Deep Learning in Depth: The Good, The Bad, and the Future

featuring Carson Sestili and Ritwik Gupta as Interviewed by Will Hayes

---

Will Hayes: Welcome to the SEI Podcast Series, a production of Carnegie Mellon University's Software Engineering Institute. The SEI is a federally funded research and development center funded by the United States Department of Defense and housed here on the campus of Carnegie Mellon University.

My name is Will Hayes. I am a principal engineer here at the Software Engineering Institute. It is my pleasure today to introduce two of my colleagues. Carson Sestili and Ritwik Gupta are here to talk about deep learning. Before we begin, could you tell us a little bit about your backgrounds, where you come from and what your interests are? Maybe can start with you Carson.

Carson Sestili: Yes, so I'm Carson. I'm from Pittsburgh Pennsylvania. I graduated from Carnegie Mellon with a degree in math in 2015. Since then, I worked at a brain imaging lab at the University of Pittsburgh for a little over a year. There I was really lucky to develop a lot of the skills that have contributed to my success as a data scientist at the SEI.

Ritwik Gupta: Yes, my name is Ritwik. I graduated from the University of Pittsburgh with a bachelor's in computer science in 2017. Carson and I have opposite paths. He started at CMU and ended up at Pitt, I started at Pitt and ended up at CMU. But my background is mainly in machine learning applied to the field of health medicine and health care. I have worked at places like the Department of Biomedical Informatics at Pittsburgh. I worked at Apple on various different things as an intern. So, very wide, varied experiences, but it has all led us to the same cool place that's the SEI.

Will: Great. You guys are both fairly early in your careers, and you've got some really, kind of high class education, and some very challenging problems to deal with I think. Why don't we start with a little bit of a definition of what deep learning is, and perhaps maybe a little bit about what it isn't?

Because I think there is a lot of material out there on the web for people to look at. Why don't we start with you, Ritwik and then go to you, Carson.

Ritwik: Sure. So, like you said, deep learning—there is a lot of information about it out there on the Internet— sometimes what people get confused about is *is deep learning different than what machine learning is?* I think it is very important to start at the start of things, which is, deep learning is a subset of the wider field of machine learning.

Generally it has been the case that a lot of people have heard of things such as SVMs and linear regressions, some basic—people call them *shallow models.* Now they are saying, *Oh, deep learning. That sounds deeper and better. So, it must obviously be more accurate and the end all, be all to everything*. That is not the case.

While traditional shallow learning, and what happens in that sense is, *I want to make some inference about some environment, or some state of the world that I want to learn about*. What I do is, I as a scientist, go out and I collect a set of features, right? An example of a set of features in this situation might be the amplitude of my voice, the tone, the pitch. I would have to specifically extract those features and put them in a data set that a computer can understand. Then we would have a simple model, or even a complicated model, that would then learn some stuff on that and do inference, regression, classification, whatever.

Compared to that, what deep learning is…Deep learning is traditionally three main things. One is that it is a composition of a series of non-linear filters. So, filter being the statistical term. So, any kind of non-linear function composed together end to end to end to end, that learns its own representation of the world.

So, unlike shallow learning, in which I would have to physically give you a set of features, "*Here is the features for you to learn on top of"*. The point of deep learning is that it will learn its own representation of the world that it's supposed to be learning. So, it would learn whatever is important for that specific environment and then do inference on top of it.

Will: So in the shallow learning example you started with, I would give you a recording of a voice and then I would indicate to you that, *when the voice sounds like this, it means this emotion is present. Or when the voice sounds like this, it means this communication is being intended*. Whereas in deep learning, you don't make that link for the system, the system derives that link?

Ritwik: So, that's a different part. What you are talking about in general is just the whole topic of dataset labeling and supervised versus unsupervised learning. What this is – I will build a simpler example. Let's say I'm observing bees flying back and forth from a hive to a flower, back and forth from a hive to a flower. I would tell the machine learning algorithm a feature set, such as the height of the bees' flight.

I would give it meters off the ground at certain points of the flight. I would tell it the color of the bee at different points – if it changes. It is a magical bee. I would tell it the distance to the flower. It is not examples of different bee flights, it is specific metrics about that environment. It can be metrics. It could be something more abstract, but I define a feature space for the environment, and then give that to the shallow learning model.

