Episodic Control in Reinforcement Learning

Why episodic control

- Sample Inefficiency
- Single Learning Model
- Slow Reward Propagation (TD(0))

- Episodic Memory
- Hippocampus in Human Brain
- Learning Efficiency

- ✓ State Relocation
- ✓ Learning from Demonstration
- ✓ Self-Imitation
- ✓ Self-supervised RL
- ✓ Curiosity-Driven RL
- ✓ **Episodic Control**

Model-Free Episodic Control

Charles Blundell
Google DeepMind
cblundell@google.com

Benigno Uria
Google DeepMind
buria@google.com

Alexander Pritzel
Google DeepMind
apritzel@google.com

Yazhe Li
Google DeepMind
yazhe@google.com

Avraham Ruderman
Google DeepMind
aruderman@google.com

Joel Z Leibo
Google DeepMind
jzl@google.com

Jack Rae
Google DeepMind
jwrae@google.com

Daan Wierstra
Google DeepMind
wierstra@google.com

Demis Hassabis
Google DeepMind
demishassabis@google.com

Arxiv 2016

Methods

- Tabular RL

$Q^{EC}(s,a)$	$a1$	$a2$	$a3$	$a4$
$s1$	100	200	-10	32
$s2$	34	152	1111	54
$s3$	424	0	132	24

- Two key problems in tabular RL
  - A large amount of memory consumption
  - Lack a way to generalize across similar states.

Methods

Update

$$Q^{\mathrm{EC}}(s_t, a_t) \leftarrow \begin{cases} R_t & \text{if } (s_t, a_t) \notin Q^{\mathrm{EC}}, \\ \max\left\{Q^{\mathrm{EC}}(s_t, a_t), R_t\right\} & \text{otherwise,} \end{cases} \tag{1}$$

where R_t is the discounted return received after taking action a_t in state s_t. Note that (1) is not a general purpose RL learning update: since the stored value can never decrease, it is not suited to rational action selection in stochastic environments.[1]

Action selection

$$\widehat{Q^{\mathrm{EC}}}(s, a) = \begin{cases} \frac{1}{k} \sum_{i=1}^{k} Q^{\mathrm{EC}}(s^{(i)}, a) & \text{if } (s, a) \notin Q^{\mathrm{EC}}, \\ Q^{\mathrm{EC}}(s, a) & \text{otherwise,} \end{cases} \tag{2}$$

