Multi-Agent Communication with Multi-Modal Information Fusion

Han Wang
(wanghan@xidian.edu.cn)
School of Computer Science and Technology,
Xidian University, 710126 Xi'an, China

Yufeng Xie
(yufeng.xie@stu.xidian.edu.cn)
School of Computer Science and Technology,
Xidian University, 710126 Xi'an, China

Bingcheng He
(22031212248@stu.xidian.edu.cn)
School of Computer Science and Technology,
Xidian University, 710126 Xi'an, China

Qingshan Li*
(qshli@mail.xidian.edu.cn)
School of Computer Science and Technology,
Xidian University, 710126 Xi'an, China

Abstract

Many recent works in the field of multi-agent reinforcement learning via communication focus on learning what messages to send, when to send, and whom to address such messages. Those works indicate that communication is useful for higher cumulative reward or task success. However, one important limitation is that most of them ignore the importance of enforcing agents' ability to understand the received information. In this paper, we notice that observation and communication signals are from separate information sources. Thus, we enhance the communicating agents with the capability to integrate crucial information from different sources. Specifically, we propose a multi-modal communication method, which modulates agents' observation and communication signals as different modalities and performs multi-modal fusion to allow knowledge to transfer across different modalities. We evaluate the proposed method on a diverse set of cooperative multi-agent tasks with several state-of-the-art algorithms. Results demonstrate the effectiveness of our method in incorporating knowledge and gaining a deeper understanding from various information sources.

Keywords: Multi-agent reinforcement learning; Multi-agent communication; Multi-modal fusion

Introduction

Humans can make effective use of communication through language to share information and coordinate on a common goal. With this motivation, there has been remarkable progress in communication based deep multi-agent reinforcement learning (MARL). Communication allows agents to share understanding on local information, which is assumed as a supplement to the unobservant environment (Hernandez-Leal, Kartal, & Taylor, 2019). Thus, agents may cooperate on shared tasks efficiently with the assistance of communication.

Several distinct lines of research on multi-agent communication can be discerned. At first, researchers focus on learning what messages to send, which enables agents to collect crucial local information to share. At each time step, each agent may make better decision on what message to send to all other agents through the communication channel. Researchers have revealed that communication could help agents form solid knowledge of cooperative strategies and reach a common goal when conducting MARL (J. N. Foerster, Assael, de Freitas, & Whiteson, 2016a, 2016b).

As we move towards complex environments, it is common that improving communication efficiency is imperative. Method in (Wang et al., 2020) allows agents to deliver compact messages. In (Singh, Jain, & Sukhbaatar, 2019), each agent has the right to choose when to communicate rather than the 'one-size-fits-all' approach of broadcasting messages to all other agents. The method in (Das et al., 2019) enables agents to deliver messages to specific recipients directly. Therefore, prior works in the field of multi-agent communication focus on the generation and transmission of communication messages (Lowe, Foerster, Boureau, Pineau, & Dauphin, 2019), but lack investigation on how to understand them.

Actually, human's experience of the world is multi-modal (Baltrušaitis, Ahuja, & Morency, 2018; Gao, Li, Chen, & Zhang, 2020; Roy & Pentland, 2002): we see objects, hear sounds, etc. The ability to interpret and reason about multi-modal messages is one of the hallmarks of human intelligence. In MARL scenarios, agents see local observations from the environment and hear communication messages from other agents. Most of the existing methods are based on the assumption that agents have a solid understanding of their input information. However, the ability of agents to process information from multiple modalities is still unclear.

In this paper, we improve agents' understanding of the communication messages and local observations. First, we view the two sources of inputs as distinct 2 modalities. We leverage on multi-modal fusion to merge information under multiple uni-modal representations, shown in Figure 1. Thus, agents may get further understanding of multi-modal inputs to improve learning efficiency.

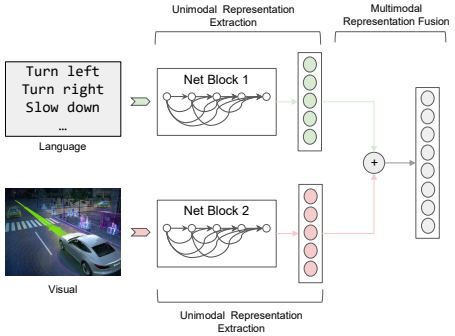


Figure 1: **An example of multi-modal fusion.** Agents may integrate the inputs from different sources and types into a global space in which both inter-modality and cross-modality can be represented in a uniform manner.

