Multiagent Learning and Coordination with Clustered Deep Q-Network

Extended Abstract

Simon Pageaud
Université de Lyon - Université Claude Bernard Lyon 1
LIRIS CNRS UMR 5205
Lyon, France
simon.pageaud@liris.cnrs.fr

Véronique Deslandres
Université de Lyon - Université Claude Bernard Lyon 1
LIRIS CNRS UMR 5205
Lyon, France
veronique.deslandres@liris.cnrs.fr

Vassilissa Lehoux
NAVER LABS Europe
Meylan, France
vassilissa.lehouxd@naverlabs.com

Salima Hassas
Université de Lyon - Université Claude Bernard Lyon 1
LIRIS CNRS UMR 5205
Lyon, France
salima.hassas@liris.cnrs.fr

ABSTRACT

Existing decentralized learning methods entail scalability issues due to the number of agents involved. Independent Q-Learning approach proposes that each agent learns its own action-values. One drawback of this method is that the non-stationarity introduced by Independent Q-Learning limits the use of experience replay memory, needed in deep reinforcement learning methods such as Deep Q-Network. This paper presents a multiagent, multi-level solution named Clustered Deep Q-Network (CDQN) to overcome this issue.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent systems**; **Reinforcement learning**; **Multi-agent reinforcement learning**;

KEYWORDS

Multiagent Reinforcement Learning; Deep Reinforcement Learning

ACM Reference Format:
Simon Pageaud, Véronique Deslandres, Vassilissa Lehoux, and Salima Hassas. 2019. Multiagent Learning and Coordination with Clustered Deep Q-Network. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Multiagent approaches overcome the scalability issue of single agent settings by using *decentralized policies* where agents choose their actions based on local action-state history. Centralized training of decentralized policies is a standard approach of Multiagent Reinforcement Learning [5] and Deep Reinforcement Learning [3] (DRL). In decentralized learning, such as Independent Q-Learning approaches [8, 9], each agent learns its own q-values based on its state-actions and considers other agents as part of the environment. It introduces a non-stationarity that reduces the use of experience replay needed for Deep Q-Network [4] in DRL. Decentralized policies face with another challenge: the multiagent credit assignment [2], where joint actions usually generate only global rewards, making it difficult for each agent to deduce its own contribution.

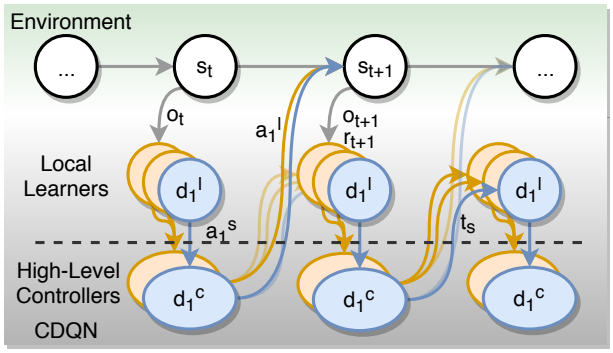
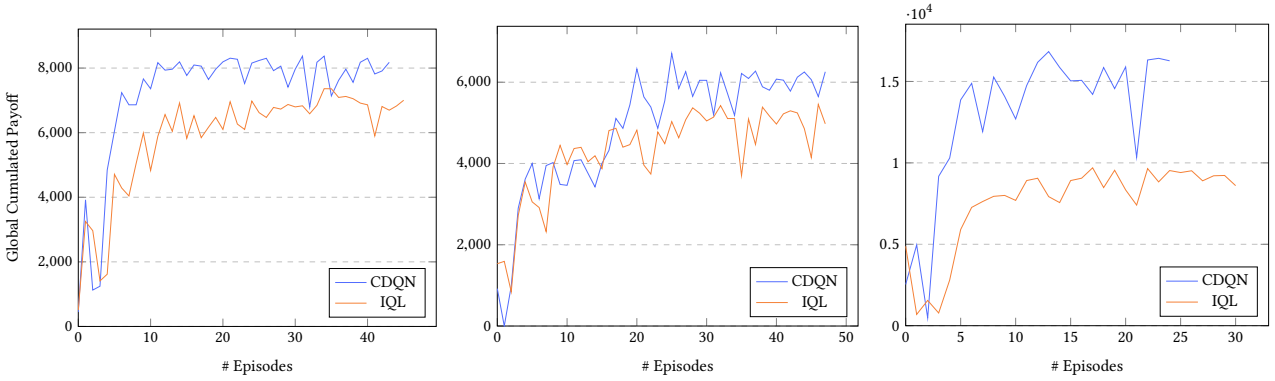


Figure 1: Three local learners are split into clusters blue and orange. A local learner d_i^l has an observation o_t from the environment and suggests an action a_i^s to its control agent d_j^c. Using a voting system, d_j^c chooses the action a_j^l applied by d_i^l. At the next iteration, d_i^l evaluates its payoff to get a reward and d_j^c updates the trust score t_s of d_i^l based on the cluster total payoff evolution.

