휴먼 플레이 데이터를 활용한 메모리 기반 모방학습

민병준 ○ 김경중*

세종대학교 컴퓨터공학과

okminkr@gmail.com kimkj@sejong.ac.kr

Memory-based Imitation learning using Human-play data

Byeongjun Min○ KyungJoong Kim*

Dept. of Computer Engineering, Sejong Univ.

요 약

최근 강화학습은 복잡한 의사결정 문제에서 좋은 성과들을 달성하고 있으며, 신경망과 강화학습을 결합한 Deep Q Network(DQN), Asynchronous Actor-Critic Agents(A3C) 와 같은 모델들이 많이 사용되고 있다. 하지만 신경망 기반 모델들은 Stochastic Gradient Descent(SGD) 의 느린 업데이트 속도로 인해 학습에 매우 오랜 시간이 소모된다. 또한 강화학습에서는 에이전트가 스스로 학습에 필요한 샘플 데이터를 수집하기 때문에, 에이전트는 매우 오랜 시간동안 환경과의 상호작용을 한다. 본 논문에서는 이러한 느린 학습속도를 개선하기 위한 메모리 기반 학습방법 Episodic Control From Demonstration(ECFD)을 제안하여 휴먼 플레이 데이터셋을 활용한 모방학습을 진행한다. 실험은 Atari 환경에서 진행하였고, 기존 모델들과 비교해 좋은 성능을 보였다.

1. 서 론

최근 다양한 의사결정 문제를 강화학습으로 해결하기 위한 연구들이 활발히 진행되고 있다. 강화학습은 학습에 주체가 되는 에이전트가 환경과 직접 상호작용하면서 현재 정책(Policy)을 개선한다. 뉴럴 네트워크와 강화학습을 결합한 Deep Q Network (DQN)[1]은 2013년 사람 전문가 수준의 성능을 달성하였으며, 2015년 소개된 알파고(AlphaGo)[2]는 인간과의 바둑 대국에서 4승1패의 전적을 획득하였다. 하지만 이러한 성취를 위해서는 많은 데이터들을 필요로 한다. 또한 신경망 기반 모델들은 Stochastic Gradient Descent(SGD) 의 느린 업데이트 방식으로 인해서 정책의 개선 또한 매우 늦다. 따라서 현재 DQN과 같은 모델들을 사용해 Atari 게임을 학습시키기 위해서는 많은 시간을 필요로 한다.

강화학습에서 학습의 속도를 앞당기는 방법으로는 모방학습(Imitation Learning)[3]이 있다. 모방 학습은 에이전트가 주어진 대상을 관찰하여, 모방을 시도하며 일련의 행동과정을 익히는 것이다. 에이전트는 전문가의 의사결정을 학습함으로써 주어진 환경에서 좋은 정책을 가지게 되는 것이다. 모방을 통한 시도는 예전부터 쭉 시도해왔지만, 컴퓨팅 자원의 발전과 학습에 대한 관심이 높아짐에 따라서 최근 더욱 많이 주목받고 있는 추세이다.

이 외에도 학습속도의 개선을 위해 메모리 기반 강화학습 방법들도 연구되고 있다. 대표적으로는 구글 Deep Mind에서 2016년에 발표한 Model-Free Episodic Control (EC, MFEC) 이 있다. EC는 신경망 모델들 보다 높은 점수를 달성하였으며, 학습속도 또한 매우 빠르다.

그림 1. MS-Pacman 게임 플레이 모습

본 논문에서는 Model-Free Episodic Control 을 모방학습에 활용가능한 ECFD 모델을 제안한다. ECFD는 과거의 성공을 재현하는데 매우 탁월하며, 이는 과거 휴먼 플레이 데이터로부터 빠른 정책수립이 가능하다. 실험은 그림 1과 같은 Atari MS-Pacman 에서 진행되었으며 DQN 및 Vanilla-MFEC와 비교해서 매우 높은 점수를 획득하였다.

