ICLR

RMAAT: Astrocyte-Inspired Memory Compression and Replay for Efficient Long-Context Transformers

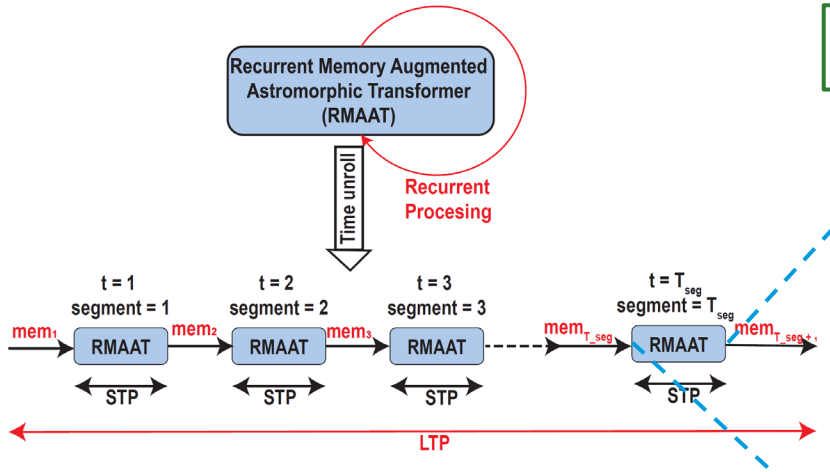
Md Zesun Ahmed Mia, Malyaban Bal and Abhronil Sengupta

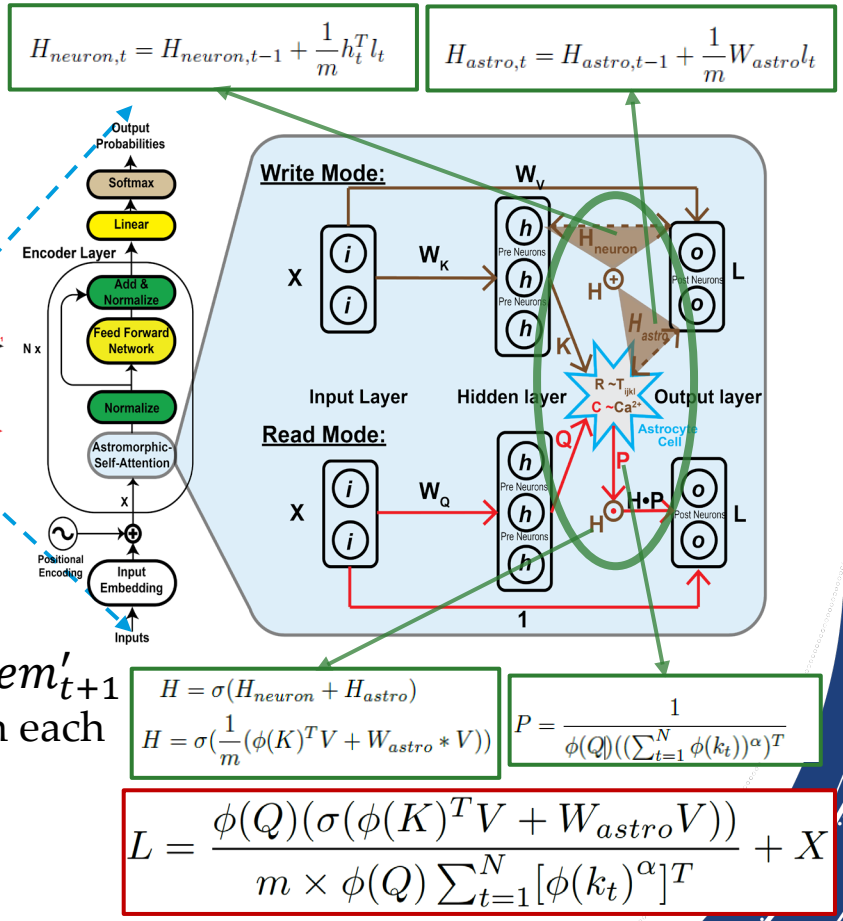
PennState

Why RMAAT? From Fixed Windows to Recurrent Long Context

ICLR

- **Problem:** Regular efficient transformers process a fixed-length segment — no mechanism to carry information across segments in recurrent processing
- **RMAAT** adds recurrent long-context processing:
  - Split input into segments of length N_{seq}
  - Prepend M persistent memory tokens per segment
  - Memory tokens mem_t flow segment-to-segment, carrying compressed context forward





