SDS PODCAST EPISODE 929: DRAGON HATCHLING: THE MISSING LINK BETWEEN TRANSFORMERS AND THE BRAIN, WITH ADRIAN KOSOWSKI

Jon Krohn: 00:00:00 What on Earth do transformer killer architectures have to do with potato latkes, with hum and Tussin? A dessert I had never even heard of before this episode and a Tower of Delicious Macon. In today's episode, you'll find out. Welcome to the SuperDataScience podcast. I'm your host, Jon Krohn. We've got Adrian Kosowski on the show today to tell us about an exciting innovation, potentially a transformer replacing architecture called BDH that him and his team at Pathway have come up with. Adrian is absolutely brilliant at blending biological neuroscience with machine learning, and you'll get to benefit from all of that intelligence in today's episode with some helpful takeaways as well. This episode of SuperDataScience is made possible by AWS Anthropic, Dell Intel and Gurobi.

00:00:58 Adrian, welcome to the SuperDataScience Podcast. Delighted to have you here. It's your third time on the show.

00:01:04 John, it's it a great to be here and thank you for having me again.

00:01:07 Yes. And for the second time we're recording in person together.

00:01:11 It's amazing. Yeah. Happy to be New York.

00:01:14 Yeah, thank you for flying out. Made the trip from San Francisco to New York very quickly. And the reason why. The reason why, so we were recording on a Monday and we decided on Friday that we were going to do this episode. You flew from San Francisco to New York on Sunday and the reason for doing all of this is that you at Pathway have had a huge breakthrough and yeah, so you have this paper that is just released out of embargo to the public. It's called the Dragon Hatchling, the missing link between the transformer and models of the Brain. And so we'll of course we'll have a link to the archive. I assume it will be an archive by the time this episode is out. Definitely. And yeah, so one of the first things to talk about is that your paper title is the Dragon Hatchling, but your abbreviation for this breakthrough, this new architecture, you're calling it BDH even though there's no B in the name. So do you want to explain the name of the breakthrough first before we get into the details of

Adrian Kosowski: 00:02:25 That? That's definitely one thing to explain because we are working on a new architecture and the new family of Reasoning models. This is a family called Baby Daggon and here we are announcing the Hatchling, which is a baby daggon, which just appears so it's no longer an egg, it's just hatched. But it has a lot to offer for the audience for a conceptual point of view. It's already something new and worth looking at, so we decided to go public with it.

Jon Krohn: 00:02:57 Nice. I love it. And so I've read your paper. It is fascinating. It's rich in both machine learning and artificial neural network concepts, transformer related concepts or ways of learning in artificial neural networks, but it also leans heavily on biological neuroscience, which is interesting.

Adrian Kosowski: 00:03:20 John, it's a pleasure to discuss with you given you have a background both in machine learning and in neur science. So happy to have this conversation.

Jon Krohn: 00:03:26 Yeah, exactly. So let's start with something called Hebbian learning, which it's kind of a big anchor in your concept. And so I actually to make sure that I didn't mess up explaining heavy and learning at all. I mean you could do it as well. So it's named after Canadian psychologist named Donald Heb. And when I was doing a PhD in neuroscience, like you just mentioned, this idea of heavy and learning is a big part of the way that we think about the way a brain in a living organism learns. And you might have some kind of other summary statistic on this or summary explanation of this. But the key idea that I take away from Hebbian learning is it's this idea that neurons, so brain cells in your brain that fire together, if something that you're thinking or something that you're seeing or something that you're hearing causes two brain cells to fire at the same time, that increases the probability that they will have a connection form between them or that that connection will strengthen. And if a brain cell is firing well another isn't, but they have an existing connection, that existing connection is likely to wane, is likely to kind of fade away. What do you think of that explanation?

Adrian Kosowski: 00:04:47 That's the way Hebbian learning works. That is a very quick introduction to what is actually a super fundamental process. One thing to take note of is the different timescales. And when thinking about verbatim or about natural models, at the back of our heads, it's always good to have an idea of how many new neurons are. We are talking about like 80 100 billion neurons and the timescales at which they operate. So if you think about all the processes, whether you like to take a perspective which is more on the physical side of dynamical systems or whether you want to go into the chemistry of what's actually going on, there are different processes which can take place in the timescale of seconds, minutes, hours, days, and those that go on throughout our lifetime. So all of the processes that you mentioned, John, which is about creating links, which is about strengthening links, which is about also links waning take place at different timescales and maybe governed by different dynamical processes.

Jon Krohn: 00:05:46 Very interesting. And so in this dragon hatchling or in this baby dragon family of models that you're developing, including in the dragon hatchling that you've now just published on, it seems like heavy in learning is a big part of the theory that you've used to develop it. Do you want to tell us more about this?

Adrian Kosowski: 00:06:12 Yes, I think we should look back a little bit about the history of the study of intelligence as such before we actually get to Baby D and Hatchling. Because in a way of the beginnings of computational science and studies of a Bain, like looking back into 1940s, the days of Alan touring, it all started together. And since then, advances in machine learning and in neural science have sometimes gone hand in hand, sometimes drifted a little bit apart. One topic, one keyword which was seen as a hope to reconcile them was around recurrent neural networks. The current neural networks were seen as a hope for at the same time the next machine learning architecture and a model which could attempt to explain what's happening in the Bain. Eventually what happened, as we all know it today, we are in an era where transformer is the keyword. We have large language models which are based on the transformer. And this architecture, this approach the transformer is harder to reconcile with what goes on in natural biological processes.

Jon Krohn: 00:07:23 I see.

Adrian Kosowski: 00:07:24 So here the approach that we took as a foundational one, we started looking at the concept of attention. Now attention is a concept which actually comes from neural science and came into the world of machine learning from the 1990s until the pattern that we see today in the context of natural language processing, it has undergone a long evolution. And the approach that we took was really to reconcile the understanding of attention in natural systems as captured for example by heavy learning and the understanding of attention as the keyword behind thefor.

