OpenAI: A brief description and list of references

Robert George Phillips

May 9, 2019

My term project will be on the non-profit research organization OpenAI, headquartered in San Francisco. OpenAI was founded in 2015 by Elon Musk of Tesla and SpaceX, Sam Altman of Y-Combinator, Ilya Sutskever of Google, and Sam Brockman. Their mission statement is "discovering and enacting the path to safe artificial general intelligence." I'm curious to look into this research organization a little more, because the reality of humans and AI living together will happen sooner rather than later, and we as a society have to discuss what steps we need to take to ensure that AI stays out of the hands of the malicious. OpenAI seems to be at the forefront of reasearch as well as industry in the field of AI.

An interesting paper from OpenAI that I'll likely delve further into is on their GPT-2 language model system. It's a complex language model that is able to detect speech and tone, identify key words, and write prose and articles. It can even outperform language models that are specifically trained on a singular work (like Wikipedia, books, or articles). Since it can be used to generate plausible-sounding fake news, generate spam and phishing attacks, and potentially impersonate other people online, OpenAI has elected not to share the specific details behind this technology. Despite that, the technology that they showcase is extremely impressive – I was shocked when I read a sample article that GPT-2 produced, as I wouldn't have been able to tell that a robot had written it had I not known before reading it.

I would also like to talk about OpenAI's Dactyl project. Dactyl is a human-like robotic hand that is able to mimic human gestures and reflexes. Interestingly, Dactyl learned how to move a hand like ours all by itself – it had no sample set to work from. Researchers at OpenAI found that Dactyl actually learned a lot of motions that humans were used to, but had some unique quirks. For example, usually humans pick things up with their thumb and index finger, but Dactyl uses its thumb and little finger. The technology behind the training model is very interesting, as well.

As a wrap-up, I would greatly like to discuss the ethics of what OpenAI and others in the field of AI are doing. It seems to me with all of these technologies that OpenAI is on the forefront of, that they are trying to build human-like androids capable of mimicing human movements, understanding our tone, being able to converse to humans in nuanced ways. I'm not sure how to feel about that, and the philosophical, technological, and ethical implications are enormous. Many jobs are also going to be automated sooner than people think, which brings its own issues.

References

[1] Andrychowicz, M., Baker, B., Chociej, M., Józefowicz, R., McGrew, B., Pachocki, J. W., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., and Zaremba, W. Learning dexterous in-hand manipulation. *CoRR abs/1808.00177* (2018).

[2] Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. Quantifying generalization in reinforcement learning. *CoRR abs/1812.02341* (2018).

[3] McCandlish, S., Kaplan, J., Amodei, D., and Team, O. D. An empirical model of large-batch training. *CoRR abs/1812.06162* (2018).

[4] Mordatch, I. Concept learning with energy-based models. *CoRR abs/1811.02486* (2018).

[5] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI* (2019).