RESEARCH-ARTICLE

An Adaptive Device-Aware Model Optimization Framework

Authors: Shuzhen Zhao, Hongwei Xie, Yan Qiang, Hao Zhang

Authors Info & Affiliations

Publication:
ISICDM 2019: Proceedings of the Third International Symposium on Image Computing and Digital Medicine • August 2019 • Pages 107–112 • https://doi.org/10.1145/3364836.3364858

ISICDM

Deep learning technology has been widely developed in all walks of life, especially in the medical research field. Recently, the deep neural network model has become a deeper and better direction, and followed by the problem of computing resources. The feasibility of a large neural network model can be evaluated by its suitability to sophisticated medical devices. With this basis, we propose an adaptive model optimization framework (AMOF). Compared to reported model compression techniques, we focus on the correlation between channels. AMOF cannot only output an accurate compression ratio, but also search for the optimal pruning channel. Specifically, evolutionary algorithms were introduced on the basis of reinforcement learning. Due to the complexity of a neural network, we propose a co-evolutionary algorithm, so as to guarantee the simultaneous evolution of multiple populations and finally output the optimal cutting channel. Notably, AMOF, combining reinforcement learning and evolutionary algorithm, can ensure the accuracy of this model applied under the full compression

30%; and the accuracy remained at 89.27%. Compared to the reinforcement learning compression method alone, AMOF can increase by 3.5 percentage points in the ResNet20 model.

References

1. Cheng, J., Wang, P. S., Li, G., Hu, Q. H., & Lu, H. Q. (2018). Recent advances in efficient computation of deep convolutional neural networks. Frontiers of Information Technology & Electronic Engineering, 19(1), 64--77.

2. Kim, Y. D., Park, E., Yoo, S., Choi, T., Yang, L., & Shin, D.. (2015). Compression of deep convolutional neural networks for fast and low power mobile applications. Computer Science,71(2), 576--584.

g, Y., Wang, D., Zhou, P., & Zhang, T. 2017. A survey of model compression and acceleration p neural networks. arXiv preprint arXiv:1710.09282.

Show All References

ISICDM

An Adaptive Device-Aware Model Optimization Framework

Machine learning

Machine learning approaches

Neural networks