Week 5 [Atari with DRL]

paper : https://arxiv.org/abs/2010.11929

Abstract

이 비디오는 **강화 학습**을 통해 고차원 감각 입력에서 제어 정책을 학습하는 첫 번째 심층 학습 모델인 **딥 Q-러닝**의 혁신적 적용을 설명합니다. 이를 통해 아타리 2600 게임에서 인간 전문 선수보다 뛰어난 성과를 달성하며, 이 기술이 데이터 효율성을 높이고 학습 알고리즘의 성능을 향상시킬 수 있는 가능성을 보여줍니다. 비디오를 시청함으로써 시청자는 **기계 학습**과 **영상 인식**의 최전선에서 이루어지는 연구 결과에 대한 깊은 통찰을 얻을 수 있습니다.

핵심 용어

- 강화 학습: 강화 학습은 기계가 주변 환경과 상호작용하면서 보상을 최대화하기위한 행동을 배우는 방법입니다.

Key Topic

1. 🎮 강화 학습과 딥 Q-러닝의 적용

- **딥 Q-러닝**은 고차원 감각 입력에서 제어 정책을 학습하는 혁신적인 심층 학습 모델이다 .
- 이 연구는 아타리 2600 게임에서 인간 전문 선수보다 우수한 성과를 달성하였다 .
- 따라서, 이러한 기술은 데이터 효율성을 향상시키고 학습 알고리즘의 성능을 높일 수 있는 가능성을 보여준다 .
- 연구는 고차원 감각 입력을 통해 **강화 학습**의 새로운 접근 방식을 제시하고 있다 .

2. 🚀 심층 강화 학습의 새로운 접근과 도전 과제

- 이 연구는 **고차원 감각 입력**을 통해 제어 정책을 학습하는 혁신적인 **딥 Q-러닝** 모델을 제안한다 .
- 제안된 모델은 **합성곱 신경망**으로 구성되어 있으며, **원시 픽셀**을 입력으로 사용하고 **미래 보상**을 추정하는 가치 함수를 출력한다 .
- 이 모델은 **아타리 2600 게임 7개**에 적용되었으며, 아키텍처나 학습 알고리즘의 조정 없이도 이전 방법들을 뛰어넘는 성과를 보여주었다 .
- 강화 학습 분야에서는 감각 데이터로부터 제어 에이전트를 학습하는 데 어려움이 있으며, 대부분의 성공적인 응용 프로그램은 수작업으로 만든 특징에 의존하고 있다 .
- 또한, 강화 학습은 **보상 신호**가 드물고, 노이즈가 있으며, 지연되는 등 심층 학습 관점에서 다양한 도전 과제를 안고 있다 .

3. 🎮 딥 Q-러닝을 통한 아타리 게임의 제어 정책 학습

- 이 논문은 **컨볼루션 신경망**이 복잡한 강화 학습 환경에서 원시 비디오 데이터로부터 성공적인 제어 정책을 학습할 수 있음을 보여준다 .
- **Q-러닝 알고리즘**의 변형을 사용하여 네트워크의 가중치를 업데이트하고, 경험 재생 메커니즘을 적용하여 과거 행동을 무작위로 샘플링함으로써 학습 분포를 매끄럽게 만든다 .
- 아타리 2600 게임은 고차원 비주얼 입력을 가지며, 다양한 난이도의 작업을 제공하여 에이전트의 훈련에 도전적인 환경을 제공한다 .
- 학습은 네트워크가 비디오 입력과 보상 신호만을 사용하여 진행되며, 인간 플레이어와 유사하게 주어진 행동 집합으로부터 선택된다 .
- 이 네트워크는 지금까지 시도한 7개 게임 중 6개에서 이전 모든 강화 학습 알고리즘을 초과하며, 3개 게임에서 전문가 인간 플레이어를 능가하였다 .

3.1. 🧠 강화 학습의 문제 영역

- 알고리즘이 새로운 행동을 학습함에 따라 **분포의 변화**가 발생하며, 이는 고정된 분포를 가정하는 심층 학습 방법에 **문제를** 일으킬 수 있다 .

3.2. 🎮 딥 Q-러닝을 통한 게임 학습 모델

- 이 연구는 **합성곱 신경망**(CNN)을 활용하여 복잡한 **강화 학습** 환경에서 원시 비디오 데이터로부터 효과적인 제어 정책을 학습하는 방법을 보여준다 .

- 네트워크는 **Q-러닝** 알고리즘의 변형을 사용하며, 확률적 경량 경량 하강법으로 가중치를 업데이트하는 방식으로 학습한다 .
- 훈련 분포를 부드럽게 하고 상관된 데이터 문제를 줄이기 위해, 기억 재생 메커니즘을 사용하여 이전 전이들을 무작위로 샘플링한다 .
- 연구는 다양한 아타리 2600 게임을 대상으로 하며, 네트워크는 오직 비디오 입력, 보상 및 터미널 신호를 통해 학습하게 된다 .
- 결과적으로, 네트워크는 시도한 여섯 개 게임 중 다섯 개에서 이전 모든 강화 학습 알고리즘을 초월하였고, 세 개의 게임에서는 전문가 인간 플레이어를 초과하는 성과를 보였다 .

