00:00
[Tony]: Welcome to Code Together, a podcast for developers by developers, where we discuss technology and trends in industry.  I'm your host, Tony Mongkolsmai.

Today we are joined by two university students that are working on a chatbot that aims to have English, human like conversations through the use of AI and ML.  Joshua Shiells, a first year university student, is the creator of the transformer based chatbot.  His collaborator is Harvey Johnson a 2nd year student studying an integrated masters in Electronics & Computer engineering at the University of Nottingham.

[Tony]: Nice to meet you Josh
[Joshua]: Nice to meet you to Tony
[Tony]: And also Harvey Johnson
[Harvey]: Hello

[Tony]: hi guys! You guys are actually using oneAPI and SYCL in a real world application. perhaps Josh you can talk a little bit about what application we're talking about.

1:02
[Joshua]: So the application itself is an artificial intelligence  chat bot that leverages transformers in order to obtain generative responses based on a prompt.  Currently, I'm using Tensorflow as the machine learning API to actually allow me to build the transformer itself, but I build it, for some stupid reason, from scratch to allow me to learn more about the inner workings of transformers but hopefully also leverage performance gains.

1:34
[Tony]: how long have you actually been working on this?

[Joshua]: God I think even started just gone October  he's about three years old now and that is surprising yeah that time kind of flew by.  I first started him in what would be the year before final high school in the US, but year twelve in the UK and it's been quite the journey.

2:07
[Tony]: so you've been working on him for three years I mean so back in 2020.  How did you come up with the idea that you wanted to build a chat bot?

[Joshua]: I was just scowling through Tensorflow documentation because I'm like I want to do  machine learning and I've done the generic cats and dogs image detection and yeah it was fun but it's been done to death and you google the issue and someone solved it and you get 99% accuracy without all that much work.  So I decided I want something more challenging.  So as I was looking through Tensorflow documentation and sites like Medium and stuff like that, I realized the natural language processing and seem quite interesting. I decided that I'd take up that and what seemed like a simple idea at the time was a chat bot and I underestimated how difficult one is by a lot.

2:56
[Tony]: let's bring Harvey in here a little bit, so Harvey why don't you talk about how you started to get in involved with the work that Josh is doing.

[Harvey]: So I met Josh in a Discord server, sentdex's Discord, he's a Python programming Youtuber, and I think I offered to train Gavin for him.

[Joshua]: yeah

[Harvey]: And while training Gavin I noticed it took thirty or forty minutes to load up the data set, and that wasn't even the full data set that was only a bit of it and so great issue of this I saw in efficiency and I couldn't help but try and fix it so I basically joined the project I rewrote the loader for the data to basically be written in C++ bindings for Python and we managed to get it down to I think it was something like ten to fifty seconds to load eight to nine gigabytes of data which was pretty much the full data set at the time.

3:58
[Tony]: so you're talking about a 60x improvement, something like that

[Joshua]: Yeah, it's insane

[Harvey]: A big, big improvement.  I can't quite remember it, but we were purely limited by the SSD speed on my end, so yeah it was a massive improvement.

4:16
[Tony]: That's cool and just to be clear you mentioned Gavin and Gavin is the name of the chat bot that you guys are working on.

[Joshua]: yeah

4:24
[Tony]: Are you guys the only two guys working on it?

[Harvey]: yeah

[Joshua]: Yes yeah we are

[Harvey]: I think we attempted to get someone else to contribute code, but it didn't work out. But yeah we're the only two.

[Joshua]: It didn't in the end. Not surprisingly, the Code little bit disorganized on my end.

[Harvey]: Same on my end, I am not professional.

4:47
[Tony]:  That’s okay, you guys are university students. I'm pretty sure that my code was not very professional for even the first couple of years I was a professional.

4:57
[Tony]: Do you guys actually deploy this anywhere, or do you just test it in house?

[Joshua]: Technically it is deployed as a Discord bot at the moment which is public although I don't advertise it much because he is currently still a little bit on the stupid end and that runs in the cloud on another one of my friends who owns some servers that he rents and he offered me some space for me to host 24/7 but other than that he's not properly deployed, no.

5:31
[Tony]: interesting is it open source?

[Joshua]: Yeah it's completely open source all his code is in a Github organization under the name Gavin development.

