Supplementary Material

Anonymous ECCV submission

Paper ID 4810

1 Experimental Results on ImageNet

We evaluate our method on the large-scale dataset, *i.e.* the ImageNet dataset [2], to further validate its effectiveness. In previous researches for robust classification against adversarial examples, this challenging dataset has been seldom evaluated on. For fair comparison, we use the network of ResNet-101. We compare with the Adversarial Logit Pairing (ALP) method [1], of which the results are shown in Table 1. Note that our experiments are conducted on a single GPU. With a large number of GPUs, *e.g.* 128 GPUs, the top-5 accuracy can be boosted to 49.7% using a very large batch size such as 4096 as in [3]. However, we do not intend to pursue the improvement through such large-scale distributed training. We concentrate on the effectiveness of the novel bidirectional likelihood regularization for improving the robustness of classification models.

Table 1. Classification accuracy (%) on the ImageNet dataset for PGD-10 adversarial attacks. The Mixup and Label Smoothing methods are baseline methods compared in [1]. The training of the ALP method and our method is conducted on the PGD-10 adversaries.

Method	Top-5	Top-5
Mixup	0.1	1.5
Label Smoothing	1.6	10.0
ALP [1]	30.2	55.8
ATBLR (ours)	**47.4**	**62.8**

References

1. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. arXiv preprint arXiv:1803.06373 (2018) 1
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Annual Conference on Neural Information Processing Systems (2012) 1
3. Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1