A random thought on multi-task learning and AI existential risks

By Freddie Yang

Executive Summary

A capable AI can greatly facilitate tasks that humans deem mundane, but fears that advanced AI endangers humanity is still looming. Many AI researchers downplay such fears by arguing that AI cannot grow to be dangerously powerful because AI's potential is constrained by factors such as the amount & quality of data, computational costs & complexity, heavy reliance on human inputs, etc. I, however, do not believe these constraints are sufficient to contain AI growth.

In Biology, scientists were able to create a superior species by cross breeding different species without breaking external constraints. Similarly, it might be possible for AI to become much more powerful under the same external constraints. An analogy to crossbreeding in AI research is multi-task learning, a methodology that trains an AI to do multiple things simultaneously. If not properly regulated, it is not impossible for someone to accidentally create a dangerous AI through multi-task learning. Similar to how people think it is necessary to regulate crossbreeding animals/plants, I believe it is important for us to begin thinking about similar regulations in AI research.

Why are people afraid of AI becoming more powerful?

A capable AI can greatly facilitate tasks that humans deem mundane, but several reasons make people scared of making AI more powerful than it is today.

First, nobody can explain how a complicated AI model works, and it is hard for humans to trust something they don't understand. Nowadays, it is not uncommon for an AI model to have millions of parameters. Attempting to understand each parameter and their interactions turned out to be an extremely difficult problem. Despite significant research effort by the Explainable AI community, the current state of research is not able to provide a satisfying solution to decipher the internal mechanisms of complicated AI models.

Second, the danger of AI is exacerbated if it is utilized by people pursuing inhumane goals. People fear AI for the same reason they fear the Nazis obtaining nuclear weapons. Even if we assume AI itself to be a benign invention, we cannot guarantee it won't be part of a malicious plan that threatens humanity.

Third, development in AI research in recent years has demonstrated AI's potential to surpass humans in a key ability that made us the most powerful species on earth – intelligence. People have comfortably accepted the fact that they can't compete with machines in terms of physical strength, because being smarter than machines gives humans a significant edge on any potential human-machine conflicts. It is, however, unacceptable for humans to live under the dominance by another species that is superior in

both the dimensions of intelligence and raw ability, because that would effectively signify the end of human's reign on this planet forever.

Fourth, AI is scary because it is not bound by the same natural constraints that humans face. An AI doesn't need to drink or sleep. It seems to have access to unlimited memory and power while a human only has a small brain to remember things and a small stomach to digest food. An exceptional AI can be trained in just a few hours and can replicate itself in seconds while it takes 18 years of good nurturing for a human to become a responsible adult. For these reasons, humans don't seem to stand a chance if they had to go to war with an AI army.

There are many more reasons to be scared, but I believe these four reasons above are sufficient to justify some immediate actions to start regulating AI development.

How can AI become more powerful? An analogy between AI and Biology

Although AI can already do amazing things, their growth potential is often constrained by factors such as the amount & quality of data, computational costs & complexity, heavy reliance on human inputs, etc. For these reasons, many AI researchers do not believe AI existential risks are serious as long as these constraints are still in place. I, however, do not believe these constraints are sufficient to contain AI growth. A simple analogy with biology might give us some insights.

For thousands of years, people have attempted to perfect the art of growing watermelons. People have tried using better soil, cleaner water, novel watering schedules, etc. Sure, these attempts resulted in slightly better-tasting watermelons, but the improvements were not significant. And for a long time, people believed that watermelon could not be made any tastier due to natural constraints in soil/water/planting methods, which, of course, was a reasonable belief until seedless watermelon was invented.

So, where did seedless watermelon come from? In 1951, Professor H. Kihara from Japan invented the seedless watermelon by crossing male pollen for a watermelon, containing 22 chromosomes per cell, with a female watermelon flower with 44 chromosomes per cell. These crossbreeding techniques allowed us to develop superior products without breaking the same previously mentioned natural constraints like water or soil.

Can we generate a significantly superior AI by crossbreeding existing AI? The answer is uncertain, but multi-task learning research is certainly trying to move us in that direction.

What is multi-task learning?

A traditional AI is trained to do one thing and do it well. And mathematically, this means that an AI model is trained to minimize a "loss function" which is a measurement of the model's error. A simple example of such a model can be one that predicts the stock price, and the "loss function" specifies how far off the predicted price is compared to the real price.

A multi-task AI, as its name suggests, is trained to do multiple things simultaneously. A classic example of a multi-task AI is one that is trained to detect both cars and motorcycles from pictures simultaneously, as

opposed to two different models that detect these two objects separately. The intuition for training such a multi-task AI is that people believe that the “skills”/ “knowledge” required to detect cars should be similar to those required to detect “motorcycles”, and it is therefore reasonable to believe that by “sharing” these common “skills”/ “knowledge”, we can obtain a more performant model.

Although it makes more sense to train multi-task models to perform similar tasks, there has been a trend in the AI research community to train AI to perform drastically different tasks through the so-called heterogeneous multi-task learning. Most recently, DeepMind announced that their new AI -Gato- is capable of performing 600 different tasks.

My own mini experiment

I am particularly scared by the possibility of people accidentally creating a much superior AI by “crossbreeding” simple AI models. And to get a sense of how realistic that possibility is, I needed to train a multi-task model myself, so I know what the major obstacles are. I eventually decided to train a simple multi-task AI that can both chat with us and play chess, because I thought this would be a good exercise of heterogeneous multi-task learning since I expect very little commonality between these tasks.