2. 배경지식

2.1 Model-Free Episodic Control

Model-Free Episodic Control(EC)[4]은 구글 딥마인드에서 2016년에 발표한 모델이다. 사람은 반복되는 일상에서 유사한 상황들을 기억에 의존해서 의사결정을 한다. 이러한 작용은 뇌 속의 해마에서 담당하며, EC는 이런 해마의 작동방식을 모델링한 알고리즘이다. 모델의 구현은 테이블 자료구조를 기반으로 만들어지며, 학습을 위해 사용되는 모델은 Q^{EC} 라는 이름을 가진다. 높은 메모리 요구량을

해결하기 위해 에이전트의 관측값(observation)들을 ϕ 함수(embedding)를 통해 저차원으로 축소하며, 이를 상태(state)로 정의한다. 실제 구현부에서는 각 액션마다 버퍼(buffer)를 하나씩 가지고 있으며, 버퍼의 전체 사이즈는 제한된다. 만약 버퍼의 크기를 초과하였을 경우엔 Least Recently Used(LRU) 기법을 통해 가장 참조되지 않은 데이터들에 새로운 정보를 갱신하게 된다. Q^{EC} 테이블은 상태를 입력으로 받아 각 액션들의 평가값을 산출하는 기능을 가지고 있다. 강화학습에서 이러한 경우 일반적인 정책은 최대값을 산출한 액션을 사용하고 있다. EC 또한 동일하게 $argmax$ 함수를 통해 결정된 액션을 현재 에이전트의 최적 행동으로 사용한다.

3. 연구 방법 및 실험

3.1 휴먼 플레이 데이터

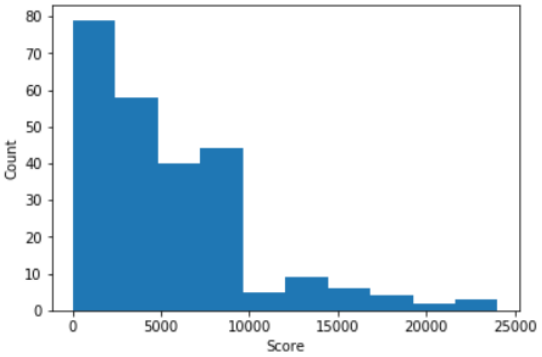

그림 2. Pacman 휴먼 플레이 데이터 히스토그램

휴먼 플레이 데이터는 에피소드 단위로 수집되었으며, 각 샘플은 1개의 생명(life)만을 사용해 플레이 되었다. 각 샘플은 84x84로 크기 변경된 흑백 이미지(observation)와, 행동(action), 보상(reward), 종료여부(terminal) 로 구성되어 있다. 각 샘플들의 최대생명은 1로 제한되어서 수집되었으며, 이유는 EC의 시연 동영상에서 모든 게임들의 에이전트들은 첫 목숨에서만 유효한 행동을 하였으며 이후 남은 생명들에서는 크게 점수를 획득하지 못하고 무작위 행동과 같은 행동을 하다가 끝나는 것을 확인하였기 때문이다. 이는 학습단계에서 생명을 1개만을 사용하여 이루어 졌다고 판단되었다. 따라서 기존 모델이 학습된 것과 동일하게 진행하기 위해 제한해 두었다.

Index	Score
Max	24031
Min	40
Average	5809

표 1. 휴먼 플레이 데이터 점수표

데이터셋은 Open AI Gym 플랫폼에서 유저 키 입력 에이전트를 만들어 수집하였다. MS-Pacman 환경에서 총 250개의 샘플 데이터를 수집했으며, 표1을 참고하면 데이터셋의 최고점, 최저점, 평균점수를 확인할 수 있다. 그림 2는 MS-Pacman 게임의 휴먼 플레이 데이터셋의 점수 히스토그램이며, 고득점으로 갈수록 플레이 데이터셋의 샘플의 수가 줄어든다.