*Corresponding author.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

In particular, separate but related modalities are construed as minimal representations effective at predicting the rewards. Such representations are then used to generate actions and communication messages to send. Inspired by the multi-modal representation learning methods, we propose a model for enhancing the understanding of multi-agent communication messages with local observations, named Multi-Modal Multi-Agent Communication (MM-MAC). Technically, MM-MAC utilizes the multi-modal factorization model (MFM) (Tsai, Liang, Zadeh, Morency, & Salakhutdinov, 2019) with conditional independence assumptions over multi-modal discriminative factors (rewards) and modality-specific generative factors (inputs from 2 sources). By approximating inference in MFM, latent representations on 2-modality inputs can be obtained, which enhances agents' understanding of complex inputs.

We evaluate and analyze the proposed solution in different experimental environments based on several recent state-of-the-art approaches. The experimental results show that

- MM-MAC can enable effective multi-source information fusion within each agent, and lead to faster convergence as compared with various baselines;
- MM-MAC can be easily combined with state-of-the-art approaches, thus extending its applicability to more challenging environments, and leading to significant improvements in performance and sample complexity.

Related Work

This work is principled on prior works in deep multi-agent reinforcement learning (MARL), the centralized training and decentralized execution (CTDE) (Lowe et al., 2017) paradigm, and focuses on improving communication efficiency in MARL (K. Zhang, Yang, & Basar, 2019).

DIAL (J. N. Foerster et al., 2016b) is proposed as a simple differential communication module that allows the gradient to flow between agents for training, and enables agents to collect crucial local information to share. (J. Foerster, Farquhar, Afouras, Nardelli, & Whiteson, 2018; Jaques et al., 2019) encourage agents to assess the quality of communication messages by capturing causal rewards through counterfactual reasoning, and generate messages with causal effects.

One line of methods on improving messages is to schedule the policy to specify who/whom and when to communicate. TarMAC (Das et al., 2019) allows each individual agent to actively select certain agents to communicate. Methods in (Das et al., 2019; Jiang & Lu, 2018; Singh et al., 2019) introduce various gating mechanisms to determine the sub-groups of communication agents. Another line of methods addressing the limited bandwidth problem is also investigated, such as downsizing the communication group via a scheduler.

However, all scheduling methods suffer from content redundancy, which is unsustainable under bandwidth limitations. Even if only a single pair of agents is allowed to communicate, a large message may fail to be conveyed due to the limited bandwidth. In addition, scheduling methods with gating mechanisms are inflexible because they introduce manual configuration, such as the predefined size of a communication group (Kim et al., 2019), or a handcrafted threshold for muting agents (Jiang & Lu, 2018).

To the best of our knowledge, none of the existing work in the field of MARL tries to improve communication efficiency by enhancing agents' understanding of their various inputs. In this work, MM-MAC enables the agent to efficiently extract information based on its own observation and communication messages from other agents.

Problem Setting

A communicative multi-agent reinforcement learning (Littman, 1994) task can be modeled as a decentralized partially observable Markov decision process (Dec-POMDP, $< N, S, A, r, P, O, \gamma >$), with an extra element M denoting the communication space. Each agent i gets local observation $c_i = [m_1, \ldots, m_n] \in M$ from all agents. Based on the observation o_i and messages c_i, agent i is supposed to generate action a_i and a new message m_i' to deliver. All agents share the same reward as a function of the states and actions $r_i : S \times A \to \mathbb{R}$. The goal of a communicative MARL task is to learn a communication protocol $\pi_i^m(m_i \mid o_i, c_i)$ and policy $\pi_i^a(a_i \mid o_i, c_i)$ so as to jointly maximize the expected discounted return $J_i = \mathbb{E}_{\pi^a}[\sum_{t=0}^{\infty} \gamma^t r_i^t(s,a)]$.

Such communication MARL tasks can be instantiated in centralized training and decentralized execution (CTDE) framework. During training, all agents are allowed to access the states and actions of other agents for the centralized critic, while decentralized execution only permits individual states and received messages without any extra information from other agents. Under the CTDE paradigm, a centralized critic guides the optimization of individual agent policies during training. The critic takes as input predicted actions $\{a_1^t, \ldots, a_N^t\}$ and observations $\{o_1^t, \ldots, o_N^t\}$ from all agents to estimate the joint action-value $\hat{Q}_t$ at each time step t. Each agent's policy π_i^a is parameterized by θ, and updated by

$$\begin{aligned} \nabla_\theta J(\pi_i^a) &= \mathbb{E}_{\pi_i^a, \pi_i^m}\left[\nabla_\theta \log \pi_i^a(a_i \mid s_i)\, \hat{Q}_t^i(s_t, a_t)\right] \\ &\approx \mathbb{E}_{\pi_i^a, \pi_i^m}\left[\nabla_\theta \hat{Q}_t^i\left(o_t^1, \ldots, o_t^N, c_t^1, \ldots, c_t^N, a_t\right)\right] \end{aligned} \tag{1}$$

