Tony [00:00:04] Welcome to Code Together, a podcast for developers by developers, where we discuss technology and trends in industry.

Tony [00:00:11] I'm your host Tony Mongkolsmai.

Tony [00:00:18] AI solutions are in the forefront of technological innovation. Beyond cool technology demos like ChatGPT and Stable Diffusion, companies are starting to integrate AI into actual products like Microsoft's recently announced Microsoft 365 Copilot, which enables AI in their office suite, to generative AI image creation in professional content creation tools. Transformer models are the powerful neural networks that have become the standard for delivering advanced performance behind these innovations. But there is a challenge, training these deep learning models at scale requires a large amount of computing power. This can make the process time consuming, complex and costly. This is actually also true of inference. It can just be really expensive to do AI in production.

Tony [00:00:56] Today we will talk about all kinds of issues around accessible production level AI solutions. To that end, we are joined by Julien Simon and Ke Ding. Julien is currently Chief evangelist at Hugging Face. He previously spent six years at Amazon Web Services, where he was the Global Technical Evangelist for AI and machine learning. Prior to joining AWS, Julien served for ten years as CTO and VP of Engineering at several large scale startups. Welcome to the podcast, Julien.

Julien [00:01:24] Thank you. Thank you for having me. It's a pleasure to be here.

Tony [00:01:27] Ke Ding is a Principal Engineer at Intel focused on applied machine learning solutions. Welcome to the podcast.

Ke [00:01:33] Thank you, Tony.

Tony [00:01:34] So when I think about Hugging Face, I think mostly about transformers and large language models and more recently, things like Stable Diffusion and really about trying to make AI accessible to a large variety of people, not just in the cloud who have a lot of hardware, but people that have some, you know, smaller hardware that they can run on their local machines. Is that how we should be thinking about Hugging Face, Julien? What is the goal of Hugging Face as a company?

Julien [00:02:01] Well, the goal of Hugging Face is to democratize machine learning. And what we mean by that is allow every developer or every machine learning engineer out there to work with state of the art models in the simplest possible way. You know, new models come out every day. At the moment we have over 160,000 models on the Hugging Face Hub. And some of them are, you know, crazy, complicated, crazy advanced. But it's only a couple of lines of code to download them and start predicting with them. And fine tuning them is reasonably simple, thanks to our open source library. So we try to abstract away a lot of the complexity and so that even if you're a junior developer, you can go and grab the latest greatest model for, let's say, you know, text to image generation and put it to work in your apps without being an expert.

Tony [00:03:03] Awesome. And just it's really funny because we actually we're trying to schedule this podcast for a couple of weeks and we were going to talk about one topic. And then yesterday you guys dropped a nice bombshell, which was the BloomZ model that starts with 176 billion parameters, and you actually ran it on a Gaudi 2 and got some really

cool performance numbers. So let's talk about that first. Jillian, talk to us about how you guys did this and what were the exciting results you got.

Julien [00:03:33] So Bloom is an open source alternative to GPT three that came out of the Big Science project, which was a community project with a Hugging Face involvement. And we built different versions of the model and so the BloomZ with a Z versions are have been fine tuned for a particular tasks. And collaborating with with Intel and Habana labs, we posted this cool blog post yesterday, go and read it, showing that comparing inference for different sizes of the BloomZ model. So we benchmarked at 176 billion parameter version than the 7 billion parameter version for for inference on Gaudi 2 and we compare that to latency that you would obtain on the on an Nvidia A100 GPU, which is what a lot of people I guess are using these days. And we showed that Gaudi 2 is 1.2x faster for the 176 billion parameter model and 3x faster for the 7 billion parameter model, which is probably the one you want, it makes more sense to me. So this is really amazing. And, the hardware optimization work we're doing with Intel on CPU and Gaudi and Gaudi 2, you know, we've been working together for a while, so it's a long list of innovations we've built is really, really impressive. And you know, every time I run those tests, you know, I honestly I have a hard time believing the numbers. I run them again, because those are large models and they they go very, very fast. You know, working with Ke, we also had a post on Stable Diffusion inference on the latest generation of Xeon CPUs. And we're below 5 seconds for image generation. So a lot of people think, oh you know it's GPU only for training, it's CPU only for insurance and you know, fair enough. I mean, let everybody use what they want to use. But I say, hey, there are really amazing alternatives there from a performance perspective, from a cost performance perspective. So, you know, do a little bit of homework and and hopefully you can you can find something that works even better than what you have today.

