EPISODE 77

[INTRODUCTION]

[0:00:10.8] SC: Hello and welcome to another episode of TWiML Talk, the podcast where I interview interesting people doing interesting things in machine learning and artificial intelligence. I'm your host, Sam Charrington.

This week on the podcasts, we're running a series of shows consisting of conversations with some of the impressive speakers from an event called the AI Summit in New York City. The theme of that event and of this series is AI in the enterprise. I think you'll find this series really interesting, and that it includes a mix of both technical and case study-oriented discussions.

Now, I won't actually be attending the AI Summit this week, because I'm in Long Beach, California attending the NIPS Conference. There are a bunch of TWiML listeners here, and I'm hoping to meet as many of you as possible. Yes, I have stickers.

If you're here at NIPS and you're actually listening to podcasts this week, please reach out to me either via the event app, the NIPS event app where there is a TWiML listeners' thread, or via Twitter where my handle is @samcharrington.

Before we proceed, let's quickly talk about the podcast schedule through the end of the year. If you subscribe to my newsletter you know that I've been on the road for a couple of weeks now. After this week's series, we've got two more series coming before we break for the year with our last show running on December 22nd.

Now if you're lamenting two weeks without your favorite Machine Learning and AI Podcast, trust me with these two series, the first from the Amazon web services reinvent conference and the next one from NIPS, you will have plenty of great content to tie you over until we get started again on January 8th.

Thanks to you, 2017 was a great year for the podcast and we plan to close it out strong. Keep your ears open the next few weeks and we hope to hear from you.

Please note that on Wednesday, December 13th we'll be holding our last TWiML online meetup of the year. Bring your thoughts on the top machine learning and AI stories of 2017 for our discussion segment.

For our main presentation, Bruno Goncalvez will be presenting the paper Understanding Deep Learning Requires Rethinking Generalization by Chiyuan Zhang from MIT and Google Brain and others. You can find more details and register for the meetup at twimlai.com/meetup.

This AI Summit series is brought to you by our friends at IBM Power Systems. IBM Power Systems offer server's design for mission critical applications and emerging workloads, including artificial intelligence, machine learning, deep learning, advanced analytics and high-performance computing.

IBM Power Systems benefit from a wide range of open technologies, many stemming from collaboration with fellow Open Power Foundation members and their design to deliver performance efficiently, whether deployed in private, public, or hybrid clouds.

To learn more about the IBM Power System AC922 platform for enterprise AI, visit twimlai.com/ibmpower.

My guest for this first show in the AI Summit series is Hillery Hunter, IBM Fellow and Director of the Accelerated Cognitive Infrastructure Group at IBM's T.J. Watson Research Center. Hillery and I met a few weeks back in New York City and I am really glad we were able to get her on the show.

Hillery joins us to discuss her team's research into distributed deep learning, which was recently released as the power AI distributed deep learning communication library, or DDL. In my conversation with Hillery, we discussed the purpose and technical architecture of the DDL, its multi-ring topology, its ability to offer synchronous distributed training of deep learning models and much more.

This is for sure a nerd alert pod, especially for the performance and hardware geeks among us. Be sure to post any feedback or questions you may have to the show notes page, which you'll find at the twimlai.com/talk/77.

Now, on to the show.

[INTERVIEW]

[0:04:32.7] SM: All right, everyone. I am on the line with Hillery Hunter. Hillery is IBM Fellow and Director of the Accelerated Cognitive Infrastructure Group at IBM's T.J. Watson Research Center. Hillery, welcome to This Week in Machine Learning and AI.

[0:04:48.1] HH: Thank you so much. Very excited to be here.

[0:04:50.9] SM: I'm excited to have you on the show as well. Folks won't know this until I tell, but we had an opportunity to meet just a few weeks ago in New York City at the NYU at a reception held in conjunction with the NYU Future Labs event, and it was certainly great to meet you in person and even better to have an opportunity to get you on the line and dig into some of the work that you've been up to.

