WomenCorporateDirectors
A Foundation Inspiring Visionary Boards Worldwide

The Surprising Brittleness of AI

M.L. Cummings

While artificial intelligence (AI) has recently been touted as very successful across a number of domains, the reality that such systems may not actually be as capable as envisioned is slowly creeping into the national consciousness. While AI can show up in many every day applications from shopping to management of home automation, it is the application of AI in safety-critical systems like transportation and medicine that is the most concerning since literally the incorrect use of AI can have deadly consequences.

Problems with automotive computer vision have been cited as contributing factors in many fatal Tesla crashes (Risen 2016, Crowe 2016) and the death of a pedestrian in an Uber self-driving car accident (Griggs and Wakabayashi 2018). Despite years of promises by many companies of full-self driving powered by AI, many companies have walked back their claims in attempt to recalibrate the public's and funders' expectations (Bubbers 2019, Elias 2019).

In concert with public backlash over AI and privacy, as well as concerns with AI embedded in social media that could be manipulating people, negative sentiment is growing about applications of AI. Many experts are concerned that this backlash could lead to another AI winter that could lead to significant distrust in legitimate AI advances and a cooling of financial support (Walch 2019). Given this potential outcome, it is important to step back and analyze just why AI is struggling to gain traction in safety-critical systems and how the roadmap to success would need to change to achieve positive outcomes.

A problem of brittleness

In safety-critical settings like transportation and healthcare, computer vision is a common application of artificial intelligence, which typically means algorithms leverage machine, sometimes called deep, learning to "perceive" the world in order to make decisions. For example, deep learning algorithms in driverless cars determine whether a car "sees" a pedestrian or in healthcare, whether a tumor exists in a grainy image of a lung. While important advancements have been made in the last 10 years in computer vision and in the deep learning algorithms that underpin these systems, such approaches to developing perceptual models of the real world are plagued by problems of brittleness.

Brittleness occurs when any algorithm cannot generalize or adapt to conditions outside a narrow set of assumptions. For example, many natural language processing algorithms are brittle when they can understand a person from New York City but fail to understand the same sentence from someone in Appalachia or who speaks English with a foreign accent (Harwell 2018). While this brittleness may be frustrating for a person attempting to navigate a phone tree, it can be deadly in a safety-critical system that relies on any kind of machine learning for perception or critical reasoning.

The source of this perceptual brittleness comes from the fact that machine learning algorithms do not actually learn to perceive the world in a way that can generalize in the face of uncertainty. For a machine learning algorithm to "learn" to see a stop sign, for example, it must "see" tens of thousands of similar images in order to understand patterns of reoccurrence. What the algorithm has "learned" is that a particular set of mathematical relationships belong together as a label for a particular object.

Algorithm brittleness occurs when the environment changes in such a way that the computer vision algorithm can no longer recognize the object due to some small perturbation. Brittleness for driverless car computer vision includes an inability to cope with changes caused by weather conditions. Lane markings that are partially covered by snow cause problems because the edges no longer match the system's internal model (Krishner 2019). Even on sunny days, when a tree branch or other vegetation partially obscures just a traffic sign, what is obvious to a human becomes impossible to interpret for a computer vision algorithm (Lewis 2019).

A common response to such brittleness is for computer scientists to gather more data in order to fill what is thought to be a perceptual gap. For example, to fix the vegetation-obscuring-a-sign problem, many engineers will say "We just need more examples to train the algorithm to correctly recognize this condition." While that is one answer, it begs the questions as to how much of this finger-in-the-dyke engineering is practical or even possible? The workload to do this is extremely high, which is one reason why there is such a talent drain in AI.

Because computer vision based on deep learning is still a relatively new area of research, new problems are coming to light in university laboratories. Researchers have only recently uncovered that neural nets are not capturing accurate depth information in images (Dijk and Croon 2019), which can have significant safety implications. A relatively new field of study has emerged in the past few years called adversarial machine learning, which examines how systems that leverage versions of deep learning algorithms can be tricked or defeated.

