Neural Machine Translation by Jointly Learning to Align and Translate

Authors:
Dzmitry Bahdanau
Jacobs University Bremen, Germany
KyungHyun Cho Yoshua Bengio
Université de Montréal

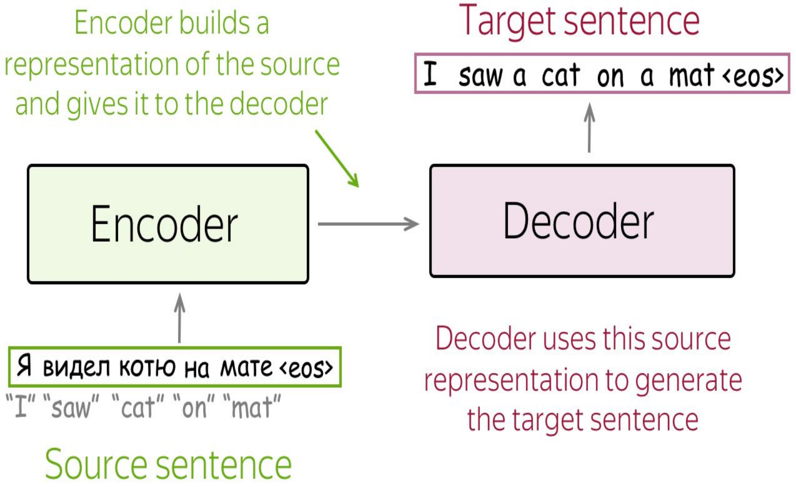
By: Mariana, Rodolfo, Fernanda, Meik

03.11.2023

Introduction to Neural Machine Translation (NMT)

- Machine translation is about **finding the most probable target sentence** (y) given a source sentence (x), using the conditional probability p(y | x).
- Neural machine translation (NMT) utilizes parameterized neural network models trained on parallel sentence pairs to learn the translation conditional distribution.
- NMT comprises of two key components: an encoder and a decoder.
- Recurrent Neural Networks (RNNs) are commonly used in NMT to handle variable-length sequences for encoding and decoding.
- **NMT has shown promising results**, achieving near state-of-the-art performance, and can even enhance traditional translation systems by incorporating neural components.

The Basic Encoder-Decoder


Encoder builds a representation of the source and gives it to the decoder
Target sentence
I saw a cat on a mat <eos>
Encoder
Decoder
Я видел котю на мате <eos>
"I" "saw" "cat" "on" "mat"
Source sentence
Decoder uses this source representation to generate the target sentence

- An **encoder** neural network **reads and encodes** a source sentence into a fixed-length vector
- A **decoder** outputs a **translation** from the encoded vector
- The whole system is jointly trained to maximize the probability of a correct translation given a source sentence
- The issue with this system is that a neural network needs to be able to compress all the information needed of a source sentence into a fixed-length vector (e.g. 10 vector length)
  - Making it difficult to deal with long sentences, especially those that are longer than what was used to train the model

Limitations of Fixed-Length Context

Limitations of fixed-length context in machine learning models, particularly in sequence-to-sequence tasks like machine translation, include:

- Inadequate for Long Sequences: Fixed-length context representations **struggle to capture the nuances of long input sequences**, as they may truncate or compress information, resulting in the loss of important details.
- Alignment Issues: When translating variable-length sentences or sequences, **fixed-length context may lead to alignment problems**, making it difficult to correctly associate source and target words.
- Context Loss: Important context, such as long-range dependencies or contextual information from earlier in the sequence, **can be lost or diluted in fixed-length representations**, affecting the model's ability to generate accurate output.
- Resource-Intensive Training: Training models with fixed-length contexts may **require more memory and computational resources**, limiting the scalability of models, especially for tasks with very long sequences.

Encoder–Decoder Model Extension

- The paper introduces an extension to the encoder-decoder model that **learns to align and translate at the same time**
  - Each time the proposed model generates a word in translation it searches for a set of positions in a source sentence where the most relevant information is concentrated
  - Model then predicts a target word based on context vectors associated with source positions and all previous generated target words

- This extension does not attempt to encode a whole input sentence into a single fixed-length vector but it **encodes it into a sequence of vectors** and chooses a subset of these vectors adaptively while it decodes the translation

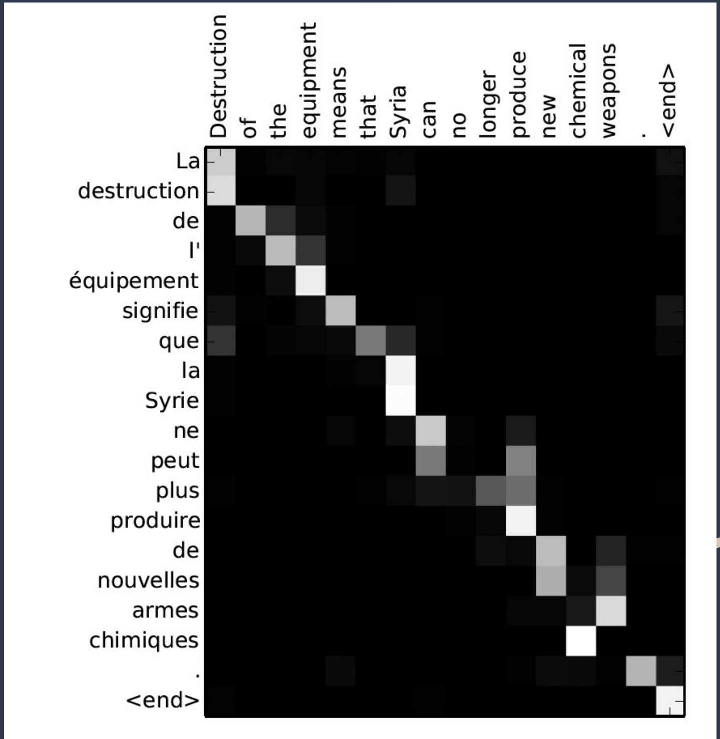
- Hence, making the model deal with longer sentences better

Alignment and Attention Mechanism

- The alignment mechanism refers to the concept of determining how words or subword units in the source sentence are aligned with words in the target sentence during the translation process.