3.2 ECFD

본 논문에서는 학습을 위한 모델의 ϕ함수는 Random Projection(RP)[5] 함수를 사용하였다. ECFD 모델은 거리비교를 통해 유사도를 측정하기 때문에, 고차원 입력벡터의 상대적인 거리를 어느정도 유지하면서 저차원으로 축소시킬 수 있는 RP 함수는 매우 효율적으로 사용된다. RP는 고차원 입력을 저차원으로 축소시키기 때문에 ECFD모델들의 높은 메모리 요구량을 해결할 수 있다.

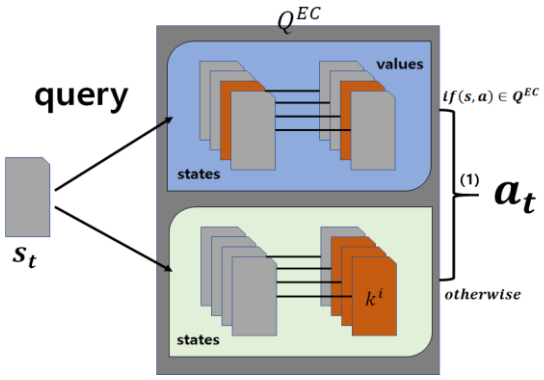

그림 3. ECFD의 쿼리 인터페이스

ECFD는 쿼리(query), 업데이트(update) 2가지 인터페이스를 가지고 있다. 쿼리는 모델에 입력을 주어 최적의 액션 선출하는 기능이며, 업데이트는 모델의 정책을 개선하는 것 이다. ECFD 모델의 업데이트는 EC모델의 업데이트를 그대로 따르고 있다. ECFD 모델은 몬테카를로 업데이트를 진행한다. 따라서 에피소드가 끝나고 나서 에피소드들을 역순으로 재구성하여, 업데이트를 진행한다. ECFD에서 t 는 time step을 의미하며, $s \in S$ 는 상태를 의미한다. $a \in A$ 는 에이전트가 수행할 수 있는 액션을 의미하며, Q^{EC} 는 학습이 진행되는 모델을 나타낸다. 쿼리기능은 그림 3에서 나타내는 것과 같이 두가지 경우가 존재하며, 그림 3에서 위의 경우는 현재 모델이 경험한 상태를 받았을 경우에 행동의 평가값을 그대로 사용하며, 아래의 경우는 경험하지 못한 상황에 대해서 비슷한 경험들을 통해 일반화 하는 것을 의미한다. 일반화는 RP 함수를 통해 나타난 상태들의 거리비교들을 통해 가까운 k 개의 상태들을 선출하고 이에 대응되는 평가값을 수식 1을 통해 평가한다. 이는 사람이 처음 겪는 상황에서 비슷한 상황을 통해 일반화 하는 것과 같다.

$$\widehat{Q^{EC}}(s,a) = \begin{cases} \frac{\alpha}{k}\sum_{i=0}^{k}\beta^i Q^{EC}(s^i,a) & if(s,a) \notin Q^{EC} \\ Q^{EC}(s,a) & otherwise \end{cases}$$

$$\beta = 1 - \frac{dis}{sdis}$$

$$sdis = \sum_{i=0}^{k} dis^i$$

수식 1. ECFD 쿼리 수식

수식 1에서 dis는 현재 상태와의 거리차를 의미하며, $sdis$ 는 k 개의 선출된 상태들의 거리합을 의미한다. ECFD에서는 β 를 가중치로서 사용하여, k 개의 모든 상태가 평가값에 동일하게 기여하지 않게 한다. α 값은 가중치로 인해 줄어든 출력값을 보정하기 위한 하이퍼 파라미터이다. ECFD 모델의 업데이트는 기존 EC와 동일하게 진행되며, 본 논문에서는 EC모델이 이미 휴먼 플레이 데이터로부터의 빠른 모방이 가능하다고 주장하며, 따라서 휴먼 학습 플레이 데이터셋을 미리 업데이트를 진행한 뒤 쿼리 기능을 수식1과 같이 변경하여 학습을 진행한다.

