Understanding and Enhancing Mamba-Transformer Hybrids for Memory Recall and Language Modeling

Hyunji Amy Lee, Wenhao Yu, Hongming Zhang, Kaixin Ma,
Jiyeon Kim, Dong Yu, Minjoon Seo

UNC NLP

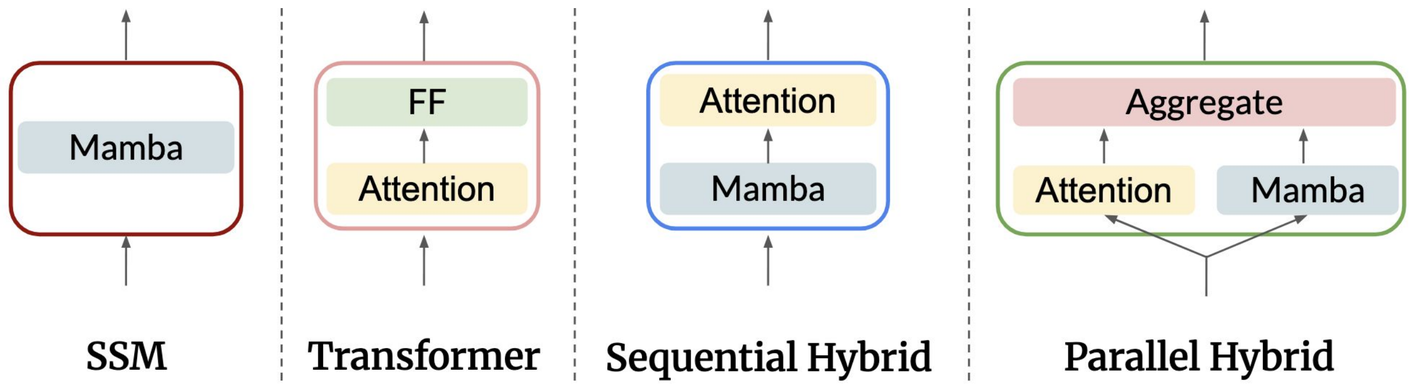
Tencent
KAIST AI
Kim Jaechul Graduate School

Motivation

Hybrid models that combine state space models (SSMs) with attention mechanisms have shown strong performance by leveraging the efficiency of SSMs and the high recall ability of attention. However, the architectural design choices behind these hybrid models remain insufficiently understood

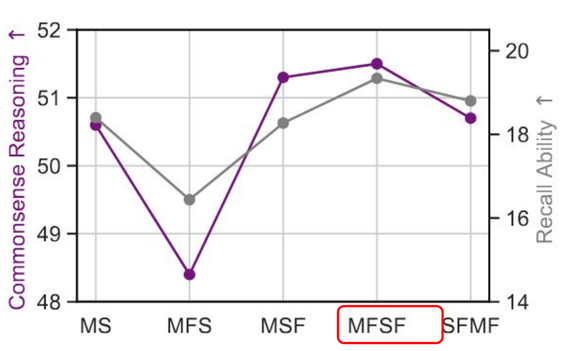
Evaluation Setup

- Four model types ⇒
- Three axes on evaluation:
  - Long Context Language modeling
  - Commonsense reasoning
  - Memory recall

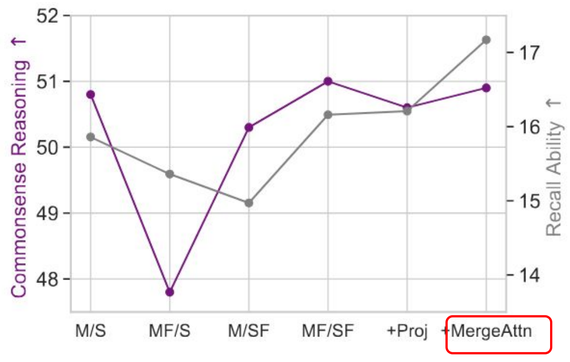


(RQ1) Aggregation Strategies:

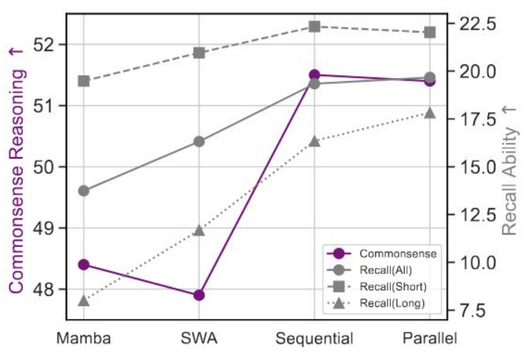
How do different ways of combining SSMs and attention affect performance and efficiency?



