Week 2 Report

2023-12675 박지호

Today's seminar was by Junyang Lin from the Qwen team, and naturally, the seminar covered the Qwen language model. The seminar started by sharing the history and the current state of the Qwen language model, along with the findings the team found during the development of the Qwen model. It was mentioned that the some versions of the model has been opensourced as of right now. The most recent model in the Qwen series is Qwen 2.5, and the specs along with several benchmark scores were shown to demonstrate the strength of the model.

Then, it was shown that smaller and smaller parameters are required to achieve the same performance as time goes on. The presenter preclaimed that smaller language models have more in store than what people give them credit for.

The importance of pretraining data was also emphasized. While the size of data greatly influences the performance of the model, ensuring the quality of data is also crucial. The presenter explained that the Qwen team utilized the Qwen 2.0 model to filter out garbage data from the set. The distribution of the data is also important, as most data exists in English. Last of all, using synthetic data can greatly enhance performance. The presenter said that at first they were skeptical about synthetic data improving the model's performance, but after realizing its importance the team decide to use synthetic data, while using language model to correct any wrong information that may be present in the data.

The presenter also introduced couple other tweaks the team tried to further improve the model's performance, like the tokenizer and model architecture. The model architecture utilizing transformer was thought to be set in stone, but the presenter discovered the possiblity that few changes could improve the model further.

The presenter then talked about how post training process, although usually glanced over, also plays a crucial part in designing language model. Supervised fine tuning, reinforcement learning from human feedback, and model merging were mentioned as few of the post training methods.

Then, the presenter concluded the seminar with the vision of their team, creating a unified multimodal multitask AI model/system, and making foundation models smarter.

One interesting question that was asked during the seminar was how students could get into language model training, as students wouldn't realistically have access to the scale of data and access that modern language models require. The presenter's answered that the best way would be to join a company that has such resources. However, if that's not available, they said it's okay to start by experimenting with smaller models and scale from there. Also, I asked what design differences there are between the base Qwen model and math or coding specific model, if there are any. However, the seminar had to be cut short, so it couldn't be answered within the time.

Listening to the seminar, I felt both inspired and slightly overwhelmed. The level of expertise that went into developing the Qwen model made me realize again how complex and difficult the process of building a language model is. It was really interesting how much thought goes into everything from pretraining data to model architecture, and post-training. The insights into synthetic data were particularly surprising, as it's a concept I had never expected to affect the model performance.

At the same time, as I realize that, I found myself a bit discouraged to get into language model training. It felt like there are just too many things to catch up on. The presenter said getting into a company would be a good way to gain access to resources required to train language model, but I don't think it's realistic to expect anyone to hire me in my current state, for I practically have nothing to show for myself. However, the presenter also did say it's a good idea to start experimenting with smaller models first, so I think that's where I should start.