- The attention mechanism is a specific implementation of the alignment mechanism within a neural machine translation (NMT) system. It is a computational mechanism that allows the NMT model to selectively focus on different parts of the source sentence when generating each word in the target sentence.

Alignment



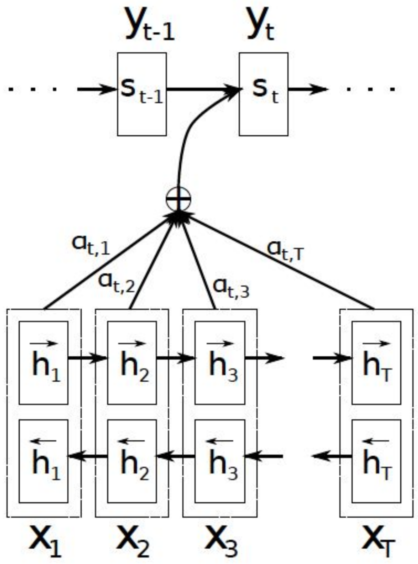
Alignment Visualization: Weights of α_{ij} , it indicates the weights associated with the annotations. We see which positions were in the source sentence were considered more important.

Soft Alignment Benefits: The use of soft alignment, as opposed to hard alignment, allows the model to make **more contextually informed translation decisions**. It naturally handles cases where source and target phrases have **different lengths**, without needing unnatural mappings of words to or from a "NULL" placeholder. This flexibility is particularly advantageous for accurate and context-aware translations.

Monotonic and Non-Monotonic Alignments: The alignment analysis reveals that the alignment between English and French is **mostly monotonic**, with **strong weights** along the **diagonal** of the alignment matrices. However, non-trivial, non-monotonic alignments are observed, particularly when translating words like **adjectives** and **nouns**, which have **different word orders** in the source and target languages.

Attention Mechanism

- Optimizes the model parameters to predict the correct alignment and translation probabilities for each word pair in the source and target sentences
- While decoding the translation, the Encoder-decoder with attention chooses a subset of these vectors adaptively



Proposed model trying to generate the t-th target word yt given a source sentence (x1,x2,...,xT).

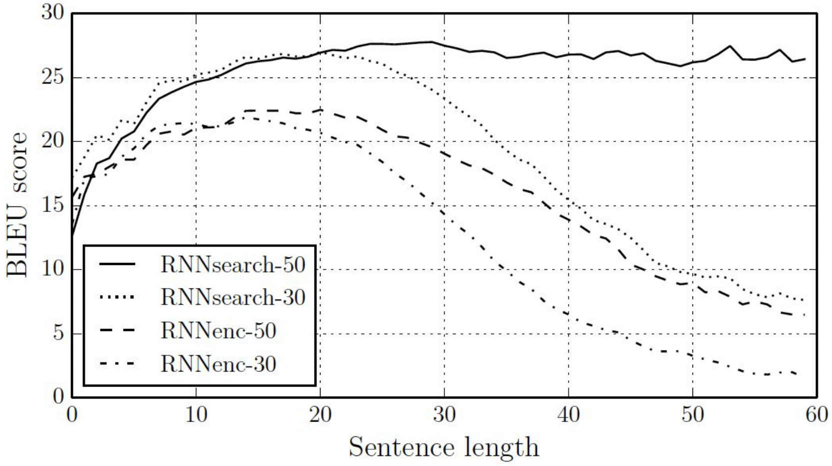
The Two Models Used

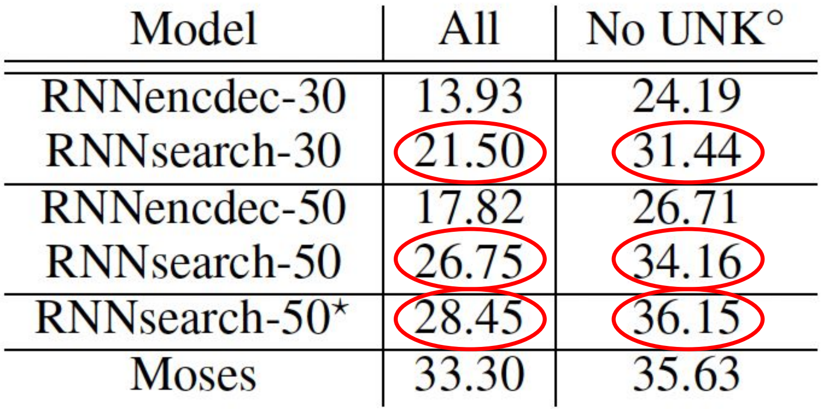
- RNN Encoder-Decoder
  - Encoder reads the input source sentence one word at a time
  - The hidden units summarizes the information in the input
  - Decoder generates the output sentence one word at a time
  - The alignment of the input and output sentences are fixed

- Proposed model: RNNsearch
  - An extension of previous model but it addresses the limitations of fixed-length context vectors by using the attention mechanism
  - The attention mechanism focuses on different parts of the input during the translation, not only word by word
  - This lets the model weigh and choose the information that is most relevant in order to translate
  - This model also learns to align the words in the input and output sentences during training
  - Dynamically focuses on the most relevant parts of the input sentence in order to properly give a translation (output)

Results and Performance

- RNNsearch did better overall compared to the RNNencdec model
- RNNsearch-50* was trained for longer
  - Until it stopped improving
- 30 is for 30 sentences
- 50 is for 50 sentences
- Moses is the conventional phrase-based translation system
  - Only sentences that have known words are considered

