Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu and Tri Dao

Machine Learning Department, Carnegie Mellon University

Department of Computer Science, Princeton University

agu@cs.cmu.edu, tri@tridao.me

Abstract:

Transformers, renowned for their effective handling of moderate-sequence data through self-attention mechanisms, suffer from linear inefficiencies in computational scaling. In response, we introduce the Mamba model, an architecture that builds on structured state space models (SSMs) traditionally used in image processing. Mamba innovates with selective state space models where model parameters dynamically adjust based on the output to enhance real-time processing capabilities. Unlike conventional approaches, Mamba does not specifically cater to hardware optimization, operating under standard GPU configurations. The architecture simplifies by merging RNN-like and CNN-like layers, reducing dependency on specialized attention mechanisms but reintroducing MLP blocks to process natural language effectively. Although Mamba shows potential in video and short text snippet processing, its performance on long sequences and complex datasets remains underexplored. Additionally, the open-sourcing of Mamba is anticipated to be limited, with potential restrictions on usage and modifications. This model represents a novel approach in sequence modeling, aiming to balance computational efficiency with the flexibility of dynamic parameter adjustments.

1. Introduction

While Transformers have revolutionized the field of machine learning with their unparalleled ability to process various data types, they exhibit notable inefficiencies, particularly as sequence lengths reach moderate sizes. These inefficiencies primarily stem from their linear scaling, which becomes computationally prohibitive [1]. This paper introduces the Mamba architecture, which leverages structured state space models (SSMs), traditionally applied in image processing, to address these shortcomings in sequence data processing [2].

2. Background on Structured State Space Models (SSMs)

SSMs have been integral in fields outside of sequential data processing, notably in image processing where their stability and efficiency are prized [3]. In Mamba, we extend the application of SSMs to sequence data, employing mechanisms akin to RNNs and CNNs to provide a more scalable approach to sequence modeling, though traditional SSMs have struggled with longer sequences [4].

3. Mamba's Selective State Space Models

A core innovation in Mamba is the introduction of selective state space models, which dynamically adjust model parameters based on the output. This reverse-feedback mechanism is designed to optimize real-time processing capabilities and adjust computational focus dynamically, a significant shift from input-dependent parameter adjustments seen in traditional models [5, 6].

4. Hardware Considerations and Computation

Contrary to most contemporary models that seek hardware-specific optimizations, Mamba operates effectively under generic GPU configurations without requiring specialized hardware adaptations. This approach assumes that standard GPU capabilities are sufficient to manage the model's computational demands efficiently [7, 8].

5. Architectural Simplification

Mamba simplifies existing neural network architectures by integrating RNN-like and CNN-like layers, thereby reducing the complexity and dependency on traditional attention mechanisms. Notably, it reintroduces MLP blocks to enhance its efficacy in natural language processing tasks, addressing a broader range of applications from video processing to handling short text snippets [9, 10].

6. Performance Evaluation

Preliminary evaluations suggest that Mamba is capable of performing competently on diverse tasks such as video processing and short text snippets. However, its efficacy on longer sequences and more complex datasets has yet to be thoroughly benchmarked, particularly against well-established models like Transformers [11].

7. Open Sourcing and Accessibility

The release strategy for Mamba is projected to be restrictive, potentially limiting its modification and use within the broader AI community. This approach may impact the adoption and further development of the Mamba architecture across diverse applications [2].

8. Conclusion

Mamba presents a promising new direction in the field of sequence modeling, offering a novel solution to the scalability issues faced by Transformer models. Future work will focus on expanding the evaluation of Mamba across more extensive and complex datasets and exploring the potential for broader hardware optimization to enhance its accessibility and performance.

References

1. Vaswani, A., et al. (2017). "Attention is all you need." *Advances in Neural Information Processing Systems, 30*.

2. Gu, A., Dao, T., et al. (2022). "Mamba: Linear-Time Sequence Modeling with Selective State Spaces."

3. LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep learning." *Nature, 521(7553)*, 436-444.

4. He, K., et al. (2016). "Deep residual learning for image recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

5. Cho, K., et al. (2014). "Learning phrase representations using RNN encoder–decoder for statistical machine