SDS PODCAST EPISODE 758 FIVE-MINUTE FRIDAY: THE MAMBA ARCHITECTURE: SUPERIOR TO TRANSFORMERS IN LLMS

(00:02):

This is Five-Minute Friday on the Mamba language model architecture.

(00:19):

Welcome back to The Super Data Science Podcast. I'm your host, Jon Krohn. Today, we're focused on language model architectures that could replace the Transformer architecture that is essentially the only serious option for Large Language Models today, from the most capable text-generating models like GPT-4 and Gemini Ultra to image-generating models like DALL-E 3 to natural-language understanding models like BERT and the vast cornucopia of other LLM applications I could list. Modern, cutting-edge A.I. basically depends entirely on the Transformer. But now, the first serious contender to the Transformer has emerged and it's called Mamba; we've got the full paper — called "Mamba: Linear-Time Sequence Modeling with Selective State Spaces", written by researchers at Carnegie Mellon and Princeton — for you to explore in all its detail in the show notes.

(01:11):

So, why would anyone want a replacement for the Transformer anyway? After all, it has unleashed the broad range of mind-blowing, cutting-edge A.I. capabilities I just listed. Well, the problem is that the computational efficiency of Transformers decreases significantly as the amount of input data increases. To be precise, the computational requirements of Transformer models increases quadratically with the length of the input sequence. This means that if we input ten times as many words into a Transformer, its compute requirements jump up by 100x. If we input a thousand times as much context into a Transformer, its compute requirements jump up by a million, a thousand times a thousand, that's quadtratic.

(01:52):

I can only assume that the name Mamba itself — an extremely long snake that can measure up to 14 feet in length — is a reference to how this new Mamba architecture addresses the long-input problem of Transformers. This problem matters because long input sequences are common across a broad range of applications from everyday ones like natural language processing to more niche, but highly impactful ones like genomics.

(02:16):

To tackle the quadratic compute issue of Transformers, researchers have developed various architectures aiming to reduce this computational burden, including things like linear attention mechanisms, gated convolution and recurrent models, and structured state space models, or SSMs for short. Despite these efforts, none had quite matched the performance of traditional attention mechanisms, via Transformers, especially in key areas like language understanding and generation.

(02:44):

This is where the Mamba model comes into play. The Mamba model introduces a revolutionary approach by allowing the parameters of structured state space models, SSMs again, to be functions of the input. This means that the model can selectively decide which information to propagate forward through its neural network and which information to forget, and this is based on the content of the current token, which you can think of as a word, so it's based on the content of that particular content or word under consideration in the sequence that's being processed. This selective memory mechanism then, is crucial for effectively handling discrete modalities like language, where the relevance of information can vary greatly depending on the context. So Mamba effectively is selectively remembering what it needs to pay attention to based on the context specifically. And this is what allows it to be so much more compute efficient

because it's not trying to keep everything in its context like a Transformer is.

(03:54):

But the innovation doesn't stop there. The Mamba model also incorporates a hardware-aware parallel algorithm that operates in recurrent mode. This allows for efficient computation even without the use of traditional attention or the Multi-Layer Perceptron, MLP blocks, that are common components in many deep learning models, including in the Transformer-based LLMs that rule the world today.

(04:17):

Ok, so we've got this new selective memory mechanism and this hardware aware parallel algorithm. So great, that's the theory. But how does it actually perform? Well, that's why there's been so much buzz about Mamba and why this is the first time I'm doing a podcast episode on a potential Transformer replacement. Namely, not only does Mamba process data five times faster than traditional Transformer models under the same conditions, but it also scales linearly with the length of the input sequence. So, where 10x-ing the input length in the example I gave earlier corresponded to a 100x compute requirement for the Transformer, with Mamba the compute requirement would only go up linearly, it would only go up by 10x instead of by a 100x. This is a game-changer for processing long sequences, where Transformers previously faced significant challenges, because the longer the input sequence we're talking about, the greater the computational efficiency of Mamba relative to a Transformer.

(05:14):

All right, so again, going back to that earlier example, if we were 1000x-ing the input with a transformer, that's a million times more compute, whereas with Mamba, that's only a thousand times more compute, so it's a

thousand times more efficient than a transformer in that circumstance. But, does that extra efficiency come with a corresponding hit in performance? Apparently not. While I haven't experimented with Mamba yet myself, the paper's authors claim exceptional performance across a variety of modalities. Whether it's language, audio, or even genomics, Mamba sets a new standard for what's possible. For instance, in language modeling, the Mamba-3B parameter model not only outperforms Transformer models of the same size but matches the performance of Transformer models twice its size. There you go. So selective attention seems to have some performance benefits as well.

(06:06):

What does this mean for the future of deep learning and sequence modeling? The implications are vast. For one, the ability to efficiently process longer sequences of data without a significant computational penalty opens up new avenues for research and application. Whether it's improving natural language understanding, advancing genomics research, or enhancing audio processing capabilities, the Mamba model represents a potentially significant leap forward.

(06:30):

Moreover, the Mamba model's approach to handling sequence data, selectively remembering and forgetting information based on its relevance, could inspire new architectures and methodologies in the field. This concept of selective memory in sequence modeling could lead to more nuanced and context-aware models in the coming months or coming years, further bridging the gap between artificial intelligence and human-like intelligence.

(06:53):

To wrap up, the Mamba model presents an exciting advancement in the field of deep learning, particularly in the realm of modeling lengthy

sequences, including natural-language sequences. By addressing the computational inefficiencies of traditional Transformer architectures and introducing a novel approach to selective information processing, Mamba sets a new benchmark for what's possible in sequence modeling. As we continue to push the boundaries of what AI can achieve, selective attention mechanisms like those employed by Mamba could play a crucial role in shaping the future of technology and its applications across countless domains.

(07:29):

All right, that's it for today. I've got links to the Mamba paper, to the Mamba Github repo and to Mamba-Chat, an LLM derived from Mamba that's fine-tuned specifically for chat applications. I hope you enjoyed today's episode. If you did, consider sharing the episode with a friend or colleague, leaving a review of the show on your favorite podcasting platform, tagging me in a LinkedIn or Twitter post with your thoughts, I will respond to public ones, or if you aren't already, subscribe to the show. Most importantly, however, just keep on listening. Until next time, keep on rockin' it out there and I'm looking forward to enjoying another round of the Super Data Science podcast with you very soon.