[0:05:16.3] HH: Yeah, it was a pleasure meeting you there and it was certainly a really interesting event and a great opportunity to see some of the exciting things going on in the New York area and AI and AI is just exploding everywhere. But it's a pleasure to be here on your podcast. I look forward to our discussion.

[0:05:32.0] SM: Awesome. Why don't we get started by having you tell us a little bit about how you – your background and how you gotten involved in artificial intelligence.

[0:05:41.1] HH: Yeah, it's interesting. I like to say that AI really has exploded for two reasons. One being the amount of data and especially publicly available data sets, and the second being the compute. I come from a background technically that's really a mix of both of those things. AI has been an opportunity to bring together a lot of different things that I've done in prior technical work and a lot of different things done by members of my team and prior technical lives before getting into AI. I come from a systems perspective, from a hardware perspective. I was an electrical engineer by training.

[0:06:18.1] SM: Go double E.

[0:06:19.7] HH: Yup. I was a double E. That background is where I'm coming from in approaching these problems that we're trying to tackle in AI. I bring to the table background and performance, a background in data movement and systems and both are proving to be really

fruitful for getting some of the grand challenge scale problems down that we're facing today in AI.

[0:06:44.6] SM: Awesome. Have you spent most of your career at IBM?

[0:06:47.9] HH: I have. Yeah. I got my PhD at University of Illinois and started at IBM in 2005. I've been with IBM research since then.

[0:06:57.3] SM: Fantastic. Have you been working on in the same group working on accelerated cognitive infrastructure, or have you done – evolved to that – this particular position?

[0:07:08.7] HH: Yeah. This position has evolved. We really ramped up our efforts around accelerated computing a number of years ago. Prior to that, I was working on things related to processor design and memory technologies. It's interesting, because again the memory relates to the data movement, relates to feeding the AI compute, feeding the accelerator. It's all come together in a really nice way.

[0:07:33.3] SM: Great. Your group recently published some really interesting research on essentially scaling deep learning performance using distributed techniques. That was one of the big things that I wanted to spend some time talking about in this interview. Can you give us an overview of that research?

[0:07:53.4] HH: Yeah. We were really excited by what we're able to publish. We're calling it the Power AI DDL, or Distributed Deep Learning capability. Basically what we showed is that we were able to create framework independence, so independent from Tensorflow, or Caffe, or PiTorch or your favorite way of doing deep learning.

We were able to create a framework independent communication library that achieves close to optimal – very, very close to optimal, up to 95% scaling efficiency. What this meant for us is that we were able to use lots of GPUs together very, very efficiently and we were able to use those mechanisms to train a neural network, to train a res-net to the highest published accuracy on a really hard problem, and we were also able to be what had been shown as a record of around an hour or 66 minutes on a smaller neural network and a smaller problem.

We really were able to show that through hardware-software integration, we are able to have world-class AI capabilities. For us, this really signaled a changed in the productivity curve for deep learning, because most of the open source today just handles a single node worth of performance, so you're stuck then. You can have two GPUs, four GPUs and maybe eight.

Scaling out to many, many servers has been really big challenge. The more servers and the more hardware you can use for a problem, the faster you get that work done and the more productive you are.

[0:09:30.0] SM: Now that last point might be one that's worth underscoring. If you look at framework like Tensorflow, there is – as part of the open source Tensorflow, there is a distributed Tensorflow. But that's more useful for scaling across GPUs than it is across servers, is that correct?

[0:09:52.3] HH: There is a version of distributed Tensorflow. What we really are looking at in that versus the DDL capabilities is the extent of scalability and ultimately the productivity of the server. The default distributions in Tensorflow haven't been shown to be able to use as many as 256 GPUs as far as I'm aware and based on our own internal studies. There are X factors to be had with every extras out of servers that you can add, you get the work done faster.

Also we have shown that our communication patterns and communication overheads are really close to optimal. The other things that are out there appear to be less efficient, and so that means at the end a longer time to solution.

[0:10:42.7] SM: Okay. The 256 GPUs that you are running was across how many systems?

