TAD
The Center for AI & Data Science
Tel Aviv University

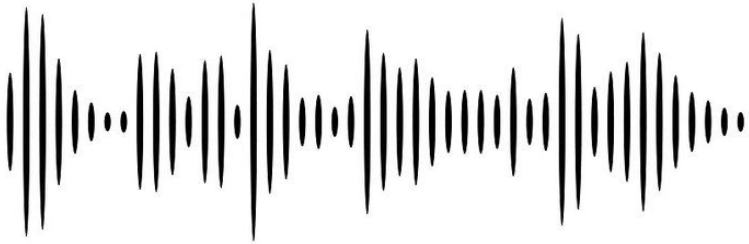
ICLR
International Conference On
Learning Representations

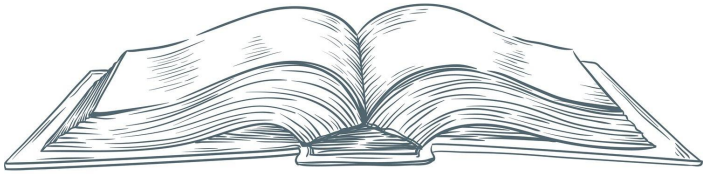
DeciMamba

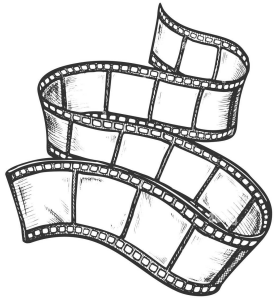
Exploring the Length Generalization Potential of Mamba

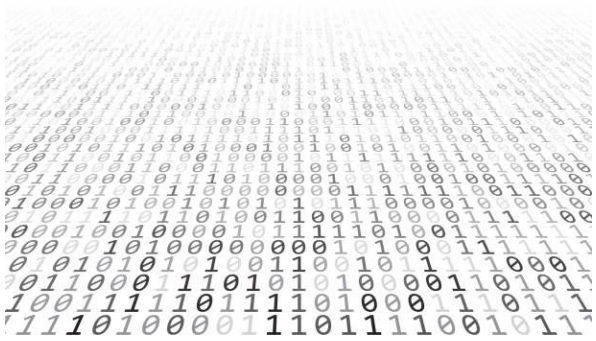
Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein,
Nadav Cohen, Amir Globerson, Lior Wolf, Raja Giryes

Long Sequences Are All Around Us









Transformers:

- Short Sequences - SOTA
- Long Sequences - Limited due to quadratic complexity w.r.t input length

The Long Sequence Architecture Zoo

Sub-Quadratic Recurrent Models

State-Space Models

You Are Here

Hyena

RWKV

Griffin

...

Transformer-Based Solutions

Flash Attention

...

Longformer

...

Activation Beacon

...

Mamba

A promising long sequence model, tailored for auto-regressive sequence prediction:

- ✓ Parallel training, sequential inference
- ✓ Scales linearly w.r.t input sequence length
- ✓ Performance on short-sequence tasks matches Transformers in important modalities like text.
- ✓ Intuitive sequence model
- ✓ Hardware-Aware implementation
- ❌ Length Generalization

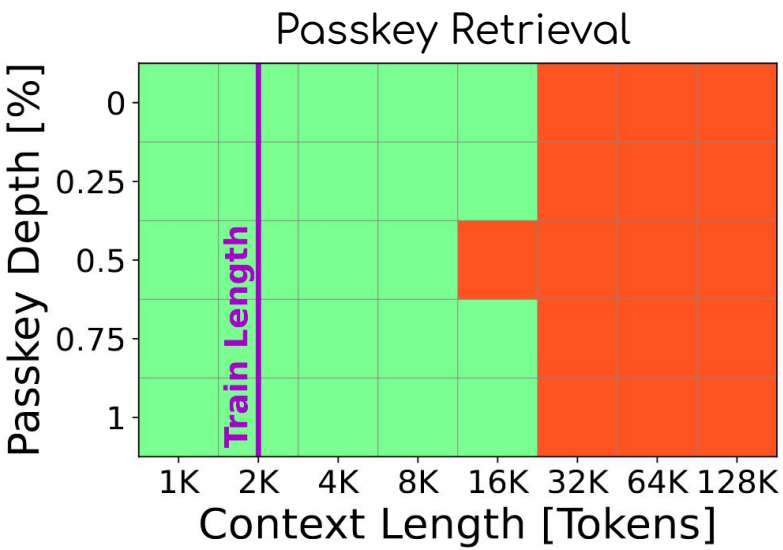
Studying Length Generalization via Passkey Retrieval

