Prioritized Experience Replay

Tom Schaul, John Quan, Ioannis Antonoglou
and David Silver

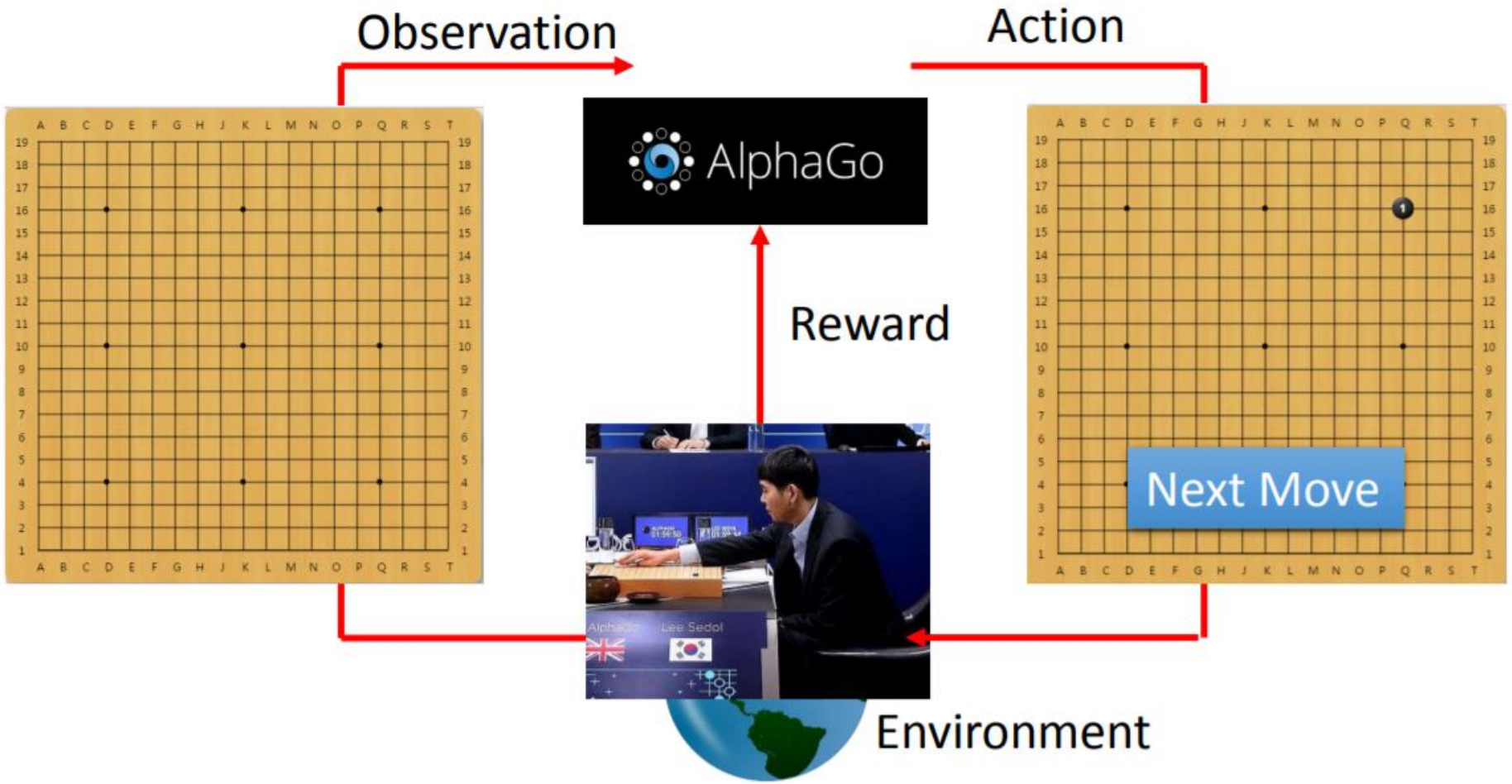
Google DeepMind

ICLR-2016

RL

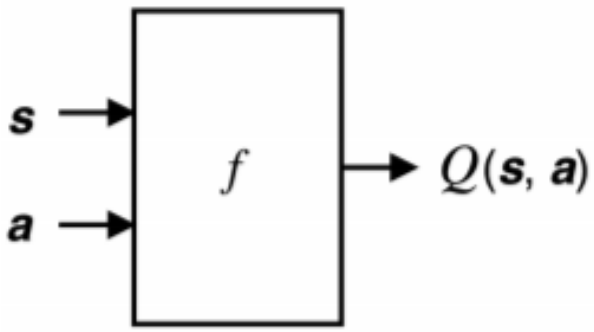
Reference: http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2017/Lecture/RL%20(v4).pdf

RL example



Reference: http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2017/Lecture/RL%20(v4).pdf

DQN



Algorithm (TD): initialize Θ arbitrarily, iterate until converge:

1. Take action a from s using some exploration policy π' derived from f_{Q^*} (e.g., ε-greedy)
2. Observe s' and reward $R(s,a,s')$, update Θ using SGD:

$$\Theta \leftarrow \Theta - \eta \nabla_\Theta C, \text{ where}$$

$$C(\Theta) = \left[R(s,a,s') + \gamma \max_{a'} f_{Q^*}(s',a';\Theta) - f_{Q^*}(s,a;\Theta) \right]^2$$

