Mitigating robust overfitting via self-residual-calibration regularization (Abstract Reprint)

Hong Liu[1] , **Zhun Zhong**[2] , **Nicu Sebe**[2] and **Shin'ichi Satoh**[1,3]
[1]National Institute of Informatics, Tokyo
[2]University of Trento
[3]The University of Tokyo, Tokyo
hliu@nii.ac.jp, zhun.zhong@unitn.it, niculae.sebe@unitn.it, satoh@nii.ac.jp

Abstract Reprint. This is an abstract reprint of a journal by [Liu *et al.*, 2023].

Abstract

Overfitting in adversarial training has attracted the interest of researchers in the community of artificial intelligence and machine learning in recent years. To address this issue, in this paper we begin by evaluating the defense performances of several calibration methods on various robust models. Our analysis and experiments reveal two intriguing properties: 1) a well-calibrated robust model is decreasing the confidence of robust model; 2) there is a trade-off between the confidences of natural and adversarial images. These new properties offer a straightforward insight into designing a simple but effective regularization, called Self-Residual-Calibration (SRC). The proposed SRC calculates the absolute residual between adversarial and natural logit features corresponding to the ground-truth labels. Furthermore, we utilize the pinball loss to minimize the quantile residual between them, resulting in more robust regularization. Extensive experiments indicate that our SRC can effectively mitigate the overfitting problem while improving the robustness of state-of-the-art models. Importantly, SRC is complementary to various regularization methods. When combined with them, we are capable of achieving the top-rank performance on the AutoAttack benchmark leaderboard.

References

[Liu *et al.*, 2023] Hong Liu, Zhun Zhong, Nicu Sebe, and Shin'ichi Satoh. Mitigating robust overfitting via self-residual-calibration regularization. *Artificial Intelligence*, 317:103877, 2023.