AIL 722: Reinforcement Learning

Lecture 26: Replay buffer and Target Network

Raunak Bhattacharyya

Recap & Today's Outline

- Serial correlation
- Experience replay
- Online Q-learning with replay buffer
- Target network
- Double Learning

Back to online Q-Learning

1. Take some action a_i and obtain (s_i, a_i, s_i', r_i)

2. $\phi \longleftarrow \phi - \alpha \cdot \frac{dQ_\phi}{d\phi}(s_i, a_i) \cdot \left(Q_\phi(s_i, a_i) - [r(s_i, a_i) + \gamma \cdot \max_{a_i'} Q(s_i', a_i')] \right)$

Why did correlation not matter in tabular Q-learning?

Q-Learning with Experience Replay

1. Collect dataset $\{(s_i, a_i, r_i, s'_i)\}$ using some policy, add to $\mathcal{B}$

2. Sample a batch (s_i, a_i, r_i, s'_i) i.i.d. from $\mathcal{B}$

K times

3. $\phi \longleftarrow \phi - \alpha \sum_i \cdot \frac{dQ_\phi}{d\phi}(s_i, a_i) \cdot \left(Q_\phi(s_i, a_i) - [r(s_i, a_i) + \gamma \cdot \max_{a'_i} Q_\phi(s'_i, a'_i)] \right)$

Note the sum

On Experience Replay

- Biologically inspired: Experiences in memory
- First time by Long Ji Lin paper: 1992
- Resurrected by DQN paper: 2015
  - With much larger buffer size
  - With off-policy algorithm
  - With neural value functions
  - With target nets

Machine Learning, 8, 293–321 (1992)

Self-Improving Reactive Agents Based On Reinforcement Learning, Planning and Teaching

LONG-JI LIN ljl@cs.cmu.edu
School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

LETTER
doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

Problem: Moving Target

3. Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| Q_\phi(s_i, a_i) - y_i \right\|^2$

- Fit the NN params to convergence

2. Set $y_i \longleftarrow r(s_i, a_i) + \gamma . \max_{a'_i} Q_\phi(s'_i, a'_i)$

- Target y has changed

Online Q-Learning

1. Take some action a_i and obtain (s_i, a_i, s_i', r_i)

2. $y_i = r(s_i, a_i) + \gamma \cdot \max_{a_i'} Q(s_i', a_i')$

3. $\phi \longleftarrow \phi - \alpha \cdot \frac{dQ_\phi}{d\phi}(s_i, a_i) \cdot \left(Q_\phi(s_i, a_i) - y_i \right)$

Replay Buffer and Target Network

1. Save target network parameters: $\phi \longleftarrow \phi'$

2. Collect dataset $\{(s_i, a_i, r_i, s_i')\}$ using some policy, add to $\mathcal{B}$

N times

3. Sample a batch (s_i, a_i, r_i, s_i') i.i.d. from $\mathcal{B}$

K times

Note the target Q

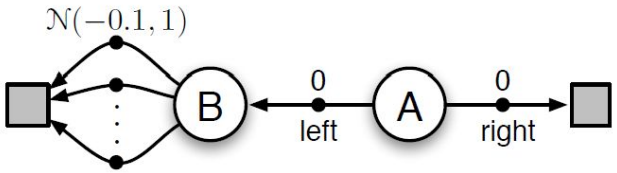
4. $\phi \longleftarrow \phi - \alpha \sum_i \cdot \frac{dQ_\phi}{d\phi}(s_i, a_i) \cdot \left(Q_\phi(s_i, a_i) - [r(s_i, a_i) + \gamma \cdot \max_{a_i'} Q_{\phi'}(s_i', a_i')] \right)$

Unified Framework: Interacting Processes

What were the interacting processes in GPI?

Can we identify similar processes here?

Problem: Overestimation



- Episodes always start in A
- Right transitions to terminal state and terminates
- Left transitions to B with reward 0
- Many possible actions from B
- All lead to termination
- Reward is drawn from N(-0.1,1)

Section 6.7, Reinforcement Learning: An Introduction, Sutton & Barto

Summary & Announcements

- Summary
  - Serial correlation
    - Replay buffer
  - Moving target
    - Target Network
  - Unified Process
  - Overestimation
    - Double Learning

- Announcements
  - If project (instead of assignment 3)
    - Proposal deadline: TBA
    - Project deadline will be same as assignment 3 deadline
    - Proposal (1 page excl. refs):
      - What is the problem?
      - What are its challenges?
      - What has been done before?
      - What do you plan to do?
  - Weightage
  - Paper selection