Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

AI Center Project in ML Research

ETH AI CENTER

Linear Language Models and their Associative Memory Limitations

Claudiu Craciun [1], Ross Roessler [1], Victor Zarzu [1], Heijhou Zheng [1],
Supervisor: Imanol Schlag [2]

[1]ETH Zurich Student, D-INFK
[2]ETH AI Center, ETH Zurich

1. Introduction

We evaluate the **associative memory** capacity of sequence decoders with different internal mechanisms — *replacing standard attention* with linear and recurrent layers, inspired by RNNs.

Goal: test how well these models generalize to longer sequences on structured tasks that require long-range recall.

2. Previous work

Transformers (Vaswani et al.) and their instances (e.g., GPT) have shown strong performance using self-attention and masked attention. However, their quadratic time complexity limits generalization over long sequences.

To address this, several methods propose linearizing attention, effectively making transformers behave more like RNNs—improving scalability and potentially restoring long-horizon associative memory.

3. Models

Starting from a generic decoder architecture (1), we have replaced the attention layer with several linear RNNs, as well as with a classic self attention in order to benchmark the linear Transformers.

We used three synthetic tasks to benchmark seven architectures: LSTM, QLSTM, LRU, Linear Transformers, GPT, DeltaNet, Mamba.

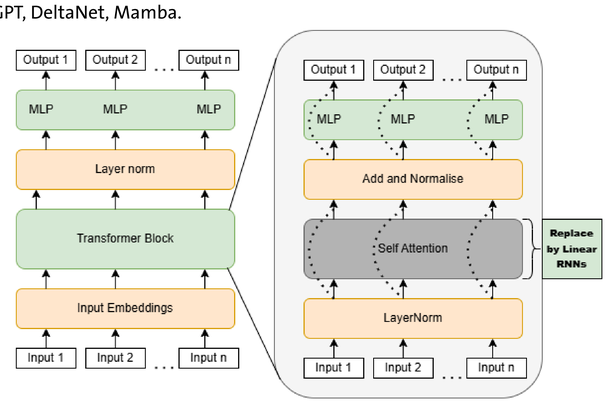

Figure 1: Transformer Decoder

4. Datasets and Methods

We have developed **3 synthetic datasets** that test the model's ability to remember previous states: bit parity, controlled bit parity, and Dyck.

Bit Parity

0 0 1 1 0 1 0 0 1 1 0 1 0 1 1

✗ ✗ ✓ ✗ ✗ ✓ ✓ ✓ ✗ ✓ ✓ ✗ ✗ ✓ ✗

Dyck - 3 paranthesis types, 6 max. depth

{ [( )] ([ ]) } [ ] ( )

✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✓ ✗ ✓ ✗ ✓

To evaluate associative memory and generalization outside of the training distribution, we trained models on short sequences of length 32 and evaluated them on progressively longer ones of length $48, 64, 128, 256$. Models were trained with identical hyperparameters.

5. Results and Conclusions

5.1 Bit parity

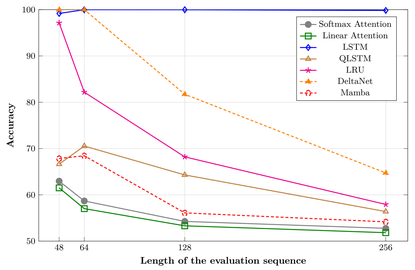

Figure 2: Accuracy on bit parity.

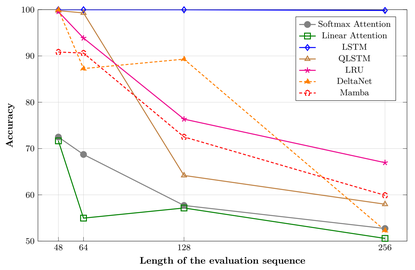

Figure 3: Accuracy on bit parity with 12 ones.

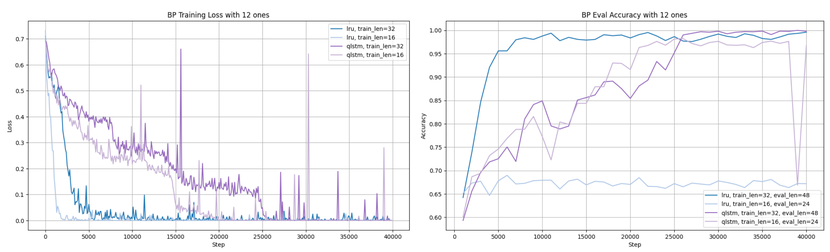

Figure 4: We compare training on sequences of length 16 and 32 with exactly 12 ones, evaluated on sequences 50% longer. Certain models such as LRU overfit when training on a high percentage of ones (Left) and do poorly when evaluated on longer sequences (Right). This does not occur when trained on longer sequences with a lower percentage of ones. Other models such as QLSTM do not exhibit this behavior.

5.2 Dyck

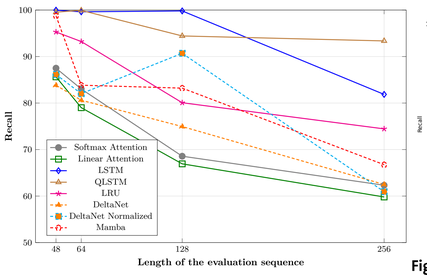

Figure 5: Recall of the models on Dyck, trained on length 32 sequences, evaluated for various lengths.

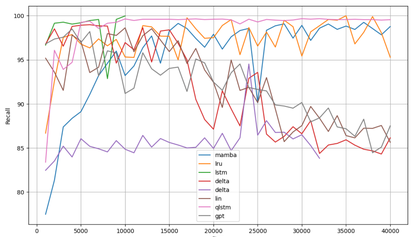

Figure 6: Evaluation recall on Dyck when trained on length 32 and evaluated on length 48 sequences. All models start comparably, but some do worse the longer they are trained.

5.3 Conclusions

- The drastic drop in evaluation accuracy on simple tasks shows that, other than the LSTM, all the other modified transformers lack a state representation of the sequences that they see.
- Most models perform better when we constrain the number of ones. We hypothesize that this is because the "1" is the signal that tells the model to change state, and so the fewer there are, the faster it can learn.
- LSTM performs the best, followed by DetaNet and the LRU. It would be interesting to explore DeltaNet and LRU in greater detail to understand what gives them the performance advantage.

6. Future Work

- Develop additional datasets on which to benchmark different models' associative retrieval capacities.
- Study the effectiveness of the models on a non-stationary task such as multi-query associative retrieval, where different (key, value) pairs are given to the model and then queried, but the values are updated over time at different rates.
- Understand in greater depth relative performance differences of some models on certain tasks; for example the QLSTM does much better than other models on Dyck, but only average on Bit Parity.