2015년도에 Google Deepmind에서 DQN을 개발한 이후 연구자들은 DQN이 안고 있는 문제와 개선점을 찾기 시작했다. 이에 2015년부터 2017년도까지의 기간동안 DQN과 관련하여 작성된 논문의 일부를 서문만 번역해서 옮겨둔다.

Deep Reinforcement Learning with Double Q-learning

Q-학습 알고리즘은 특정 상황하에서 행동값을 이상하게 예측하는(overestimate) 경향이 있는 것으로 알려져 있다. 또한 이러한 이상예측이 흔히 일어나는지, 성능에 악영향을 미치는지, 아니면 전반적으로 예방될 수 있는지에 대해 알려져 있지 않았다. 이 논문에서는 위의 세 질문에 모두 '그렇다'고 답한다. 특별히 우리는 최근(2015)에 연구된 Q-학습을 deep 뉴럴네트워크와 결합한 DQN 알고리즘이 몇몇 아타리 2600게임에서 심각한 이상예측에 시달렸음을 최초로 보였다. 그 뒤 우리는 Double Q-학습 알고리즘의 기반에 있는 tabular setting으로 소개된 대규모 함수 근사에 사용될 수 있는 아이디어를 보인다. 또한 이 논문에서 우리는 특수한 형태로 변형된 DQN알고리즘이 관측되는 이상예측을 줄일 수 있을 뿐 아니라 이러한 알고리즘이 몇몇 게임에서 월등한 성능을 보였음을 설명한다.

Continuous control with deep reinforcement learning

우리는 Deep Q-학습 알고리즘의 성공의 기반이 되는 아이디어를 continuous action 분야에도 적용해봤다. 이에 우리는 continuous action분야에서 동작할 수 있는 deterministic policy gradient에 기반한 actor-critic, model-free알고리즘을 선보인다. 동일한 학습 알고리즘과 네트워크 구조, 하이퍼 파라메터들을 사용하는 우리의 알고리즘은 cartpole swing-up, dexterous manipulation, legged locomotion과 차량 운전등을 포함하는 20개가 넘는 물리 시뮬레이션 문제를 해결할 수 있었다. 또한 우리의 알고리즘은 환경에 대한 모든 정보를 갖고 있는 상황에서 잘 설계된 알고리즘의 성능에 필적하는 행동정책을 찾아낼 수 있음을 보였고, 더 나아가 다양한 과제에서 우리의 알고리즘이 화면 픽셀 입력값만으로도 단대단 학습을 이용, 행동정책을 학습할 수 있음을 보였다.

Continuous Deep Q-Learning with Model-based Acceleration

최근 다양한 분야에 Model-free 강화학습이 성공적으로 적용되는 한편 무거운 뉴럴네트워크와 value function도 사용할수 있는 수준까지 확장이 이뤄지고 있다. 그러나 특히 고차원 함수 근사를 이용할때 문제가 되는 model-free 알고리즘의 샘플 복잡도가 물리학적 시스템에 적용되는걸 막고 있었다. 이 논문에서 우리는 연속 통제(continuous control) 과제에 적합하도록 deep 강화학습의 샘플 복잡도를 줄이는 알고리즘과 대표값을 연구했다. 이에 우리는 이러한 알고리즘의 효율성을 높일 수 있는 두가지의 상호 보완적인 기술을 보인다. 첫째는 흔히 쓰이는 policy gradent와 actor-critic mothod를 대체할 NAF(normalized adantage functions)라고 불리는 Q-학습 알고리즘의 continuous varient를 만드는것이다. NAF representation은 Q-학습의 experience replay를 연속적인 과제에 적용할 수 있게 해주고, 이는 로봇 통제 시뮬레이션 과제에서 큰 성능 향상을 가져다 줬다. 더욱 더 접근법의 효용성을 높이기 위해 우리는 사전에 학습된 모델을 model-free 강화학습을 가속하는데 사용하는것을 연구해봤다. 여기서 우리는 iteratively refitted local linear 모델이 특히 유용한것과 그러한 모델이 적용될 수 있는 분야에서 학습속도가 현저히 빨라짐을 보였다.

Learning Tetris Using the Noisy Cross-Entropy Method

cross-entropy method는 효율적인 보편 최적화 알고리즘이다. 그러나 강화학습에 있어 사용처는 알고리즘이 빠름에도 불구하고 제한되어 있고, 이는 cross-entropy method가 자주 suboptimal policy에 수렴하기 때문이고, 이러한 조기 수렴을 막기 위해서 기본적으로 쓰이는 방법은 잡음을 삽입하는것이다. 이에 우리는 noisecross-entropy method를 테트리스에 적용하여 효율성을 선보이기로 했고, 알고리즘이 생성한 최종 행동정책은 이제껏 존재했던 강화학습 알고리즘들과 거의 두배의 성능차를 보이며 평균 300,000점을 기록했다.

Dueling Network Architectures for Deep Reinforcement Learning

최근 들어(2015년 후반) 강화학습에 deep representation을 성공적으로 적용한 사례가 나오고 있지만 대부분은 convolutional networks, LSTMs, auto-encoders등 종래의 구조를 답습하고 있다. 이 연구에서 우리는 새로운 model-free 강화학습을 위한 뉴럴 네트워크 구조를 제안한다. 우리의 dueling 네트워크는 두개의 개별적인 estimators로 대표된다. 하나는 상태값 함수(state value function), 다른 하나는 상태에 종속된 행동 우위 함수(action advantage function)다. 이러한 접근의 주된 이득은 행동에 대한 학습을 기저에 깔린 강화학습 알고리즘

을 크게 바꾸지 않고도 일반화 시킬 수 있다는 점이다. 이 연구의 결과는 이 구조가 대량의 비슷한 값을 갖는 행동 목록이 있을때 더 나은 행동 평가가 가능함을 보였고, 더 나아가 dueling 구조가 우리의 강화학습 agent가 아타리 2600을 수행하도록 설계된 최신 agent를 능가하는 성능을 낼 수 있게 해줌을 보였다.

Asynchronous Methods for Deep Reinforcement Learning

이 연구에서 우리는 deep뉴럴 네트워크 컨트롤러를 최적화 하기 위해 astnchronous gradient descent를 사용하는 개념상 간단하고 가벼운 deep 강화학습 프레임워크를 제안한다. 이에 우리는 asynchronous한 종류의 강화학습 알고리즘 4개를 선보이고(?) pararallel acrot-learners가 훈련에 있어 안정화 효과가 있어 4개의 방법이 모두 뉴럴 네트워크 컨트롤러를 학습시키는데 성공했음을 보인다. 또한 그중 가장 성능이 좋았던 actor-critic의 asynchronous한 종류는 최신 아타리 agent의 성능을 뛰어넘는 수준을 GPU를 사용치 않고 오직 한개의 멀티코어 CPU를 사용해 반절의 시간동안만 훈련하여도 달성할 수 있음을 보였다. 더 나아가서 우리는 asynchronous actor-critic의 경우 다양한 연속 모터 통제 과제와 무작위적인 3D 미로를 화면 입력만을 가지고 풀어나가는 과제에도 적응할 수 있음을 보일 수 있었다.