ as independent variable. Under this premise, $(\cos\theta_{d_i} - \cos\theta_{\hat{d}_i}) \geq 0$ means that the model makes correct classification result with our analysis; with the increase of $\|y\|$, the loss function value decreases accordingly. Conversely, if $(\cos\theta_{d_i} - \cos\theta_{\hat{d}_i})<0$ holds, the model would make a false prediction and the decrease of $\|y\|$ would also result in the decrease of loss function. As a result, the magnitude of sample vector $\|y\|$ has no influence on the model making a decision, and in fact, the decreasing trend of $\|y\|$ may well just induce the model to make false predictions accumulating around the origin while the correct predictions spread far away from the origin. For validation, Fig. 1 a, b shows the sample distributions of training and test data embedded in a two-dimensional feature space obtained by a typical angular loss. From the figure, we can see that the samples nearby the origin are more likely to be misclassified, while the samples far away from the origin have a higher probability to be accurately classified, manifesting the validity of our analysis. Thus, to decrease the undesirable behavior of features due to the magnitude difference among different types of datasets, we here fix the $\|y\|$ as a constant to reduce the radial variance during training in an empirical manner.



(a) Training visualization result (b) Test visualization result

Fig. 1. Feature visualization of training and test set under the way of A-Softmax training on MNIST [1].

APPENDIX B

Table I shows the time elapsed during inference. Similar to accuracies in Tables, here, each experiment was repeated 5 times, and then the mean time along with its corresponding standard deviation was recorded in tables for increasing reliability of the results. The results indicate that the computations involved in feature and weight normalizations result in negligible time overhead.

Different from CosFace-based L_B definition, hypersphere transformation (embedding) based SCE (HT-SCE) for L_B uses naive linear layer to map high-dimensional normalized features to output space and then exploit feature and weights normalization, angular margin operation on these features to obtain final latent features. Fig. 2 shows the classification accuracies on both normal samples and adversarial examples using the same setting as the Fig. 6 in the body of this paper by enabling HT-SCE mode in the ABCL framework. We observe that most models trained based on the SCE-based L_B with different coefficients obtains high prediction accuracy on normal images except in the case of $(0, 0)$ parameter configuration. The cause of $(0, 0)$ degradation can be attributed to the hypersphere transformation applied to the last layer's features, and there are two reasons accounting this phenomenon. First, hypersphere transformation may change the feature distribution. Since the Softmax classifier is a probability model based on the data distribution, if the feature distribution is changed, it will affect the model's fitting ability and lead to a decrease in model performance. Second, the transformation introduces additional non-linear transformations, which generates many exceptions of negative (or abnormal) feature vectors that cannot be self-consistently revised by the subsequent HT-SCE loss computing mechanism.The interference negatively affects the selection of gradient optimization direction, and this abnormality also appears in adversarial test experiments, highlighting the problem of model degradation. The revised learning mechanism, HT-SCE+**AB**, significantly reduces this issue by addressing the feature difference originating from angular transformation and enhancing the focus on hard samples, consequently improving adversarial robustness. In addition, the figure illustrates the sensitivity of robustness to the change of two coefficients, highlighting the difference between euclidean and cosine optimization spaces corresponding to L_B and L_I/L_O. Empirical observations suggest that the hyperparameter combination settings of $\{(64, 64), (128, 128), (256, 256), (512, 512), (512, 64), (256, 4)\}$ can be effective.

APPENDIX C

The sensitivity analysis of s is concisely discussed in the rest of this section. Its value is restricted in $[0.125, 0.25, 0.5, 1, 2, 3, ..., 8, 9, 10, 15, 16, 20]$. Fig. 3 shows

H. Zhao, H. Yang, S. Zheng are with the Department of Electronic Engineering in Shanghai Jiao Tong University, and H. Su is with the Department of Computer Science and Technology in Tsinghua University.

