Review of “Grounded Language Learning in a Simulated 3D World”

In this paper, several researchers out of DeepMind aim to improve on previous efforts to learn grounded, embodied language. The need for grounded language understanding is well laid out in the paper as in order to obtain a human-level AI, an agent must be able to operate in a physical world from the perspective of itself using only information it can gather from its own sensors.

Many methods are used during the training of these agents, but I think it is most important to focus our attention on the fact that this paper found that teaching the agent on a new task from scratch takes much longer than teaching an agent that knows only a couple words. Much like a human, the agent learns its foundations in these first two words and then is able to fit its current understanding around the next words or tasks. When a human baby is born, it takes nearly a year for it to learn its first word, but afterwards it will learn its second, third, fourth, and so on exponentially faster. In addition to this, the agent is actually able to learn complex tasks much faster than an agent trained from scratch. In many tasks, the agents that were trained from scratch didn’t learn anything by the time the agent trained off of previous knowledge had mastered it.

I found this to be a great experiment as many improvements in AI typically stem from our understanding of ourselves as humans, and they provide more motivation to learning more and more about the brain and how humans work.

While this paper is a bit limited in its scope, it is still excellent work towards an intelligent system. I would like to see similar agents trained on more tasks and more words, especially with modern language-modelling performance in transformer-based

systems. It would even be great to see how this same model performs with an attention mechanism in place as there are situations where the agent appears to be looking at things humans typically wouldn't.

One thing in particular that differs from a human approach to this agent is that when asked to find something in a room, a human would understand where it might be, i.e. a lamp should be on a table, microwave should be on a counter, and so on. With this knowledge, we can quickly rule out areas that we no longer need to look in. In the agent's case, the things it is looking for are typically found along the walls meaning that once the agent has seen all of the places that an object could be, it should immediately turn around and check elsewhere, i.e. the other room.