Carson: If I can unpack that a little bit with your audio understanding example. To think about the idea of what a feature is here, it would be reasonable for a human engineer to want to make a claim to the computer about maybe, *When the pitch of your voice takes this certain contour. Here is a set of 20 different contours that we, the humans, believe is meaningful for this language of speech*. Or maybe *this kind of filtering of white noise or something is useful for speech in this way. This kind of sentence structure is useful in this particular way*.

You are still going to give data to a machine learning algorithm or a deep learning algorithm. You are going to say, *Here is the data set, and here is the right answer*. But the difference is going to be, in a deep learning system, the reason they exist and the reason they shine is what is very hard to describe what the right features might be.

In this case, say I don't know anything about the language that I'm studying. I don't know that maybe a rising pitch at the end of the sentence has any kind of semantic meaning. But I do have a million labeled examples of a sentence that ends like that, and I know that there is something in common between them. The point of a deep learning algorithm is that it can infer the fact that that rising pitch was important. You did not have to tell the system that in the first place.

Will: Perhaps another example to make it even more obvious for our audience, when my mother saw my firstborn child, she looked down at the crib and said, "*Yes, that's one of mine"*. Was it because my child has the same nose I have? Was it the shape of the eyebrows? Was it the shape of the jawline? Those are things that we could talk about as features.

The difference between shallow learning and deep learning is, in shallow learning, we would tell the system, "*These are aspects that you need to care about"*. In deep learning, that's inferred. Just as my son had to infer what his grandmother looks like, never being told before he was verbal, that these are features that you care about. His way of learning about who that person is looking down at the crib was much more in line with deep learning than shallow learning, where I say, *When you're studying for this book report*, now that he's in sixth grade, *these are the kinds of thing you should attune to* and so I'm telling him features. It's a different kind of learning.

Ritwik: Correct. And what your son did, was learn a representation of his grandma, AKA representation learning, which is what deep learning models do, representation learning.

Carson: There's something else that Ritwik mentioned in his characterization of deep learning that I think is really good, which is this sequence of non-linear transformations. For me, as a mathematician, I perfectly understand what that means, but in case that isn't clear, what you are doing is…Passing through a filter will take your input data and change it slightly to exaggerate important parts of it. In an image, what that might do is to find edges, horizontal edges or vertical edges, to pick out what is the boundary of things. Or, it might actually be a blurring. If it turns out that your image has a lot of noise in it like TV static, blurring might be a really useful way to transform that image to get rid of the stuff that doesn't matter. You can do that once, but you can also do it 100 times. Every time that you are applying this transformation, you are pulling out the parts of the data that are interesting, that matter for the problem at hand. So that is really what a transformation is, but the non-linearities really gives it mathematical power to create good computation.

Will: So you get a wider range with each of the filters than you would with the linear application?

Ritwik: Correct. That is not to say that shallow learning doesn't do non-linear transformations. It is just that, let's say I have one non-linear transformation, like a sigmoid function, right. It kind of looks like that. Let's say we just have that as a shallow learning model. It can only represent data that kind of looks like that. But imagine I stacked, composed, a whole bunch of those together. The more non-linear filters that are composed together…you can imagine me putting kinks in a ruler, right? The more kinks I put in, the more I can approximate much more convex, very varied shaped functions, right?

That is the idea. As you compose more and more non-linear functions together, you can represent a much wider function space than you could with just one non-linear function. That is why deep learning is different from shallow learning. Shallow learning doesn't compose multiple things together. Deep learning does.

Will: As I was preparing for this podcast, one of the things I thought of is the process of communicating to somebody how someone looks. So if I was describing over the telephone to a distant cousin of mine what my mother's appearance is, because this distant cousin is meeting her at a train station, I would start by talking about features of the face that are known to be germane to such a description. That might be shallow learning. How does deep learning differ from that?

Ritwik: So in shallow learning—if I am using something like shallow learning to explain to you what my mom looks like—I would say, *She has a chin that looks like a 75-degree V. She has eyebrows that are about three centimeters wide and two centimeters thick, and a nose that is*

about this sharp. In deep learning, I would say, *Oh, she looks like me. She looks like my brother. She looks like my sister*.

