RFC-HyPGCN: A Runtime Sparse Feature Compress Accelerator for Skeleton-based GCNs Action Recognition Model with Hybrid Pruning

Abstract

Skeleton-based Graph Convolutional Networks (GCNs) models for action recognition have achieved excellent prediction accuracy in the field. However, limited by large model and computation complexity, GCNs for action recognition like 2s-AGCN have insufficient power-efficiency and throughput on GPU. Thus, the demand of model reduction and hardware acceleration for low-power GCNs action recognition application becomes continuously higher.

To address challenges above, this paper proposes a runtime sparse feature compress accelerator with hybrid pruning method: RFC-HyPGCN. The hybrid pruning approach includes dataflow reorganization and mixed-grained pruning means. By reorganizing the multiply order, this method skips both graph and spatial convolution workloads. Following spatial convolution’s channel-pruning dataflow, a coarse-grained pruning method on temporal filters is designed, together with sampling-like fine-grained pruning on time dimension. Later, we come up with an architecture with all convolutional layers mapped on chip to pursue high throughput. The scale of storage elements and computing units for each layer are specifically tuned to match the pruned model. To further reduce storage resource utilization, online sparse feature compress format is put forward. Feature is divided and encoded into several banks according to presented format, then bank storage is split into depth-variable mini-banks. In this way, the runtime compress method not only decreases useless storage, but also gets better storage regularity over compressed sparse columns format (CSC). Furthermore, this work applies quantization, input-skipping and intra-PE dynamic data scheduling to accelerate the model. In experiments, pruning method above is conducted on 2s-AGCN, acquiring 3.0x~8.4x model compression ratio and 73.20% graph-skipping efficiency with no accuracy loss. Moreover, our pruning method ensures a hardware-friendly weight-static dataflow. Implemented on Xilinx XCKU-115 FPGA, the proposed architecture has the peak performance of 1142 GOP/s and achieves 9.59x and 2.56x speedup over high-end GPU Nvidia 2080Ti and Nvidia V100, respectively. Compared with latest accelerator for action recognition GCNs models, our design reaches 22.9x speedup and 28.93% improvement on DSP efficiency.