In this work, we focus on information fusion from observation o_i and communication messages c_i of agent i, which aims at improving learning performance in complex tasks. From a multi-modal perspective, the observation o_i and communication messages c_i can be viewed as different modalities. Thus, agents can integrate information extracted from different uni-modal sources into a single compact multi-modal representation. We propose to learn joint embeddings of o_i and c_i to leverage the complementarity of multi-modal information to exploit the comprehensive semantics.

Multi-Modal Multi-Agent Communication

This section describes the proposed method Multi-Modal Multi-Agent Communication (MM-MAC). Suppose that there are N agents jointly performing a cooperative task with policies $\{\pi_1^a, \ldots, \pi_N^a\}$ parameterized by $\{\theta_1, \ldots, \theta_N\}$, and commu-

nication protocols parameterized by $\{\omega_1, \ldots, \omega_N\}$. At every time-step t, each agent i gathers a local observation o_t^i and communication messages c_t^i from others and must select an action a_t^i guided by π_t^i and send a continuous communication message m_t^i followed by π_i^m. Since no agent has access to the complete state of the environment s_t, there is an incentive in communicating to utilize observations and communication messages to recover the unobservant environment.

Multi-Modal Information Fusion

As mentioned above, separate networks tackling local observations and communication messages are responsible for integrating information from 2 sources, and then used to generate integrated latent representations. Inspired by (Tsai et al., 2019), the information fusion can be performed as a Multimodal Factorization Model (MFM). We define $S = \{o_i, c_i\}$ as the multi-modal inputs from 2 modalities, and the reward r as the labels, with joint distribution $P_{S,r} = P(S,r)$. In MFM, mutually independent latent variables $Z = \{Z_r, Z_s\}$ are defined to generate generative factors F_r and F_S, where F_r is provided to generate $\hat{r}$ while $\hat{S}$ generated from F_r and F_S. Thus, the joint distribution $P(\widehat{S}, \hat{r})$ can be factorized as

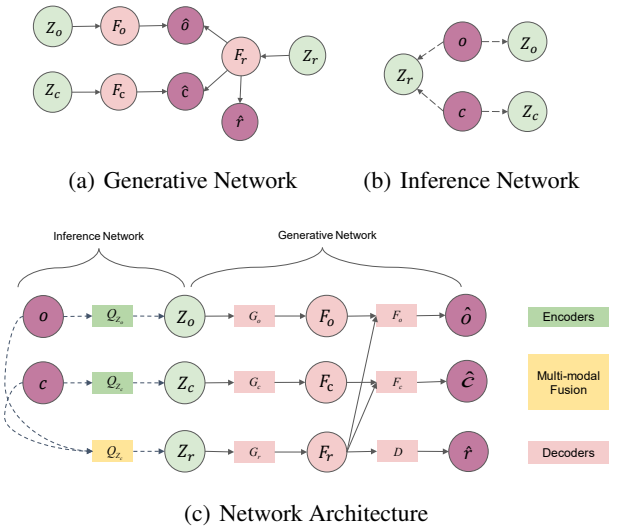
$$
\begin{aligned}
P(\hat{S}, \hat{r}) &= \int_{F,Z} P(\hat{S}, \hat{r} \mid F) P(F \mid Z) P(Z) \mathrm{d}F \ \mathrm{d}Z \\
&= \int_{F,Z} \left(P\left(\hat{r} \mid F_r\right) P\left(\hat{O} \mid F_O\right) P\left(\hat{C} \mid F_C\right)\right) \cdot \\
&\quad \left(P\left(F_r \mid Z_r\right) P\left(F_O \mid Z_O\right) P\left(F_C \mid Z_C\right)\right) \cdot \\
&\quad \left(P\left(Z_r\right) p\left(Z_O\right) P\left(Z_C\right)\right) \mathrm{d}F \ \mathrm{d}Z
\end{aligned} \tag{2}
$$

