Weight-pruning optimization is a new learning mechanism for neural networks. By this mechanism, we can train neural networks while keeping it as quantized and sparse ones. However a major challenge of weight-pruning optimization is its memory & computational cost during training. In this study, we developed a novel technology called iterative randomization to greatly reduce the costs. We both empirically and theoretically showed that our technique resolves the memory & computational challenge of weight-pruning optimization. By advancing this study, we will make AI technologies more affordable and energy-efficient.



References

[1] D. Chijiwa, S. Yamaguchi, Y. Ida, K. Umakoshi, T. Inoue, "Pruning randomly initialized neural networks with iterative randomization," Advances in Neural Information Processing Systems 34, 2021.

[2] V. Ramanujan, M. Wortsman, A. Kembhavi, A. Farhadi, M. Rastegari, "What's hidden in a randomly weighted neural network?," in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11893–11902, 2020.

Contact

Daiki Chijiwa / Computer and Data Science Laboratories Email: cs-openhouse-ml@hco.ntt.co.jp