Will: Ahh. Reference points.

Ritwik: Yes. What she would have to do is pull all of you together in her head and build her own representation of what your mom should look like. What are common features between you, your siblings, etc. that your mom would share that she can then identify your sister. That is shallow learning versus deep learning.

Will: You gentlemen both work for different elements of the Software Engineering Institute. You [Carson] are at CERT, and you [Ritwik] are with the Emerging Technology Center. But deep learning has applications in both of these. I imagine you are not necessarily just working on one project together. Could you talk a little bit about maybe what is happening at CERT, and then what's happening at ETC?

Carson: Yes. So, I think it is actually really important to impact the distinction because that took a long time for me to understand the meaning. CERT was actually the—please fact check me on this—but CERT was the United States' first cybersecurity group. It was formed in response to the Morris worm, which was a super bad computer virus.

All the projects that we do, in order to get funding, need to be pitched under the lens of, *This is useful for software security for the defense of the United States*. So all of my data science work, and all my machine learning work is involved with a cybersecurity application. Right now I am working on projects that investigate the utility of machine learning for code in various parts of its development cycle, for analysis of software in various parts of its development cycle.

Will: So, new frontiers that you push are frontiers relating to cybersecurity and really have a nice focusing effect on what you are doing. And you still get to work with a colleague who has a different filtering effect based on his affiliation.

Carson: Exactly. I think because I had this image-processing background from the brain imaging lab, I was fortunate to have the opportunity to work with Ritwik who can now tell you about what the ETC does.

Ritwik: Sure, yeah. Again, CERT does some really, really cool world class cybersecurity stuff. So, the ETC is a bit different. We are very new to SEI. We were founded in 2013. Our focuses lie in three main areas, them being human machine interaction/machine emotional intelligence *(How do we get a machine to better understand emotion, the emotional state of a human, we can work better together) and* applied artificial intelligence/machine learning. So, this is AI applied to some task in the world. So, that is a very broad description. We tend to not do cybersecurity,

because that in CERT's domain. So, we do things like satellite imagery, voice articulometry, etc. And then, the third one is advanced computing. So, *how do we push the fabric of computing further*? So, as the paradigms of computing change from CPUs to SIMD architectures and GPUs and TPUs, what's next? We are working on the *what's next* as well. So, that's our focus areas. You can imagine, there is a lot of applications for deep learning in all of those aspects.

Will: One of the most interesting recent applications comes from the Intelligence Advanced Research Project Activity's work on Functional Map of the World Challenge. And you gentlemen were both part of that?

Carson: That is correct.

Will: Tell us a little bit about that because that sounds like a neat project.

Carson: We refer to it as IARPA. If you are interested, you can always check out a link in the transcript of the IARPA Functional Map of the World Challenge. This was an image-recognition challenge on steroids. The problem was the United States has a lot of satellites that are fixed above the earth and has an overwhelming amount of satellite imagery of portions of the earth's surface.

What they are interested in doing is finding out what is going on on these plots of land. There are various functions that can be ascribed to buildings or facilities on the surface of the earth, like airport, amusement park, nuclear facility, or hospital. It is really important to be able to tell what is in this chunk of land for lots of intelligence-related reasons. There were several complications to the challenge that made it more interesting than just the standard, *Does this photo have a dog or a cat in it?* But, it was in spirit a very similar approach.

Ritwik: I do want to add that it is not just identifying the functional buildings, it is also identifying a random piece of land. *So, is this a crop field, or is this a flooded road ?* Again there are a variety of applications for this. IARPA obviously would be intelligence based, but there are also large amounts of humanitarian use cases for this stuff.

One distinct thing I can imagine for a really heavy use case of satellite imagery was there was a city, Fort-something in Canada, which was ravaged by wildfires a year or two ago. One of the things that happened was, when humanitarian workers went in to do rescue operations and to provide aid, they did not know where structures stood anymore. They had to use satellite imagery to identify what…like this square on the map now was actually a hospital before, et cetera. There are a lot of use cases for stuff like this.

Will: So it could help them better navigate hazards. It can help them understand where there might not be anything solid to stand on, even though it is not apparent here where there might be more people in need of rescue.

