Effective Methods for Long-Context Language Modeling

Matvey Kairov, Aydar Bulatov
Neural Networks and Deep Learning Lab, MIPT, Dolgoprudny, Russia
kairov.mv@phystech.edu, bulatov@deeppavlov.ai

Abstract

A major limitation for the broader scope of problems solvable by transformers is the quadratic scaling of computational complexity with input size. Recurrent Memory Transformer (RMT) solved this problem efficiently by employing a recurrent memory mechanism on a sequence separated into segments. However, a significant issue faced by RMT is its relatively high computational expenses during fine-tuning due to a convoluted training curriculum employed in the original architecture. This paper aims at solving the inefficiency of RMT by incorporating layer-wise memory updates and mechanisms for memory retrieval.

1 Introduction

Transformer-based models have demonstrated remarkable effectiveness across numerous domains and tasks. However, a significant limitation remains in their ability to process long sequences due to the quadratic computational complexity of self-attention with respect to input length. While several approaches have been proposed to address this challenge, many still struggle with effectively storing and retrieving information over extensive contexts.

The Recurrent Memory Transformer (RMT) (Bulatov et al., 2022) offers a promising solution to this problem by augmenting transformer models with memory tokens that enable linear scaling with sequence length. RMT has demonstrated unprecedented capabilities in retaining information across sequences of up to two million tokens. However, a critical limitation of RMT is its reliance on curriculum learning during training, which necessitates a gradual increase in the number of segments. This requirement significantly increases training complexity and computational demands.

In this work, we propose several enhancements to the RMT architecture aimed at enabling effective training without extensive curriculum learning:

1. We introduce a layer-wise memory update mechanism that allows more granular information flow between transformer layers, enhancing the model's ability to maintain contextual relevance across segments.

2. We develop an improved memory retrieval system that more effectively identifies and accesses relevant information stored in previous segments.

3. We augment the memory mechanism with Rotary Positional Embeddings (RoPE) applied to cached hidden states, preserving the sequential ordering information as content moves across segments.

Our approach aims to maintain the computational efficiency and memory retention capabilities of RMT while simplifying the training process. By eliminating or reducing the need for curriculum learning, our method has the potential to make long-context modeling more accessible and practical for a wider range of applications.

The preliminary experiments suggest that our approach offers a promising direction for improving the training efficiency of recurrent memory-augmented transformer models without sacrificing their ability to process extremely long sequences.

2 Related Work

Our work revolves around the concept of memory in neural architectures. Memory has been a recurrent theme in neural network research, dating back to early works (McCulloch and Pitts, 1943; Kleene, 1951) and signifcantly advancing in the 1990s with the introduction of the Backpropagation Through Time learning algorithm (Werbos, 1990) and Long-Short Term Memory (LSTM) neural architecture (Hochreiter and Schmidhuber, 1997). Contemporary memory-augmented neural networks (MANNs) typically utilize some form

of recurrent external memory separate from the model's parameters. Neural Turing Machines (NTMs) (Graves et al., 2014) and Memory Networks (Weston et al., 2014) are equipped with storage for vector representations accessible through an attention mechanism. Memory Networks were designed to enable reasoning through sequential attention over memory content.

NTMs, followed by Differentiable Neural Computer (Graves et al., 2016) and Sparse DNC (Rae et al., 2016), are implemented as recurrent neural networks capable of writing to memory storage over time. All these models are differentiable and trainable via backpropagation through time (BPTT). Parallel research lines extend recurrent neural networks, such as LSTM, with data structures like stacks, lists, or queues (Joulin and Mikolov, 2015; Grefenstette et al., 2015). MANN architectures with more advanced addressing mechanisms, such as address-content separation and multi-step addressing, have been proposed in (Gulcehre et al., 2016, 2017; Meng and Rumshisky, 2018). The Global Context Layer model (Meng and Rumshisky, 2018) employs address-content separation to address the challenge of training content-based addressing in canonical NTMs. Memory is often combined with Transformers in a recurrent approach. Long inputs are divided into smaller segments, processed sequentially with memory to access information from past segments. Transformer-XL (Dai et al., 2019) preserves previous hidden states for reuse in subsequent segments, while Compressive Transformer (Rae et al., 2019) adds new compressed memory. Ernie-Doc (Ding et al., 2020) enhances contextual information fow by employing same-layer recurrence instead of attending to previous layer outputs of preceding segments. Memformer (Wu et al., 2020) introduces a dedicated memory module to store previous hidden states in summarized representations. Using a similar approach to Memformer, MART (Lei et al., 2020) and Block-Recurrent Transformer (Hutchins et al., 2022) adopt memory update rules analogous to LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014). FeedBack Transformer (Fan et al., 2020) implements full recurrence beyond the segment level and merges low and high layers representations into a memory state. A drawback of most existing recurrent methods is the need for architectural modifcations that complicate their application to various pre-trained models. In contrast, the Recurrent Memory Transformer can be built upon any model that uses a common supported interface.