[Tony]: Okay cool yeah we'll have to link that. I think I actually googled it because I knew we were going to be talking about it, so I think I found it but I'll have to confirm with you it's the right place. And then Harvey you were talking about kind of how you accelerated it and obviously you work as an Intel student ambassador and you use oneAPI so talk a little bit about how you were able to use the different pieces or whatever pieces of oneAPI used in order to accelerate Gavin.

[Harvey]: So we have the tokenizer which takes in the text and converts it into a one dimensional or two dimensional vector.

[Joshua]: one dimensional vector

[Harvey]: It's numbers, yeah, one dimensional vector and he was using, I think it's deprecated, the one you were using now right as the builder. The Tensorflow one.

[Joshua]: yes

[Harvey]: It's bad enough it got deprecated basically, so I figured I have a crack at making one and so I did a implementation of byte pair encoding. It's not a full implementation because it doesn't, I think because byte pair encoding will keep combining until it gets to a certain vocab size whereas mine will only combine one time. Essentially I wrote the algorithm on the CPU and it was fast and then I thought well we could parallelize this because it's basically just two four loops and so I thought, GPU.

And that's when I started investigating how to do on the GPU. So we took one look at CUDA and I was like maybe not. I tried Vulkan because I have a decent amount of experience in Vulkan trying to render engines also wait too much boiler plate code and so then I went around looking and SYCL came up and tried it. And very easy. So we used SYCL to accelerate the creation of our tokenizers the vocab building and it's pretty big acceleration I think it's like my 12900KS is about eleven to twelve times slower than a 3090 and this is with SYCL. So I don't think the codes particularly well optimized like native CUDA would be but it's a very big speed up basically.

8:05
[Tony]: So you're actually running… you wrote it in SYCL but just to be clear for our listeners, you're running it on an NVIDIA 3090

[Harvey]: yep

8:14
[Tony]: Cool! And what compiler, how did you actually get it to compile for that?

[Harvey]: It was probably the easiest thing I've ever done involving Github. It was literally just pull the LLVM git repo, follow the instructions, which was like three things in command line and then twenty minutes later I had myself a compiler which after I had played with it for five minutes to work my way around how to use the command line for it which was actually, I don't know why I did that because it was standard. I basically did a test build of it and then it worked and we tested it and then it was like hurray, I've got it working. It was really easy actually

[Joshua]: yeah

8:58
[Tony]: has that made your life easier having the shorter load time? Is it making your development cycles any better?

[Joshua]: Yes, one hundred percent

[Harvey]: Yeah, cause when we have issues with the training right and to be able to actually encounter those issues we need to be able to start training so the fact that we can actually load the data up in a coupple of seconds now versus thirty or forty minutes it actually exposes us to a lot more bugs and a lot more issues that we're trying to fix but that's good it's progress.

9:32
[Tony] I'm curious what topics does your chat bot cover. So if it's in Discord there's a lot of different types of conversations going in Discord. You said it's not very good. What actually do I get out of talking to Gavin?

[Joshua]: That's where this gets a little bit interesting. When I first started Gavin, the easiest data set or the easiest way for me to collect a bunch of text data without having to break TOS and scrape a bunch of sites was to already use a data set out there.

So I ripped Reddit it because there's pushshift/io. The CDN has all the files they used for the last, back into like 2015 of all comments. So I just downloaded all of them. Took its time. And processed them into a databases which I've since somehow managed to delete but that's a future me problem. And then from there I've used Harvey's, other parts of Harvey's tools, using Intel, converted from that JSON data into custom file formats. Which is how the load times are crazy, crazy fast now.

But because he's trained on Reddit, the issue is one, I've got to filter a lot. So on the Discord bot there is a list of words I just tell it to filter out of the message, because I'm not dealing with Discord knocking on my door for my bot saying things it shouldn't.

Another issue with Reddit is it's probably not the best data set to use purely because it's just a massive website with so many sub-reddits, that there's not a continuous style of typing because it's probably millions of different users is using that. So Gavin himself is highly incoherent after three or four messages trying to follow the same conversation he just gives up, purely due to Reddit.

I'm trying to switch to what GPT-3 uses, which is the pile but that is about a terabyle of data, compressed and currently I don't have the time to be able of write the code to be able to leverage Harvey's tools and actually process it. But that is to come.

11:55
[Tony]: That's interesting. Have you tried other, have you tried Google chat bots? Some of the ones obviously there was the story earlier t is year right where the Google engineer claimed that the chat bot was sentient. Have you tried any other chat bots and how do you feel about your Gavin effort versus what's out there?