Tony [00:06:01] One interesting question that I actually got from Twitter this morning was why did the scaling for 176 billion parameters, why was that only 1.2x A100, whereas the 7 billion parameter model was 3x of the A100 on Gaudi 2. Do we actually have any data to suggest why the scaling was better, the performance better on Gaudi 2 compared to A100 on the smaller model?

Julien [00:06:29] I'm honestly I'm not sure I didn't you know, I didn't I don't know Ke if you know more, but...

Ke [00:06:34] Yeah, so this is an interesting question. So this is actually related to like the ratio between compute and memory and the some of the like the fabric, communication piece, right? So that is the ratio will be different for different sizes of the models and the for 170 billion 176 billion model, so it requires the like advanced techniques even for inference, like tensor parallelization and pipeline parallelization in order to utilize like 8 Gaudi in one box kind of setup, right? So that leads to like as a ratio would be different. So it is like a the motivate us to design the better and general hardware so that they can support multiple models so nicely.

Julien [00:07:23] One thing I can say is when we work with customers very, very quickly, we, we find and we agree on the fact that they don't need such a huge model. They have a particular use case they want to solve. And the domain for the use case is actually, you know, much narrower than a general purpose text generation model has. And so there are lots of benefits in actually using smaller models. And what I mean, smaller I mean, you know, maybe 20 billion or 10 billion or a 6 billion, they will in many cases, we find if you fine tune them on domain specific data, they will outperform the larger models. So, you

know, right sizing models is important and bigger is not necessarily better, especially for enterprise use cases.

Ke [00:08:12] So I fully agree. I fully agree on that. So picking the right model for the right problem to solve. So that's actually very important. And because at the end of the day, so cost efficiency and they are they are one of the major considerations if you want to really deploy a model in a production environment. And so the right thing here is like Intel provides the right hardware, actually a broader portfolio of different hardware for different use cases and then together with Hugging Faces, right? So software on top of that and to provide both rise of productivity with performance and it was very easy for use with the customer.

Julien [00:08:54] Yeah exactly. I mean none of us I mean certainly not me, you know our hardware experts. And so, you know, I'm excited as an engineer. I'm always excited to read about the, you know, the spec sheets for the latest Xeons, the latest Gaudis. These are amazing chips. But I wouldn't have the first idea on, you know, where to get started to optimize my code. So and I think it's the same for most most developers and engineers out there. So it's you know we build we leverage the hardware acceleration and some of the software tools that Intel has. And we you know, we packaged them into our hardware acceleration library is called Optimum. And there's a version for Intel CPUs and there's a version for Habana, etc.. And again, one or two lines of code, you just accelerate your models just like that. And and I think that's the way it should be, right? The experts let the experts figure it out, let the model experts figure out how to build great models, let the hardware experts figure out how to build great chips and and let's let the developers enjoy all that good stuff in two lines of code. You know, that's my mission.

Tony [00:10:07] Yeah. And that brings us kind of to the, where you're talking about the the webinar that I mentioned earlier where when we look at this portfolio that we have of hardware, we have CPU, we have Gaudi, we have GPU, right? We have A100 and everybody has has cool hardware. But the one thing that we we look at is kind of the cost benefit, right? What what behavior, what performance can I get, what problems can I solve? And that's where, you know, I want to mention the webinar that you guys had, which was how do I take a relatively big model and optimize it for CPU? So I don't want to go over the entire webinar. I mean, we'll put a link for the podcast, but maybe you can talk about kind of the tools that Intel was kind of promoting there to kind of make this, this dream of the one or two line, you know, transformation of your code into optimized performance on CPU. Can you talk about that a little bit?