Recurrence equation:

$$mem'_{t+1} \leftarrow Model(x_t, mem_t)$$

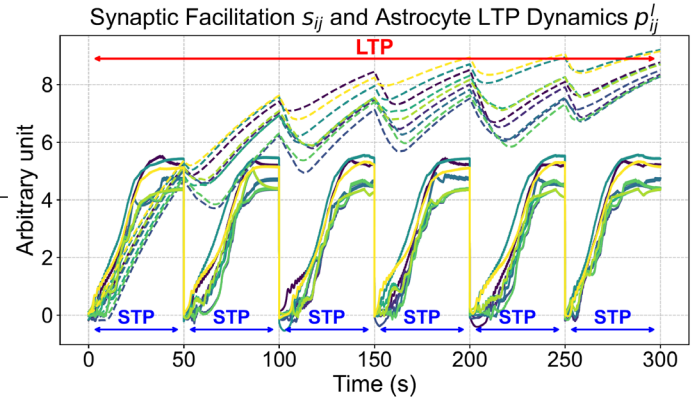
$$mem_{t+1} \leftarrow RetentionFactor(t, T) \times mem'_{t+1}$$

- Same astromorphic STP attention within each segment
- LTP dynamics govern memory across segments
- O(N) complexity per segment

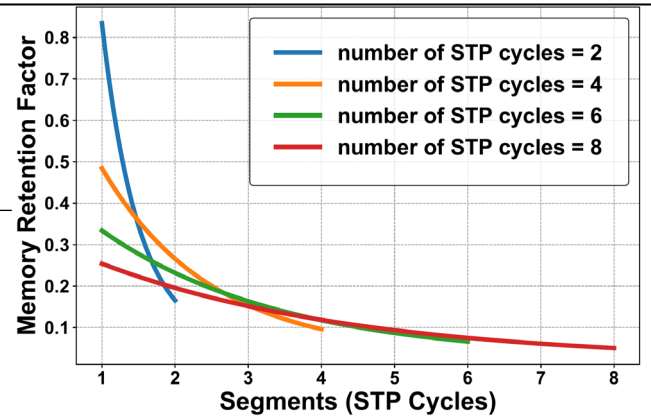
RetentionFactor(t, T): a principled compression schedule derived from astrocyte LTP simulation — explained next slide

PennState

Memory Retention Factor: LTP-Derived Compression



Solid = fast STP dynamics (s_{ij});
Dashed = LTP process (p_{ij}^l) gradually integrating and saturating across STP cycles.



Retention factor decreases per segment as total sequence length increases — adaptive, bio-inspired compression.

$$\tau_p^l \frac{dp_{ij}^l}{dx} \propto -\gamma^l p_{ij}^l + s_{ij}$$

p_{ij}^l: long-term astrocyte process — integrates synaptic activity (s_{ij}) over slow timescales