[0:10:48.6] HH: Yeah. We ran across 64 servers and we were using the Pascal P100 GPUs. The work that we did was right before, right on the cusp of Volta coming out and this is the very latest GPUs, and I like to always describe why this is a hard problem, because the GPUs are so fast that they all learn very quickly. If you think of there being 256 learners in the system, the hard part is keeping everybody synced up.

That process of keeping them synced up, it's really critical that that be done with as we'll wait and see the communication as possible. That's really the core of the technology that we showed is the communication time is really, really low. That enables you then to do a fully synchronous

training, meaning everyone is updating everyone else with all the information that's being learned. All the waits are being updated as they should after every batch.

That means at the end that you're doing a type of deep learning that is opposed to asynchronous, where they're occasionally updating each other in order to lower the communication overhead. When you're able to do fully synchronous, you're able to keep everything moving forward a little bit more predictably and you're able to get a higher accuracy result at the end in general.

[0:12:06.0] SM: Okay. In introducing this you said it's called Power AI DDL. For those who aren't familiar with power AI, what is power AI and to what extent are the results that you demonstrated here tie to that power AI architecture?

[0:12:23.7] HH: Yeah. For us, this is very much an effort of hardware and software co-design. Power AI runs on the IBM SC822LC servers – sorry, actually S822LC servers. Those servers have two power processors and they have four GPUs. Everything is connected by doubled up NVLinks. Nvidia has this high bandwidth interconnect capability and we have that high bandwidth connectivity, not just between the GPUs, but also back to the host processor.

That provides extra provides performance in moving data, in moving wait updates and everything in the system. Our communication library also leverages all of that bandwidth to its max and to its full potential.

When we talk about this whole space of deep learning and what we're doing from a system's perspective and with things like DDL, we're talking about matching the software to fully utilize the hardware capabilities. For us, this is also a storyline around collaboration between our research division that I actually report in and our product division, our development division, because we were able to actually put this codebase out for IBM customers to try and download and try themselves, try DDL on their servers, or on the cloud at the same time that we made the publication and the announcement of our leadership deep learning capabilities using this framework.

Power AI, it's a download and go, instead of distributions of frameworks that run on our accelerated servers. We now have the distributed deep learning capabilities available there as well for customers to try themselves.

We really from a research perspective were very excited about this, because it means that we can take this rarified skillset of distributed deep learning, this grand challenge thing everyone is competing over and put it in the hands of our customers to go ahead and try out and see what they can do with it on their data, with their text and neural networks.

[0:14:33.3] SM: You mentioned that the results that you saw are framework independent. Is that true in the strictest sense, or is it rather that the software that you wrote was written to adapt to some fixed number of frameworks?

[0:14:53.3] HH: It's pretty true in a strict sense, but let me define strict to make sure we're not talking past each other. What DDL is is it's a communication library. It lets things in a system talk to each other. We have been able to use that communication library across many different frameworks. What we released is already is part of Power AI was our integration into Tensorflow and into Caffe. But we also published results using our integration into Torch and we have other integrations that we have shown work just fine as well for our internal use currently.

We have shown I think at this point integration into enough frameworks that I'm pretty comfortable saying that this library can be integrated into pretty much anything that you want.

[0:15:39.0] SM: Right, right. Yeah, I think in the context of my question the answer is yes and yes, right?

[0:15:43.3] HH: Yes. Yes and yes. There we go.

[0:15:47.1]SM: Interesting. How does this compare to some of the previous and even subsequent work, like you refer to Facebook and Microsoft work in the paper or in a blog post. I think since you posted this, Uber published a open source project called Horva that seeks to do something similar. Are you familiar with that one?

[0:16:15.6] HH: Yeah, I am. That's obviously a great result that they put out there. We love seeing all the different efforts that are going on in this space, because it really is such an important area for productivity. The Horva team showed a great set of scaling experiments with their integration essentially of an MPI reduce.

[0:16:37.3] SM: It as open MPI, right.

[0:16:39.4] HH: MPI reduce and the Nickel libraries into Tensorflow. We do believe that our communication topologies are a bit better than what is there, so then that scaling efficiency will be better and we look forward to being able to talk about that in little bit more concrete detail pretty soon here.