Q: long-term reward

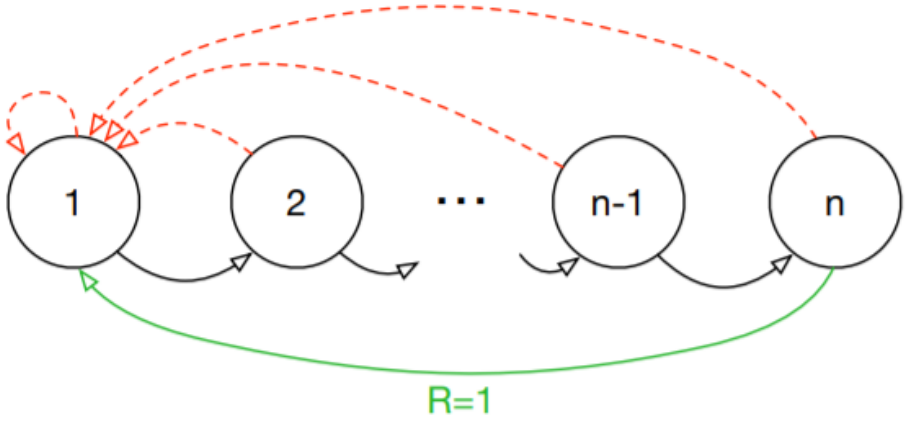
Training data: rollout the trajectory. from s to s'…
Avoid iid.

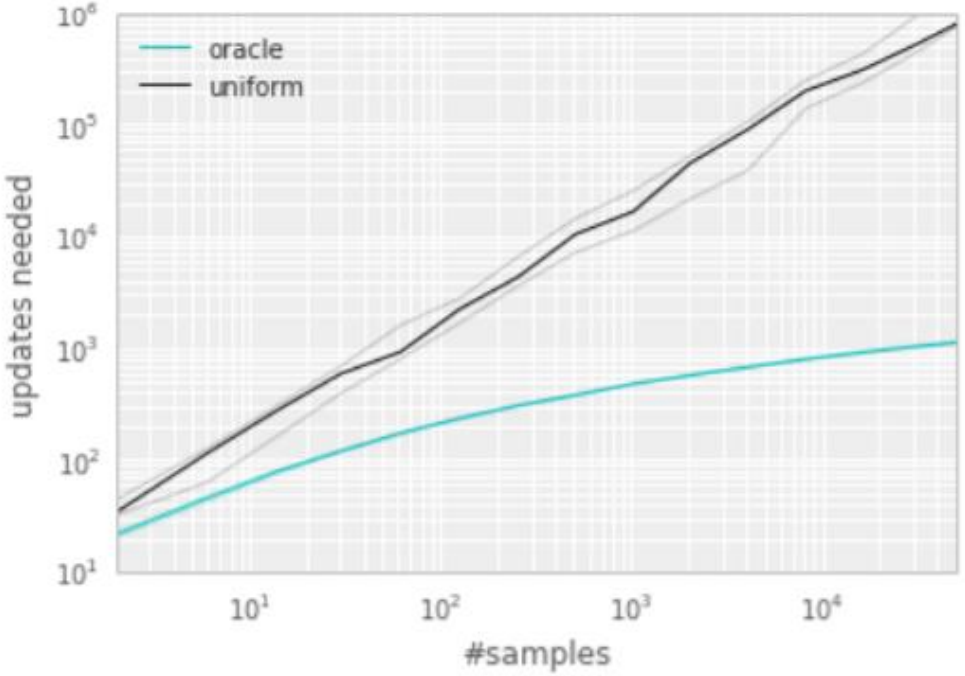
Experience replay

- Use a replay memory D to store recently seen transitions (s,a,r,s')(experience)
- Sample experience from the buffer and train.

Prioritized Experience Replay

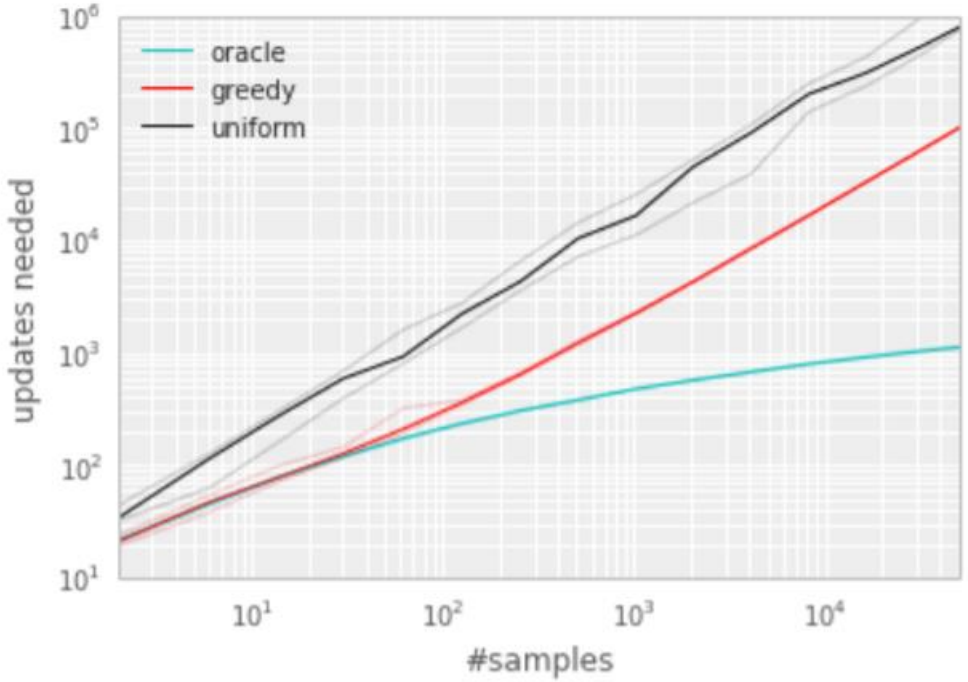
- Motivation: some transitions are more important.
- Example: Blind Cliffwalk





