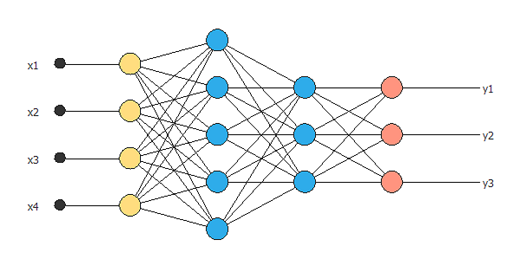
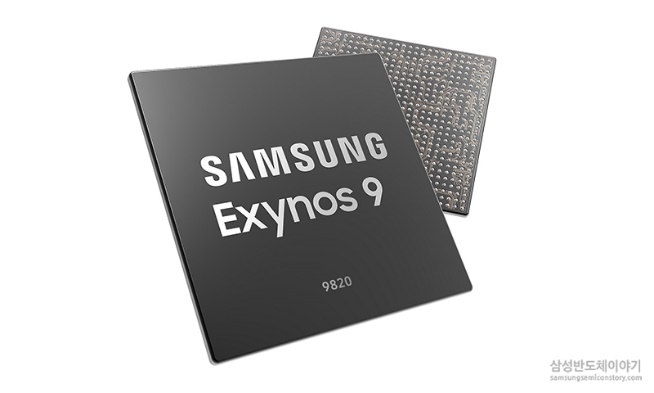
**2020 Multimedia&Labs HW#1**