Ke [00:10:59] Yeah, sure, Sure. So Intel and the Hugging Face has been working like for almost three years. And so we work together and they try to like optimize to the transformer and Hugging Face community so with the latest Intel's hardware features and so the main I would say like a use cases right in this space, it's like how to make sure so I can take that a pretrained model to like there was a fine tuning for my particular task. So this is one of the main thing and another thing is how to make the inference to run much faster on that particular platform like so Intel Xeon platform. So those are the two main focus for our collaboration. And as for the for the fine tuning piece, so we work together and we support like a distributed fine tuning features directly into the transformer libraries and take advantage of like the latest Intel Sapphire Rapids features like AMX and low precision formats like bf16. And just a very simple command line option. Like dash dash IPEX and the bf16 and you get a job done. And we demonstrate like almost a linear scaling with, with multiple nodes of Sapphire Rapids machines, right? and on the inference side so we also like as as Julien mentioned, so Optimum library with Optimum Intel and

Optimum Habana and we integrate our Intel tools like Intel Neural Compressor, OpenVINO and IPEX right and make sure so like model quantization, tuning, sparsity and attestations are much easier for the end use and the users right. They just need to set very simple options and then code the very familiar train API and they get the job done by inference and runs best on Intel platform.

Tony [00:12:59] Yeah and then we talk about kind of how people are actually starting to deploy AI into real world solutions that a lot of people will use. When I mentioned things like Microsoft Copilot for Office 365 and Adobe Firefly, these are kind of the practical use cases that people have around these demos that we're seeing for large language models or generative AI image generation. But those are kind of from a big corporate perspective. Are there other practical applications that other companies are starting to use? Julien, that perhaps we aren't thinking about, that aren't coming from these big companies?

Julien [00:13:34] Sure. You know, the deep learning is is great at learning stuff hidden in unstructured data and predicting new data and transformers, you know, even more so. So a lot of I would say everybody has use cases for transformers, particularly in natural language processing, right? Tasks like extracting information from documents, translating documents, summarizing documents. A lot of the customers we work with start from there. You know, they have tons of information. Information overload definitely is a problem for everyone. And and they need to extract the right information from those 50 page documents or from those earning calls or from those interviews, etc., or social media and find the signals that like, let them take the right decisions. Right. So it's really okay. There's too much data for everybody to actually read from A to Z or parse from A to Z so can we just extract the right signals from that data and then, you know, use that to act. So it's the number one thing. Document, document processing. And of course, you know, all the test types like working with audio and speech, you know, speech to texts and then obviously using the text for additional NLP processing or computer vision, etc.. The transformer architecture as really generalized to, I would say almost every machine learning use case out there for unstructured data. We have customers working on, you know, protein sequence prediction and drug discovery. You know, that's really exotic, but it works. And so, yeah, it's you know, you can start small like, you know, raise your hand if you don't have a document processing issue in your company. Right. Okay. I see no hands. So there you go. Start with the small models. Like I said, it's a couple of lines of code to get started and you can start small and experiment and then move on to scale and deploy on prem or in the cloud. Find the right hardware platform for that, find the right, you know, cost performance ratio. For that, we try to give developers as many options as we can.

Ke [00:16:09] Transformer is kind of a foundational model, right? So it can support a different modality, not only the text about the others as well as like a multi modality, but you can have multiple input to that. And the other thing is the task it can support like a it's also a varied set, variable set and starting from simpler language task all the way to like a very complicated language task or other task. So that's why we need a different set of models, different size of models to tackle like those problems.