Prioritizing with TD-error

- Using TD-error as priority $R(s,a,s') + \gamma \max_{a'} f_{Q^*}(s',a';\Theta) - f_{Q^*}(s,a;\Theta)$
- Indicates how far is from target value.



Stochastic Prioritization

- Issues with greedy TD-error prioritization:

some transitions with low TD-error may not be replayed. Lack of diversity.

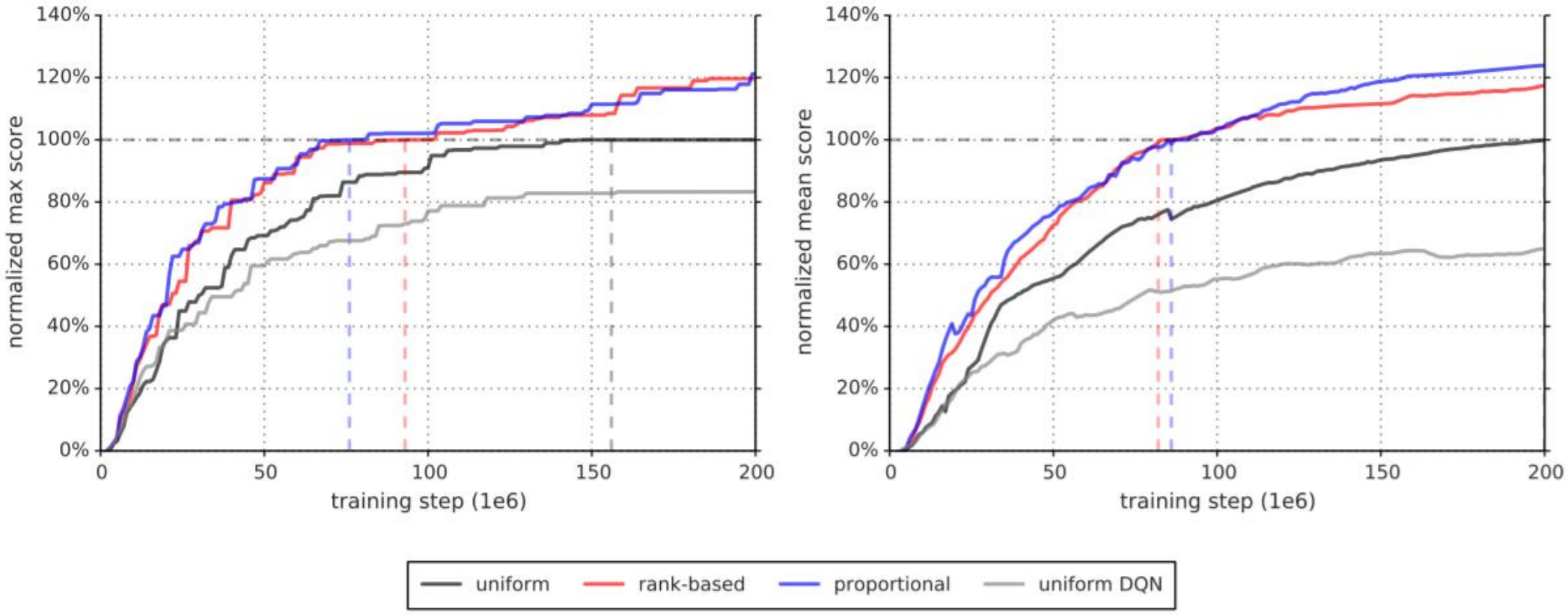
- Add stochastic

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$$