**Dept: Software**

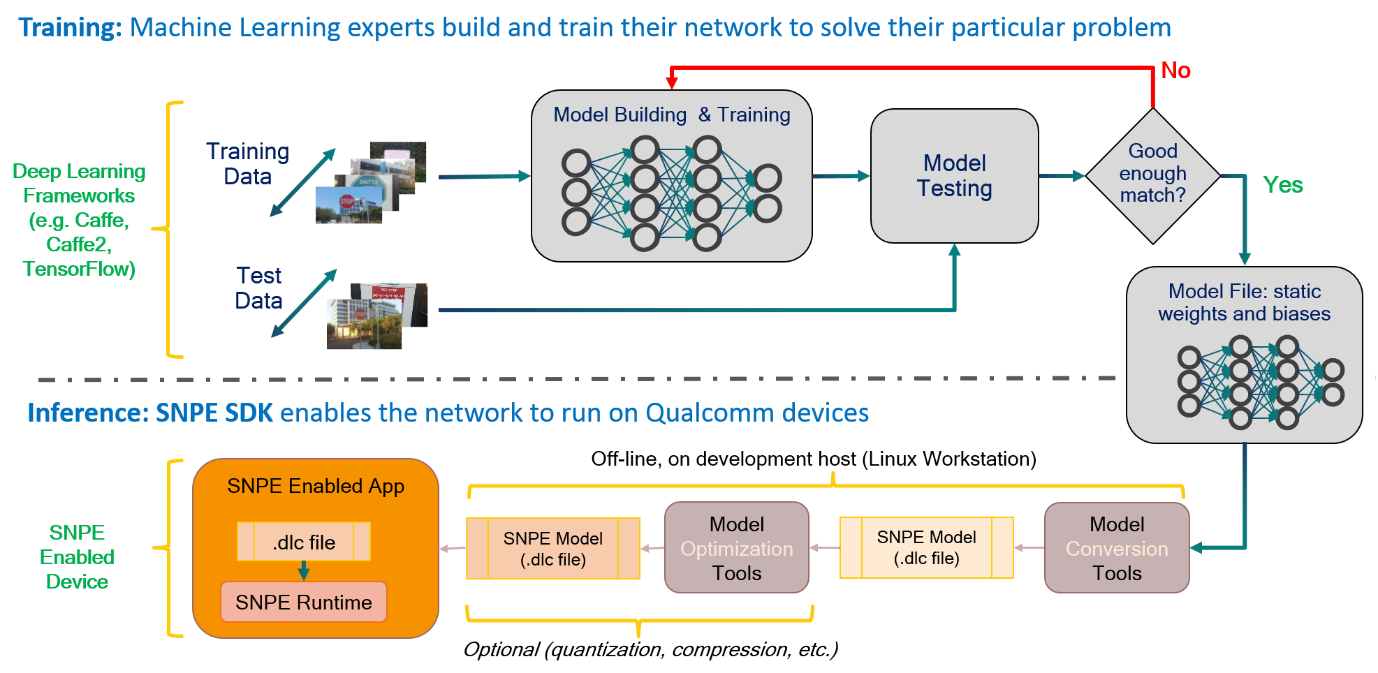
**ID/Name: 201533661 이승수**

AI(Artificial Intelligence) accelerator is a class of specialized computer system designed to accelerate complicated operations on AI applications like artificial neural networks and machine learning. It’s typical usage is for robotics, IoT, and other data-intensive or sensor-driven tasks. There are GPU, DSP, and NPU as a type of AI accelerator. GPU(Graphic Processing Unit) process image and graphic processing DSP(Digital Signal Processor) compress analogue signal into digial signal efficiently.

NPU(Neural Processing Unit) is a chipset used to perform neural network processing. Present mobile smart phone market is era of on-deviced AI products. NPU realize AI functions like Google ‘NightShot’ can be processed on mobile device. Apple’s iPhone XR use NPU, named A11 bionic, on smart phone first and it’s the trigger of NPU attached mobile phones. Major smart phone producers develop their own NPU. Apple have A11/A12 bionic, Hwaweii have Kirin 970/980, and Samsung mobile phone use Exynos 9820. At this point, we have to know why smart phone use NPU on mobile device.

[ Neural Network(left) and NPU(right) ]



[ phase of DNN based AI processing with SNPE model ]

Basic machine learning phase can be devided into training and inference phase. At Training phase, we build learning model and train the model with training dataset. At DNN(Deep Neural Networking) model, accuracy of model improved itself by back-propagation. After model is trained, Inference phase can be performed on mobile device. At IoT device, processer ‘decide’ solution for decisions required at data from sensors, mike, and camera.

Usually, Training phase is performed with TPU at cloud/data center and Inference phase is performed at NPU on device. on device Inference phase have advantage at service delay and security problem. Because it doesn’t need to transmit data on network. And it can reduce storage using fee because it doesn’t need cloud storage.

Qualcomm’s SNPE(Snapdragon Neural Processing Engine) is SW framework for each cores on SoC(GPU, DSP, CPU) can executed. It is composed by converting tool that transplant existing DNN into Qualcomm SoC, and optimizing and debugging tool for each core. Developers can use SNPE SDK to optimize their applications at heterogeneous environment.

NPU’s appearance on smart phone is because of upward-leveled demand of customers for mobile phone’s functions. NPU process operations for machine-learning arithmetic operations at device. We have used TPU(Tensor Processing Unit) at cloud-level to perform complicated processing, but it needs network, and it becomes bottle-neck at processing time. As a result, for fast bio-certification like faceID(Apple) or Imoge(Samsung), We need NPU at smart phone. It’s used for not only personal certification, but to image revision. NPU used to moderate dynamic range more naturally and 3D-scanning.

|  |
| --- |
|  |

Another usage of NPU on smart phone is Real-time video rendering to object recognization. Famous camera application like ‘snow’(Line), and ‘cheese’(Kakao) use NPU for motion recognition. NPU is used to decide situation like in-door/out-door and performing HDR, too.

[Real-time video object recognintion at Hwaweii Kirin 980]

NPU’s roll at future will be more bigger because of expansion of IoT products, and auto-driven car. NPU have advantage at security and throughput then TPU because NPU process calculations at AI on device, so it doesn’t need help of network. So NPU market is expected to be expand more at recent future.

**[References]**

About Deep Learning Model:

<https://developer.qualcomm.com/docs/snpe/overview.html>

<https://m.post.naver.com/viewer/postView.nhn?volumeNo=11764014&memberNo=20717909>

About AI accelerator:

<https://en.wikipedia.org/wiki/AI_accelerator>

About NPU:

<https://byline.network/2018/09/14-28/>

<https://byline.network/2018/11/20-32/>