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

Efficient DNN Training on Heterogeneous Clusters with Pipeline Parallelism

Authors : Lizhi Ma, Dan Liu

Abstract : Pipeline parallelism has been widely used to accelerate distributed deep learning to alleviate GPU memory bottlenecks and to ensure that models can be trained and deployed smoothly under limited graphics memory conditions. However, in highly heterogeneous distributed clusters, traditional model partitioning methods are not able to achieve load balancing. The overlap of communication and computation is also a big challenge. In this paper, HePipe is proposed, an efficient pipeline parallel training method for highly heterogeneous clusters. According to the characteristics of the neural network model pipeline training task, oriented to the 2-level heterogeneous cluster computing topology, a training method based on the 2-level stage division of neural network modeling and partitioning is designed to improve the parallelism. Additionally, a multi-forward 1F1B scheduling strategy is designed to accelerate the training time of each stage by executing the computation units in advance to maximize the overlap between the forward propagation communication and backward propagation computation. Finally, a dynamic recomputation strategy based on task memory requirement prediction is proposed to improve the fitness ratio of task and memory, which improves the throughput of the cluster and solves the memory shortfall problem caused by memory differences in heterogeneous clusters. The empirical results show that HePipe improves the training speed by 1.6×−2.2× over the existing asynchronous pipeline baselines.

Keywords : pipeline parallelism, heterogeneous cluster, model training, 2-level stage partitioning
Conference Title : ICSLP 2024 : International Conference on Speech and Language Processing
Conference Location : San Francisco, United States
Conference Dates : November 04-05, 2024