# Recurrent Action Transformer with Memory


Egor Cherepanov[1]*, Alexey Staroverov[1]*, Dmitry Yudin[1,2], Alexey K. Kovalev[1,2] and Aleksandr I. Panov[1,2]

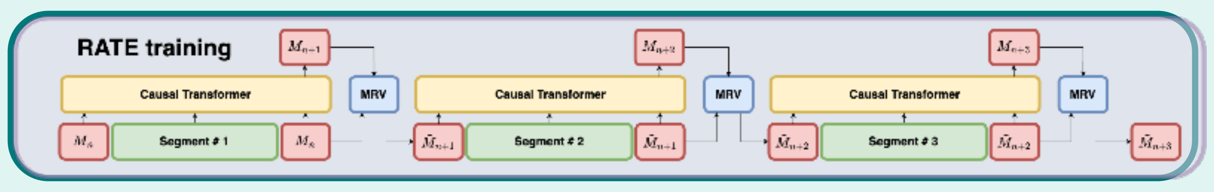
[1]AIRI, [2]MIPT

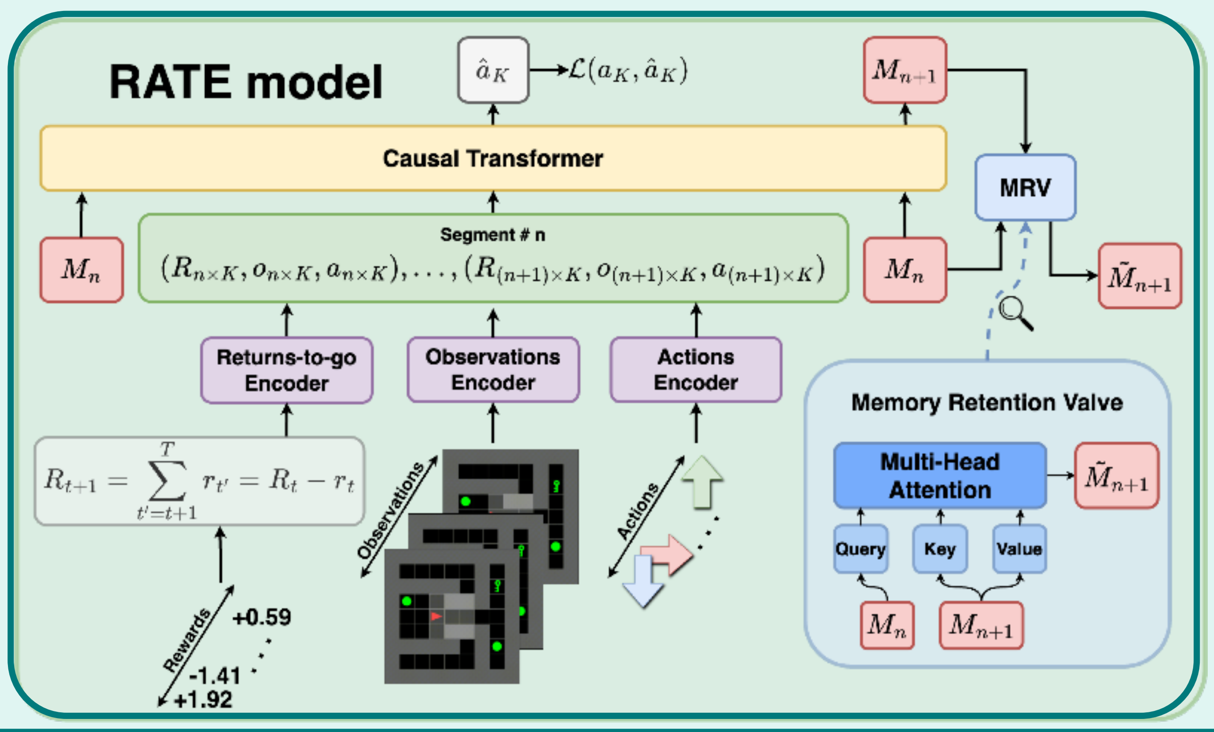
Contact me: cherepanovegor2018@gmail.com