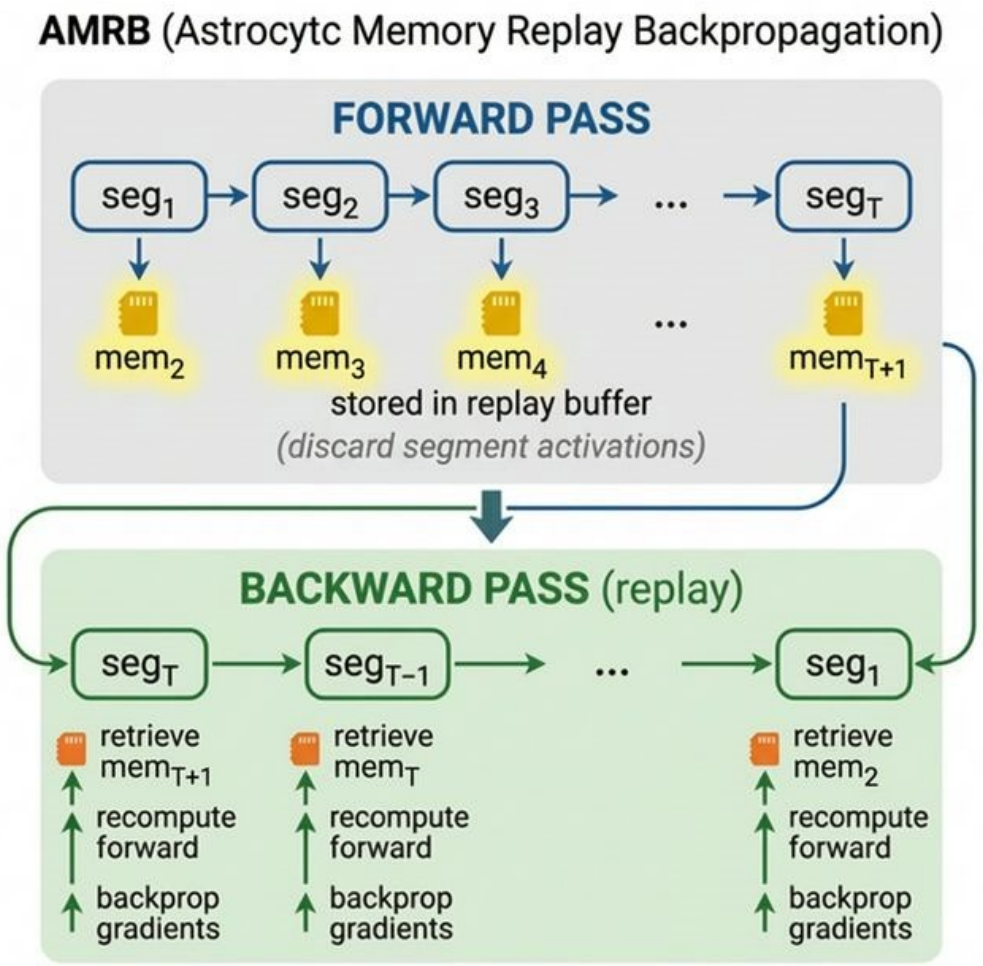
- Simulate this LTP dynamics over multiple STP cycles
- Key behavior: gradual integration → accumulation → saturation (finite Ca^{2+} resources)
- Normalize total capacity to 1 unit
- Measure increase in p^l per segment → fraction of total

$$RetentionFactor(t,T) = \frac{\Delta p_t^l}{\sum_{i=1}^{T} p_i^l}$$

(fraction of LTP capacity gained at segment t)

- Principled compression schedule —derived from neuroscience simulation, not learned
- Ablation: removing retention factor drops Retrieval accuracy 83.2% → 80.5%

AMRB: Memory-Efficient Training via Astrocytic Replay



AI generated figure

- Standard BPTT: stores all activations for entire sequence — memory-prohibitive
- AMRB forward pass:
  - Process segments sequentially; store only memory token states($mem_1, mem_2, \ldots\ldots, mem_{T+1}$) → discard all segment activations
- AMRB backward pass: For each segment t (reverse order):
  1. Retrieve mem_t from buffer
  2. Recompute forward pass for segment t only
  3. Backpropagate gradients

Critical synergy: compression + AMRB are synergistic —bio-inspired retention factor makes AMRB effective (ablation: removing compression degrades accuracy significantly)

RMAAT: Competitive Accuracy, Lower Memory, Higher Throughput

ICLR

Table 1: Accuracy and Memory Comparison on Long Range Arena (LRA) Benchmark Tasks.