Tony [00:16:48] And it's interesting when we think about how how we focus on these kind of larger models, we talk about the 100 and hundred plus billion models that everybody's using. But then in the in reality, to actually use them in production is not very cost effective. Right. It's it's just really not you can't be running inference on 8 A100s every time you need to to do any type of inference. We really need to make it deliverable to people. Right? You need to be able to run it on your end user system potentially on your phone, things like that. So it's pretty cool the way when I think of Hugging Face, I think of it probably

incorrectly, as a repository for...as a technical person, I go there and I can see kind of what models are available and what technology I can use. That is obviously not a business model. That's just a cool tech demo, but Hugging Face has a business. So Julien, what type of support do you give to developers who actually want to take the cool models that you guys make available and provide them in a consumable customer facing way?

Julien [00:17:54] Sure. So the core of the company is open source and that's that's important to all of us. But obviously, you know, we try to keep the company going. We we need to make money, right? Simple as that. So we have different commercial services and we also do consulting. So basically the vision we have is models should be free to use to experiment with because you don't want to put every you know, when you have an idea, you want to experiment as fast as you can. You want to validate the idea, you want to get feedback on your idea and any commercial deal, any procurements, anything like that that stands in the way just slows you down. And you've been kind to say early on, I've been a CTO for ten years, so I know, I know, I know the frustration of having to go through, you know, legal and procurement and and complex sales processes to even get started. So here that's not the case. Go to the hub, grab a model, start playing in minutes. But of course, at some point you want to move to production. Right? And so concerns like how can I deploy my models in a simple and scalable and secure way. If I decide to fine tune my models? You know, how do I do that in the simplest possible way without having to build all my machine learning infrastructure and machine learning platform myself? How do I get to the performance level that I want? Right? I have every application as a latency budget, right? So if you're doing batch prediction, it's okay, you don't need to worry so much. But if you're doing product search or semantic search or conversational apps, you have a very tight latency budget. And as mentioned before, those are big models. And this is where, you know, hardware acceleration comes into play. So all this production concerns, I don't want to say MLOps, you know, let's stay out of buzzwords but generally all the challenges that stand between your cool PoC and scalable production is where we have commercial services and commercial support and, you know, custom model training and engineering. I mean, we have a crazy bunch of experts, whatever use case, we can probably, whatever the model, we can probably find the right person for you to to talk to and accelerate your path to production. I mean, I've seen customers get going from literally, you know, PoC to two fulls full prod in weeks and you know, like two or three weeks because we have the expertise, we put you on the right track and then we have we save you from all those silly mistakes. So yeah, there are you know that's that's beneficial and and we work with a lot of companies on from startups to really large ones.

Tony [00:20:54] So I'm going to pivot a little bit. So you mentioned Julien how Hugging Face feels that models should be open and people should be able to try them and understand them as much as possible. The interesting thing, and there's a lot of ethical issues and questions around AI. The interesting thing about the latest GPT-4 is that it's closed source and a lot of the inputs that went into it are not well known versus previously OpenAI had kind of a more open model where they, you know, everything was known about GPT-3. What went into it and people could understand why they potentially were getting answers out of GPT-3 in a certain way. How does Hugging Face, I guess maybe it's not Hugging Face, how do you feel around what OpenAI is doing? Is that something that you think makes sense in the long run? Obviously, it's a business decision for them, and I would probably not want to be poking at other people's business too much. But how do you feel about kind of where that's going? Do you feel like AI models are going that direction, or do you think that we can continue to keep them open?

Julien [00:22:01] Well, businesses are businesses and, you know, you have to make money. So, I mean, OpenAI is a fantastic team of people.

Tony [00:22:13] Absolutely.