* – equal contribution


## 1. Introduction

- Most real-world problems require an agent to have memory because it has to deal with **partial observability, where full information about the system is not available** at the time of decision making
- Transformers perform well in Reinforcement Learning domain, but can only effectively solve **memory** and **credit assignment** tasks if the entire trajectory **fits within the model context**,
- You **can't increase the context infinitely** because of quadratic attention complexity - important information may "fall out" of context
- **Memory mechanisms** offer a **promising solution** to consider past information **without increasing context size**
- We propose the **Recurrent Action Transformer with Memory (RATE)**, a model that uses memory mechanism: **memory embeddings** and **Memory Retention Valve (MRV)** for **Offline RL**




## 2. Method

- We use methodology outlined in the **Decision Transformer (DT)** paper [1] to represent the trajectory: $\tau = [(R_0, o_0, a_0), \dots, (R_t, o_t, a_t), \dots, (R_T, o_T, a_T)]$
- In RATE we utilized both recurrently trained **memory embeddings** [2] and the preservation of previous **hidden states** [3]. For training of RATE memory embeddings $M$, we split trajectories into $N$ segments of length $K$. Thus, RATE processes sequences $N$ times shorter than DT, but still sees the same trajectory information – **effective context length** $K_{eff}^{RATE} = N \times K = K^{DT}$
- To control the process of forgetting information, the **Memory Retention Valve** processes memory embeddings after each processed segment.

**Algorithm 1** RATE

**Require:** $R \in \mathbb{R}^T, o \in \mathbb{R}^{d_o \times T}, a \in \mathbb{R}^T$

1: $\tilde{R} \leftarrow \texttt{Encoder}_R(R)$
   $\tilde{o} \leftarrow \texttt{Encoder}_o(o)$
   $\tilde{a} \leftarrow \texttt{Encoder}_a(a)$
2: $\tau_{0:T-1} \leftarrow \{(\tilde{R}_t, \tilde{o}_t, \tilde{a}_t)\}_{t=0}^{T-1}$
3: $M_n \leftarrow M_0 \sim \mathcal{N}(0, 1)$
4: **for** $n$ in $[0, T//K - 1]$ **do**
5: $S_n \leftarrow \tau_{nK:(n+1)K}$
6: $\tilde{S}_n \leftarrow \texttt{concat}(M_n, S_n, M_n)$
7: $\hat{a}_n, M_{n+1} \leftarrow \texttt{Transformer}(\tilde{S}_n)$
8: $M_{n+1} \leftarrow \texttt{MRV}(M_n, M_{n+1})$
   **Output:** $\hat{a}_n \rightarrow \mathcal{L}(a_n, \hat{a}_n)$, $M_{n+1}$
9: **end for**

