End to End: Part 5

Earl Wong

Subjects

- Optics
- Sensor
- **ISP**
- GPU
- **NPU**

Different Naming Conventions

- ANE (Apple) - Apple Neural Engine
- AI Engine (Qualcomm) - Hexagonal Tensor Processor
- APU (Mediatek) - AI Processing Unit
- NPU (Samsung) - Neural Processing Unit
- TPU (Google) - Tensor Processing Unit

Objective

- Deliver fast, low latency inference throughput under strict power and thermal constraints.

How

1) Dedicated ML operations

- Convolution (CNN’s)
- Matrix Multiplication (MLP’s)
- Activation (Non Linear Functionality)

2) Minimize DRAM fetch

Details

- One big MAC that fuses a series of CNN architecture operations - convolution, mlp, activation.

- Like a multi-stage pipeline, etc. without the need to frequently access DRAM.

Example

- Insert here.

CNN vs Transformer
On NPU

- Answer the question: Why are transformers not well suited for the current accelerators?

NPU vs GPU

- Why NPU / Why not GPU?
- Pro’s and cons.
- Essentially, GPU was designed for gaming, while NPU was designed for CNN architecture based ML inference.
- GPU = high power and data fetches from L1, L2 caches and DRAM.
- NPU = low power and input / output type architecture.