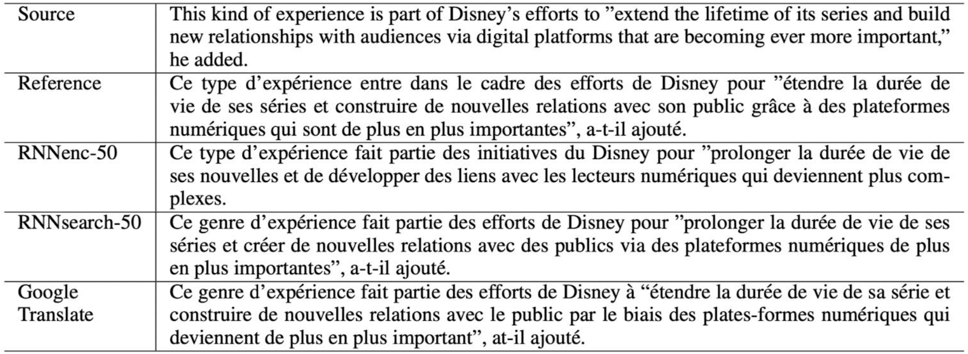




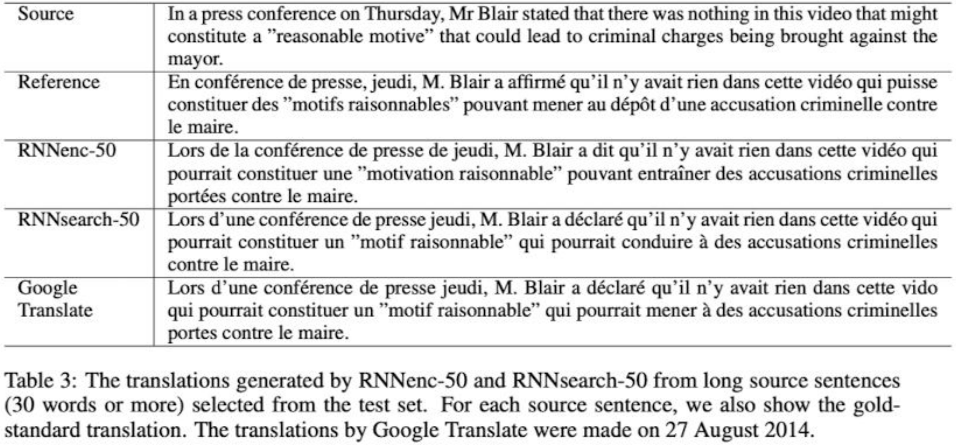
Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50⋆	28.45	36.15
Moses	33.30	35.63

Second column is score for all sentences and the third column is for sentences without any unknown words in them

Results and Performance



Source	This kind of experience is part of Disney's efforts to "extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important," he added.
Reference	Ce type d'expérience entre dans le cadre des efforts de Disney pour "étendre la durée de vie de ses séries et construire de nouvelles relations avec son public grâce à des plateformes numériques qui sont de plus en plus importantes", a-t-il ajouté.
RNNenc-50	Ce type d'expérience fait partie des initiatives du Disney pour "prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.
RNNsearch-50	Ce genre d'expérience fait partie des efforts de Disney pour "prolonger la durée de vie de ses séries et créer de nouvelles relations avec des publics via des plateformes numériques de plus en plus importantes", a-t-il ajouté.
Google Translate	Ce genre d'expérience fait partie des efforts de Disney à "étendre la durée de vie de sa série et construire de nouvelles relations avec le public par le biais des plates-formes numériques qui deviennent de plus en plus important", at-il ajouté.



Source	In a press conference on Thursday, Mr Blair stated that there was nothing in this video that might constitute a "reasonable motive" that could lead to criminal charges being brought against the mayor.
Reference	En conférence de presse, jeudi, M. Blair a affirmé qu'il n'y avait rien dans cette vidéo qui puisse constituer des "motifs raisonnables" pouvant mener au dépôt d'une accusation criminelle contre le maire.
RNNenc-50	Lors de la conférence de presse de jeudi, M. Blair a dit qu'il n'y avait rien dans cette vidéo qui pourrait constituer une "motivation raisonnable" pouvant entraîner des accusations criminelles portées contre le maire.
RNNsearch-50	Lors d'une conférence de presse jeudi, M. Blair a déclaré qu'il n'y avait rien dans cette vidéo qui pourrait constituer un "motif raisonnable" qui pourrait conduire à des accusations criminelles contre le maire.
Google Translate	Lors d'une conférence de presse jeudi, M. Blair a déclaré qu'il n'y avait rien dans cette vido qui pourrait constituer un "motif raisonnable" qui pourrait mener à des accusations criminelles portes contre le maire.

Table 3: The translations generated by RNNenc-50 and RNNsearch-50 from long source sentences (30 words or more) selected from the test set. For each source sentence, we also show the gold-standard translation. The translations by Google Translate were made on 27 August 2014.

Conclusion

- This paper adds on to the basic encoder-decoder by letting the model soft-search for input words when generating the output
- Allowing the model to not have to encode the whole input into a fixed-length vector and lets it focus on the information that is most relevant to the next translated word
- This model makes it possible to translate longer sentences with more accuracy
- RNNsearch outperformed the RNNencdec model
- This model can correctly align the input and output sentences as it generates a translation
- Challenges that will be left for the future would be to have the model better handle unknown or rare words

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

By: Mariana, Rodolfo, Fernanda, Meik

A bit of background

- Foundations models or large pretrained models for downstream tasks are now predominantly of the Transformer architecture and its core attention layer.
- Some drawbacks:
    a. Inability to model anything outside of a finite window
    b. Quadratic scaling wrt. Window length.

Introduction of MAMBA

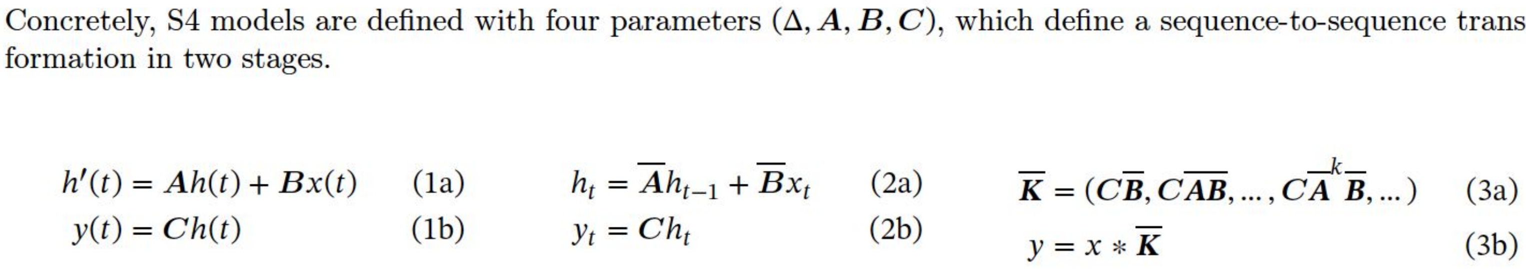
- Mamba builds on the state space model approach with key contributions.
- Selection mechanism added for efficient data selection.
- Hardware-aware algorithm for faster computation.
- Simplification of prior deep sequence model architectures into a single block.