Ritwik: Exactly.

Carson: There is not a barn here anymore because it burned down, but there used to be one and, we need to know that.

Ritwik: We can tell you that because we know what the function of the land is. Therefore the Functional Map of the World.

Will: Could you talk a little bit more about the collaboration you had on the Functional Map of the World project?

Ritwik: I can start off on that. Our goal for entering the challenge was kind of twofold. One was to basically work on a fun little problem that had real use cases for the United States government. The other one was to basically try out very new methods ourselves and find limitations within our own infrastructure, within our own methodology. There is kind of a two-pronged way. It was not necessary to win the challenge.

Again, the task is, you are given 62 different possible functions of the world, plus one false detection category. That means that this was just bad label data. That task was that you are given about five terabytes of data of which about 65 percent is training data. Training data does not contain false detections. Can you using that data in the best manner possible identify the land use of the satellite imagery? What we did was, we tried a variety of methods to do that: variety of existing methods, deep learning methods, you know, using models such as DenseNets, using models such as SENets, that's squeeze-and-excitation networks.

Carson: So if I can interrupt for a second, a very high-level view of this is that, this is just an image recognition challenge, which is what deep learning kind of came to fruition in proving to the world that it was good at. Image recognition is like the slam-dunk success of deep learning so far. We said, not only we, the people who instantiated the challenge said, *Here's a starter deep learning architecture*. We used a variety of different deep learning methods that have been used for image recognition in the past. Other such challenges were given on the set of labeling all the images that made it on to Google Images. The technology behind this is convolutional neural networks, which is a really good way to get into deep learning if you are interested in learning more.

Ritwik: Again, Carson previously mentioned that this is not standard image recognition tasks, it is not standard image recognition classification tasks. That is because generally when we are

talking about categories, like classification, it is like dogs versus cats, or it might be something that's very close, like Porsche versus Lamborghini, or certain sports cars. This one is unique because you have satellite imagery, which for a large part of the earth is very homogeneous, right? One patch of forest in the USA might look exactly like a patch of forest in the Amazon, right? Or a city from like a zoomed-out view might look the same as the next city.

The idea is how can you take these very minute differences, not only in scale, but also in landscape, the buildings on there, etc., and identify different land functions. This makes it very different from just a traditional image classification problem because you have to take in not only the object of interest, which is like let's say a building, but also its entire surroundings.

Will: For example, what a farm looks like in Kansas versus what a farm looks like in New Delhi? They are very different things. Not only are they growing different crops, but the size of the fields, the machines used, the seasonal activity that happens are very different. This challenge needs to be able to accommodate those kinds of sources of variation.

Ritwik: Yes, like crop fields in Kansas have irrigation circles, right? It is a very modern technology, these irrigation circles. Crop fields in New Delhi don't have that. They are usually hand-watered or they have pipes running in the field. On satellite imagery, those would look very different from each other. So, identifying something like that is important, but you do not really get that in just general deep learning tasks.

Will: The feature learning task is what this challenge addresses.

Ritwik: This is the most critical part of this challenge, yes.

Carson: If I can pull that back to our feature representation discussion from earlier, unless you are a world expert in what satellite imagery should look like, or you know a lot about farming techniques or whatever, it could take you years to determine this is what a crop field in Kansas is going to look like, this is what one in New Delhi is going to look like. You have terabytes of data, you don't have enough time in order to make that happen. That was the whole reason that we needed to use this deep learning strategy is because we are not experts on what satellite images look like, but the computer can become one.

Ritwik: It is not to say that we are completely clueless about satellite imagery. We have some idea, which we try to incorporate in the models. The idea is, it can take our cluelessness, or some naive knowledge of it and build on top of it and become really good at identifying all these things.

Will: It is the power of what the algorithms and the techniques are able to do beyond whatever the human brings to it, that we are really trying to test with this challenge.

Ritwik: Correct. I think one of the biggest things is not only what can it do beyond humans, but the computation power and the infrastructure that lies behind the deep learning, which really empowers it. We are really good at doing all these cool abstract tasks, we being humans, like identifying what your grandmother looks like with one look at your grandmother, or reasoning about complex things, like *Oh, that is a bookshelf*', and *Looking at those books, we're probably in a Software Engineering Institute*. What the computers are really good at is doing one thing really well, and fast, really fast, much faster than humans can.

