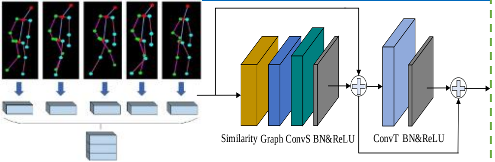
RFC-HyPGCN: A Runtime Sparse Feature Compress Accelerator for Skeleton-Based Action Recognition Model with Hybrid Pruning

Dong Wen, Jingfei Jiang, Jinwei Xu, Kang Wang, Tao Xiao, Yang Zhao, Yong Dou,

National Laboratory of Parallel and Distributing Computing, College of Computer, National University of Defense Technology

Background



2s-AGCN: A skeleton-based graph convolutional neural network for action recognition. A human skeleton is modeled as a graph with 25 points, skeleton graph and global relationship graph are introduced. Graph computation, spatial and temporal convolution, BN and shortcut path are embed in a block. Ten blocks and a FC layer consists the whole network.

Motivation

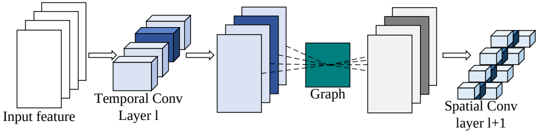
Pose estimation extracting human skeleton features from video stream and actual circumstances. GCN action recognition model depends on such algorithm to provides network input.

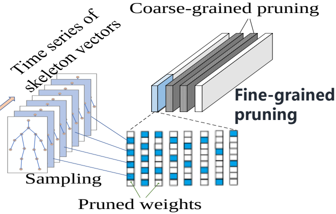
Challenge:

- **Gap of computing performance between fronted-end algorithm and GCN action recognition models.**
- **GCN action recognition models need high-end GPU to deploy, its complexity puts challenge on embedded device.**

Model	Platform	Throughput	Power-efficiency
Mobile-pose	Snapdragon 845	60fps	44.4fps/W
2s-AGCN	Nvidia 2080Ti	28fps	0.11fps/W

Method



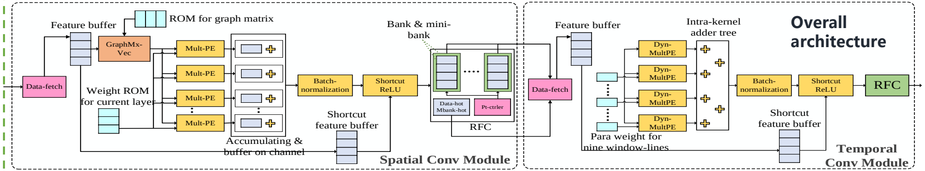


Data reorganization & coarse-grained pruning

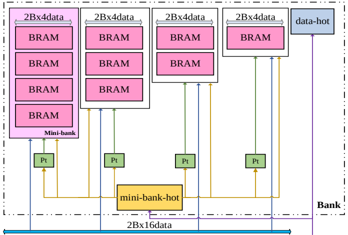
$$X(h,w,oc)=\sum_{i=1}^{ic}\left(\sum_{p=1}^{25} f_{in}(h,p,i)*G(p,w)\right)*W(1,1,i,oc)$$

$$X(h,w,oc)=\sum_{i=1}^{ic}\left(\sum_{p=1}^{25} G(p,w)*f_{in}(h,p,i)*W(1,1,i,oc)\right)$$

Architecture & result



Overall architecture



Compared with former GCN action recognition accelerator

	dsp	bram blocks	LUT	dsp efficiency	peak perf	frequency	fps
ours	3544	1806	176776	0.322GOP/s/DSP	1142GOP/S	172Mhz	271.25
[10]	228	151	44457	0.202GOP/s/DSP	46GOP/S	188Mhz	11.99

Compared with high-end GPU

	ours	2080Ti-original	V100-original	2080Ti(w/o C)	V100(w/o C)	2080Ti-skip	V100-skip
throughput	271.25	29.53	69.38	45.42	98.87	104	199.09
speed-up		9.19	3.91	5.97	2.74	2.61	1.36

Runtime feature compress module