State Space Models

Structure state space sequence models (S4) are a recent class of sequence models for deep learning related to RNNs and CNN and classical state space models.

Inspired by a particular continuous system that maps a 1-d function or sequence x(t) -> y(t) through an implicit latent state h(t).

- Discretization: first stage
- Computation: computed in two ways: linear recurrence (autoregressive inference) or global convolution (parallelizable training)
- Linear Time-Invariant (LTI): equations are constant through time



Concretely, S4 models are defined with four parameters $(\Delta, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, which define a sequence-to-sequence transformation in two stages.

$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t) \tag{1a}$$

$$y(t) = \boldsymbol{C}h(t) \tag{1b}$$

$$h_t = \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t \tag{2a}$$

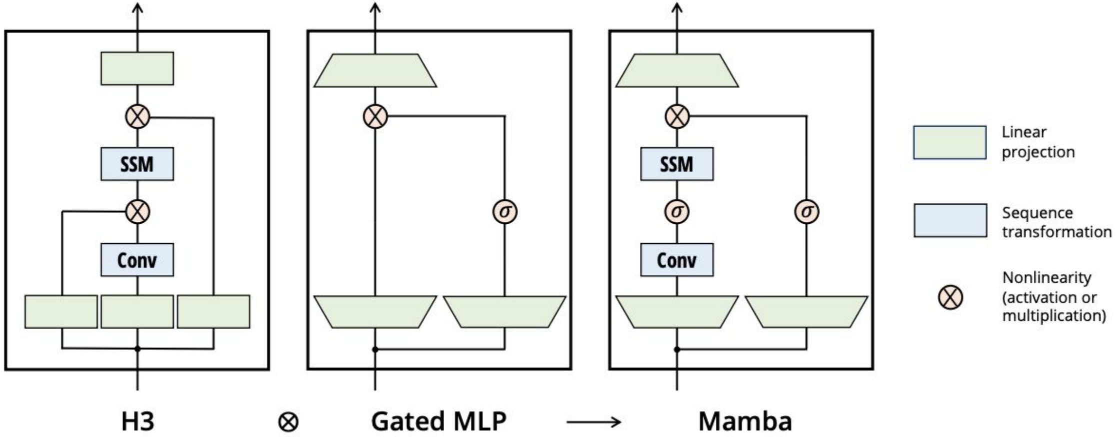
$$y_t = \boldsymbol{C}h_t \tag{2b}$$

$$\overline{\boldsymbol{K}} = (\boldsymbol{C}\overline{\boldsymbol{B}}, \boldsymbol{C}\overline{\boldsymbol{AB}}, \dots, \boldsymbol{C}\overline{\boldsymbol{A}}^k\overline{\boldsymbol{B}}, \dots) \tag{3a}$$

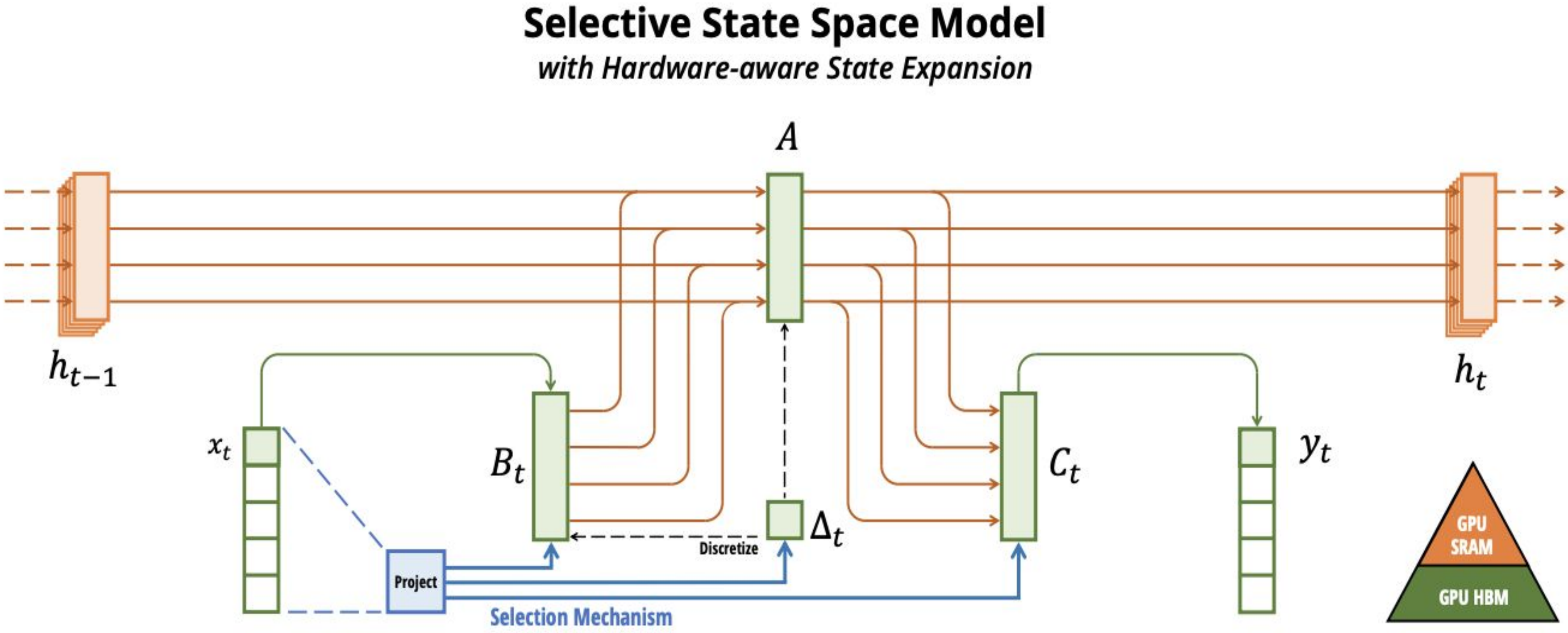
$$y = x * \overline{\boldsymbol{K}} \tag{3b}$$

Model Architecture – MAMBA

- Combines Hungry Hungry Hippos (H3) block with a gated MLP block.
- Selective State Spaces implemented via a scan instead of convolution.
- Employs parallel scan, kernel Fusion, and recomputation to address challenges.