I think it's great to see these different efforts to our distributed deep learning happening across different types of frameworks, because ultimately the community needs to have this type of productivity in order for deep learning really to take off. For us though, the positioning of Power AI is really about taking open source and taking the complexity of managing it, of installing it, tuning up the performance, optimizing it for our systems and then ultimately providing support on it for our customers.

We love to see improvements in open source and much of what it's provided in Power AI is open source and is based on open source. Then we're providing improvements on top of that and optimizations on top of that as well.

We want to both take the time to get going with open source way down, which is the just download and go. As well as we want to then be able to provide support and optimizations and improvements that are really significant on top of what's going on in the space.

[0:18:03.0] SM: In terms of DDL, I'm curious if you can walk through the next level of detail and why – what are some of the architectural elements that you feel given an advantage relative to other things that one could do in general?

[0:18:19.2] HH: Yeah, absolutely. There is two things. There is one that the advantages of the hardware which are that the GPUs are double NVLink connected to one another, and also double NVLink connected back to the host processor.

Those features provide performance if you use them with an appropriate communication topology, which we do do with DDL. In addition, then DDL at the overall system level when you're talking about connecting together a bunch of different learners, a bunch of different system nodes, uses a multi-ring topology. Not a single ring, but a multi-ring topology. What this results in is our ability to use all of the links to the greatest advantage possible.

If you do a naïve mapping onto a system, or if you have a system with PCIE interconnect, there are going to be bottlenecks. What we do is maneuver around the bottlenecks and we use all the different bandwidths, so the bandwidth between the GPUs, the bandwidth on a node, the bandwidth getting out to the network, etc., we use those to their best possible efficiency.

Ideally you want to use all of the hardware and all the interconnect that you have and software doesn't naturally do that. We have taken all that work away from the developer and said, "We're going to max out the system, and if you buy the hardware it's going to work really well, because of the software and the way the software is using all the system bandwidths."

[0:19:47.2] SM: When you talk about these rings, where in the system architecture do they exist? Are these at the interconnect level, or are they in-memory, or are they someplace else?

[0:19:57.6] HH: These are all within a different interconnects of the system.

[0:20:02.3] SM: Okay. Part of the double NVLink that you mentioned?

[0:20:08.2] HH: Yup. A part of the double NVLink, a part of the connections out in the network, part of the nodes being connected to each other with network connections, that kind of layer. There is another thing we could talk about. You brought up memory. There is another thing that we could talk about, which is the large model support that we do, which is also a feature in Power AI. That is a situation where you really want to try to use the different pieces of memory in the system.

If you think of it philosophically with distributed deep learning, we're using all of the links in the system, all the network bandwidth, all the bandwidth available in a given system. We're large memory support, we're providing an out of core capability, meaning a capability of accessing other resources, other memory in the system greater than just what the GPU has all by itself.

The way to think about that is if I have a system and the GPU has a small amount of memory today about 16 gigabytes, but the host processor has a lot more memory, a 128 gigabytes or 256 or up to a terabyte these days, right? We provide function that enables people to explore bigger model sizes, bigger data sizes, etc., by accessing and using actively all of the memory and not just the GPU's memory in the system.

Generally, I would say our philosophy is to use all the resources available in the system and let the AI developer explore things as bigly – as bigly – as largely and as large as they like to go, as big as they like to go. That includes dimensions of both memory and the number of systems that you have to tie together.

[0:21:50.0] SM: We've talked about the framework transparency nature of this. Does that mean that all of the thinking that has to go into taking advantage of these elements happens at the DDL and/or framework layer and there is no aspects of the problem that has to be thought through by the developer, or are there still things that the developer has to be aware of in order to be able to take advantage or what you've done here?

[0:22:27.6] HH: Yeah. It's such an important point in productivity, because you can run as fast as possible. But if you're going to sit there for multiple days scratching your head, trying to figure out how to run fast it doesn't do anybody any good, right?

