Stochastic Variance Reduction for Deep Q-learning

Extended Abstract

Wei-Ye Zhao
Carnegie Mellon University, Robotic Institute
Pittsburgh, PA
weiyez@andrew.cmu.edu

Jian Peng
University of Illinois at Urbana-Champaign
Champaign, IL
jianpeng@illinois.edu

ABSTRACT

Recent advances in deep reinforcement learning have achieved human-level performance on a variety of real-world applications. However, the current algorithms still suffer from poor gradient estimation with excessive variance, resulting in unstable training and poor sample efficiency. In our paper, we proposed an innovative optimization strategy by utilizing stochastic variance reduced gradient (SVRG) techniques. With extensive experiments on Atari domain, our method outperforms the deep q-learning baselines on 18 out of 20 games.

KEYWORDS

Deep Q-learning; Stochastic variance reduction; Gradient variance

ACM Reference Format:
Wei-Ye Zhao and Jian Peng. 2019. Stochastic Variance Reduction for Deep Q-learning. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

With the help of deep learning, reinforcement learning (RL) [7] has recently achieved remarkable success on massive real-world applications, such as human-computer interaction [5], video games [6], visual navigation [10], goal-oriented autonomous decision making [3] and autonomous driving [2]. Q-learning [9] is one of the most popular reinforcement learning algorithms. A standard method to solve optimization problems in Q-learning is gradient descent [4]. Since it is expensive to compute the full expectation in the gradient, stochastic methods are often used to optimize the loss function based on gradients of small batches of samples. Despite these successes, the inaccurate estimation of gradient as well as huge variance arisen from RL training procedure is still the key problem of these stochastic optimization methods, the inexact approximate gradient estimation can be viewed as the distorted gradient direction.

In large scale deep Q learning problem, the Q value is represented with deep Q network with proper tuned network parameters. The DQN learning process can be viewed as iteratively optimizing network parameters process according to gradient direction of the loss function at each stage. Therefore, the inexact approximate gradient estimation with a large variance can largely deteriorate the representation performance of deep Q network by driving the network parameter deviated from the optimal setting, causing the large variability of DQN performance. On the other hand, if we assume the network parameter of DQN is θ, the core learning step of deep Q learning is to minimize the gap between the estimated maximum Q value ($y_(s,a)$) given state s and action a and current Q value($Q(s,a;\theta)$) using the operation that $\hat{\theta} = \text{argmin}_\theta \mathbb{E}\|y_(s,a) - Q(s,a;\theta)\|^2$. It is noteworthy that $\hat{\theta}$ is obtained with gradient descent, thus if the gradient estimation has a large variance, it requires more iterations of argmin operation such that θ could reach $\hat{\theta}$, which means large gradient variance will postpone the process when DQN gets local optima.

In this work we address issues that arise from Approximate Gradient Estimation (AGE), and propose Stochastic Variance Reduction for Deep Q-learning (SVR-DQN) optimization, as shown in algorithm 1, to accelerate the convergence for deep Q-learning by reducing the AGE variance. We conduct the AGE variance analysis and theoretically explain how the proposed algorithm addresses them.We evaluate our proposed algorithm using Arcade learning environment [1]. Our experiments show that SVR-DQN optimization algorithm can significantly reduce the delay before the performance gets off the ground, and further lead to aggressive sample efficiency at initial training stage. Our new strategy outperforms Adam in both reward scores and training time on 18 out of 20 games.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

	Mean	Median
SVR-DQN	**139.75%**	**118.02%**
Double DQN	92.48%	63.13%

Table 1: Mean and median normalized scores.

2 EXPERIMENTAL RESULTS ON ATARI GAMES

To demonstrate our method's effectiveness, we evaluate our proposed algorithm on a collection of 20 games from Arcade Learning Environment [1]. This environment is considered as one of the most challenging datasets because of its high-dimensional state representation [8]. We utilize the similar neural network [6] as the approximation of action value, taking raw images as input. The network architecture is a convolutional neural network with three convolutional layers and a fully-connected layer. All experiments are performed on an NVIDIA GTX Titan-X 12GB graphics card. In this paper, we utilize the tunned version of Double DQN algorithm [8], as it somehow resolves the over-estimation issue in Q-learning.

3 RESULTS AND DISCUSSION

Our evaluation procedure follows the description by [6], we apply '30 no-op evaluation' to provide different starting points for the

Algorithm 1 Stochastic Variance Reduction for Deep Q-learning Optimization

procedure STOCHASTIC VARIANCE REDUCTION FOR DEEP Q-LEARNING OPTIMIZATION($B, \eta, m, \alpha, \beta_1, \beta_2, b$)
  Inputs:
  B: Training sample batch size; η: SVRG learning rate; m: Number of SVRG inner loop iteration; b: mini-batch size
  for $s = 0, 1, 2, \ldots$ **do**
    $B^s = B$ elements sampled without replacement from all training samples ▷ training sample batch
    Calculate the anchor point:
    $\tilde{\mu}^s = \frac{1}{B} \sum_{i \in B^s} \nabla f_i(\tilde{w}^s)$
    $w_0 = \tilde{w}^s$
    for $t = 1, 2, \ldots, m$ **do** ▷ SVRG variance reduction
      Draw a mini-batch b^t uniformly random from B^s ▷ mini-batch
      Reduce variance and update parameter with mini-batch b_t:
      $w_t = w_{t-1} - \eta(\frac{1}{b} \sum_{i \in b^t} \nabla f_i(w_{t-1}) - \frac{1}{b} \sum_{i \in b^t} \nabla f_i(\tilde{w}^s) + \tilde{\mu}^s)$
    end for
    Calculate variance reduced gradient: $g_s = w_m - \tilde{w}^s$
    Adam optimization procedure
  end for
  return $\tilde{w}^s$
end procedure

