Ising Networks for Deep Hierarchical Reinforcement Learning

1 Introduction

2 Related Work

2.1 Hierarchical Reinforcement Learning

2.1.1 Temporal Abstraction

Various methods have devised hierarchical techniques to abstract the temporal correlation in MDPs [12, 13, 18, 19]. The Hierarchical Deep Q-Network (hDQN) [12] proposes an integration of temporal abstraction and intrinsic motivation in order to solve long-horizon problems. hDQN demonstrates improved long-horizon optimality at the cost of manual feature engineering of the states and reward function. This is addressed by making use of deep successor networks [13] which extract relevant goal embeddings for the agent and direct its behavior towards the goal. A more sophisticated temporal scheme consists of learning abstraction of hierarchies based on expert demonstrations and regularized latent space constraints [20]. However, sub-policies rely on memory-based controllers for learning fine-grained abstraction and often cripple the lower levels of the hierarchy once the task has been accomplished. This leads to limited transfer of skills between policies and hence, results in sub-optimal convergence [6]. Transfer of skills between different policies operating in sub-MDPs is extensively studied using exploration [11]. [17] presents MAVEN, a hierarchical architecture for variational exploration in the case of multi-agent settings. Policies carry out exploration in the latent space using variational inference. MAVEN demonstrates improved coordination between agents in the long-horizon. However, policy behavior distributed across agents which forces sub-policies to rely on other agent policies for optimal convergence. [4] proposes the trajectory autoencoder which enables the agent to exhibit self-consistent behavior and tackle sparse rewards in the long-horizon. Long-horizon suitability can further be improved by jointly training all policies and utilize the skills learned individually [14]. Such kind of sub-policy adaptation stabilizes the behavior of agent and provides efficient training of hierarchies at all levels.

Abstraction may be carried out among states [2] or using a different medium such as language [10] which allows the agent to solve temporally-extended tasks in a diverse manner. [5] presents the MAXQ algorithm which proposes an efficient method for training state-based abstractions in the form of hierarchies. Q-values corresponding to each lower-level task are aggregated at the higher-level policy nodes to yield a modular framework for abstraction. While the MAXQ demonstrates suitability for low-dimensional tasks, its performance cannot be assessed in high-dimensional state and action spaces due to the larger number of abstractions produced by the framework. [1] tackles the problem of high-dimensional and long-horizon learning by presenting a theoretical perspective of abstracting states using transitive behavior which helps in optimal computations. Such techniques have motivated active research in the form of modified abstraction algorithms for performance improvement and schemes leading to richer abstract representations [16].

2.1.2 Option-based Hierarchies

Hierarchies can be constructed on the basis of skills or possible options of policies which the agent may choose from during its interactions in the environment [19, 3]. [3] presents the Options-Critic Architecture which is a collection of various policy options. Agent can select from the collection of sub-policies at various timesteps to yield a policy which is robust and optimal in the long-horizon. The provision of options enables the agent to acquire a wide variety of skills during exploration and obtain a policy which makes best use of these skills. Furthermore, the architecture incorporates learning of termination conditions of a particular option which is essential for switching option policies [9]. The architecture is extended to the Double Actor-Critic architecture which comprises of two parallel MDPs [21]. A dual architecture enables the learning of intra-option policies with termination conditions of each option. While, option-based learning provides diversity in skills and control of temporal extension to the agent, frequent fluctuations between options results in instability in updates and hence, sample-inefficient learning [15]. [15] addresses the instability in training by combining the options framework with off-policy maximum entropy reinforcement learning [8, 7]. Maximization of entropy allows the agent to effectively explore all options and consistently select the optimal sub-policies based on selected options.

2.1.3 Hierarchical Control

2.2 Ising Models

3 The Ising Model

4 Intuition for Spin Values

5 Implementation Details

References

[1] D. Abel, D. Arumugam, L. Lehnert, and M. Littman. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*, 2018.

[2] D. Andre and S. J. Russell. State abstraction for programmable reinforcement learning agents. In *AAAI/IAAI*, 2002.

[3] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture, 2016.

[4] J. D. Co-Reyes, Y. Liu, A. Gupta, B. Eysenbach, P. Abbeel, and S. Levine. Self-consistent trajectory autoencoder: Hierarchical reinforcement learning with trajectory embeddings. *arXiv preprint arXiv:1806.02813*, 2018.

[5] T. G. Dietterich. State abstraction in maxq hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, 2000.

[6] C. Florensa, Y. Duan, and P. Abbeel. Stochastic neural networks for hierarchical reinforcement learning, 2017.

[7] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies, 2017.

[8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

[9] A. Harutyunyan, W. Dabney, D. Borsa, N. Heess, R. Munos, and D. Precup. The termination critic, 2019.

[10] Y. Jiang, S. S. Gu, K. P. Murphy, and C. Finn. Language as an abstraction for hierarchical deep reinforcement learning. In *Advances in Neural Information Processing Systems 32*. 2019.

[11] N. K. Jong, T. Hester, and P. Stone. The utility of temporal abstraction in reinforcement learning. In *AAMAS (1)*, 2008.

[12] T. D. Kulkarni, K. R. Narasimhan, A. Saeedi, and J. B. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation, 2016.

[13] T. D. Kulkarni, A. Saeedi, S. Gautam, and S. J. Gershman. Deep successor reinforcement learning, 2016.

[14] A. Li, C. Florensa, I. Clavera, and P. Abbeel. Sub-policy adaptation for hierarchical reinforcement learning. In *International Conference on Learning Representations*, 2020.

[15] C. Li, X. Ma, C. Zhang, J. Yang, L. Xia, and Q. Zhao. Soac: The soft option actor-critic architecture, 2020.

[16] S. Lin and R. Wright. Evolutionary tile coding: An automated state abstraction algorithm for reinforcement learning. In *Proceedings of the 8th AAAI Conference on Abstraction, Reformulation, and Approximation*, 2010.

[17] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson. Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems 32*. 2019.

[18] D. Precup. Temporal abstraction in reinforcement learning, 2000.

[19] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. 2018.

[20] W. Wang, Y. Hu, and S. Scherer. Learning temporal abstraction with information-theoretic constraints for hierarchical reinforcement learning, 2020.

[21] S. Zhang and S. Whiteson. Dac: The double actor-critic architecture for learning options, 2019.