Multi-View Deep Attention Network for Reinforcement Learning (Student Abstract)

Yueyue Hu,[1] **Shiliang Sun,**[1] **Xin Xu,**[2] **Jing Zhao**[1]
[1]School of Computer Science and Technology, East China Normal University, Shanghai 200062, China.
[2]College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China.
{slsun, jzhao}@cs.ecnu.edu.cn

Abstract

The representation approximated by a single deep network is usually limited for reinforcement learning agents. We propose a novel multi-view deep attention network (MvDAN), which introduces multi-view representation learning into the reinforcement learning task for the first time. The proposed model approximates a set of strategies from multiple representations and combines these strategies based on attention mechanisms to provide a comprehensive strategy for a single-agent. Experimental results on eight Atari video games show that the MvDAN has effective competitive performance than single-view reinforcement learning methods.

Introduction

The main intuition behind deep reinforcement learning (DRL) is that the agent relies heavily on the observations, which are often represented by a deep model (Mnih et al. 2015). There, the representation of the data in DRL directly determines the performance of the model. Specifically, multi-view representation learning aims to exploit the specific statistical property of each view to learn a more comprehensive representation. Multi-view representation could improve model performance by manually generating multi-view data from the original single-view data (Zhao et al. 2017). The nonlinear function approximator of DRL can be regarded as the visual representation extractor with respect to the value function.

In this abstract, we propose a novel multi-view deep attention network (MvDAN) for the RL framework. The proposed model overcomes the limitation of data observation via introducing multi-view representation learning into the value function approximation. The highlights of our model as follows: 1) implementing collaborative learning among multiple policies, and 2) providing a more comprehensive final policy for the single-agent. Each view is processed by one mapping function from states to view-specific policies, and then all the policies are combined through attention mechanisms, exploiting complementarity across multiple policies. However, one of the biggest challenges is the presence of view disagreement, which may lead to model degradation. To tackle this issue, we apply regularizations to the view-specific policies for view consistency, thus decisions are made in a multi-view aligned space. We demonstrate that the MvDAN outperforms competitively DQN in several Atari games.

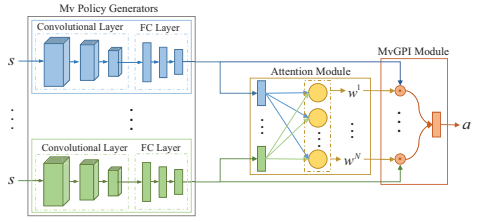


Figure 1: The MvDAN consists of multi-view policy generators, an attention module and a MvGPI module.

Methodology

Multi-view RL Formulation The RL agent learns control strategy with the goal of maximizing the expected discounted return at time-step t: $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$, where $\gamma \in [0, 1)$ is the discount factor, $r \in \mathcal{R}$ is the immediate reward and T is the termination step. The deep Q-network is employed as an embedding function to represent state and approximate the Q-value from the observations $s \in \mathcal{S}$. The key idea is to find the embedding state representation, which determines the mapping relationship from the state to the action. The visual representations $\Psi = \{\psi^k(s)\}_{k=1}^N$ of each view is extracted from N different Q-networks, and the action-value functions is defined as:

$$Q^{\Pi}(\Psi, a; \Theta) = \mathbb{E}\left[R_t | \Psi_t = \Psi, a_t = a\right]. \quad (1)$$

The policies $\Pi = \{\pi^k\}_{k=1}^N$ are implicitly defined by their value functions, with actions are selected by maximizing these functions, which are denoted by $Q^{\Pi} = \{Q^{\pi^k}\}_{k=1}^N$.

The RL algorithms based on the value function are optimized alternately between *policy evaluation* and *policy improvement*. The action-value function is used to evaluate the policy, and the optimal action is selected to improve the policy by acting greedily: $\Pi' = \mathrm{argmax}_a \, Q^{\Pi}(\Psi, a; \Theta)$.

Multi-view Policy Generation To generate the multiple policies corresponding to multiple views, we minimally modify the Q-network architecture with multiple policy networks to satisfy the multi-view reinforcement learning. As shown in Figure 1, each multi-view policy generator constructs the mapping function from the observed state to the view-specific action-value function. Our work focuses on the case of the Q-networks extracts multi-view representations, as long as putting the original observation as the input of nonlinear function approximators at the same time.