Some approaches redesign the self-attention mechanism to reduce computational complexity while minimizing input coverage loss. Star-Transformer (Guo et al., 2019), Longformer (Beltagy et al., 2020), GMAT (Gupta and Berant, 2020), Extended Transformer Construction (ETC) (Ainslie et al., 2020), and Big Bird (Zaheer et al., 2020) limit attention distance and employ techniques such as global representations to preserve long-range dependencies. Memory Transformer (Burtsev et al., 2020) introduces memory by extending the unchanged model input with special memory tokens. A common constraint of these methods is that memory requirements grow with input size during both training and inference, inevitably limiting input scaling due to hardware constraints. In contrast, recurrent approaches have constant memory complexity during inference. The longest Longformer, Big Bird, and Long T5 (Guo et al., 2021), LongNet (Ding et al., 2023) models reported in their respective papers have a maximum length of less than 33,000 tokens. CoLT5 (Ainslie et al., 2023) can handle up to 64,000 tokens. LongNet with sparse delated attention can potentially handle up to 1B tokens per batch in distributed setting. However, CoLT5 and LongNet hit the GPU limit and run out of memory with longer inputs. Furthermore, LongNet is unable to scale to millions of tokens on single GPU due to memory constraints (Ding et al., 2023). Memorizing Transformers (Wu et al., 2022) and Unlimiformer (Bertsch et al., 2023) further extend memory through k-NN. Another line of recent related work focuses on models with an alternative to traditional attention mechanism: S4 (Gu et al., 2021), Hyena (Poli et al., 2023), RWKV (Peng et al., 2023), RetNet (Roy et al., 2023). They aim to combine the best of convolutions and recurrence – high parallelism during training and linear scaling with sequence length. As architectures change, these methods require training models from scratch and are unable to reuse plethora of already pre-trained transformer-based models.

3 Recurrent Memory Transformer

Starting from the initial Recurrent Memory Transformer (Bulatov et al., 2022) (RMT), we adapted it for a plug-and-play approach as a wrapper for a range of popular Transformers. This adaptation

augments its backbone with memory, composed of m real-valued trainable vectors (Figure 1). The lengthy input is divided into segments, and memory vectors are prepended to the frst segment embeddings and processed alongside the segment tokens. For encoder-only models like BERT, memory is added only once at the beginning of the segment, unlike (Bulatov et al., 2022), where decoder-only models separate memory into read and write sections. For the time step τ and segment H_τ^0, the recurrent step is performed as follows:

$$\tilde{H}_\tau^0 = \left[H_\tau^{mem} \circ H_\tau^0\right], \quad (1)$$

$$\bar{H}_\tau^N = \text{Transformer}\left(\tilde{H}_\tau^0\right), \quad (2)$$

$$\left[\bar{H}_\tau^{mem} \circ H_\tau^N\right] := \bar{H}_\tau^N, \quad (3)$$

here N is a number of Transformer layers.

After the forward pass, $\bar{H}_\tau^{mem}$ contains updated memory tokens for the segment τ.

Segments of the input sequence are processed sequentially. To enable the recurrent connection, we pass the outputs of the memory tokens from the current segment to the input of the next one:

$$H_{\tau+1}^{mem} := \bar{H}_\tau^{mem}, \tilde{H}_{\tau+1}^0 = \left[H_{\tau+1}^{mem} \circ H_{\tau+1}^0\right].$$

Both memory and recurrence in the RMT are based only on global memory tokens. This allows the backbone Transformer to remain unchanged, making the RMT memory augmentation compatible with any Transformer-based model.

