Taylor Livingston

2019-05-03

CMPT*301*01

Homework 6

Simulating Reality for Data Production

The business of data collection is a 21st century goldrush. Not more than a decade ago it was questionable that companies at the heart of such practices like Facebook, Twitter, and Amazon would produce sufficient revenue streams for operation. At the time big data was still valued, but the current advances in artificial intelligence and GPU hardware have taken it to the next level. Now Amazon is a valued at over one trillion dollars while Facebook and Twitter have woven themselves into the fabric of global discourse. As the desire for more and more data increases, the barriers to entry for the average person in accessing large datasets become ever higher.

Although these barriers for non-synthetic data collection are extremely high, a paper authored by researchers affiliated with Nvidia were able to produce a synthetic alternative to big data which allowed their deep learning networks to produce amazing results when used in the real world. (Prakash) The current big data landscape is innovating just as the algorithms that train models dependent on it are, but those artificially intelligent algorithms can also help to innovate big data. Simulating the real world in terms of data is an inherently partially observable problem. There is no way to capture a state of reality in its entirety, and thus there is no way to produce a realistic state of reality entirely. Some algorithms that simulate large real-world datasets are single agent, but those of the GAN variety include one agent to produce, and one agent to verify so these should be considered mutli-agent algorithms. The methods discussed in

this paper are mostly sequential, where the data that entered the network previously also effects the predictions on data that come after it. Deep learning algorithms can constantly adapt to the data that is ingested by them. This dynamism allows them to be effective across a wide variety of scenarios. Undoubtedly it can never be stated that synthetic data producing algorithms are discrete, they can merely imitate real data in a continuous fashion.

Prakash and his team's research is important for a variety of reasons. Firstly, big data is not only hard to acquire in the context of vehicle driving data, as explored in their research, it's also very difficult to annotate/label. (Prakash) Just because a video of a car's driving is recorded, does not mean the individual frames are understood by the computer to be composed of different objects and environments as it is perceived by people. In a simulated driving environment, the data comes automatically labeled, as it is composed of virtual objects and environments as people perceive them. The intermediary step between data collection, and big data can be thought of as one in a simulated environment that is built to represent the real word. This generation process is referred to as Structured Domain Randomization or SDR, and it functionally eliminates the process of data labeling required in a training dataset for many deep learning algorithms. The results are so promising that they concluded pretraining on SDR data improves the results when real data is subsequently presented. (Prakash)

Apple researchers also interested in computer vision come to a similar conclusion in a paper from 2016, titled Learning from Simulated and Unsupervised Images through Adversarial Training. They find that the collection and annotation of training datasets is increasingly cumbersome when training deep neural networks. (Ashish Shrivastava) This takes the previously mentioned research and builds upon it in many ways. The Apple team concludes that essentially training a network using simulated supervised learning is enough to prepare it for

unsupervised real-world data. To quote directly, “We described Sim-GAN, our method for S+U learning, that uses an adversarial network and demonstrated state-of-the-art results without any labeled real data.” (Ashish Shrivastava). These breakthroughs in synthetic data creation mean more people will have access to the tools of big tata than ever before. If what Apple researchers suggest is correct there is tremendous promise for the innovating and development of AI across a huge amount of areas.

During the past decade the world surrounding big data and the technology fueled by it have evolved to an almost unrecognizable state. The introduction of methods like SDR and Sim-Gan for computer vision applications is a massive step in the right direction. In an article from Towards Data Science by George Seif it’s said, “Being able to create high quality data so quickly and easily puts the little guys back in the game. Many early-stage startups can now solve their cold start by creating data simulators to generate contextually relevant data with quality labels in order to train their algorithms.” For the first time in recent history there is a good chance that companies and institutions other than the Amazons and Facebooks of the world will have access to the amount of data required to train new and complex models. Seif says that, “Simulating data will level the playing field between large technology companies and startups.” (Seif) Going forward it seems that the incredible capability of current AI algorithms might be enough to push the next generation into fruition.

References

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, Russ Webb. "Learning from Simulated and Unsupervised Images through Adversarial." Apple Inc, 2016. <https://arxiv.org/pdf/1612.07828.pdf>.

Prakash, Aayush & Boochoon, Shaad & Brophy, Mark & Acuna, David & Cameracci, Eric & State, Gavriel & Shapira, Omer & Birchfield, Stan. "Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data." 2018. <https://www.researchgate.net/publication/328494394_Structured_Domain_Randomization_Bridging_the_Reality_Gap_by_Context-Aware_Synthetic_Data>.

Seif, George. *Deep learning with synthetic data will make AI accessible to the masses*. 10 Jul 2018. 03 May 2019.