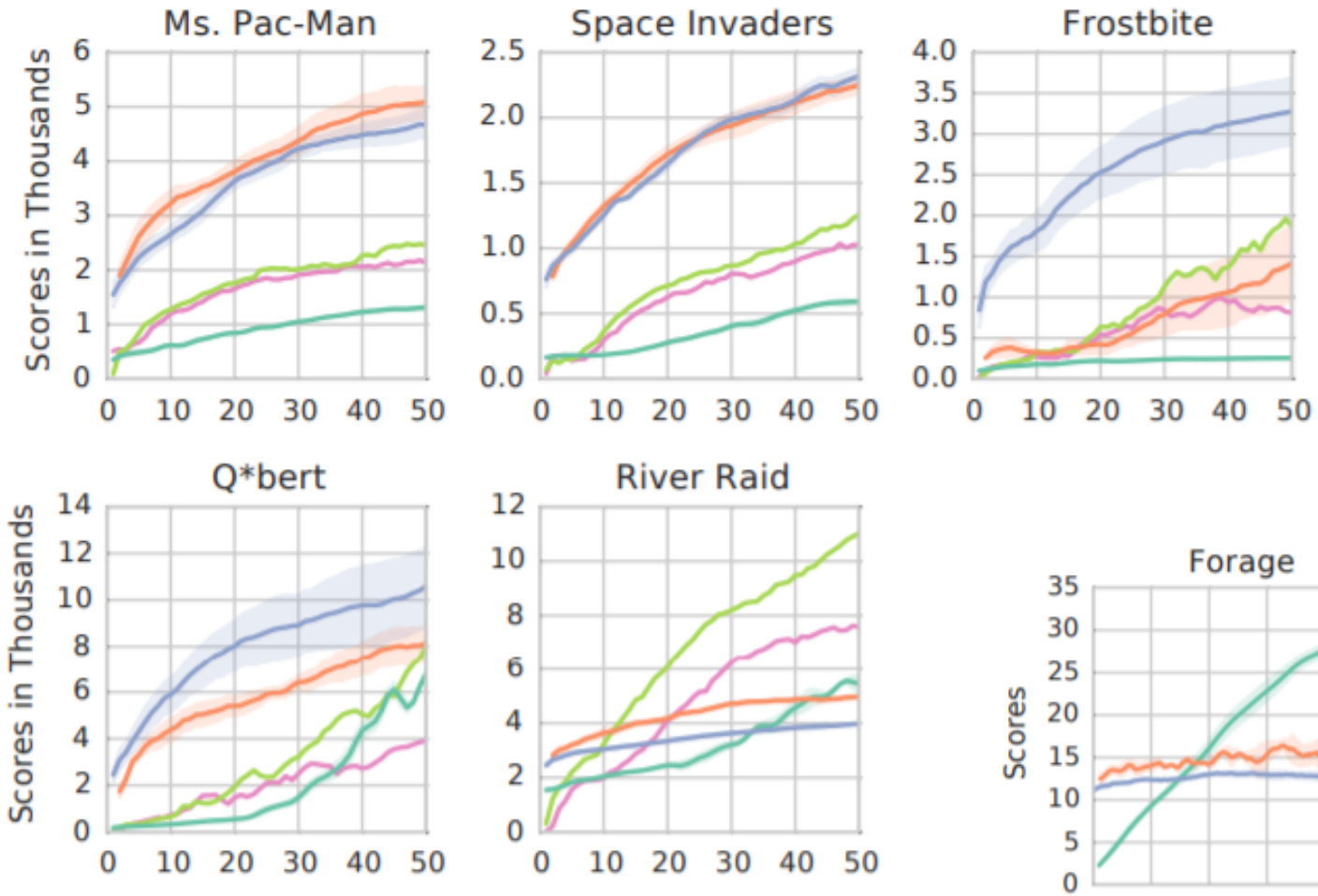
where $s^{(i)}$, $i = 1, \ldots, k$ are the k states with the smallest distance to state s.[2]

Methods

Algorithm 1 Model-Free Episodic Control.

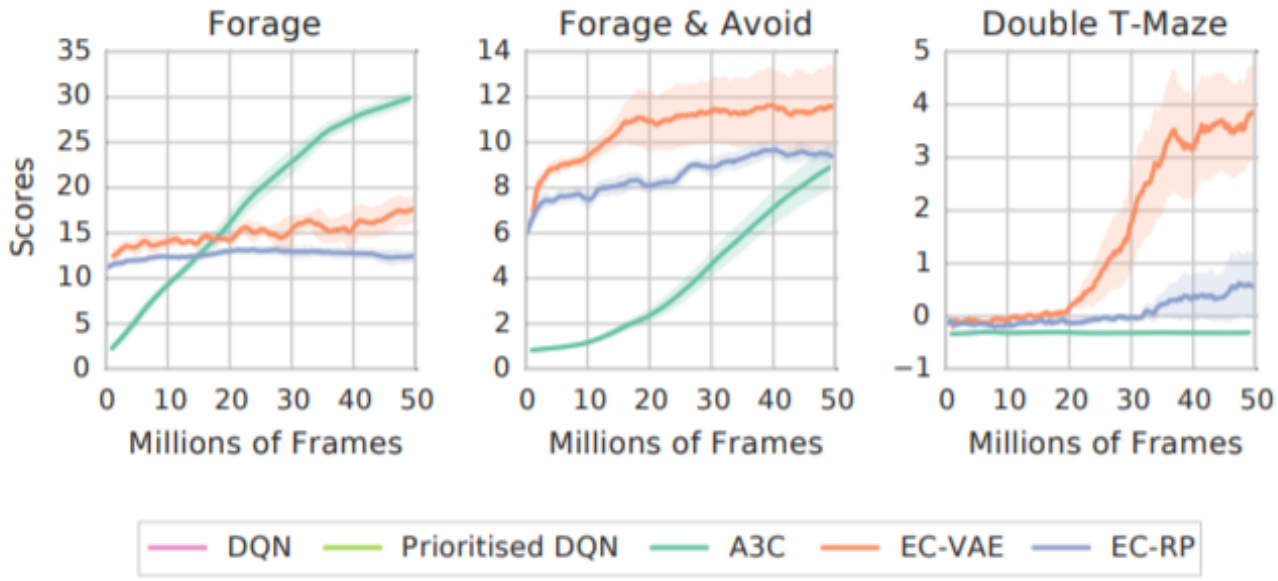
for each episode **do**
  for $t = 1, 2, 3, \ldots, T$ **do**
    Receive observation o_t from environment.
    Let $s_t = \phi(o_t)$.
    Estimate return for each action a via (2)
    Let $a_t = \arg\max_a \widehat{Q^{\mathrm{EC}}}(s_t, a)$
    Take action a_t, receive reward r_{t+1}
  end for
  for $t = T, T-1, \ldots, 1$ **do**
    Update $Q^{\mathrm{EC}}(s_t, a_t)$ using R_t according to (1).
  end for
end for