Bulatov et al. (Bulatov et al., 2024) demonstrated that RMT scales linearly for any model size if the segment length is fxed. This is achieved by dividing an input sequence into segments and computing the full attention matrix only within segment boundaries. Larger Transformer models tend to exhibit slower quadratic scaling with respect to sequence length because of computeheavy FFN layers (which scale quadratically with respect to hidden size). However, on extremely long sequences > 32,000, they fall back to quadratic scaling. RMT requires fewer FLOPs than non-recurrent models for sequences with more than one segment (> 512 in this study) and can reduce the number of FLOPs by up to ×295 times. RMT provides a larger relative reduction in FLOPs for smaller models, but in absolute numbers, a ×29 times reduction for OPT-175B models is highly signifcant. Additional experimental comparison of computational effciency can be found in the Appendix.

3.1 Block-Wise Memory and Retrieval

We augment the original RMT (shown in 1) with two memory mechanisms, which can be used separately or at the same time. The first mechanism is block-wise retrieval, which replicates the general idea of RMT memory updates, but performs it on each transformer block for each segment instead of on each segment only. The other mechanism is block-wise memory retrieval (shown in 3).

We also have augmented RMT-R by incorporating Rotary Positional Embeddings (Su et al., 2024) to encode the positions of stored memory tokens during retrieval.

4 Experiments

To evaluate the long-context reasoning capabilities of our model, we employed the BABILong benchmark (Kuratov et al., 2024). We selected BABILong due to its explicit design for assessing models' ability to reason across extended textual contexts, a critical requirement for our model. Unlike existing benchmarks such as LongBench (Bai et al., 2023) and L-Eval (An et al., 2023) that are limited to shorter contexts, BABILong offers scalability to millions of tokens, aligning with the expanded input capacity of contemporary language models. Furthermore, the benchmark's "needle-in-a-haystack" approach, utilizing natural language from the PG19 corpus (Rae et al., 2020), presents a more realistic and challenging evaluation compared to synthetic tasks that can be easily overfitted (Kuratov et al., 2024).

Specifically, we focused on tasks QA1 and QA3 of the BABILong suite. These tasks involve question answering based on a single or multiple facts, providing a thorough assessment of the model's comprehention of provided long contexts. As demonstrated by Kuratov et al. (2024), performance on BABILong can reveal limitations in context utilization, with many models struggling to effectively leverage more than a small fraction of the input. By focusing on the QA1 and QA3 tasks, we aim to provide a detailed analysis on these limitations within our long-context model.

The experiments on QA1 4 have shown the models' ability to converge on both training and validation, and the model that converged the fastest was ARMT (Rodkin et al., 2024).

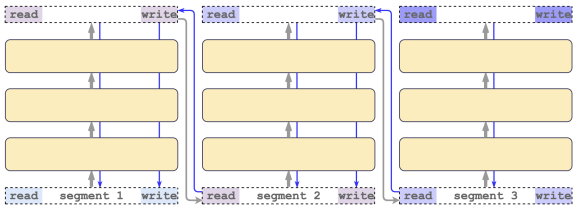

Figure 1: Recurrent Memory Transformer.

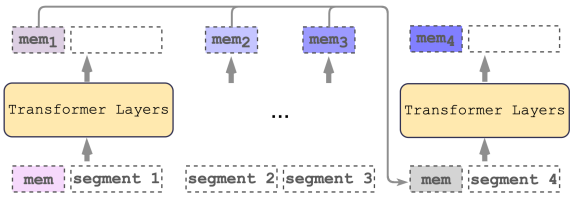

Figure 2: Recurrent Memory Transformer with segment-wise memory retrieval.

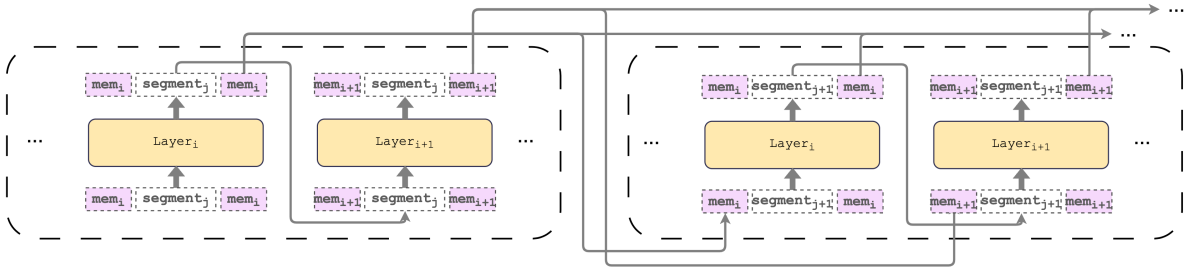

Figure 3: Recurrent Memory Transformer with block-wise memory retrieval.