One of the decisions that we made when we did the integration for example into Tensorflow was to leverage the slim library, which is a library that creates the capability to run a particular neural network. In that case, we have hidden the DDL capabilities completely from the developer and they just have to use slim.

We're trying to hide under in general where we can, abstractions that have been created at higher levels, so that people don't have to spend days and weeks trying to figure out how to use this technology. I think we're very focused on developer productivity. It's part of the reason why we went from research endeavor to very quick engagement with our development team and wanted to get it out there in the hands of people very quickly.

Why we're focused on speed, because speed – this is one of the very few areas of competing today where people sit around waiting for days to develop the capability. It's really crazy. We want to go from days down to hours and we want to get people there as quickly as possible to download and go, and then also through use of some of these higher level abstractions like slim.

[0:23:48.6] SM: Okay. One of the critiques, I guess of some deep learning work is that you had described as overfitting on image net and you referenced image net in your results as well.

What gives you confidence that these results will be generally applicable beyond a specific data set and problem?

[0:24:11.5] HH: Yeah. It's a great thing to talk about, because it gets to why we are so passionate about the key thing here being that we were showing the capabilities of what you can do with the framework, only that that's the only thing that it does, right?

As you start to look at other classes of neural networks, the kind of computation to communication balance changes. Some classes in neural networks are known not to scale very far today, not known – they won't run it 256 GPUs, they run it maybe a smaller number of GPUs.

What we see though is that if you're using the type of really close to optimal communication methods that we have that no matter what the inherent capability that neural network to scale is going to be, we're going to guarantee that you get – out to as much scale as possible.

If you're using a sub-optimal communication method, maybe you can only get to 8 or 10 GPUs for other classes in neural networks. We're going to push that number up by getting the communication link and see to be as absolutely short as possible.

In our view, this is really about whatever the current state is of a neural network and the scientific understanding of it and the ability to scale, we're going to push that out and get that training done faster.

[0:25:32.9] SM: Awesome. Awesome. What are the things that you're working on in your group?

[0:25:37.9] HH: I mentioned the large model support, which is something that again, like I said comes from the same perspective of wanting to use the available system resource and not just off letting work to a GPU and being constrained by what that provides. So we want to go multi-GPU, we want to go multi-system and we also want to be able to leverage all of the capability of a system, the full system as memory capacity.

We are in general along those lines also looking at acceleration in the machine learning space. We have some work that we've published and shown some really great results on algorithms that are a bit more chatty between the GPU and the CPU and how they leverage that NVLink capability, that fat bandwidth pipe of communication between the CPU and the GPU. How they get in and out of memory and use the larger capability, the CPU memory. Those are a couple of

different examples. But we're really trying to use the system to the max of its capability and tackle problems faster and enable people to explore problems that are larger.

[0:26:44.9] SM: What's an example, or some examples of particularly chatty machine learning task or libraries?

[0:26:52.1] HH: Yeah. In general, things like nearest neighbor computations, word2vec technology or glove, things that are used in semantic analysis, those are a couple of different types of none, very deep neural network things.

[0:27:09.5] SM: The chattiness relates to in those network architectures the way that state needs to be shuffled around?

[0:27:18.4] HH: Yeah. It relates to how state needs to be shuffled around, but it also relates to how quickly the compute happens and how quickly the GPU needs to be supplied with more data, for example. That's another big component in it. The faster the compute on the GPU happens, the more likely it is that you need to be feeding it data more quickly, and getting that data fed to it more quickly can help it just be brick-walled in its execution time and not sitting around waiting for the next data to get to it.

[0:27:53.5] SM: Short of actually profiling the execution of your training jobs, how do you develop an intuition for what's going to be chatty, like it doesn't sound like it's just the depth of the network and the deeper it is, the harder it is, or is it? Or are there other – is it with, is it something else? Is it memory features? Are there things that you can look for to get a sense for the chattiness of your network architecture?

[0:28:29.4] HH: We like to talk about the algebra of how deep learning happens, or the pipeline stages. We look at how much data is needed in order to start a computation on the GPU. How long does it take to move that data from the storage and then into the GPU? Then how long does the GPU take to communicate – sorry, to compute? How long does the GPU take to compute? Then how long does it take to communicate results, either back to the CPU or back to other GPUs?