Experiments



Two different embedding:

- Random Project
- VAE



Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Episodic Memory Deep Q-Networks

Zichuan Lin[13], Tianqi Zhao[2], Guangwen Yang[1], Lintao Zhang[3]
[1]Tsinghua University
[2]Microsoft
[3]Microsoft Research
linzc16@mails.tsinghua.edu.cn, tianqi.zhao@microsoft.com,
ygw@tsinghua.edu.cn, lintaoz@microsoft.com

IJCAI 2018

Framework

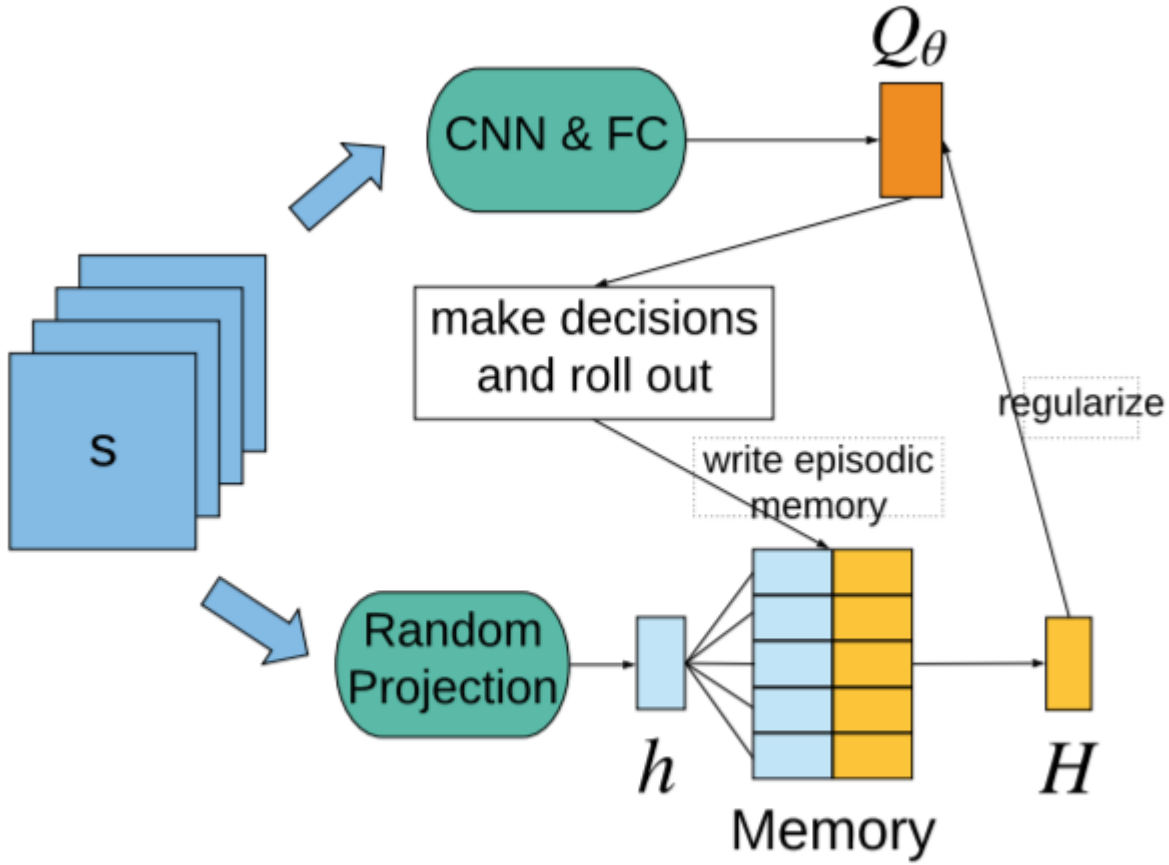


Figure 1: EMDQN architecture on a single action.