In such architecture, the fusion net can be viewed as an auto-encoding structure that consists of encoder (inference) and decoder (generative) modules (Baltrušaitis et al., 2018). The encoder module easily samples Z from an approximate posterior, while the decoder module is parameterized according to $P(\hat{S}, \hat{r} \mid Z)$. Posterior inference in Equation (2) may be analytically intractable due to the integration over Z. Therefore, we resort to using an approximate inference distribution $Q(Z \mid S, r)$. For the encoder $Q(Z \mid S, r)$, we maintain a deterministic mapping $Q_{\text{enc}} : S, r \to Z$. For the decoder, the generation process from latent variables $G_{\text{r}} : Z_r \to F_r$, $G_S : Z_S \to F_S$, $D : F_r \to r$ and $F_S : F_r, F_S \to \hat{S}$ can be defined as deterministic functions parameterized by neural networks.

To obtain better latent factor disentanglement and sample generation quality, Wasserstein Auto-encoder (WAE) (S. Zhang et al., 2019) is used for approximating inference, which derives an approximation for the primal form of the Wasserstein distance (W-distance). The joint-distribution on W-distance $W_c\left(P_{S,r}, P_{\hat{S},\hat{r}}\right)$ over multi-modal samples under squared cost $c(S,\hat{S}) = \sum_{t=1}^{T} \left\| S^t - \hat{S}^t \right\|_2^2$ can be derived as

$$
\begin{aligned}
\mathbb{E}_{P_{S,r}} \mathbb{E}_{Q(Z|S,r)} \big[\sum_{i=1}^{M} c_{S_i}(S_i, F_i(G_i(Z_i), G_r(Z_r)) + \\
c_r(r, D(G_r(Z_r)))\big]
\end{aligned} \tag{3}
$$

The prior P_Z is chosen as a centered isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$, which implicitly enforces independence between the latent variables $Z = \{Z_r, Z_s\}$. The neural architecture of MFM is illustrated in Figure 2.



(a) Generative Network (b) Inference Network

(c) Network Architecture

Figure 2: **MFM architecture.** MFM factorizes multi-modal representations into multi-modal discriminative factors F_r and modality-specific generative factors $\{F_o, F_c\}$. Figure 2(a) illustrates generative network with latent variables $\{Z_r, Z_o, Z_c\}$, factors $\{F_r, F_o, F_s\}$, and the reconstructive multi-modal samples $\{\hat{o}, \hat{c}, \hat{r}\}$. MFM defines a joint distribution over multi-modal samples, and by the conditional independence assumptions in the assumed graphical model.

Multi-Modal Communication

Establishing complex collaboration strategies requires efficient communication, which is essentially based on the ability to understand the local observations and received messages to generate suitable actions, as well as multi-round communication. Each agent is modeled as a Dec-POMDP augmented with communication messages. We enhance agents' understanding of communication messages with local observations and propose a Multi-Modal Multi-Agent Communication (MM-MAC) framework. MM-MAC operates under the centralized learning and decentralized execution paradigm where a centralized critic guides the optimization of individual agent policies during training. The critic takes as input predicted actions and internal observations with augmented message representations from all agents to estimate the joint action Q-value at every time-step t.

The communication protocol for agents can be flexibly designed. We follow the scheme of several state-of-the-art algorithms for training the communication protocol and optimizing the policy. For the informative multi-agent communication via the information bottleneck method (IMAC), the scheduler follows the same principle for learning a weight-based mechanism. Variational information bottleneck can be applied in

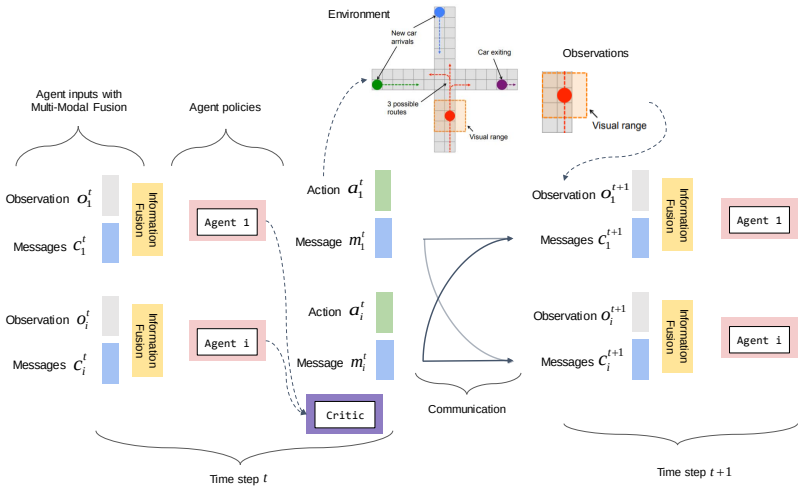


Figure 3: **Overview of MM-MAC.** Left: At every time-step, each agent's policy gets a local observation o_i^t and aggregated message c_i^t as input, and predicts an environment action a_i^t and a communication message m_i^t. Right: Communication between agents can be implemented based on specific mechanisms, such as signature-based soft attention. Each agent broadcasts a message m_i^t. At the next time-step, each receiving agent gets as input a combination of message values.

