Model-Based Episodic Memory Induces Dynamic Hybrid Controls

Hung Le, Thommen Karimpanal George, Majid Abdolshah, Truyen Tran, Svetha Venkatesh
thai.le@deakin.edu.au

APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY

Introduction

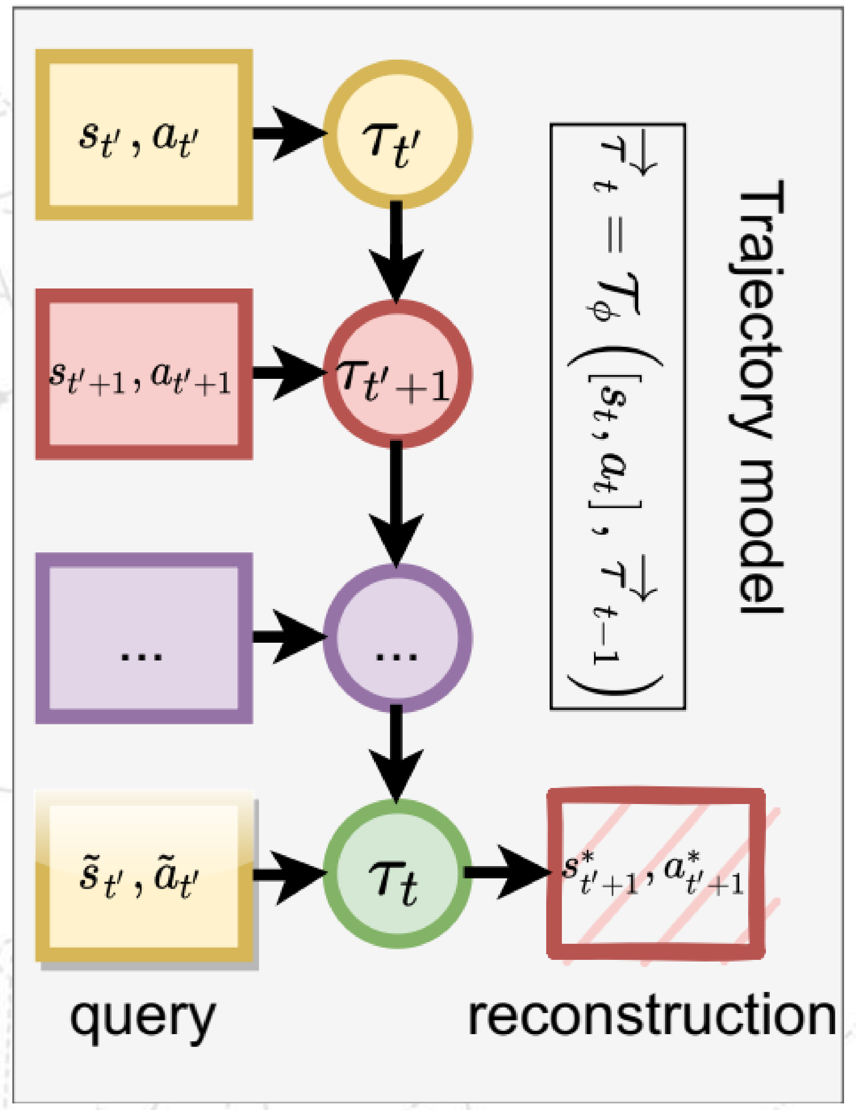
Episodic control [1] enables sample efficiency in reinforcement learning by recalling past experiences from an episodic memory. We propose a new model-based episodic memory of trajectories addressing current limitations of episodic control. Our memory estimates trajectory values, guiding the agent towards good policies (MBEC). Built upon the memory, we construct a complementary learning model via a dynamic hybrid control unifying model-based, episodic and habitual learning into a single architecture (MBEC++).