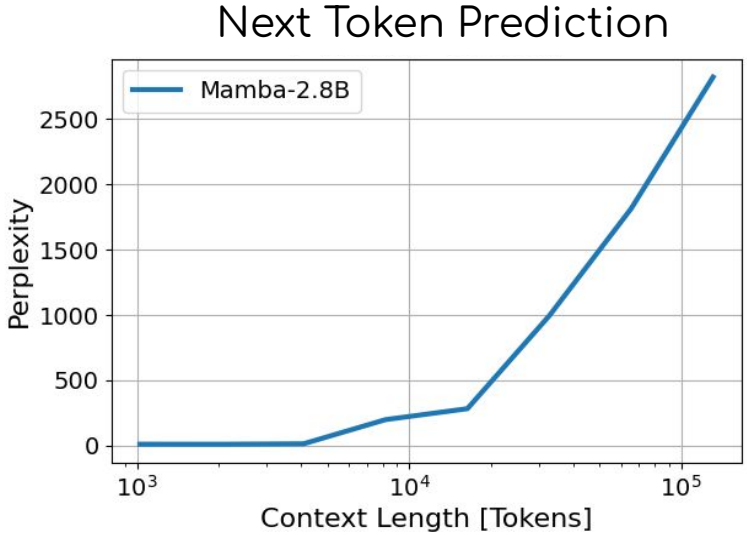
Sample from the Passkey Retrieval task:

There is an important pass key hidden inside a lot of irrelevant text. Find it and memorize it:
The Boston Celtics are an American professional basketball team based in Boston. The Celtics compete in the National Basketball Association (NBA) as a member of the Atlantic Division of the Eastern Conference. Founded in 1946 as one of the league's original eight teams, the Celtics play their home games at TD Garden…..The pass key is 51928. Remember it. 51928 is the pass key.….The Celtics are regarded as one of the most successful teams in NBA history and hold the records for most NBA championships won, with 18, and most recorded wins of any NBA franchise. Answer: 51928

Mamba Does Not Length Generalize Well

When trained on sequences of 2K tokens:

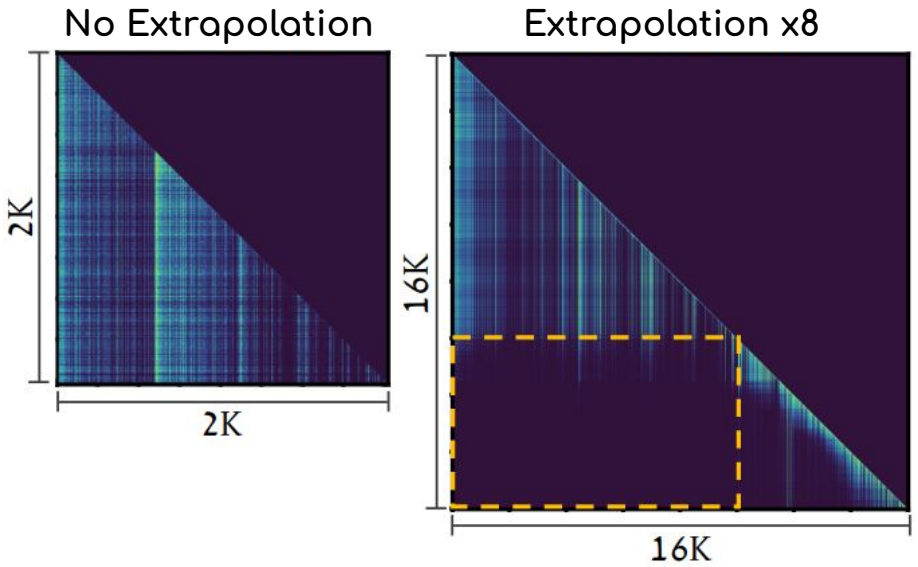




Identifying The Problem

Mamba's Hidden Attention During Extrapolation

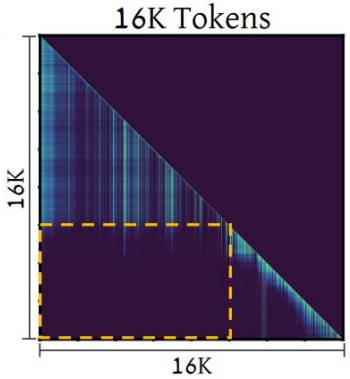
Attention Map
Layer 17



Limited Effective
Receptive Field

Why Do Limited ERFs Happen?

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix} = \begin{bmatrix} C_1\bar{B}_1 & 0 & \cdots & 0 \\ C_2\bar{A}_2\bar{B}_1 & C_2\bar{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ C_L\Pi_{k=2}^{L}\bar{A}_k\bar{B}_1 & C_L\Pi_{k=3}^{L}\bar{A}_k\bar{B}_2 & \cdots & C_L\bar{B}_L \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix}$$



$$\prod_{k=j+1}^{L} \bar{A}_k = \prod_{k=j+1}^{L} \exp(A\Delta_k) = \exp(A \sum_{k=j+1}^{L} \Delta_k)$$

The delta sum is bounded by the training sequence length L →
The learnt decay rate A is not constrained to support longer sums.