[Joshua]: Well I haven't had first-hand experience with the latest state of the art models because they either they either cost to get or you need permission from the researchers themselves and I don't have the time nor money to be able to access that at the moment.

However I have played with GPT-2, which is an outdated model now, and it is much better. I'm purely going to put that down to it because OpenAI has a research budget and I do not, or we do not.

So we are using consumer GPUs, very, very powerful consumer GPUs, but nevertheless they're consumer GPUs and we obviously have university on top of on top of everything else as well to finding the time to be able to knuckle down and fix all the bugs that come up with Gavin is challenging at times

[Harvey]: Yeah that is a real issue actually

13:10
[Tony] Have you played about with other AI models or are there other areas of AI that you guys are interested in or are you focused mostly on Gavin and your university work.

[Harvey]: Hmm I guess I'll go first with this one. I got into AI through Tensorflow tutorials. Then I spent a long time learning how DC GANNs work and learning different generative models. I played around with resolution for a bit I played around with recoloring images and then actually generating faces from scratch which is half the reason why I have the computer set next to me, is to do that.

So I mostly did a lot of feed forward conv-nets. Nothing anywhere near as complex as transformers which I think Gavin is based on and I don't really know how transformers work but luckily I don't have to deal with that, that's Josh's department and expertise.

14:11
[Joshua]: I have played around with other models, natural language processing is 100% my passion, but from time to time I need a break because it makes me want to rip my hair out. So I do, recently I've done a little bit on image recognition. So recently I did some old Kaggle competitions the generic ones for detecting certain diseases and that because it seems pretty interesting to realize there's actually a model, some team of people, out there saving lives as we speak.

In sentdex's Discord, one of the moderators there, I was lucky enough that he presented the master thesis and I happened to be around at the time, and his model is currently being used to detect a very specific issue with the heart but in and around hospitals in the US, it's saving lives while Gavin definitely won't be doing that, I'm hoping that maybe in some way, my skills that I build beyond natural language processing but also natural language processing will also be able to help in similar ways.

15:22
[Tony]: it's actually interesting. I'm doing a podcast, I'm recording it at least this week with a company that's focused on radiology and chest imaging. So I spent some time looking at that this week and there was there's been a couple things. Google has done some work around identifying abnormalities in chest x rays. And there was another paper, maybe it wasn't a paper, but there was a news article yesterday where somebody published that through a chest x-ray, they claim they were able to determine if you would have cardiovascular disease in the next ten years based on how your arteries looked in the chest x ray. They said that their AI algorithm was able to detect essentially if your arteries were hardening having some type of arteriosclerosis and they could identify that up to ten years ahead of time. We'll see how that comes out, I couldn't find any like details or research paper behind it but they did have a link to something, it just didn't have the paper itself so we'll see what happens.

[Joshua]: Interesting, yeah.

[Harvey]: Sounds very impressive and very much far ahead of

[Joshua]: yes yeah, it's childs play in comparison

[Tony]: At the end of the day it's just a different same, same type of workflow. Different models, different level of data you know.

[Joshua]: yeah

[Harvey]: yeah

[Joshua]: Gavin's entire data set, in the compressed form, that Harvey uses, the entire data set is 20 gis, 10 gig files each which is about sixty five million samples. So a lot of samples, well a total of a hundred and twenty, but sixty five million useful samples because you need to pass two samples at the same time to the model.
A lot of samples for consumer but on a scale as big as that, chest xray or GPT-3 tiny.

17:13
[Tony]: Yeah when I was talking to the guys who are doing Aurora, so we did a podcast around that, I mean they were talking about petabytes of data, right? As inputs.

[Harvey]: God I could only dream of having that much available.

[Joshua]: I feel sorry for

[Tony]: or you could suffer trying to ingest that much data into your processing engine.

[Harvey]: yeah

[Joshua]: I feel like what we've run into issues with, Harvey may have to correct me here, but I think we've run into issues with PCIE bandwith in the past. I imagine petabytes would start properly throttling that.

[Harvey]: I mean our GPU accelerated algorithm actually has really lack luster core utilization because it is not bottlenecked by the core is not even bottlenecked by the memory. It's bottlenecked by the PCIE bus and unfortunately there's nothing I can do to fix that.

[Joshua]: yeah

[Harvey]: Other than try really hard to make it more efficient but I don't think there's actually a way to do that unfortunately.