Proportional: $p_i = |\delta_i| + \epsilon,$

Rank-based: $p_i = \frac{1}{\text{rank}(i)}$

Atari experiment



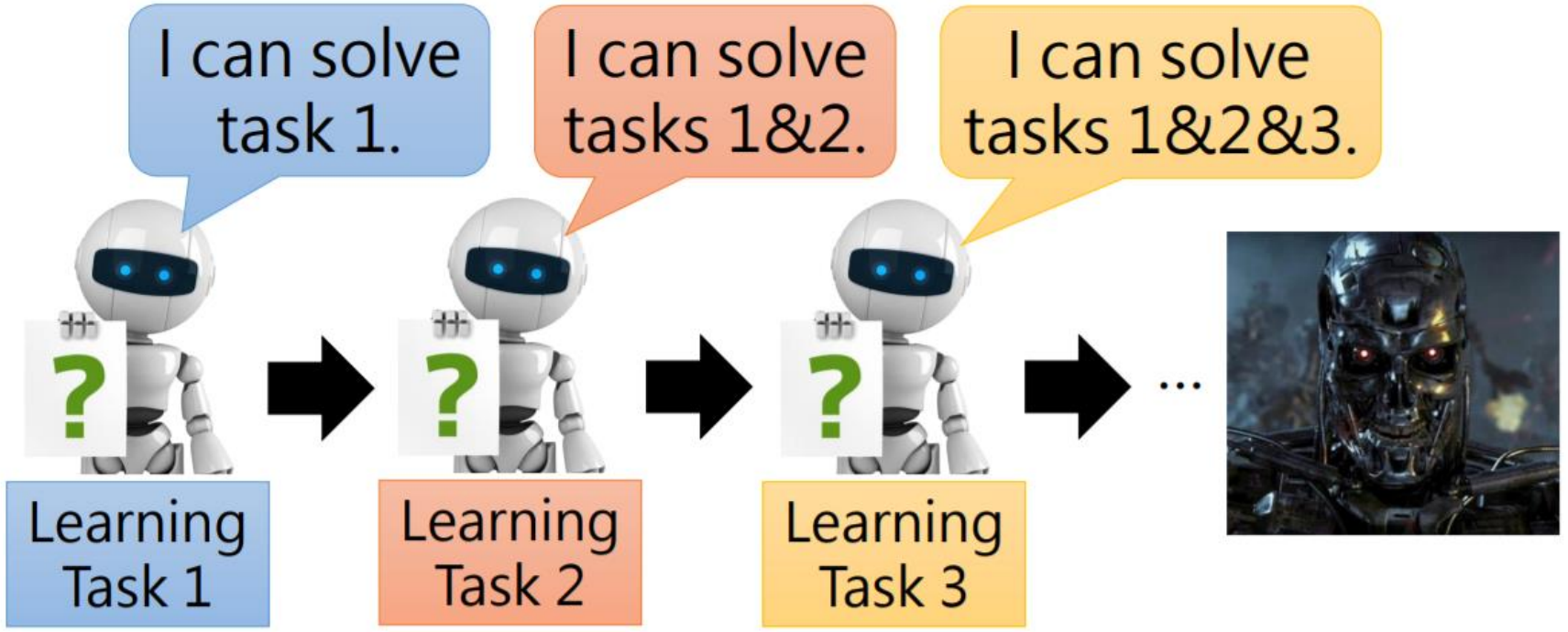
Selective Experience Replay for Lifelong Learning

David Isele, Akansel Cosgun

The University of Pennsylvania, Honda Research Institute

AAAI-2018

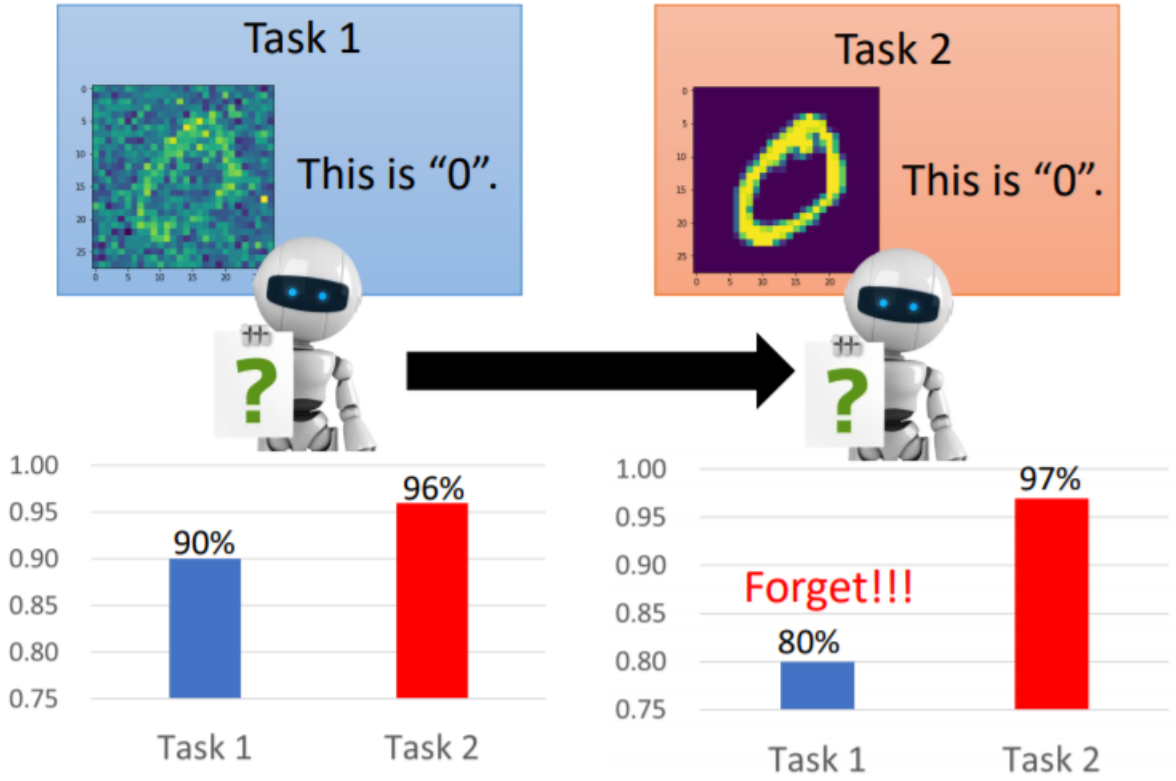
Lifelong Learning



Learn in sequence

Reference:
http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Lifelong%20Learning%20(v9).pdf