Progress in adversarial machine learning has been eye-opening as one set of researchers demonstrated that putting four innocuous black and white stickers on a stop sign could trick a computer vision algorithm to see a 45mph speed limit sign (Evtimov et al. 2017). Another set of researchers then went on to show only a single pixel needed to be changed to cause such an algorithm to mislabel an object (Su, Vargas, and Sakurai 2019). These recent efforts show just how vulnerable these machine learning-based approaches are in computer vision applications, and ultimately how nascent this field actually is.

Conclusion

Artificial intelligence in the form of machine learning has the potential to transform elements of many safety-critical applications and offer up new forms of human-computer collaboration that previously were out of reach. For example, one effort recently demonstrated that an AI-enabled robotic arm could assist the pilot of an airplane in non-essential mundane tasks (Aurora Flight Sciences 2016). This is especially important since there is currently a global pilot shortage and so this kind of human augmentation could free co-pilots to take captain roles and effectively double the workforce.

Even though AI has limits, particularly in safety-critical systems with potentially deadly edge cases, demanding perfection could limit the benefits of developing such technology. As in the case of the robot pilot arm or in the case of slow-speed driverless shuttles that operate in protected environments, there may be very advantageous uses of AI-enabled systems, even though the technology is not flawless. This then motivates the need to develop clear criteria and testing protocols so that companies and governments buying or approving AI-enabled systems can be sure that the proposed systems are capable of operating in their intended operational domains. History is replete with examples of how a lack of understanding of actual vs. desired operational readiness ended in costly system failure, Boeing is feeling this pain right now. Going forward, companies need to ensure that they have clear strategies for not just technology development but also testing across a set of edge cases so they can avoid AI brittleness surprises.

References

Aurora Flight Sciences. 2016. Aurora Demonstrates DARPA Aircraft Autonomy Program. Manassas, VA: Aurora Flight Sciences,.

Bubbers, Matt. 2019. "Don't hold your breath - fully autonomous cars are still decades away." The Globe and Mail.

Crowe, Steve. 2016. Tesla Autopilot Causes 2 More Accidents. Robotics Trends.

Dijk, Tom van, and Guido C.H.E. de Croon. 2019. "How do neural networks see depth in single images?" *arXiv:1905.07005*.

Elias, Jennifer. 2019. "Alphabet exec says self-driving cars 'have gone through a lot of hype,' but Google helped drive that hy." CNBC, accessed 22 NOV. https://www.cnbc.com/2019/10/23/alphabet-exec-admits-google-overhyped-self-driving-cars.html.

Evtimov, Ivan, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. 2017. "Robust Physical-World Attacks on Deep Learning Models." *arXiv preprint 1707.08945*.

Griggs, Troy, and Daisuke Wakabayashi. 2018. "How a Self-Driving Uber Killed a Pedestrian in Arizona." *New York Times*.

Harwell, Drew. 2018. "The Accent Gap." The Washington Post, accessed 2 Dec. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

Krishner, Tom. 2019. "5 reasons why autonomous cars aren't coming anytime soon." AP News, accessed 2 Dec. https://apnews.com/b67a0d6b6413406fb4121553cdf0b95a.

Lewis, R.K. 2019. "Reality is going to stall for some time the advent of driverless cars." The Washington Post. https://www.washingtonpost.com/realestate/reality-is-going-to-stall-for-some-time-the-advent-of-driverless-cars/2019/08/01/343c9458-afa8-11e9-a0c9-6d2d7818f3da_story.html.

Risen, Tom. 2016. Tesla Updates Radar in Wake of Autonomous Car Crashes. US News & World Report.

Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. "One Pixel Attack for Fooling Deep Neural Networks." *IEEE Transactions on Evolutionary Computation* 23 (5):828--841.

Walch, Kathleen. 2019. "Are We Heading For Another AI Winter Soon? ." Forbes, Inc. , accessed 15 DEC. https://www.forbes.com/sites/cognitiveworld/2019/10/20/are-we-heading-for-another-ai-winter-soon/#5347c1f256d6.