Those are the canonical pieces. We look at how much compute is there and how long will that take, and then we look at how much data has to be moved. We look at those as clear pipeline stages or phases of the work you're trying to get done.

[0:29:17.3]SM: It sounds like those are again kind of empirical observations of a training task, as opposed to being able to look at a picture of a network and tell by some characteristic of the network that, "Oh, this is probably going to be chatty and this kind of technology will apply well."

[0:29:38.9] HH: Yeah. It's a good question. I mean, I think that we do have the ability to estimate pretty well based on the characteristics in the neural network in static sense. It's definitely a mix of observation in the empirical characteristics. But in general, from looking at the neural network, you should be able to figure out the volume of data for example that needs to be moved to do wait updates; you know how many nodes are in it and all that other kind of stuff.

There is some – you have to figure out what the many batch sizes are that are appropriate for that. It's probably I guess the – it's probably a mix of some things that can be determined, but then other things that you have to know what the training characteristics of that are, and some of that is still somewhat experimental today.

[0:30:25.1] SM: Awesome. When you think about this work and all of the – you're working on it, there are other folks working on it, like what do you see as the impact overall of it, and what's the timeframe? Do you have a crystal ball vision around how this impacts the way folks develop deep learning models?

[0:30:54.5] HH: Yeah. It's interesting, because I think that there has been a lot of concern as you mentioned that folks are overfitting image net. The question really is the rate and pace that deep learning will take off on other types of data – I mean, it's been proven as highly successful for image and speech, but we're seeing so many other use cases happen.

We're seeing use cases across risk and fraud and predictions and forecasting in lots of different areas that to some extent, the same thing is happening where there are classical machine learning techniques that are being subsumed by the capabilities and the ease of use of deep learning.

The thing that excites me in this capability conversation around DDL is the thought that with this type of speed-up that we will see those fields and those use cases develop much more quickly, because with a more productive system and a more productive hardware and software solution, people will be able to discover and explore their data and define the right models for those data sets more quickly.

If you are able to get through more of your data more quickly and you are able to turn through more models and get to higher accuracy on these new types of outcomes that you're looking for, around like I said risk and fraud and predictions and forecasting and those things, then those fields will develop and mature and those use cases will develop, mature that much more quickly, right?

This has an inflection point in the rate and pace of enterprises ultimately to apply deep learning and increase their confidence in it and increase the accuracy and things that aren't just images and speech, is ultimately what we're really excited about.

Again, that goes back to why we wanted to get it out into the hands of customers, because as a productivity enabler, we hope that it will change that rate and pace of adoption of these techniques and a productivity of data scientist teams working on their own data sets.

[0:33:02.3] SM: Yeah. I think that's a really important point and one that I hear a lot in terms of deep learning becoming deeply, pun intended I guess. Associated with image types of data, but they're being these much broader applications and implications. Are there any particular use cases that you think of as the next big killer app for deep learning?

[0:33:28.2] HH: I think it's hard to necessarily forecast the next big killer app. I mean, I would say that we're seeing more than you might imagine use cases across enterprise environments of even speech and image, because people want to be able to interact more; interact with us in a more natural way and such.

I would say that two things, one the speech and image, I think is taking off in other industries that have other data types but that want to use more natural interaction capabilities and such. That's one thing that is great to see is the a little bit more creative use in other context, especially as it relates to interacting with people beyond just talking to your mobile phone, example. That's one thing.

Then I think, there is a lot of discovery and exploration right now. I think it's really hard to make the call as to what the next killer app and what the next data set is, but I think we are seeing a good degree of productivity on a lot of different types of data.

There are very interesting things happening as well in automation techniques. We have a tool set that helps to label data and other things like that. As more folks adopt those techniques, they'll be able to get access to use of deep learning on more data sets. As we use automation and tool sets like we have and AI vision to enable more quick labeling of data, then that really should also help accelerate the rate at which people enter into these new spaces.