18:08
[Tony]: Now you just need to find a third person who knows how to do it.

[Harvey]: yeah, hopefully

[Joshua]: I don't suppose you know Tony. Want to join the Gavin team?

[Tony]: I'm not sure I'm the right expert for you guys there.

18:24
[Tony]: Other interesting things so Harvey I know that you recently got an Arc card.

[Harvey]: yep

[Tony]: I'm curious

[Harvey]: Do you want to see it?

[Tony]: Well I would love to see it.  I mean, I obviously, I have one too.  Nice!

[Harvey]: yeah

[Tony]: so you didn't get a, you got a 750, so you didn't get the RGB one?

[Harvey: I have the A770, it's sat in a NUC, just behind the laptop I'm on.

[Tony]: you've got two. I'm curious that you try running Gavin through them using your SYCL model, how'd that work?

[Harvey]: Yes, I don't have numbers to compare to over GPUs, but I confirm the data set deploys and runs no bugs because I pretested it because our plan was to try and deploy Gavin to the NUC 11, which is the A770 because were having some issues with the 3090s, something, something it doesn't like to work after an hour or two.

[Joshua]: yeah

[Harvey]: and Intel, this is just a neat little thing that I noticed when I first got my Arc card. You don't need a display plugged in to get display out which is amazing for headless servers, but NVIDIA last time I checked and tried, you need to have a dummy plug. So my hunch was that the reason why we couldn't train Gavin for more than ten hours at a time was NVIDIA just being like nope you're doing data center work and killing the process.

It was my hunch I have no proof at all to be clear but, you know, they do that thing with virtual displays or at least they used to, so I wanted to get a build going ASAP on the NUC so I've gone ahead and tested it all and code fully deploys and it performs pretty well as far as I'm aware.

[Tony]: Okay, cool. I mean that's good news. That's what we're trying to do.

20:20
[Joshua]: My only issue is the Tensorflow itself because the NUC is on windows at the moment because of our personal issues with not being able to get SSH set up because reasons with each of our networks. Being university students they don't really like us fiddling with the network settings.  So Tensorflow doesn't appear to support Arc cards on windows yet which, an issue.

While obviously it does train on the CPU, that isn't exactly fair to the CPU to do that to. So I haven't pushed to do that but when I eventually find the time to be to develop the process and code for the pile data set I will be using the NUC and be leveraging both the CPU and GPU in order to actually ingest that data, tokenize it and save it to the format and everything that needs to be done before it gets on to a model.

And I see that can be done on whatever GPU it needs to be done as soon as the data is processed. It is pretty much plug and play.

21:19

[Tony]: Yeah, no that makes sense. So I'd like to talk to Harvey a little bit about, so for those who don't know and most of you listening probably don't know, Harvey I met through, he's one of the moderators for our Discord, Intel Insiders. He also is active in the Intel DevMesh channel. and he is part of our Intel Student Ambassador program. So lots of ties to Intel there, and I can imagine that that's cool because when I was a university student, which is now twenty years, twenty plus years ago I would have thought that was really cool. But that was before the internet really took off and I would have had no idea how to do it, but I'm curious how did you get involved in all of these things and what's your experience?

[Harvey]: Oh gosh, I got involved a little over a year or year and a bit ago. I joined the Intel discord because I purchased some shares of Intel on a stock trading app and it was a nice community. To give them credit they were, I hate to go on NVIDIA again, nicer than NVIDIA discord and I just kind of kept chatting and talked to people in there and I just kind of became part of the community.

[Harvey]: And then I just happened to be obviously doing Gavin and stuff alongside that through conversations with people there and some of the employees in that I kind of got shown the different programs that Intel offers and I got made a student ambassador. And then I think later two weeks later I got made a moderator on the Discord and so I've been doing the two of them for pretty much the same amount of time in parallel which is actually quite fun.

[Harvey]: Because I get the hardware and that from Insiders which then I also get to use for stuff with Gavin and then I get the support on the technical side from ambassadors and I also get to showcase my work and and share why I like SYCL with a lot of people here at university because a lot of people here don't even know that Intel makes GPUs, so they look at me kind of funny when I come in and I'm like oh yeah I got an A750 here would you like to see it.

[Harvey]: It's nice to because a lot of people are very apathetic towards the situation of heterogenous compute and it's just cute or don't bother. It is nice to go in and show them that here you know you can you can do SYCL and I've done it and I've been able to do it because I've had the assistance and the training that Intel provided. And it's allowed me to go and do what I want relatively easily, which is the most important part for me because if it's too hard I ain't got the time unfortunately.