Julien [00:22:13] So, you know, we're really all impressed by, you know, what they're doing. And if they think the way they're doing it right now is the best for them, then, you know, who am I to say otherwise? You know, I'm not going to go and preach. It's not who I am. Now, from from a customer perspective, I try to see things from the customer angle. I think it's important to have diversity when it comes to models. I think it would be dangerous for innovation and ethics and pretty much everything, if you know, a few years down the line, we have, you know, 3 to 5 humungous models that overpower everything. You know, I think innovation comes from open competition in every single business in the world. So I think closing things down, not disclosing the training data or the high level training process or the high level architecture or, you know, what are the moving parts is a concern. And, you know, we've seen another example of that in a few months ago. You know, when text to image models became wildly popular and broke the Internet and we saw some some players closing down their models again for their own reasons. I'm sure, you know, it makes sense to them. I'm not I'm not judging. But what was really interesting is the reaction of the community. And within a few weeks, the community joined forces and released the the diffusers library that that we host, that we that we steward. And, you know, like stable diffusion models are now open. And honestly, I think that's what people use instead of the closed models. So to me, you know, there's hope. Again, companies are companies. They take their own decisions, let you know, let may the best company when may the best model win. But fighting the community. Ignoring the community. Seems like a dangerous place to be for me. Long term, we've seen this. You know, I'm old enough to remember Windows versus Linux. I'm not going to repeat the immortal words of the Microsoft leaders of the time. We've seen closed languages versus open languages, closed databases versus open source databases, literally everything. And and look where we are now. So OpenAI is doing their best, the best they can. They have some they have you know, they have some lead for sure. Long term, let let's see where this lands. But for sure, when it comes to Hugging Face, we know where we stand.

Ke [00:25:08] Yeah. So maybe I want to add the one thing people now talk about, like that so-called iPhone moment. The narrative is an iPhone moment because it is kind of revolutionary. So it's just increasing a large lots of attention and interest. But I would also say like, what is that equivalent enjoyment? And so that is the open ecosystem, open source, open platform, right? Intel being a strong believer on this side. I think Hugging Face also shares the same like on this perspective. And so like I would say, so Intel and Hugging Face, our collaboration we should demonstrate, is the greatest moment, Android moment for generative AI.

Tony [00:25:57] That is a great soundbite. I can't wait to figure out how to put that in the title Ke, for Generative AI. That's a really good analogy though, because that's really true. There's the what's the technology that pushes everything forward and then what's the technology that makes it actually accessible to everyone? Because not everybody wants to pay 1300 dollars for the cutting edge iPhone, but they want the same the same benefits without having that that cost around it. For the challenges of making sure that AI is ethical and usable in a lot of spaces where there are concerns around how the AI is coming to the conclusion that it is. There's a lot of efforts. Intel has some explainable AI efforts. I know...

Ke [00:26:50] Yeah, sure, sure. So we have those. Yeah, of course we have some effort on this part because this is very important, and especially with those large language models. Well, how can we explain it on the other side? In some particular cases, the regulations, there's a requirement so people need to explain how and why the model behaves like that. So for that part, Intel, so we are developing a toolkit called XAI, explainable AI and support, like those algorithms like Sharp and others and a just a for this purpose. And we have some features like a model card generation and then we have explainers and we are working together with with open source with other partners and to integrate it into our Intel AI software offerings.

Tony [00:27:46] Cool. And then on the Hugging Face side, Julien, can you talk about how Hugging Face is making sure that the solutions that they're providing and promoting kind of meet these ethical rules around AI, really most likely to the benefit of their customers.

