TOP 10 RESEARCH ACHIEVEMENTS

GANPU: An On-Device Training Processor for Generative Adversarial Networks

Department
School of Electrical Engineering

Principal Investigator
Hoi-Jun Yoo

Homepage
http://ssl.kaist.ac.kr

This research is regarding an energy-efficient AI processor for generative adversarial networks (GAN), targeting inference and training of the models on mobile platforms. GANs are highly praised for their ability to generative new data, but they not only require larger number of computations, but also consist of multiple networks, making it difficult to optimize the hardware architecture. This research proposed a reconfigurable processor architecture to efficiently handle multi-network and a method to accelerate sparse convolution. It achieved 4.8 times higher energy-efficiency compared to the previous state-of-the-art, enabling on-device AI on performance-limited and battery-limited mobile devices.

1. Background

Generative Adversarial Networks (GAN) can generate and recreate new images, so they are used in a wide range of applications, from image style transfer to synthetic voice generation. Deepfakes, which have become a big social issue by synthesizing other people's faces on top of the existing videos, is also a GAN-based technique. GANs can be used in various applications of mobile devices where a lot of video and photo contents are produced as well as consumed. Therefore, they have attracted great attention not only from academia but also from industry.

However, GANs have complex algorithmic architectures, incorporating multiple Deep Neural Networks (DNN) to be trained in a single model. Moreover, each individual DNN in a GAN have different characteristics, making it difficult to optimize the accelerator architecture. In addition, GANs require more computations than conventional AI models to generate high-resolution images due to the high video fidelity requirement of the recent displays and image sensors. Therefore, until this research, mobile devices were regarded unsuitable to implement GANs on due to limited speed and power. Moreover, there are even greater limitations to realize on-device training for advanced GAN operations.

2. Contents

Most of the previous DNN accelerators only supported inference. Although a few DNN training accelerators have been introduced, they only supported a single DNN. In this research, we proposed the world's first low-power GANPU (Generative Adversarial Networks Processor Unit) capable of not only inference but also training on mobile devices and supporting multi-DNN workloads such as GANs.

More specifically, adaptive spatio-temporal workload multiplexing (ASTM) is proposed for versatile accelerator architecture which can be rearranged based on variable multi-DNN workloads. ASTM enables

efficient allocation of limited resources such as external memory bandwidth and computational resources. Also, there are lots of zeros in the data due to the nonlinear function of the DNN. By designing a processor architecture that skips input·output activation sparsity, the speed and energy efficiency in the inference and training are maximized. The proposed architecture achieved a 28.53× throughput increase when input·output activation sparsity is 90%.

The GANPU with the above technology achieved 4.8 times higher energy efficiency than the previous state-of-the-art training accelerator. Additionally, a face modification system built based on GANPU has been implemented and demonstrated. The application allows users to directly modify face images taken from a tablet camera by adjusting facial features such as hair, glasses, and eyebrows.

(Demonstration Video – https://www.youtube.com/watch?v=HnNWsgqkEU0)

3. Expected effects

GANPU is the first silicon implemented DNN accelerator which support both inference and training of multi-DNNs on a single chip. Through its high energy efficiency, it enables the implementation of GANs on mobile devices, which can perform new applications of AI such as face modification, voice generation, and video synthesis. In addition, the proposed AI chip preserves privacy by eliminating the need for sending user-specific data to datacenter through improving the AI model's performance on-device. The proposed research can be the stepping stone of opening up a new era of creative AI on user's edge devices.

Figure 1. Demo System with GANPU

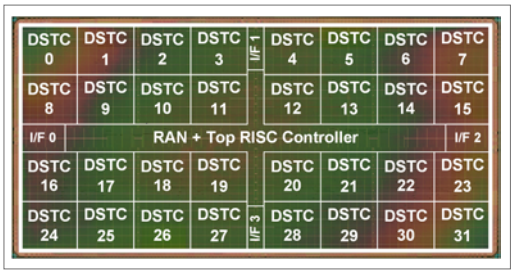

Figure 2. Chip Photo

Research outcomes

[Paper] S. Kang, D. Han, J. Lee, D. Im, S. Kim, S. Kim, H-J. Yoo, "GANPU: A 135TFLOPS/W Multi-DNN Training Processor for GANs with Speculative Dual-Sparsity Exploitation", 2020 IEEE International Solid-State Circuits Conference (ISSCC 2020).
S. Kang, D. Han, J. Lee, D. Im, S. Kim, S. Kim, J. Ryu, H-J. Yoo, "GANPU: A Versatile Many-Core Processor for Training GAN on Mobile Devices with Speculative Dual-Sparsity Exploitation", 2020 IEEE Hot Chips Symposium.

[Patent] Domestic patent · US patent application completed

[Award] 26th Samsung Electronics Human Tech Thesis Award - Gold Prize Award
2020 Microsoft Research Asia Fellowship Award

[Press release] YTN news report, "Development of self-drawing AI semiconductor chip"
https://www.ytn.co.kr/_ln/0115_202004070308303992
Reported on about 30 domestic media including Munhwa-Ilbo and Donga-Science
http://www.munhwa.com/news/view.html?no=2020050401031805000002&mobile=false