Catastrophic forgetting

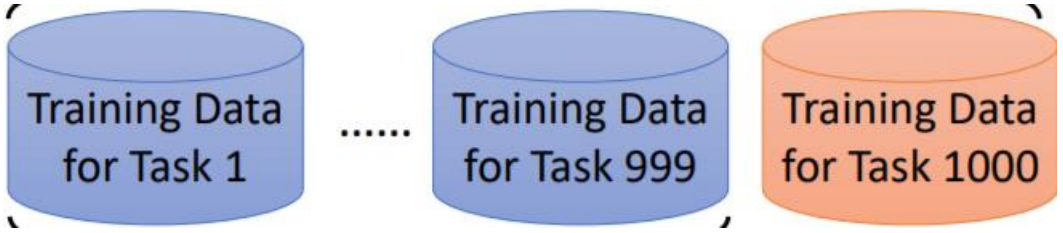


学了后面，忘了前面

Why?

Reference:
http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Lifelong%20Learning%20(v9).pdf

How to solve



- Suppose memory is limitless so that it can store all the data.
- Review data for task1. when learning task2.
- But not realistic

Selective experience replay

- Try to store experience for all the tasks. But only part of them.
- Selective strategy:
  - favoring surprise,
  - favoring reward,
  - matching the global training distribution,
  - maximizing coverage of the state space

Selective strategy

1. surprise

$$\mathcal{R}(e_i) = |r_i + \gamma \max_{a'} Q(s'_i, a') - Q(s_i, a_i)|$$

2. Reward

$$\mathcal{R}(e_i) = |R_i(e_i)| \quad .$$

Selective strategy

3. Global distribution

$$\mathcal{R}(e_i) \sim N(0,1)$$

Assign a random value to each experience, store experiences with highest value.

4. Coverage maximum

$$\mathcal{N}_i = \{e_j \; s.t. \; \text{dist}(e_i - e_j) < d\}$$
$$\mathcal{R}(e_i) = -|\mathcal{N}_i|, \text{order according to rank}$$