[0:35:01.2] SM: The tools that you mentioned that helps with data annotation, is that open source?

[0:35:06.5] HH: It's not open source, but it's also available through the Power AI frameworks and folks can try it on the Nimbix Cloud, they have the Power AI stuff available there. Yeah, that's available for people to try.

[0:35:20.1] SM: Yeah. It seems like that is also – that clearly comes up constantly in these kind of conversations. When I think about the full life cycle of these AI deep learning development projects, there is certainly a lot of emphasis on the deep learning framework and the model development and training, but there is a much broader set of activities that has to take place and very – we're just starting to see projects – open source projects come online to provide broader platform for handling all that.

The data annotation is one big area, model life cycle and versioning and production, performance management and tracking is another whole set of areas. Any thoughts on how all this stuff evolves?

[0:36:19.8] HH: Yeah. I think that the – this is about maturity of the field of deep learning. It's about people moving past just fully labeled, public data sets that they get online. I think the overall topic of management of what models are being done, when, where and how, overall life cycle, integration with various data sources, all those kind of things are what we see as being necessary from enterprise environment perspective, because people would be wanting to improve the AI specific to their use cases and specific to their data.

We have a good number of IBM tools that's out there today; again, folks can try them. The data science experience is a really nice Jupiter notebook spaced environment that has a lot of the features we just discussed. We also have other tool sets that are available that help with data integration from various sources through our spectrum family, for example.

Then within Power AI, we have that integrated as well with the data science experience, the DSX that I just mentioned and we have additional capabilities that include the clicker data scientist-type capabilities, including your own data labeling and choosing your models and starting with some default models, other things like that. We see these things that you mentioned as absolutely essential to do a more mature deep learning, which is going beyond just the fully labeled data sets that we have in a public domain.

[0:37:59.0] SM: Right. Certainly, enterprises need to get there in terms of maturity. Perhaps to wrap things up, any words for enterprises that are earlier on in the cycle and just getting started?

[0:38:15.7] HH: Yeah. I think that one of the things I've always like to try to clarify in these discussions is that our intention in showing distributed deep learning capabilities at 256 GPUs was to show the capability of what you can get if you do hardware and software optimization to show the flexibility across of what we have created etc., and put in in people's hands.

Everything that we did was done with that enterprise consumer in mind, knowing that many may also be early in their journey. The DDL environment supports any number of notes, any number of systems that someone would like to use it on. It supports different types of networks, so you don't have to use what we published our results on.

What we really want to do is to start with where people are at, start with the data volume that they have, start with the neural network capabilities that they have and provide that ability to grow over time as they hit that stride of productivity of really figuring out how they want to do their deep learning and they want to apply it to the rest of their data sets as they discover new areas within their enterprise that they want to apply deep learning and replace classical techniques.

Perhaps, they can grow out, they can scale out their deep learning environment, add more systems to it and still get that productivity. I think that's always important to point out, because

it's not just the taking the really long days, long jobs down to hours for use of a huge system. It's that no matter what the system size is, no matter what the data set size is, we want to meet people where they're at and provide this flexibility and elasticity and capability.

[0:39:56.9] SM: Great. Great. Well Hillery, thank you so much. I really appreciate you taking the time to chat with us. I learned a ton about what you're doing with DDL and I am looking forward to following the effort.

[0:40:08.3] HH: It was a pleasure talking to you. Thanks so much for the opportunity.

[END OF INTERVIEW]

[0:40:15.1] SC: All right everyone, that's our show for today. Thanks so much for listening and for your continued feedback and support. For more information on Hillery or any of the topics covered in this episode, head on over to twimlai.com/talk/77. To follow along with this AI Summit series, visit twimlai.com/aisummit.

Of course, you're encouraged to send along your feedback or questions to us by leaving a comment right on the show notes page or via Twitter to @twimlai or @samcharrington.

Thanks once again to IBM Power for their support of this series. For more about the IBM Power Systems Platform for Enterprise AI, visit twimlai.com/ibmpower.

Thanks once again for listening and catch you next time.

[END]