Learning to communicate

Hugo Bérard, Tom Bosc

04/13/2017

Introduction

- How does cooperation emerge in a POMDP with several agents and shared rewards?
- Settings: Several agents, POMDP, same reward for all agents, a communication channel.
- How does direct RL and indirect RL methods compare on these kind of tasks? (recurrent DQN vs recurrent PG)
- How does sharing the model affect convergence speed?

Deep Q-Learning

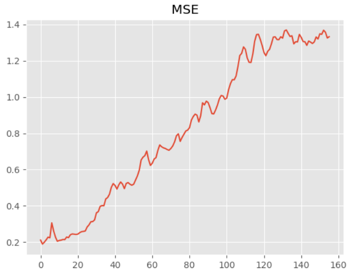
- **Experience Replay**: We store all the transitions (s_t, a_t, s_{t+1}, r_t) in a memory, and we sampled a random batch for updating.
- **Target Network**: We use a second network $Q_{\hat{\theta}}$, to compute the update:

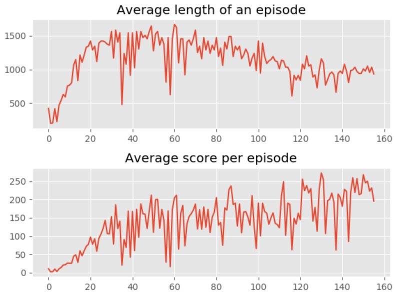
$$\theta_{t+1} = \theta_t + \alpha(r + \max_a Q_{\hat{\theta}}(s_{t+1}, a) - Q_{\theta_t}(s_t, a_t))\nabla_{\theta_t} Q_{\theta_t}(s_t, a_t)$$

where $\hat{\theta}$ is updated to θ every 10 000 iterations.

Experiment on Breakout

We trained on the Atari game *Breakout*.





Deep Distributed Recurrent Q-Networks

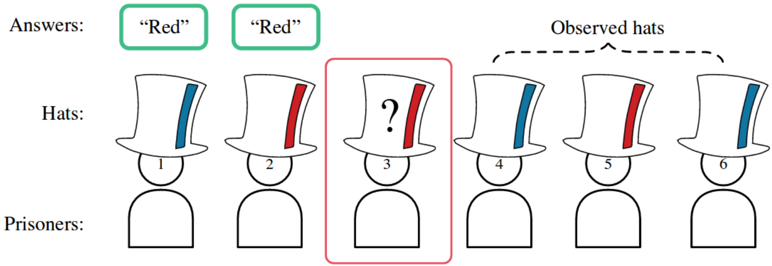
Extension of DQN to POMDP with multi agents.

- **"Naive" solution**: Use a Recurrent Layer and one network per agent.
- **Deep Distributed Recurrent Q-Networks (DDRQN)**: Share the parameters between agents.

3 Tricks:

- Last-action input
- Inter-agent weight sharing
- Disabling Experience Replay

The Hats Riddle



Recurrent Policy Gradients

- Direct RL method for POMDP.
- **Experience Replay** is extended to entire trajectories $h = (o_1, a_1, r_1, o_2, a_2, r_2, ..., o_T, a_T, r_T)$.
- Gradient update on parameters theta $\nabla_\theta J \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{T} \nabla_\theta log \pi(a_t | h_t^n)(R_t^n - b)$
- $\pi(a_t | h_t^n)$ is implemented as an RNN (LSTM), R_t^n is the empirical return at time t and b is a baseline
- Choice of baseline is crucial. Average return or use a separate RNN conditioned on the sequence of observations and actions.
- Discounting in episodic tasks introduces bias but reduces variance.

References

- *Solving Deep Memory POMDPs with Recurrent Policy Gradients, Wierstra et al. 2007*
- *Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." Nature 518.7540 (2015): 529-533.*
- *Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep Reinforcement Learning with Double Q-Learning." AAAI. 2016.*
- *Hausknecht, Matthew, and Peter Stone. "Deep recurrent q-learning for partially observable mdps." arXiv preprint arXiv:1507.06527 (2015).*
- *Foerster, Jakob N., et al. "Learning to communicate to solve riddles with deep distributed recurrent q-networks." arXiv preprint arXiv:1602.02672 (2016).*