Networked Multi-Agent Reinforcement Learning with Emergent Communication

Extended Abstract

Shubham Gupta*
Indian Institute of Science
Bangalore, India
shubhamg@iisc.ac.in

Rishi Hazra*
Indian Institute of Science
Bangalore, India
rishihazra@iisc.ac.in

Ambedkar Dukkipati
Indian Institute of Science
Bangalore, India
ambedkar@iisc.ac.in

ABSTRACT

We develop a Multi-Agent Reinforcement Learning (MARL) method that finds approximately optimal policies for cooperative agents that co-exist in an environment. Central to achieving this is how the agents learn to communicate with each other. Can they together develop a language while learning to perform a common task? We formulate and study a MARL problem where cooperative agents are connected via a fixed underlying network. These agents communicate along the edges of this network by exchanging discrete symbols. However, the semantics of these symbols are not predefined and have to be learned during the training process. We propose a method for training these agents using emergent communication. We demonstrate the applicability of the proposed framework by applying it to the problem of managing traffic controllers, where we achieve state-of-the-art performance (as compared to several strong baselines) and perform a detailed analysis of the emergent communication.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**; *Cooperation and coordination*;

KEYWORDS

multi-agent reinforcement learning; emergent communication; traffic

ACM Reference Format:
Shubham Gupta, Rishi Hazra, and Ambedkar Dukkipati. 2020. Networked Multi-Agent Reinforcement Learning with Emergent Communication. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

We consider a multi-agent setting where a certain number of cooperative agents co-exist in an environment that can only be partially observed by each of them. Further, we assume that these agents are connected via a fixed network topology and that they can communicate with their immediate neighbors along the edges of this network to achieve cooperation. The objective of agents is to learn a protocol to communicate with each other to cooperatively maximize the rewards provided to them by the environment.

Note that: **(i)** Agents *learn* to communicate with each other which enables global cooperation by supplementing the information content of agents' local observations. We use discrete communication to facilitate the analysis of the emergent language. **(ii)** Agents only communicate along the edges of a fixed underlying network. This a more practical scenario and it allows us to study the relationship between emergent communication and the network topology. Together, these two features distinguish our work from existing approaches which are either very stylized [1, 7], or don't perform an in-depth analysis of communication [5, 6], or don't consider a network topology [4, 12, 13].

Many real-world problems can be cast in this framework. We consider the problem of intelligently managing traffic in a city. The nodes in the network (i.e., the agents) correspond to traffic controllers and the edges correspond to roads. The controllers must act cooperatively to ensure a smooth flow of traffic by maximizing an appropriate notion of reward. Our main contributions are: formulation of the MARL problem with networked agents and emergent communication; demonstration of the effectiveness of the proposed approach using traffic management as a case study and; most importantly, analysis of the emergent communication to investigate: **(i)** utility of communication; **(ii)** grounding of language; and **(iii)** interplay between network topology and emergent language.

2 PROPOSED MARKOV GAMES WITH EMERGENT COMMUNICATION

A Markov game [11], specified by the tuple $(\mathcal{S}, \{\mathcal{O}_i, \mathcal{A}_i, r_i\}_{i=1}^N, \mathcal{T}, \gamma)$, models an environment with N intelligent agents. Here, $\mathcal{S}$ is the state-space, $\mathcal{O}_i$, $\mathcal{A}_i$ and r_i are the observation-space, action-space and reward function respectively for agent i, and $\mathcal{T}$ is the transition function. At time t, we denote the state by $\mathbf{s}^{(t)} \in \mathcal{S}$, the observation made by agent i by $\mathbf{o}_i^{(t)} \in \mathcal{O}_i$, and the action taken by it as $\mathbf{a}_i^{(t)} \in \mathcal{A}_i$. The goal of all agents is to find their respective optimal policies $\pi_i : \mathcal{O}_i \to \Delta(\mathcal{A}_i)$ that maximize the expected long term reward, $\mathcal{R}_i = \mathrm{E}_{\boldsymbol{\pi}}[\sum_t \gamma^t r_i^{(t)}]$. Here $r_i^{(t)}$ is the reward received by agent i at time-step t and $\gamma \in (0, 1]$ is the discount factor. We model the problem as a Markov game with two additional assumptions: **(i)** let $\mathcal{V} = \{1, 2, \ldots, N\}$ be the set of all agents, we assume that the agents are connected to each other via an underlying network whose edge set is given by $\mathcal{E}$; and **(ii)** agents can communicate with their immediate neighbors in the underlying network. To communicate, at each step, agents broadcast a message which is received by the neighbors at the next time-step. The observation

*Equal Contribution

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

made by each agent is augmented to consider the messages received from all of its neighbors.

In the traffic management problem, as mentioned earlier, nodes represent agents and edges represent roads. We used a traffic simulator known as Simulation of Urban MObility (SUMO) [9] to simulate the traffic flow. Each agent observes an image representation of the traffic junction obtained by cropping a square patch of size 140px centered at that agent from the simulation window. Actions correspond to valid configurations of traffic lights [10]. Rewards depend on factors like queue length, waiting time of vehicles, number of vehicles executing emergency deceleration and so on.

Learning Policies with Communication: The policy of each agent is composed of three modules: **(i)** *Observation encoder:* It encodes the observation of an agent into a form suitable for the other two modules. **(ii)** *Communicator:* It takes the encoded observation as input and produces a discrete message to be broadcasted $\mathbf{m}_i^{(t)} \in \{0,1\}^d$ as output. Here d is the number of bits in the binary vector $\mathbf{m}_i^{(t)}$. We use the straight through Gumbel-Softmax trick [8] to retain differentiability. Aside from sending messages, communicator is also responsible for processing the received messages. **(iii)** *Action selector:* It takes the encoded observation and processed received messages as input and produces a probability distribution over actions in $\mathcal{A}_i$ as output. We parameterize these modules using neural networks and train them using policy gradient [14].

