MSc(CS) Dissertation Public Seminar

****

Title: Optimize the Inference of Transformer-based Models on DaVinci AI Chips

Speaker: He Debin

Date & Time: April 15 2021, Thursday, 01:30pm

Zoom Meeting Link: https://hku.zoom.us/j/95488647812

Meeting ID: 954 8864 7812

Password: 976970

Abstract:

The transformer is the most critical algorithm breakthrough in the field of Natural Language Processing. DaVinci is a powerful architecture for neural network computing designed by Huawei. It provides huge computing resources. However, there are few studies of optimizing inference latency on the DaVinci chips. In this work, firstly, we profile the BERT on DaVinci chip. Based on the profiling result, we analyze the key bottleneck of the BERT on the DaVinci chips. There are two key bottlenecks: (1) memory-bound operators cost the most inference time (2) the utilization of cube unit is low. Secondly, based on the profiling results, we propose the kernel fusion framework for the DaVinci chips.

There are two main components: (1) special fusion rules matching and (2) general fusion rules searching. Experiment on kernel fusion demonstrated about 2x speedup. With the kernel fusion framework, it brings the performance gains and more deep learning model can be turned into reality.

About the Speaker:

He Debin is currently a full-time MSc(CS) student of the Department of Computer Science in the University of Hong Kong. His supervisor is Prof. CL Wang.

All are welcome!

Tel: 3917-1828 for enquiries

MSc(CS) Dissertation Public Seminar

****

Title: Learning Voxel Feature Encoding for 3D Object Detection

Speaker: Jia Gaoguo

Date & Time: April 15 2021, Thursday, 02:15pm

Zoom Meeting Link: https://hku.zoom.us/j/95488647812

Meeting ID: 954 8864 7812

Password: 976970

Abstract:

LiDAR-based 3D object detection plays an important role in auto driving. OpenPCDet is a toolbox of 3D object detection methods and it supports the fixed-size voxel encoding methods.

However, the fixed ratio of voxels and the limited number of points lead to information loss inevitably. The LearningVoxel is an extension of OpenPCDet, which is used for 3d detection from Lidar points cloud. It solves the problem of the voxel size, which supports dynamic voxel sizes, multi-view voxel fusion, and Hybrid voxel sizes. The extension is based on C++ and I adjusted the original OpenPCDet to fit the requirement of data. The main contribution of the extension is that it provides a platform to test different voxel encoding methods. Experiments of some voxel encoding methods show that The Dynamic Voxel has the highest efficiency and takes the least time, which is around half of the time cost of PointPillars. The Multi-view Voxel achieves the best result, and it is a good balance of efficiency and accuracy.

About the Speaker:

Jia Gaoguo is currently a full-time MSc(CS) student of the Department of Computer Science in the University of Hong Kong. His supervisor is Dr. Ping Luo.

All are welcome!

Tel: 3917-1828 for enquiries

MSc(CS) Dissertation Public Seminar

****

Title: A Packet Level Simulator for Routerless Network on Chip

Speaker: Meng Yizhuo

Date & Time: April 15 2021, Thursday, 03:00pm

Zoom Meeting Link: https://hku.zoom.us/j/95488647812

Meeting ID: 954 8864 7812

Password: 976970

Abstract:

As technologies continue to advance, chip multi-processor (CMPs) or multi-core architecture have become mainstream in the processor market. Whereby multiple processing units are integrated onto a single monolithic integrated circuit or multiple dies in a single packet. So, efficient and reliable communication among several chips must be precisely designed and implemented.

There are two kinds of network-based connections including Router-based and Routerless Network-on-Chip (NoC). We focus on the latter one, which is capable of terminating the costly routers used in the Router-based NoC and places them with multiple rings across cores to achieve high scalability. However, different ring algorithm and configurations are hard to be tested on Routerless NoC because it is usually implemented by Field-programmable gate array (FPGA). So for the first part of this study, we design and implement the first Routerless NoC Simulator according to the newest research. Meanwhile, we optimize buffer management of its network interface and routing algorithm. At last, various experiments will be shown and packet level performance, latency and throughput, will be evaluated, which indicates the validity and high performance of our simulator.

About the Speaker:

Meng Yizhuo is currently a full-time MSc(CS) student of the Department of Computer Science in the University of Hong Kong. His supervisor is Prof. Lawrence Yeung.

All are welcome!

Tel: 3917-1828 for enquiries

MSc(CS) Dissertation Public Seminar

****

Title: Accelerating Deep Neural Networks Inference on Atlas200 DK

Speaker: Zhang Haoxin

Date & Time: April 15 2021, Thursday, 03:45pm

Zoom Meeting Link: https://hku.zoom.us/j/95488647812

Meeting ID: 954 8864 7812

Password: 976970

Abstract:

More and more deep learning applications have begun to be deployed on the mobile devices, companies in the industry have proposed chips specifically for AI computing, Huawei’s DaVinci architecture is one of them. In order to deploy real-time deep learning applications on mobile devices with limited computing power and power, it is necessary to accelerate its inference performance as much as possible. This dissertation explored the performance of matrix multiplication and convolution operators on DaVinci AI Core, and explored the impacts of tiling, double buffering, and core parallelism on its performance, the best optimization gain reached 48.5%. It also studied the impact of quantification and kernel fusion on the performance of object detection network and face recognition applications deployed on DaVinci AICore, the best performance gain reached 223%.

About the Speaker:

Zhang Haoxin is currently a full-time MSc(CS) student of the Department of Computer Science in the University of Hong Kong. His supervisor is Prof. CL Wang.

All are welcome!

Tel: 3917-1828 for enquiries