Summary for cs231n 14

강화학습(Reinforcing Learning)이란 environment와 agent가 state, action, reward를 반복적으로 주고받으며 문제 또는 목표를 수행하도록 학습하는 것을 의미한다. 예를 들어 알파고 문제에서 목표는 agent가 바둑 게임을 이기는 것이다. 바둑판 위에서 바둑알의 위치라는 상태가 주어졌을 때 agent는 다음 바둑알을 놓는 action을 취하게 되고, 모델이 잘 훈련되어 최적의 action을 수행했을 때 얻을 수 있는 보상은 게임을 이기는 것이다. Markov Decision Process는 이러한 상황을 수학적으로 정의한 것이 된다. Markov Decision Process에서 일반적으로 정의한 상황에서는 각 action을 취했을 때 가장 많은 reward의 합을 얻는 policy를 찾는 것을 목표로 하고, 이때 이 policy를 the optimal policy라고 한다. Markov Decision Process의 상황에서는 초기 상태, 확률 등으로 인해 임의성이 항상 발생한다. 이를 해결하고 optimal policy를 찾을 수 있는 방법은 reward의 기대값을 최대화하는 optimal policy를 찾는 것이다! State가 주어졌을 때의 reward를 정의하는 Value function과 State-action 쌍이 주어졌을 때 reward를 정의하는 Q-Value function가 있다. Q-Value function은 Bellman equation이라고도 한다. Bellman equation은 state-action이 주어졌을 때의 optimal policy는 reward의 기댓값을 최대화하는 것임을 시사한다. 문제는 Bellman equation이 계산 불가능한 식이라는 것이다. 이러한 문제를 해결하기 위해 neural network를 이용하고, 이를 Q-Learning이라고 한다. Q-Learning은 Bellman Equation을 통한 forward pass와 backward pass를 가능하게 하며, Q-learning 모델에 state로 사용되는 이미지의 feature를 추출해주면 Q-network를 통해 강화학습이 가능해진다.

그러나 Q-learning에도 단점은 있다. 매우 복잡해질 수 있다는 것이다. 따라서 policy 중에서도 최적의 policy를 학습하는 방법인 policy gradient 방법이 사용되고 있다. Reinforce algorithm은 policy gradient 식을 적분하여 policy가 주어졌을 때의 reward를 계산 한 것, 이를 다시 정리하여 gradient를 업데이트하기 위한 식으로 만든 것이 gradient estimator이고 이때 발생하는 고분산 문제를 해결하기 위해 variance reduction을 시행한다.

강화학습은 게임뿐만 아니라 Recurrent Attention Model처럼 이미지 인지 분야에도 쓰이고 있다.