3.3. 📊 강화 학습의 공식 구조

- **강화 학습**은 에이전트가 **환경**과 상호작용하며, 행동, 관찰 및 보상으로 이루어진 일련의 작업을 고려한다 .
- 에이전트는 각 시간 단계에서 **합법적인 게임 행동**의 집합에서 하나의 행동을 선택하고, 이 행동은 **에뮬레이터**의 내적 상태와 게임 점수에 변화를 준다 .
- 에이전트는 현재 화면의 **픽셀 값**을 나타내는 벡터를 관찰하고, 점수의 변화인 보상을 수신하며, 이는 **부분 관찰된 작업**으로 이해된다 .
- 상태의 전체 이력을 통해 **마르코프 결정 과정(MDP)**을 정의하고, 행동 가치 함수인 **Q-함수**를 사용하여 미래의 보상을 극대화하는 전략을 학습한다 .
- 이 과정에서 **벨만 방정식**을 통해 최적의 행동 가치 함수를 추정하고, 이를 활용한 **Q-네트워크**는 신경망을 활용하여 심층 학습된 모델을 통해 정확한 반환을 예측하면서 학습한다 .

3.4. 📚 강화 학습의 역사적 발전과 현재

- **TD-Gammon**은 전적으로 강화 학습과 자기 플레이를 통해 학습하여 초인적 수준의 바둑을 달성한 가장 잘 알려진 사례이다 .
- 그러나 체스, 바둑, 체커에 같은 방법을 적용한 초기 시도는 성공적이지 않았으며, 이는 TD-Gammon 접근법이 특정 게임에만 효과적이라는 믿음을 불러일으켰다 .
- 최근에는 심층 신경망을 결합한 강화 학습이 부활하였고, 이러한 방법들은 비선형 함수 근사기를 사용하는 경향이 있다 .
- 경험 재생(experience replay) 기법을 이용하여, 데이터 세트에서 에이전트의 경험을 저장하고 이를 통해 학습하는 이점이 있으며, 이는 **데이터 효율성**을 높이는 데 기여한다 .

- 마지막으로, 경험 재생을 통한 오프 정책 학습의 필요성이 대두되면서, Q-러닝을 선택하는 방향으로 발전하고 있다 .

3.5. 🖥️ Atari 게임을 위한 데이터 전처리 및 신경망 아키텍처

- **원시 아타리 프레임**은 210x160 픽셀 크기의 이미지로서, 기본적인 전처리 단계를 거쳐 입력 차원을 줄이게 된다 .
- 원시 프레임은 먼저 RGB 표현에서 **흑백(gray-scale)**로 변환되고, 110x84 이미지로 다운샘플링된다 .
- 마지막 입력 표현은 플레이 영역을 대략적으로 포착하는 84x84 영역을 잘라내어 얻어진다 .
- 이 논문에서는 **Q-함수**에 대한 입력을 생성하기 위해 마지막 4프레임의 전처리를 수행하고 이를 스택한다 .
- 신경망은 가능한 모든 행동을 위한 별도의 출력 유닛을 통해 Q값을 예측하며, 이를 통해 각 상태에서 모든 행동의 Q값을 단일 **순전파**로 계산할 수 있는 장점이 있다 .

Conclusion

1. 🎮 DQN의 실험적 성과와 평가 방법

- **Deep Q-Networks (DQN)**은 아타리 게임에서 다양한 실험을 통해 잘 작동하며, 일곱 개의 인기 게임에서 동일한 네트워크 아키텍처와 학습 알고리즘을 사용하였다 .
- 게임 훈련 중 **보상 구조**를 수정하였으나, 모든 긍정적인 보상을 1로 고정하여 학습 과정의 안정성을 높여 오류의 파생물 규모를 제한하였다 .
- DQN은 **RMSProp 알고리즘**을 사용하여 훈련하는 동안 에이전트의 행동 정책을 조정하였고, 총 1000만 프레임을 통해 훈련하였다 .
- 결과적으로, DQN은 아타리 게임들에서 다른 학습 알고리즘보다 성능이 우수하며, **사람 플레이어**에 비해 높은 평균 보상을 달성하였다 .
- 훈련 중 최대 예측 Q 값의 변화를 통해 DQN의 안정적인 개선을 측정할 수 있었으며, 이 방법은 이론적 수렴 보장을 결여하고 있음에도 불구하고 안정적으로 동작하였다 .

2. 🕹️ 딥 Q-러닝의 성과와 발전 방향

- 이 연구는 **강화 학습**을 위한 새로운 **딥 러닝 모델**을 소개하고, 아타리 2600 컴퓨터 게임에서 도전적인 제어 정책을 마스터하는 능력을 입증하였다 .
- 연구팀은 **온라인 Q-러닝**의 변형을 제시하였으며, 이는 확률적 미니배치 업데이트와 경험 재생 메모리를 결합해 깊은 네트워크의 학습을 용이하게 한다 .
- 이 방법은 테스트된 일곱 개 게임 중 여섯 개에서 최첨단 성과를 보여주었고, 아키텍처와 하이퍼파라미터 조정 없이도 결과를 달성하였다 .
- 그러나 여전히 Q*bert, Seaquest, Space Invaders와 같은 게임들은 인간 성능에 미치지 못하며, 이러한 게임들은 네트워크가 장기간 전략을 발견해야 하는 도전 과제가 있다 .
- 이러한 실험 결과들은 이론적 기여를 기반으로 향후 연구 개발에서의 기회를 강조하고 있으며, 필요성이 부각된다 .