Q-value 기반 트리밍 및 데이터 증강을 활용한 강화학습

유승찬, 주석훈, 이정우
서울대학교

scyu@cml.snu.ac.kr, seokhunju@cml.snu.ac.kr junglee@snu.ac.kr

Reinforcement Learning with Q-value based Trimming and Data Augmentation

Seung Chan Yu, Seok Hun Ju, Jung Woo Lee
Seoul National Univ.

요 약

본 논문에서는 기존의 전체 데이터에 대한 데이터 증강 대신, 학습에 도움이 될 것으로 예상되는 데이터를 선별하여, 선별된 데이터에 대해서만 데이터 증강을 적용한다. 이를 위해 보상의 기대값인 Q-value 를 기준으로 너무 크거나 작은 값을 트리밍 하고 중간에 위치한 값을 선별한다. 이렇게 트리밍과 데이터 증강을 활용한 방법과, 전체 데이터에 대한 데이터 증강을 활용한 방법을 비교하였다.

Ⅰ. 서 론

이미지 기반의 데이터 증강 기법은, LeNet-5[1]에서부터 시작하여, 한정된 양의 데이터를 극복하고 과적합(Overfitting)을 방지하기 위해 광범위하게 사용되었다. 또 강화학습 분야에서도 단순 이미지 데이터 증강만으로 Deepmind Control Suite 환경[2]에서 성능을 향상시키는데 성공하였다[3]. 우리는 여기서 더 나아가 선별된 데이터에 대해서만 데이터 증강을 적용하여, 연산량을 줄임과 동시에 학습 성능 또한 향상되는지 확인하였다. 데이터 선별에는 보상(Reward)의 기대값인 Q-value 를 사용하였다. Q-value 가 작은 데이터는 좋은 학습 방향이 아니므로 제외(Trimming)하고, 큰 값 또한 과추정(Overestimation) 일 가능성이 높기 때문에 데이터 증강에서 제외하였다.

Ⅱ. 본론

제안 방법

먼저 데이터 트리밍 설계를 위해 학습중에 데이터들의 Q-value 분포가 어떻게 되는지 살펴보았다. Figure 1 이 한 학습 사이클에서 보이는 Q-value 분포이다. 학습 초반에는 가우시안에 가까운 분포를 보이지만, 학습이 진행될수록 조금씩 평평하게 퍼지는 경향을 볼 수 있었다. 특히 학습 후반으로 가면 높은 Q-value 의 분포가 커지는 과추정(Overestimation) 경향 또한 확인할 수 있었다.

기반 알고리즘으로 Double DQN[4] 을 사용했기 때문에, 한 step 에서 두개의 Q-value 분포가 그려지는 것을 확인할 수 있다. Double DQN 은 역시 과추정 문제를 해결하기 위한 기법으로, 두개의 네트워크를 사용하여 계산한 두개의 Q-value 중에서 더 작은 Q-value 를 취해서 과추정을 방지하는 기법이다.

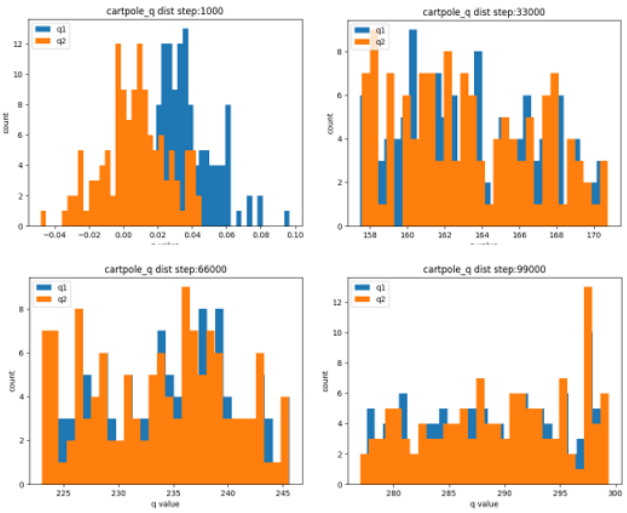

Figure 1 학습 중 Q-value 분포

Q-value 분포에서 작은 값들은, 보상(Reward)의 기대값이 작기 때문에 데이터 증강에서 제외하고, 큰 값들은 과추정 되었을 가능성이 높기 때문에 역시 제외하였다. 이런 트리밍 과정을 거쳐 남은 중앙에 위치한 50%의 데이터에 대해서만 데이터 증강을 적용하였다. 이에 더하여, 매 step 에서 트리밍을 진행하면 반대로 underfitting 문제에 직면 할 수 있기 때문에, 홀수 step 에서는 트리밍 없이 모든 데이터에 데이터 증강을 적용하고, 짝수 step 에서는 데이터 트리밍 후 중앙 50% 데이터에만 데이터 증강을 적용하는 soft 트리밍 방법을 적용하였다. 데이터

증강방법으로는 검은색 배경 위의 무작위 위치에 원본 이미지를 붙여넣는 Random Translation 을 사용하였다.

