Neural Optimization

부동 주제

Basic-Optimization for DL

How to measure the KPIs of a Model

- How to Calculate Model Metrics
  - FLOPs
    - 분해는 1s에 얼마나 많은 floating point 연산관련 지표
    - 딥러닝에서는 단위 시간이 아닌 절대적인 연산량의 횟수
      - Pulling
      - Fully Connected Layer — i x j
      - Convolution Layer — N(kernel) x K x K x Cin x Cout x Hout x Wout
        - H, W, C
      - Activation Function — n
  - MACs — Multiply-Accumulate Computations (1 Addition + 1 Multiple)
    - a <-- a + (b x c)
    - Activation Function
    - 들어가기 전 모습
    - 1MAC ≈ 2 FLOPs
  - FLOPS — Number of Operations per Second
- Practice(Utils)
  - Get the Size of Model
  - Get the Inference time
  - Calculate the size of the model files
  - Calculate FLOPs, FLOPS, and MACs

Separable Convolutions

- Normal Convolution — 5 x 5 x 3 x 256 x 8 x 8 = 1,228,800
- DepthWise Convolution — 5 x 5 x 3 x 8 x 8 = 4800
- PointWise Convolution — 256 x 1 x 3 x 8 x 8 x 1 = 49152

Model Pruning

- Structures
  - Unstructured
  - Structured

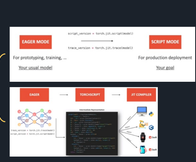
L1, L2 Norm을 이용하여 mask를 만들어 사용한다

- Criterion
  - Weight and Gradient Magnitude
    - L1 norm: |W| = sigma(W)
    - L2 norm: |W|^2 = sigma(W**2)
  - Global and Local
    - 각 layer에서 일정 퍼센티지 만큼 줄인다면 local
    - 전체 layer에서 일정 퍼센티지 만큼 줄인다면 global
- Technique
  - Classic
  - Learning Rate Rewinding
  - Soft Pruning
  - Initialization Pruning
  - Sparse Training
  - Mask Learning
  - Penalty-Based Methods
- Problem — Pytorch can't make sparse model but Keras can do. That doesnt mean this is better. So Find way to How to make Sparse Model

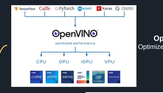
Model Deployment

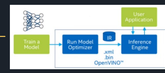
4 Types of Deployment Hardware
CPU, GPU, VPU, FPGA

- Prototyping
- Industrializing — Optimization & Industrialization
  - Use Pytorch C++ Library to read model fast
  - Model Inference with Pytorch
    1. **Evaluation Mode**
       * model.eval()
    2. **Autograd Off**
       with torch.no_grad()
       ....
    3. Convert the model to **TORCHSCRIPT**
       Uses the Just In Time (**JIT**) Compiler
       torch.jit.save(torch.jit.script(model), "model.p")



  - ***Using Platform***
    It's upon the **Device**
    - OpenVINO
      Optimized Performance





서브 주제

    - Deep Stream

Advanced Tactics

11 Advanced Tactics to Bosst up!

Weight Quantization

An ***Int model*** is:
* 4x smaller (model size)
* 2~4x faster (inference speed)
* 2~4xlighter (memory bandwidth)

5 Things to know about Quantization
- Inference Only
- Not every layer can be quantized
- Not every layer should be quantized
- Not every model reacts the same way to quantization
- Most available implementations are CPU only

6 Steps for Quantization
- Module Fusion

- Formula Definition

$$Q = \ lnt(\frac{-3.57}{S}) + Z$$

$$Z = 2^b - Int\left(\frac{\alpha}{s}\right)$$

$$S = \frac{\beta - \alpha}{2^b - 1}$$

$$r\left(Q\right) = Clip\left(\ lnt\left(\frac{Q}{S}\right)\right) - Z$$

$$Q\left(r\right) = Clip\left(\ lnt\left(\frac{r}{S}\right)\right) + Z$$

Q: Quantized value
R: real value
S: scale
Z: zero-point integer
Clip: range(0~255 ...)

- Hardware Deployment — Choose your CPU Intel vs ARM
- Dataset Calibration

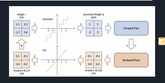
- Weight Conversion

3 Quantization

- Dynamic vs Static Quantization
  - In dynamic quantization, the clipping range is dynamically calculated for each activation map during runtime
  - In static quantization, the clipping range is pre-calculated and static during inference
- Symmetric vs Asymmetric Quantization
  - symmetric quantization: [-3.0 3.0]
  - asymmetric quantization: [-3.0 3.8]
  - Z = 0

Figure 3: *Illustration of symmetric quantization and asymmetric quantization. Symmetric quantization with restricted range maps real values in [-127, 127] and full range maps to [-128, 127] for 8-bit quantization.*

- Training-Aware Quantization — Retraining Model with quantized parameters



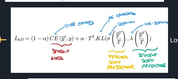
- Other techniques

Knowledge Distillation

- What is Soft Labels?
  - Hard vs Soft Targets
  - In soft labels, we want to use soft targets

  - SOFTMAX vs Temperature SOFTMAX to Dark Knowledge

- How does training happen?
  - Loss Function
    - What is KL Divergence? difference between the CE and E 즉 선생과 학생 사이의 확률적 분산을 측정하는 것
    - Log SOFTMAX vs SOFTMAX



- Is this the Only type of Architecture?