scheduling for agent i with regularization on the mutual information between the scheduling messages c_i and all agents' messages. For communication protocols trained with the influence reward, we use the influence reward to directly train agents to use an explicit communication channel to get more meaningful and effective collective outcomes. For targeted multi-agent communication, we use a signature-based soft-attention mechanism in the communication structure to enable targeting to ensure the entire communication architecture is differentiable, and message vectors are learnt through back-propagation.

Experiments

We evaluate MM-MAC on a variety of tasks and environments. All our models were trained with a batched synchronous version of the multi-agent Actor-Critic described above, using RMSProp with a learning rate of 1×10^{-4}, batch size 32, discount factor $\gamma = 0.99$, and entropy regularization coefficient 0.01 for agent policies. All results are averaged over 5 independent seeds (unless noted otherwise), and error bars show the standard error of means.

In each domain, the baselines are TarMac (Das et al., 2019), SchedNet (Kim et al., 2019), Causal Influential Communication (Causal) (Jaques et al., 2019) and IMAC (Wang et al., 2020). TarMac permits multi-round communication and uses a signature-based soft-attention mechanism to enable targeting. SchedNet forms a new deep MARL framework with scheduled communications for handling the constraints of limited bandwidth and shared medium access. Causal Communication enhances the agents incentivized to communicate via the social influence reward learn faster, and achieve significantly higher collective reward for the majority of training. IMAC is conducted by an information-theoretic regularization on the mutual information between the messages and input features.

Traffic Junction

In the traffic junction (Sukhbaatar, Szlam, & Fergus, 2016), cars may enter a junction from entry points with a probability p_{arr}. The task has three difficulty levels, varying in the number of possible routes, entry points, and junctions, shown in Table 1. Traffic junction is a common benchmark for testing whether the communication mechanism is working by setting the vision of cars to be 0. The reward is the sum of the waiting time and the penalty for collisions. Therefore, more entry points, vehicles, and routes will increase the probability of the vehicle's collision, which leads to low rewards.

Table 1: Different settings correspond to 3 difficulty levels.

Difficulty	P-arrive		Grid Size	N-total	Arrival Points	Routes Per Entry Point	Two-Way	Junctions
	Start	End						
Easy	0.10	0.20	7	5	2	1	F	1
Medium	0.02	0.05	14	10	4	3	T	1
Hard	0.02	0.05	18	20	8	7	T	4

We evaluate MM-MAC in three difficulty levels followed (Singh et al., 2019), listed in Table 1. The results in Table 2 indicate that a communication model trained with multi-modal information fusion consistently outperforms the baselines in various difficulty levels. We attribute this to the fact that MM-MAC utilizes multi-modal information to enhance the agent's decision-making by providing a comprehensive understanding of the environment.

StarCraftII

MM-MAC and baselines are applied to decentralized StarCraftII micromanagement benchmark to show that MM-MAC can facilitate various multi-agent methods in complex game domains. We use the setup introduced by SMAC (Samvelyan et al., 2019) and consider four combat scenarios.

Table 2: Comparative results on 3 difficulty levels, where success rates are listed. Bold font indicates the best result in each paired method (baselines and baselines with multi-modal information fusion).

Method	Easy	Medium	Hard
TarMac	91.7±3.1	86.3±3.6	74.7±4.7
MM-TarMac	**93.1±3.7**	**88.0±4.1**	**75.6±5.2**
SchedNet	90.6±4.7	87.3±5.2	73.7±5.7
MM-SchedNet	**91.2±4.6**	**88.6±5.3**	**77.9±5.4**
Causal	92.7±4.2	91.0±4.7	78.0±7.6
MM-Causal	**93.4±3.7**	**92.1±4.4**	**79.6±5.1**
IMAC	93.1±3.4	92.3±3.7	81.7±4.4
MM-IMAC	**94.2±3.9**	**92.7±4.1**	**85.6±5.1**

3m and 8m. Both tasks are symmetric battle scenarios, where marines controlled by the learned agents try to beat enemy units controlled by the built-in game AI. Agents will receive some positive (negative) rewards after having enemy (allied) units killed and/or a positive (negative) bonus for winning (losing) the battle.