TABLE I

INFERENCE TIME COMPARISON (/S) BY USING DIFFERENT MODELS, WHICH ARE TESTED IN GPU PARALLEL ENVIRONMENT, HERE, THE TEST BATCH SIZE IS SET TO 20. IN TERMS OF INFERENCE TIME, A LOWER NUMERICAL VALUE INDICATES BETTER PERFORMANCE.

Method	PGD	TRADES	ALP	CosFace	CosFace+**AB**	ArcFace	ArcFace+**AB**	SCE	HT-SCE+**AB**
Clean	2.67±0.02	**2.64**±0.06	2.72±0.07	2.82±0.14	2.82±0.03	2.87±0.03	2.89±0.07	2.67±0.03	2.84±0.05
FGSM	2.52±0.06	**2.43**±0.03	2.52±0.01	2.67±0.09	2.74±0.04	2.79±0.08	2.77±0.01	2.54±0.05	2.70±0.13
PGD	2.53±0.05	**2.43**±0.02	2.53±0.03	2.59±0.02	2.72±0.04	2.74±0.14	2.77±0.01	2.54±0.07	2.61±0.02

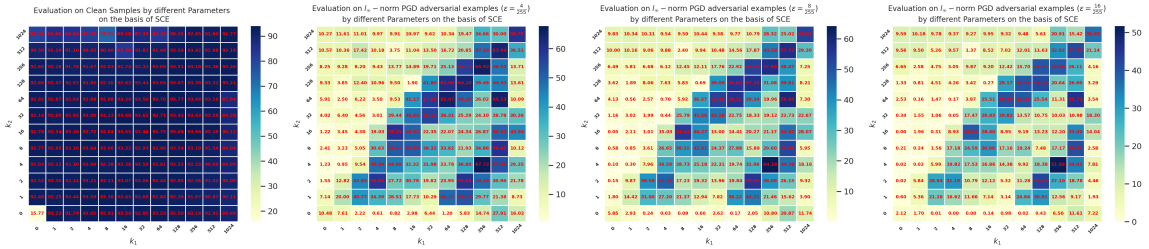


Fig. 2. Classification accuracy of joint clean samples and adversarial examples on CIFAR-10 [2] using different settings of parameters based on HT-SCE.

TABLE II

COMPARISON BETWEEN THE PROPOSED FRAMEWORK AND I-SCE [10] ON TWO DATASETS UNDER THE UNTARGETED ATTACK, THE STANDARD VARIANCE OF OUR METHOD IS 0.

	MNIST			CIFAR10		
Method	Clean	$\epsilon = 0.02$	$\epsilon = 0.04$	Clean	$\epsilon = 0.01$	$\epsilon = 0.04$
I-SCE	**99.57**	63.47	46.41	89.09	14.82	5.51
CosFace+AB	99.3	99.25	99.19	**93.62**	**56.16**	**56.15**
ArcFace+AB	99.41	**99.4**	**99.34**	90.99	51.51	51.51

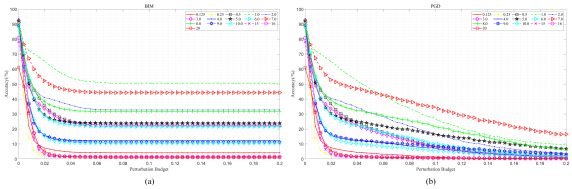


Fig. 3. The accuracy versus perturbation budget curves of the 16 models trained with different s values for HT-SCE as the basis of loss on CIFAR-10 against BIM and PGD white-box attacks under the ℓ_∞ norm constraint. For all given abnormal samples, it indicates that the overall performance including test accuracy and robustness under this mode at $s = 1, 7$ can obtain relatively good results.

the quantitative results of the ABCL framework with different numerical parameters. It's obvious that different kinds of basis functions make the trained models behave differently. As for the SCE-based naive functions (HT-SCE), the obtained models have negligible influences on standard accuracy (perturbation budget is 0), except for s values of 15, 16 and 20, indicating that a large s can negatively affect model performance. Furthermore, changes in s have an impact on model robustness against adversarial perturbation. Considering the overall effects of different s settings, we observe that $s = 1$ or 7 for the HT-SCE+**AB** allows models to achieve relatively better accuracy on normal samples and higher robustness against adversarial examples.