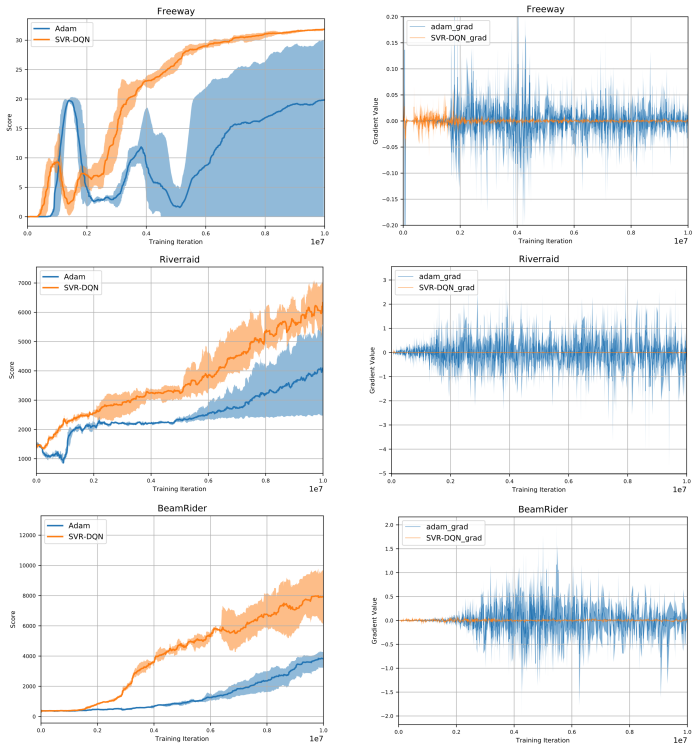

Figure 1: The bold lines are averaged over 6 independent learning trials (6 different seeds). The shaded area presents one standard deviation.

agent. Our agent is evaluated after a maximum of 5 minute gameplay, which contains 18,000 frames, with the usage of ϵ-greedy policy where $\epsilon = 0.05$. The rewards are the average from 100 episodes. For each game, our agent is evaluated at the end of every epoch (160 epochs in total). To compare the performance of our algorithm to the Double DQN baseline across games, we apply the normalization algorithm proposed by [8] to obtain the normalized improvement score in percent as follows:

$$\text{score}_{\text{normalized}} = \frac{\text{score}_{\text{agent}} - \text{score}_{\text{random}}}{|\text{score}_{\text{Double DQN}} - \text{score}_{\text{random}}|} \quad (1)$$

In summary, we adopt the 'Double DQN' and 'random' score reported by [6]. We observe a better performance on 19 out of 20 games, which demonstrates the effectiveness of our proposed algorithm. We also give the summary statistics in terms of mean and median score in Table 1. Also, the performances of 3 representative games are reported in Fig.1. The three games include 'BeamRider', 'Freeway', 'Riverraid'. As can be seen in Fig.1 that out proposed SVR-DQN method results in significant lower average gradient estimates, and the variance of gradient is largely reduced. We also observe that our method outperforms the baselines with a significant margin on the majority of the games, and SVR-DQN leads to less variability between the runs of independent learning trials. For the game of Freeway, we see that the divergence of Double-DQN can be prevented by SVR-DQN. On the other hand, the performance of Double-DQN with Adam optimizer has a sudden deterioration at 4M iteration where the gradient variance suddenly increases.

4 CONCLUSION

In this paper we proposed an innovative optimization algorithm for Q-learning which reduces the variance in gradient estimation, our proposed optimization algorithm achieves significantly faster convergence than the Adam optimizer. Our method significantly improves the performance of Double DQN on the Atari 2600 domain. In the future, we plan to investigate the impact advanced constrained optimization and explore the potential synergy with other techniques.

REFERENCES

[1] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The Arcade Learning Environment: An evaluation platform for general agents. *J. Artif. Intell. Res.(JAIR)* 47 (2013), 253–279.

[2] Xiaohui Dai, Chi-Kwong Li, and Ahmad B Rad. 2005. An approach to tune fuzzy controllers based on reinforcement learning for autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems* 6, 3 (2005), 285–293.

[3] Michael J Frank and Eric D Claus. 2006. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological review* 113, 2 (2006), 300.

[4] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[5] Pattie Maes and Robyn Kozierok. 1993. Learning interface agents. In *AAAI*, Vol. 93. 459–465.

[6] 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.

[7] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.

[8] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning.. In *AAAI*. 2094–2100.

[9] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.

[10] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 3357–3364.