Multimodal Reinforcement Learning with Effective State Representation Learning

Extended Abstract

Jinming Ma[1], Yingfeng Chen[2], Feng Wu[1*], Xianpeng Ji[2], and Yu Ding[2]
[1]School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
[2]Netease Fuxi AI Lab, Hangzhou, China
{jinmingm}@mail.ustc.edu.cn,{wufeng02}@ustc.edu.cn,{chenyingfeng1,jixianpeng,dingyu01}@corp.netease.com

ABSTRACT

Many real-world applications require an agent to make robust and deliberate decisions with multimodal information (e.g., robots with multi-sensory inputs). However, it is very challenging to train the agent via reinforcement learning (RL) due to the heterogeneity and dynamic importance of different modalities. Specifically, we observe that these issues make conventional RL methods difficult to learn a useful state representation in the end-to-end training with multimodal information. To address this, we propose a novel multimodal RL approach that can do multimodal alignment and importance enhancement according to their similarity and importance in terms of RL tasks respectively. By doing so, we are able to learn an effective state representation and consequentially improve the RL training process. We test our approach on several multimodal RL domains, showing that it outperforms state-of-the-art methods in terms of learning speed and policy quality.

KEYWORDS

Deep Reinforcement Learning; Multimodal Learning

ACM Reference Format:
Jinming Ma[1], Yingfeng Chen[2], Feng Wu[1*], Xianpeng Ji[2], and Yu Ding[2]. 2022. Multimodal Reinforcement Learning with Effective State Representation Learning: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Deep reinforcement learning has made significant progress recently in many tasks [5, 9, 11, 14, 17, 20], which can train in the end-to-end fashion with raw sensory inputs. Cognitive and psychology studies [18] reveal humans are able to use multiple sources of information (e.g., vision, audio, and tactile) to build a better understanding of the physical world and make decisions. Generally, it is believed that multimodal information is crucial for agents to make robust and deliberate decisions. Multimodal RL becomes an active RL topic and many applications based on it [3, 4, 10, 13, 21] have been successfully developed.

Although multimodal information is indeed beneficial for agents to make decisions, it also brings several challenges to RL algorithms. Firstly, the *heterogeneity* nature of multiple modalities makes it difficult to form a consistent representation in deep neural networks [2]. This issue is especially challenging for multimodal RL because the states (i.e. representation goals) are usually hidden and must be learned implicitly from reward signals instead of directly from supervised labels. Secondly, modalities may play different *importance* for decision making in different situations. Therefore, an agent should be able to dynamically bias towards more informative modalities and enhance the importance of themselves.

Against this background, we propose a novel multimodal RL with effective state representation learning, targeting at the *modal heterogeneity* and *dynamic importance* issues. Respectively, our approach consists of two main modules, i.e., *modality alignment* and *importance enhancement*. By combining these two modules together, we are able to learn an effective state representation, which is the key to the performance of RL given multimodal information.

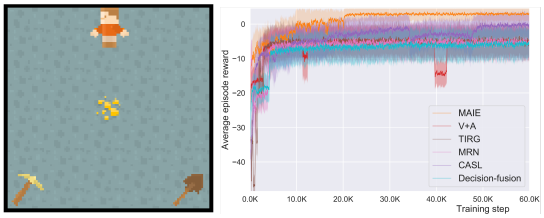
2 THE METHOD

Here, we propose our multimodal RL (named MAIE, which stands for Modality Alignment and Importance Enhancement) with effective state representation learning. As aforementioned, state representation is challenging in multimodal RL due to modal heterogeneity and dynamic importance of different modalities. Respectively, we devise the modality alignment and importance enhancement modules to address these issues. We put them together to learn an effective state representation that can be used by deep RL methods (e.g., A2C). Specifically, we introduce a novel technique to produce a better state vector that can be used by the value or policy network training in an end-to-end manner with multimodal information.

Modality Alignment: To begin with, we use CNN and LSTM to extract high-dimensional and temporal features of each modality. Then, we put them together to learn an effective state representation. Due to modal heterogeneity, different modalities are usually distributed inconsistently by the feature extractors. Therefore, simply concatenating the vectors of all modalities may not be helpful for the RL training. Intuitively, it would be helpful if we can find relationships and correspondences between sub-components of instances from two or more heterogeneous modalities. To achieve this, we use a *similarity measurement* (e.g., Euclidean, cosine, and KL distance), which are commonly used in multimodal machine learning [1], to do multimodal alignment through a loss. Specifically, let f^i and f^j be two vector representations of modalities i and j that need to be aligned in the feature space. Given this, we define the loss function as follow: $\mathcal{L}_{sim}(\phi) = \sum_{i=1}^{m} \sum_{j \neq i} \psi[f^i \mid f^j]$, where $\psi[f^i \mid f^j]$ is a distance function for vectors f^i and f^j.