Julien [00:28:02] Sure. So at this point, we think it's still very much an awareness and transparency. So we introduced the concept of model card on the model hub. Where we, for all the models that we manage and obviously we encourage everybody else to do the same, we provide as much information as we can on how the model was trained, what data it was trained on, what are the potential use cases or non use cases for the model, What are known limitations and bias issues probably caused by the training sets, etc., etc.. So we try to you know, it's powerful technology, it's complex technology. We're very far still from being able to have, you know, standard tools and standard test suites that we could just run on those models to to get a full picture of, you know, what's good and what's bad about them. So we think the starting point for that is awareness, explaining, explaining again and again and again that some of those models, you know, are unpredictable. I mean, generative AI is designed to output unpredictable data, right? It's what it's for building, generate an image or generate an answer. You know, you don't know what you're going to get. So hopefully you're going to get the correct answer to your prompt, but more importantly, you're going to get something that's acceptable from a business and an ethical and maybe even legal point of view. Right. And and so awareness and transparency and analyzing data sets and understanding what's what's wrong about them is still very much needed, I think. But we don't we still don't fully understand what the issues are. We certainly not don't fully understand how to fix them. And there is a lot of research that goes into that. So, you know, I hope we can collectively make some progress on this. But for now, it's still very much, you know, let's let's be safe and and and test those models and figure out, you know, when to put them in production and when not to put them in production. And again, you know, coming back to GPT-4, their 80 page technical report has this paragraph that says for basically for commercial, for competitive reasons, which everybody will understand and for safety reasons, that I personally don't understand, we're not going to disclose anything about the model. And I think it's a bit shocking. If it's unsafe in certain ways, then why release it? You know, why not be transparent on on the on the problems and again, help the community figure out how to fix them.

Tony [00:31:02] Yeah. And I think we also have to be careful about how people tend to use things. I, I was reading somewhere where somebody said that they and again this was Twitter, so who knows? But multiple people had similar experiences where somebody said, I submitted a paper to research conference and the reply was, please go look at these other references to make sure that they don't conflict with what you're already writing and the person's multiple people said, I did that. I couldn't find these papers that were referenced. And it turns out that the assumption is that somebody was using a generative AI to try to find similar references and the generative AI spit out a bunch of references that looked reasonable, but as far as people could tell, don't exist. So so there's all kinds of

things like that, too, where people are trying to use this new technology without having the proper guardrails around it.

Julien [00:31:49] A hallucination is a problem. And, you know, the the numbers that OpenAI shared on on hallucination are, you know, concerning and the you know, and just general generally, you know, like the ability to generate fake news convincing, very convincing fake data and so on. You know, it's the list goes on. So I'm not saying we should you know, we should pull the brakes on this. Quite the contrary. I think it's it's it's important technology, but we need to discuss the issues in the open and decide when, you know, what's the risk benefit ratio, just like drugs. You know, when you design a new drug, you analyze the risk benefit ratio and and we should do the same here for every every piece of new tech. And for some use cases, it's fine. For some it's not. And then we work together to find solutions. But like, you know, putting the, locking the issue away in a box is not is not going to make it disappear. And it's not helping with awareness for sure. So, yeah, that's concerning.

Tony [00:33:01] Yeah. And hallucination in this context, what you mean is it's generating output that is not correct, factually correct...

Julien [00:33:07] Yeah, it's it's syntactically and grammatically correct and very convincing, you know, with a chain of thought, but it's totally wrong. So if you're not an expert, you're going to swallow this, you know, bait, hook and sinker, I think is the expression.

Tony [00:33:28] Hook, line and sinker?

Julien [00:33:29] Hook, line and sinker. Yep. So that, you know, that's the thing. You know, the credibility of the of the output is important, you know. So we need to we need to find ways to improve that. You know, I don't have the solution, but I know the the solution will come from collective research and collective intelligence. And the more we collaborate and the more we open this can of worms and fix it, the better. And, you know, I think that's the way.

Tony [00:33:58] We always like to ask our guests at the end of each podcast, what are they looking forward to? And in this space, there's a lot of different things to look forward to. I'll go to Ke first from your perspective, either whether from an Intel perspective or just from your perspective. What are you looking forward to the most as we look at where AI is going in the next couple of years?