For the chess part of the model, I simply followed Logan Spears’ training procedure, where he trained a four-layer fully connected model using chess moves data played by experts from the Lichess website. For the chatbot part, I followed Microsoft’s DialoGPT’s training process, where they used transfer learning on the Transformer model with data scraped from Reddit. And I combined these two models by adding a few more common fully connected layers on top of both tasks.

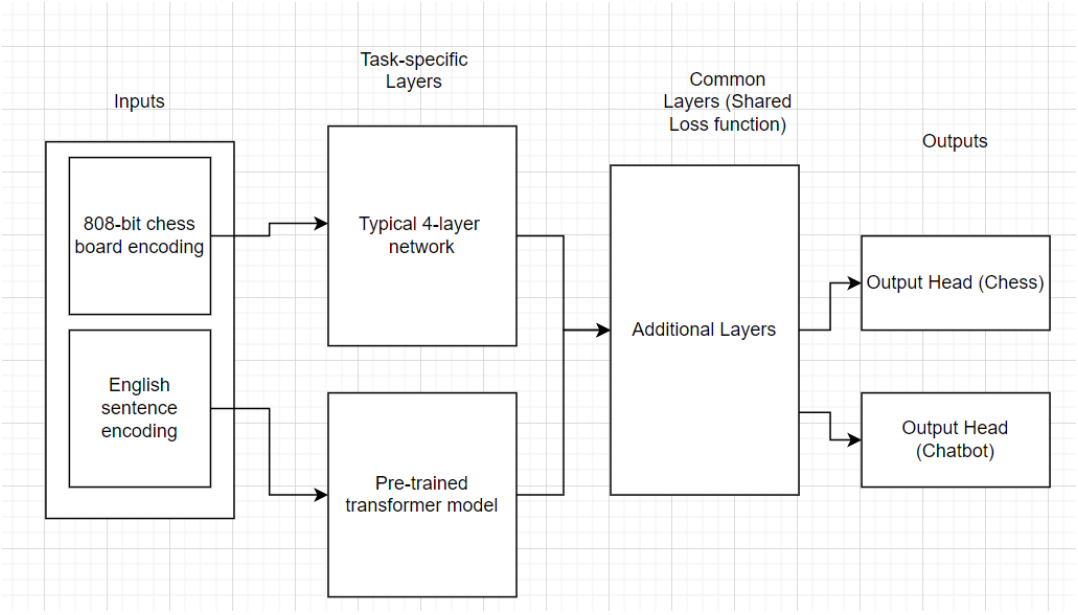

Figure 1. Chat-Chess Model Architecture

Results & lessons learned

So, how did my model perform? Unfortunately, it performed very poorly. I could "sort of" chat with it, but it gives nonsensical responses. For example, when I asked the model "what's your name?", the output was something like "I am lala … blue …". The model's chess ability was poor as well, as it could not even beat a beginner like me.

Although I did not get the miraculously superior model like the seedless watermelon that I was hoping for, I don't think of this as a complete failure. On the bright side, as someone who's just begun learning about deep learning this summer, I did gain some valuable insights as to why the model didn't perform well.

The first reason, of course, is that the two tasks are so different that there wasn't much "common knowledge" to exploit by multi-task learning. I also only trained the model on a single GPU for less than 2 hours which were not long enough. I also needed to find a better way to balance the two tasks better, probably by assigning greater weight to one task than the other. I could also try a more complicated architecture, such as those with recurrent networks and more appropriate optimizers, than the one I used.

Multi-task learning and AI alignment research

Is a more powerful AI necessarily harmful to humans? After all, even if AI machines can outperform humans in nearly every task, they are not necessarily dangerous as long as they exist to serve human needs. But how do we guarantee that AI and human needs will forever remain consistent? This is the core research question being tackled by the AI alignment research community. Unfortunately, as of now, we are still far from getting a satisfying answer.

Although my mini experiment failed to generate an AI that was sufficiently capable of multitasking, I do believe multi-task learning provides a novel approach to the alignment question. Had my mini experiment succeeded in creating a chatbot & chess AI, my original plan was to use transfer learning to train two more multi-task models based on it. For the first model, I was going to fine-tune the chatbot to generate only positive responses by fine-tuning it with Reddit data that only carry positive sentiment. Examples of such data could be conversations like "you are awesome" or "good job". For the second model, I was going to fine-tune it with only "negative data" like "you did horribly" or "you are fired". The first model can be thought of as one that is aligned with human needs because it is trained to "say nice things to humans" while the second model is misaligned.

Then I would ask the two models to compete in the game of chess. If we observe any consistency in the game results, then we would have a way of indirectly answering the alignment question. For example, if it turns out that the first model consistently beats the second model, then we can reasonably argue that a chatbot & chess AI is aligned with human interests, and there is no need to be scared of it getting better. On the other hand, if the results were reversed, then we could argue that this AI model is misaligned and we should be wary of making it more powerful.

Conclusions

AI risks can materialize in various forms, and I think multi-task learning, seemingly harmless, is one of them. Although my own experiment did not support this argument, I don't believe it is impossible for humans to accidentally create a much more powerful AI by crossbreeding multiple benign AI models. Similar to how people think it is necessary to regulate crossbreeding animals/plants, I believe it is important for us to begin thinking about similar regulations in AI research.

References

https://www.watermelon.org/the-slice/where-does-seedless-watermelon-come-from/

https://towardsdatascience.com/train-your-own-chess-ai-66b9ca8d71e4

https://arxiv.org/abs/1911.00536