Enhancing Context-Based Meta-Reinforcement Learning Algorithms via An Efficient Task Encoder* (Student Abstract)

Feng Xu [1†] , **Shengyi Jiang** [1†] , **Hao Yin** [1], **Zongzhang Zhang** [1‡],
Yang Yu [1], **Ming Li** [1], **Dong Li** [2], **Wulong Liu** [2]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2] Noah's Ark Lab, Huawei Company
{xufeng, jiangsy, yinh, zhangzz, yuy, lim}@lamda.nju.edu.cn, {lidong106, liuwulong}@huawei.com

Abstract

Meta-Reinforcement Learning (meta-RL) algorithms enable agents to adapt to new tasks from small amounts of exploration, based on the experience of similar tasks. Recent studies have pointed out that a good representation of a task is key to the success of off-policy context-based meta-RL. Inspired by contrastive methods in unsupervised representation learning, we propose a new method to learn the task representation based on the mutual information between transition tuples in a trajectory and the task embedding. We also propose a new estimation for task similarity based on Q-function, which can be used to form a constraint on the distribution of the encoded task variables, making the task encoder encode the task variables more effective on new tasks. Experiments on meta-RL tasks show that the newly proposed method outperforms existing meta-RL algorithms.

Introduction

Humans can adapt to new tasks from small amounts of exploration in the environment by leveraging their prior knowledge. However, this step of learning raises a big challenge for AI agents. Meta learning frameworks aim at tackling this problem by capturing shared knowledge across different tasks. In the reinforcement learning environment setting, agents capture knowledge by interacting with the environment and learning different tasks. One state-of-the-art meta-RL method is PEARL (Rakelly et al. 2019), which is an off-policy context-based meta-RL algorithm. PEARL uses a task variable learned from the context to guide the policy. The key to context meta-RL algorithms is how to learn an efficient task variable for the policy to utilize it. In this paper, we try to enhance the previous works on context-based meta-RL algorithms by focusing on the task encoder.

*This work is in part supported by the National Natural Science Foundation of China (No. 61876119) and Huawei Noah's Ark Lab (No. HF2019105005).

†Equal Contribution.

‡Corresponding author.

Methods

Contrastive Method Inspired by the unsupervised mutual information estimation method mentioned in Deep InfoMax (DIM) (Hjelm et al. 2019), we maximize the mutual information (MI) between the tuples sampled from the replay buffer and the encoded task variable. Let $x = \{s, a, r\}$ be a tuple sampled from the replay buffer, and z be the encoded task variable. Let E_ϕ be a neural network with parameters ϕ, which takes input x and produces the parameters of the posterior Gaussian distribution of z. Following the formulation in (Nowozin, Cseke, and Tomioka 2016), we use an MI estimator based on Jensen-Shannon Divergence (JSD):

$$\hat{I}^{JSD}_{\omega,\phi}\big(X; E_\phi(X)\big) = \mathbb{E}_x\Big[-sp\Big(-T_{\omega,\phi}\big(x, E_\phi(x)\big)\Big)\Big] - \mathbb{E}_{x,x'}\Big[sp\Big(T_{\omega,\phi}\big(x', E_\phi(x)\big)\Big)\Big],$$

where x' is sampled from the replay buffer of another task and $sp(a) = \log(1 + e^a)$ is the softplus function. We would like to optimize E_ϕ by estimating and maximizing the mutual information between x and z. Thus, the optimal parameters for the task encoder E_ϕ and the discriminator T_ω are

$$\arg\max_{\omega,\phi} \hat{I}^{JSD}_{\omega}\big(X; E_\phi(X)\big) + \arg\min_{\phi} D_{\mathrm{KL}}(\mathbb{Z}||\mathbb{N}),$$

where $\mathbb{Z}$ is the distribution of the encoded variables, and $\mathbb{N}$ is the prior Gaussian distribution.

Similarity Estimation Observing the high variance in the reward curves, we analyze how the task encoder maps the task variable from the source distribution to the target distribution. After projecting the 5-dimensional variable to a 2-dimensional plane with PCA, we find that the ordered source distribution becomes relatively unordered, as shown in Fig. 1(b). To address this issue, we propose an estimator of task similarity to assist the encoder. We use the Wasserstein distance and a new estimator inferred by the Q-function to compute the similarity between the encoded task variables. The formula for the 2-Wasserstein distance between two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ is

$$d^2_W\big(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)\big) = ||\mu_1 - \mu_2||^2_2 + ||\Sigma_1^{\frac{1}{2}} - \Sigma_2^{\frac{1}{2}}||^2_F,$$