Multimodal alignment based on similarity is particularly useful when all the modalities are informative and corresponding to the underlying state. Note that RL is a sequential decision-making

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.



(a) (b)

Figure 1: Illustration and training curves for the Mining.

problem. When one of the modalities is totally not informative or even purely noisy, forcing other modalities to align with it will reduce their temporal discrimination and make them less and less sensitive. Therefore, we expect that the feature vectors of identical modalities are discriminative in the time scale. To this end, *temporal discrimination* is introduced for increasing the difference between the feature vectors of each modal at different moments with the loss function defined as: $\mathcal{L}_{td}(\phi) = -\sum_{i=1}^{m}\sum_{t=1}^{T-1}\psi[f_t^i \mid f_{t+1}^i]$. To put the loss functions above together, we have the overall loss function for the state representation learning: $\mathcal{L}_{SRL}(\phi) = c_{sim}\mathcal{L}_{sim}(\phi) + c_{td}\mathcal{L}_{td}(\phi)$, where c_{sim}, c_{td} are scaling constants.

Importance Enhancement: Now, we proceed to our importance enhancement module, which is inspired by work [6]. We first normalize feature vector f^m of each modality m as follow: $\hat{f}^m = (f^m - \mu^m)/\sqrt{\sigma^m + \epsilon}$ where μ^m, σ^m are the mean and variance computed by soft-update. Noting that both the mean and variance are regarded as additional parameters saved in the network when training and retrieved during inference.

Here, we consider the normalized features $\hat{f}^m$ is more informative if the feature deviates further from its mean since it occurs with a lower probability and often provides more information [7]. Based on this property, we use softmax to calculate the importance coefficient λ^m for each modality m as follow: $\lambda_{[l]}^m = e^{|\hat{f}_{[l]}^m|}/\sum_{i=1}^{|\mathcal{M}|} e^{|\hat{f}_{[l]}^i|}$ where $\hat{f}_{[l]}^m$ is l-dimensional of feature $\hat{f}^m$ and $|\mathcal{M}|$ denotes the set of modalities. During forward inference, we perform the inner-product of λ^m and f^m to get the weighted features: $\tilde{f}^m = \lambda^m \cdot f^m$. Finally, we concatenate the weighted features, $\tilde{f} = [\tilde{f}^1, \tilde{f}^2, ..., \tilde{f}^m]$, as a state representation for the standard RL.

3 EXPERIMENTS

We conduct our experiments in two benchmark domains: Mining and autonomous driving. The Mining domain is challenging because the agent needs to make full use of multiple modalities and the modality importance will change during the task. In addition, we performed a case study in a more challenging and realistic domain: self-driving car control, which aims to show the potential and usefulness of our approach in real-world RL applications.

Mining Domain: This domain is originally introduced by the **CASL** paper [16] for testing multimodal RL. As shown in Figure 1(a), an agent wants to mine either gold or iron ore. Specifically, it must determine the type of ores based on their unique audio cue and then pick the right tool for mining. Here, visual input is useful for navigation, and the audio is necessary when deciding which tool to be picked. We then compared our approach with several methods, including: 1) **V+A** [4]; 2) **TIRG** [19]; 3) **MRN** [8]; 4) **CASL** [16]; 5) **Decision-fusion**[15].

Table 1: Results of Self-Driving Car Control.

Method	Average	Worst-Case
MAIE	**2611.25 ± 556.99**	**920.18**
CASL	2144.52 ± 660.24	158.82
Image-Lidar	2457.21 ± 876.57	397.75
Image-Only	1997.12 ± 749.67	202.89
Lidar-Only	2153.20 ± 642.75	508.53

As shown in Figure 1(b), our method substantially outperformed all the compared methods, both in the speed of convergence, the stability of learning, and the quality of policy. Again, this confirms that our method can effectively align the modalities and dynamically enhancement them based on their importance. The reason why these state-of-the-art machine learning methods (i.e. **TIRG** and **MRN**) did not achieve good results in this environment is that modalities may play different importance at some periods in dynamic environments. The competitive performance of CASL shows that the attention mechanism is indeed useful in this domain. However, the training of the attention mechanism requires a large number of samples. Therefore, CASL converged much slowly than ours. All in all, we advances the state-of-the-art with a more effective and efficient multimodal RL approach.

