World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:11, 2024

SAP-Reduce: Staleness-Aware P-Reduce with Weight Generator

Authors : Lizhi Ma, Chengcheng Hu, Fuxian Wong

Abstract : Partial reduce (P-Reduce) has set a state-of-the-art performance on distributed machine learning in the heterogeneous environment over the All-Reduce architecture. The dynamic P-Reduce based on the exponential moving average (EMA) approach predicts all the intermediate model parameters, which raises unreliability. It is noticed that the approximation trick leads the wrong way to obtaining model parameters in all the nodes. In this paper, SAP-Reduce is proposed, which is a variant of the All-Reduce distributed training model with staleness-aware dynamic P-Reduce. SAP-Reduce directly utilizes the EMA-like algorithm to generate the normalized weights. To demonstrate the effectiveness of the algorithm, the experiments are set based on a number of deep learning models, comparing the single-step training acceleration ratio and convergence time. It is found that SAP-Reduce simplifying dynamic P-Reduce outperforms the intermediate approximation one. The empirical results show SAP-Reduce is 1.3× −2.1× faster than existing baselines.

Keywords : collective communication, decentralized distributed training, machine learning, P-Reduce
Conference Title : ICSLP 2024 : International Conference on Speech and Language Processing
Conference Location : San Francisco, United States
Conference Dates : November 04-05, 2024