2s3z. In this scenario, each group consists of 2 stalkers and 3 zealots. They are required to move closely to enemy units to attack. Additionally, stalkers are required to learn kiting to consistently move back in between attacks to keep a distance between themselves and enemy zealots to minimize received damage while maintaining high damage output.

5m_vs_6m. This is an asymmetric battle scenario, in which the friendly side controls 5 marines to compete with the competitor side controls 6 marines. Each agent observes its own states within its field of view and also observes other units' statistics such as health, location, and unit type. Agents can only attack enemies within their shooting range.

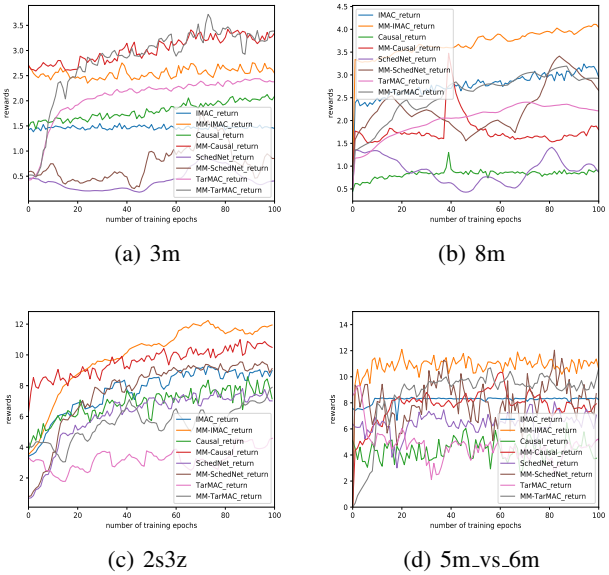
The average performances of 4 algorithms are drawn within each scenario, shown in Figure 4. The results suggest that methods with multi-modal information fusion procession demonstrate learning ability in speed and performance. This prompts us to rethink how to address the lack of information understanding during learning, and this paper works in this direction to foster multiagent communication.

Sequential Social Dilemmas

Sequential Social Dilemmas (SSDs) are spatially and temporally extended multiagent games (Hughes et al., 2018; Jaques et al., 2019). Each agent i receives its own reward $r_i(s_t, a_t)$, which may depend on the actions of other agents. Thus, an individual agent can obtain a higher reward by engaging in defecting, non-cooperative behavior, but the average reward per agent will be higher if all agents cooperate in SSDs. We experiment with Cleanup and Harvest scenarios.

Cleanup and Harvest. In both games, Apple offered rewards but only limited. Agents can exploit other agents for immediate reward but at the expense of long-term collective reward of the group. Agents must coordinate harvesting apples with the behavior of other agents to achieve cooperation.

As shown in Figure 5, we visualize the learning curves

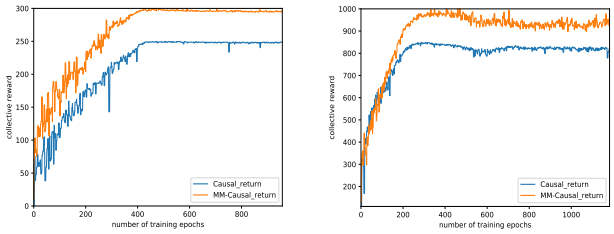


(a) 3m (b) 8m

(c) 2s3z (d) 5m_vs_6m

Figure 4: The average performances of 4 algorithms on 5 independent StarCraftII experiments.

across 5 random episodic runs using Causal Influential Communication (Causal) (Jaques et al., 2019) and Causal communication with multi-modal information fusion. We observe that the agent task performance improves most substantially in Causal communication with the addition of improved understanding of multi-modal inputs.



(a) Cleanup (b) Harvest

Figure 5: Total collective reward obtained by agents trained by causal communication algorithm on 5 independent experiments in Cleanup and Harvest.

