Technical Details

Data

Software

- for all implementations we use pytorch and pytorch lightning
- pytorch lightning provides high-level functionalities on top of pytorch
  - like checkpointing, logging, distributed training, etc.
- we do not need to implement them in pytorch manually (to some extend)
  - pytorch already provides a high-level API but it is more prone to errors
- we save time and can focus on the actual implementation
- errors in vanilla pytorch are likely and hard to debug
- research will enevitably involve a lot of trial and error (experimentation)
- to keep track of all experiments, we use the experiment tracking tool Weights & Biases[1]

Hardware

To train the models and store the data, we used the GPU cloud platform runpod.io[2]. This platform provides access to a wide range of GPU types, including the NVIDIA RTX 4090 24GB, which we use for training almost all of our models. The reason for choosing this platform over traditional cloud providers like AWS or GCP is that the price per GPU hour is significantly lower, which allows us to train more models for the same budget. For example, as of September 2024, the price per GPU hour for an NVIDIA RTX 4090 24GB on runpod.io is $0.69. A comparable GPU on AWS, e.g. the NVIDIA V100 16GB, costs $3.06[3] per GPU hour. This price difference is despite the fact that the V100 was released in 2017, while the RTX 4090 was released in 2022. The RTX 4090 is also faster than the V100, with a higher memory bandwidth and more CUDA cores, so training on the RTX 4090 is more cost-effective by a margin. This evaluation is significant, as there is no external funding for this work.

To store all of our data, which is a total of >900 GB, we use a network volume provided by runpod.io. This volume is mounted on a virtual machine on start-up, allowing to access the data for training on the GPUs and provides high flexibility.

The RTX 4090 instances we used have 61 GB of memory and 8 virtual CPUs for one GPU, and around 132 GB of memory with 16 virtual CPUs if using two GPUs, which is similar to that of AWS.

runpod.io also provides on-demand vm instances, as well as spot instances. The latter can be automatically terminated if demand is high, but the price is significantly lower. On-demand instances are more expensive but are non-interruptible, which is, even though we create a model checkpoint after each training epoch, important for long-running experiments, which is the case for most training runs in this work. As of September 2024, the price for an on-demand instance with a single RTX 4090 is $0.69 per GPU hour, while the price for a spot instance is just $0.35 per GPU hour. The price for an instance increases proportionally with the number of GPUs used, so for us, a two-GPU instance costs $1.38 per GPU hour.

Costs Breakdown

[1] https://wandb.ai

[2] https://www.runpod.io

[3] Price is taken from the official AWS pricing page (https://aws.amazon.com/de/ec2/instance-types/p3/) and based on the on-demand price for the region us-east (North-Virginia).