APPENDIX D

Fig. 4 illustrates the effectiveness of using different s values to defend against the Jitter attack [8].

TABLE III

COMPARISON BETWEEN THE PROPOSED FRAMEWORK AND I-SCE [10] ON TWO DATASETS UNDER THE TARGETED ATTACK.

	MNIST		CIFAR10	
Method	$\epsilon = 0.02$	$\epsilon = 0.04$	$\epsilon = 0.01$	$\epsilon = 0.04$
I-SCE	12.74	6.78	5.26	1.95
CosFace+AB	99.28±0.004	99.29±0.006	**56.16±0.0**	**56.16±0.0**
ArcFace+AB	**99.42±0.005**	**99.43±0.007**	51.51±0.0	51.51±0.0

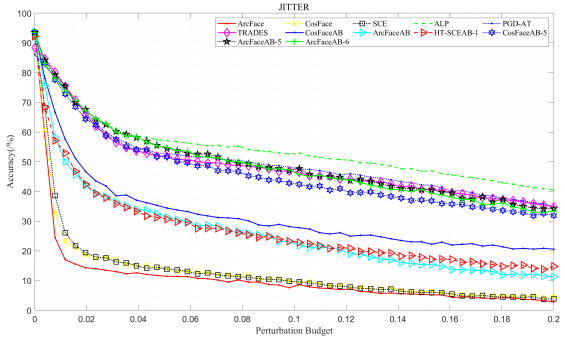


Fig. 4. The accuracy versus perturbation budget curves of the 12 models trained with different s values for ABCL framework on CIFAR-10 against Jitter white-box attack under the ℓ_∞ norm constraint. As for abnormal samples generated via Jitter attack, it indicates that the overall performance including test accuracy and robustness under this mode at $s = 5, 6$ for ArcFace based AB loss method can obtain good results, which are similar to adversarial training methods.

APPENDIX E

We also use the t-SNE [5] algorithm to visualize the latent features. In the experiments, 10-dimensional features learned by Alexnet [6] are then projected onto 2−dimensional plane, and shown in Fig. 5. Here, each point symbolizes a sample and its color represents the corresponding category. In terms

TABLE IV
COMPARISON BETWEEN THE PROPOSED FRAMEWORK AND I-SCE [10] ON TWO DATASETS UNDER SIMBA [11] BLACK-BOX SETTING.

	MNIST				CIFAR10			
Method	clean	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$	clean	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$
I-SCE	99.40	98.73	97.90	97.10	**89.63**	85.93	83.33	82.07
CosFace+AB	**99.90±0.01**	**99.90±0.02**	**99.89±0.03**	**99.87±0.02**	89.56±0.08	**89.53±0.08**	**89.56±0.03**	**89.56±0.06**
ArcFace+AB	99.75±0.01	99.73±0.03	99.77±0.01	99.76±0.03	85.37±0.09	85.48±0.17	85.49±0.08	85.56±0.16

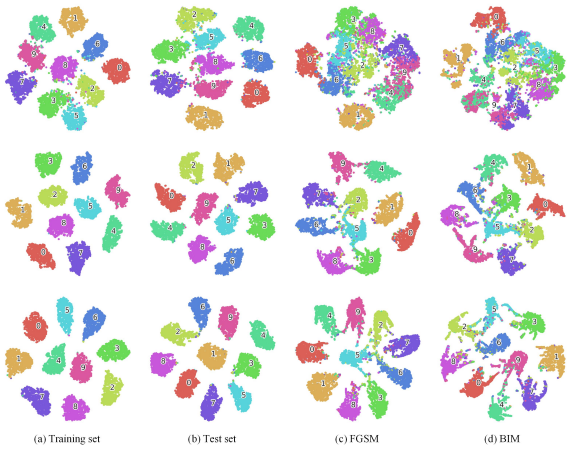


Fig. 5. Latent features visualization of training, test and adversarial test (attacked by BIM [9] under ℓ_∞ norm and FGSM [7] under ℓ_∞ norm constraints in sequence) sets under different losses guidance on MNIST [1], the experiment results from the first to third rows are corresponding to SCE, CosFace [3] and ours in sequence.

of the feature distribution of both training and test samples, our method and CosFace achieve better discriminability among different categories compared to the SCE method. The main reason accounting for it is that the SCE-based method cannot benefit from DNNs to obtain a discriminative distribution of latent features due to the existence of radial variance, whereas cosine transformation improves the learning ability of linear layers in latent representation, which, in turn, has an impact on the convolutional layers through the back-propagation process.