Trajectory representation learning

- Trajectory model is LSTM. Hidden state $\vec{\tau}$ is the representation
- **Self-supervised learning:** recall past events given a query as the preceding event (reconstruction loss)
- 2 trajectories having more common transitions are closer in the representation space
- **Trajectorial recall loss:**

$$\mathcal{L}_{tr} = E\left(\|y^*(t) - [s_{t'+1}, a_{t'+1}]\|_2^2\right)$$

$$y^*(t) = \mathcal{G}_\omega\left(\mathcal{T}_\phi([\tilde{s}_{t'}, \tilde{a}_{t'}], \vec{\tau}_t)\right)$$



Memory operations

Given a key-value episodic memory: $\mathcal{M} = \{\mathcal{M}^k, \mathcal{M}^v\}$

1. Memory read: given a query $\vec{\tau}$, randomly choose taking either (a) average or (b) max value of query's neighbors as the estimated value of the query:

$$\text{read}(\vec{\tau}|\mathcal{M}) = \begin{cases} \sum_{i \in \mathcal{N}^{\mathcal{K}_r}(\vec{\tau})} \frac{\langle \mathcal{M}_i^k, \vec{\tau}\rangle \mathcal{M}_i^v}{\sum_{j \in \mathcal{N}^{\mathcal{K}}(\vec{\tau})} \langle \mathcal{M}_j^k, \vec{\tau}\rangle} & (a) \\ \max_{i \in \mathcal{N}^{\mathcal{K}_r}(\vec{\tau})} \mathcal{M}_i^v & (b) \end{cases}$$

2. Memory write: the values of the query $\vec{\tau}$'s neighbors approach the written value with speeds relative to the distances:

$$\forall i \in \mathcal{N}^{\mathcal{K}_w}(\vec{\tau}): \mathcal{M}_i^v \leftarrow \mathcal{M}_i^v + \alpha_w\left(\hat{V}(\vec{\tau}) - \mathcal{M}_i^v\right)$$

where the written value: $\hat{V}(\vec{\tau}_t) = \sum_{i=0}^{T-t-1} \gamma^i r_{t+1+i}$

3. Memory refine: at any step, we perform memory read to estimate bootstrapped value Q' of next $\vec{\tau}_t'(a)$, which is written to update the values of the current $\vec{\tau}_{t-1}$'s neighbors:

$$Q' = \max_a r_\varphi(s_t, a) + \gamma \text{read}\left(\vec{\tau}_t'(a) \middle| \mathcal{M}\right)$$