Method

TD error

$$\min_{\theta} \sum_{(s_i,a_i,r_i,s_{i+1}\in D)} \Big[(Q_\theta(s_i,a_i) - S(s_i,a_i))^2 + \lambda(Q_\theta(s_i,a_i) - H(s_i,a_i))^2\Big], \tag{6}$$

$$S(s_t,a_t) = r_t + \gamma \max_{a'} Q_\theta(s_{t+1},a'). \tag{3}$$

$$H(s_t,a_t) = \max_i R_i(s_t,a_t), i \in \{1,2,...,E\}, \tag{4}$$

Experiments

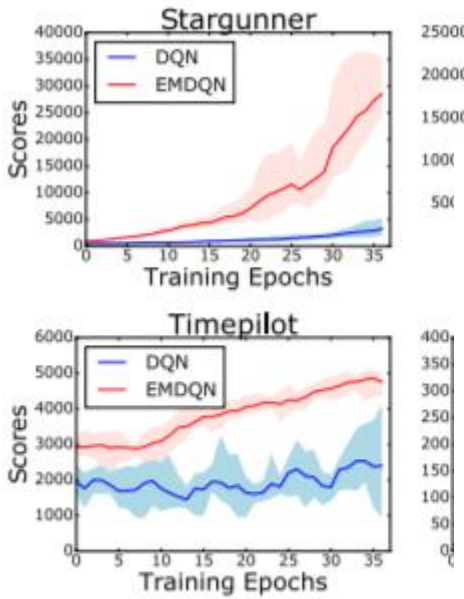


Figure 2: Testing scores f
tative games. The scores
epochs. Each game is rur

	Mean	Median
DQN(40M)	151.2%	52.7%
MFEC(40M)	142.2%	61.9%
NEC(40M)	144.8%	83.3%
EMDQN(40M)	**528.4%**	**92.8%**
DQN(200M)	227.9%	79.1%
DDQN(200M)	330.3%	114.7%

Table 1: Mean and median human-normalized scores at 40 Million frames over 57 Atari games.