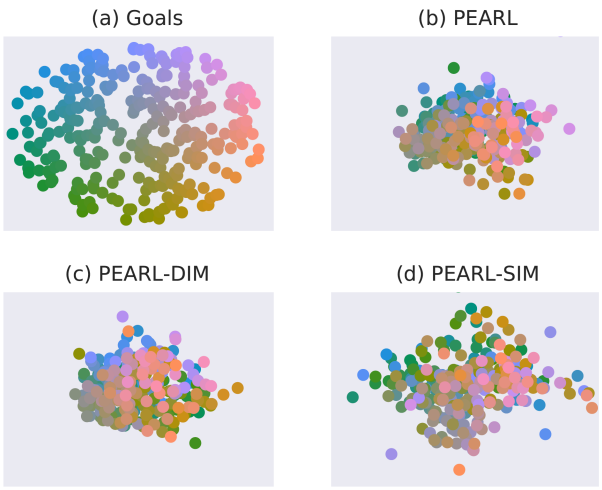

Figure 1: The behavior of the context encoders. (a) shows the source distribution. (b∼d) show the behavior of PEARL, PEARL with Deep InfoMax (PEARL-DIM) and PEARL with Similarity Estimation (PEARL-SIM), respectively.

where $||M||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|m_{ij}|^2}$. We normalize the Wasserstein distances by the maximum Wasserstein distance in the prior distribution, denoted $d_{\max}$. The similarity estimation given by the Wasserstein distance estimator is

$$S_W\big(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)\big) = 1 - \frac{d_W^2\big(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)\big)}{d_{\max}}.$$

The second estimator is inferred from the Q-function. Since Q-function gives how good the action is in the state, completing the current task. Intuitively, in the same state, under different tasks, the difference in the $Q(s,a,z)$ should reflect the similarity between tasks. Based on this intuition, we propose a task similarity estimator calculated from the Q-value difference. First, we normalize the Q-values as

$$\bar{Q}(s,a,z) = [Q(s,a,z) - \mu_Q]/\sigma_Q,$$

where μ_Q and σ_Q are the mean and standard deviation of the Q-values in the current epoch. The difference between the encoded task variables can be calculated as

$$Q_{\texttt{diff}}(s,a,z,\tilde{z}) = ||\bar{Q}(s,a,z) - \bar{Q}(s,a,\tilde{z})||_1.$$

The similarity between the encoded task variables is

$$S_Q(z,\tilde{z}) = 1 - \frac{\mathbb{E}_{s,a}Q_{\texttt{diff}}(s,a,z,\tilde{z})}{2},$$

where $\mathbb{Z}$ is the distribution of the encoded task variable z. Finally, we use the L2-norm distance as the loss:

$$L_{\texttt{similarity}}(\mathbb{Z}) = \mathbb{E}_{z,\tilde{z}\sim\mathbb{Z}}\big(S_Q(z,\tilde{z}) - S_W(z,\tilde{z})\big)^2.$$

Thus, the optimal parameters for the encoder is

$$\hat{\omega} = \arg\min_\omega \big(D_{\text{KL}}(\mathbb{Z}|\mathbb{N}) + L_{\texttt{similarity}}(\mathbb{Z})\big).$$

Metric	Goal	PEARL	PEARL-DIM	PEARL-SIM
EV_D	404	301	294	361
EV_S	3321	3272	3281	3286

Table 1: Results on quantified metrics

Experiments

In this section, we compare the performance of PEARL with our enhanced context encoder methods, including the contrastive method and the task similarity estimation method, on a physics-based control task named ant-goal in the MuJoCo physics engine. The dimension of the encoded task variable is 5. For each method, we visualize the 5-dimensional encoded task variables by projecting them to a 2-dimensional plane with PCA, so that the distances between the projected variables reflect their original spatial features.

The behavior of our methods is presented in Fig. 1. From it, we can see that the posterior distribution generated by PEARL-SIM discriminates different tasks effectively, and matches better to the source distribution.

To quantify the results, we define two metrics to capture the spatial features of the distribution. The first metric, which calculates the dispersion degree, is

$$EV_D = \mathbb{E}_{z_1,z_2}(D_E(z_1,z_2)),$$

where D_E is the Euclidean distance. To describe the distribution similarity EV_S, we divide EV_D by the standard deviation σ_Z of the variables, which is

$$EV_S = \frac{EV_D}{\sigma_Z}.$$

The result is shown in Table 1. From it, we can see that the closer the value is to the goal distribution, the more similar they are.

Conclusion

In this paper, we propose two methods to enhance the performance of the context-based meta-RL algorithm PEARL. We use contrastive methods to discriminate different tasks and propose a new estimator for the task similarity, so that the task encoder could discriminate different tasks and learn the similarity between tasks better.

References

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning Deep Representations by Mutual Information Estimation and Maximization. In *ICLR*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *NIPS*, 271–279.

Rakelly, K.; Zhou, A.; Finn, C.; Levine, S.; and Quillen, D. 2019. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *ICML*, 5331–5340.