Ke [00:34:19] Yeah, so like for Intel, so fundamentally we are like a hardware platform. We produce the most advanced hardware with a big portfolio to support all these use cases. So we definitely want to continue to innovate with, with the silicon that's hardware to meet this demanding AI workload and for example as I earlier mentioned compute, and memory and the communications they are the most important aspects when we designe the hardware, right? So then how to balance those things and to make it a flexible and that can support different AI use cases, right? There's no one, one model for one rules for kind of thing, right? So we need to make sure it is balanced and the hardware it has performance, has efficiency and has a good TCO story. So that's the that's the part I think it will continue need to develop how accelerators and for AI in particular we need to add a features for example like a how is sparsity support right and to continue to innovate on the on the low precisions. Right that's a very very low level it's maybe it's not it's invisible to the end users, so I that's the the thing I think we need to think from hardware aspect and then

hardware is not enough, right? So it's the software makes things different and eventually the user interface and the we need to a collaborate with like domain leader ecosystem leaders like Hugging Face. And so our principle I would say is like we should follow for that by software defined, hardware accelerate kind of principle and and the really like when we did in hardware I really think about what are the end users and use cases. Another thing is like we will continue to innovate right and our XPU strategy so we have oneAPI and that we really want to make sure they can make the life easier for, for the developers for, for the, for the, for the software world so like a one thing can cover different pieces, abstract away those the differences.

Tony [00:36:30] Yeah and I think the one thing that actually popped into my head that we didn't talk about earlier when we talked about Gaudi and Gaudi 2, you were talking about hardware innovation and balancing the compute and the memory and the communication. One of the great things for working at Habana, because of the Gaudi solution, is actually built for that communication aspect, right? Because it actually allows all of the cards to communicate. One of the cool things that we didn't talk about within the BloomZ experiment was it actually is noted that you're able to do model parallelism on the Gaudi because of that type of architecture. So having that kind of novel architecture actually enables new ways to accelerate our AI compute. So, it was just fun. I really enjoyed that part of working at Habana, I want to make sure I give that, that shout out there because it just gets me excited every time I hear it. But then we go next, let's go to Julien. Julien, what are you looking for in the next couple of years from Hugging Face the AI industry?

Julien [00:37:29] Well, I think, you know, I expect. You know, I think we should continue building models that apply to more and more use cases. You know, there are still lots of problems that are, you know, not solved or can be solved by assembling different models. But doing the task types we have today are still very low level. So, you know, I'm really looking forward to seeing industry level models, you know, models for health care, models for education, models for I don't know anything out there, not just, hey, we have a model for, you know, entity extraction and you can go and fine tune it. And I think it's still you know, there's still a lot of manual work to do to get to to get to end to end solutions. And I think I'm really looking forward to bringing those models to the to the edge, so to speak, you know, to smaller devices. For now, you know, it's still challenging for size and latency reasons, but the ability to deploy those super high accuracy models on smaller devices, on phones, etc., well, it will certainly help help a lot, you know, potentially working completely disconnected from from the cloud, etc.. And that that would be that'd be great. That would enable a lot of a lot of use cases. So, yeah, I think it's you know, everybody wants super accurate models that are small enough to fit everywhere and and that you can customize to your own business problem. So there's still a lot of work to do there.

Ke [00:39:13] So that's AI everywhere, right? So AI everywhere story everybody can use AI,. benefit from AI and the and eventually it's a truly AI everywhere kind of vision.

Julien [00:39:24] Yeah exactly. It's applying applying, you know, it's transformative technology. I think we've only still scratched the surface on what's possible. And and I really, really want to see AI finding its way into, you know, every aspect of our lives. You know, God knows there's a lot of stuff to fix in, in society from, you know, education to health care to traffic to energy consumption. We have a few issues out there. So I'm all for fun and, you know, generating, you know, cat pictures and whatnot. But there's also some very serious work to do. And and I hope we can get there.

Tony [00:40:13] And that concludes our time for today. I'd like to thank our listeners for joining us, and I'd like to thank Julien.

Julien [00:40:19] Sure. My pleasure. Thanks for having me.

Tony [00:40:22] And thank you Ke for taking time out of your day as well.

Ke [00:40:25] Thank you, Tony.