Jon Krohn: 00:08:04 I see, I see. Lemme try to recap back to you the history that you've just given so far. So you were saying that some artificial neural network architectures like the recurrent neural network were designed with a biological inspiration in mind. And the idea was that that could actually end up being a good model of how our brains work, of how biological brains work, of how human brains work. And then we ended up having this architecture with the attention is all you need paper getting close to a decade ago now actually, which is wild, that attention is all you need. Paper Paper described an architecture called the transformer and at the time it was all people who worked at Google that had come up with this idea of this transformer. And what you're saying is that the attention mechanism that the transformer has and that has now become widespread in all of the large language models of the frontier labs that are kind of at the cutting edge, whether open source or proprietary, that attention mechanism is it's designed with GPUs in mind and efficient compute. It's not designed to be a model of the way that attention works in biological brains.

Adrian Kosowski: 00:09:26 That's it. And I think what we can discover is actually it's worth studying attention as a concept because as you mentioned was like a GPU friendly implementation of attention, which has been since the attention is all you need paper, maybe even a little bit earlier than that. But the actual question of what attention is it, it's a somewhat more fundamental concept because you can look at attention as not just an implementation, it's a mechanism which allows you to efficiently manage your context to think in a contextualized manner. So based on our introduction, our listeners will be able to also approach the next things that'll be discussing differently. For example, even the word attention by now it's clear that we are discussing it in the context of machine learning and national systems and not for example, some other meaning as in attention, attention, whatever else. So context is important. And...

Jon Krohn: 00:10:25 So it seems to me like it could be a good opportunity now to give people a bit of an idea of what attention is in a biological organism. And it's something that we experience every moment right now. Hopefully people are listening to this podcast, hopefully their attention is mostly focused on the sound of my voice and then your voice when it comes on. But we also have, my understanding is that in terms of at least attending to conversations, the human brain on average has the ability to listen to 1.6 conversations at the same time. And so there are things like the cocktail party effect where you're deeply engaged in conversation with some person or a group of people and then in another corner of the room someone says your name. And so there was some, you weren't consciously attending to that other conversation happening, but it was being processed a kind of lower level of attention than the main conversation you're focused on. But when your name came up, boom, your attention pops over there.

00:11:31 So attention is I guess where your conscious experience is focused on any given time and we only have so much bandwidth of attention to provide. And so I could for example be using that say 0.6 if I am 100% focused on the conversation that you and I are having right now, Adrian, I could still have that extra 60% maybe kind of running on what am I going to say next, attending to internal conversations. And so yeah, this attention can be external, it can be internal and it doesn't necessarily, in fact, it's hard to imagine how it could have anything to do with the transformer architecture.

Adrian Kosowski: 00:12:15 Indeed, it's a path which needs bridging. We've made some steps in this direction. But one thing to look at if we look at the natural kind of systems at the Bain is that you can look at attention at two levels, you can look at the level of the entire system. Here we speak of what we consciously feel or what a Bain as the whole is actually attending to or listening to or focused on. And we can also look at it at the micro level for each new and specifically what is this new and focusing on? And the spoiler here, the hint is that actually what nuns care the most about is their connections to their neighbors, which means that this is basically what they are, who they connect to is in some sense what they are and if they want to pay attention to some of their neighbors, to what some of their neighbors are saying more than to others, this is the attention mechanism which is encoded through the new connections, a visa called synapses, which change in time, which can open up communication, can say they potentiate and allow communication between neurons more directly at super short timescales.

00:13:28 And as we speak of the nuance nuances in the parts of a brain which are responsible for processing conversation, for understanding, for reasoning around it are all the time acting on these synapses. So here we have this picture, imagine it as a massively dynamical system, a bit like a social network of neurons which just decide actively which of their fans, which of their neighbors they're listening to and which they're not listening to. So this is natural attention and indeed at the other extreme we have attention as it is understood in machine learning. So these interpretations of attention have changed historically. The one that perhaps listeners will be most familiar with is attention as it is explained in the context, a transformer implementation on GPU. So you have a certain number of vectors of attention in each layer of the model which is running. And then you search given the current, we say token, the current element of conversation, we search for elements in the past which somehow relate to it.

00:14:39 And at a very intuitive level, at the basic level of an architecture like the transformer, we are listening, we are connecting to words sounds that have heard in the past and in higher levels of the architecture. It's a bit like we are connecting to something that's not as well defined. Some higher order concepts which appear. So this is basically the understanding of the tension, which is very much about context lookups and searching. So it's like a search data structure which is very different from at least in description from the local mechanism that we know appears in the brain.

Jon Krohn: 00:15:20 Really well explained there Adrian. That was fantastic, very easy for me to follow. And so hopefully this now gives a bit of context around the history, neuroscience machine learning of those two different branches, kind of where they overlap, where they don't. And so your paper now, the dragon hatchling paper that's out, it is described as the missing link between the transformer models of the brain. So how does this paper blend together what we've been talking about so far and form a missing link?

Adrian Kosowski: 00:15:55 We introduce a post transformer architecture, it's an architecture which lies on attention and which fundamentally has a properties of a massively parallel system of noons. So you can look at it as a system, a network of a large number of noons which also communicate with each other. They're artificial noons, but they do have some of the properties of natural noons, especially in what pertains to how they address the problem of attention. And in some sense, this architecture is the missing link in the sense that it is more biologically plausible. It is closer to the brain and it also, hopefully this is something to be verified in experiments but hopefully explains some of the mechanisms of the functioning of reasoning in the Bain or at the very least provides a plausible explanation of how the Bain could use certain mechanisms to achieve, to achieve or performance known from machine learning models like the transformer.

