Selecting Training Hyper-Parameters And Model Initializations

The easiest way to find a good hparam and model init starter set is to steal it from a similar training that you know has succeeded. Here is a collection of public training LLM/VLM logbooks to get you started. The other common source is papers if they disclose that information. You can also try to reach out to the authors and ask them for these details if they didn't publish it.

Glossary

Training jargon uses a multitude of abbreviations and terms, so here are some important for this chapter.

- BS: Batch Size - here we mean batch size per gpu, often it is also referred to as MBS (micro-batch-size)
- GBS: Global Batch Size - total batch size per iteration - may include gradient accumulation
- GAS: Gradient Accumulation Steps - how many forward/backward cycles to perform before one full iteration is complete
- TFLOPs: Trillion FLOPs per second - FLOPS
- PP: Pipeline Parallelism

Global Batch Size Ramp Up

If you intend to train with a very large GBS, with say 1024, or 2048 samples and even higher, when you just start training, it's very wasteful to feed such large batch sizes to the model. At this point it's totally random and can't benefit from having too refined data. Therefore to save data and resources, one often ramps up the global batch size over some period of time.

It's also important to not start with GBS that is too small, since otherwise the progress won't be efficient. When there is too little data the compute (TFLOPS) is inefficient and will slow everything down. This is especially so when Pipeline Parallelism (PP) is used, since the most important thing about PP tuneup is a small GPU idleness bubble, and the smaller the GBS the larger the bubble is.

For example, for BLOOM-176B, where we did use PP, after doing throughput benchmarking we found that starting with GBS=16 was incredibly slow (8 TFLOPs), so we eventually started with GBS=192 (73 TFLOPs) and then we ramped up to GBS=2048 (150 TFLOPs) - we increased GBS by 16 every 9_765_625 samples.

STD Init

This hyper parameter is super-important and it requires math to get it right. For details see STD Init.