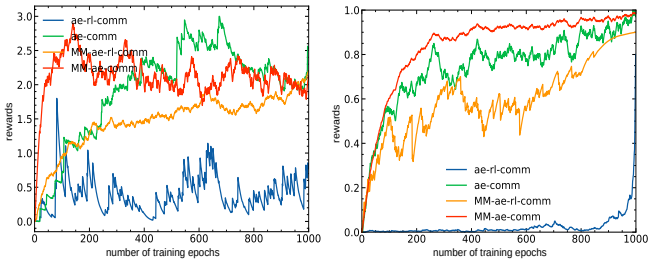
MarlGrid Environments

In (Lin, Huh, Stauffer, Lim, & Isola, 2021), the authors introduce the ae-comm algorithm and two new grid environments: FindGoal and RedBlueDoors, which are adapted from the GridWorld environment. States are set randomly at every episode and are partially observable to the agents. In FindGoal, the goal is to reach the target location colored in green. Each agent receives a reward of 1 when they reach the goal, and an additional reward of 1 when all 3 agents reach the goal within the time horizon. RedBlueDoors is a sequential task. A

reward of 1 is given to both agents if and only if the red door is opened first and then the blue door.

The baselines include: (1) ae-comm, use a standard representation learning by auto-encoding for arriving at a grounded common language, which enables agents to understand and respond to each other's utterances and achieve strong task performance across a variety of multi-agent communication environments; (2) ae-rl-comm, also uses auto-encoding representation but maintains an additional communication network.

We compare the task performance of baselines and baselines with multi-modal information fusion. Figure 6 demonstrates that agents with MM-fusion are able to complete the episode much faster than other agents. The results further suggest that MM-fusion is useful for agents' learning to understand the communication messages.



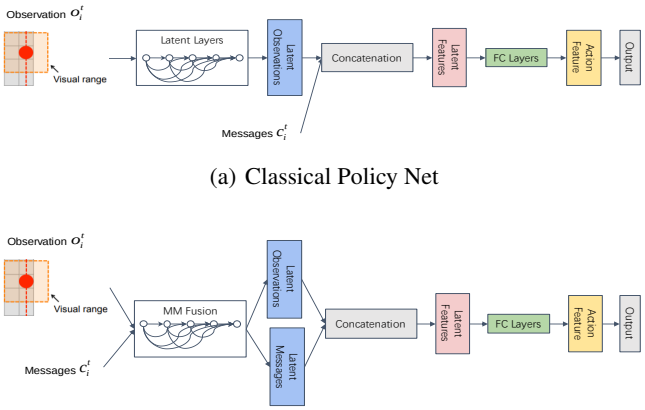
(a) FindGoal (b) RedBlueDoors

Figure 6: Comparison between baselines and MM-fusion. We observed that training policy with multi-modal information fusion improves algorithm performance across all environments.

Similarity of Latent Representations

Recent work has sought to understand the behavior of neural networks by comparing representations between layers and between different trained models. We compare the two neural network representations based on centered kernel alignment (CKA) (Kornblith, Norouzi, Lee, & Hinton, 2019). Specifically, we choose the policy of agent 0 in StarCraftII 3m task and reveal CKA between latent features (colored in pink), 2 fully-connected (FC) layers (colored in green), and the action features (colored in yellow) from the networks in Figure 7.

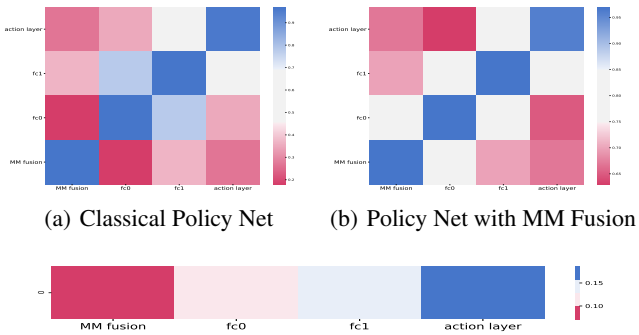
In 3m task, we initialize the architecturally identical networks (latent features, FC-layers, and the action features) in IMAC and MM-IMAC with different random parameters. After training, we first investigate the similarity between the 4 sub-layers both in IMAC and MM-IMAC. As depicted in Figure 8, we find that performing multi-modal information fusion leads to more similar representations between layers. These results demonstrate that after observations and messages information fusion, the latent multi-modal features are more informative, and thus agents can transform these features to generate appropriate actions. In Figure 8, we also show similarity between the same layer of 2 policy nets on randomly sampled trajectories. Overall, layers in 2 different nets are generally less similar to each other, among which the latent features yield a clear difference.



(a) Classical Policy Net

(b) Policy Net with MM Fusion

Figure 7: 2 policy structures. The traditional policy net only extracts information from local observations and concatenates latent observations with messages as the latent input feature. The MM fusion policy net generate both latent observations and messages to create latent input features.



(a) Classical Policy Net (b) Policy Net with MM Fusion

(c) Similarity Between 2 Nets

Figure 8: Linear CKA between layers of policy models on the 3m task with random trajectories.