00:17:03 So this is the connection to the brain and the connection to the transformer is such that we rely on an attention mechanism which is at a very general level the same, it is implemented in a different way. A couple of elements which we may dive into which are technical, it is what is known as a state-based model. State-based models are a form of reconciliation of some of the concepts of recurrent neural networks ance form. The fact that it is a state-based model means that attention does not have to be viewed like a lookup structure. You can also view it in this local perspective of actually paying attention to certain concepts. You don't have to look at it as looking back in time. So you have the state space interpretation.

Jon Krohn: 00:17:58 Excellent. Alright, so with this kind of state space interpretation and this dragon hatchling architecture that you have that could potentially replace a transformer, it sounds like it's exciting not only because it could be a more efficient replacement for the transformer and could be something that some years from now is kind of the go-to thing When we build a large language model, instead of going to the transformer as your choice, you go to some animal from the baby dragon family. And so that seems like one part of it that's really exciting. And then the other part you mentioned there is that it could end up being that this could help us understand the way that learning happens in biological systems better. And so it could potentially not only be a machine learning breakthrough, but a neuroscience one as well.

Adrian Kosowski: 00:18:50 That's exactly the hope and that's why we are approaching it. There's one more aspect which is perhaps worth mentioning, which is that there's kind of pointless replacing the transformer in places where the transformer performs well, but there are places where the transformer does have its limitations and the human brain is able to overcome 'em. These pertain to lifelong learning to reasoning over long periods of time learning with experience. So visa areas where a human mind can spend years potentially diving into a subject to perfect the state of the art and to push the state of the art forward. This does not have to be scientific state of the art. It can concern any kind of work activity, any specialized task in which the human mind is able to do it, whereas the transformer has its limitations. There's a lot of ongoing discussion about specifically reasoning models, synthetic reasoning models today, whether they're able to extend reasoning beyond patterns that they have seen in retaining data, whether they're able to generalize reasoning to more complex reasoning patterns and longer reasoning patterns. The evidence is largely inconclusive with the general no as the answer. Currently machines don't generalize reasoning as humans do, and this is the big challenge where we believe architectures that we are proposing may make a real difference.

Jon Krohn: 00:20:32 Fantastic. So the idea is that the baby dragon family, starting with the baby dragon hatchling, BDH, we could have a model that has no limits on its context window it sounds like. So you could theoretically have as much learning as you want and be able to attend efficiently all of that information.

Adrian Kosowski: 00:20:59 That's true, that's true. And the answer is yes. One thing to always bear in mind is that there's no such thing as a free lunch in the sense that especially for those of the audience who are familiar with post transformer architectures which have appeared in the past, there's often this attempt to somehow make context infinite, but at the same time compare it, just have a memory of the size of a peanut and hope for it to be able to somehow compare long, long reasoning chains of thought or long, long sequences of text. In our case, the way the architecture is designed, and obviously we invite everyone to have a look at a paper and take a look into the details, but the idea is really that there so much space, so much flexibility in manipulating storing context that this is not a bottleneck and at the same time it's efficient.

00:22:10 So here for the sake of analogy, if you take again the human being, one parameter that I mentioned the beginning is the scale of it. It's like it's 80 100 billion neurons. But also if you look at the way the Bain represents its state here we are talking about synapses. They're like a hundred trillion of them, not clear exactly how many, but it's 100 trillion. And I can assure you that none of us post 100 trillion tokens in the lifetime. It's orders and orders of magnitude less. So we are talking about a system which does have the storage space, it does have the state insufficient quantity to be able to process long context but needs to do so efficiently. So it's not to waste time on doing some operations which don't move it forward.

Jon Krohn: 00:22:59 And so we'll talk about that in a moment later in the episode. What you're starting to allude to there is this idea that something that's a characteristic of transformer based architectures is that they tend to be densely activated. So you're activating either all the neurons or more recently it's become invoked to kind of have modules of neurons, but then you still have dense activation within those modules. And when I say dense, I mean that you're basically, you're flowing information through all of the neurons in the network or all of the neurons in that module. And this is computationally expensive energy, expensive. If the human brain were to do that with the trillion connections that we have in it, we wouldn't have enough energy to support that. And so we have a much sparser activation system in our brain, in the human brain, which is that, yeah, so only a relatively small portion of neurons are being activated at any one time and there's all kinds of things.

00:24:00 So then you can do things like functional magnetic resonance imaging, FMRI studies or EEG electroencephalogram studies to be able to see what parts of the brain are active during particular types of thought. And that you can do that is demonstrative of the fact that we have a sparsely activated brain. And so a big thing that you've done with BDH is that this now allows for sparse activation. So you can correct me if I'm wrong on the stat, but it looks like 95% of the artificial neurons in BDH are silent at any given time. They're not firing and yet the model is able to rival. So right now you've only been experimenting with it being a little baby dragon hatchling. So it's a relatively small network. It's about a billion parameters, which is about comparable to GT two. GT two is densely activated, so all of the neurons are being activated at any given time to any given prompt, whereas in yours only about 5% are active at that time, much more like a brain. Maybe explain a bit more about that and how central that efficiency principle is to BDH as well as the implications for biological brains.

Adrian Kosowski: 00:25:22 Absolutely. So John, indeed, the idea of working with models at a billion parameter scale here is largely for the sake of demonstration of certain tasks which we can work with. And here we were really looking at the core tasks that can be considered advanced from the point of view of cognition related to language related to translation of tasks that need attention. We are demonstrating it at this scale. Indeed, the aspect of sparse activation is a fascinating one. It's an aspect which is it's quite deep because there's a bit of a chicken and egg story to unravel here in fact that two worlds which work well, which can be compared to each other, however there's a gap between them. So there's no middle ground. You can be in one world or the other world. One world is the world of dense activations, which as the world transformer is in the other world is the world of sparse positive activations.

