Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls and Thore Graepel
DeepMind Technologies, London, United Kingdom
sunehag@google.com

ABSTRACT

We study the problem of cooperative multi-agent reinforcement learning with a single joint reward signal. This class of learning problems is difficult because of the often large combined action and observation spaces. In the fully centralized and decentralized approaches, we find the problem of spurious rewards and a phenomenon we call the "lazy agent" problem, which arises due to partial observability. We address these problems by training individual agents with a novel value-decomposition network architecture, which learns to decompose the team value function into agent-wise value functions.

KEYWORDS

reinforcement learning, DQN, Q-learning, collaborative, multi-agent, value-decomposition, neural networks

1 INTRODUCTION

We consider the cooperative multi-agent reinforcement learning (MARL) problem [4, 17, 25], in which a system of several learning agents must jointly optimize a single reward signal – the *team reward* – accumulated over time. Each agent has access to its own ("local") observations and is responsible for choosing actions from its own action set. Coordinated MARL problems emerge in applications such as coordinating self-driving vehicles and/or traffic signals in a transportation system, or optimizing the productivity of a factory comprised of many interacting components. More generally, with AI agents becoming more pervasive, they will have to learn to coordinate to achieve common goals.

Although in practice some applications may require local autonomy, in principle the cooperative MARL problem could be treated using a *centralized* approach, reducing the problem to single-agent reinforcement learning (RL) over the concatenated observations and combinatorial action space. We show that the centralized approach consistently fails on relatively simple cooperative MARL problems in practice. For several tasks, the centralised approach fails by learning inefficient policies with only one agent active and the other being "lazy". This happens when one agent learns a useful policy, but a second agent is then discouraged from learning because its exploration would hinder the first agent and lead to worse team reward.[1]

An alternative approach is to train *independent learners* to optimize for the team reward. In general, each agent is then faced with a non-stationary learning problem because the dynamics of its environment effectively changes as teammates change their behaviours through learning [14]. Furthermore, since from a single agent's perspective the environment is only partially observed, agents may receive spurious reward signals that originate from their teammates' (unobserved) behaviour. Because of this inability to explain its own observed rewards naive independent RL is often unsuccessful: for example Claus and Boutilier [5] show that independent Q-learners cannot distinguish teammates' exploration from stochasticity in the environment, and fail to solve even an apparently trivial, 2-agent, stateless, 3×3-action problem and the general Dec-POMDP problem is known to be intractable [3, 16].

We introduce a novel **learned additive value-decomposition** approach over individual agents. Implicitly, the value-decomposition network aims to learn an optimal linear value-decomposition from the team reward signal, by back-propagating the total Q gradient through deep neural networks representing the individual component value functions. The implicit value function learned by each agent depends only on local observations, and so is more easily learned. Our solution also ameliorates the coordination problem of independent learning highlighted in Claus and Boutilier [5] because it effectively learns in a centralised fashion at training time, while agents can be deployed individually.

Further, in the context of the introduced agent, we evaluate weight sharing, role information and communication channels as additional enhancements that have recently been reported to improve sample complexity and memory requirements [8, 11, 22]. However, our main comparison is between three kinds of architectures: value-decomposition across individual agents, independent learners and centralized approaches. We investigate and benchmark these techniques applied to a range of new interesting two-player coordination domains. We find that value-decomposition is a much better performing approach than centralization or fully independent learners, and that when combined with the additional enhancements, results in an agent that consistently outperforms centralized and independent learners by a big margin. In addition, both value-decomposition by itself as well as any of the other evaluated agents using a value-decomposition layer, performed better

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. Andre, S. Koenig (eds.), July 2018, Stockholm, Sweden

https://doi.org/doi

[1] For example, imagine training a 2-player soccer team using RL with the number of goals serving as the team reward signal. Suppose one player has become a better scorer than the other. When the worse player takes a shot the outcome is on average much worse, and the weaker player learns to avoid taking shots [11].

