Midterm Review

General Information

- In-class exam on April 2nd 5:10pm-6:10pm
- 1 hour
- Close book, bring calculator
- I will join online, TAs will invigilate

Recap (Lecture 1: P1-P27)

- Neural Network Basics
  - MLP
  - Forward and backward propagation of MLP
  - Weight decay, dropout
  - The training optimizer: SGD, RMSProp, Adam
  - Multistage learning rate scheduler

Recap (Lecture 2: P1-P54)

- Conv2D operation
  - How the computation is performed
  - Input dimension, weight dimension, output dimension
  - Computational cost
- BatchNorm
  - Parameter folding-in during inference
- ResNet, MobileNet, ShuffletNet, SqueezeNet, DenseNet
  - Depthwise Separable Conv
  - Groupwise Convolution

Recap (Lecture 3: P1-P65, P73-P76,P79-P80)

- Transformers
  - How the computation is performed and why
  - Multi-headed attention, FFN
  - LayerNorm, RMSNorm, GeLU
  - Positional embedding, Word embedding
- Vision Transformer
  - How to convert image into visual tokens
- LLM
  - Prefilling, decoding
  - KV cache
- SSL
  - Contrastive learning
  - MAE

Recap (Lecture 4: P1-P70)

- Computational cost saving with pruning
  - CNN & Transformer
- Sparse matrix encoding
  - Bitmap, Run-length encoding, COO
- General pruning techniques
  - Magnitude pruning, gradient-based, Hessian-based pruning
  - Lasso
  - Taxonomy of Pruning
  - Network Slimming, N:M sparsity
  - Lottery ticket hypothesis
  - Cascade effect of pruning
- Transformer pruning
  - Token pruning
  - Head pruning

Taxonomy of Pruning

- Pruning techniques can be classified from different perspectives
  - Iterative pruning, zero-shot pruning
  - Structured pruning, unstructured pruning, N:M pruning
  - Weight pruning, activation pruning
  - Static pruning and dynamic pruning
  - Pruning for inference, pruning for training

Recap (Lecture 5: P1-P63)

- Basic Data Formats
  - Fixed point (INT), Floating point (FP), Block floating point (BFP)
- Quantization
  - Unsymmetrical & Symmetrical
  - Why fixed, FP, BFP & logarithm quantization can save computation?
- STE
- Taxonomy of Quantization
- Qunatization during training
  - Stochastic quantization
- Learnable adaptive quantization scheme

Taxonomy of Quantization

- Quantization techniques can be classified from different perspectives:
  - Weight quantization, activation quantization
  - Quantization aware training, post training quantization
  - Tensor-based quantization, vector-based quantization, group-based quantization
  - Quantization for inference/training
  - Deterministic quantization, stochastic quantization

Recap (Lecture 6: P1-P37)

- Distillation
  - Feature-Based Knowledge Distillation
  - Online distillation
  - Self distillation
  - Multi-teacher, multi-student, cross-modal

Recap (Lecture 7: P1-P32, P35-P37, P40-P56)

- Efficient training of DNNs
  - Training data sampling
  - Parameter sampling
  - Pruning during training
  - Quantization during training
  - Efficient storage
- Parameter efficient finetuning
  - Lora, Bitfit, Adapter
- Federated Learning
  - Concept, Non-iid, system issue

Recap (Lecture 8: P10-P30, P39-P45)

- Distributed DNN Training
  - Parameter server, all-reduce
  - Data parallelism, model parallelism
- Distributed DNN Inference
  - Layerwise partition, spatial partition
- Speculative Decoding
  - Basic concept