So how we best leverage all that architecture, the infrastructure that has been built around, or has already existed for various different tasks to work for deep learning. A part of the challenge that we really focused on was, *How do you best create an infrastructure that facilitates this deep learning task?* There are a lot of small research centers. There are a lot of big research centers, which have a lot of heterogeneous hardware out there. All of it may not be best suited for machine learning. One of the best things that we did focus on for this research was, how do we best help people in those situations create deep learning stacks that would work really well at performing and best facilitate the learning task at hand.

Carson: I think this actually calls to mind a really interesting part of this project and something that was a great idea that the ETC data science group and the CERT data science group were able to come together on. So Ritwik faced this. *We have got terabytes of data.* It took a week to download. You were running up the bare metal. The electrons were causing you problems sometimes. From my perspective as a mathematician and in the CERT data science group, we mostly have a statistics background.

So, I don't know how computers work at all, but I know kind of a lot about how linear algebra works and how the theoretical gains that need to be addressed and made to do better on this project. It was actually really awesome to be able to realize that we needed both of those perspectives on the project. A group that is trying to do a project like this themselves needs to have people that have both of those skills. If your data is huge, you can't go without someone who knows about the electrons. Also, if your data is huge, you can't go without someone who knows about the math. You need both.

Ritwik: You can see that big groups like Google, Microsoft, Apple, do really well at this, right? They have people who are cloud engineers, who are really good at the infrastructure, and you have people who are just deep learning scientists, or machine learning scientists, who are really good at the math. They work together in one lab to make themselves the best AI labs in the world. Even if you are not at that skill, even if you are not at Google, even if you are just a three-person team, it is essential to be able to identify your underlying infrastructure limitations and identify workarounds to those problems so that both tasks can be best facilitated. There are

tradeoffs that come with both of those, both on a learning side and on an infrastructure side. You have to identify the best tradeoff for your situation.

Will: So there are features to be discovered in what infrastructure is needed?

Ritwik: Right. Part of the research that we did focus on is, *What kind of tradeoffs would you have to make to do this?* Again, as Software Engineering Institute, we focus on the software engineering part of it. What engineering challenges are there to architect a system that would work on…This is a relatively small dataset, small being five terabytes, right? Because there are data sets that are much bigger than that. But even at five terabytes, how do you handle a dataset of that size? We focused heavily on that research.

Will: You have talked a lot about involving different kinds of folks, people with different perspectives. One of the great things about being at Carnegie Mellon is people come to an institution like this to offer their insights. You have got lots of experience being a student here. Could you talk a bit more about the wider net in this field?

Carson: Yes. First, I guess what I'd like to say, is if you are coming out of a technical background, especially a math background or one where if you don't consider yourself a computer scientist or a programmer, it is OK. There is hope for you. As I was coming out with my math degree, I was going to these career fairs and I was lining up to these big tech companies and feeling like nobody wanted me because their skill set that they were focusing on was one that was a little bit different from what I had.

If you feel like, *I am an OK programmer, but I am a good deep thinker and a good mathematician,* that is actually one of the corners of what it takes to be a successful data scientist. Again, in regard to our previous conversation, you cannot get away with only knowing math. But if you do know math, you are going to be useful to people in a way that other people will not be. Anyway, there is hope.

Will: All the math majors out there?

Ritwik: Right. We want people like Carson. To build on top of that too, there is a wide spectrum of data science. As you said, it is getting really, really hot recently. There is a wide spectrum of data science work, right? There is a lot of, *Let's just sit down and program a lot of stuff and glue a bunch of libraries together and do some really basic data analysis*, which is really, really fundamental work that needs to be done for a lot of companies. Then there is all the stuff like, *Let's invent a whole new field of math to solve this one specific problem.* And there is a whole spectrum in between.

As you can obviously tell, what that needs is a whole variety of people with a whole spectrum of talents to work on that. So, for example, in our team, the Emerging Technology Center, I have a computer science/biochemistry background. We have other people on our team who have a Ph.D. in physics. We have people who have a bachelor's in political science. We all work together to create a really, really impactful machine learning/computer science team. What is important is that even though deep learning is supposed to learn all these things by itself, it can't.