Sequential Hybrid



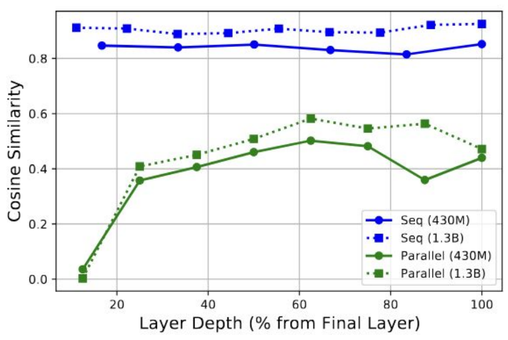
Parallel Hybrid



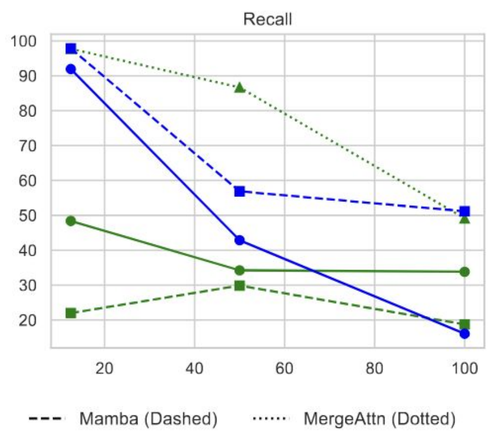
sequential models tend to perform better in relatively shorter contexts whereas parallel combinations general show superior performance in longer contexts

(RQ2) Component Roles:

What are the respective contributions and characteristics of SSMs and attention layers in hybrid models?



cosine similarity of their output embeddings across block depths: *Sequential hybrids show high similarity whereas parallel hybrids show much lower similarity*



performance degradation on recall tasks when removing blocks by depth:

- *Removing the first block causes the steepest drop (crucial role of early layers)*
- *In sequential models, early subcomponents are most critical; in parallel models, the aggregation layer matters most.*

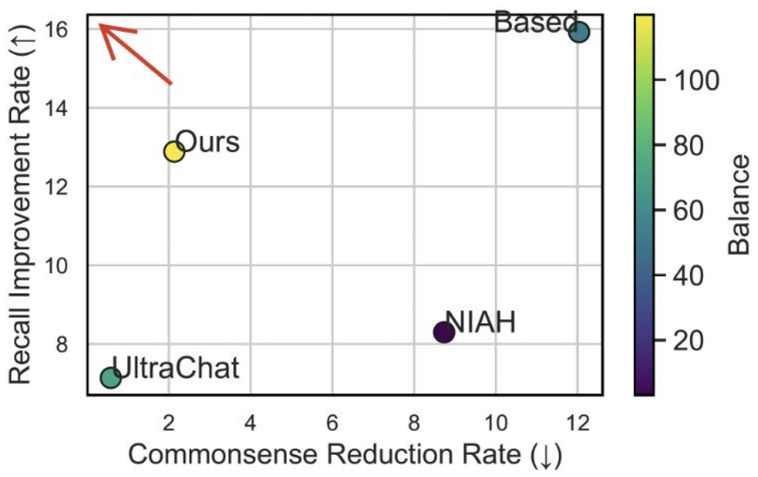
(RQ3) Data-Centric Enhancements:

Can performance be further improved through data-centric methods, beyond architectural design alone?

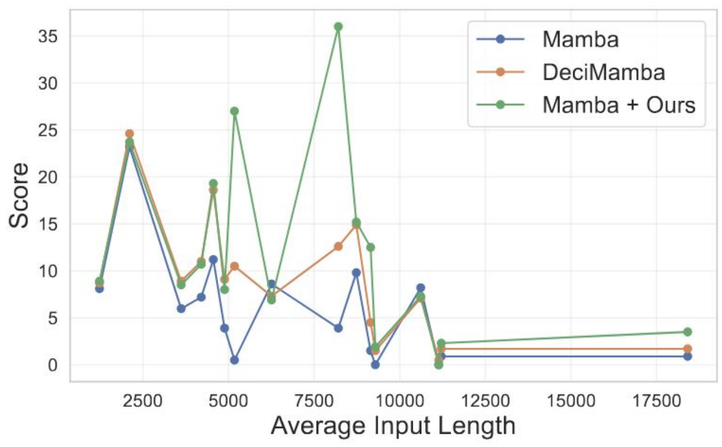
Motivation: Prior work has improved recall mainly through *architectural changes*. Here, we take a *data-centric approach* designed to complement and further enhance those advances.

- Add paraphrasing
- Data should remain close in distribution to the original pretraining corpus

> *continually training using a subset of the paraphrased training corpus (SlimPajama)*



Our Dataset Strikes the Best Balance



our approach tends to consistently outperform DeciMamba esp. on medium and long input lengths.

Additional results:

- our method generalizes across different released variants of Mamba-2.8B
- Longer chunk sizes yield stronger results
- Performance improves as the size of the training dataset increases