Suppose that the multi-view policy generators parameterized by Θ obtain N representations Ψ given the same observation, and the action-vlaue functions are approximated by $Q^{\Pi}(\Psi, a|\Theta)$, yielding policies $\Pi(a|\Psi;\Theta)$. Specifically, the optimal Q-value functions approximated by the multi-view policy generators is written as:

$$Q^*(\Psi, a; \Theta) = \max_{\Pi} \mathbb{E}\left[\, r + \gamma \max_{a'} Q^*\left(\Psi', a'; \Theta\right) \mid \Psi_t = \Psi, a_t = a, \Pi \,\right]. \tag{2}$$

Attention Based Policy Integration Due to the given same observation, the action-value functions generated by the multi-view policy generators are complementary. We apply the attention mechanism to model the complementarity across multiple views. The attention module is utilized to assign the attention weights,

$$w^k = \frac{\exp\left(Q^{\pi^k}\right)}{\sum_{i=1}^{N} \exp\left(Q^{\pi^i}\right)}, \quad k \in \{1, 2, \cdots, N\}, \tag{3}$$

which automatically determines the importance of view-specific action-value function. Then we integrate multi-view strategies at the decision level through element-wise product operation and obtain a comprehensive action-value function,

$$Q^{\pi^{mv}} = \sum_{i=1}^{N} \left(w^i Q^{\pi^i}\right). \tag{4}$$

Multi-view Generalized Policy Improvement We extend general policy improvement to multi-view generalized policy improvement (MvGPI), which uses a set of policies to implement the policy improvement. The MvGPI achieves the goal of co-learning based on the complementarity across multi-view policies.

Let $\{\pi^k\}_{k=1}^N$ be N decision policies under each view and let $\{Q^{\pi^k}\}_{k=1}^N$ be approximations of action-value function, respectively. The MvGPI follows ϵ-greedily and guides the agent to select the optimal action with the highest Q-value:

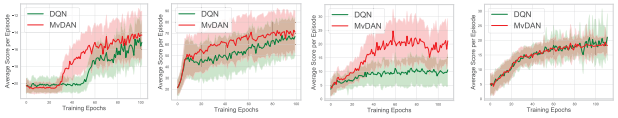
$$\pi^{mv} = \operatorname*{argmax}_{a} Q^{\pi^{mv}}, \tag{5}$$

where $Q^{\pi^{mv}}$ is the final action-value function, which yields a final policy. Moreover, we utilize consistency to add regularizations between each pair of view-specific action-value functions. To achieve this, we force each action-value function to make the same decision by adding a penalty term,

$$\delta = \left| Q^{\pi^u}\left(\psi^u, a; \theta^u\right) - Q^{\pi^v}\left(\psi^v, a; \theta^v\right) \right|. \tag{6}$$

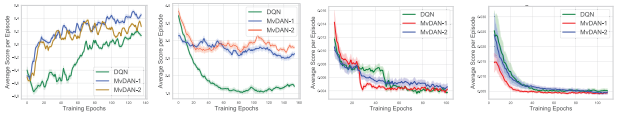
Table 1: The average total reward ($\pm std$) on Atari games.

	Pong	MsPacman	Seaquest	Q*bert
DQN	−13.8 (±4.0)	56.6 (±13.0)	14.4 (±5.5)	18.1 (±6.2)
MvDAN	−11.1 (±2.2)	75.4 (±11.1)	21.6 (±7.2)	17.8 (±3.2)
	Gopher	Atlantis	BeamRider	Breakout
DQN	68.0 (±23.2)	17.7 (±6.5)	17.9 (±3.4)	25.6 (±8.9)
MvDAN	75.9 (±20.3)	30.0 (±14.8)	15.7 (±3.7)	22.3 (±4.8)



(a) Pong (b) MsPacman (c) Seaquest (d) Q*bert

Figure 2: The average reward per episode on Atari games.



(a) Pong (b) Seaquest (c) Pong (d) Seaquest

Figure 3: The left two plots show the average maximum view-specific Q-value per epoch, and the right two plots are the average loss per epoch, respectively.

Empirical Experiments and Results

The performance is illustrated in Table 1 and Figure 2, and we improvement MvDAN on eight Atari 2600 video games with two view representations in all experiments. Several observations are obtained as follows: 1) Overall, MvDAN is significantly better than DQN on most benchmark games. 2) Figure 2(d) shows that MvDAN has a smaller standard deviation, which reflects the stability of our model. 3) Furthermore, Figure 3 illustrates faster convergence against DQN. This is mainly due to the introduction of constraint between multi-view policies to maintain consistency of decisions.

Conclusion

In this abstract, we proposed MvDAN which generates multi-view policies in a novel multi-view learning setting. We have demonstrated the superiority of our method by the effectiveness of multi-view representation on decision-making and the competitive performance of the model.

References

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Zhao, J.; Xie, X.; Xu, X.; and Sun, S. 2017. Multi-view learning overview: Recent progress and new challenges. *Information Fusion* 38:43–54.