NEURAL INFORMATION PROCESSING SYSTEMS

Understanding Square Loss in Training Overparameterized Neural Network Classifiers

Tianyang Hu*[1], Jun Wang*[2 3], Wenjia Wang*[2 3], Zhenguo Li[1]

[1]Huawei Noah's Ark Lab, Shenzhen China

[2]Hong Kong University of Science and Technology, Hong Kong, China

[3]Hong Kong University of Science and Technology(Guangzhou), Guangzhou, China

HUAWEI

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

香港科技大學(廣州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Which loss shall we chose for classification task: Cross Entropy Loss or Square Loss?

- Cross Entropy (CE) loss has always been the default choice

- However, cross entropy has several drawbacks： lack interpretability[1], adversarial vulnerability[2], over-confidence[3], …

- Recent experiential evidence[4], [5] shows Square Loss (SL) has better performance in some NLP/CV application

[1] Yu et al. Learning diverse and discriminative representations via the principle of maximal coding rate reduction; NIPS 2020.
[2] Pang et al. Rethinking softmax cross-entropy loss for adversarial robustness; arXiv 2019.
[3] Guo et al. On Calibration of Modern Neural Networks; ICML 2017.
[4] Hui et al. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks; arXiv 2020.
[5] Jacobson et al. Excessive invariance causes adversarial vulnerability; ICML 2020.

We theoretically and empirically investigate SL from following three aspects:

- **Generalization error bound**

  ——The convergence rate on misclassification error

- **Adversarial robustness (margin property)**

  ——The robustness of the decision boundary to perturbation on input data

- **Calibration error**

  ——The distance between the predicted confidence and the underlying condition probability