한국컴퓨터정보학회 하계학술대회 논문집 제31권 제2호 (2023. 7)

OpenAI Gym 환경의 Acrobot에 대한 DQN 강화학습

강명주O
O청강문화산업대학교 게임콘텐츠스쿨
e-mail: mjkkang@ck.ac.krO

DQN Reinforcement Learning for Acrobot in OpenAI Gym Environment

Myung-Ju KangO
OSchool of Game, Chungkang College of Cultural Industries

● 요 약 ●

본 논문에서는 OpenAI Gym 환경에서 제공하는 Acrobot-v1에 대해 DQN(Deep Q-Networks) 강화학습으로 학습시키고, 이 때 적용되는 활성화함수의 성능을 비교분석하였다. DQN 강화학습에 적용한 활성화함수는 ReLU, ReakyReLU, ELU, SELU 그리고 softplus 함수이다. 실험 결과 평균적으로 Leaky_ReLU 활성화 함수를 적용했을 때의 보상 값이 높았고, 최대 보상 값은 SELU 활성화 함수를 적용할 때로 나타났다.

키워드: Acrobot(Acrobot), 활성화함수(Activation function), DQN(Deep Q-Networks)

I. Introduction

인공지능 분야에서 강화학습은 자율자동차, 로봇, 게임 등에서 많이 적용되는 알고리즘으로 에이전트의 현재 상태와 상호작용하며 보상을 최대화하는 행동을 학습하는 머신러닝의 한 분야이다.

본 논문에서는 OpenAI Gym 환경의 Acrobot-v1 게임[1]에 DQN(Deep Q-Network) 강화학습을 적용하였고, DQN에서 학습에 적용되는 활성화 함수의 성능을 비교 분석하였다.

II. Preliminaries

1. Deep Q-Networks

DQN은 [2]에서 처음 소개한 알고리즘으로 기존의 Q-leaning 알고리즘의 단점을 보완한 Deep Q-learning 알고리즘을 신경망에 접목한 강화학습 알고리즘이다.

기존의 Q-learning 알고리즘은 모든 상태(state)와 행동(action)을 Q-table로 정의함으로써 state와 action space가 커질수록 Q-table을 저장하는 메모리가 커져 탐색시간이 오래 걸리는 단점을 가지고 있다. 이러한 Q-learning의 단점을 보완하기 위해 Deep Q-learning 알고리즘에서는 (state, action)으로 구성된 Q-table에서 replay 메모리로 샘플링 추출하여 업데이트하는 방법을 사용한다.

DQN의 학습에 적용되는 손실함수는 샘플링 데이터를 타겟으로 하고 샘플링 데이터를 Q-network를 통해 학습한 데이터를 예측값으로 하여 타겟과 예측값의 오차를 최소화하도록 설계되어 있다[2]. [그림 1]은 DQN의 네트워크 구조이다.

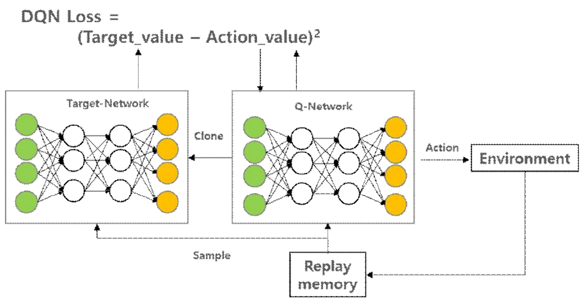

Fig. 1. structure of DQN

2. Acrobot-v1

Acrobot-v1[1]은 두 개의 링크에 선형적으로 연결된 체인 구조이다. 체인의 한쪽 끝은 고정되어 있고, 두 링크 사이에는 조인트가 작동된다. 이 게임의 목표는 체인이 아래쪽으로 매달린 초기 상태에서 시작하여 조인트에 토크를 가하여 끝에 있는 체인이 주어진 높이 이상으로 스윙하도록 하는 것이다. 두 링크 사이의 조인트에 적용되는 Action space는 [표 1]과 같고, Observation space는 [표 2]와 같다.

