Week 8 Report

2023-12675 박지호

The topic of today's lecture was Solar Pro and workspace LLMs. Today's presenter, Sungrae Park, shared stories about how Solar model came to be, and what their team intends to do in the future.

They started off the seminar by setting some basis about the structure of B2B business with AI technology. The B2B customers wants an AI technology that effectively increase their business value. Therefore, AI tech companies aims to develop and maintain an AI model which would be too much for a single customer company to handle, and aid them in achieving their goal. In this business, the customer wants a secure, reasonably priced, and well performing AI solution.

The presenter first discussed how to achieve the performance that customer requires. The customer usually doesn't mind if the model is general or specialist. Therefore, if the generalist solution's performance fails to achieve the customer's desired performance, fine-tuning the model to create a specialist model can help breach the gap. When even that falls short, we can employ some additional engineering to increase the performance further. The preseneter went on and shared couple examples in which their team utilized these strategies to match the customer's needs.

Even after reaching the customer's desired performance, the answer is not completely over. Usually, the customer's required baseline grows as time passes, so the model's performance must constantly improve to keep up. Furthermore, the difficulty of matching the customer's requirements may vary depending on the specific task. Therefore, it is essential that we constantly optimize for a higher performance.

To achieve this long-term performance improvement. Developing new specialist model for every task would not suffice. Although specialist models perform better at a task they are specifially trained to solve, they do not scale well. Eventually, the generalist model will outperform any specialist model in the long run. Therefore, the presenter claimed that their team wishes to constantly improve the general model, all the while fine-tuning the models to provide short-term solutions that matches the customers' needs.

Then, the presenter moved on to their newest model: the Solar Pro. The presenter first explained some of the difficulties they faced during the development of the model. The fine-tuning process can be difficult as the model can easily forget about the injected knowledge. This problem happens more frequently when attempting to scale up a trained model. To counteract this, we can scale the model using a method called depth up scaling. In this method, we obtain the larger model by inserting the exact copy of the layers in the original models in between the other layers, which we train further. One interesting thing about this approach is that once the training is over, the copied layers usually end up being altered heavily from their original state.

Another challenge when fine-tuning the model is choosing which knowledge to inject to the model. For this task, we use the data preprocessing pipeline. In this process, we should find the balance between quality and cost. The presenter introduced us to a recent paper that discusses this process further, saying anyone interested should give it a look.

The presenter lapped up the seminar by showing us the benchmark metrics for the Solar Pro model. One thing that I do not understand about these benchmark metrics is that every presenter seems to claim that their model performs best. Benchmark metrics are often carefully chosen to highlight strengths while downplaying weaknesses. Each team typically selects datasets, tasks, or evaluation methods that align well with their model's capabilities, which can lead to inflated or overly favorable comparisons. Hence, while benchmarks are a useful tool for gauging performance, they should be interpreted with caution and viewed as part of a broader context that includes practical applicability, robustness, and the ability to meet specific customer needs.