Experiment

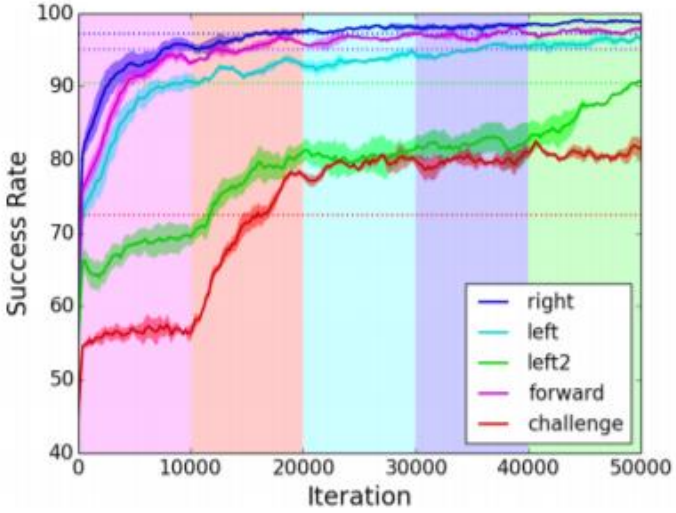
(a) *Right*

(b) *Left*

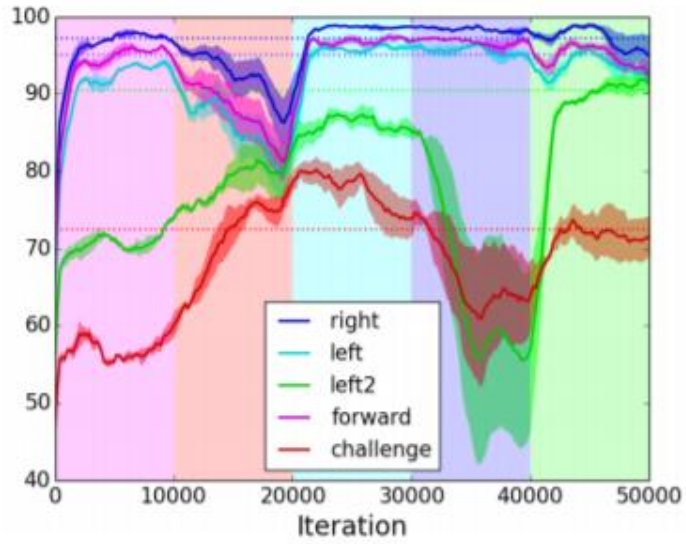
(c) *Left2*

(d) *Forward*

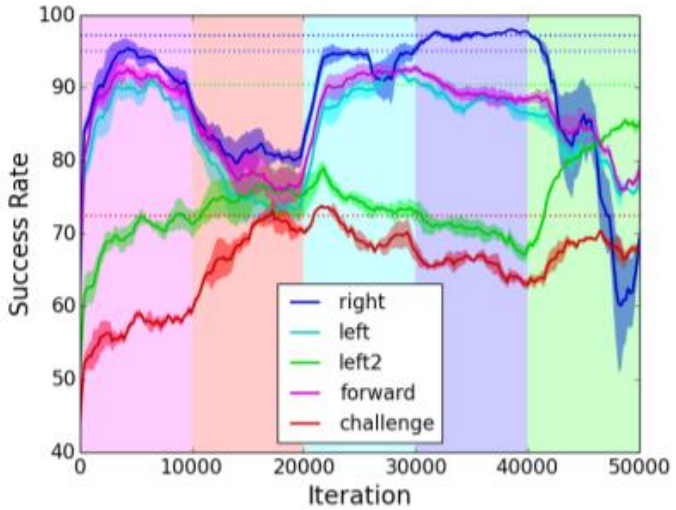
(e) *Challenge*



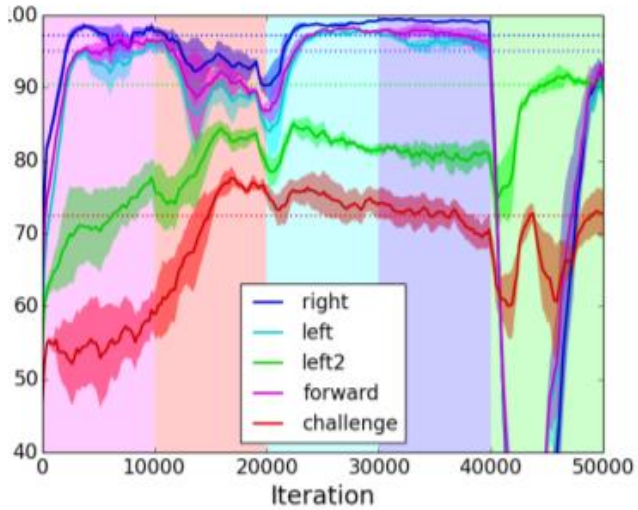
(a) Unlimited capacity



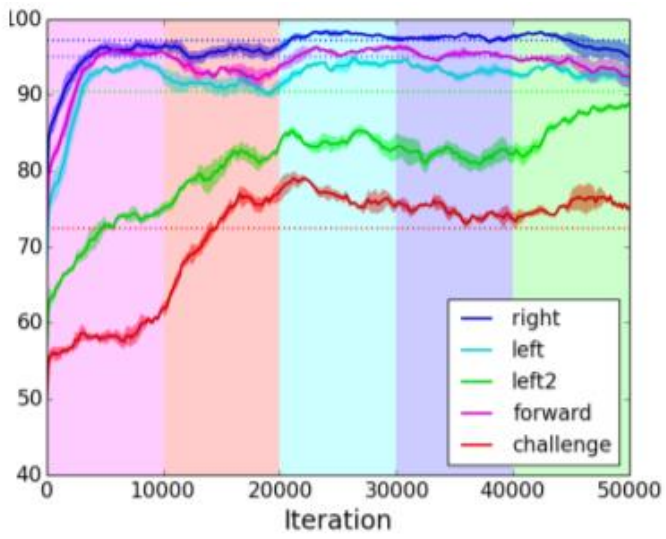
(b) Limited capacity (FIFO)