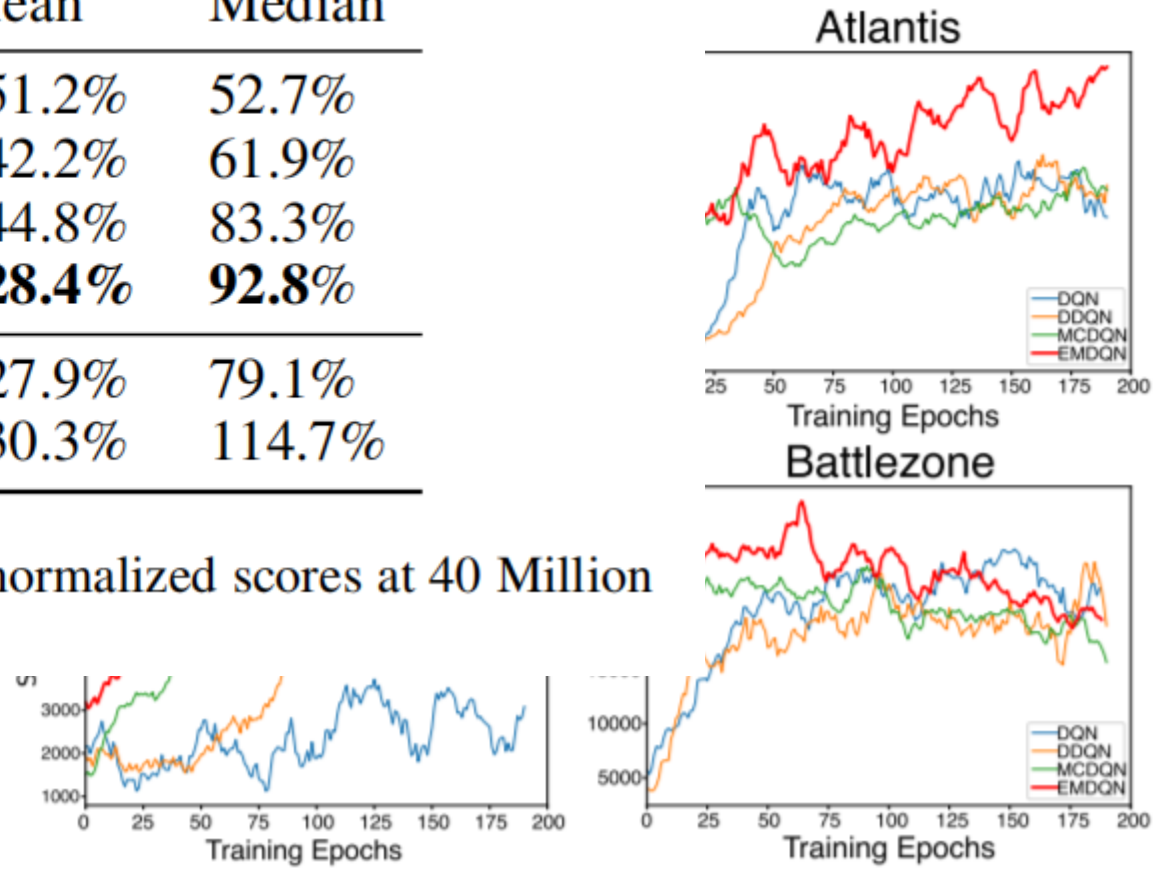


Figure 3: Training curves on 200M frames.

EPISODIC REINFORCEMENT LEARNING WITH ASSOCIATIVE MEMORY

Guangxiang Zhu[1]*, **Zichuan Lin**[2]*, **Guangwen Yang**[2], **Chongjie Zhang**[1]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
guangxiangzhu@outlook.com, linzc16@mails.tsinghua.edu.cn,
ygw@tsinghua.edu.cn, chongjie@tsinghua.edu.cn

ICLR 2020

Methods

- Motivation
  - Previous work on episodic reinforcement learning neglects the relationship between states and only stored the experiences as unrelated items.
  - Studies in psychology and cognitive neuroscience discover that **associative memory found in hippocampus plays an important role in human activities, which associates past experiences by remembering relationship between them.**

Methods

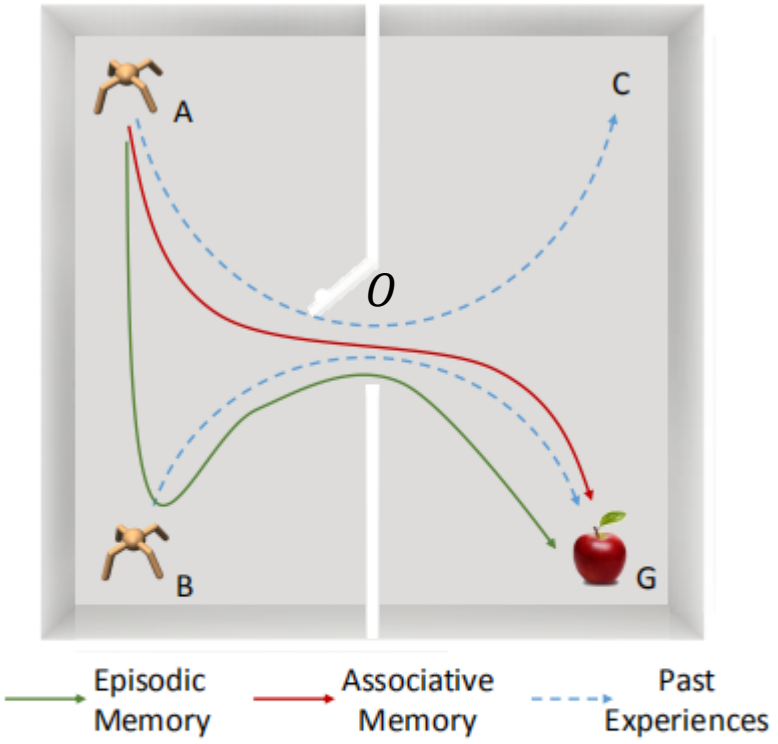


Figure 1: Comparison of selected policies based on episodic memory and associative memory. An agent starts from two place A and B to collect two experiences.

- Trajectory "AC" with low reward.
- Trajectory "BG" with high reward.

