note of 《SIGMA: A Sparse and Irregular GEMM Accelerator with Flexible Interconnects for DNN Training》

zxp

July 27, 2024

1 background

《SIGMA: A Sparse and Irregular GEMM Accelerator with Flexible Interconnects for DNN Training》 is based on matrix multiplication and is aimed at sparse matrices. Currently, the proportion of reduction is becoming more flexible (different layers in a network may be pruned by 10% to 90%), and the paper points out that existing methods lead to low utilization of memory and computing resources after reduction, and the calculation consumes increasingly larger and larger matrix sizes, and the existing methods are not flexible enough.

1.1 experiment and method

On a GPU, as the sparsity of the matrix increases, efficiency gradually decreases, even only 25%. The paper proposes a structure of flexible size and relies on the network to compute sparse matrices. The paper uses many variable-sized computational units and, by analyzing the characteristics of sparse matrices, arranges the computational units into a network to achieve the computation of sparse matrices. To do this, a mapping is required, and detailed routing analysis is necessary. The paper proposes an adder-tree topology named Forwarding Adder Network (FAN), which seems to determine which part is selected based on priority.
Because it is flexible in structure, the method proposed in the paper will consume less memory and have a higher memory utilization rate.

1.2 evaluation index

The paper considers the matrix scale and sparsity of Transformer, Google Neural Machine Translation (GNMT), Neural Collaborative Filtering (NCF), and

”Baidu DeepBench”. (not actually executed these all network). The layer with the highest computational efficiency reaches 100%.

1.3 conclusion of the paper

Carefully analyzed via a place-and-routed design is needed。

2 Feelings

Didn’t understand much, know too little about network clipping, this paper is sparse matrix related.If we need to cut the network and calculate a particularly large matrix (too big to fit the memory), See the paper