3 EXPERIMENTAL RESULTS

Through our experiments, we wish to establish the following claims: **(i)** the proposed approach outperforms baseline methods, **(ii)** communication is useful as the agents are exchanging meaningful information, **(iii)** emergent communication is grounded in the actions taken by the agents, and **(iv)** network topology plays an important role in determining the nature of emergent communication.

Comparison with baselines: We compare our approach with the following baselines: **(i)** Fixed-time control: The agents periodically switch between actions in a round-robin fashion after every five steps. **(ii)** Self-Organizing Traffic Light control (SOTL): SOTL [3] switches between actions when the queue length at an adjoining lane exceeds a predefined threshold. **(iii)** Deep-Q Learning (DQN): Agents are training independently and each agent has its own deep Q-network. **(iv)** IntelliLight: [16]; **(v)** Fixed communication protocol: Agents share all the parameters needed to compute rewards with their neighbors directly. It can be seen that our method outperforms all baseline approaches (Fig. 1).

Utility of communication: We provide a qualitative analysis of the communication: **(i)** We modified the setup presented to mask all communication messages in the system with an all zeros vector (*blank message*). We observed that post convergence rewards were lower as compared to the original setting (the difference was $\approx$ **85**, also see Fig. 1). **(ii)** We define an agent to be visually impaired (or blind) if it does not use its local observation while taking an action. Note that a visually impaired agent can still receive messages from its neighbors. We observed that, after convergence, the rewards were same as the rewards obtained in the original setup. This indicates that the visually impaired agent learned to receive necessary information from its neighbors through communication. To test this hypothesis further, we made two neighboring agents blind so they

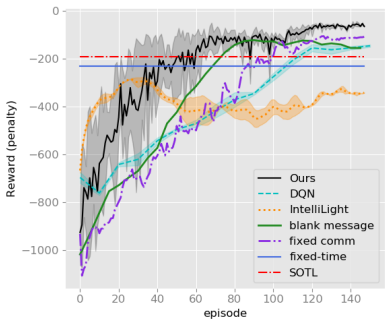

Figure 1: Comparison of our method with the baselines. Figure shows mean and standard deviation over five independent runs.

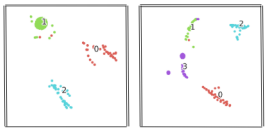

Figure 2: t-SNE plot for neighbors of a three-way (left) and four-way (right) junction. Points have been colored based on the action with which the symbol was highly correlated. Agents on three and four-way junctions can take three and four actions respectively.

can no longer supplement each other's missing information using communication and observed that the performance decreased.

Grounding in communication: To establish groundedness of communication, we constructed a Pointwise Mutual Information (PMI) [2] matrix for each pair of agents. The rows of this matrix correspond to the actions of one agent (say i) and the columns correspond to the discrete symbol sent by the other agent (say j). If two columns of this matrix are similar, it indicates that the corresponding symbols spoken by j have a similar effect on the actions taken by i. Fig. 2 shows the t-SNE [15] plot of columns of the PMI matrix for a pair of agents on a three-way and a four-way junction. It can be seen that symbols cluster together and, moreover, these clusters align with the actions taken by i.

Effect of network topology: We obtain a tf-idf matrix where rows correspond to agents and columns correspond to the words in the vocabulary. Similarity between two rows implies similarity in the language spoken by the agents. We observed that agents that broadcast to a common neighbor tend to be have similar rows in the tf-idf matrix.

4 CONCLUSION

In this paper, we formulated a networked multi-agent reinforcement learning problem where cooperative agents communicate with each other using an emergent language. As future work, we intend to extend the proposed setup to address the following **(i)** scalability of our setup; **(ii)** robustness of our setup to randomness in the underlying network. It would be interesting to try out a continuous communication version of our setup.

REFERENCES

[1] Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent Communication through Negotiation. In *International Conference on Learning Representations*.

[2] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* 16, 1 (1990), 22–29.

[3] Seung-Bae Cools, Carlos Gershenson, and Bart D'Hooghe. 2013. *Self-Organizing Traffic Lights: A Realistic Simulation*. Springer, 45–55.

[4] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. TarMAC: Targeted Multi-Agent Communication. *Proceedings of the 36th International Conference on Machine Learning, PMLR* 97, 1538–1546.

[5] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2137–2145.

[6] Shubham Gupta and Ambedkar Dukkipati. 2019. On Voting Strategies and Emergent Communication. *CoRR* abs/1902.06897 (2019).

[7] Serhii Havrylov and Ivan Titov. 2017. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2149–2159.

[8] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkE3y85ee

[9] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker-Walz. 2012. Recent Development and Applications of SUMO - Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements* 5, 3&4 (2012), 128–138.

[10] J. S. Linkenheld, Rahim F. Benekohal, and J. H. Garrett. 1992. Knowledge-Based System for Design of Signalized Intersections. *Transportation engineering journal of ASCE* 118 (1992), 241–257.

[11] Michael L. Littman. 1994. Markov Games As a Framework for Multi-agent Reinforcement Learning. In *International Conference on Machine Learning*. 157–163.

[12] Igor Mordatch and Pieter Abbeel. 2018. Emergence of Grounded Compositional Language in Multi-Agent Populations. In *Thirty-Second AAAI Conference on Artificial Intelligence*. 1495–1502. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17007

[13] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning Multiagent Communication with Backpropagation. In *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2244–2252.

[14] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*. Vol. 13. 1057–1063.

[15] L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[16] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. 2018. IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining (KDD '18)*. 2496–2505.