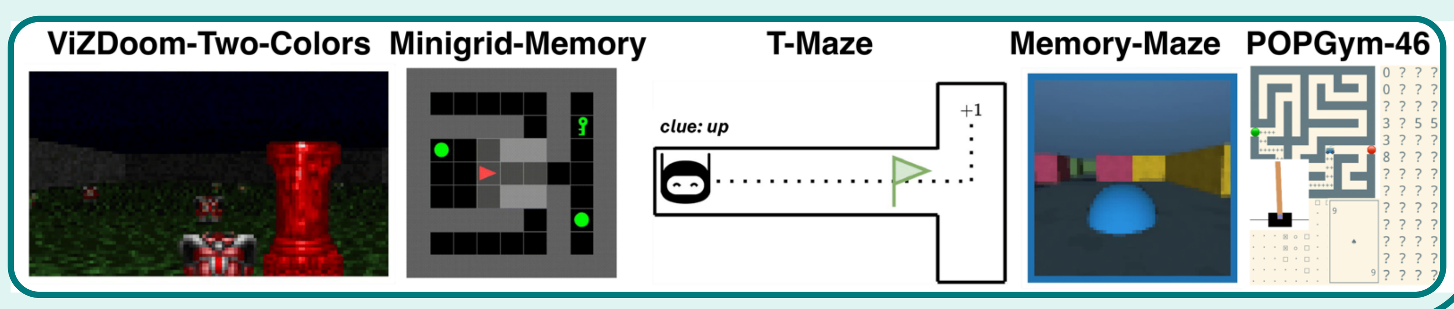
**Algorithm 2** Memory Retention Valve

**Require:** $M_n, M_{n+1} \in \mathbb{R}^{m \times d}$

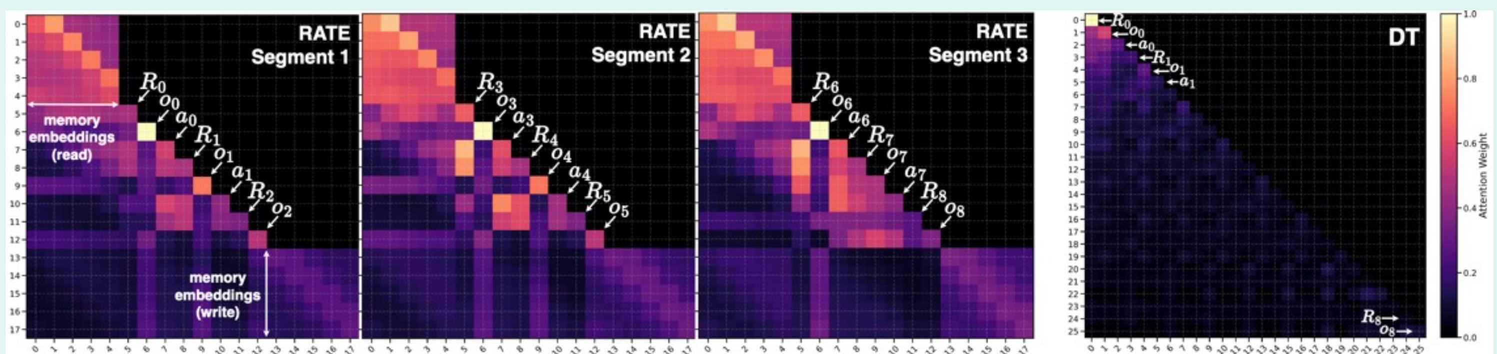
1: $\mathbf{Q}_h \leftarrow M_n \mathbf{W}_Q^{h\,T}$
2: $\mathbf{K}_h \leftarrow M_{n+1} \mathbf{W}_K^{h\,T}$
3: $\mathbf{V}_h \leftarrow M_{n+1} \mathbf{W}_V^{h\,T}$
4: $M_{n+1}^h \leftarrow \texttt{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d}}\right) \mathbf{V}_h$
5: $M_{n+1} \leftarrow \texttt{concat}(M_{n+1}^0, \dots, M_{n+1}^h)$
6: $M_{n+1} \leftarrow M_{n+1} \mathbf{W}_M^T$
**Output:** $M_{n+1}$

## 3. Environments

We designed experiments to demonstrate the success of our RATE model in **memory-intensive environments** (ViZDoom-Two-Colors, T-Maze, Minigrid-Memory, Memory Maze, 48 POPGym tasks) and to proof the versatility of the proposed model on **classic benchmarks** (Atari games and MuJoCo control tasks).



