Chris Grimm

STAYING INSIDE THE ADVERSARIAL LOOP

Deepfake videos and their ability to create realistic fake news have recently drawn attention due to the numerous negative ramifications they could have on American and global society. These faked videos could spawn disinformation campaigns capable of disrupting the security of nations, the legitimacy of voting processes, or trust in national leaders (Harwell, 2019). Before Deepfakes, experts in deep learning were warning of the ease with which these algorithms could be tricked. In a seminal paper called "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", Nguyen et al demonstrated that high accuracy deep neural networks would classify images that looked like static as various objects such as backpacks and soccer balls with high confidence (2015). In just three years, this seemingly harmless insight into deep learning was being applied as adversarial stickers which could be used to trick self-driving cars into thinking a stop sign is a forty-five mile per hour speed limit sign (Eykholt et al, 2018). This technology would allow outwardly meaningless stickers to fool autonomous vehicles into behaving erratically and causing injury to others. These same sorts of tactics could be applied to a plethora of problems that are of concern to the Department of Defense (DoD).

Generative Adversarial Networks (GANs) are the primary method for producing Deepfake videos. Introduced in the paper "Generative Adversarial Networks" by Goodfellow et al., a GAN is described as two mirroring models trained by a common data set. These models are

a generative model which produces new data such as images and a discriminative model which determines whether the data fed to it is part of the original training data set or produced by the generative model. While the paper itself did not explore any applications beyond the ability to generate data that closely resembled the original dataset, Goodfellow et al. provided some theoretical areas of application including the possibility of semi-supervised learning. Semi-supervised learning could "improve performance of classifiers when limited labeled data is available" (2015). Instead of using GANs to augment the performance of classifiers, in "Generating Adversarial Examples with Adversarial Networks", Xiao et. al. demonstrated that GANs can be used to produce adversarial images capable of tricking deep learning algorithms (2018). These same techniques are now currently used to create Deepfake videos capable of tricking humans. Fortunately, similar deep learning algorithms can just as easily be trained to detect adversarial content. Guera and Delp demonstrated that deep learning methods could detect Deepfakes with an accuracy of 97% (2018). GANs and other adversarial methods have created a fight between those trying to implement deep learning to improve their performance and those trying to implement deep learning to degrade their adversary's performance. The United States must prepare for potential algorithmic challenges from their near-peer competitors. In the coming age of algorithmic warfare, the U. S. will have to pursue any advantage necessary to stay inside our enemy's adversarial loop.

"We are in an AI arms race". This remark was made by the Chief of the Algorithmic Warfare Cross-Function Team, Marine Corps Col. Drew Cukor. The Algorithmic Warfare Cross-Function Team is a specialized unit stood up under the Office of the Undersecretary of Defense for Intelligence to integrate artificial intelligence (AI) into the way the U.S. military fights (Pellerin, 2017). Algorithmic warfare must not only focus on how to implement AI for the DoD,

but also how these algorithms will defend against the numerous adversarial techniques that our competitors will surely use. These techniques could either prevent the DoD's application of cutting-edge deep learning algorithms or cause artificially intelligent systems to perform in harmful ways. As an example, AI algorithms used to detect and track targets in intelligence, surveillance, and reconnaissance imagery could be fooled into redirecting sensors to track open fields due to adversarial images. In order to defeat our enemies, the U.S. must look across the full spectrum of algorithmic warfare and ensure the integration between its various elements.

Like any other form of warfare, algorithmic warfare is ultimately fought using tactics. These tactics must be supported by algorithmic doctrine and strategy which are guided by the policies and laws of our country. Ultimately, algorithmic warfare is constrained by the feasibility of the technology. This technology includes the software, algorithms, operating systems, and virtual environments by which algorithmic warfare is conducted. Finally, software and algorithms are confined to the performance of the actual hardware it runs on. It is of critical importance that research, integration, and collaboration across all these elements occurs in order to win in an algorithmic warfighting domain. Algorithmic tactics without knowledge of the algorithms themselves could result in defeat if algorithms are expected to perform beyond their limits. Likewise, hardware and software must be developed that meets the requirements of the warfighter. As an example, hardware can enable strategic advantages. This can be demonstrated by the strategic guidance of John Boyd combined with hardware advances in the field of deep learning.

John Boyd's Observe, Orient, Decide, and Act (OODA) loop has often been used to describe the decision-making process required for defeating an enemy in aerial combat (Feloni & Pellison, 2017). This same decision-making process could easily apply to deep learning-based

algorithmic warfare. Deep learning has two major parts: training and inference. On the training side, deep learning algorithms are fed data to improve their performance on a task. Training often takes many weeks if not months and is very computationally intensive. After the algorithm has been trained, it can be implemented in a process called inference were data in real-time is passed into the system and a decision is made in a matter of milliseconds. Applied to the OODA loop of algorithmic warfare, training would be equivalent to orient in that our algorithms should be adapted to new data. Inference then would serve as the new decide where the faster the algorithm can be implemented, the faster the decision can lead to an action. If deep learning has an OODA loop of its own, in what ways are we able to increase the speed of our algorithmic OODA-loop?