$$\mathcal{M} \leftarrow \text{write}(\vec{\tau}_{t-1}, Q'|\mathcal{M})$$

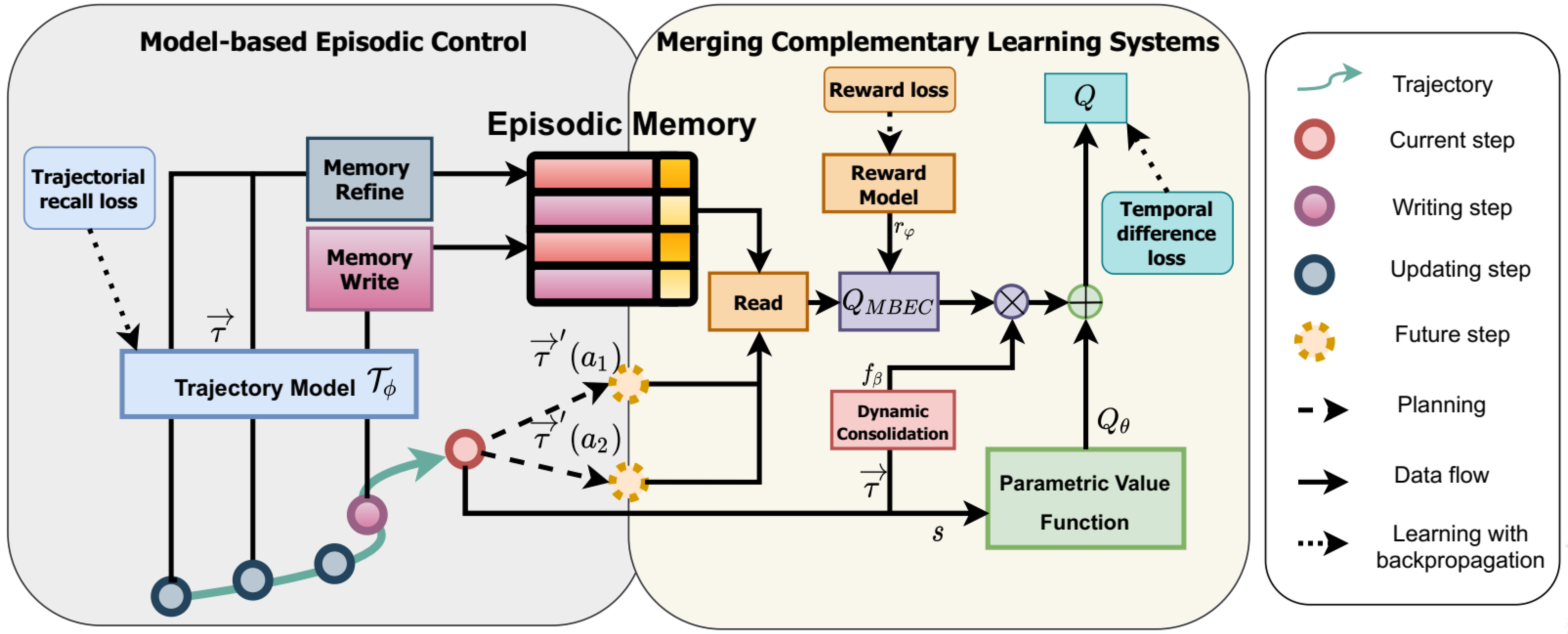
Memory-based planning

Step 1: estimate episodic value of taking an action a from state s: $Q_{MBEC}(s,a) = r_\varphi(s,a) + \gamma\text{read}\left(\vec{\tau}'(a)\middle|\mathcal{M}\right)$

Step 2: combine episodic value with DQN's value through gating: $Q(s_t,a_t) = Q_{MBEC}(s_t,a_t) f_\beta(\vec{\tau}_{t-1}) + Q_\theta(s_t,a_t)$

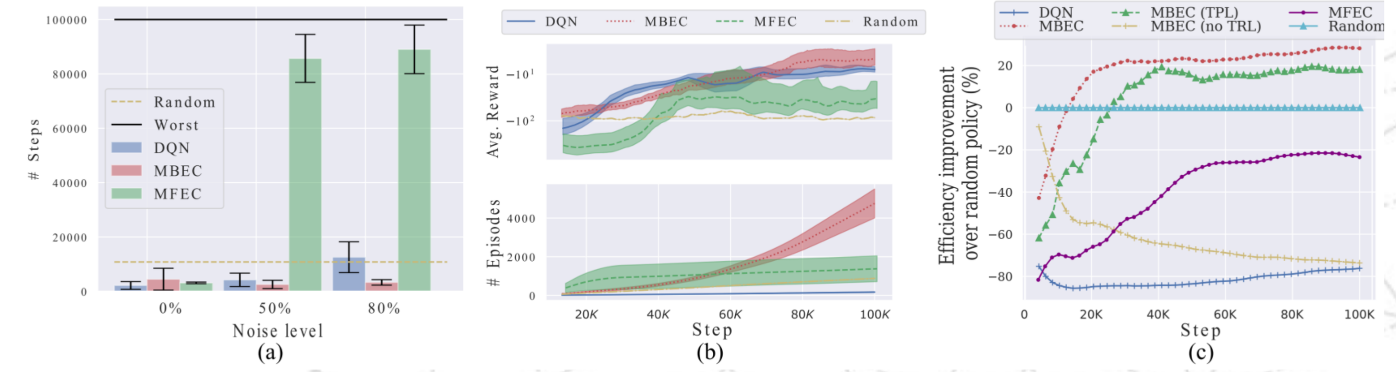
Step 3: train the networks via minimizing TD error: $\mathcal{L}_q = E\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)^2$

