SDS PODCAST

EPISODE 978: A POST-TRANSFORMER ARCHITECTURE CRUSHES SUDOKU (TRANSFORMERS SOLVE ~0%)

Jon Krohn: 00:00 This is episode number 978 on the Post Transformer Architecture Crushing Sudoku Extreme. Welcome back to the SuperDataScience podcast. I am your host, Jon Krohn. Today's topic is a puzzle. Literally, we're going to be talking about Sudoku and why a game that millions of people around the world knock out every morning over their coffee is exposing a fundamental weakness in the most powerful AI models on the planet. So here's what happened. Pathway, a company I've had on the show before back in episode number 929 most recently. In that episode, I had their co-founder and chief scientific officer, Adrian Kosovsky in the studio to talk about their post-transformer architecture called BDH. Pathway recently published a research article alongside a benchmark called Sudoku Extreme that consists of roughly 250,000 of the hardest Sodoku puzzles available. And the results are striking. Pathways BDH architecture solved these extreme Sodoku puzzles with 97.4% accuracy.

01:01 The leading large language models, and we're talking 03 mini, deepSeqar1, Clot 3.7 Sonnet, scored effectively 0%, not low, zero. On a task that any reasonably practiced human can do with a pencil and some patients. Now, before we dig into why this matters and what BDH is doing differently, let's quickly recap what makes Sudoku such an interesting test for AI. Sudoku is a constraint satisfaction problem. In a nine by nine grid, every move has to satisfy multiple rules simultaneously. The numbers one through nine must appear exactly once in each row, once in each column, and once in each three by three box. A completed grid is trivially easy to verify, but producing that grid from a partially filled board requires searching through interacting possibilities without violating the rules. It's that combination of search, constraint, and backtracking that makes Sudoku a clean test of whether a system can reason under constraints rather than merely describe them.

02:00 And this is exactly where today's transformer-based LLMs start to struggle. Here's the core issue. Large language models turn every problem into text and then solve it by predicting the next token one step at a time. That works brilliantly when language is the right medium for a task, writing an email, summarizing a report, generating code, but Sudoku doesn't live in language. Forcing it into a chain of text is painfully inefficient because the transformer architecture processes information token by token with a limited internal state at each step. The data that represent the model's thinking, its latent space are constrained to roughly a thousand floating point values per token. And critically, each decision gets locked in as text is generated. Transformers can't hold multiple candidate strategies in parallel. They don't have the ability to step back and reconsider earlier moves without verbalizing every intermediate thought. Now, if you've been a listener for a while, this constraint on the transformer's internal representation might ring a bell.

02:53 When Adrian Kosovsky was on the show, he made this exact point. The size of the attention head vector dimension in a transformer has essentially stopped scaling, even as models get larger and larger in terms of parameters. It's converged to around a thousand dimensions, which means all concepts that the transformer works with have to be mapped into a vector space of about that size. Adrian described this as a fundamental ceiling on the transformer's capacity for nuanced reasoning, and during his interview on the show last year, I found that argument compelling. The Sudoku benchmark data now provide concrete quantitative evidence for it. So what is Pathways BDH Architecture doing differently that allows it to crush this Sudoku benchmark? BDH, which by the way, stands for Baby Dragon Hatchling, I probably should have mentioned that earlier in their episode, in this episode, is a playful name for the first model in Pathways Baby Dragon family.

03:43 And it's what Pathway describes as a native reasoning model. BDH maintains a much larger internal reasoning space, what they call a latent reasoning space that isn't constrained to verbalizing every thought as text. The analogy pathway uses is a chess grandmaster playing 20 simultaneous games with her eyes closed. She's not whispering each move to herself in words. She's internalized the patterns and can navigate the search space seamlessly. BDH is designed to enable that kind of internalized reasoning in a machine. There are a few key technical ingredients here that are worth understanding. First, BDH uses what are called sparse positive activations, meaning that at any given time, only about 5% of the artificial neurons in the network are firing. This is radically different from a transformer where you're flowing information through essentially all the neurons, dense activation on every single input. Mixture of experts models are somewhere in between, but we're not covering that in this episode.