3.3 실 험

실험에 사용된 파라미터로는 $k = 7$, $frame\ skip = 4$, $\gamma = 1$, $\alpha = 1.5$ 그리고 $epsilon = 0.001$을 사용하였다. $epsilon$ 은 에이전트의 탐험(exploration)과 착취(exploit)의 비중을 담당하는 파라미터로 e-greedy 탐색을 위해서 사용된다. $epsilon$ 값은 1에서 시작하여 1만 프레임에 거쳐 0.001 값에 도달한다. $frame\ skip$ 은 이전 상태(s_t)에서 다음 상태(s_{t+1}) 사이의 무시하고 넘어가는 프레임들의 수를 의미하는 파라미터이다. 스킵핑(skipping)되는 구간의 에이전트의 액션 (a) 은 이전상태 (s_t)에서 수행한 액션과 동일한 액션 (a_t)으로 진행한다. γ은 할인율(discount factor)로서 미래가치를 현재가치로 환산하는데 사용되는 파라미터다. γ 는 일반적으로 0에서 1사이의 값을 사용한다.

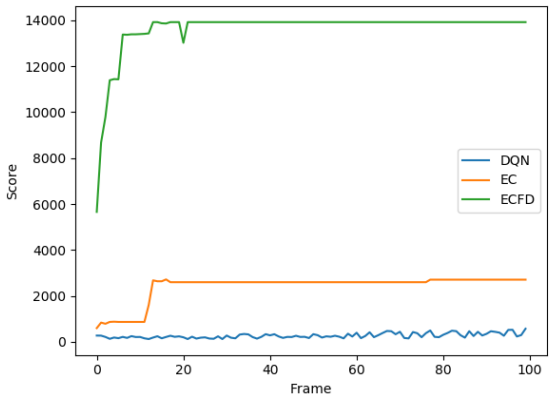

그림 4. ECFD 의 학습결과 및 비교

Model	Max Score
DQN	410
EC	2710
ECFD	13921

표 2. 학습된 모델들의 최고점

실험은 Atari MS-Pacman에서 진행되었으며, 그림 4는 100만 프레임을 학습한 모델들의 점수 비교를 보여준다. 학습된 모델들의 최고점은 표2를 통해 확인할 수 있다. DQN 모델은 100만 프레임에 도달하여서도 학습이 천천히 진행되고 있는 반면에 ECFD는 DQN 모델과 비교하여서는 13000점을 앞서 나가고 있으며, EC와 비교하여서는 약 5배 정도의 성능을 보였다.

4. 결론 및 향후 연구

본 논문에서는 기존의 느린 강화학습들의 대안으로 메모리 기반 모방학습이 가능한 ECFD 모델을 제안하였다. 휴먼 플레이 데이터셋을 선행학습한 ECFD 모델은 기존 모델들과 비교해 매우 좋은 성능을 보여주었으며, 기존 모델들이 학습에 많은 시간을 소모하는 것을 고려하면 ECFD의 빠른 학습은 매우 효율적이다. 이것은 휴먼 플레이 데이터로부터 ECFD가 매우 빠르게 성공적인 정책을 수립한 것을 확인할 수 있다. 하지만 기대와 달리 ECFD는 휴먼 플레이 데이터셋의 최고점에는 도달하지 못하였다. 따라서 후속 연구로는 수렴속도를 늦추며 좀 더 휴먼 플레이 데이터셋의 스펙을 도달할 수 있는 연구와 신경망 모델의 결합 가능성에 대해서 연구할 예정이다.

감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2017R1A2B4002164)

참고문헌

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

[2] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484.

[3] Ross, S., Gordon, G., & Bagnell, D. (2011, June). A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 627-635).

[4] Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., ... & Hassabis, D. (2016). Model-free episodic control. arXiv preprint arXiv:1606.04460.

[5] Bingham, E., & Mannila, H. (2001, August). Random projection in dimensionality reduction: applications to image and text data. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 245-250). ACM.