00:26:38 This is the world that baby dagan hatchling is in incidentally, the use of sparse positive activations in biologically inspired models. And their discovery in brain function is a topic which has been ongoing at least since the nineties with brilliant work regarding visual function of factory function, but basically sense of smell even for the food fly, which lies on sparse positive activation. So these topics have been explored from the point of view of a peasants in the brain, but the use of sparse positive activation for reasoning is something that's new. We believe we have the first work that has actually scaled it to transformer or beyond transformer performance. And the fact that we had 1 billion means it scales onwards, meaning all the key points have been achieved and we have achieved the scale, the implementation is GPU efficient by the way. So just reassure everyone with no, especially for infants, it has a lot of points which allow us to outperform the transform on certain hardware configurations quite significantly. So this is actually super encouraging, but maybe one more thing to say while introducing the topic, and maybe we can dive in just a little bit on the technical side, what these two worlds mean, but first one main difference between the two worlds, which I'd like to kind of highlight.

Jon Krohn: 00:28:14 And so just kind of recap those two worlds. So on the one hand you have the densely activated world of the transformer where as I was talking about earlier, when any prompt goes through in a fully dense network like GPT two, all of the neurons we run compute through every single neuron, every single connection on every input that goes into the model. Whereas the other world is the world that the baby dragon hatchling, BDH lives in where some small percentage are activated.

Adrian Kosowski: 00:28:46 So I would say to analyze the transformer, you can actually dive deeper because the here for the transformer, essentially if you look at the architecture which is behind the transformer, and for most open source models, the basic part is the GP two T two architecture, nothing much has changed. If you look at all the family of LAMA models, LA free lama for all of them, this is like GP T two almost unchanged.

Jon Krohn: 00:29:13 Yeah,

Adrian Kosowski: 00:29:14 Just scaled up. Just scaled up. So here the scaling of a transformer is actually a bit of an interesting story because there's no one single recipe for scaling the transformer. It ends up being done differently. You have a different number of attention heads, you have a different number of layers, et cetera, et cetera, et cetera. So the transformer scales in different ways and it's a little bit of know-how, how to scale it to actually, it's from a point of view of a, as somebody who would want to apply something like computational complexity and ask what's the limit of a transformer, it's actually rather hard to decide how we should scale it up. But first one dimension, which appears to be fixed to have converged and that's the size of the attention head, vector dimension in a transformer. So this does not scale even as the models get larger, it has stopped scaling. And here the intuition is that basically all concepts that the transformer works with have to be mapped into a vector space of about 1000 dimensions.

Jon Krohn: 00:30:28 I see, I see.

Adrian Kosowski: 00:30:30 Yeah.

Jon Krohn: 00:30:31 So we have this constraint that's appeared in the transformer that will limit its ability to have nuanced reasoning because it doesn't seem like we can scale the dimensionality of the vector space of the attention heads beyond a thousand.

Adrian Kosowski: 00:30:47 So this is actually a point that we this a bit more mathematically in our paper because we take the opportunity to analyze both baby D and hatchling and also kind of take the reverse approach to history and look at transformers and approximation of baby D and hatchling because it's actually easier to take this direction to say that we are coming up with a simpler and somewhat cleaner mathematically architecture cleaner in terms of the number of moving blocks. So here the advantages, we can kind of look back at what the transformers doing from this perspective and somehow see the way it organizes concepts and maximum them into a vector space. The kind of question which comes up, and this is I think for a large part of the audience who are familiar with vectors, with vector spaces, with notions even quite far from the transformers, such as let's say preference vectors, you can manipulate them as if it was a linear space.

00:31:51 So you can add vectors together, you can sub duct vectors, you can have the opposite elements, you can have a negative vector and this is the essence of a vector space. By contrast, if you go into the sparse positive spaces, the way you actually compose concepts is somewhat different. You don't work so much with Nina, combinations of vectors, you work more with bags of concepts, so bags of words, combinations. It works a bit like a tag cloud, it looks a bit like an association set. So a number of elements put together which form a whole, it's a bit how you form sentences by putting together words to get the meaning correctly. Or in some languages which have a bit more of a tendency to play, especially Germanic languages, German in particular where you compose multiple words to create a new noun for example. This is a place where you have this kind of compositional effect and there are a number of differences.

00:33:08 One difference to this type of representation apart from the points that you discussed, John, that efficiency and so on, is actually that you don't look so much towards negatives or opposites. Starting with a simple example, if you show somebody a piece of work that's been done badly and you tell them, look, this is what you're supposed to do, but do the opposite to get a good effect, there's no such thing as take the opposite and find the opposite to it. Likewise, in reasoning patterns, there's no symmetry between being attacked towards a certain reasoning pattern and being repelled from it. So if I tell you now don't think about the color blue, don't think about the color blue. You'll be consciously thinking about the color blue just to try to compensate for it and avoid it. But it's not the mechanism of switching off of DAMing down. So we are in a different vector space in different space.

Jon Krohn: 00:34:18 And so is what you're saying there is that, so transformer architecture, it doesn't behave in the same way as a brain with these kinds of negative activations, but your but BDH does your new architecture. Is that, so there's a term that you mentioned earlier and we kind of glossed over it. I wouldn't mind trying to dig into this a bit more. You described the BDH architecture as positive sparse, and so now we've been talking about negatives. Can you explain a bit more about this positivity thing?

Adrian Kosowski: 00:34:50 The question is actually a profound one because the correspondence between the two worlds, the world of dense vector spaces versus the world of sparse positivity, it's like two worlds, which is sometimes complimentary. Sometimes one is seen as a view of the other because it's possible to go mathematically between the two worlds. For the audience who is familiar with, again with hands-on machine learning, the vector world is the world where the L two norm rules, that's the kind of king or queen of norms. The L two, if you go the sparse positive world, you suddenly go into the world of probabilities, the world of probabilities, the world of chance, because the concepts that you're working with start to have interpretations of likelihoods or at least some value between zero and one, which reflects how much you are doing to a given concept. So in L two you don't have it in the probabilities, you have like an L one norm

Jon Krohn: 00:36:06 Interpretation then. So just to quickly say this, so in either of these cases, whether we're talking about L two norm or L one norm, these are ideas that have been around for decades in machine learning and in either case they're an additional factor that we add into our models that allows our models to generalize better to data that they haven't seen before.

