🧡

[11주차] Playing Atari with Deep Reinforcement Learning

0. Abstract

Deep Q-Network (DQN) 모델

- DeepMind에서 발표한 논문에 기반한 첫 번째 딥 러닝 모델
- 고차원 감각 입력으로부터 직접 제어 정책을 학습할 수 있는 강화 학습 모델
- 구성 요소
  - **Convolutional Neural Network (CNN)**: 원시 픽셀 입력을 처리하는 합성곱 신경망
  - **Experience Replay**: 에이전트의 경험을 저장, 무작위로 샘플링하여 학습에 활용
  - **Target Network**: 안정적인 학습을 위해 별도의 타깃 네트워크를 사용
- DQN 모델은 Atari 2600 게임 7종에 적용되었으며, 기존 접근 방식을 능가하는 성능을 보였고 3개 게임에서는 인간 전문가를 능가하는 결과를 달성함

DQN 모델의 핵심 특징

1️⃣ **고차원 감각 입력 처리**: DQN은 원시 픽셀 입력을 직접 처리할 수 있는 CNN 기반 모델로, 기존 접근 방식보다 고차원 감각 입력을 효과적으로 다룰 수 있음

2️⃣ **강화 학습을 통한 제어 정책 학습**: DQN은 강화 학습 기법 중 하나인 Q-learning을 활용하여 최적의 제어 정책을 학습함

3️⃣ **Experience Replay**: DQN은 에이전트의 경험을 저장하고 무작위로 샘플링하여 학습에 활용함으로써 데이터 효율성과 안정성을 높임

4️⃣ **Target Network**: DQN은 안정적인 학습을 위해 별도의 타깃 네트워크를 사용하여 Q-값 업데이트의 변동성을 줄임

DQN 모델의 응용 및 확장

- 단순한 게임 플레이 외에도 다양한 분야에 응용될 수 있음
- 로봇 제어, 자율 주행, 자연어 처리 등의 분야에서 DQN 기반 모델이 활용될 수 있음
- DQN 모델은 다양한 확장 연구를 통해 성능 향상 및 응용 범위 확대가 이루어지는 중

1. Introduction

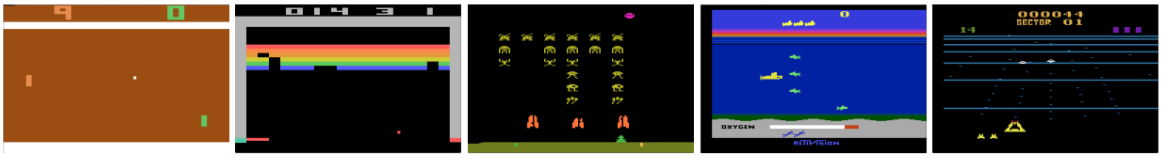

Figure 1: Screen shots from five Atari 2600 Games: (*Left-to-right*) Pong, Breakout, Space Invaders, Seaquest, Beam Rider

- **고차원 감각 입력 처리의 어려움**
  - 비전, 음성 등 고차원 감각 입력으로부터 제어 정책을 학습하는 것은 강화 학습의 오랜 과제
  - 기존 시스템은 수작업으로 설계한 특징을 사용했지만, 성능이 특징 표현의 품질에 크게 의존했음
- **딥러닝의 등장과 가능성**
  - 최근 딥 러닝 기술의 발전으로 원시 감각 데이터에서 고수준 특징을 추출할 수 있게 됨
  - 이에 따라 이러한 기술이 강화 학습에도 적용될 수 있을지 궁금해짐
- **강화 학습의 고유한 어려움**
  - 그러나 강화 학습은 딥 러닝이 직면한 몇 가지 고유한 어려움 有
  - 예를 들어 보상 신호의 지연, 데이터의 상관관계, 분포의 비정상성 등
- **DQN 모델의 제안**
  - 본 논문은 이러한 어려움을 극복하기 위해 합성곱 신경망 기반의 DQN(Deep Q-Network) 모델을 제안
  - DQN은 Q-learning 알고리즘을 사용하여 원시 비디오 데이터로부터 제어 정책을 학습함