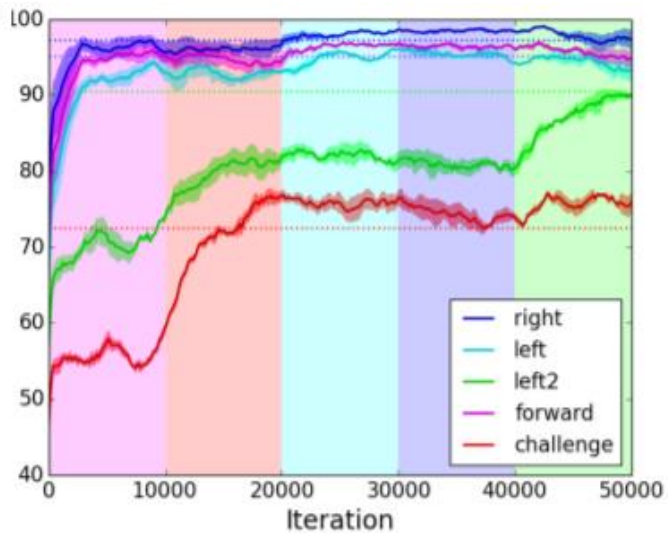
(a) Surprise



(b) Reward

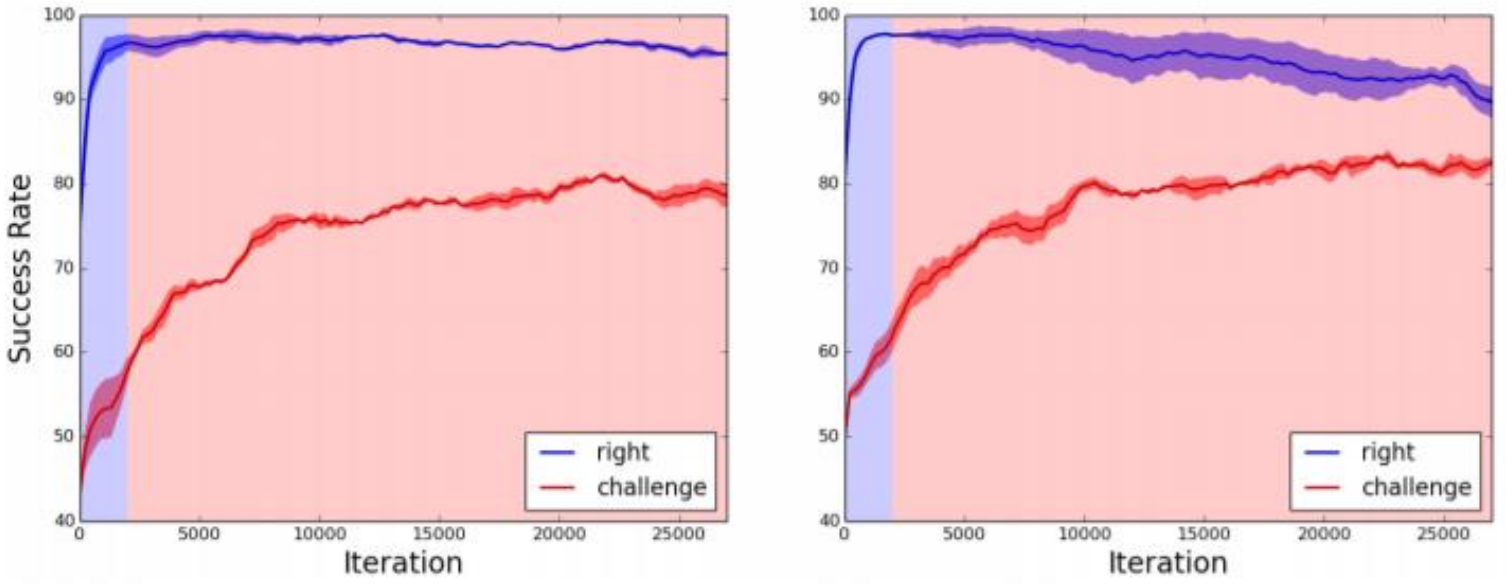


(c) Coverage maximization



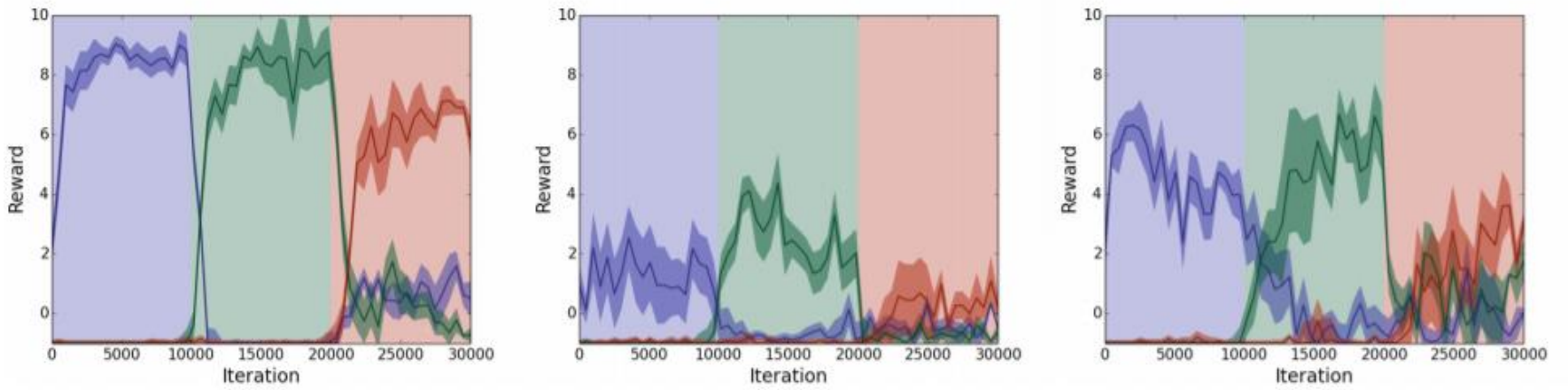
(d) Distribution matching

Unbalanced training

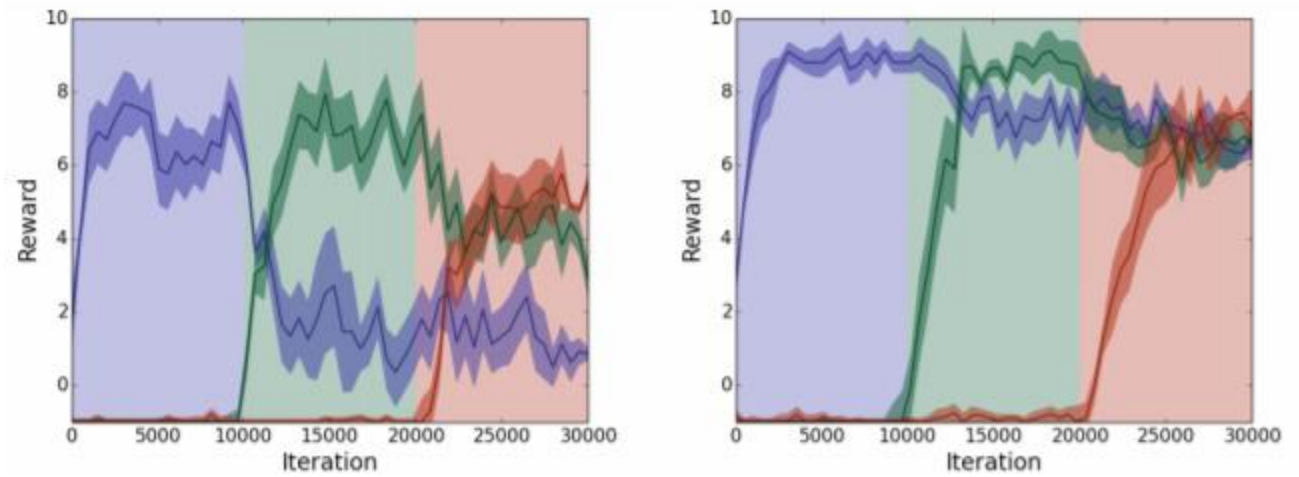


(a) Coverage Maximization (b) Distribution Matching

Grid World

