Week 3 Report

2023-12675 박지호

Today's topic was the Phi language model, and the importance of mixture of expert in the advanced training.

The seminar started by discussing about the small language model. Small language model refers to a smaller, less compute intensive models that perform well at specific tak. Compared to LLM, it is cost effective, faster, and easier to manage and fine tune. The presenter demonstrated that phi 3 mini can produce the answers of similar quality to the GPT while being a lot faster. Moreover, as they are less computationally intensive, they can run in mobile devices. Phi 3.5, the latest model of the Phi series, outperforms models twice its size, can run on any CPU/GPU, and also has multilingual support. The model is the best performing model when taking into account the number of parameters.

Then, the Phi-3.5-MoE model was introduced, which uses MoE approach. MoE stands for mixture of experts, which refers to a model that consists of multiple 'expert' modules, each specializing in certain field. When given an input, the module picks which of these models to use to generate an output. This method allows the model to achieve the best performance per computational requirement. Also, traditionally, the expert modules existed in parallel and independently, but recent approaches tend to use an approach where the expert modules share the inital and final layers.

Some of the major downsides is that increasing the number of parameter doesn't result in as much performance gain as expected. Another important aspect is load balancing. If only few expert modules gets selected most of the time, other models wouldn't have much chance to train.

After introducing these downsides, the presenter introduced the GRIN model, which attempts solve these problems with the MoE method. First, in order to utilize the additional parameters more efficiently, the SparseMixer method, which uses some mathematical concepts not explained thoroughly to get a better gradient and therefore more efficient training, was used. Also, in order to balance the load more evenly, global load balancing, which uses load balancing across multiple GPU, instead of in a single GPU, was used. As a result, this model was able to achieve high performance metrics, remarkable for it's size, across the board. Also, the model showed great performance in reasoning problems.

One thing I found interesting was the emergence of residual connection in the GRIN model. The presenter explained that one way to achieve better utilization of parameters were to have a single dense model which handles general inputs, with multiple expert blocks that occasionally help out the main expert block. However, when applying the gradient method mentioned before along with many expert models, this pattern naturally emerged without any modification to the module. A while ago, I attended an interesting seminar on the emergence patterns that appeared from ai training. This seminar was less focussed on the ai aspect, but rather the implication it had on the emergence patterns in living organism's intelligence. However, it was still interesting to see a pattern like this emerge without outside interference.

As such, I asked the presenter why they think this pattern emerged only after utilizing the new gradient method. As this gradient method wasn't really explained in detail, I naturally had no idea what impact it would have on the model. However, as I only came up with the question too late, it couldn't be answered in time. If I have the time, I might try and look into the paper the presenter mentioned, that is, once the midterm is over.