- **Atari 2600 게임 적용 및 성능 평가**
  - DQN 모델을 Atari 2600 게임 7종에 적용
  - 기존 접근 방식을 능가하는 성능을 보였고, 3개 게임에서는 인간 전문가를 능가하는 결과를 달성

2. Background

- 강화 학습 개요
  - 에이전트가 환경과 상호작용하며 행동을 선택하고 보상을 받는 과정
  - 에이전트의 목표는 미래 보상을 최대화하는 것
- **Markov Decision Process (MDP)**
  - 강화 학습 문제를 모델링하는 프레임워크
  - 상태 s, 행동 a, 상태 전이 확률 $P(s'|s,a)$, 보상 $R(s,a,s')$로 구성
  - 에이전트는 상태 s에서 행동 a를 선택하고, 다음 상태 s'로 전이되며 보상 r을 받음
  - 에이전트의 목표는 미래 보상의 기대값을 최대화하는 것
- **최적 행동-가치 함수 (Q*)**
  - 특정 상태 s에서 행동 a를 선택했을 때 기대되는 미래 보상의 최대값
  - $Q^*(s,a) = \max_\pi E[R_t|s_t=s, a_t=a, \pi]$
  - 벨만 방정식을 만족: $Q^*(s,a) = E[r + \gamma \max_{a'} Q(s',a')|s,a]$
- **Q-learning 알고리즘**
  - 벨만 방정식을 이용해 Q-함수를 반복적으로 업데이트
  - 선형 또는 비선형 함수 근사기를 사용하여 Q-함수 근사
  - 확률적 경사 하강법으로 Q-네트워크 학습
  - 모델 프리(model-free) 및 오프 정책(off-policy) 알고리즘

3. Related Works

- **TD-gammon**: 강화 학습의 성공 사례

    - **TD-gammon**은 백게몬 게임을 완전한 강화 학습과 자기 대결을 통해 인간 수준을 뛰어넘는 성과를 달성했음
    - Q-learning과 유사한 모델 프리 강화 학습 알고리즘을 사용했고, 다층 퍼셉트론으로 가치 함수를 근사함
- 강화 학습의 한계
    - TD-gammon의 성공에도 불구하고, 체스, 고, 체커 등 다른 게임에서는 동일한 방법론이 성공적이지 못했음
    - 이는 **TD-gammon의 접근법이 특수한 경우에만 효과적**이라는 인식을 낳음
    - 특히 주사위 굴림으로 인한 상태 공간 탐험과 가치 함수의 부드러움이 TD-gammon의 성공 요인으로 여겨졌음
- 강화 학습과 비선형 함수 근사의 문제
    - **Q-learning**과 같은 모델 프리 강화 학습 알고리즘을 비선형 함수 근사기와 결합하면 Q-네트워크가 발산할 수 있음
    - 이후 대부분의 연구는 수렴 보장이 더 좋은 **선형 함수 근사기**에 집중했음
- 최근 동향: **Deep Reinforcement Learning의 부활**
    - 최근 다시 딥러닝과 강화 학습을 결합하는 연구가 활발해졌음
    - 환경 모델 추정, 가치 함수 추정, 정책 추정 등에 딥 신경망이 활용되고 있음
    - 발산 문제도 gradient temporal-difference 방법 등으로 부분적으로 해결되고 있음
- 유사 선행 연구
    - **Neural Fitted Q-learning (NFQ)**: 배치 업데이트 방식으로 Q-네트워크 학습
    - **Q-learning + 경험 재현 + 단순 신경망**: 저차원 상태 표현 사용
- Atari 2600 플랫폼
    - Atari 2600 게임을 강화 학습 테스트베드로 활용하는 연구가 진행되어 왔음
    - 선형 함수 근사와 일반적인 시각 특징을 사용한 접근, 더 많은 특징과 해싱 기법 활용, HyperNEAT 진화 아키텍처 등이 시도되었음

