Paper Review

Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variable Model

[1]David Pomerenke, JingYang Zeng
Maastricht University
Advanced Concepts in Machine Learning
08.12.2020

Reviewed work:

Lee, A., Nagabandi, A., Abbeel, P., & Levine, S. (2020). *Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model.* Advances in Neural Information Processing Systems, 33.

Summary

Reinforcement learning, that is learning a policy of applicable actions for each state based on rewards related to the state transitions, is a sort of problem that is especially useful in real-world applications such as robotics. Q-learning is a popular existing model-free technique for reinforcement learning where state transitions and rewards are not learned separately but combined to a single Q-value. The actor-critic algorithm is a variety of Q-learning, where an actor component selects samples based on the current Q-values, and a critic component updates the Q-values based on the temporal-difference error of these samples. A problem of state-of-the-art reinforcement learning algorithms is that they require the manual engineering of a suitable state representation, and otherwise are very slow and and dependent on hyperparameter tuning. The paper addresses this issue by separating the reinforcement learning problem into a *representation learning* component and a *task learning* component.

For **task learning**, the paper uses a state-of-the-art technique called soft actor-critic, which is built upon a maximum-entropy reinforcement learning framework. The latter means that the loss function that is minimized also includes a term for the entropy of the policy. This leads to a better exploration of the state space by incentivizing exploration, exploring similarly promising states with similar intensity, and yet ignoring clearly unpromising states.[1] Furthermore, soft actor-critic improves upon earlier maximum-entropy approaches, which were using so-called double Q-learning, by instead learning a single "soft" Q-value.[2]

The **representation learning** component is concerned with encoding the state of a partially observable Markov decision process (POMDP; a formalization suitable to many stochastic real world environments) in a way that makes learning easy for the task learning component. For this purpose, the paper is inspired by variational autoencoders. Autoencoders learn an efficient encoded representation of any kind by training a model with a latent space that is smaller than the original space with the objective of reconstructing the original input. Variational autoencoders improve the robustness of autoencoders by adding noise in the training process. The paper transfers the idea underlying variational autoencoders to the specific problem of learning POMDPs, by using a technique called amortized variational inference. Variational inference is a probabilistic modeling technique where the explicit graphical model of the latent variables is replaced by a deep neural network for the purpose of achieving tractability.[3]

The main contribution of the paper is bringing together the representation learning and task learning components in a unified algorithm. The algorithm produces stunning results in comparison to other state-of-the-art algorithms, especially for more complex image-based reinforcement learning tasks.

Relevance

The paper is highly relevant to the machine learning community, because it addresses the issue of reinforcement learning for real-world scenarios. Reinforcement learning, in general, is already a suitable representation for many real-world problems, since —unlike labels—, rewards and punishments can often be found in the real world. However, according to the framing of the authors, reinforcement learning has so far not been directly applicable to many real-world tasks; only when a good representation of the state space is available do reinforcement learning algorithms achieve good results. A good representation may require human effort, as well as additional sensors and additional data preprocessing steps.

The paradigmatic application domain presented in the paper is for image-based robotic tasks. Here, financial costs could be reduced by replacing multiple specialized sensors by one single camera; and no special image preprocessing would be necessary. The approach is probably also transferable to other automation tasks, although this is not explored in the paper.

Conceptually, the paper transfers the idea behind variational autoencoders to the domain of Q-learning. To this purpose, tools from Bayesian modeling and deep learning are employed. Thus, the reinforcement learning community is introduced to useful concepts from the statistics and deep learning communities.

Significance

The experimental evaluation of the SLAC algorithm indicates highly significant results: For image-based robotic control tasks from the Deepmind benchmark[4], SLAC performs slightly (but rather insignificantly) better than the best state-of-the-art model-free Q-learning algorithms, both in terms of the speed and the value of the convergence (cf. Figure 3 in the paper). The core strength of the paper is with the OpenAI Gym benchmark[5], where image-based robotic control tasks have been used as well, but where according to the authors they are more challenging for a number of reasons (cf. p. 7): In the relevant categories of this benchmark, the average return after convergence is up to 4 times higher than for other state-of-the art algorithms (cf. Figure 4 in the paper).

The significance is also reflected in the reception by the research and coding communities: Albeit the paper has only been published this year, it has already been cited 44 times according to Google Scholar at the time of the submission of this review, with 6 highly influential citations according to Semantic Scholar; the implementation has received 94 stars and has been forked 19 times on Github.[1] This is a clearly above-average reception, which is probably due to the excellent benchmark results for solving a relatively fundamental machine learning problem.

Novelty

This paper has general novelty because it is a combination of representation learning and reinforcement Learning. Although the paper claims that the algorithms can deal with high-dimensional input, the core of the algorithm is an off-policy model-free Q-learning algorithm with the combination of representation learning, and the latent layer works as a conjunction.