$Q^{EC}(s,a)$	$a1$	$a2$
A	low	low
C	low	low
B	$high$	$high$
G	$high$	$high$
O	$high$	$high$

$Q^{EC}(s,a)$	$a1$	$a2$
A	$high$	$high$
C	low	low
B	$high$	$high$
G	$high$	$high$
O	$high$	$high$

Methods

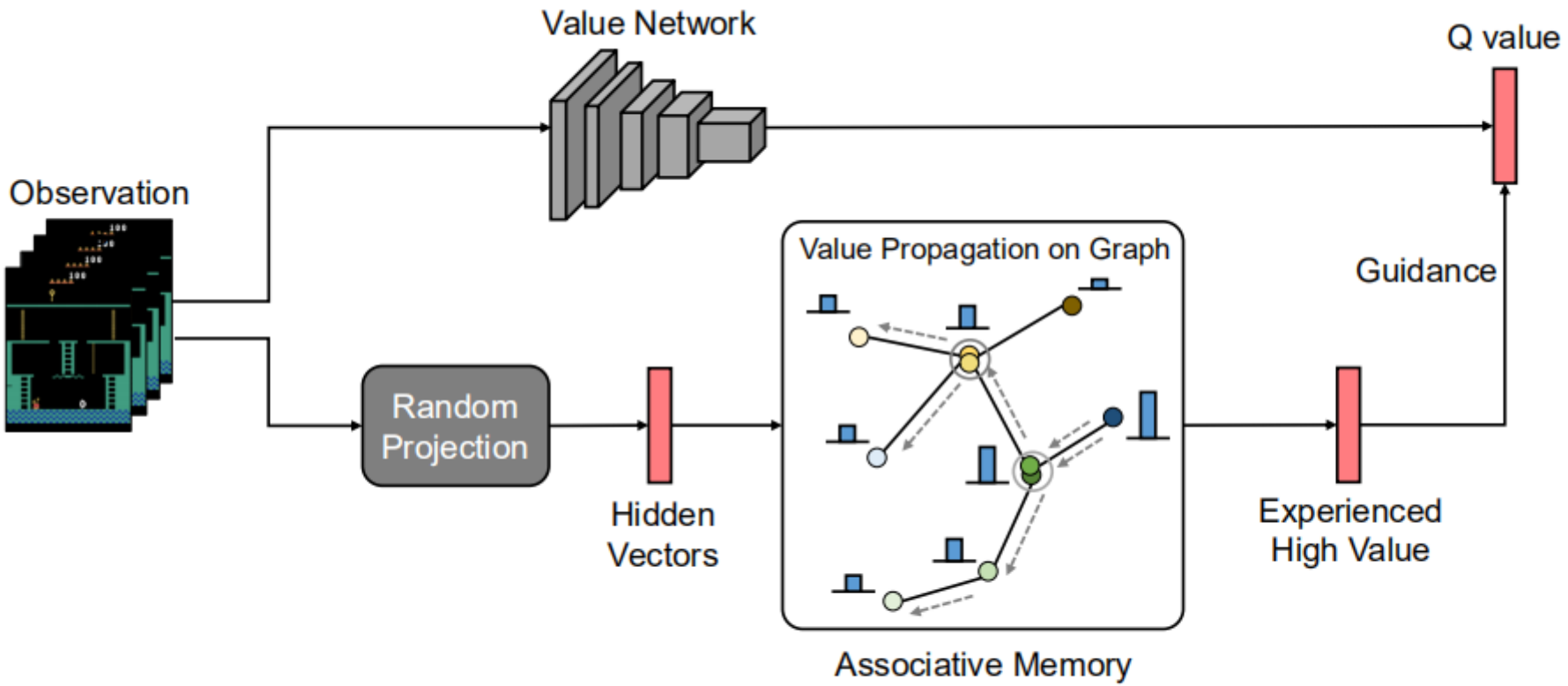


Figure 3: Overall framework of ERLAM.

$$\mathcal{G} = (V, E),\ V = \phi(s), E = \{s \to s' \mid (s, a, s') \text{ is stored in memory}\}. \tag{2}$$

Methods

$$Q_{\mathcal{G}}(\phi(s), a) \leftarrow r + \gamma \max_{a'} Q_{\mathcal{G}}(\phi(s'), a'). \quad (3)$$

