Deep Learning in Depth: Adversarial Machine Learning

featuring Ritwik Gupta and Carson Sestili as Interviewed by Will Hayes

---

Welcome to the SEI Podcast Series, a production of the Carnegie Mellon University Software Engineering Institute. The SEI is a federally funded research and development center sponsored by the United States Department of Defense and housed here on the campus of Carnegie Mellon University. A transcript of today's podcast is posted on the SEI website at sei.cmu.edu/podcasts.

Ritwik Gupta: Ian Goodfellow and Nicolas Papernot, who were two scientists who kind of founded—maybe this is a hyperbolic claim—but they are two leading researchers in the field of adversarial machine learning. I know Ian Goodfellow his invention general adversarial network—was named one of the top 10 breakthroughs by MIT or something.

They just had critical insights that, *Hey, maybe these deep neural networks aren't as non-linear as we think they are. They are actually very linear, somehow*. Having that insight in mind and applying all sorts of math, computer science optimization, numerical computation knowledge to it, they basically made breakthroughs in the field of adversarial machine learning. And now they're doing amazing…Nicolas Papernot hasn't even finished his Ph.D. yet at Penn State. The way it is advancing, the way people can just come on in and decide to do amazing things is breathtaking.

Will Hayes: OK, so I can't let you go without explaining a little bit more about adversarial machine learning. What is that about?

Carson Sestili: Actually, could I take this? For the audience or maybe for the more lay person, or someone of my level right now, what Ritwik is talking about when he mentions adversarial machine learning is this really interesting way of poking a hole in the system that we thought was doing really well. If you take, this is an example, a neural network with very high accuracy labels images correctly on your image dataset. It turns out you can just alter a single pixel of an image and then convince it with extremely high accuracy that it is in a totally different label.

So, you can go from a dog to an airplane by taking one pixel and just messing it up.

Will: You covered this in your blog post. I remember.

Carson: Yes. OK, this is really important for people to know about because if you are going to make a self-driving tank that has to make a decision about where to go or where to shoot, it needs to make the right decision. It needs to make a decision that you can trust. Speaking to the adversarial nature, your opponents know that you are using this technology. And they are going to do everything in their power to corrupt that one pixel in order to block your gun, right?

Will: Or the pizza delivery drone that is coming to…

Carson: Yes! Sure. Yes, I want my pizza.

Ritwik: No, I'll take your pizza by messing with the pixels.

Will: That's a new form of security concern.

Ritwik: Yes, concerns everywhere. There was recently a paper which showed that if you stick a sticker on a stop sign—and these happen all the time, right? Sometimes you will see vandals or graffiti or a sticker put on something somewhere. By that sticker existing, image recognition algorithms were fooled 100 percent of the time that that is not a stop sign. They thought it was a speed limit sign or something else. So, you can imagine. It was this tiny sticker, and it changed the way it thought about the entire world. You can imagine a self-driving car driving and thinking the stop sign is actually a speed limit sign and speeding up and just causing a massive four-way collision. Basically what we are saying, and what Ian Goodfellow and all of the people in the field are saying, is that we make these overblown claims about the robustness and veracity of these algorithms.

There are so many flaws. People who make the claims that we are at the state of general artificial intelligence don't know what they're talking about. It is kind of like using a colander to scoop buckets of water out. There are so many holes that, even if it works well for one bucket, there is sure as hell not…

Will: A bucket of water moving will never be finished.

Ritwik: Yes, there's a lot of ways to break it.

Will: You are really pushing how we understand the utility of deep learning in the cybersecurity space. Given this as a background, can you talk a little more about your work there?

Carson: Yes, absolutely. So again, I work for a cybersecurity organization, and my research involves analyzing security in software at many stages in its development. This can be software

that you are writing that you don't want people to attack or it could be software that comes from somebody else who is trying to attack you. I can say that since deep learning has had such success in image recognition, which it really has, as much as we have succeeded in poking holes in it over the last couple of minutes, it really is doing great in that field. People are very excited. And they say, *How can we put this into other problem domains? I've got a terabyte of data, please, please, please*.

It works really well in certain scenarios. It does not work well in all scenarios. Some of my work in cybersecurity machine learning research is to see can we take this technology that has been working great and make it work on this problem domain? Code is a lot different from images. It turns out code is actually a lot different from even natural language, which is the dataset that people are claiming is similar enough to code for the same techniques to work.

My current work right now is to investigate where is that line? What is deep learning good at? What is it not good at? I believe that it is unethical to continue propagating the claim that deep learning is going to solve every problem. Because what that does is it wastes time basically. If you get your grant for a year for funding, and you claim that you are going to solve this problem, and you can't because the problem is not going to work that way.

Will: We might see a security firm who sells virus software look at a malware catalog and try to use these kinds of techniques to come up with a new understanding of what their products should contain.