Adrian Kosowski: 00:36:28 This is the idea that you introduce a concept that hasn't seen before, hasn't been seen before. And the question is where do you place it compared to other concepts? So we can walk through this, I will be using as pops a number of pastries. The first

Jon Krohn: 00:36:46 Apologies to our audio only listeners, this is segment. Adrian brought out a set of delicious treats which smell fantastic and it's very hard not to eat his props. So we have three different plates or serving dishes of desserts or of food. And they're critical to this explanation, I

Adrian Kosowski: 00:37:10 Suppose. Well, they are, at least they set the scene and I will explain as best I can to our listeners. So the debate, what's better like L one or L two from the point of view of working with basically any kind of dynamical system representation, this is a debate which is prevalent

Jon Krohn: 00:37:38 In many fields. And so what would better look like if one is better than the other? What is the

Adrian Kosowski: 00:37:45 Outcome? So there are differences, and I'll explain the differences. It's hard to say which one is better and that's why I'm bringing in cakes because everybody has their own preference for cakes and here it's there'll be applications, there'll be situations in which every cake is suitable. The first two selections, cakes in the selection are actually not due to me. They're due to one of the more renowned figures in quantum information quantum computing, Scott Aon who had this idea that the L two norm world compares to around potato cake like latke. And the L one norm compares to angular cake ham. And the explanation a little bit is to understand how you can, of course it's still taking a high dimensional space and deducing it into what we can see, what we can feel. But you can try to do it. So in a vector space you'll have this kind of round thing, it's like a ball in which you have vectors and you move around inside this kind of circle you have positive, you have negative. In fact, if you move to the L one spaces, then at this point you have corners, you start to have sharp corners. I see. And these are the kind of concepts that you are connecting. So the kind of well-defined entities are in the corners of a triangle. For example, if you take the lowest example possible, the smallest example possible, and the combinations of these are in the center. So you have this effect of combining corners to mix for men.

Jon Krohn: 00:39:38 I see. So the potato latkes are representative of the L two norm, and so it's not a coincidence then that potato lockkey is relatively homogeneous that it's all kind of the same kind of substance throughout. Whereas with this other kind of treat, which I must say is new to me, Lockkey is very familiar with, think they're delicious. These are humin. Tussin, Adrian Kosowski: 00:40:03 Yes.

Jon Krohn: 00:40:04 Nice. And then that's, so the tussin part, at least that's pocket in German.

Adrian Kosowski: 00:40:08 So I think both of 'em obviously come from the Dipo cuisine, but you can get both in central Europe in and then of course in New York and New York obviously, I mean it's like we can get everything in New York. But anyway, I'm bringing up these two because as long as you stay in discussion between these two, it's a bit of a debate as some of you may have seen that debates between the two kinds of gigs, academic debates at most, academic debates, academic debates. So you have to reduce your problem to the debate between the two. But the kind of thinking here is that if you are in this L one norm world, tusan world with the point corners with corners, then that's kind of like, okay, it's fine. But it only starts to get interesting when you go to higher dimension because sparsity sparsity needs higher dimension because you want to be choosing a few concepts from a very large group of concepts. And at this point we would need high dimensional pastry to be able to present this. You need

Jon Krohn: 00:41:20 A baker that can bake in more than three dimensions.

Adrian Kosowski: 00:41:22 That's it. That's it. So technically here we are talking about three concepts and we have a triangle, three corners on triangular

Jon Krohn: 00:41:31 Ion.

Adrian Kosowski: 00:41:32 The best I could do was go one dimension higher and here we are escaping central European cuisine to the best of French, still available in New York. Of course. Of course. So here we have a structure which is a triangle, but one dimension higher at least.

Jon Krohn: 00:41:52 Okay, yeah. So in this case here, for people who aren't watching the video version, we have a stack, it's kind of like people order those seafood platters with lots of layers of seafood where the bottom layer is the biggest and maybe there's some lobster, some crab on there. And then you have a medium tier that's a little bit smaller and you have the shrimps on there and then a top tier with scallops, that's the smallest or something.

Adrian Kosowski: 00:42:17 It is a kind of representation of a pyramid at least that was the objective to have a pyramid. And the interesting thing is that when you look at this kind of structure, first of all, the first thing you notice is that you don't eat vice center, vice center, you just have a plate to it. It's the things that are outside on the

Jon Krohn: 00:42:35 Walls

Adrian Kosowski: 00:42:36 That are interesting. So you have combinations of smaller numbers of corners, which give you the interesting elements. In fact, although we went from what is essentially a structure in two dimensions to a structure in three dimensions, we are not combining four concepts, but we're still combining usually three concepts to get the desired outcomes. Also, one thing about the specific type of construction,

Jon Krohn: 00:43:08 I also, I don't think we've even mentioned here, is that this is, so Adrian didn't bring a seafood tower into the studio, which might not have been the most considerate thing to do, that would've been pretty interesting. But it's a tower of Macon,

Adrian Kosowski: 00:43:22 It is a tower of Macon, and these macaron are beautiful ground objects. So you can think of each macaron as having a certain radius because indeed it does have a certain radius and you can have concepts which somehow have a certain radius of detection. Again, for those of you who are more into machine learning, this has a vibe of nearest neighbors kind of detection field where you want,

Jon Krohn: 00:43:53 And that sounds similar to the idea earlier of neighboring brain cells being the thing that are most interest to a given local brain cell.