Algorithm 1 Value propagation in Associative Memory

h: embedded vector of state, $h = \phi(s)$
$\mathcal{G} \leftarrow$ Sort nodes in graph $\mathcal{G}$ by sequential step ID t in descending order
repeat
  for $m = 1 \ldots |\mathcal{G}|$ **do**
    Get current state-action pair $(s, a) = (s_m, a_m)$
    Get successor state embedding s' and action a' using graph $\mathcal{G}$.
    Update graph augmented memory using Eq. 3.
  end for
until $Q_{\mathcal{G}}$ converges

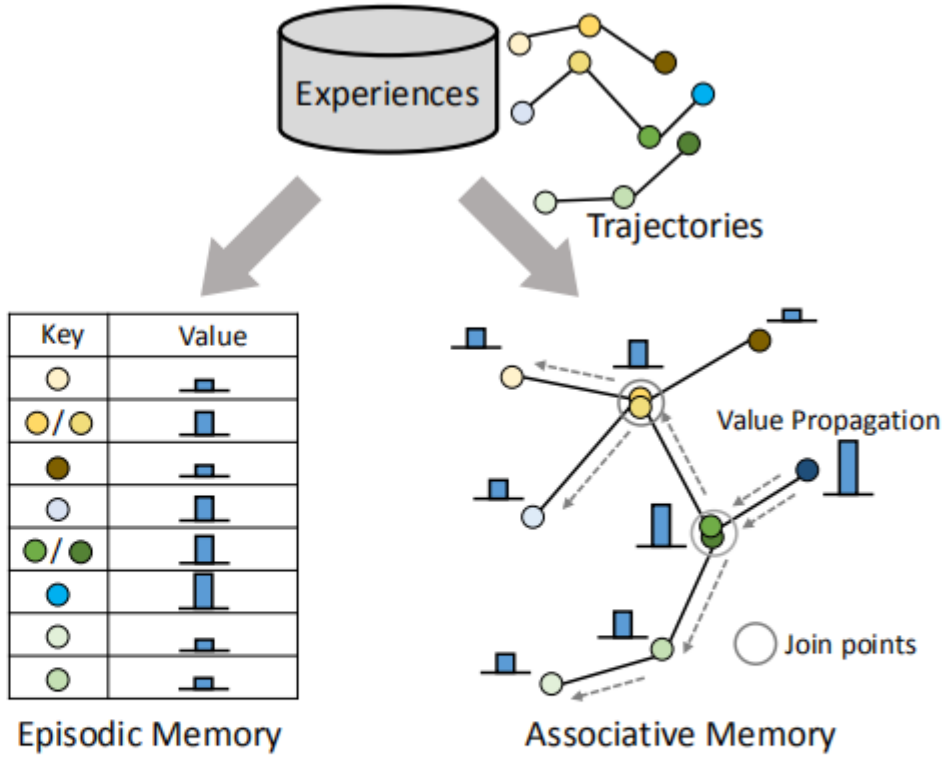


Figure 2: Comparison of episodic memory and associative memory.

$$L_\theta = \mathbb{E}_{(s,a,s',r)\sim\mathcal{D}}\Big[\big(r + \gamma \max_a Q_{\hat{\theta}}(s', a) - Q_\theta(s, a)\big)^2 + \lambda\big(Q_{\mathcal{G}}(\phi(s), a) - Q_\theta(s, a)\big)^2\Big], \quad (4)$$

Methods

Algorithm 2 ERLAM: Episodic Reinforcement Learning with Associative Memory

$\mathcal{D}$: Replay buffer
$\mathcal{G}$: Graph (Associative memory)
T_e: Trajectory length of e-th episode
K: Associate frequency

for Episode number $e = 1 \ldots E$ **do**
  for $t = 1 \ldots T_e$ **do**
    Receive initial observation s_t from environment with state embedding $h_t = \phi(s_t)$
    $a_t \leftarrow \epsilon$-greedy policy baesd on $Q_\theta(s_t, a)$
    Take action a_t, receive reward r_t and next state s_{t+1}
    Append (s_t, a_t, r_t, s_{t+1}) to $\mathcal{D}$
    if t mod $update\_freq$ == 0 **then**
      Sample training experiences (s, a, r, t) from $\mathcal{D}$
      Retrieve $Q_\mathcal{G}(\phi(s), a)$ from associative memory
      Update parameter θ using Eq. 4　利用带约束的优化目标进行网络更新
    end if
  end for
  for $t = T_e \ldots 1$ **do**
    $R_t \leftarrow r_t + \gamma R_{t+1}$, if $t < T_e$; $R_t \leftarrow r_t$, if $t = T_e$
    Append (h_t, a_t, r_t, t, R_t) to $\mathcal{G}$ if $(h_t, a_t) \notin \mathcal{G}$
    Update $Q_\mathcal{G}$ using Eq. 1 if $(h_t, a_t) \in \mathcal{G}$　计算采样的episode的各(s,a)的累积奖励，去更新episodic memory中的值
  end for
  if e mod K == 0 **then**
    Run Algorithm 1 to update $Q_\mathcal{G}$　每隔一定episode利用图关系来更新累积奖励值
  end if
end for