Therefore, it is a straightforward idea that if the agent needs to learn with high-dimensional input, it needs to learn some representation to filter the input; but this algorithm provides a very prominent performance. Moreover, after transferring the high-dimensional input to the limiting valid input, they apply the POMDP to learning the remaining problem. Here, POMDP works well but it is a very standard choice, therefore we believe this algorithm is implemented in a normal track. In our view, it's a successful extension of the off-model Q-learning algorithm, so we believe this algorithm is a great innovation rather than an invention.

Soundness

The paper is technically sound. In the Introduction section, the authors introduce the challenge they address, and they mention the bottleneck reinforcement learning problem in the Related Work section. Then, the algorithm comes up after the statement of the bottleneck problem. Besides the technical soundness in the structure, the paper also provides strict math formulae to prove the mechanism of the algorithm.

These formulae justify the inference and logic in the mechanism: All variables are defined to match the concept in the mechanism, then functions apply the logic of the mechanism so we can make sure the mechanism is perfectly proved in math. When the formula is introduced, the authors start with some small sections of the algorithm, then combine these small sections together by the mechanism. Meanwhile, the Pseudocode at the side does help us to understand the flow of the SLAC. Therefore, we believe this paper states its contribution in a very technical way.

Evaluation

This paper set up many experiments and fully evaluates the results of the experiments. They evaluate the SLAC in 8 tasks, four of the tasks are cheetah run, walker walks, ball-in-cup catch, finger spin of the DeepMind Control Suite banchmark[4], the others task are cheetah, walker, ant, and hopper from the OpenAI Gym benchmark[5]. So it is no doubt that SLAC can solve different reinforcement learning problems. Furthermore, in order to prove the priority of SLAC, they compare the performance of SLAC with other successful algorithms such as SAC, DVRL, PlaNet. Therefore, the prior performance of SLAC is exactly confirmed. By watching these charts, I do believe SLAC can converge in earlier epochs and achieve better optimal performance.

However, the authors make many strong claims, but not all of the claims are fully supported. For example, the authors claim that their approach performs infinite horizon policy optimization, but I don't find any statement about the infinite horizon policy optimization of SLAC. Therefore, this paper has a great evaluation overall, but there remain some claims that are not fully evaluated.

Clarity

Most of the paper is clear and straightforward. When we look through the whole paper, we can understand the main idea of SLAC, and get the notion of the complicated mechanism. However, the paper does have some confusing parts that lack sufficient explanation.

- The authors claim that their approach does not use the model for prediction, but SLAC does calculate the next action and next latent state in the soft Q-function, and the generative model also considers the possibility of the next action based on the current action. Therefore, the mechanism makes some predictions indeed, hence it would be better if there gives more explanation between the claims and formula about the prediction.

- Another unclear point is we don't know how the input images x tease out the relevant information into a compact and disentangled representation z. This is a significant point in the latent variable model because the disentangled representation z will decide the performance of

[1] https://github.com/alexlee-gk/slac

the latent actor-critic part, but there should be more explanation about how the disentangled representation z is generated.

Thus, although this paper is sufficiently clear and persuasive, it still has some tiny flaws in clarity.

Detailed Comments

- One of the comments is about architecture. It would be better to go back to the bottleneck problem in reinforcement learning after experiments and evaluation. Since the addressed problem is high-dimensional observation spaces in reinforcement learning, and we confirm the prior performance of SLAC comparing other algorithms. However, we don't know whether better performance solves the bottleneck problem or not. Therefore, after the experiments, go back to the addressed point and give more interpretation on how SLAC addresses the bottleneck problem would be better in our opinion.
- Some terms relating to existing state-of-the-art methods have not been introduced in sufficient detail; thus, we needed to look up these terms in order to follow the paper. Some of these are: *Control as inference (p. 2), variational inference (p. 2), amortized variational inference (p. 3), evidence lower bounds (p. 4)*. We assume this is due to space constraints.
- In the first paragraph of the Related Work section, 10 citations of prior work on representation learning for reinforcement learning are given. The authors contrast their work in a single and not very clear sentence from all these 10 approaches at once. The problem of representation learning for reinforcement learning seems to be quite fundamental and we would expect that there are already some important achievements in this field, so the authors should engage in more depth with these approaches.

Questions for the authors

1. Variational autoencoders are mentioned to illustrate the role of amortized variational inference in the SLAC algorithm. Why do the authors not use variational autoencoders themselves?
2. When variational autoencoders are applied properly to learning image encodings, human-interpretable visual features are learned. Is the same true for using amortized variational inference? If not, can it be considered a drawback that there is a loss of interpretability in comparison to traditional Q-learning where a suitable encoding is engineered manually?

References

[1] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.

[2] J. Duan, Y. Guan, S. E. Li, Y. Ren, and B. Cheng, "Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors," *arXiv preprint arXiv:2001.02811*, 2020.

[3] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[4] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. d. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, *et al.*, "Deepmind control suite," *arXiv preprint arXiv:1801.00690*, 2018.

[5] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.