Stanford University

Generalization of QMDP-Net on Atari Games

Xiaobai Ma, Zhenkai Wang, Sheng Li

WN18 CS230 Final Project

{maxiaoba, zackwang, lisheng} @stanford.edu

Problem

Partially Observable Markov Decision Process (POMDP) is a model for sequential decision making problems under state and transition uncertainties. QMDP is a traditional algorithm for solving simple POMDPs. The QMDP-net [1] is a neural network that embeds the QMDP [2] algorithm in its network architecture. It combines the advantages of model-free learning and model-based solution structure. Based on the work of [1], we generalize the QMDP-net structure to continuous state space problems and train it with reinforcement learning method.

Training and Testing Environment

Openai Gym [3] provides a rich pool of environments for reinforcement learning. One of the most famous types is the Atari games. Atari games are a series of simple 2D video games developed by Atari, Inc. As a RL environment, its state is fully described by a ram of 128 bits. Although it is a discrete state space, the total state number is so large (2^128) that we could almost treat it as a continuous state space environment. To add the partial observability, we add a random mask to the ram such that certain number of random bits are set to 0 and use it as the observation to the RL agent. For the same reason, we could treat the observation space as continuous. The action space for Atari games is discrete. We test our network structure on 3 interesting Atari games (carnival, space invaders, and star gunner) for their rewards are relatively dense.

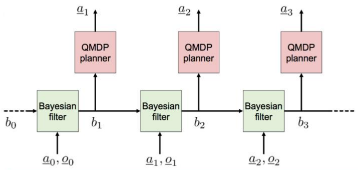
Carnival-ram-v0

SpaceInvaders-ram-v0

StarGunner-ram-v0

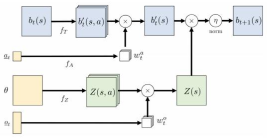
Neural Network Architecture

QMDP-net Architecture



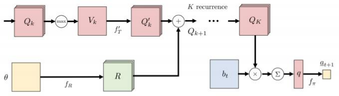
The QMDP can be decomposed as a belief filter and a planner which are replaced by neural network layers in QMDP-net. The belief state works as the hidden state in a recurrent network.

Bayesian Filter Module



QMDP uses Bayesian filter to update the belief state. In QMDP-net, this procedure is mimicked by a convolutional network. We replace the convolution layer by a fully-connected layer to generalize the original 2D state representation to 1D abstract states. Although the state and observation representation is discrete in the network, but with soft-indexing, the network is able to work on continuous state and observation spaces.

QMDP Planner Module



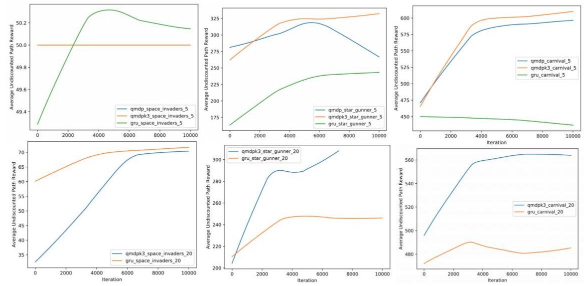
The QMDP-net use a loop of sum and max pool operations to mimic the value iteration loop in QMDP. We make the initial value and reward (θ) as trainable variables to improve the efficiency and performance.

Training Algorithm

We use Vanilla Policy Gradient [4] to train the network which is an on-policy optimization method with the objective function:

$$J = \frac{1}{Batch\ Size} \sum_{(s,a,ad,s')\in paths} ad \cdot \Pr(s,a,s' \mid \pi_\theta)$$

Results



Three different network structures are compared. The baseline is the GRU-net. The second one is the QMDP-net with random constant b0 (initial belief states), v0 (initial values), and θ (environment parameters) and performing value iteration for k = 30 times at each forward propagation. The third one is the QMDP-net using trainable b0, v0, and θ and k = 3. Making b0, v0, and θ trainable allows the network to accumulate model knowledge. Thus even with a smaller k, which means faster forward and backward propagation, the network could actually perform better.

Discussion and Future Work

- QMDP-net generally outperforms the baseline.
- QMDP-net is generalizable to various environment.
- Trying state-of-the-art policy optimization methods like Proximal Policy Optimization, and Asynchronous Advantage Actor-Critic for further improve the performance.
- Using Fast Informed Bound instead of QMDP to estimate the state value which gives a tighter upper bound of approximation

References:

[1] Karkus, P.; Hsu, D.; and Lee, W. S. 2017. Qmdp-net: Deep learning for planning under partial observability. *arXiv preprint arXiv:1703.06692*.

[2] Cassandra, A. R., and Kael- bling, L. P. 2016. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995: Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, July 9-12 1995*, 362. Morgan Kaufmann.

[3] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

[4] Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057– 1063.