Self-Driving Car Control: To test our methods in more realistic environments, we use a simulator for self-driving car control [12], which is developed for research on RL agent with multi-sensory inputs (i,e, camera image and lidar data). Our results are summarized in Table 1. As expected, the average reward of Image-Lidar (i.e., simply concating image and lidar data) has a higher value of 2457.21 than Image-Only (1997.12) and Lidar-Only (2153.20), which confirms the benefit of multimodal learning. However, directly combining image and lidar data (i.e., Image-Lidar (397.75)) makes even worse than Lidar-Only (508.53) in some situation, i.e., a sudden lane change may cause a collision if the agent cannot attention the approaching vehicles behind. As aforementioned, this is because learning state representation from multimodal inputs directly is challenging for RL.

Our method substantially outperformed all the compared methods, and achieved the best average reward (2611.25), the lowest deviation (556.99), and the best reward in worst case (920.18). Most importantly, our method is 80.9% higher than Lidar-Only in worst case. CASL did not achieve good results in the case study, due to the attention mechanism has more complex structure and requires a large number of samples to train. These results show the potential of our method to improve the safty issues of self-driving car when controlled by RL agent with multi-sensory inputs, thanks to our modality alignment and importance enhancement modules.

ACKNOWLEDGMENTS

This work is supported in part by the Major Research Plan of the National Natural Science Foundation of China under Grant 92048301.

REFERENCES

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2017. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932* (2017).

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[3] Devendra Singh Chaplot, Lisa Lee, Ruslan Salakhutdinov, Devi Parikh, and Dhruv Batra. 2019. Embodied Multimodal Multitask Learning. *arXiv preprint arXiv:1902.01385* (2019).

[4] Florian Henkel, Stefan Balke, Matthias Dorfer, and Gerhard Widmer. 2019. Score Following as a Multi-Modal Reinforcement Learning Problem. *Transactions of the International Society for Music Information Retrieval* 2, 1 (2019).

[5] Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. 2020. Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

[6] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[7] E. T. Jaynes. 1957. Information Theory and Statistical Mechanics. *Phys. Rev.* 106 (May 1957), 620–630. Issue 4.

[8] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. *arXiv preprint arXiv:1606.01455* (2016).

[9] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. 2016. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research* 17, 1 (2016), 1334–1373.

[10] Guan-Horng Liu, Avinash Siravuru, Sai Prabhakar, Manuela Veloso, and George Kantor. 2017. Learning end-to-end multimodal sensor policies for autonomous navigation. *arXiv preprint arXiv:1705.10422* (2017).

[11] Jinming Ma and Feng Wu. 2020. Feudal Multi-Agent Deep Reinforcement Learning for Traffic Signal Control. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Auckland, New Zealand, 816–824.

[12] Kyushik Min, Hayoung Kim, and Kunsoo Huh. 2018. Deep q learning based high level driving policy determination. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 226–231.

[13] Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. *arXiv preprint arXiv:1704.08795* (2017).

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[15] Emilie Morvant, Amaury Habrard, and Stéphane Ayache. 2014. Majority vote of diverse classifiers for late fusion. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 153–162.

[16] Shayegan Omidshafiei, Dong-Ki Kim, Jason Pazis, and Jonathan P How. 2017. Crossmodal attentive skill learner. *arXiv preprint arXiv:1711.10314* (2017).

[17] Avi Singh, Larry Yang, Kristian Hartikainen, Chelsea Finn, and Sergey Levine. 2019. End-to-end robotic reinforcement learning without reward engineering. *arXiv preprint arXiv:1904.07854* (2019).

[18] Elizabeth Spelke. 1976. Infants' intermodal perception of events. *Cognitive psychology* 8, 4 (1976), 553–560.

[19] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6439–6448.

[20] Yixiang Wang and Feng Wu. 2020. Policy Adaptive Multi-Agent Deep Deterministic Policy Gradient. In *Proceedings of the 23rd International Conference on Principles and Practice of Multi-Agent Systems (PRIMA)*. Nagoya, Japan.

[21] Jiaping Zhang, Tiancheng Zhao, and Zhou Yu. 2018. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. *arXiv preprint arXiv:1805.03257* (2018).