Understanding the Development of State Space Models (SSMs) and their Performance in Language Applications

Manav Rajiv Moorthy[1], Prateek Mahajan[1]
University of Washington, Seattle[1]

What are SSMs?

- Mathematical framework used to represent and process signals, inspired from control theory
- Describe a system based on state and input parameters
- Maintain a **hidden state** that evolves over time, with the output at each step being a function of the hidden state and the input

$$h_t = \mathbf{A}h_{t-1} + \mathbf{B}x_t$$

$$y_t = \mathbf{C}h_t + \mathbf{D}x_t$$

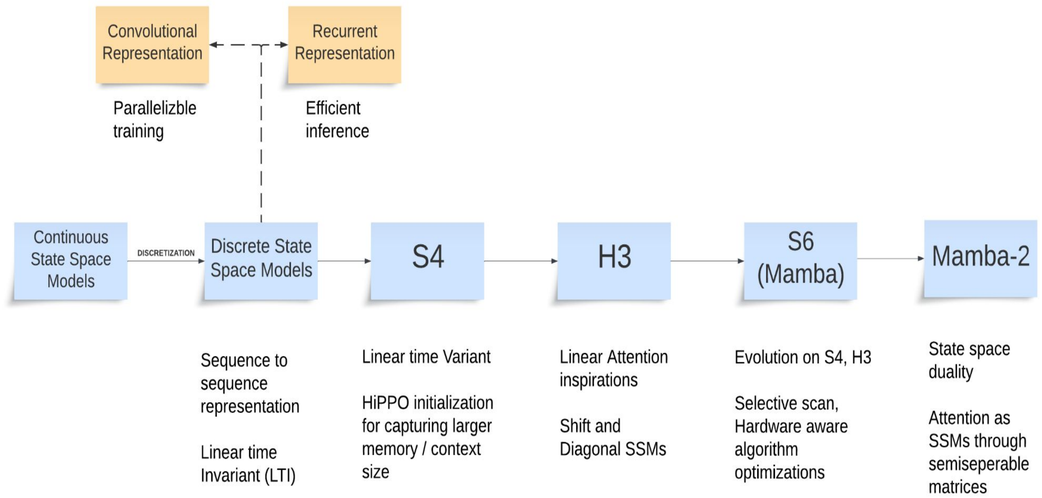
Why and How were they adapted for ML applications

- Matrices A, B, C, and D were discretized and parameterized using learned neural network weights and structured representations

Methodology Motivation

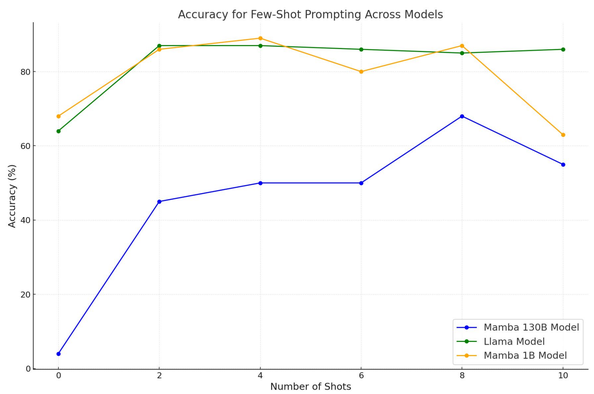
- These papers did very minimal testing in language applications (only zero-shot evaluation)
- We performed 4 experiments:
  - Few Shot Prompting: IMDB Review Dataset
  - SSMs in RAG applications : SSM Papers!
  - CoT Prompting : GSM8K Dataset
  - Long-ish Context : NarrativeQA Dataset
- Experiments ran on **AMD MI2104x GPU**
- Transformer Reference Model : Llama 3.2-1B

Evolution of SSM Architectures

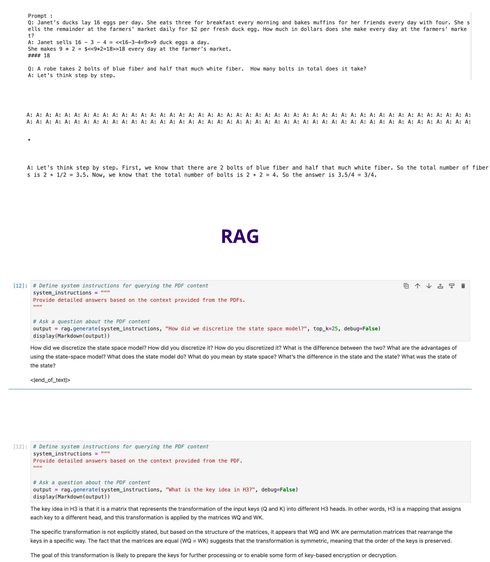


Results

Few Shot Prompting



CoT Prompting



RAG

Key Takeaways

- Model development is now also adopting a bottom-up approach where efficient systems are used as a base for new architectures
- Fine-tuning the model did not provide a lot of gain in terms of performance
- Mamba provided a comparable performance to Llama in few-shot prompting in classification
- SSMs in language applications still perform poorly in reasoning tasks like CoT
- Mamba-1B does not do a great job at retaining long-term dependencies (long-ish context)
- While SSMs address the quadratic computational inefficiencies in attention models, they are not feasible for all language tasks

Further Investigation

- Exploring more nuanced Vision, Audio and Forecasting applications (as they were covered in all these papers)
- Exploring newer SSM-based architectures that do a better job of using context information
- Comparing the perf of transformers to other alternative architectures (Griffin, Hawk, RWKV)