컴퓨터 그래픽스와딥러닝

DQN Lunar Lander assignment report

학번: 202021698

학부: 컴퓨터정보공학부

이름: 변한빛

**목 차**

**실험 개요 2**

**추가 환경4**

**학습 과정(러닝레이트)5**

**학습 과정(엡실론)7**

**학습 과정(엡실론 디케이)9**

**학습 과정(감마)11**

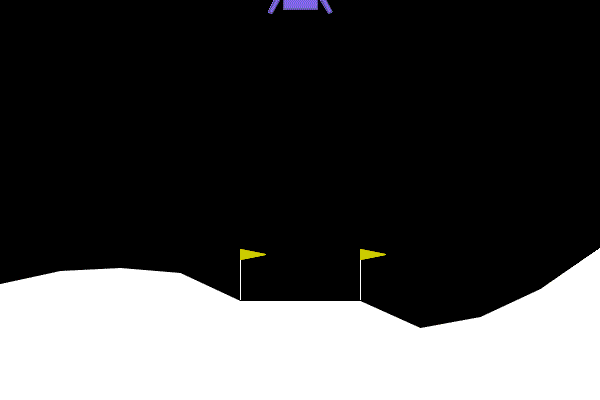
**학습 과정(뉴럴 네트워크)13**

**학습 결과(영상 비교)15**

**학습 후기17**

**실험 개요**

* Lunar Lander 소개



우주선을 조종해 깃발 사이에 안착하는 것을 목표로 함

|  |  |
| --- | --- |
| Action Space(4) | 오른쪽 ( 사이드 엔진 )  왼쪽 ( 사이드 엔진 )  위 ( 메인 엔진 )  정지 |
| Observation Shape(8) | 우주선 좌표(x, y)  우주선 속도(x, y)  각, 각속도  착륙 성공 여부(leg-left, leg-right) |
| Reward | 깃발 내 착륙 시(+100 ~ +140)  땅에 충돌 시(-100)  아무것도 안하고 정지 시(+100)  착륙 성공 시 우주선 다리 당(+10)  메인 엔진(위로 이동) 점화 시 프레임 당(-3)  사이드 엔진(옆 이동) 점화 시 프레임 당(-0.03) |

\* 순서는 임의로 지정함 실제 데이터 값과 순서는 다름

* 실험 목표

이번 실험의 목표는 DQN을 가지고 Lunar Lander를 풀기 위해

하이퍼 파라미터를 수정하면서 접근하되

빠른 컴파일 시간과 평균적으로 높은 보상을 중점으로 두고

최적의 하이퍼 파라미터값을 하나 하나 찾는 것을 목표로 한다.

\* 해당 실험에서는 빠른 학습 시간을 위해 Experience Replay 기법과 Double DQN 기법을 적용하지 않는다

* 실험 계획

1. 처음 설정을 EX03 CartPole DQN(<http://cg.catholic.ac.kr/~mgchoi/CourseDocs/CGDL/CartPoleDQN/>)을 초기 환경으로 두고 \*추가 환경을 세팅한다.

\* 추가 환경은 다음 페이지에 정리함

1. 러닝 레이트 값을 조정한 후 각각의 결과 값을 비교한다.
2. 엡실론 값을 조정한 후 각각의 결과 값을 비교한다.
3. 엡실론 디케이 값을 조정한 후 각각의 결과 값을 비교한다.
4. 뉴럴 네트워크 레이어 값을 조정한 후 각각의 결과 값을 비교한다.

\* 뉴럴 레이어는 학습 시간을 빠르게 하기 위해 2차원으로 한정한다.