DeciMamba

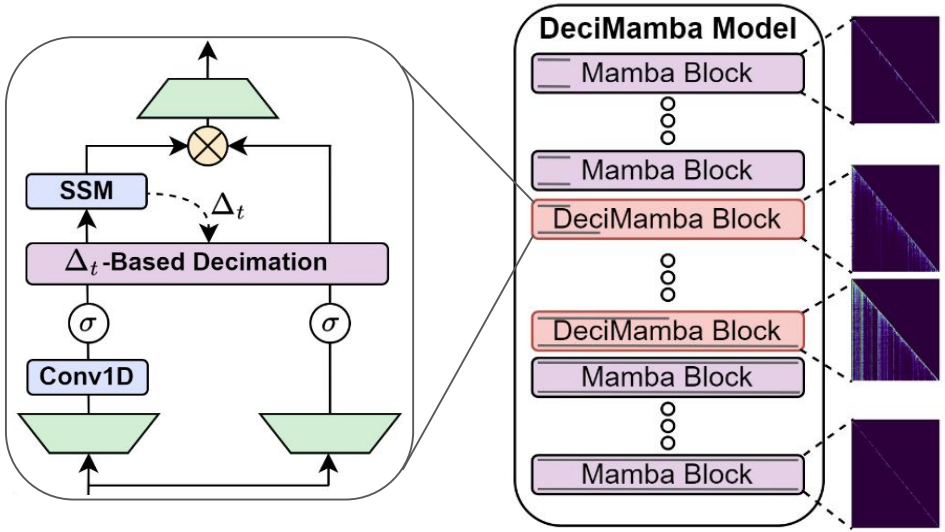
(Decimating-Mamba)

DeciMamba - Algorithm

(i) Assign token importance score

(ii) Perform global pooling (Decimation)

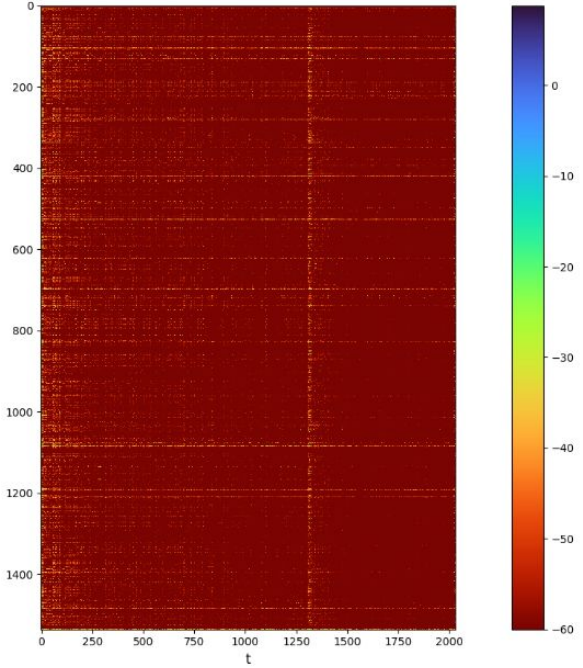
(iii) Feed the State-Space Model (SSM) with the pooled sequence



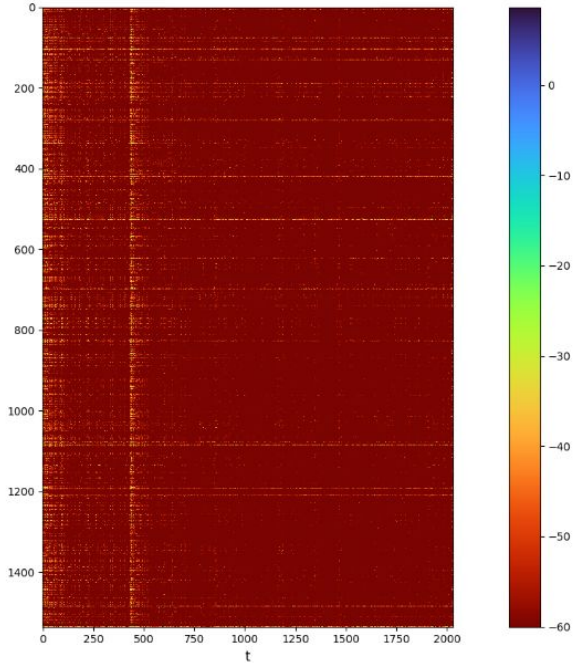
Δ_t = Token Importance Score

Values of Δ_t per channel:

Passkey at 1270



Passkey at 400



Layer 17,
Mamba-130M,
Passkey Retrieval

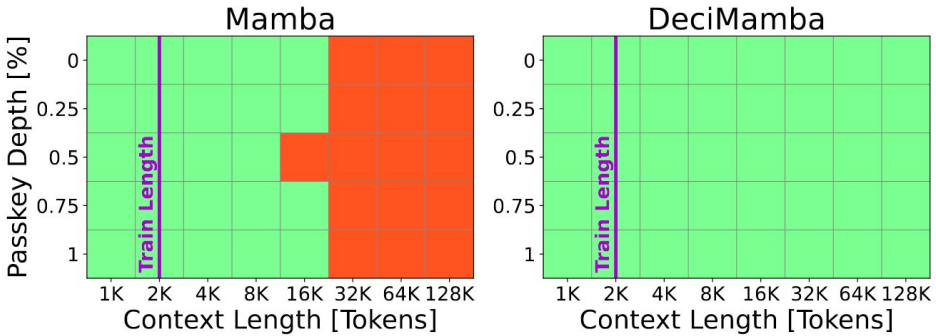
Results

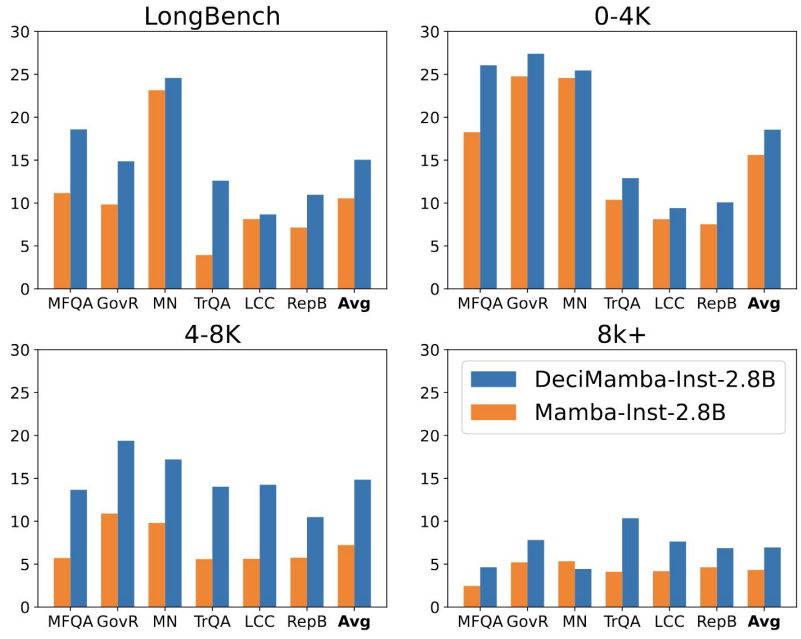
Results

Computation Speed [Sec]

Model \ Context Length	8k	16k	32k	64k	128k	256k	512k
Mamba-130m	0.17	0.3	0.69	1.27	2.12	4.03	8.17
DeciMamba-130m	**0.14**	**0.19**	**0.31**	**0.59**	**1.08**	**2.01**	**4.22**

- ✓ Extension by more than a magnitude
- ✓ Real-world long-range NLP tasks
- ✓ Works in zero-shot
- ✓ Less computational resources
- ✓ No re-training

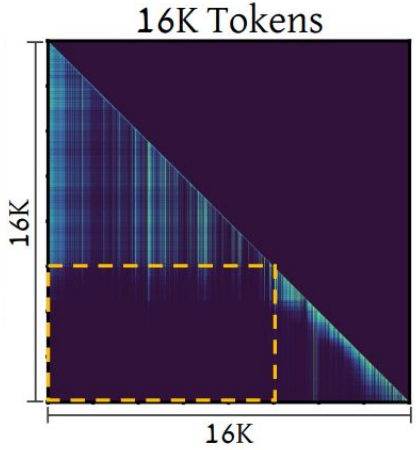


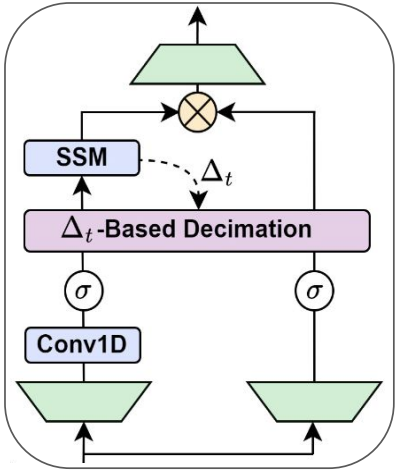


Recap

Recap

- Mamba is a promising long context model
- Mamba does not length generalize well due to chronic overfitting to the training length
- DeciMamba overcomes this limitation by pooling the sequence in global layers
- DeciMamba extends the context in both synthetic and real-world long context tasks





Thank You!

DeciMamba repo on Github