In Fig. 5, the 3_{ed} and 4_{th} columns represent the visualization of latent features of adversarial examples, generated using FGSM [7] and BIM [9] attacks, respectively. Compared with SCE, there are fewer overlaps among different categories obtained via the other two losses. However, CosFace suffers from category adhesion as the large margin cosine loss computation method may have a weaker ability to guide inter-class discrepancy. Different from previous approaches, the proposed method jointly optimizes the minimum inter-class angular distance and intra-class angular distance in cosine space to achieve a more rational feature distribution. The interim structural features can guide good classification decisions, which are consistent with quantitative comparison shown in Table IV in the body of this paper. The interim structural features effectively guide good classification decisions, consistent with the quantitative comparison shown in Table IV in this paper. Based on experimental results, we also observe that using the AB loss is more effective for defending against adversarial attacks than the other two methods.

APPENDIX F

In this section, we present a comprehensive experimental comparison of our proposed ABCL method with the state-of-the-art Inference-Softmax Cross Entropy (I-SCE) method [10] on MNIST [1] and CIFAR-10 [2]. To maintain consistency with the setting in [10], we use ResNet-32 as the network architecture to implement and train our model. Similarly, we employ the white-box attack and black-box attack, including untargeted and targeted PGD [12] and SimBA [11], to investigate the effectiveness of the proposed method. We first conduct white-box attack experiments under ℓ_2-norm constrained untargeted and targeted PGD attacks on the two datasets and compared results, which are listed in Table II and Table III, respectively. The comparison shows that the ABCL method outperforms the I-SCE method, and particularly, on the MNIST dataset, the proposed framework significantly boosts the adversarial robustness of deep learning models. As for black-box attack experiments, similar to [10], we use SimBA [11] to test the robustness of different methods. The comparison results are listed in Table IV, from which we can observe that under this setting models from ABCL framework can have higher robustness and higher or approximating to the best value on standard accuracy compared to I-SCE.

REFERENCES

[1] Y. Lecun *et al.*, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.
[2] A. Krizhevsky, G. Hinton, "Learning multiple layers of features from tiny images," tech. rep., University of Toronto, 2009.
[3] H. Wang *et al.*, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *Proc. CVPR.*, 2018, pp. 5265-5274.
[4] J. Deng, J. Guo, N. Xue *et al.*, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. CVPR.*, 2019, pp. 4685-4694.
[5] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579-2605, 2008.
[6] A. Krizhevsky *et al.*, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. NeurIPS.*, 2012, pp. 1106-1114.
[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *Proc. ICLR.*, 2015.
[8] L. Schwinn, R. Raab, A. Nguyen, D. Zanca, and B. Eskofier, "Exploring misclassifications of robust neural networks to enhance adversarial attacks," *arXiv:2105.10304*, 2021.
[9] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security: Chapman and Hall/CRC*, 2018, pp. 99-112.
[10] B. Song, R. Wang, W. He, and W. Zhou, "Robustness Learning via Inference-Softmax Cross Entropy in Misaligned Distribution of Image," Mathematics, vol. 10, no. 19, p. 3716, Oct. 2022.
[11] C. Guo, Jacob R. Gardner, and Y. You, "Simple Black-box Adversarial Attacks," in *Proc. ICLR.*, 2019.
[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras *et al.*, "Towards Deep Learning Models Resistant to Adversarial Attacks," in *Int. Conf. Learn. Represent., ICLR - Conf. Track Proc.*, 2018.