Model Architecture cont.

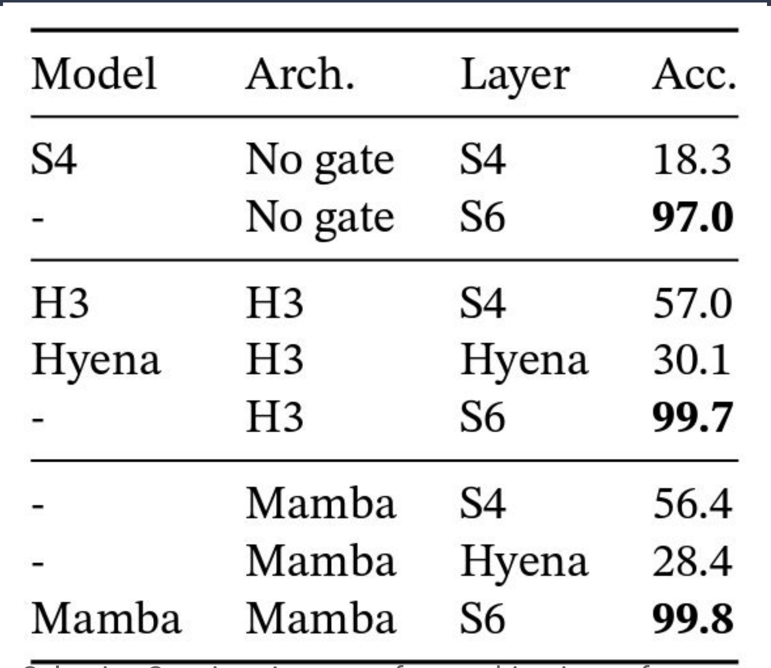


Empirical Evaluation

- Synthetic tasks, including selective copying and induction heads.
- Language modeling on the Pile dataset.
- DNA modeling on human genome data.
- Audio modeling and generation experiments.
- Speed and memory benchmarks, showing Mamba's efficiency

Copying Task and Selective Copying Task

- Copying Task
  - Testing how well the model remembers sequences
  - Linear models excel by tracking time but they struggle when token spacing is randomized
  - Gated architectures like H3 have limited success
  - **Mamba Advantage**: The selection mechanism in Mamba outperforms others
- Selective Copying Task
  - Prevents shortcuts in Copying Task by using randomized spacing between tokens
  - Gated architectures struggle
  - **Mamba Advantage**: Selection mechanism in Mamba does well but especially with powerful architectures

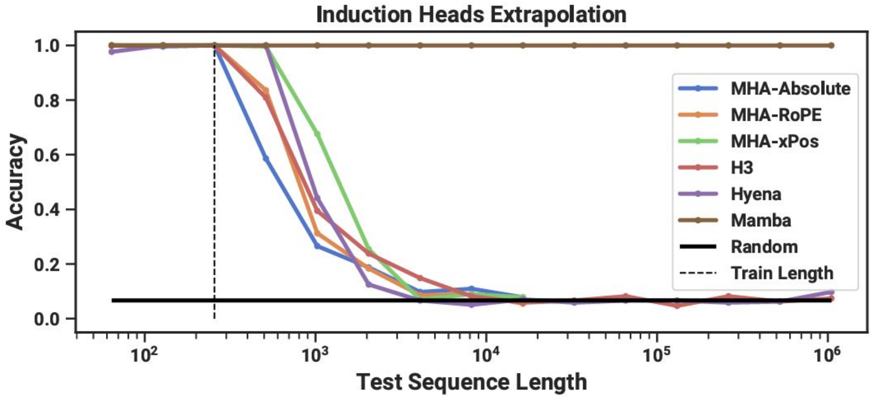


Model	Arch.	Layer	Acc.
S4	No gate	S4	18.3
-	No gate	S6	**97.0**
H3	H3	S4	57.0
Hyena	H3	Hyena	30.1
-	H3	S6	**99.7**
-	Mamba	S4	56.4
-	Mamba	Hyena	28.4
Mamba	Mamba	S6	**99.8**

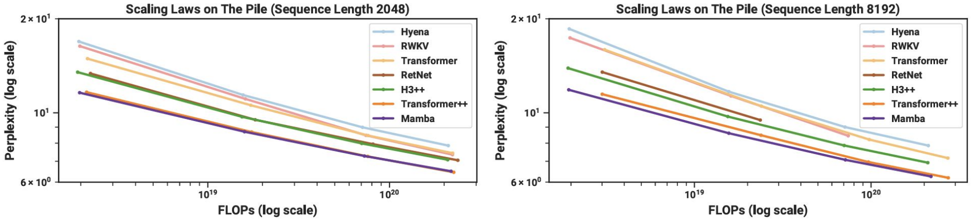
Selective Copying: Accuracy for combinations of architectures and inner sequence layers. Mamba with S6 (selection mechanism SSM with scanning (processing elements within a sequence) does well.