04:38 Anyway, as Adrian explained when he was on the show, this sparse activation is far more biologically plausible. It's much closer to how a human brain actually works. We have around 80 to a hundred billion neurons and roughly a hundred trillion synaptic connections, but only a tiny fraction are active at any given moment. If our brains were densely activated the way a transformer is, we wouldn't have enough energy to power them. Second, like the best known post-transformer architecture Mamba, BDH is a state-based model, meaning it doesn't rely on the standard transformer attention mechanism that looks back through your entire input sequence to find relevant context. Instead, it maintains and updates an internal state, somewhat analogous to how biological neurons continuously update their synaptic connections based on what they're processing. This is closely related to heavy and learning, the foundational neuroscience principle, the neurons which fire together, wire together.

05:28 That's how we learn how biological animals learn. And Asia discussed this at length back in episode number 929, and he explains how BDH draws inspiration from these biological mechanisms. And third, and this is particularly relevant to the Sudoku results, BDH achieves what Pathway calls continual learning. It learns from every interaction and internalizes that learning over time. According to Pathway, BDH can pick up the rules of a new game and reach an advanced beginner level in as little as 20 minutes, then improve through repeated play. This is a far cry from a transformer which has a fixed set of weights after training and relies on in- context learning or chain of thought prompting to tackle novel problems. And now here's a detail that should get the attention of anyone thinking about the economics of AI. BDH achieves its 97.4% accuracy on Sudoku Extreme at materially lower cost than the leading LLMs achieve their near zero scores.

06:23 Because BDH reasons in its internal latent space rather than generating long chains of text, it doesn't burn GPU cycles verbalizing every intermediate step. Pathway reports that the cost is roughly 10 times lower compared to 03 mini, deep CKR1 and Sonnet 3.7 with no chain of thought required at all. Now, you might be thinking. Okay. Sudoku is interesting, but is this just a parlor trick? I think not, and here's why. The ability to solve Sudoku is really a proxy for the ability to navigate constraint satisfaction problems more broadly, holding multiple possibilities in parallel, backtracking when needed, and converging on solutions that satisfy all rules simultaneously. These are precisely the skills needed for countless real world challenges in medicine, law, operations, planning, and many other spaces, domains where you're balancing competing constraints under uncertainty. A system that can reason through these spaces natively rather than forcing everything into a text-based chain of thought could eventually do more than summarize information.

07:19 It could help generate strategy. Pathway, indeed, calls this generative strategy, looking at a problem, understanding the constraints, and creatively proposing what should be done rather than merely remembering what has been done before. That's an exciting frontier. Now, I do want to be balanced here. BDH is still early. When Adrian was on the show, we discussed how the architecture has been demonstrated at about a billion parameter scale comparable to GPT-2 from now many years ago. And Pathway hasn't yet released a massive frontier scale model. Adrian was clear that there's nothing stopping them from scaling much larger, but their current focus is on entering the reasoning model space where this architecture's advantages are most pronounced. That said, I find the Sudoku extreme results that I've covered in today's episode compelling as evidence that the transformers limitations are real and that alternative architectures can address those limitations. The data are clear.

08:13 0% from transformer-based architectures versus 97.4% from BDH is not a marginal difference. It's a categorical one. And if you combine this with the theoretical arguments Adrian laid out on the show previously about sparse positive activations, about the biological plausibility of BDH, about its potential for lifelong learning and reasoning over long time horizons, the picture emerges that this is an architecture that could meaningfully push AI capabilities beyond what transformers alone can achieve. All right. We've got links in the show notes to Pathways full research article on the Sudoku extreme benchmark, as well as to the previous BDH paper on archive, so you can dig into the technical details to your heart's content. The Transformer has been the undisputed king of AI architectures for the better part of a decade, and it's exciting to now see credible challenges emerge that are fundamentally rethinking how machines reason. All right, that's it.

09:08 If you enjoyed today's episode or know someone who might consider sharing this episode with them, leave a review of the show on your favorite podcasting platform or on YouTube. Tag me in a LinkedIn post with your thoughts, and if you haven't already, be sure to subscribe to the show. Most importantly, I just hope you'll keep on listening. 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.