Conclusion

In this paper, we propose the Multi-modal Communication framework for communicative MARL methods, named MM-MAC. MM-MAC utilizes multi-modal learning techniques to transfer knowledge across modalities. All experiments have demonstrated that MM-MAC improves the agents' understanding of the information from different sources, such as communication messages and local observations. MM-MAC can be utilized in environments with both homogeneous and heterogeneous agents. Our future work will explore extensions of MM-MAC for real-world tasks, where the input information may involve more modalities. We believe that MM-MAC sheds light on the advantages of learning multi-modal representations and may potentially open up new horizons for communicative MARL.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. U21B2015), the Fundamental Research Funds for the Central Universities (Grant Nos. XJSJ23033, YJSJ24015), and the Innovation Fund of Xidian University.

References

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423–443.

Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., & Pineau, J. (2019). Tarmac: Targeted multi-agent communication. In *Proceedings of the 36th international conference on machine learning* (pp. 1538–1546). Long Beach, CA.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. In *Proceedings of the 32nd aaai conference on artificial intelligence* (pp. 2974–2982). New Orleans, LA.

Foerster, J. N., Assael, Y. M., de Freitas, N., & Whiteson, S. (2016a). Learning to communicate to solve riddles with deep distributed recurrent q-networks. *arXiv preprint arXiv:1602.02672*.

Foerster, J. N., Assael, Y. M., de Freitas, N., & Whiteson, S. (2016b). Learning to communicate with deep multi-agent reinforcement learning. In *Annual conference on neural information processing systems* (pp. 2137–2145). Barcelona, Spain.

Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, *32*(5), 829–864.

Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 750–797.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Duéñez-Guzmán, E. A., Castañeda, A. G., … Graepel, T. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In *Annual conference on neural information processing systems* (pp. 3330–3340). Montréal, Canada.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., … De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th international conference on machine learning* (pp. 3040–3049). Long Beach, CA.

Jiang, J., & Lu, Z. (2018). Learning attentional communication for multi-agent cooperation. In *Annual conference on neural information processing systems* (pp. 7265–7275). Montréal, Canada.

Kim, D., Moon, S., Hostallero, D., Kang, W. J., Lee, T., Son, K., & Yi, Y. (2019). Learning to schedule communication in multi-agent reinforcement learning. In *Proceedings of the 7th international conference on learning representations.* New Orleans, LA.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *Proceedings of the 36th international conference on machine learning* (pp. 3519–3529). Long Beach, CA.

Lin, T., Huh, J., Stauffer, C., Lim, S., & Isola, P. (2021). Learning to ground multi-agent communication with autoencoders. In *Annual conference on neural information processing systems* (pp. 15230–15242). Virtual Event.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th international conference on machine learning* (pp. 157–163). New Brunswick, NJ.

Lowe, R., Foerster, J. N., Boureau, Y., Pineau, J., & Dauphin, Y. N. (2019). On the pitfalls of measuring emergent communication. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 693–701). Montreal, Canada.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Annual conference on neural information processing systems* (pp. 6379–6390). Long Beach, CA.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, *26*(1), 113–146.

Samvelyan, M., Rashid, T., de Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G. J., … Whiteson, S. (2019). The starcraft multi-agent challenge. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 2186–2188). Montreal, Canada.

Singh, A., Jain, T., & Sukhbaatar, S. (2019). Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *Proceedings of the 7th international conference on learning representations.* New Orleans, LA.

Sukhbaatar, S., Szlam, A., & Fergus, R. (2016). Learning multiagent communication with backpropagation. In *Annual conference on neural information processing systems* (pp. 2244–2252). Barcelona, Spain.

Tsai, Y. H., Liang, P. P., Zadeh, A., Morency, L., & Salakhutdinov, R. (2019). Learning factorized multimodal representations. In *Proceedings of the 7th international conference on learning representations.* New Orleans, LA.

Wang, R., He, X., Yu, R., Qiu, W., An, B., & Rabinovich, Z. (2020). Learning efficient multi-agent communication: An information bottleneck approach. In *Proceedings of the 37th international conference on machine learning* (pp. 9908–9918). Virtual Event.

Zhang, K., Yang, Z., & Basar, T. (2019). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *CoRR*, *abs/1911.10635*.

Zhang, S., Gao, Y., Jiao, Y., Liu, J., Wang, Y., & Yang, C. (2019). Wasserstein-wasserstein auto-encoders. *CoRR*, *abs/1902.09323*.