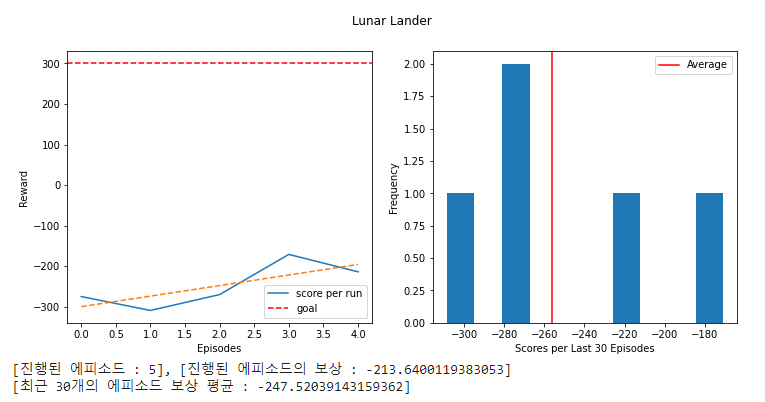
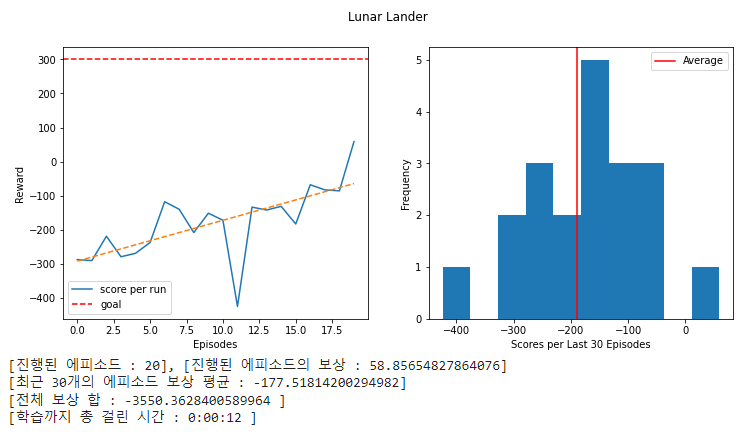
1. 에피소드 횟수를 조정한 후 각각의 결과 및 최종 영상을 비교한다.

**추가 환경**

* Episode 250 -> Episode 200으로 횟수 차감

\* 빠른 컴파일 결과를 위해 횟수를 차감한다.

* CartPole -> Lunar Lander 변경
* 최근 30개의 결과의 평균, 전체 보상 합, 학습까지 총 걸린 시간을 표시하는 창 추가

\* 최고 목표는 보상 값 계산에 의해 300에 거의 수렴하기에 300으로 둔다

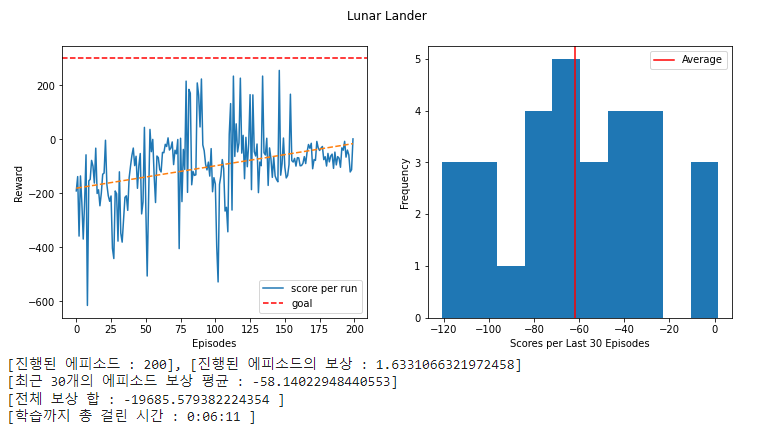
\* 학습 진행중

\* 학습 완료

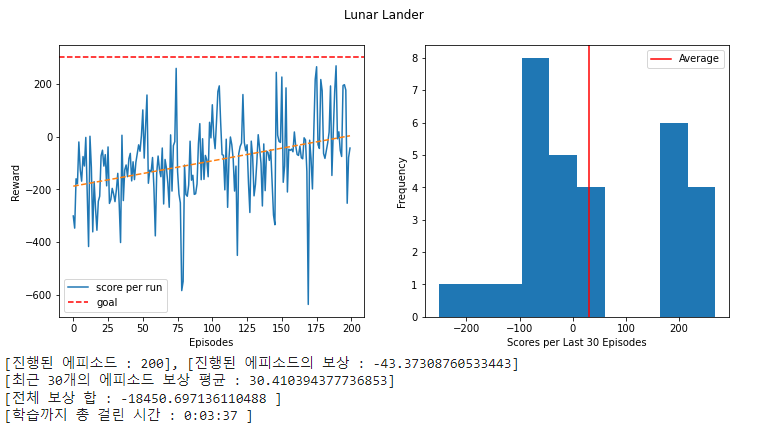
* 애니메이션 시드를 100부터 1500까지 50씩 더해가며 얻은 결과값들 중 가장 높은 보상 값이 나온 영상 결과를 표시



