Model-Based Episodic Memory Induces Dynamic Hybrid Controls
Hung Le, Thommen Karimpanal George, Majid Abdolshah, Truyen Tran, Svetla Venkatesh
thai.le@deakin.edu.au

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 model is \( Q = \sum_{t} r^t \) Trajectory, which is \( \mathcal{G}_t = \sum_{t} r^t \) where the written value: \( \mathcal{V}_t = \sum_{t=-1}^{t+1} y^t \) Trajectory, self memory: \( m = \sum_{t} r^t \tau^t \) where the estimated value of the query:

\[
\text{read}(\mathcal{T}, \mathcal{M}) = \sum_{t \in \mathcal{N}^\alpha(t)} (\mathcal{M}^{a_{t^{-1}}, \tau}(\mathcal{T}), \mathcal{M}^{a_{t+1}, \tau}) (a) \\
\max_{t \in \mathcal{N}^\alpha(t)} (\mathcal{M}^{a_{t-1}, \tau}(\mathcal{T})) (b)
\]

Memory operations
Given a key-value episodic memory, \( \mathcal{M} = \{ \mathcal{M}^{a, \tau}, \mathcal{M}^{\tau} \} \)
1. Memory read: given a query \( t \), randomly choosing either (a) average or (b) max value of query’s neighbors as the estimated value of the query:

\[
\text{read}(\mathcal{T}, \mathcal{M}) = \sum_{t \in \mathcal{N}^\alpha(t)} (\mathcal{M}^{a_{t^{-1}}, \tau}(\mathcal{T}), \mathcal{M}^{a_{t+1}, \tau}) (a) \\
\max_{t \in \mathcal{N}^\alpha(t)} (\mathcal{M}^{a_{t-1}, \tau}(\mathcal{T})) (b)
\]

2. Memory write: the values of the query \( t \)'s neighbors approach the written value with speeds relative to the distances:

\[
\forall i \in \mathcal{N}_\tau(\mathcal{T}): \mathcal{M}^{a_i, \tau} = \mathcal{M}^{a_{t^{-1}}, \tau} + \alpha (\mathcal{V}(\mathcal{T}) - \mathcal{M}^{a_{t^{-1}}, \tau})
\]

3. Memory refine: at any step, we perform memory read to estimate bootstrapped value \( Q' \) of next \( t' \), which is written to update the values of the current \( t' \)'s neighbors:

\[
Q' = \max_{a_i} r_{t'}(a_i) + \gamma \text{read}(\mathcal{T}, \mathcal{M}(t')) (a) \\
\mathcal{M} \leftarrow \text{write}(\mathcal{T}, Q') (b)
\]

Memory-based planning
Step 1: estimate episodic value of taking an action \( a \) from state \( s \):

\[
Q_{MBEC}(s, a) = r_s(a, s) + \gamma \text{read}(\mathcal{T}, \mathcal{M}(s')) (a) \\
\]

Step 2: combine episodic value with DQN's value through gating:

\[
Q(s, a) = Q_{MBEC}(s, a) f_{\alpha}(s, a) + Q(s, a)
\]

Step 3: train the networks via minimizing TD error:

\[
L_q = E \left( y^s(t) - \left[ r^s(t+1) + \gamma \text{read}(\mathcal{T}, \mathcal{M}(s')) \right] \right)^2
\]

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.

References

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 \( \bar{t} \) 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
- Trajectory recall loss:

\[
L_{rec} = E \left( y^s(t) - \left[ r^s(t+1) + \gamma \text{read}(\mathcal{T}, \mathcal{M}(s')) \right] \right)^2
\]

Memory-based planning
Step 1: estimate episodic value of taking an action \( a \) from state \( s \):

\[
Q_{MBEC}(s, a) = r_s(a, s) + \gamma \text{read}(\mathcal{T}, \mathcal{M}(s')) (a) \\
\]

Step 2: combine episodic value with DQN's value through gating:

\[
Q(s, a) = Q_{MBEC}(s, a) f_{\alpha}(s, a) + Q(s, a)
\]

Step 3: train the networks via minimizing TD error:

\[
L_q = E \left( y^s(t) - \left[ r^s(t+1) + \gamma \text{read}(\mathcal{T}, \mathcal{M}(s')) \right] \right)^2
\]

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.

References