Methods

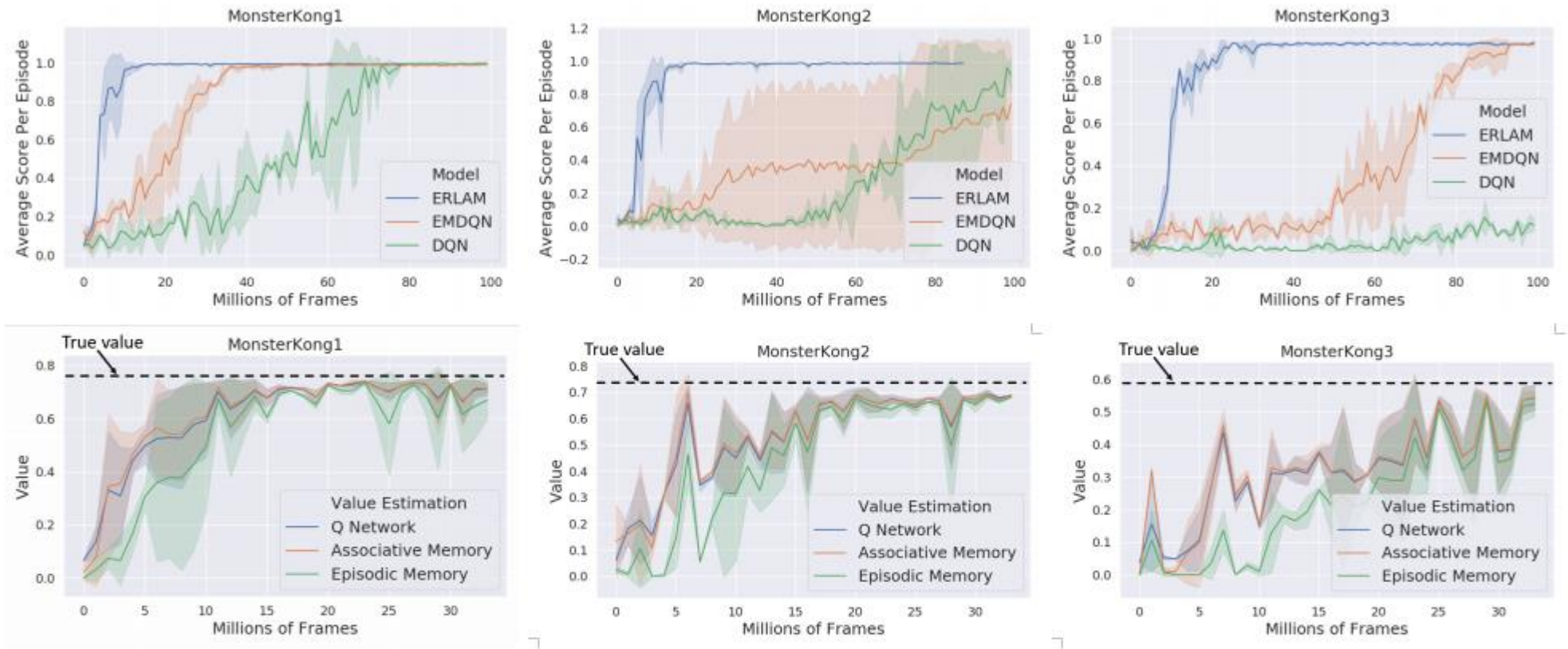


Figure 5: Learning curves of ERLAM, EMDQN and DQN on *Monster Kong*. The top row compares the average scores per episode between all models. The bottom row shows state-action value estimates by associative memory, episodic memory, and Q networks when running ERLAM. The black dash line represents the actual discounted state-action values of the best learned policy.

Thanks