4. Deep Reinforcement Learning

- Deep Reinforcement Learning의 동기

    - 컴퓨터 비전과 음성 인식 분야에서 깊은 신경망을 이용한 접근이 큰 성과를 거두었음
    - 이를 바탕으로 강화 학습에도 깊은 신경망을 적용하고자 함
    - TD-Gammon과 같은 기존 접근법을 현대적인 깊은 신경망 아키텍처와 강화 학습 알고리즘으로 발전시키고자 함
- Deep Q-learning 알고리즘의 제안
    - **경험 재현**(Experience Replay) 기법 사용
        - 에이전트의 경험(상태, 행동, 보상, 다음 상태)을 데이터셋에 저장
        - 무작위로 샘플링하여 Q-learning 업데이트 수행
    - 장점

      1️⃣
      데이터 효율성 향상: 하나의 경험이 여러 번 사용됨

      2️⃣ **상관관계 감소**: 무작위 샘플링으로 연속적인 샘플의 강한 상관관계 해결

      3️⃣ **발산 방지**: 온-정책 학습의 피드백 루프 문제 해결
- 구현 상의 제한사항
    - 메모리 버퍼 크기 N으로 인해 최근 경험만 저장되고 **오래된 경험은 삭제**됨
    - 균일 샘플링으로 모든 경험에 동일한 **중요도를 부여**함
    - **더 발전된 샘플링 전략**(우선순위 sweeping 등)이 필요할 수 있음

4.1 Preprocessing and Model Architecture

- **Preprocessing**
    - Atari 게임 프레임은 210 x 160 픽셀의 128색 팔레트 이미지로, 계산량이 많음
    - 따라서 입력 차원을 줄이기 위해 다음과 같은 전처리 과정을 거침

      1️⃣ RGB 이미지를 그레이스케일로 변환

      2️⃣ 110 × 84 크기로 다운샘플링

      3️⃣ 84 × 84 영역을 잘라내어 입력으로 사용
    - 전처리 과정은 GPU 기반 2D 컨볼루션 연산을 위해 필요
- **모델 아키텍처**
    - Q 함수를 신경망으로 모델링하는 방식에 대한 접근법

1️⃣ **상태와 행동을 입력**으로 받는 방식: 각 행동에 대한 Q값을 계산하려면 별도의 순전파가 필요해 **비효율적**

2️⃣ **상태만을 입력**으로 받고 **각 행동에 대한 Q값을 출력**하는 방식: 한 번의 순전파로 모든 행동의 Q값을 계산할 수 있어 **효율적**

➡️ **본 논문에서는 두 번째 접근법을 사용**

1️⃣ **입력**: 84 x 84 × 4 이미지 (전처리된 최근 4프레임)

2️⃣ **은닉층 1**: 16개의 8 x 8 필터, stride 4, ReLU 활성화

3️⃣ **은닉층 2**: 32개의 4 x 4 필터, stride 2, ReLU 활성화

4️⃣ **은닉층 3**: 256개의 완전연결 ReLU 유닛

5️⃣ **출력층**: 행동 개수(4~18개)에 해당하는 선형 출력 유닛

5. Experiments

- 7개의 인기 있는 Atari 게임(Beam Rider, Breakout, Enduro, Pong, Q*bert, Seaquest, Space Invaders)에 대한 실험을 수행
- 모든 게임에 동일한 신경망 아키텍처, 학습 알고리즘, 하이퍼파라미터를 사용
- 이를 통해 게임 특화 정보 없이도 다양한 게임에서 강건하게 작동하는 것을 보여줌
- 다만 보상 구조에 대해서는 게임 간 스케일 차이를 줄이기 위해 학습 중에만 양의 보상은 1, 음의 보상은 -1로 클리핑

5.1 Training and Stability

- 지도 학습과 달리, 강화 학습에서는 학습 중 에이전트의 성능을 추적하기 어려움
- 두 가지 지표 사용
  - 에피소드 당 평균 총 보상: 노이즈가 심해 학습 진행을 파악하기 어려움
  - 예측된 Q값의 평균: 더 안정적으로 증가하는 것을 확인할 수 있음
- 실험 결과, 이 방법은 강화 학습 신호와 확률적 경사 하강법을 사용하여 큰 신경망을 안정적으로 학습할 수 있었음

5.2 Visualizing the Value Function

Figure 3: The leftmost plot shows the predicted value function for a 30 frame segment of the game Seaquest. The three screenshots correspond to the frames labeled by A, B, and C respectively.