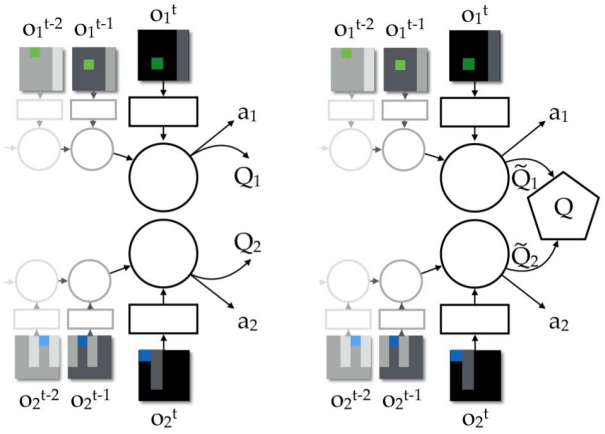

Figure 1: Independent agents (left) and value-decomposition architecture (right); In both architectures, observations enter the networks of two agents, pass through the low-level linear layer to the recurrent layer, and then a dueling layer produces individual Q-values. In the value-decomposition architecture these "values" are summed to a joint Q-function for training, while actions are produced independently.

than individual learners and centralization. Please see Sunehag et al. [23] for more detailed descriptions than can be provided here.

1.1 Other Related Work

Schneider et al. [21] and Russell and Zimdars [20] optimize the sum of individual rewards by learning value functions from those individual rewards. Our approach works with only a team reward, and *learns* the value-decomposition autonomously from experience. [10] and the max-plus algorithm [13, 26] also rely on specified individual rewards. *Difference rewards* [24] measures the impact of an agent's action on the team reward, but comes with practical difficulties [1, 6, 18]. A recent Deep-RL approach to difference rewards [9] learns a centralized value function (critic), which we here show is hard without a simplifying architecture like value-decomposition or the new further generalization Rashid et al. [19]. Other approaches are Babes et al. [2], Devlin et al. [7], HolmesParker et al. [12].

2 ARCHITECTURES FOR DEEP COOP-MARL

Building on purely independent DQN-style agents (see left in Figure 1), we add enhancements to overcome the identified issues with the MARL problem. Our main contribution of value-decomposition is illustrated by the network on the right in Figure 1.

The main assumption we make and exploit is that the joint action-value function for the system can be additively decomposed into value functions across agents,

$$Q((h^1, h^2, ..., h^d), (a^1, a^2, ..., a^d)) \approx \sum_{i=1}^{d} \tilde{Q}_i(h^i, a^i)$$

where the $\tilde{Q}_i$ depends only on each agent's local observations. We learn $\tilde{Q}_i$ by backpropagating gradients from the Q-learning rule using the joint reward through the summation, i.e. $\tilde{Q}_i$ is learned implicitly rather than from any reward specific to agent i, and we do not impose constraints that the $\tilde{Q}_i$ are action-value functions for any specific reward. Although learning requires some centralization, the learned agents can be deployed independently, since each agent acting greedily with respect to its local value $\tilde{Q}_i$ is equivalent to a central arbiter choosing joint actions by maximizing the sum $\sum_{i=1}^{d} \tilde{Q}_i$.

Figure 2: Heat map showing each agent's final performance, averaged over the last 5,000 episodes of 50,000 and across ten runs, normalized by the best architecture per task. The agents are ordered according to average over the domains, which can be seen in the right most column. Value-decomposition (IL+V+ ...) strongly outperform Individual Learners (IL) and Centralization (C Comb(inatorial)). C Lin(ear), centralized with a value-decomposition layer, is much better than C Comb and IL but clearly worse than more individual value-decomposition (IL+V+...).

3 EXPERIMENTS

We introduce a range of two-player domains, and experimentally evaluate the introduced value-decomposition agents with different levels of enhancements, evaluating each addition in a logical sequence. We use two centralized agents as baselines, one of which is introduced here again relying on learned value-decomposition, as well as an individual agent learning directly from the joint reward signal. We perform this set of experiments on the same form of two dimensional maze environments used by Leibo et al. [15], but with different tasks featuring more challenging coordination needs. Agents have a small $3 \times 5 \times 5$ observation window, the first dimension being an RGB channel, the second and third are the maze dimensions, and each agent sees a box 2 squares either side and 4 squares forwards, see Figure 1.

3.1 Results and Conclusions

We compare nine approaches on seven tasks (see Sunehag et al. [23]). The very clear conclusion is that architectures based on value-decomposition, with any combination of other techniques or none, outperforms the centralized approaches and individual learners.

REFERENCES

[1] A. K. Agogino and K. Tumer. 2008. Analyzing and Visualizing Multiagent Rewards in Dynamic and Stochastic Environments. *Journal of Autonomous Agents and Multi-Agent Systems* 17, 2 (2008), 320–338.

