Week 6 Report

2023-12675 박지호

The topic of the today's seminar was "Don't teach, incentivize". The presenter first told us how they do not intend to focus on the technical results and experimental details, but rather a genral approach on how to train AI models. The presenter stressed out that although it's easy to get caught up with how to solve the problem, the more important aspect is identifying the problem correctly, rather than solving it. They said they wished to share their point of view on the matter, and .

The presenter said that in many scenarios, simply scaling up is a better way to train a model than utilizing some technical optimization strategy. The uniqueness of the AI field means that often times many optimization that seems to bring some short-term gain could end up becoming a bottleneck long-term. However, researchers used to underestimate the importance of scaling up, as this was not technically satisfying.

The presenter than moved on to the large language models (LLM), and how the philosophy apply to them. They first briefly explained the structure of LLM. The natural langauge data is tokenized and used as a training data. Then, the model uses these pretraining tokens to predict the next token.

The presenter mentioned how astounding it is that with this simple of a process, the model can learn, understand, and speak the natural language with a highly complex set of rules and nuance. This ability is emergeant; the natural language was never explicitly taught the natural language. Rather, it was simply incentivized to learn it. From this observation, the presenter argued that this is a general approach we should take when training AI: instead of teaching the model how to perform the task, the model should be weakly incentivized to learn the task.

The downside of this approach, however, is that it requires a lot more computing power compared to teaching the model how to do things specifically. However, the presenter explained how we shouldn't think that machines operate in the same way humans do. Unlike humans, machines have no upper bound in the computations they can make, and while humans perform better when they focus on one ability, abilities can emerge in machines when they are instructed to perform a more general task. Although the cost of hardware still function as the limiting factor, this cost decreases exponentially with the cost of time.

The presenter ended off the lecture by emphasizing the importance of unlearning when researching AI. They pointed out how newcomers with not much basis in the AI field tend to come up with groundbreaking idea that allows a breakthrough in the entire field.

I really liked this seminar compared to other ones. Like they promised, the presenter focused less on inconsequential details on the model and benchmark results that do not mean much to me, they focused more on insights and general approach and methodology.

However, this was also very disheartening as well, and made me not want to research AI more than before. The presenter mentioned how technically satisfying revelations do not necessary bring about performance gains. While this is quite interesting and likely true, I have a feeling that I might not enjoy researching AI, for I would like to do something technically satisfying and mentally stimulating.

This left me with mixed feelings. I appreciated the insight into focusing on problem identification and incentivizing learning, but it made me question whether my passion for technically intricate work fits in a field driven by scaling and emergent behavior over precise optimizations. The idea that breakthroughs often come from simplicity and scaling rather than deep technical finesse made me wonder if my desire for hands-on, technically satisfying challenges aligns with the direction of AI research.