Techology Trend of Edge AI

Yen-Lin Lee, Pei-Kuei Tsung, and Max Wu
MediaTek Inc.
{yenlin.lee, pei-kuei.tsung, max.wu}@mediatek.com

***Abstract* — Artificial intelligence (AI), defined as intelligence exhibited by machines, has many applications in today's society, including robotics, mobile devices, smart transportation, healthcare service, and more. Recently, lots of AI investment in both big companies and startups have launched. Besides cloud-based solution, AI on the edge devices (Edge AI) takes the advantages of rapid response with low latency, high privacy, more robustness, and better efficient use of network bandwidth. To enable Edge AI, new embedded system technologies are desired, including machine learning, neural network acceleration and reduction, and heterogeneous run-time mechanism. This paper introduces challenges and technologies trend of Edge AI. In addition, it illustrates edge AI solutions from MediaTek, including the dedicated AI processing unit (APU) and NeuroPilot technology, which provides superior Edge AI ability in a wide range of applications.**

I. Introduction

In the recent years, artificial intelligence (AI) appears in every technology field. From home electronics to the complex simulation experiment for protein structure, AI or machine learning has been launched to enhance the quality of computation and create possibilities for new applications, such as face unlock of mobile phone or autonomous driving. However, high performance of machine learning or deep learning requires huge computation capability to deal with complex training and inference methodologies and large dataset [1]. That is, in order to satisfy the computation thirst, cloud servers have to provide very powerful computational capabilities. Hence, more and more non-traditional alternative solutions have shown up in recent years to effectively execute the AI computation tasks. For example, Google provides the tensor processing unit (TPU) solution as the specialized computing unit for AI processing tasks [2]. NVidia also innovates new GPU server architecture to favor AI characteristic [3].

The cloud-based eco-system has demonstrated itself as a practical platform to serve some AI applications. However, the cloud-based solution has many limitations that might prevent the adoption on all AI applications. Taking the autonomous driving for example, the connection robustness and its latency from the server seriously impact the safety of the vehicle due to the time-to-collision. In addition, uploading the personal information or the record of street-view video to cloud brings the privacy issue. Furthermore, there is not always existing internet connectivity everywhere. These issues lead to the requirement that AI computation must be on the edge devices (Edge AI). In this paper, design challenges and technology trend of Edge AI are discussed, and how MediaTek overcomes the challenges stated above is also introduced. By developing the dedicated AI processing unit (APU) and the NeuroPilot technology, MediaTek provides ready-to-product solution for Edge AI.

Fig. 1. Edge AI opportunities different from cloud-based AI

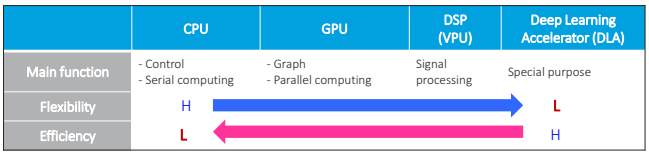

Fig. 2. Processors comparison for AI processing

The rest of the paper is organized as following: First, the design challenges and current technology progress for Edge AI is illustrated in section II. Then, MediaTek approaches for Edge AI is discussed in section III. Finally, section IV concludes this paper.

II. Edge AI Design Challenges and Technologies

Figure 1 describes the opportunities and key requirements of Edge AI computing comparing to the challenges of cloud-based AI frameworks. For different applications, there are different critical requirements including latency, efficiency, availability, privacy, and so on. For the vehicles or drones, moving speed limits the latency tolerance of response. Otherwise, a crash or accident will happen. Home assistant systems or devices always bring privacy concerns because processed content touches on personal information. Furthermore, power efficiency is extremely important for Edge AI devices, especially for wearable, in order to have longer duration usage. All these needs make the Edge AI computing necessary and have brought it to the forefront. However, Edge AI also has its own design challenges that need to address:

A. Power Efficiency and Different Types of AI Processors

Fist and the most important challenge is that the edge devices have to provide enough computational capacity within specific limitations, such as thermal or form factor size. Due to these limitations, the Edge AI prefers to focus on the inference part and leave the training stage in the cloud as usual. The inference computation in Edge AI can be handled by various computation units inside the device. Figure 2 shows the various embedded processors for AI computing and their