Adrian Kosowski: 00:44:00 Yeah. So it is the kind of representation that we would be looking at. So again, this is kind of a little bit trivialized, but the other thing that we are trying to represent is how you can have a combination of different concepts and how it gives guys to kind of a new concept and how you can represent it in this past representation. And perhaps for most visual way we can think of it is through colors. Not like these concepts, positive or negative. I would want to leave the vector space world in which we talk about location and length and so on, but more about mixing colors. If you mix yellow with red with blue, you'll probably get something very brownish, but in some cases if you mix too many colors it's a bad thing. But if you mix just the right colors and just the right amounts, you get exactly what you want to get. So you have the whole kind of effect. The visual impact comes from the fact that you are just not mixing the whole palette together, but you actually picking a small number of colors to mix in the palette to place in one place in the painting. And that way these parse mixtures of colors allow you to represent the concepts that you want to represent and then you can only then you kind of place 'em

Jon Krohn: 00:45:23 In the space. And so is that why there's some specific color ordering to your tower of Macon to your four tier? Is that providing us with a fourth dimension of information?

Adrian Kosowski: 00:45:34 So I would say that the two, ideally in an ideal world with let's say perfectly, perfectly arranged towers, that would be a certain duality between the color and the location.

Meaning every color would have its location on the pyramid. So you could use one information over the other. I see. But somehow, the thing I want to hint at is that the location is less important then the actual, the color that this carries, I expect you could make the same analogy with taste and mixing taste. Again, if you're a good chef, you don't want to mix all the different possible tastes, but you either want to have a sparse combination of taste and you won't have a macaron with everything in it from chocolate to blueberry. But you will be trying to isolate individual one or two tastes and put them together. So

Jon Krohn: 00:46:27 Nice would've mess with your tower if I took a bite of a, I could take one from the back so it doesn't change anything for the audience. Nice.

Adrian Kosowski: 00:46:35 That's one of the things, and this is how we go towards sparsity. You don't have to represent all of them, but I can still tell you if the one that you've taken should have had that color. So

Jon Krohn: 00:46:47 Those are very good Mac and all thank you Adrian for bringing those into the studio. I'll dig into the lot keys and the ham in later.

Adrian Kosowski: 00:46:54 We'll be demonstrating sparsity.

Jon Krohn: 00:46:56 Nice. Fantastic. Okay, so I think we've hit on a lot of the key ideas around your new architecture, the BDH, and so hopefully we'd have a lot of theory now under our belts. So then maybe some of the questions I have next can maybe we'll see if they can be a bit more rapid fire. In the biological brain there's been an idea, there was an idea that was in vogue a few decades ago. It's fallen out of vogue a little bit now, but it was the idea that you would have a grandmother brain cell that would fire when you saw your particular grandmother or heard her voice that this particular neuron would fire. And in subsequent years it's become clear that for something that complex, you don't just have one brain cell amongst the 90 billion in your brain that fires, but you have a set of brain cells and those set of brain cells firing together gives you this representation of your grandmother. And so this seems similar, this kind of having a grandmother cell seems similar to something that you talk a lot about in your BDH paper, which is your discovery that there are specific neurons at fire for the idea of a currency or for the idea of a country. And the implication there is that it might be much easier with BDH to interpret what the neural network is processing than with a dense activation like a transformer based architecture.

Adrian Kosowski: 00:48:34 Definitely. And definitely this is two, the way to look at it first of all is again comes back to positive activations. One of the things to note about positive activations is that combinations are easy to express. You don't have to work so much with doing things like projections or finding negative and positive coefficients to say that I want this part of my network to activate in a positive way and that one in a negative way and balance it all together. You have full interpretability because you just say, I want this set, I want this combination to fire and then this represents my grandmother. The interestingly enough, if you dive into optimization of this type of architectures and find the natural patterns that they optimize for, those are pattern that the more important concepts are represented by generally smaller sets. So for those in the audience who are familiar with concepts of network science of systems versus this general search for power laws and systems, power law distributions in which the more important concepts are somehow more compactly represented and you can actually for the more important concepts, even in a relatively small network, even at 100 million scale or even below that find in our network an individual synapse which is sufficient evidence for a concept being mentioned.

00:50:20 So this touches on notion known as ity or being responsible for one concept and ticking on one concept. And this we see perhaps one more thing John to say is when you mentioned the notion of GaN mother cells and how they appear, they appears spontaneously and in architecture it's not architected. There's no genetic code that says, Hey, I wanted to have cells responsible for different operations. They evolve, they emerge in the course of training. We have no control over where they will be, but we find that they emerge and that you have this very clear location of signals, of signals passing through the artificial BA as a function of what we

Jon Krohn: 00:51:10 Are talking about. So kind of like your multicolored tower of Macon, you don't have control over where kind of the orange and yellow Macon end up in the tower, but you can bet that they will kind of end up aggregating together somewhere in the structure you just don't know where.

Adrian Kosowski: 00:51:29 It's pretty fascinating. And as we continue with this tower and it becomes sparser, you also end up with patterns in which the most important kind of locations are filled. So those more important concepts are filled and the less important ones are not that much filled. Maybe one thing to say for the audience who's kind of interested in technical details, but one fascinating technical detail here is that in our model, which we've been able to find and would actually shed some light on how these things work, is the concept of a GaN mother synapse rather than the GaN mother nun. Which means that if you think of how the state of the system works, the state of the system, the context that we are listening to is represented by synapse activation potentiation and those synapses that are responsible or that are related to specific contexts activate in those settings. So we have specific synapses which react to specific specific notions.

Jon Krohn: 00:52:35 And just for our listeners who don't have a neuroscience background, a synapse is where two neurons meet and chemicals pass between them and that is where learning must happen. Chemicals cross over this tiny little gap, the synapse between two brain cells and it's kind of analogous to the idea of the parameter in an artificial neural network model.

Adrian Kosowski: 00:53:01 Yeah, this is how we see it and being able to align the two is actually

Jon Krohn: 00:53:09 Nice. Alright, so next question for you related to this. Something that I found fascinating about your paper, about your BDH paper is you were able to concatenate literally just like a concatenate operation. You could have one neural network trained on one language, let's say English, and you could have another language trained on let's say French in honor of the Mackin all here and with your architecture. And this seems like a rare thing to be able to do with an architecture that could be the building block of a large language model. You can just concatenate those English and French language models together and because of the sparse activation, it just works and it's a multilingual model.