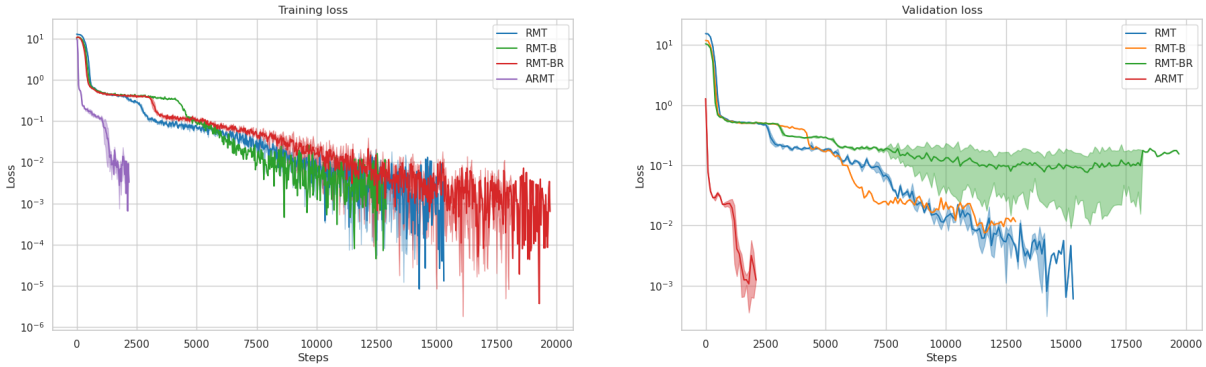

Figure 4: Single-pass training (left) and validation (right) loss on BABILong QA1.

5 Results

6 Conclusion

We have proposed new augmentations to the architecture of RMT – RMT-R with incorporated Rotary Positional Embeddings, RMT-B with Transformer block-wise memory updates, and RMT-BR which combines the principles behind both those architectures. We conducted single-pass training experiments on two tasks from BABILong – QA1 and QA3. We observed the benefits of memory retrieval in RMT-R, however, further augmentations have failed to produce better models.

Limitations

While the proposed updates to the RMT architecture show promising results on long-context memorization tasks, there are several key limitations to our work that require further research. First, our experiments involve fine-tuning the proposed architectures on the BABILong benchmark, which, while representative of the model's ability of memory retrieval, may not fully represent its potential in the solving of more general language modeling tasks. Testing our approach on a broad selection of datasets could offer important insights into the model's ability for generalization.

Another limitation observed in our experiments is the model's tendency to overfit during single-pass training. While this may be partially caused by the relatively small size of BABILong, further investigation must be conducted in order to determine the exact reason for this difference from the models trained using the curriculum-based approach.

Despite the incorporation of memory aggregation in RMT-R and an increased amount of memory tokens in RMT-B, these models still require

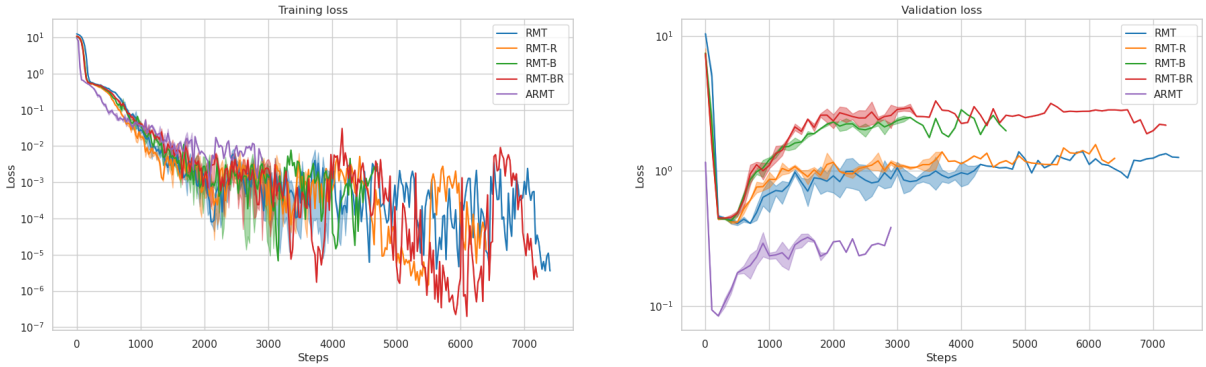

Figure 5: Single-pass training (left) and validation (right) loss on BABILong QA3.