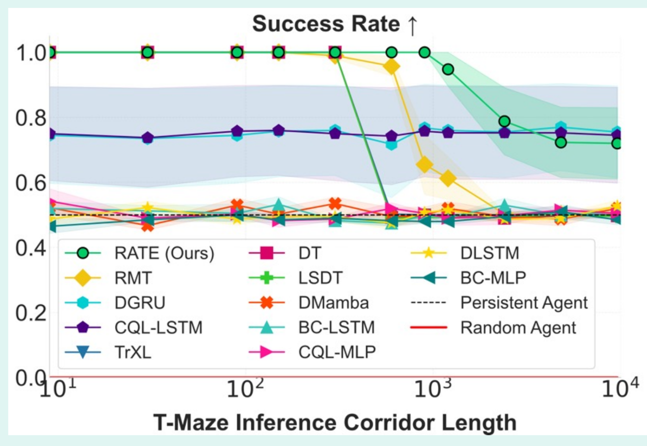

## 4. Visualization of attention maps



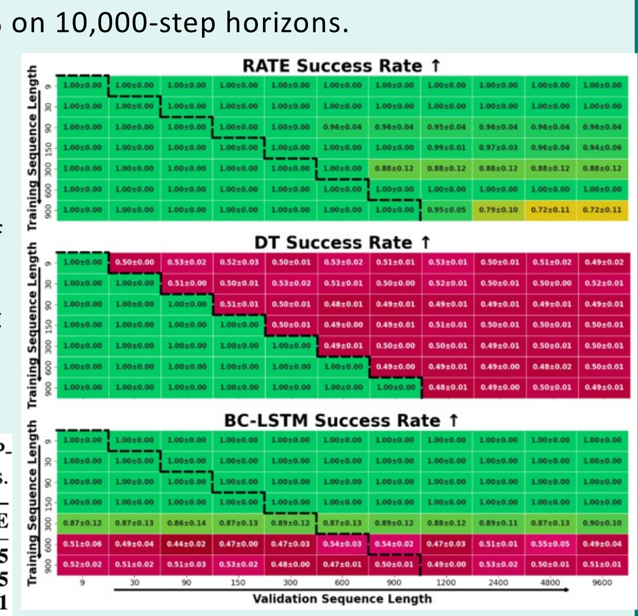

- Attention maps of RATE and DT on the T-Maze task with corridor length T = 8. DT is trained on full 8-step trajectories, while RATE processes the sequence in three segments of length 3 recurrently, passing information between segments through memory embeddings.

## 5. Experiments

- In **ViZDoom-Two-Colors**, the agent must remember the color of a pillar (red or green) for the first 45 steps, and then collect items of the same color as the pillar to survive and get a reward
- RATE demonstrated the best and most balanced results compared to other memory-intensive baselines, demonstrating its ability to retain important information in memory for a long time (up to 2100 steps)


- In the **T-Maze** task, the agent must remember a cue (up or down) at the beginning of a T-shaped corridor, and then it must turn at the junction at the end of the corridor according to this clue
- RATE, trained on 90-step corridors with a 30-step context, demonstrated the longest memory retention, showing an SR of 100% on 1,000-step horizons and an SR of 75% on 10,000-step horizons.




- In T-Maze experiments on **interpolation** to corridors of shorter length than the agent saw during training, and on **extrapolation** to corridors of longer length than the agent saw during training, RATE demonstrates the ability to exploit the best of **both worlds: recurrent and transformer**


Figure 7: Minigrid-Memory generalization task.

Table 2: Aggregated average returns on 48 POPGym tasks, split into memory and reactive subsets.

| Tasks | Rand. | BC-MLP | DT | BC-LSTM | RATE |
|---|---|---|---|---|---|
| All (48) | -12.2 | -6.8 | 5.8 | 9.0 | **9.5** |
| Memory (33) | -14.6 | -11.9 | -3.5 | -0.2 | **0.5** |
| Reactive (15) | 2.3 | 5.1 | **9.3** | **9.1** | **9.1** |

- On the **Minigrid-Memory** and **POPGym-48 tasks**, RATE also shows the best results compared to other baselines, which demonstrates the high generalizability of the agent to different memory-intensive tasks.