2 CLUSTERED DEEP Q-NETWORK

We propose the Clustered Deep Q-Network architecture (CDQN) to answer both challenges, experience replay in non-stationary environments and credit assignment problem, with a hierarchical approach where high-level agents manage clusters of low-level learning agents and efficiently coordinate them to improve urban policies (figure 1). A fully cooperative multiagent task can be described as a Dec-POMDP [1] where agents must collaborate to maximize the sum of the joint rewards they receive over multiple iterations. A factored Dec-POMDP [6] represents states as a vector

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.



(a) Fusion (b) Separation (c) Fusion & Separation

Figure 2: Global cumulated reward for IQL and CDQN on three different scenarios.

of state variables and the reward function is the sum of local reward functions. The CDQN extends the Factored Dec-POMDP with the addition of a multi-level approach. The low level contains local learners with partial observation of their environment. The high level contains control agents that manage clusters of local learners, performing the same action to the environment at a time step. This aims at reducing non-stationarity by decreasing the number of different actions applied each time on the environment. Local learner actions are scenario-related and suggested to their cluster. Within a cluster, the most suggested one is applied. Local learners have an individual reward based on modifications of the payoff between two iterations and control agents assign a trust score to each local learner of their cluster based on specific rules. Control agents manage clusters using two actions: *separation* which splits one cluster into two new clusters that distribute local learners based on their trust score and *fusion* that merges two clusters together based on a sequence of similar applied actions. We show that some improvements on independent DQN agents along with two independent and decentralized levels of autonomous agents are able to tackle the problem of limited communication in multiagent partially observable settings and allow the use of both experience replay and individual reward assignment.

3 EXPERIMENTATIONS

The main framework for the environment is based on the SmartGov model [7] to produce a multi-level simulation for urban policy regulation, where agents with different personalities interact in real-world environments based on Open Street Map data. The CDQN model is built on top of the simulator with local learners perceiving factored states and suggesting scenario related actions to be applied by control agents. The efficiency of control actions is evaluated on a set of real world pricing problems in a city through three experiments with the objective of maximizing the global cumulative payoff and to minimize the number of clusters. Two scenarios describe the behavior of clusters when possible actions are performed independently, one scenario when they are performed together, and CDQN is compared with the IQL in each case (figure 2). The fusion scenario relies on equivalent sequence identification and has one control agent per districts with commuters with the same personality. The separation scenario uses one control agent managing two districts each with commuters with different personalities. The objective is to identify the local learners associated with each district using trust score assignment and separate them in two clusters, increasing the global cumulative payoff. The last scenario combines both actions with two control agents that both have two districts with commuters with different personalities. The objective is to identify each district and increase the cumulative payoff by regrouping the local learners with similar commuters personalities together. Results show how a combination of both control actions leads to increase the global cumulative payoff and efficiently manage clusters in deep multiagent reinforcement settings using Independent Q-Networks and experience replay.

4 CONCLUSION

The proposed model is composed of a high-level population of agents that use a custom trust score assignment to manage clusters of low-level agents. Low-level agents learn an action-value function using individual local reward and efficient experience replay. The multi-level, multiagent setting allows coordination even with no communication and no interaction or feedback from other low-level agents. Efficient management of clusters of local learners through high-level agents helps increase local and global payoff. A simulation based mechanism constructs groups of agents at both levels through fusion and separation actions and their corresponding Q-value functions at high levels. Using experiments on a pricing policy problem, we show that combining the use of trust scores and individual local rewards enables efficient learning and coordination between low-level agents. We compare the results of DQN with an Independent Q-Learning to demonstrate how we can overcome non-stationarity in decentralized learning.

ACKNOWLEDGMENT

This project was funded by the French Region Auvergne-Rhône-Alpes (ARC7 contract). This project is a collaborative effort between LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information) in Lyon (France) and the formerly Xerox Research Center Europe, now NAVER LABS Europe in Meylan (France).

REFERENCES

[1] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research* 27, 4 (2002), 819–840.

[2] Yu-Han Chang, Tracey Ho, and Leslie Pack Kaelbling. 2004. All learning is local: Multi-agent learning in global reward games. *Advances in neural information processing systems* (2004), 807–814.

[3] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate to Solve Riddles with Deep Distributed Recurrent Q-Networks. *arXiv preprint arXiv:1602.02672* (2016).

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-Level Control Through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529.

[5] Frans A. Oliehoek, Matthijs T.J. Spaan, Shimon Whiteson, and Nikos Vlassis. 2008. Exploiting Locality of Interaction in Factored Dec-POMDPs. *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems* (2008).

[6] Frans A. Oliehoek, Shimon Whiteson, and Matthijs T.J. Spaan. 2013. Approximate Solutions for Factored Dec-POMDPs with Many Agents. *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems* (May 2013).

[7] Simon Pageaud, Véronique Deslandres, Vassilissa Lehoux, and Salima Hassas. 2017. Co-Construction of Adaptive Public Policies using SmartGov. *29th International Conference on Tools with Artificial Intelligence* (2017).

[8] Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations.* Cambridge University Press.

[9] Ming Tan. 1993. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. *Proceedings of the tenth international conference on machine learning* (1993), 330–337.