a significant amount of computational resources. Additionally, while memory retrieval has shown its potential to improve the model's memorization ability, the size of stored memory states grows linearly with the length of the input sequence, and the effects of storing long-term memory states remains unexplored.

Finally, while we have shown the potential improvements provided by the proposed modifications to RMT, further research is required to optimize the model for a lower chance of overfitting along with improving its scalability in case of the storing of memory states.

Acknowledgments

References

Joshua Ainslie, Tao Lei, Michiel de Jong, Santiago Ontañón, Siddhartha Brahma, Yury Zemlyanskiy, David Uthus, Mandy Guo, James Lee-Thorp, Yi Tay, and 1 others. 2023. Colt5: Faster long-range transformers with conditional computation. *arXiv preprint arXiv:2303.09752*.

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers. *arXiv preprint arXiv:2004.08483*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36:35522–35543.

Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, and Mikhail Burtsev. 2024. Beyond attention: breaking the limits of transformer context length with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17700–17708.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091.

Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. 2020. Memory transformer. *arXiv preprint arXiv:2006.11527*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

Siyu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-doc: A retrospective long-document modeling transformer. *arXiv preprint arXiv:2012.15688*.

Angela Fan, Thibaut Lavril, Edouard Grave, Armand Joulin, and Sainbayar Sukhbaatar. 2020. Addressing some limitations of transformers with feedback memory. *arXiv preprint arXiv:2002.09402*.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, and 1 others. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. 2015. Learning to transduce with unbounded memory. *Advances in neural information processing systems*, 28.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Caglar Gulcehre, Sarath Chandar, and Yoshua Bengio. 2017. Memory augmented neural networks with wormhole connections. *arXiv preprint arXiv:1701.08718*.

Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. 2016. Dynamic neural turing machine with soft and hard addressing schemes. *arXiv preprint arXiv:1607.00036*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. *arXiv preprint arXiv:1902.09113*.

Ankit Gupta and Jonathan Berant. 2020. Gmat: Global memory augmentation for transformers. *arXiv preprint arXiv:2006.03274*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

DeLesley Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, and Behnam Neyshabur. 2022. Block-recurrent transformers. *Advances in neural information processing systems*, 35:33248–33261.

Armand Joulin and Tomas Mikolov. 2015. Inferring algorithmic patterns with stack-augmented recurrent nets. *Advances in neural information processing systems*, 28.

Stephen Cole Kleene. 1951. Representationof events in nerve nets and finite automata. *CE Shannon and J. McCarthy*.

Yury Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*.

Warren S McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.

Yuanliang Meng and Anna Rumshisky. 2018. Context-aware neural model for temporal information extraction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 527–536.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, and 1 others. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR.

Jack Rae, Jonathan J Hunt, Ivo Danihelka, Timothy Harley, Andrew W Senior, Gregory Wayne, Alex Graves, and Timothy Lillicrap. 2016. Scaling memory-augmented neural networks with sparse reads and writes. *Advances in Neural Information Processing Systems*, 29.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.

Ivan Rodkin, Yuri Kuratov, Aydar Bulatov, and Mikhail Burtsev. 2024. Associative recurrent memory transformer. *arXiv preprint arXiv:2407.04841*.

Amit Roy, Riasat Abdullah, Fahim Ahmed, Shahriar Mashfi, Sazid Hayat Khan, and Dewan Ziaul Karim. 2023. Retnet: retinal disease detection using convolutional neural network. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Paul J Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Qingyang Wu, Zhenzhong Lan, Kun Qian, Jing Gu, Alborz Geramifard, and Zhou Yu. 2020. Memformer: A memory-augmented transformer for sequence modeling. *arXiv preprint arXiv:2010.06891*.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. *arXiv preprint arXiv:2203.08913*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang,

Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.