[2] Monica Babes, Enrique Munoz de Cote, and Michael L. Littman. 2008. Social reward shaping in the prisoner's dilemma. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Volume 3*. 1389–1392.

[3] Daniel S. Bernstein, Shlomo Zilberstein, and Neil Immerman. 2000. The Complexity of Decentralized Control of Markov Decision Processes. In *UAI '00: Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence, Stanford University, Stanford, California, USA, June 30 - July 3, 2000*. 32–37.

[4] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions of Systems, Man, and Cybernetics Part C: Applications and Reviews* 38, 2 (2008).

[5] Caroline Claus and Craig Boutilier. 1998. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA*. 746–752.

[6] M. Colby, T. Duchow-Pressley, J. J. Chung, and K. Tumer. 2016. Local Approximation of Difference Evaluation Functions. In *Proceedings of the Fifteenth International Joint Conference on Autonomous Agents and Multiagent Systems*. Singapore.

[7] S. Devlin, L. Yliniemi, D. Kudenko, and K. Tumer. 2014. Potential-Based Difference Rewards for Multiagent Reinforcement Learning. In *Proceedings of the Thirteenth International Joint Conference on Autonomous Agents and Multiagent Systems*.

[8] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 2137–2145.

[9] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2017. Counterfactual Multi-Agent Policy Gradients. *CoRR* abs/1705.08926 (2017). arXiv:1705.08926 http://arxiv.org/abs/1705.08926

[10] Carlos Guestrin, Michail G. Lagoudakis, and Ronald Parr. 2002. Coordinated Reinforcement Learning. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML '02)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 227–234. http://dl.acm.org/citation.cfm?id=645531.757784

[11] Matthew John Hausknecht. 2016. *Cooperation and Communication in Multiagent Deep Reinforcement Learning*. Ph.D. Dissertation. The University of Texas at Austin.

[12] C. HolmesParker, A. Agogino, and K. Tumer. 2016. Combining Reward Shaping and Hierarchies for Scaling to Large Multiagent Systems. *Knowledge Engineering Review* (2016). to appear.

[13] Lior Kuyer, Shimon Whiteson, Bram Bakker, and Nikos A. Vlassis. 2008. Multiagent Reinforcement Learning for Urban Traffic Control Using Coordination Graphs. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I*. 656–671.

[14] Guillaume J. Laurent, Laëtitia Matignon, and N. Le Fort-Piat. 2011. The World of Independent Learners is Not Markovian. *Int. J. Know.-Based Intell. Eng. Syst.* 15, 1 (2011), 55–64.

[15] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*. Sao Paulo, Brazil.

[16] Frans A. Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs*. Springer.

[17] Liviu Panait and Sean Luke. 2005. Cooperative Multi-Agent Learning: The State of the Art. *Autonomous Agents and Multi-Agent Systems* 11, 3 (2005), 387–434.

[18] S. Proper and K. Tumer. 2012. Modeling Difference Rewards for Multiagent Learning (Extended Abstract). In *Proceedings of the Eleventh International Joint Conference on Autonomous Agents and Multiagent Systems*. Valencia, Spain.

[19] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witta, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. [n. d.]. MIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. ([n. d.]).

[20] Stuart J. Russell and Andrew Zimdars. 2003. Q-Decomposition for Reinforcement Learning Agents. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*. 656–663.

[21] Jeff G. Schneider, Weng-Keen Wong, Andrew W. Moore, and Martin A. Riedmiller. 1999. Distributed Value Functions. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*. 371–378.

[22] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning Multiagent Communication with Backpropagation. *CoRR* abs/1605.07736 (2016). http://arxiv.org/abs/1605.07736

[23] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. *CoRR* abs/1706.05296 (2017).

[24] K. Tumer and D. Wolpert. 2004. A Survey of Collectives. In *Collectives and the Design of Complex Systems*, K. Tumer and D. Wolpert (Eds.). Springer, 1–42.

[25] Karl Tuyls and Gerhard Weiss. 2012. Multiagent Learning: Basics, Challenges, and Prospects. *AI Magazine* 33, 3 (2012), 41–52.

[26] Elise van der Pol and Frans A. Oliehoek. 2016. Coordinated Deep Reinforcement Learners for Traffic Light Control. *NIPS Workshop on Learning, Inference and Control of Multi-Agent Systems* (2016).