Ritwik: Hopefully it is antivirus software.

Will: Yes, thank you, antivirus software. That seems a fairly straightforward surface-level place where we would expect it to work. Where might we be trying to make it work, and it is not working? How far away from this very vanilla example I misstated would we go with applying machine learning without revealing…

Carson: Yes, well sure. So even there, I didn't do this work, so I can say it. There are people who are using deep learning in the domain of malware detection. In classification, you get a new file. You want to know if it is malware. *Does it look like any other malware I have seen before? It kind of works, but there's also some limitations*. In particular, a couple of studies that I have read say we've got 98 percent accuracy. They used, I think, 15 different malware families. There is a lot more than 15 different malware families.

Ritwik: Again, it is important to state that these subject matter experts in malware analysis have developed a very good set of tools historically to tackle these problems and their own understanding. That is *combined* with deep learning to solve these problems. Deep learning is not

the only thing that they are using. By far it is not the only thing they are using. It comes with the toolkit.

Will: There might be knowledge about what other techniques, when combined with deep learning applications, have the utility of greater or less outcomes in particular fields. So, pairing with other tools and other perspectives is something we can learn about here.

Ritwik: Or even, what features do we know as researchers about that the deep learning model that it just physically cannot learn about by itself? We can kind of add that to the deep learning model and say, *OK, you have learned all of the stuff on your own but here is some really important stuff you should be looking at as well that you couldn't learn otherwise*.

Carson: I think the claim that it is always going to learn all the features that are useful is absolutely not supported. In fact, a good deep learning researcher will say, *Listen, what features can I give you? Please take that*, and also learn some more in case those were not good enough. I think it is really, you cannot only view yourself as a deep learning researcher. You need to view yourself as a machine learning researcher, a data scientist, and just say, *If I've got some features, I can give them to you if I know they matter.*

Ritwik: I do have to say along with that point is that, here at CMU, there is world-class research being done, not only in all sorts of deep learning techniques, but also in this case of representation learning. How do we learn better features, etc.? I don't think that a single researcher at the [CMU] School of Computer Science would disagree that we should just not include features that we already know about. If you know about it, put it in there in some way, shape or form. Even though with deep learning the claim to fame is it can learn everything by itself, why would you want to, right? If I could bootstrap you to do something, then by all means, please bootstrap me, right? Exactly what Carson said.

Carson: If nothing else, these models can learn to ignore the feature that you give it. So, it is fine.

Will: Is it reasonable to say one of the ways that we can take advantage of deep learning is to go beyond the intuition or past knowledge that we bring to the problem. It can help extend and perhaps create new reason to alter our intuition about what is going on.

Ritwik: Yes, and you can see this happening live with a game of Go, right? DeepMind with Google challenged Lee Sedol, maybe I am saying his name wrong, to a game of Go that was televised everywhere on the news. The machine beat, he was number two I think, the player of Go. And because of the moves that the machine made, people playing Go now are learning from the machine and changing the way that their intuition that Go works. They have had to learn the game of Go, not from a friend, not from a teacher, but from data.

Will: There is a blessing and a curse there when we talk about Carson's work in cybersecurity. If we pushed the boundaries of our understanding of vulnerabilities, do we then fence off some sort of vulnerabilities that the bad guys no longer pursue, knowing that these other things that the machine learning has helped us to uncover, that our intuition didn't previously cover, are we getting an immunity to a certain set of bacteria and allowing others to thrive by what we are doing? It is an interesting philosophical question maybe.

Carson: I think no matter what tool you are using, if you tell your enemies, *Here is how I am doing what I am doing*, they will come up with a way to use that against you. Maybe that's not a super insightful observation, but just…

Ritwik: Another point about learning immunity is, I would even make a stronger claim. By using deep learning to learn about vulnerabilities automatically, it goes both ways. Attackers can learn to exploit a system in ways that are not even intuitive to us at all. An adversary could learn a representation of a cyber physical system or whatever that is well supported by data, but would make no sense to us, and automatically find vulnerabilities and attack patterns that would not make sense to a human but would work together in tandem with each other. So, it goes both ways.

Will: People who build those systems might want to apply such a technique to assessing the robustness of the system they are building.

Ritwik: Yes, so red teams and pen testing teams around the world do this, right? They test themselves to make sure that they are as robust as possible. Deep learning, again, is another cybersecurity tool to be used in all sorts of applications.

Thank you for joining us. Links to resources mentioned in this podcast are available in our transcript. This podcast is available on the SEI website at sei.cmu.edu/podcasts and on Carnegie Mellon University's iTunes U site, as well as the SEI's YouTube channel. And as always, if you have questions, please do not hesitate to send us an email at info@sei.cmu.edu