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

Hongtian Zhao, Hua Yang, Hang Su, Shibao Zheng

APPENDIX

APPENDIX A

To better understand the reasons why the norm of $\boldsymbol{y}$ is supposed to be fixed as a constant, we make the following analysis. First, in $L_{Cos-SCE}$, if s is replaced with $\|\boldsymbol{y}\|$, and the θ_{d_i} and $\theta_{\hat{d}_i}$ are both constants, the loss is a function only concerning $\|\boldsymbol{y}\|$ 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 $\|\boldsymbol{y}\|$, 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 $\|\boldsymbol{y}\|$ would also result in the decrease of loss function. As a result, the magnitude of sample vector $\|\boldsymbol{y}\|$ has no influence on the model making a decision, and in fact, the decreasing trend of $\|\boldsymbol{y}\|$ 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. 3 in the body of this paper 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 $\|\boldsymbol{y}\|$ as a constant to reduce the radial variance during training in an empirical manner.

APPENDIX B

Different from CosFace-based L_B definition, hypersphere transformation (embedding) based SCE, hereinafter referred to as HT-SCE, for L_B uses naive linear layer to map high-dimensional normalized features to output space and then exploit feature&weights normalization, angular margin operation on these features to obtain final latent features. Fig. 1 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. One reason can account for the phenomenon: exploiting hypersphere transformation on the features from the last layer, usually generates a large number of exceptions of negative (or abnormal) feature vectors, and the subsequent HT-SCE loss computing mechanism cannot revise the inconsistent numerical sign exceptions in a self-consistent manner, and further interfere with the selection of gradient optimization direction. It can be seen that the abnormal phenomena also appears in adversarial test experiments, which further demonstrate the fact of model degradation. When modifying the loss criterion, the new criterion is defined as HT-SCE+**AB**. The revised learning mechanism can effectively relieve the problem of the features difference originating from the angular margin-based transformation, meanwhile, the DNNs are imposed to focus more on the hard samples, which further improves their adversarial robustness. In the figure, we can see that the robustness is sensitive to the change of two coefficients due to the difference between euclidean and cosine optimization spaces corresponding to L_B and L_I/L_O. Empirically we consider the combination of $k_1 = 512$, $k_2 = 64$ as an appropriate parameter setting.

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.

APPENDIX C

As for the SCE-based basis function, with results shown in Fig. 2, we find that the numerical sensitivity of s has more impact on the robustness of deep learning models compared with cosine-based basis functions; when $s>8$, the model is apt to degrade, showing poor performance on clean samples and adversarial examples, and one possible reason for this phenomenon is that large s will force model to excessively concentrate on angular separation and ignore other information, which would limit the model representation capability. Considering the overall effects of different s settings, we observe that $s = 4$ or 5 for the SCE-based loss allows models to have comparatively better accuracy on normal samples, and higher robustness against the adversarial examples.

APPENDIX D

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. 3. Here, each point symbolizes a sample and its color represents the corresponding category. As for the feature distribution of training and test samples, the results of our method and CosFace have better discriminativeness among different categories compared with 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, while the

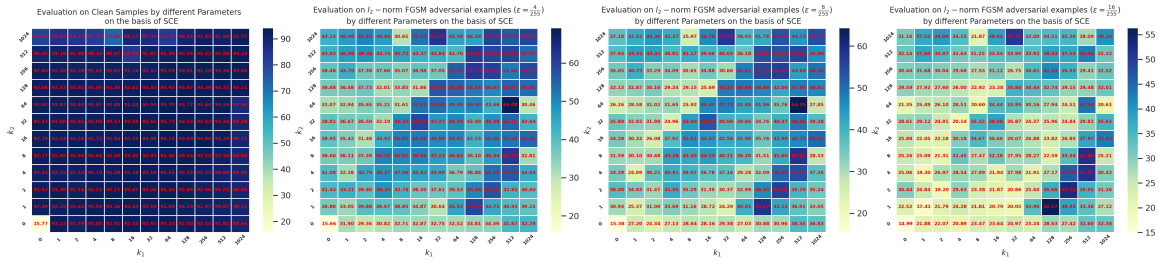

Fig. 1. Classification accuracy of joint clean samples and adversarial examples on CIFAR-10 [2] using different settings of parameters based on analogous cosine-transformed SCE.

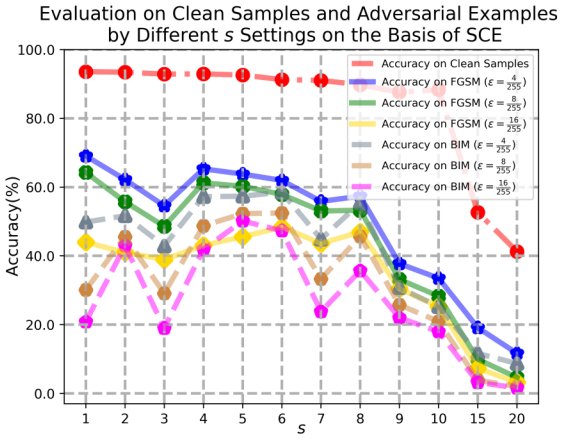

Fig. 2. The test accuracy decreases with the increase in s of the SCE-based method. The numerical values of s have some impact on the accuracy of different adversarial examples in this method, and the trends are similar among different adversarial examples. For all given abnormal samples, it indicates that the overall performance including test accuracy and robustness under this mode at $s = 5$ can obtain good results.

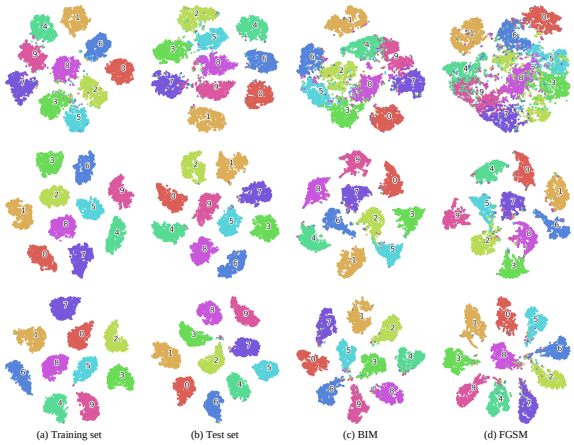

Fig. 3. Latent features visualization of training, test and adversarial test (attacked by BIM [8] under ℓ_∞ norm and FGSM [7] under ℓ_2 norm constraints in sequence) sets under different losses guidance on MNIST [1], the experiment results from the first to third rows are corresponding to SCE, CosFace [3] and ours in sequence.

cosine transformation does improve the latent representation learning ability of linear layers, and further have an impact on the convolutional layers via the back-propagation process.

In Fig. 3, the 3_{ed} and 4_{th} columns are the latent features visualization of adversarial examples, attacking by BIM [8] and FGSM [7], respectively. Compared with SCE, there are fewer overlaps among different categories obtained via the other two losses. As for the comparison between CosFace [3] and the proposed method, there exists adhesion phenomenon of different categories for CosFace results, since the computation method of large margin cosine loss may have a weak ability to guide inter-class discrepancy. Different from previous approaches, the presented method considers jointly optimizing the minimum inter-class angular distance and intra-class angular distance in cosine space to obtain a more reasonable features distribution. The interim structural features can guide good classification decisions, which are consistent with quantitative comparison shown in Table V in the body of this paper. Based on the experimental results, we further obtain the trend that using the AB loss is more capable of defending against adversarial attacks compared with the other two methods.

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] 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.