978-1-5386-4260-3/18/$31.00 ©2018 IEEE

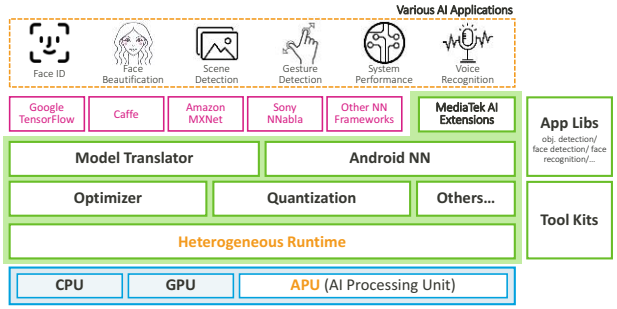

Fig. 3. MediaTek NeuroPilot framework introduction

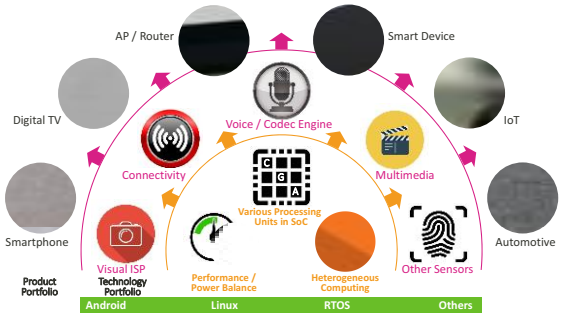

Fig. 4. MediaTek provides broad-range electronic devices with the latest AI technologies running on different platforms and OSes

pros and cons. Although CPU and GPU have been used for AI process as computation resources for a long period of time, the general purpose architecture still causes higher power consumption. For those applications that only require short burst of computation, the general purpose processor can provide enough performance and take the advantage of time-to-market. However, for long duration or sustainable scenarios, such as AI post-processing on social video streaming, energy efficiency becomes very important and necessary. The specialized processor, like DSP-based processor, has been adopted to have better power efficiency for specific application with less flexibility. For computation and power efficiency, fixed-point format is widely used, and the bit-width might be configured to 8 bits or even lower in DSP [4]. Today, deep learning accelerator (DLA) starts to be used in edge devices. DLA can achieve the highest power efficiency by implementing the key computation operations as hardwired logic [5].

B. *Computational Complexity and Efficiency*

For Edge AI applications, the computational requirements are represented in two different ways. First, the algorithm throughput needs to be constantly sustainable period to meet the real-time constraint without frame drop. Second, the processing latency should be low enough going through overall algorithm pipeline. Taking autonomous driving as an example, the latency of video processing should be less than 100ms for safety [6].

Challenges lead to the innovation of AI processor architecture approaches. The conventional parallel computing architecture usually leverages multiple threads to hide memory access latency. New on-chip memory architecture and data flow management needs to be re-considered because the latency cannot be simply hidden in the AI computation. In order to have better throughput on pre-defined hardware, SoC vendors usually provide efficient pre-built libraries for the key AI computation operations. Getting benefit from AI libraries, developers can reduce lots of effort with programing but still get good performance as close-to-metal coding.

C. *Privacy/Security*

Edge device without any internet connection has high privacy protection. However, most consumer electronic devices need to support on-line applications other than the Edge AI application only. Therefore, it is necessary to enable the dedicated security zone to protect the private information, such as finger print, voice, and face recognition data. The security zone might require isolated hardware or software environment to avoid pollution.

III. MediaTek Solutions for Edge AI

Facing the Edge AI challenges and opportunities, MediaTek prepares the corresponding solution. As shown in Fig. 3, MediaTek provides both hardware and software environment to optimize Edge AI performance [7]. First of all, the dedicated APU is designed to have better power efficiency. Comparing to CPU operation, up to 95% of energy consumption can be eliminated. Second, the heterogeneous runtime in NeuroPilot software development kit (SDK) manages the task scheduling among the CPU, GPU, and APU. Moreover, NeuroPilot supports current state-of-the-art AI frameworks, including Caffe, Tensorflow, MXNet and NNabla. The toolchains including model translator in NeuroPilot allow programmers to enable AI application on devices. Finally, the application libraries are offered to fast connect the PC-based prototyping to the close-to-metal performance in the Edge AI development. As the result, Fig. 4 shows the application scope with MediaTek solution. With the core technology supporting, Edge AI can be realized in a wide range of applications.

IV. Conclusion

In this paper, the design challenges and technology trend in Edge AI are introduced. Compared to cloud-based AI, Edge AI takes the advantage of low latency and privacy protection. Power efficiency and computation efficiency become the necessary requirements for edge AI. Finally, the MediaTek solutions is proposed. The dedicated APU design and NeuroPilot SDK enable Edge AI ability in a wide range of applications.

Reference

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning.” In Nature, 2015.
[2] Norman P. Jouppi, et. al., “In-Datacenter Performance Analysis of a Tensor Processing Unit.” arXiv:1704.04760
[3] “NVidia Tesla V100 GPU Architecture” Nvidia website, 2017
[4] M. Courbariaux, et. al., “Training deep neural networks with low precision multiplications.” arXiv:1412.7024
[5] Y.-H. Chen, et. al., "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks". IEEE International Solid-State Circuits Conference, ISSCC 2016
[6] S. Mochizuki et. al.,“A 197mW 70ms-Latency Full-HD 12-Channel Video-Processing SoC for Car Information Systems”. IEEE International Solid-State Circuits Conference, ISSCC 2016
[7] https://www.MediaTek.tw/features/artificial-intelligence