You need experts who are really good at all of these varied fields, subject matter experts, right? If were trying to solve a bio problem, we need a bio-subject-matter expert to be able to tell you when things go wrong. Let's say a deep learning model learns something, and it performs well, there is an accuracy. But let's say it just learned the completely wrong…behind the scenes, it is learning some confounding variable, you need an expert to say, *Hold up. That's not it. There is something else going on here that you should be looking at*. It takes a whole team to do this. It may be hard for someone who is not experienced in computer science or deep learning or machine learning to break into this stuff. But with the democratization of machine learning and the speed at which it is going, it is very easy to get into. I would say—again, this kind of builds on your earlier point of this field is advancing very rapidly—deep learning is not the solution to all problems.

There may be researchers out there that disagrees with me. They say, *No deep learning is the way*. If I may recommend to the audience, and we will include this in the transcript, a paper by Gary Marcus from New York University talking about the limitations of deep learning. There is lots of work to be done to get truly to artificial general intelligence. Deep learning is, at least in my opinion, a very small piece of that. We need people from all backgrounds—cognitive psychology, biology, chemistry, math, physics—everyone to come together and work towards solving this problem.

Will: That helps to link the evolving techniques and technology to the correct things, but it helps to focus which direction they evolve towards as well.

Ritwik: Yes. We don't live in this hubris like, *Computer science and machine learning can solve everything*. No, these fields exist for a reason, right? We need subject matter experts from these fields to guide us, to tell us what we are supposed to be doing, and to actually tell us what are the problems we need to be solving, right? It takes a village. It takes a village to do something well.

Will: OK. I want to focus in on an effort that you are undertaking, Carson, where you are looking really at in the realm of cybersecurity, what are the places where there is a sweet spot for machine learning and maybe where there are less fruitful avenues. Can you talk a little about your ongoing research there?

Carson: Yes, I will actually start a little bit broader than that. To jump off of something that Ritwik was mentioning, which was the relationship between deep learning research and artificial intelligence research that is happening today. Artificial intelligence is a word that people love to use, and I feel like every person who uses it uses it slightly differently depending on what science fiction books you read, when news broadcast you watch, what blogs you follow.

What is great about it is there are so many experts in the field who disagree with each other. It is really exciting to read the latest either artificial intelligence blog post or deep learning blog post and see, *Here is the way in which these experts are disagreeing today*. Research advances are happening every single day. It is kind of breathtaking and wild and quite difficult to follow. I think if anyone tells you they know how AI is shaping up, I believe they are lying to you. I believe that no one person really understands what's happening in the field right now.

Will: It is impossible to keep up with everything. There is so much going on.

Ritwik: You might very well know how your local field is shaping up, but it is very hard unless you are a luminary in the field who talks to everyone, to really know which way AI is going. It is very hard to know because AI again has many definitions across many different fields.

Will: One of the best minds Carnegie Mellon ever had was a gentleman named Herb Simon, you may have heard of him. I had the great pleasure of taking a class with him. He said, *I don't read anything. All my friends do. They tell me what the most important recent publications are*. I think you're getting on what Herb was talking about.

Carson: Yes, it is good to have good content aggregators. There's a lot of noise. It is actually really good to have smart friends who can tell you where the signal is. What is really interesting—and I talk about this in the blog post that I wrote a few weeks ago—is that knowledge is coming out of places where you might not expect.

There are people who were previously totally unheard of researchers who discovered something really interesting about deep learning and wrote a blog post about it on Medium. That is the article that gets a shared and then gets cited in somebody's research paper two months later. It is interesting because there is this haystack of ideas that people think work, and there are a few needles in them of actually good ideas. It requires skill and discretion to sort through it I think.

Ritwik: And diversity of thought behind it, right? You need to be able to think about a lot of things and how all things interact together to get that new insight. One of my favorite—I do not want to call it underdog story—but as you said like scientists who were not really popular, and all of the sudden they have made a discovery that is amazing, 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: OK, so I can't let you go without explaining a little bit more about adversarial machine learning. What is that about?

Carson: 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.