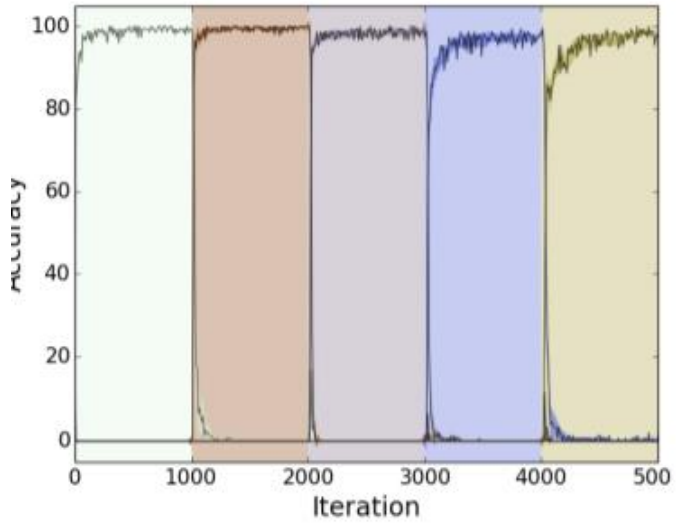


(a) No Selection (b) Suprise (c) Reward

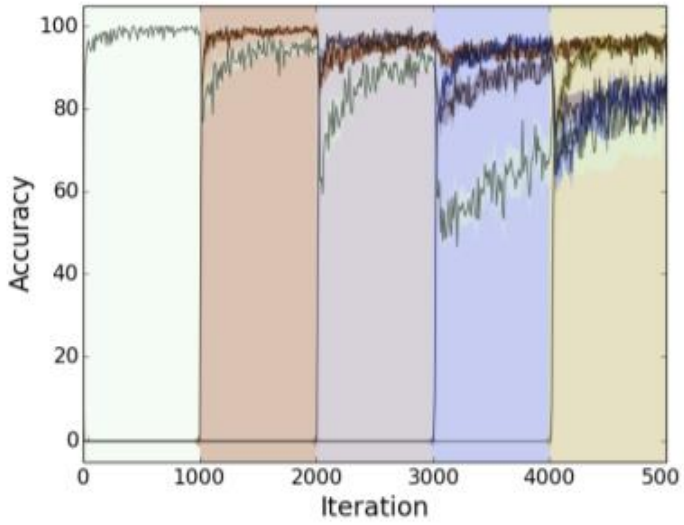


(d) Coverage Max. (e) Distribution Matching

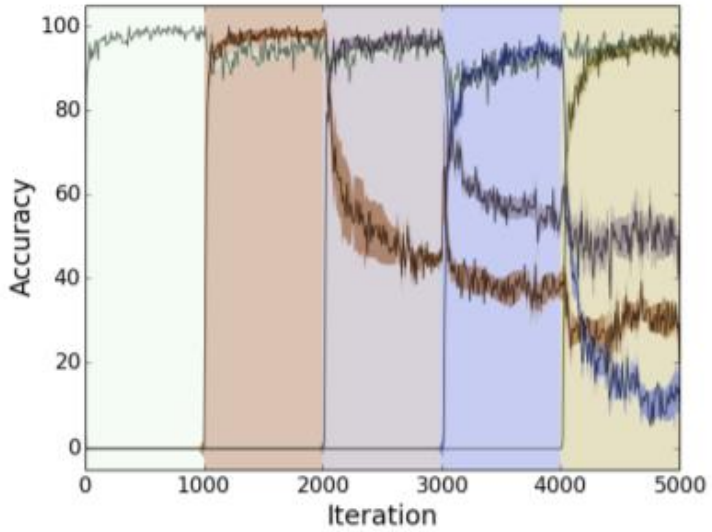
MNIST



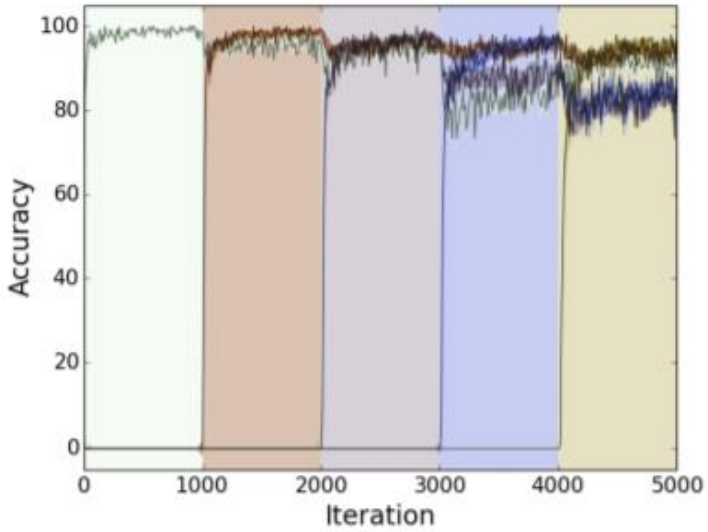
(a) No Selection



(b) Suprise



(c) Coverage Maximization



(d) Distribution Matching