Table 3: Normalized scores on MuJoCo tasks from the D4RL benchmark (Fu et al., 2021). Although RATE is designed for memory-intensive environments, it **performs competitively** – and often surpasses – methods tailored for standard MDP control. Top-1 and Top-2 results are highlighted.

| Dataset | Environment | CQL | DT | TAP | TT | DMamba | MambaDM | RATE (ours) |
|---|---|---|---|---|---|---|---|---|
| ME | HalfCheetah | **91.6** | 86.8±1.3 | **91.8±0.8** | **95.0±0.2** | **91.9±0.6** | 86.5±1.2 | 87.4±0.1 |
| ME | Hopper | 105.4 | **107.6±1.8** | 105.5±1.7 | **110.0±2.7** | **111.1±0.3** | **110.5±0.3** | **112.5±0.2** |
| ME | Walker2d | **108.8** | **108.1±0.2** | **107.4±0.9** | **101.9±6.8** | **108.3±0.5** | **108.8±0.1** | **108.7±0.5** |
| M | HalfCheetah | 44.4 | 42.6±0.1 | **45.0±0.1** | **46.9±0.4** | 42.8±0.1 | 42.8±0.1 | 43.5±0.3 |
| M | Hopper | 58.0 | **67.6±1.0** | 63.4±1.4 | 61.1±3.6 | **83.5±12.5** | **85.7±7.8** | **77.4±1.4** |
| M | Walker2d | 72.5 | 74.0±1.4 | 64.9±2.1 | **79.0±2.8** | **78.2±0.6** | **78.2±0.6** | **80.7±0.7** |
| MR | HalfCheetah | **45.5** | 36.6±0.8 | **40.8±0.6** | **41.9±2.5** | **39.6±0.1** | **39.1±0.1** | **39.0±0.6** |
| MR | Hopper | **95.0** | **82.7±7.0** | **87.3±2.3** | **91.5±3.6** | **82.6±4.6** | **86.1±2.5** | **83.7±8.2** |
| MR | Walker2d | **77.2** | 66.6±3.0 | 66.8±3.1 | **82.6±6.9** | 70.9±4.3 | **73.4±2.6** | **73.7±1.4** |
| | **Average** | **77.6** | 74.7 | 74.8 | **78.9** | **78.8** | **79.0** | **78.5** |

Table 4: Raw scores on Atari games. RATE outperforms DT in 3 out of 4 environments.

| Environment | CQL | BC | DT | DMamba | MambaDM | RATE (Ours) |
|---|---|---|---|---|---|---|
| Breakout | 62.5 | 42.8 | 76.9±27.3 | 70.6±9.3 | **106.9±5.8** | **111.0±2.9** |
| Qbert | **14013.2** | 2862.0 | 2215.8±1523.7 | 5786.0±1295.2 | 10052.5±1116.5 | **12486.9±280.4** |
| SeaQuest | 782.2 | **992.1** | **1129.3±189.0** | **992.1±57.7** | **1286.0±42.0** | **1037.9±53.7** |
| Pong | **18.8** | 6.4 | **17.1±2.9** | 1.6±15.3 | **18.4±0.8** | **18.8±0.3** |

- On **MuJoCo** classic control tasks, RATE performs better or no worse than SOTA state-space models agents on 7 out of 9 tasks
- On classic **Atari** game tasks, RATE shows the best results on 2 problems and comparable results on two more problems
- Together, these results show that RATE performance does not degrade when running on tasks without memory, demonstrating the versatility of the proposed model

## 5. Conclusions

- RATE integrates learnable memory embeddings, recurrent hidden-state caching, and a Memory Retention Valve (MRV) into a single architecture, enabling stable long-horizon memory-retention in partially observable and sparse-reward environments
- RATE maintains near-perfect success rates on extrapolation tasks (e.g., T-Maze with inference up to 9.6k steps), where all other transformers struggle
- Despite being designed for POMDPs, RATE matches or exceeds the performance of strong MDP-oriented baselines

Paper

Website

## References


[1] Chen L. et al. Decision transformer: Reinforcement learning via sequence modeling //Advances in neural information processing systems. – 2021.

[2] Bulatov A. et al. Recurrent memory transformer //Advances in Neural Information Processing Systems. – 2022

[3] Dai Z. et al. Transformer-xl: Attentive language models beyond a fixed-length context //arXiv preprint arXiv:1901.02860. – 2019.