Induction Heads and Language Modeling

- Induction Heads Task
  - Testing in-context learning abilities
  - Requires associative recall and copy, predicting based on previously seen sequences (Harry Potter example)
  - Models trained on sequences of length 256 and evaluated for generalization and extrapolation (how well the model applies learned and new knowledge)
  - **Mamba Advantage**: Outperforms others by ability to remember and reproduce specific elements based on associations in a sequence
- Language Modeling
  - Mamba competes with strong Transformer recipes based on different architectures
  - Evaluating Mamba architecture on autoregressive language modeling shows that it outperforms other architectures
  - **Mamba Advantage**: Highlights efficiency in long sequences under the Chinchilla protocol (standard evaluation model used to test the scaling performances of models by comparing their results across varying parameters and model sizes)



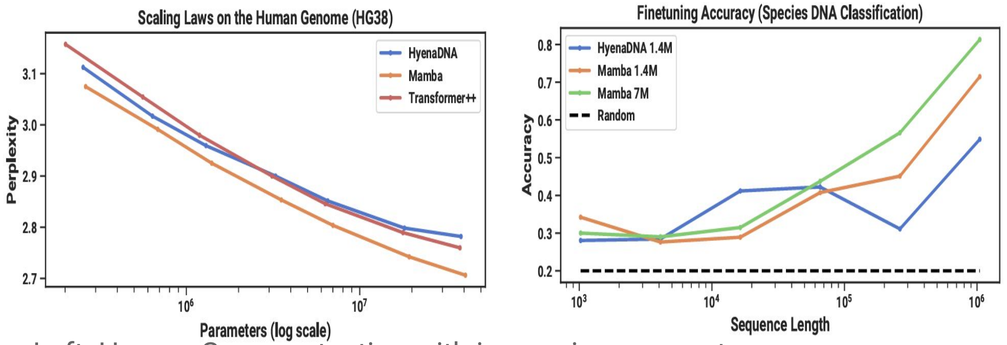
Models trained on sequence length 256 and tested on increasing sequence lengths of 64 and up to 1048576



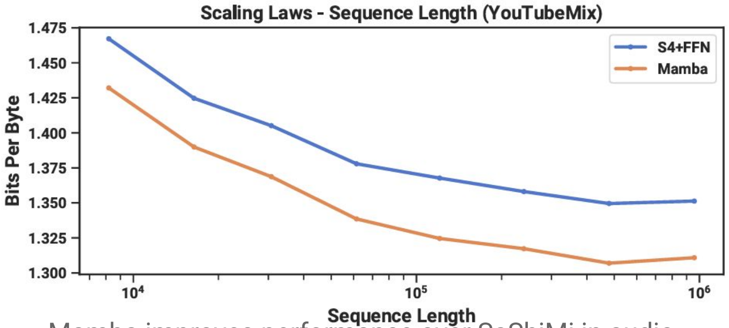
Comparison of model performance specifically in terms of size and complexity to understand how well they scale in handling challenging tasks within The Pile dataset. FLOPS measure a computer's performance

DNA Modeling and Audio Modeling and Generation

- DNA Modeling
  - Investigating Mamba as a foundation for genomics since DNA is a sequence of tokens with a finite vocab
  - Mamba scales well with model size and sequence length
  - Instead of using species with very different DNA, they used different ape species that share up to 99% of their DNA
  - **Mamba Advantage**: Outperforms others in synthetic species classifications (Monkey DNA)
- Audio Modeling and Generation
  - Comparing Mamba to SaShiMi architecture (Neural Network for audio waveform modeling)
  - Achieving better bits per byte in pretraining on piano music
  - **Mamba Advantage**: Does better performance in speech generation



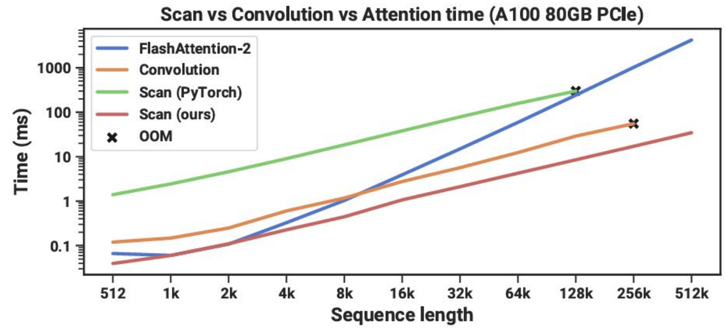
Left: Human Genome testing with increasing parameters
Right: Great Apes DNA Classification with increasing sequence lengths



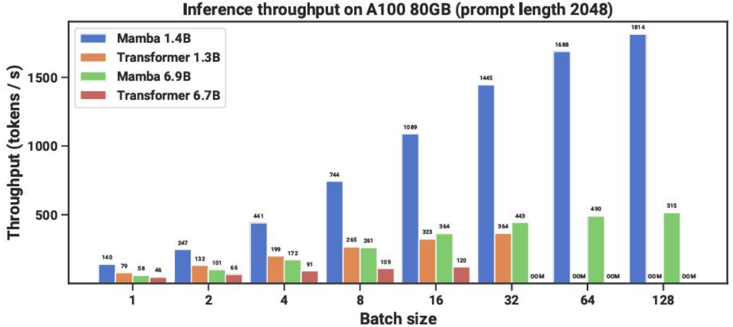
Mamba improves performance over SaShiMi in audio modeling while improving up to million-length sequences

Mamba’s Speed and Memory Benchmarks

- Efficiency in Sequence Scanning
  - State Space Model scan operations (processing elements within a sequence) outperforms attention implementations
  - Faster than FlashAttention-2 and standard scan implementation in PyTorch
  - **Mamba Advantage**: 4-5x more predictions from a model based on input data than a comparable Transformer
- Memory Efficiency
  - Benchmarked for speed and memory consumption
  - Achieves higher inference throughput (efficiency) than a smaller Transformer
  - **Mamba Advantage**: Efficient use of memory which enables higher batch sizes for faster inference (predictions from a model based on input data)