Model	ListOps (2K)		Text (4K)		Retrieval (8K)		Image (1K)		Pathfinder (1K)		Average
	Acc.(S.)	Mem.*	Acc.(S.)	Mem.*	Acc.(S.)	Mem.*	Acc.(S.)	Mem.*	Acc.(S.)	Mem.*	Acc.
Transformer (Vaswani et al., 2017)	36.4(1)	4.7	64.3(1)	6.7	57.5(1)	5.2	42.4(1)	7.8	71.4(1)	5.4	54.4
Sparse Trans.[a](Child et al., 2019)	17.1(1)	–	63.6(1)	–	59.6(1)	–	44.2(1)	–	71.7(1)	–	51.2
Longformer[a](Beltagy et al., 2020)	35.6(1)	–	62.9(1)	–	56.9(1)	–	42.2(1)	–	69.7(1)	–	53.5
Linformer[a](Wang et al., 2020)	35.7(1)	–	53.9(1)	–	52.3(1)	–	38.6(1)	–	76.3(1)	–	51.4
Reformer[a](Kitaev et al., 2020)	37.3(1)	–	56.1(1)	–	53.4(1)	–	38.1(1)	–	68.5(1)	–	50.7
BigBird[a](Zaheer et al., 2020)	36.1(1)	–	64.0(1)	–	59.3(1)	–	40.8(1)	–	74.9(1)	–	55.0
LT (Katharopoulos et al., 2020)	16.1(1)	4.7	65.9(1)	5.7	53.1(1)	3.9	42.3(1)	6.2	75.3(1)	6.2	50.5
Performer[a](Choromanski et al., 2020)	18.0(1)	–	65.4(1)	–	53.8(1)	–	42.8(1)	–	77.1(1)	–	51.4
FNet[c](Lee-Thorp et al., 2021)	35.3(1)	–	65.1(1)	–	59.6(1)	–	38.7(1)	–	77.8(1)	–	55.3
Nyströmformer[c](Xiong et al., 2021)	37.2(1)	–	65.5(1)	–	79.6(1)	–	41.6(1)	–	70.9(1)	–	59.0
Luna-256[c](Ma et al., 2021)	37.3(1)	–	64.6(1)	–	79.3(1)	–	47.4(1)	–	77.7(1)	–	61.3
AT (Mia et al., 2025)	18.1(1)	4.7	61.5(1)	5.8	77.3(1)	4.1	47.3(1)	6.2	77.9(1)	6.3	56.4
RMT (Bulatov et al., 2022)	37.4(8)[b]	20.4	65.0(8)	24	79.3(16)	18.3	54.6(2)	22.7	81.5(4)	12.7	63.6
RLT (Kozachkov et al., 2023)	18.4(8)[b]	14.4	64.8(8)	22.6	78.4(16)	12.1	55.0(2)	21.6	74.9(4)	13.6	58.3
RMAAT (Ours)	**38.9(8)**[b]	**5.2**	**65.9(8)**	**5.1**	**83.2(16)**	**3.4**	**64.8(2)**	**5.3**	**87.1(4)**	**4.7**	**68.0**

Table 2: Detailed Throughput/Speed Comparison on Long Range Arena (LRA) Tasks.

Model	ListOps	Text	Retrieval	Image	Pathfinder
Transformer (Vaswani et al., 2017)	1×	1×	1×	1×	1×
LT (Katharopoulos et al., 2020)	1.24×	1.01×	1.03×	1.03×	1.03×
AT (Mia et al., 2025)	1.26×	1.26×	1.05×	1.08×	1.03×
RMT (Bulatov et al., 2022)	1×	1×	1×	1×	1×
RLT (Kozachkov et al., 2023)	1.05×	1.13×	1.37×	1.21×	0.95×
RMAAT (Ours)	**1.5×**	**1.5×**	**1.73×**	**1.3×**	**0.95×**

LRA benchmark | RMAAT achieves high accuracy + lowest memory + up to 1.73× throughput

PennState

Thank You

Questions?