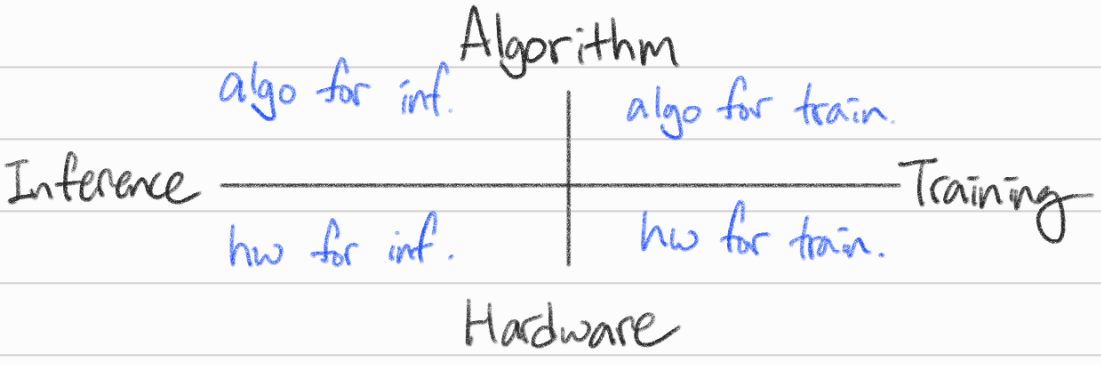
Efficient Methods and Hardware for Deep Learning

- Deep Learning is changing our lives
- Models are getting larger

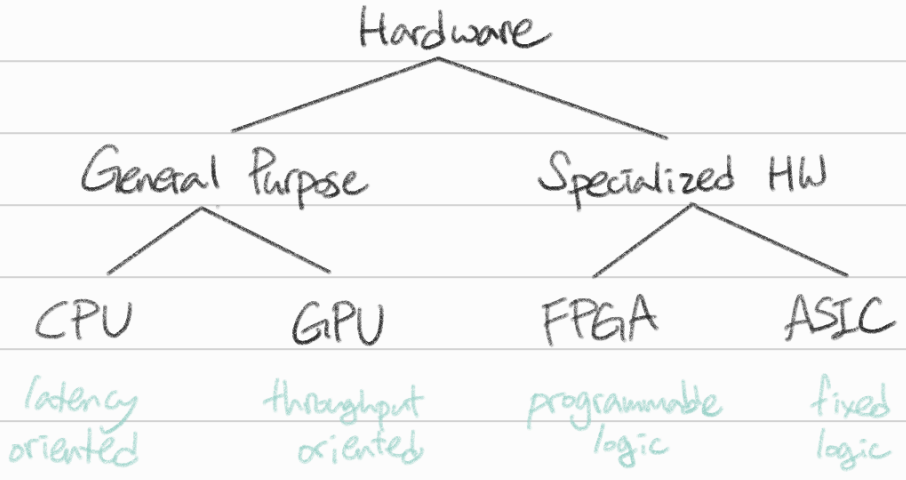
- The first challenge: Model Size - Hard to distribute large models through over-the-air update
- The second challenge: Speed - Such long training time limits ML researcher's productivity
- The third challenge: Energy Efficiency

Larger model → more memory reference → more energy

Agenda



Hardware 101 : the family



Algorithms for Efficient Inference

1. Pruning : redundant parameter 제거 ⇒ Computation ↓
   pruning + retraining iteration 통해 accuracy ↑

2. Weight Sharing : 근사해 single number 사용
   2.09, 2.12, 1.92, 1.87 ⇒ 2.00
   Huffman Coding

3. Quantization (양자화) : run the NN and train it with the normal floating point numbers.
   quantize the weight and activation

4. Low Rank Approximation (행렬분해)

5. Binary / Ternary Net

6. Winograd Transformation

0 weight 는 skip

Hardware for Efficient Inference

Eyeriss
DaDianNao
TPU
EIE

Efficient Training Algorithm

1. Parallelization

2. Mixed Precision with FP16 and FP32

3. Model Distillation

4. DSD : learn the trunk first
then learn the leaves

Hardware for Training Algorithm