Progress Report

Xuanhang Diao, Letong Han, Yifan Chen

May 16, 2023

1 Schedule

- days before 2/4
  - Form the team
  - Determine the direction of topic selection to improve the performance of neural networks
  - Proposal report submission(26/3)
- Week1(2/4-8/4)
  - Confirm that the selected topic is related to NeRF
  - Proposal presentation preparation
  - Proposal presentation(7/4)
- Week2(9/4-15/4)
  - Re-discuss the topic selection, and clarify that the main workload is neural network compression
- Week3(16/4-22/4)
  - Understand mainstream network compression strategies: pruning, quantization, tensorization, etc.
  - Discuss the revised topic with instructor(19/4)
- Week4(23/4-29/4)
  - Choose to use a hybrid compression method(23/4)
  - Confirm the theoretical feasibility of using ADMM-based vectorization strategy, K-Means-based quantization strategy, GPU-based parallel acceleration
- Week5(30/4-6/5)
  - presentation preparation

  - complete presentation(6/5)
  - baseline construction
- Week6(7/4-13/5)
  - Confirm technical details: distributed ADMM, parallel K-Means
- Week7(14/4-20/5)
  - Implementation
  - Test part design (scene selection, parameter setting)
- Week8(21/4-27/5)
  - Experimental testing, fine-tuning
  - Result analysis
- Week9(28/5-3/6)
  - Final report writing
- days after 3/6
  - Tidy up the materials
  - Submit

2 Work Summary

We have designed a set of GPU-based neural network compression workflows that run fast, have significant volume compression, and less performance loss.

The high compression rate and low performance loss come from the design of the compression algorithm. We did not simply choose an existing compression strategy, but carried out scheme design and hyperparameter selection according to the nature of MLP in NeRF. We can achieve up to 97% PSNR of the original model with a model size of 30%.

The running speed comes from the GPU implementation. In the design phase, we considered the need for parallelization acceleration, and used an optimization strategy that has been proven to be parallelizable. In the implementation stage, it is based on GPU, which greatly improves the running speed.

3 Progress on Goals

We have confirmed the feasibility of the entire workflow, as well as the optimization algorithm applied, the implementation is in progress.

We found that the choice of scene is very important. Due to the limitation of computing power, we cannot test too large scenarios, but it is also difficult to reflect the advantages of our algorithm in smaller scenarios. Note that while we will choose appropriate scenarios, we will not deliberately avoid relevant questions.

4 Result to Present

We will show the PSNR and visualization results under different hyperparameter settings (different quantization bit numbers, different low-rank constraints).

5 Issues

The problem of scene selection has been mentioned earlier.

Theoretically our workflow can be generalized to neural networks of arbitrary structures, but we notice that neural network structures of different structures are sensitive to compression differently. For example, in the quantization process, the quantization of NeRF may only slightly affect the image quality, while the same strategy is put into the neural network of the classification task (which often has strong sparsity) and it is likely to cause serious performance degradation. Therefore, the practicability of extending to other deep learning models needs further research.