Adrian Kosowski: 00:53:58 That's the spirit and I think touches on so many different aspects, which I think are good to highlight because it's something new. It's new in many senses. As I mentioned before, the transformer while obviously being an amazing breakthrough in the focus of machine learning and AI in general does have its limitations in the way we understand it scaling. So if you have two transformers and you put them side by side, there's no really clear way how to connect them in. BDHV is much easier in the sense that the model scales in one dimension, we call it the number of new ones N and it's like a size of a brain. And then if you want to put two such veins together, you can do it depending on what you do, it'll be a little bit like a mix of the skills that you had or you can also do some post staining for the combined vein and make sure it coordinates properly. But definitely if you just put for Bain side by side, you have a model which out of the box has understanding for the different languages or is able to map them into concepts in English, for example and to work with them.

Jon Krohn: 00:55:12 That is very cool. Alright, so with all of these incredible novel capabilities of BDH relative to transformer, so the positive sparse activation that we've talked about, this ability to concatenate that comes out of that, the energy efficiency that comes out of it and compute efficiency that comes out of it. Where are you today? It kind of sounds like you've with this paper with BDH, with the baby dragon hatchling paper, we're talking about a billion parameter model, which is about the size of GT two from OpenAI, which is now some years old and it performs comparably to GT two despite requiring far less

Adrian Kosowski: 00:55:54 Compute. So just to reassure the readers, listeners, to this point, we are looking at models which at a given scale are on par with models of a given scale. So really it's given all the pocus but has happened in the state of the art. We use that progress obviously as the one B models that we produce are comparable or outperform the one B models out there. The kind of focus and the reason why we focus on this one B scale for demonstrations is that this is a scale at which we are able to achieve instruction following and to start testing other capabilities of a model which is able to actually follow instructions and to have a basic capabilities that we would expect a language model. And this is really for the ease and speed of experimentation. There's nothing particularly stopping us from releasing a super large model like in the 70, 80 billion scale larger.

The kind of question which is super pertinent is why do it? Because if you are in the world of language models, just language models versus certain market, which we could call a bit of a commodity market for the kind of chatbot like applications,

Jon Krohn: 00:57:25 Discussions and so on, right? So your clawed, your Gemini, your chat, GPT, they're competing in the same space.

Adrian Kosowski: 00:57:34 I think for switch that most of us are most aware of this. If you are working with a reasoning model or not, usually you're explicitly aware of the switch, especially with models like GPT versus a 1 0 3 with Claude, et cetera. You have this option to go into reasoning mode and this is the place where we don't want to just yet launch a non reasoning model, which is super large because was actually not our objective here. What we are doing is we are entering reasoning models, we are entering it from the moderate scale obviously, but this is a scale where we can display the advantage of this architecture. I see. Notably, yeah.

Jon Krohn: 00:58:26 Yeah. So the most promising avenue for you we're moving forward with this baby dragon family is into reasoning models. So models where you don't just have tokens output being spit out to your screen immediately, but there's multiple phases of reasoning happening in the background, refining your answer, ensuring accuracy. Yeah, that's where you see the most potential.

Adrian Kosowski: 00:58:51 That's it. Lots of consideration, lots of inspection. And also something that we see as extremely pertinent is the ability of reasoning models to work with contextualized inputs and to post them. So if you think of baking for barriers, the limits of 1 million token context, but you have reasoning model which goes through billions of tokens of context. Here you're in a space in which you can for example, ingest a contextualized dataset private to enterprise like a documentation of an entire technology, which is like 1 million pages of paper, 1 million sheets of paper, that's 1 billion tokens. You ingest it in a matter of minutes given enough hardware on this architecture. And with that in hand you can start actually making sense of large data sets in the way you would expect of reasoning models. Again, maybe for the developer audience out there, I'm sure you're familiar with use case of AI assisted coding in general, and this is perhaps for currently the frontier use case.

01:00:08 We are looking at the next generation of use cases like this, but to focus on this use case for a moment, the complexity of having an AI code assistant increases with the amount of preexisting code with the size of the code base. And usually it's much easier to have a model which contributes a piece of new code or just invents things without actually having internalized everything that was created before its actions. So it's basically doing a project on the side of its own then to have basically a model which is able to control and context contextually operate in an environment which requires understanding of a large code base. And again, code base is perhaps be the frontier example, but they're still the easiest kind of example that we are looking towards.

Jon Krohn: 01:01:06 Exciting. So for people who want to get their hands on this architecture or on models based on the BDH architecture today, can they do that? Are you providing these publicly?

Adrian Kosowski: 01:01:18 We are providing the architecture publicly a simplified version of the architecture, which nonetheless performs reasonably well comparable to the transformer and is provided in our paper. We have our internal enhancements of course, and especially those that allow this architecture to work faster. Transformer, especially in the influence generation and so on. These we keep internal. So you can lay your hands on the architecture, you can play with it. I believe it's an excellent playground for anybody interested in understanding models. So all topics of introspection, of observability, of getting a feel of how the model works and what we will be releasing in the near future. So follow on announcements, just hinting at them will be related to infant efficiency and reasoning on especially enterprise

Jon Krohn: 01:02:15 Data sets. Excellent and so fantastic for you to provide this for our audience so that they can implement the baby dragon hatchling themselves. So there's been hubbub before about a kind of architecture that could potentially replace the transformer as the go-to building block within a large language model. So for example, mamba was something that was pretty big a couple years ago and there was hype around this being superior to the transformer. I can't even now remember why was the mamba considered superior? Was it compute efficiency or

Adrian Kosowski: 01:02:53 It was compute efficiency in the case of lung context and bypassing limitations on lung

Jon Krohn: 01:02:59 Context, right? But yet today I haven't heard somebody talk about mamba in over a year. As far as I know, there's not a single frontier model that uses mamba instead of a transformer. So what makes BDH different in your view? How could BDH actually be a transformer killer?