Training; efficient scan is 40x faster than standard implementation



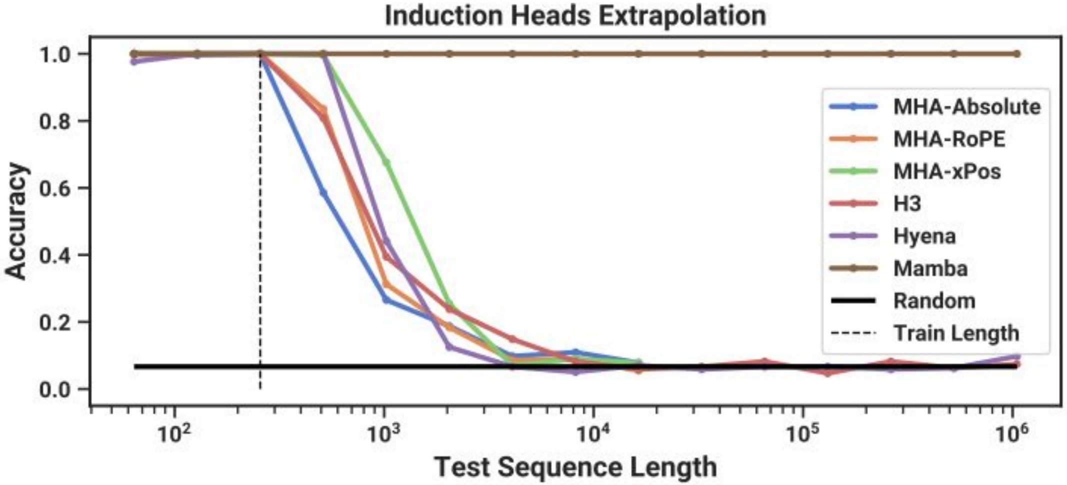
Inference: Mamba achieves 5x higher throughput than Transformers

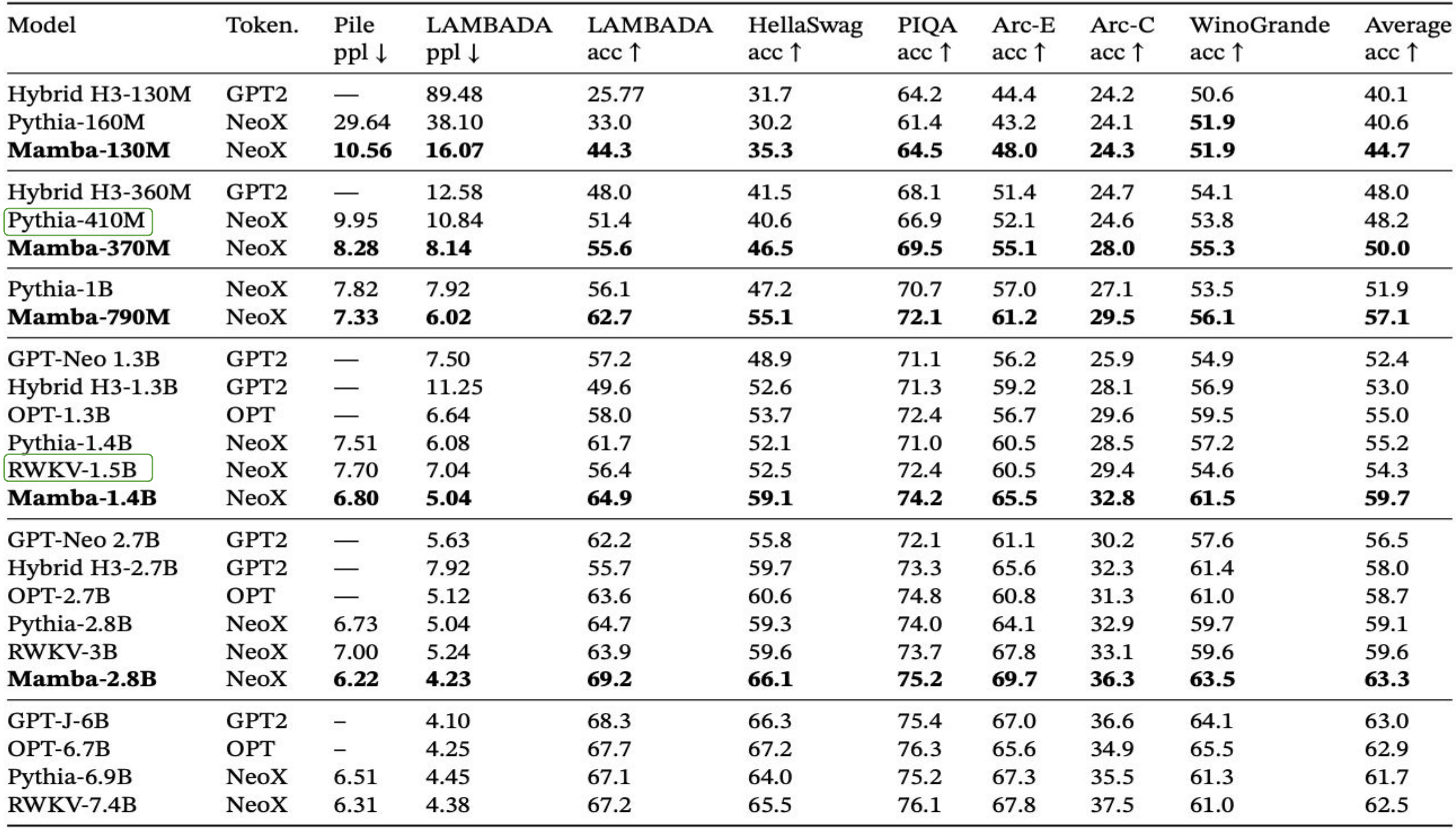
Results

- Mamba outperforms baselines in language modeling.
- Scales better than hyena DNA and Transformer Plus+ in DNA modeling.
- Good performance on audio tasks, with a note on potential limitations in certain cases.

Benchmarks

- Mamba's scan operation implementation is faster than convolution and attention on an A100 GPU.
- Achieves higher inference throughput than Transformers, benefiting from its recurrent nature.