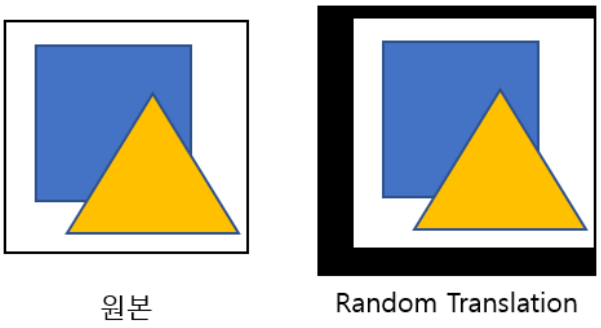

Figure 2 데이터 증강 방법

실험 결과

실험은 Deepmind Control Suite 의 Cartpole-Swingup 과 Cheetah-run 에서 진행하였다. Cartpole 환경은 수레에 막대기가 연결되어있는 형태로 관절이 1 개, Cheetah 환경은 관절이 총 5 개로 이루어져있어, Cheetah 환경이 더 복잡하고 어려운 환경이다. Cartpole-Swingup 에서의 결과는 아래

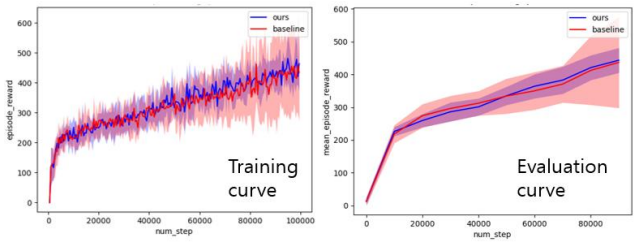

Figure 2 와 같다.

Figure 3 Cartpole-Swingup 실험결과

파란색은 데이터 트리밍과 데이터 증강 모두 적용한 방법이고, 붉은색이 모든 데이터에 대해 데이터 증강을 적용한 대조군이다. 5 개의 무작위 시드(Random seed)로 학습을 진행하고, 평균값을 진한 선으로 표시하였다. 본 논문에서 제시한 트리밍 방법이 약간이지만 더 좋은 성능을 보였고, 특히 산포가 줄어든 또 다른 효과도 보여주었다.

Cheetah-run 에서의 결과는 아래 Figure 3 와 같다.

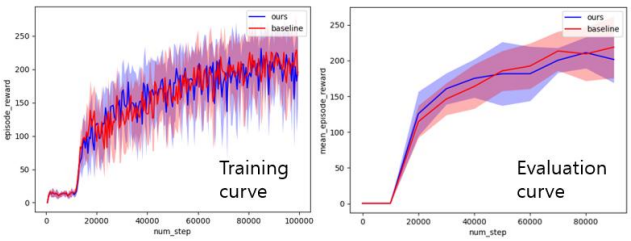

Figure 4 Cheetah-run 실험결과

마찬가지로 파란색이 데이터 트리밍 적용 방법, 붉은색이 대조군이다. 마찬가지로 5 개의 무작위 시드로 실험을 진행하였다. 대조군과 본 논문에서 제시한 방법 사이에 유의미한 차이를 찾기 힘들었다. Cartpole 에서 처럼 산포가 줄어드는 효과도 찾아볼 수 없었다.

Ⅲ. 결론

본논문에서는 강화학습 환경에서 성능향상과 연산 효율성을 위해, 데이터 트리밍을 통해 선별한 데이터에 대해 데이터 증강 기법을 적용해 보았다. Q-value 가 작은 값 뿐만 아니라, 큰 값까지 트리밍함으로서, 과추정 문제까지 함께 해결하고자 하였다. 실험 결과 비교적 단순한 환경인 Cartpole-Swingup 환경에서는 성능의 개선 뿐 아니라, 무작위 시드별 산포의 감소까지 확인할 수 있었다. 하지만 상태가 훨씬 복잡한 Cheetah-run 환경에서는 유의미한 차이를 보기 어려웠다. 복잡한 환경에서는 soft 트리밍의 하이퍼파라미터를 좀 더 세밀하게 조절할 필요성이 있는 것으로 보인다.

.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(25%)), Institute of Information & communications Technology Planning & Evaluation (IITP, 2021-0-00106(25%), 2021-0-02068(25%)) grant funded by the Ministry of Science and ICT (MSIT), National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(2021M3F3A2A02037893(25%)), INMAC, and BK21-plus.

참 고 문 헌

[1] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[2] Tassa, Yuval, et al. "Deepmind control suite." arXiv preprint arXiv:1801.00690 (2018).

[3] Laskin, Misha, et al. "Reinforcement learning with augmented data." Advances in neural information processing systems 33 (2020): 19884-19895.

[4] Hasselt, Hado. "Double Q-learning." Advances in neural information processing systems 23 (2010).