CNN 추론을 위한 이기종 컴퓨팅 장치의 조화

이제민
한국전자통신연구원 지능형반도체연구본부
e-mail : *leejaymin@etri.re.kr*

The Harmony of Heterogeneous Computing Units in CNN Inference

Jemin Lee
AI SoC Research Division
Electronics and Telecommunications Research Institute

Abstract

We investigate current aspects of commodity NPUs and discover their limitations of NPUs. As a result, NPUs could not support all kinds of operations from neural networks because NNs have rapidly evolved. To remedy the challenge, we present a concept for a harmonious system.

I. Introduction

Convolutional neural networks (CNNs) have changed the technology landscape, ranging from computer vision to machine translation. The resource requirements prevent CNNs from being used in mobile and embedded devices because recent CNNs require huge computational resources to handle hundreds of megabytes of weight parameters. To afford costly CNNs on resource-constrained mobile devices, many researchers have tackled this challenge by trimming down CNN complexity with various compression techniques and accelerating CNNs using special hardware such as neural processing units (NPU). Big-tech companies have already rolled out their products for CNN acceleration such as Google TPU [1], Intel Movidius NCS [2], NVIDIA-NVDLA [3], and Huawei-Kirin [4].

In this paper, to investigate the limitations of these commodity NPUs, we perform testing on NVIDIA-NVDLA and VTA (similar to Google-TPU) [5]. As a result, NPUs could not support all kinds of operations from neural networks. To support diverse CNNs and augment the synergy of heterogeneous units, we present a harmonious system that automatically optimizes operation schedules for CPUs, GPUs, and NPUs. Figure 1 shows an overall concept of harmony. Towards efficient inference of modern CNNs, the harmony takes into account the following: i) an optimal intermediate representation (IR) generator for orchestration of heterogeneous computing units, ii) a method of scheduling primitive operations, and iii) precision-aware computing to support diverse data types on NPUs.

We believe that our concept could make deep learning-related tasks on a local device faster and more power-efficient.

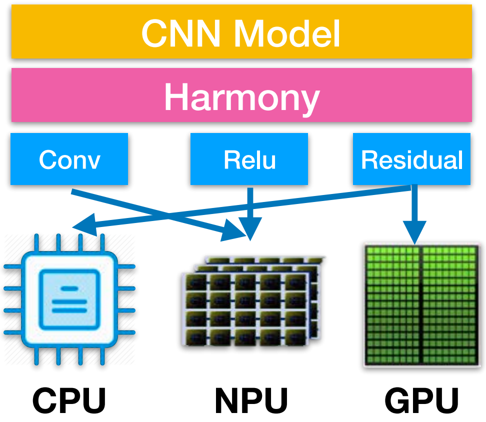

Figure 1 Overview of the harmony

II. Experimental Results

Table 1 shows inference time by running ResNet models on different NPUs. Considering low power consumption, NPUs achieved comparable performance to server-side GPGPU. However, NPUs do not completely support all operations of ResNet because it contains a residual layer that skips a few layers, unlike trivial CNNs. It leads to a necessity to use general-purpose computing units (CPU and GPU). Therefore, in order to support the latest CNNs such as ResNet, NPU system should be combined with general-purpose processors.

Table 1 Inference results of ResNet models on Jetson Xaiver and VTA PYNQ-Z1 boards

Computing Unit	NN Model	Precision	Power	Latency
NVDLA + Volta GPU	Resnet 50	FP16	30W	0.048 s
VTA + Cortex-A9	Resnet 18	INT8	2.5W	0.50 s

Ⅳ. Conclusion

We discovered the limitation of commodity NPUs. The limitation was that all kinds of operations from neural networks were not executed in NPUs. For end-to-end inference, traditional computing units like CPU or GPU should be involved. For the heterogenous computing, we presented the concept of the harmonious system. Also, we are planning to implement our concept in real hardware.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2018-0-00769,Neuromorphic Computing Software Platform for Artificial Intelligence Systems)

참고문헌

[1] Jouppi, Norman P., et al. "In-datacenter performance analysis of a tensor processing unit." Proceedings of the 44th annual international symposium on computer architecture. 2017.
[2] https://software.intel.com/en-us/movidius-nc
[3] http://nvdla.org/
[4] http://www.hisilicon.com/en/Solutions/Kirin
[5] T. Moreau, T. Chen, and L. Ceze. Leveraging the VTA-TVM hardware-software stack for FPGA acceleration of 8-bit ResNet-18 inference. In ReQuEST-DNN, 2018.