Specialized hardware for the inference component of deep learning has been demonstrated to provide numerous benefits. In 2008, the Defense Advanced Research Project Agency (DARPA) started a project called Systems of Neuromorphic Adaptive Plastic Scalable Electronics (SyNAPSE) to create computer hardware that replicated the human brain. The result of this project was the creation of the IBM TrueNorth chip in 2014 which can run deep learning inference at orders of magnitude lower power consumption than Graphic Processing Units (Cassidy et al., 2016). The U.S. is not the only country that has dabbled in specialized chips for deep learning. Recently, China's Tianjic chip aided a driverless bicycle to act autonomously. The chips small size and power consumption critically enabled the implementation of deep learning on a platform as small as a bicycle (Ray, 2019). State actors are not the only ones investing money into deep learning specialized hardware. Google has developed a specialized chip that performs low-precision deep learning operations while "resulting in faster processing, less power consumption and, potentially more importantly, a dramatically smaller surface area for the actual

chip" (Lynley, 2018). A small start-up called Mythic has even developed specialized inference chips that exploit low-precision deep learning operations at incredibly low power consumption and high computational speeds (Hemsoth, 2018). On the extreme end, advanced research is being performed on binary neural network accelerators that would enable AI algorithms to make ultra-fast decisions in devices as small as smartphones (Valavi et. al., 2018). These and other specialized hardware implementations have enabled the use of deep learning inference in a variety of situations from bicycles and smartphones to the enormous amount of data coming through Google's datacenters. The benefits of implementing specialized hardware for deep learning training would potentially be just as advantageous if not more so for algorithmic warfare.

While these specialized inference chips help AI algorithms make decisions faster, at lower powers, and ultimately closer to the source of the data, defeating GANs and other adversarial methods will require faster training. The U.S. will have to orient and train its algorithms to false data we capture and identify the patterns associated across classes of false data. In order to close our algorithmic OODA loop ever tighter, we will need specialized hardware for training deep learning algorithms.

References

Cassidy, A. S., Sawada, J., Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Akopyan, F., … Modha, D. S. (2014). *TrueNorth: a High-Performance, Low-Power Neurosynaptic Processor for Multi-Sensory Perception, Action, and Cognition*. IBM Research.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., … Song, D. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. In *Computer Vision and Pattern Recognition*.

Feloni, R., & Pelisson, A. (2017, August 13). A retired Marine and elite fighter pilot breaks down the OODA Loop, the military decision-making process that guides 'every single thing' in life. *Business Insider*. Retrieved from https://www.businessinsider.com/ooda-loop-decision-making-2017-8

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Nets. In *Neural Information Processing*.

Guera, D., & Delp, E. J. (2018, November). Deepfake Video Detection Using Recurrent Neural Networks. In *IEEE International Conference on Advanced Video and Signal-based Surveillance*. Retrieved from https://engineering.purdue.edu/~dgueraco/content/deepfake.pdf

Harwell, D. (2019, June 12). Top AI researchers race to detect ‘deepfake’ videos: ‘We are outgunned’. *The Washington Post*. Retrieved from

https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/

Hemsoth, N. (2018, August 23). A Mythic Approach To Deep Learning Inference. *Next Platform*. Retrieved from https://www.nextplatform.com/2018/08/23/a-mythic-approach-to-deep-learning-inference/

Lynley, M. (2018, July 25). Google is making a fast specialized TPU chip for edge devices and a suite of services to support it. *TechCrunch*. Retrieved from https://techcrunch.com/2018/07/25/google-is-making-a-fast-specialized-tpu-chip-for-edge-devices-and-a-suite-of-services-to-support-it/

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Computer Vision and Pattern Recognition*.

Pellerin, C. (2017, July 21). Project Maven to Deploy Computer Algorithms to War Zone by Year's End. *U.S. Department of Defense*. Retrieved from https://www.defense.gov/Newsroom/News/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/

Ray, T. (2019, August 3). China pumps up the hype about A.I. with oddball computer chip. *ZDNet*.

Valavi, H., Ramadge, P. J., Nestler, E., & Verma, N. (2018, June). A Mixed-Signal Binarized Convolutional-Neural-Network Accelerator Integrating Dense Weight Storage and Multiplication for Reduced Data Movement. In *VLSI Symposium on Circuits*. Retrieved from http://www.princeton.edu/~nverma/VermaLabSite/Publications/2018/ValaviRamadgeNestlerVerma_VLSI18.pdf

Xiao, C., Li, B., Zhu, J., He, W., Liu, M., & Song, D. (2018). Generating Adversarial Examples with Adversarial Networks. In *Twenty-Seventh International Joint Conference on Artificial Intelligence*. Retrieved from https://www.ijcai.org/proceedings/2018/0543.pdf