Theoretical Expressivity of (A)RMT: Bridging Transformers and RNNs

Sergei Kudriashov

AIRI Summer School 2025

Abstract

This paper explores the theoretical expressivity of Recurrent Memory Transformers (RMT) [1] and their associative variant (ARMT) [3], positioning them as architectures that bridge the gap between Transformers and Recurrent Neural Networks (RNNs). We analyze their Turing-completeness, computational bounds, and limitations, particularly in handling algorithmic tasks on large graphs. A hierarchical RMT modification is proposed to address these limitations, offering logarithmic depth and improved memory efficiency. Our work provides theoretical guarantees and outlines future directions for scaling transformers in sequence length.

1 Introduction

The expressivity of neural architectures is a cornerstone for understanding their applicability to algorithmic tasks[**?**]. While Transformers and RNNs have been proven Turing-complete under varying conditions, their practical limitations in memory and depth remain underexplored. This paper investigates the theoretical expressivity of RMT and ARMT, architectures that combine the strengths of Transformers and RNNs. We highlight their ability to express arbitrary computational circuits, discuss their limitations, and propose a hierarchical modification to enhance their efficiency. Our contributions include: (1) a unified analysis of RMT's reducibility to Transformers and RNNs, (2) worst-case bounds on memory capacity, and (3) a path forward for empirical validation.

2 Related Work

The computational power of neural networks has been extensively studied, from Siegelmann and Sontag's [5] proof of RNNs' Turing-completeness to recent results on bounded-precision Transformers [2, 6]. The Universal Approximation Theorem (UAT) provides guarantees for continuous functions, but modern applications demand discrete, algorithmic reasoning. Recent work by Sanford et al. [4] and Yehudai et al. [7] explores Transformers' capabilities in graph

algorithms, emphasizing depth-width tradeoffs. Our work builds on these foundations, focusing on RMT's unique position between Transformers and RNNs [1].

3 Method

RMT is reducible to both Transformers and RNNs, inheriting their theoretical guarantees. For Transformers, this is trivial; for RNNs, we assume hard attention with a specified pattern. The key limitation of RMT lies in its effective depth, bounded by the segment length L, which restricts expressivity for long sequences. To address this, we propose a hierarchical RMT with logarithmic depth, where segments run in parallel and neighboring memory outputs serve as inputs for shared-weight networks. This modification maintains expressivity while reducing computational complexity and memory bottlenecks.

4 Experimental Setup

Future work will evaluate the proposed hierarchical RMT on synthetic graph tasks of increasing dimensionality. The experiments will compare default (A)RMT with the hierarchical variant, testing generalization and scalability. Metrics will include memory efficiency, depth requirements, and performance on tasks with sublinear memory constraints.

5 Results

Theoretical analysis shows that hierarchical RMT is capable of achieving logarithmic depth without sacrificing expressivity, addressing the linear memory growth issue of vanilla RMT. Empirical results (to be obtained) will validate these claims, demonstrating improved performance on long-sequence and graph-based tasks.

6 Conclusion

RMT and ARMT offer a promising avenue for unifying the strengths of Transformers and RNNs. Our hierarchical modification mitigates their limitations, enabling scalable and efficient computation. Future work will derive explicit bounds on memory capacity and validate the approach on algorithmic tasks, paving the way for broader applications in discrete domains.

References

[1] Aydar Bulatov, Yuri Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave,

K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[2] Stephen Chung and Hava Sieglemann. Turing completeness of bounded-precision recurrent neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[3] Ivan Rodkin, Yuri Kuratov, Aydar Bulatov, and Mikhail Burtsev. Associative recurrent memory transformer, 2025.

[4] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth, 2024.

[5] Hava T. Siegelmann and Eduardo D. Sontag. On the computational power of neural nets. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 440–449, New York, NY, USA, 1992. Association for Computing Machinery.

[6] Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *arXiv preprint arXiv: 2107.13163*, 2021.

[7] Gilad Yehudai, Clayton Sanford, Maya Bechler-Speicher, Orr Fischer, Ran Gilad-Bachrach, and Amir Globerson. Depth-width tradeoffs in algorithmic reasoning of graph tasks with transformers, 2025.