- Seaquest 게임에서 학습된 가치 함수(Value Function)를 시각화한 결과
- 화면 왼쪽에 적이 나타나면(A 지점) 예측된 가치가 급격히 상승
- 에이전트가 적에게 어뢰를 발사하면 어뢰가 적중하는 순간(B 지점) 가치가 최고조에 도달
- 적이 사라지면(C 지점) 가치가 다시 원래 수준으로 떨어짐
- 이를 통해 제안한 방법이 비교적 복잡한 일련의 사건에 대한 가치 함수의 변화를 학습할 수 있음을 보여줌

5.3 Main Evaluation

	B. Rider	Breakout	Enduro	Pong	Q*bert	Seaquest	S. Invaders
Random	354	1.2	0	−20.4	157	110	179
Sarsa [3]	996	5.2	129	−19	614	665	271
Contingency [4]	1743	6	159	−17	960	723	268
DQN	**4092**	**168**	**470**	**20**	**1952**	**1705**	**581**
Human	7456	31	368	−3	18900	28010	3690
HNeat Best [8]	3616	52	106	19	1800	920	**1720**
HNeat Pixel [8]	1332	4	91	−16	1325	800	1145
DQN Best	**5184**	**225**	**661**	**21**	**4500**	**1740**	1075

Table 1: The upper table compares average total reward for various learning methods by running an ϵ-greedy policy with $\epsilon = 0.05$ for a fixed number of steps. The lower table reports results of the single best performing episode for HNeat and DQN. HNeat produces deterministic policies that always get the same score while DQN used an ϵ-greedy policy with $\epsilon = 0.05$.

- 제안한 방법(DQN)의 성능을 기존 강화 학습 방법들과 비교
- Sarsa와 Contingency 방법은 게임 화면에 대한 사전 지식을 활용했지만, DQN은 raw RGB 스크린샷만을 입력으로 사용했음
- DQN이 모든 게임에서 다른 학습 방법들을 크게 앞섬
- DQN은 Breakout, Enduro, Pong 게임에서 전문가 인간 플레이어를 능가하고, Beam Rider에서는 인간 수준에 근접함

- Q*bert, Seaquest, Space Invaders 게임에서는 인간 수준에 미치지 못하는데, 이는 장기적인 전략이 필요한 게임이기 때문
- DQN은 진화 알고리즘 기반 방법(HNeat)보다도 대부분의 게임에서 우수한 성능을 보임

6. Conclusion

- 본 논문은 강화 학습을 위한 새로운 딥러닝 모델을 소개함
- 이 모델은 Atari 2600 게임에서 원시 픽셀만을 입력으로 사용하면서도 어려운 제어 정책을 마스터할 수 있음을 보여주었음
- 또한 경험 리플레이 메모리와 결합된 온라인 Q-learning의 변형을 제시했는데, 이를 통해 강화 학습을 위한 딥 네트워크 학습을 용이하도록 함
- 제안한 접근 방식은 7개의 게임 중 6개 게임에서 최신 기술 수준을 능가하는 결과를 보였으며, 이는 모델 아키텍처나 하이퍼파라미터를 조정하지 않고도 달성한 것

논문에 대한 의견 및 의문점(꼭지)

➡️ 다양한 환경 및 과제로의 확장이 필요해보임. 현재는 Atari 게임과 같은 제한적인 환경에서 주로 연구가 진행되고 있지만, 향후에는 실제 세계의 복잡한 과제와 환경에서의 강화 학습 적용이 중요해질 것이라고 생각함.