CSE 493 S / 599 S - Advanced ML
Homework 2

1 Overview

The goal of this homework is to give you hands-on experience developing modern machine learning models. In short, your task will be to **train language models from scratch**. You will have the freedom to make decisions regarding the architecture, optimization strategies, and other aspects of the model. This homework is deliberately open-ended, in order to simulate the experience of a practitioner working on a real-world project.

For this assignment, you should form **groups of three** people. If possible, we recommended that groups are diverse in terms of the students' backgrounds and experience with ML.

By the end of this assignment, you will have developed a solid understanding of how to train a language model, gaining hands-on experience in processing data and making decisions about the model architecture, optimization strategies, and hyperparameters. This assignment is due June 4th, 11:59pm.

2 Training your own language model

Your starting point will be the LLaMA codebase,[1] a recent large language model from Meta AI. This codebase provides *inference-only* code for their models. Your job is to transform this into a fully-functional codebase for *training*.

From inference to training. To add training support, there are a few things you'll need implement. First, you should implement the data loading logic, forming batches of tokens to feed the model. If you're using padding, you should be careful not to compute the loss on these tokens. We highly encourage teams to carefully look into the data that is being fed into the model, since that can often cause silent bugs.

Next, you should implement the training loop. You should select an optimizer and appropriate hyper-parameters. The LLaMA paper can serve as a guide, although it's likely that the models you will be training are going to be smaller than the ones in the paper, so we expect some hyper-parameter optimization will be needed. You should evaluate your model on validation data frequently throughout training. You should also implement monitoring and logging, making sure to track the loss values during both training and validation. Don't forget to save checkpoints.

Data. You will use the Pile, an open-source dataset that consists of 22 smaller, high-quality text data sources combined. The dataset is available at `https://the-eye.eu/public/AI/pile/`. You do not need to download or train on all of the training data, a small subset is fine. For validation, please use the first 10,000 samples from the validation set. If you want to use other data sources as ablations, that is also allowed.

Ablation study. In order to gain hands-on experience on what it is like to work on a real-world project, one of your tasks is to create an ablation study where you change some experimental configuration. You can change anything, including the source of data, architecture, or hyper-parameters. The change should be clearly motivated and described. Only one ablation is required, but more are welcome.

[1] `https://github.com/facebookresearch/llama/`

In order to understand whether your change reliably improved the model, you will conduct a scaling experiment, where you test your change at various levels of compute. You should create a scaling plot where the y-axis shows validation loss and the x-axis shows the amount of compute as measured by either total training FLOPs or GPU hours. For each line of experiment, you should conduct multiple training runs varying the amount of training compute by at least an order of magnitude.

Discuss the results of this study. Did the change improve the model or did it make it worse? How did that change with scale? Would you recommend your change to practitioners interested in building a much larger version of your model?

3 Deliverables and Deadlines

The main deliverable is a document describing your work on this assignment. **The deadline for this homework is June 4th, 11:59pm**, and document should contain:

- A description of how data processing was implemented. You should show example batches in your report.
- A clear description of the modifications done to the codebase to support training.
- A clear description of all experiments, including hyper-parameters and compute used.
- Main results, including plots for the training loss and evaluation loss along training for your best model. You should include examples of unconditional and conditional generation in your report (any prompts will do).
- A clear description of the ablation study you conducted, and the corresponding scaling plots. There should also be a discussion on the results of the study.
- A link to your code and checkpoints.
- A brief description of what was done by each person in the team.
- (Optional) A description of what you found least and most challenging when working on this assignment.

4 Grading

Grading will follow the rubric below.

- Data processing (10 points)
  - A clear description of data processing is provided.
  - The report shows example batches from the Pile
- Training implementation (20 points)
  - The code for training the model is provided, clear, and bug-free.
- Experiments and results (30 points)
  - Experiments are clearly described, including hyper-parameters, architecture, hardware and the time each experiment took.
  - The trained models are of expected quality. We will compare your model to those of other teams, taking model size and hardware constraints into account. We highly encourage you to try to train the best model you can.
- Reproducibility (20 points)
  - Code and checkpoints are submitted
  - There is clear documentation on how to run training and evaluation.
  - There is clear documentation on how to run generation with a custom prompt.
  - TAs are able to reproduce results from training and inference without problems.
- Ablation and scaling plots (20 points)
  - A clear explanation of what is being ablated is provided, along with rationale for the experiment
  - Scaling plots effectively compare the choices.
  - Results of the study are properly discussed.

- Extra points (max 10 points)

  Groups that go above and beyond will be rewarded. In general, extra points will be awarded if you do something cool =)

5 FAQs

Are groups with different sizes allowed? You should form groups of three people whenever possible. In exceptional cases, please reach out to TAs.

Can I use ChatGPT/GPT-4? Yes. However, please clearly state when and how it was used, anytime that happens.

Am I allowed to use other codebases? Yes. However, clearly state which ones are used, and for which purpose.

Can I use this homework as a starting point for my final project? Yes.

Where am I supposed to get compute for this? There are a few free options for GPUs. For example, `https://colab.research.google.com/` and `https://www.paperspace.com/gradient/free-gpu`. While we don't discourage students to pursue larger-scale experiments, we also will not penalize students that only run small scale experiments due to hardware constraints.

Will individual students get separate grades? In most cases, there will be a single grade per group. This can be revisited for exceptional cases.