Table 1. Action space

Num	Action	Unit
0	apply -1 torque to the actuated joint	torque(N m)
1	apply 0 torque to the actuated joint	torque(N m)
2	apply 1 torque to the actuated joint	torque(N m)

Table 2. Observation space

Num	Observation	Min	Max
0	Cos(theta1)	-1	1
1	Sin(theta1)	-1	1
2	Cos(theta2)	-1	1
3	Sin(theta2)	-1	1
4	Angular velocity of theta1	~(-4*pi)	~(4*pi)
5	Angular velocity of theta2	~(-9*pi)	~(9*pi)

III. Experiments

1. Experiments Environments

본 논문에서는 OpenAI Gym에서 제공하는 Acrobot-v1[1]에 대해 DQNAgent를 이용하여 DQN 강화학습을 진행하였다. 학습에 적용된 활성화 함수는 [표 3]과 같다[3].

Table 3. Activation functions

Name	Equation
ReLU	$f(x)=\max(0,x)$
Leaky_ReLU	$f(x,\alpha)=\max(\alpha x,x)$
ELU	$f(x,\alpha)=\max(\alpha(e^x-1),x)$
SELU	$f(x,\alpha)=\lambda\times\max(\alpha(e^x-1),x)$
Softplus	$f(x,\beta)=\frac{1}{\beta}\log(e^{\beta x}+1)$

이 게임의 목표는 가능한 적은 스텝으로 끝 조인트가 지정된 목표 높이에 도달하는 것이다. 목표에 도달하지 못한 모든 스텝은 -1의 보상을 받고, 목표에 도달하면 0의 보상을 받아 종료된다. DQN을 통한 학습 횟수는 100,000회 진행하였다.

2. Experiments Results

본 논문에서는 활성화 함수가 학습에 끼치는 영향을 평가하기 위해 각 활성화 함수에 따른 학습에서의 보상에 대한 평균, 최소, 최대 값을 비교하였다. 실험 결과는 [표 4]와 같다. 실험 결과 평균적으로 Leaky_ReLU 활성화 함수를 적용했을 때의 보상 값이 높았고, 최대 보상 값은 SELU 활성화 함수를 적용할 때로 나타났다.

Table 4. Mean/Min/Max Reward for each activation function

Activation func	Mean	Min	Max
ReLU	-170.25	-500.0	-96.0
Leaky_ReLU	-162.79	-500.0	-91.0
ELU	-171.82	-500.0	-93.0
SELU	-168.39	-500.0	-79.0
Softplus	-200.48	-500.0	-90.0

[그림 2]는 각 에피소드에 따른 활성화함수에 대한 보상을 비교한 그래프이다.

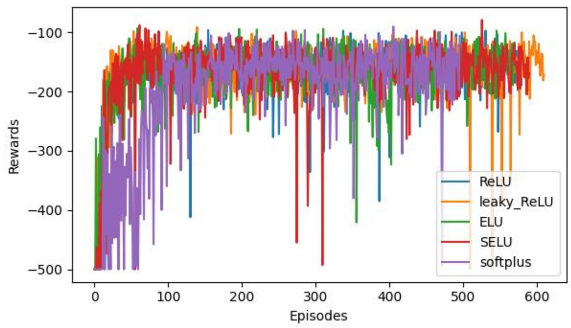

Fig. 2. Comparison of Rewards for each activation function according to episodes

IV. Conclusions

본 논문에서는 Acrobot-v1 게임에 대해 에이전트가 DQN 강화학습으로 학습할 경우 학습에 적용되는 활성화함수의 성능을 비교분석하였다. 실험 결과 평균적으로 Leaky_ReLU를 적용했을 때 보상 값이 높았고, 최대 보상 값은 SELU 활성화 함수를 적용할 때임을 알 수 있었다.

REFERENCES

[1] https://gymnasium.farama.org/environments/classic_control/acrobot/
[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstr, Martic Riedmiller, "Playing Atari with Deep Reinforcement Learning", arXiv:1312.5602v1, 2013
[3] https://keras.io/api/layers/activations/