[Harvey]: But it's been it's been a blast.

24:27
[Tony]: It's great and also there was, you put together an event say, what is it a month ago? A month and a half ago at the University of Nottingham, is that right?

[Harvey]: Yes it was on October the eighth, we had some researches up from Cambridge's open zettascale compute lab I think it was called. They came up to give a presentation on SYCL and oneAPI but the toolset as a whole so it wasn't just SYCL or there was OpenMP, I think it was something called MPI. It was OpenMPI. We did stuff on VTune. And we did, yeah, that was about it.

[Harvey]: The only thing I'm really good at or though from that is VTune and SYCL. I haven't touched anything else. I think we did something on, it was that framework for, it wasn't openVNN, but it was something, it was not the machine learning framework it was MKL, that's it. The Math Kernel Library is what we did some stuff on as well. And it was quite fun, got lots of swag and gave people things and we had a lot of people turn up

[Harvey]: I think initially we were up around fiftypeople but then twenty more trickled then over the like first thirty minutes and we got up to about seventy but then, of course after the free pizza we lost a few people but it was fun.

26:00
[Tony]: Yeah that's pretty cool. The key learning there is pizza will bring people in the door.

[Harvey]: It does, 100% of the time.

[Tony]: Okay and so well as we kind of wrap things up, I'll ask each of you kind of the same question. You guys are both university students, and I'll give you a little framing, when I was a university student I wanted to write video games. Because this was the late nineties early two thousands and I liked playing games like quake and Unreal Tournament and I really wanted to go work for Tim Sweeney and make cool video games which I ended up never doing, but that's what really excited me about where the industry was going.

26:42
[Tony]: You guys are you know learning and going to university in this great age essentially of AI, parallelism, this huge growth in accelerated computing. What are you guys looking forward to as you kind of get done with your university studies and kind of move out into industry?

[Harvey]: Gosh, I always reply with embedded systems, but I don't mean edge embedded I mean high performance like very highly specialized silicon. Like when someone says tensorcores or ray tracing cores, that kind of peaks my interest. When you mention accelerating things through doing like special hardware that is what really interests me.

[Harvey]: And hopefully over the next week I have enough time to fully implement something, I've been working for the past two days on building out my own matrix library for C++ and I'm like 50% of the way there for the CPU code and then hopefully I'm going to then start porting it to SYCL to use XMX and start testing that. And that's because my ultimate goal is essentially to go lower and lower level until eventually I'mwriting everything and everything needs to be fast in assembly and having a full knowledge of it.

[Harvey]: So my future is X86 software dev manual, which is in my drawer. I finished reading that. I get good at writing assembly better than a compiler and I also know how to design hardware. That's what i'm looking forward to doing. Something that brings all of those together where I can make a big impact on performance.

28:28
[Tony]: Wow, I will tell you that our job that we are trying to do at Intel is trying to make sure that your job doesn't exist and that things are easy enough that people can do them at a higher level.

[Harvey]: yeah

[Tony]: we have not figured that out yet, totally, but we are trying to get there.

[Harvey]: Hopefully I'll be part of the making the solution I guess.

28:54
[Tony]: Yeah, it seems very likely from talking to you the last couple of months. How about you Josh?

[Joshua]: My dream is to abuse both of you and for the first few years, so in theory I would like a doctorate but I'm still in my first bachelors so I've got some time to say the least but I want to do some academic research. Preferably with AI to actually offer something to the field that I've been so interested in for so long at this point.

[Joshua]: Because I just think they'll be awesome even if no one knows my paper exists, at least I know it exists and that's enough. After that hopefully by some sheer miracle a company seems to like what I do and I go ahead and I continue research but under the pay and the embrace of a much, much larger company using tools and things that are designed by individuals like Harvey and even possibly tools that Intel have made.

To continue doing artificial intelligence, I've mentioned early about my interest with medical models and hopefully in the future once I've got in more knowledge we'll be able to implement and produce my own models that are used in hospitals.

[Tony]: That’s cool. That's some good goals and I hope you guys get there. That's all we have today for our Podcast. Thanks listeners for tuning in and thanks Josh and Harvey for joining us today

[Joshua]: thank you for having me

[Harvey]: same here it's been a pleasure