Adrian Kosowski: 01:03:16 So there are two aspects at least, which I'd like to say in which we are more fundamentally different than the transformer. The first is something that I hinted at before that in order to get a non transformer architecture to work, you have to make a number of changes to make it work. You can't just say I'm changing one thing. So the change that's most often talked about is taking attention as we discussed attention before. Attention is an operation which viewed in vector spaces allows finding approximate closest neighbors due to the soft max operation used in the thisfor, at least as it is implemented in the transform. And one attempt was basically to take the soft max out of it. That's an operation which is called linear attention. And linear attention brings us into the world of state-based models. So this is one change, but they're in fact five changes that you have to put together to make it work.

01:04:18 Not one but five. So being in sparse, positive, large dimension and all this has to be put together to make these models work and to find ourselves out of the ridge costing the valley, so to speak, between two, from one place of local optimum, which is the transformer to a place further out and there things start to happen. So this is kind of one thing. A second point is also that you can be completely radical with use of our architecture. There's a lot of talk of supplementing the transformer with one or two layers, which are different, which help a little bit. So like the hybrid transformer and maba for example, a hybrid transformer and something here, there's no need for this. You can call all go all the way baby dragon. You can have this entire architecture all consistent. And what this means is that you don't complicate your picture compared to the transformer.

01:05:17 You simplify it. When you simplify it, you simplify the way it works with hardware, whether it is GPU, whether it is a GPU alternative. So one of the AI accelerators one out there, the D tensor based processing, it's GPU or lium or others, the layout of memory is radically simplified with BDH. And you find that some of the bottlenecks in memory transfers of the transformer disappear because of this architecture. So in some sense, even if you take an improvement of the transformer, which doesn't have a bottleneck, but you still stick to the transformer, you will have the transformer based bottlenecks with BDH, you can start to fresh and start sharding, start ranging things differently for more optimal performance.

Jon Krohn: 01:06:09 Wow, really exciting times. Adrian, you and the Pathway team must be delighted to have made this discovery, this invention.

Adrian Kosowski: 01:06:18 It's a first. We are super excited about it. More news coming up soon, but definitely, definitely it's a good place to be.

Jon Krohn: 01:06:26 Congratulations. It's a big deal. It incredible for me to think about all of the disparate knowledge that you and your team, a pathway concentrated together to be able to come up with this breakthrough. It's inspiring and yeah, I look forward to seeing BDH in Frontier LLMs all over the world a few years from now. Exciting times. Indeed. Before I let you go, you'll surely remember this because you've been on the show before. I do always ask my guests for a book recommendation.

Adrian Kosowski: 01:06:59 I have no choice this time, John. I mean, I could think of different books, but since we are talking about ENSs, I have to tell you the backstory, like white ENSs. So these ENSs are due to Terry Patchett,

Jon Krohn: 01:07:14 Terry Patch. Gotcha.

Adrian Kosowski: 01:07:16 Yeah, so it's for this Quil series and here it'd be kind of hard best to pick one specific book. I think you can start from the beginning. So it's the Color of Magic. Light, fantastic. The

Jon Krohn: 01:07:25 Color of Magic. Light,

Adrian Kosowski: 01:07:27 Fantastic. The first two books, the Color of Magic and Second Part.

Jon Krohn: 01:07:31 Very nice. All right, thank you for that recommendation. I'm sure we have some Terry Prt lovers out there listening already, and yeah. And then finally, where should people be following you or pathway to get the latest on your breakthroughs or pathways breakthroughs, your latest thinking After today's episode,

Adrian Kosowski: 01:07:52 We'll be sharing on our website a series of updates. Please don't hesitate to subscribe to that. That's pathway.com and of course, follow us on social media.

Jon Krohn: 01:08:04 Fantastic. Yeah, we'll have the links in the show notes for you. Yeah, what an exciting time. Crazy how quickly these innovations come, and it must be a real trip to be experiencing those innovations come up within your consciousness, within your biological neural network and your attention mechanisms.

Adrian Kosowski: 01:08:24 It is exhilarating, and the interesting part is that you actually get to understand yourself better as you work on this type of topics. So that's kind of attom meta level. It's kind of all in falling as well.

Jon Krohn: 01:08:41 Love it. Well, I hope to get the opportunity to check in with you and Pathway again on the podcast in the near future and see how these brilliant innovations are coming along. Thanks for joining me here in New York today, Adrian,

Adrian Kosowski: 01:08:52 Pleasure, anytime, John. Thank you.

Jon Krohn: 01:08:56 What a mind expanding episode with Adrian Kosowski on the brain on machine learning and how pathways exciting new BDH architecture could provide the missing link between the two fields. As always, you can get all the show notes, including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Adrian's social media profiles, as well as my at superdatascience.com/929. Thanks of course to everyone on the SuperData Science Podcast team, podcast manager, Sonja Brajovic, media editor, Mario Pombo, partnerships manager, Natalie Ziajski, researcher Serg Masís, writer Dr. Zara Karschay, and our founder Kirill Eremenko. Thanks to all of them for producing another outstanding episode for us today for enabling that outstanding team to create this free podcast for you. We are deeply grateful to our sponsors. You can support the show by checking out our sponsors links, which are in the show notes, and if you yourself are interested in sponsoring an episode, you can get the details on how to do that by making your way to johnkrohn.com/podcast. Otherwise, help us out by sharing this episode with folks that would love to hear about it, review it on your favorite podcasting app or on YouTube, subscribe if you're not a subscriber. But most importantly, just keep on tuning in. I'm so grateful to have you listening and hope I can continue to make episodes you love for years and years to come. Until next time, keep on rocking it out there, and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.