Dynamic hybrid control with the episodic memory at its core

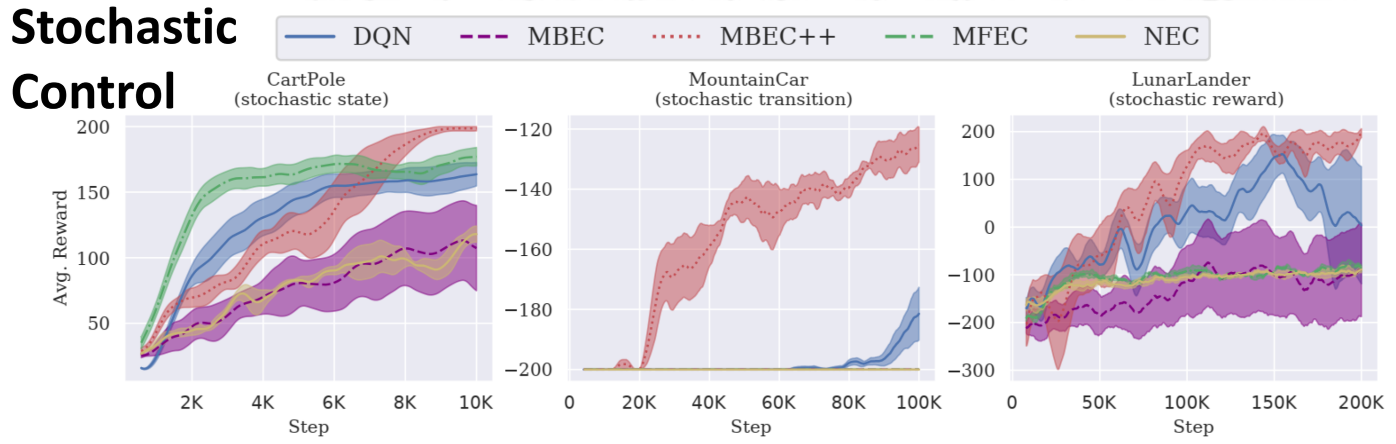


Experimental results

2D Maze: (a) Noisy, (b) Trap and (c) Dynamic mode.

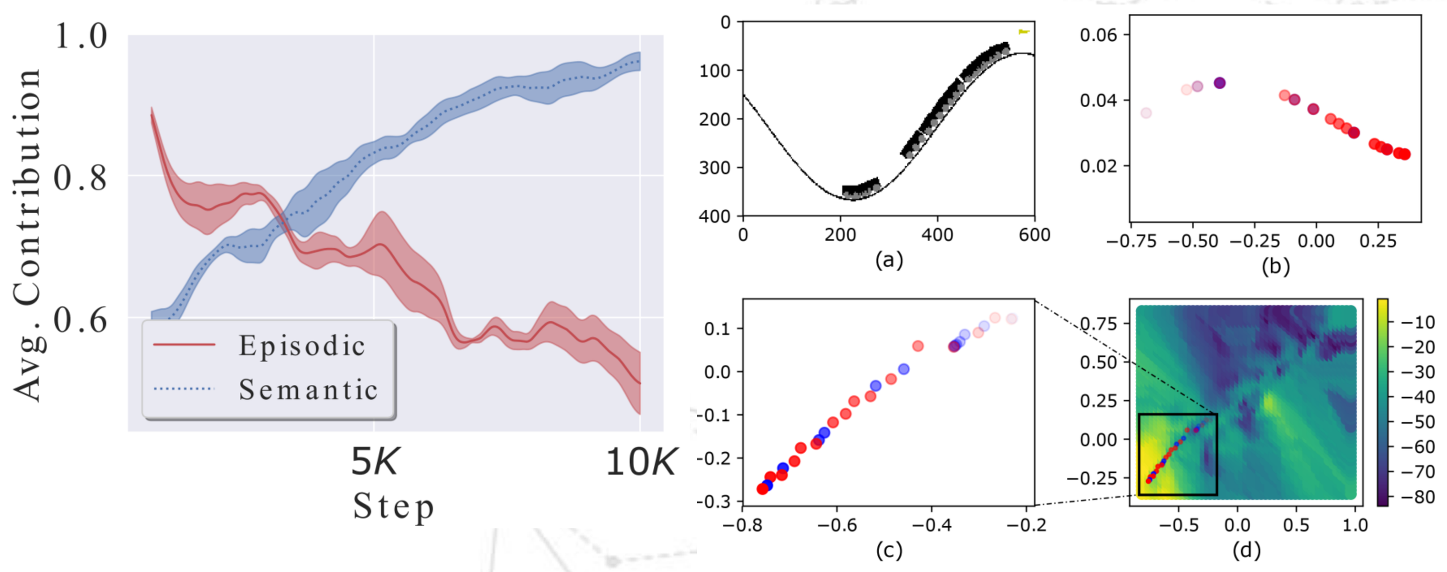


Stochastic Control



Atari games: Human normalized scores (mean/median) at 10 million frames for all and a subset of 25 games.

Model	All	25 games
Nature DQN	15.7/51.3	83.6/16.0
MFEC	85.0/45.4	77.7/40.9
NEC	99.8/54.6	106.1/53.3
EMDQN*	528.4/92.8	250.6/95.5
EVA	-	172.2/39.2
ERLAM	-	515.4/103.5
MBEC++	**654.0/117.2**	**518.2/133.4**



Noisy CartPole: Automatically reduce episodic contribution overtime

Noisy MountainCar: Trajectory Model learns smooth representations despite noisy states

References

[1] Blundell et al., Hassabis. Model-free episodic control. arXiv preprint arXiv:1606.04460, 2016.