Will: You guys kind of add a focal point of some really interesting advancements here. What is about to come out? What is about to break? What should our audience be looking for?

Carson: Can we tell them anything?

Ritwik: I would say this again. We make this point again. If anyone says, *Here is what is nex*t, they are probably not going to tell you. I will tell you what I believe is next in specific things that I am interested in. This is a disclaimer to anyone who is watching out there. These are the things that I am interested in, and what I think are big. My personal focus has always been how can we use machine learning, deep learning, and statistical techniques to better improve the human condition. That usually turns to health, medicine, and healthcare systems. I believe that there are going to be revolutions in using, and there already have been, in using deep learning and machine learning to automatically detect cancer from CT scans, from radiology scans.

There are going to be revolutions in automatically discovering drugs to cure a certain specific disease and automatically detecting and discovering again mechanisms of drug behavior or disease behavior that are not intuitive to humans, but a machine can discover with data that will change the way that we look at treating diseases or curing things. There will be massive changes in electronic health record processing and continuity of care.

Will: I heard somebody say that they can detect with some percentage of confidence opioid addiction looking at the eye of a person walking by a camera.

Ritwik: Correct, and Google just made a claim recently. Jeff Dean, he is the chief scientist, he is the guy for machine learning at Google, just tweeted out some great work that his team did, which is they make a claim (they are still exploring this) that by looking at retinal images, which Google has a good history of doing work with, they can detect cardiovascular disease. There is an insane amount of work and disruption to be done in the field of health, medicine, healthcare. I think it is a great way to apply machine learning. There is tons of research being done here, not only at CMU, but also at SEI, in fields that are not only computational biology and health, but also related fields.

Will: That is a pretty awesome vision for the future. Are you going to try to top it?

Carson: Actually, instead of good cop/bad cop, what I am going to do is exciting cop/boring cop. I am such a wet blanket here, but I spend a lot of time talking to my team who has an extensive background in statistics. We are frequently talking about—not only we, but a lot of the community at large—is focusing on, *How can we quantify the uncertainty that comes out of these models, and what do we do with that uncertainty?*

It is common to just interpret the output of machine learning models as probabilities. They say, *Oh, my model says 70 percent. I guess that is pretty confident*. But that is just a number. And it is a number with a super sketchy statistical and mathematical underpinning. So I think now, there has actually been a paper—I think it's the one you referenced earlier saying, *Hang on a second. It has been five years. We have had a great success. Let's take a step back and think about how we do be responsible with the results of these machine learning techniques. How do we be socially responsible in making sure that they are deployed in ways that are trustworthy, that are investigatable, that you can ask them why they made such a decision*? Also how do you be mathematically responsible? How do you report uncertainty? How do you make sure that your audience knows…Again, it is about how much they can trust you. I don't think that people need to be worried yet about A*re the robots going to kill me or take my job*? I do think people need to worry about *Are the policymakers going to make a decision based on a bad statistics paper or something or a bad machine learning paper?*

Ritwik: One thing, again so tying this back to the health point, one thing that Carson said here, is that one of the biggest things holding deep learning back in the field of medicine and health is, everything a doctor does has to be backed by his decision. He has to say, *I am doing this because*. The reason they can't do that with deep learning models yet, is because a deep learning model cannot tell you why it did that. So if a doctor says, *Oh I performed the surgery because a deep learning model told me to*, that is when they get sued for medical malpractice.

Will: You cannot audit the deep learning models the same way you can audit…

Ritwik: And there again, there is amazing research being done all over the world on how we make what is called explainable AI. How do we explain what the deep learning is trying to do? This is a problem across all fields. We have kind of approached this point where we have gone so much in deep learning that we have kind of gone away from solid statistical underpinning. Not to say that there isn't, but there's just some things that are being done without saying like, *Here is all the statistical proof behind it*. I think it is very important to step back a bit and just look at just deep learning itself and say, *Hold up. How do we make you better and not apply any of it to any other field, just how do we make the field itself better*?

Will: Thanks very much guys. This has been really interesting.

Carson and Ritwik: Thank you for having us.

Will: As always, a transcript of this podcast is available along with the podcast recording itself on the SEI's website at sei.cmu.edu/podcasts. You can also find this 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