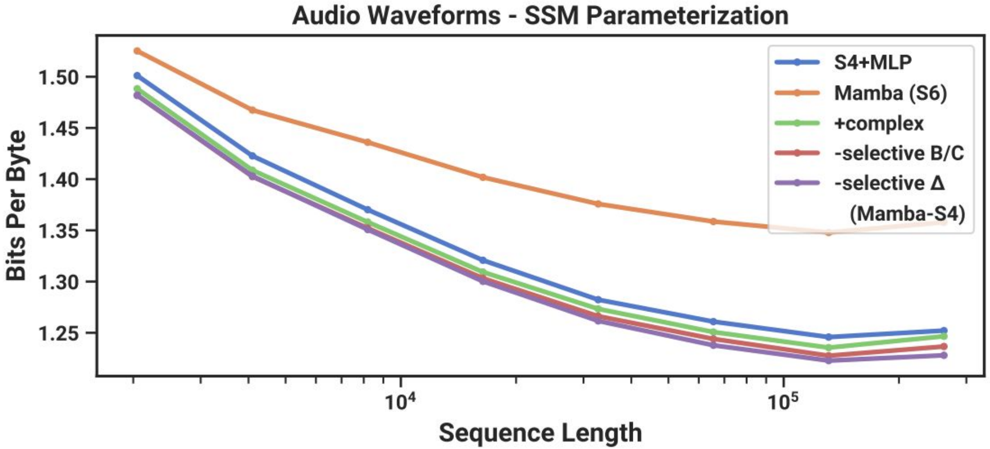




Model	Token.	Pile ppl ↓	LAMBADA ppl ↓	LAMBADA acc ↑	HellaSwag acc ↑	PIQA acc ↑	Arc-E acc ↑	Arc-C acc ↑	WinoGrande acc ↑	Average acc ↑
Hybrid H3-130M	GPT2	—	89.48	25.77	31.7	64.2	44.4	24.2	50.6	40.1
Pythia-160M	NeoX	29.64	38.10	33.0	30.2	61.4	43.2	24.1	**51.9**	40.6
Mamba-130M	NeoX	**10.56**	**16.07**	**44.3**	**35.3**	**64.5**	**48.0**	**24.3**	**51.9**	**44.7**
Hybrid H3-360M	GPT2	—	12.58	48.0	41.5	68.1	51.4	24.7	54.1	48.0
Pythia-410M	NeoX	9.95	10.84	51.4	40.6	66.9	52.1	24.6	53.8	48.2
Mamba-370M	NeoX	**8.28**	**8.14**	**55.6**	**46.5**	**69.5**	**55.1**	**28.0**	**55.3**	**50.0**
Pythia-1B	NeoX	7.82	7.92	56.1	47.2	70.7	57.0	27.1	53.5	51.9
Mamba-790M	NeoX	**7.33**	**6.02**	**62.7**	**55.1**	**72.1**	**61.2**	**29.5**	**56.1**	**57.1**
GPT-Neo 1.3B	GPT2	—	7.50	57.2	48.9	71.1	56.2	25.9	54.9	52.4
Hybrid H3-1.3B	GPT2	—	11.25	49.6	52.6	71.3	59.2	28.1	56.9	53.0
OPT-1.3B	OPT	—	6.64	58.0	53.7	72.4	56.7	29.6	59.5	55.0
Pythia-1.4B	NeoX	7.51	6.08	61.7	52.1	71.0	60.5	28.5	57.2	55.2
RWKV-1.5B	NeoX	7.70	7.04	56.4	52.5	72.4	60.5	29.4	54.6	54.3
Mamba-1.4B	NeoX	**6.80**	**5.04**	**64.9**	**59.1**	**74.2**	**65.5**	**32.8**	**61.5**	**59.7**
GPT-Neo 2.7B	GPT2	—	5.63	62.2	55.8	72.1	61.1	30.2	57.6	56.5
Hybrid H3-2.7B	GPT2	—	7.92	55.7	59.7	73.3	65.6	32.3	61.4	58.0
OPT-2.7B	OPT	—	5.12	63.6	60.6	74.8	60.8	31.3	61.0	58.7
Pythia-2.8B	NeoX	6.73	5.04	64.7	59.3	74.0	64.1	32.9	59.7	59.1
RWKV-3B	NeoX	7.00	5.24	63.9	59.6	73.7	67.8	33.1	59.6	59.6
Mamba-2.8B	NeoX	**6.22**	**4.23**	**69.2**	**66.1**	**75.2**	**69.7**	**36.3**	**63.5**	**63.3**
GPT-J-6B	GPT2	–	4.10	68.3	66.3	75.4	67.0	36.6	64.1	63.0
OPT-6.7B	OPT	–	4.25	67.7	67.2	76.3	65.6	34.9	65.5	62.9
Pythia-6.9B	NeoX	6.51	4.45	67.1	64.0	75.2	67.3	35.5	61.3	61.7
RWKV-7.4B	NeoX	6.31	4.38	67.2	65.5	76.1	67.8	37.5	61.0	62.5

Limitations of Mamba

- Empirical evaluation **limited to small model sizes**, and further assessment is needed at larger sizes.
- Data that is **uniformly sampled and very smooth**, benefits from continuous linear time-invariant (LTI) methods, not Mamba S6
- Scaling SSMs could need more engineering work and modifications.



Conclusion

- Mamba presents an efficient approach to modeling long.

- On a wide range of domains, Mamba **produces state-of-the-art results** when integrated into a simple attention-free architecture, **matching or surpassing** the performance of powerful Transformer models.

- Released code available on GitHub for exploration.

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu*[1] and Tri Dao*[2]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Computer Science, Princeton University
agu@cs.cmu.edu, tri@tridao.me

Foundation models, now powering most [illegible] deep learning, are almost universally based on the Transformer architecture [illegible]. Many subquadratic-time architectures such as linear attention, gated convolu[illegible]odels, and structured state space models (SSMs) have been developed to address Tra[illegible] efficiency on long sequences, but they have not performed as well as attention on i[illegible] language. We identify that a key weakness of such models is their inability to pe[illegible] several improvements. First, simply letting the SSM parameters be fu[illegible] with discrete modalities, allowing the model to selectively propag[illegible]ormation along the sequence length dimension depending on the current token. Second, eve[illegible] change prevents the use of efficient convolutions, we design a hardware-aware parallel algo[illegible]t mode. We integrate these selective SSMs into a simplified end-to-end neural network arc[illegible]out attention or even MLP blocks (Mamba). Mamba enjoys fast inference (5× higher throughp[illegible]ormers) and linear scaling in sequence length, and its performance

State Space Models

- Discretization
- Computation
- Linear Time Invariance (LTI)
- Structure and Dimensions
- General State Space Models
- SSM Architecture

Q&A
Kahoot Time