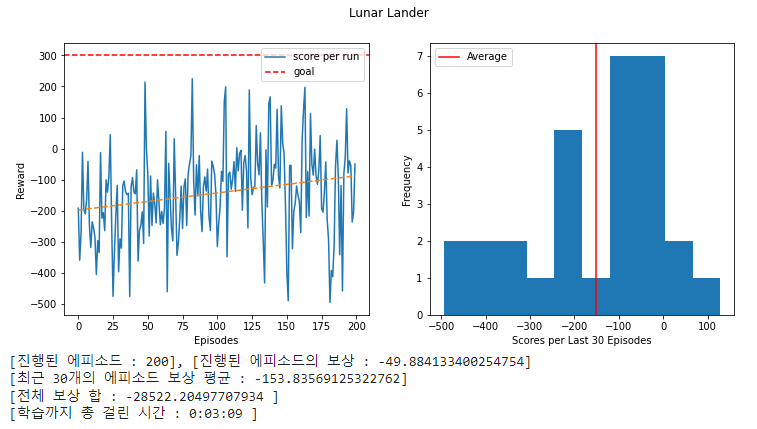
**학습 과정 ( 러닝 레이트 )**

* 러닝 레이트(LR) 변경

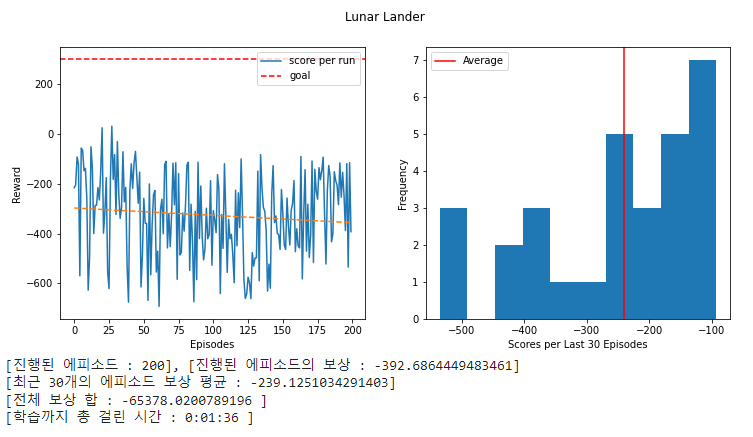
|  |  |
| --- | --- |
| LR : 0.001 | |
| 최근 30개의 에피소드 보상 평균 | -58.1 |
| 전체 보상 합 | -19685.5 |
| 학습까지 총 걸린 시간 | 00:06:11 |



|  |  |
| --- | --- |
| LR : 0.005 | |
| 최근 30개의 에피소드 보상 평균 | 30.4 |
| 전체 보상 합 | -18450.6 |
| 학습까지 총 걸린 시간 | 00:03:37 |



|  |  |
| --- | --- |
| LR : 0.01 | |
| 최근 30개의 에피소드 보상 평균 | -153.8 |
| 전체 보상 합 | -28522 |
| 학습까지 총 걸린 시간 | 00:03:09 |



|  |  |
| --- | --- |
| LR : 0.1 | |
| 최근 30개의 에피소드 보상 평균 | -239.1 |
| 전체 보상 합 | -65378 |
| 학습까지 총 걸린 시간 | 00:01:36 |

* 러닝 레이트(LR) 변경 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LR 0.001 | LR 0.005 | LR 0.01 | LR 0.1 |
| 최근 30개의 에피소드 보상 평균 | -58.1 | 30.4 | -153.8 | -239.1 |
| 전체 보상 합 | -19685.5 | -18450.6 | -28522 | -65378 |
| 학습까지 총 걸린 시간 | 00:06:11 | 00:03:37 | 00:03:09 | 00:01:36 |

러닝 레이트 변경 결과로 LR 0.005를 채택하기로 했다.

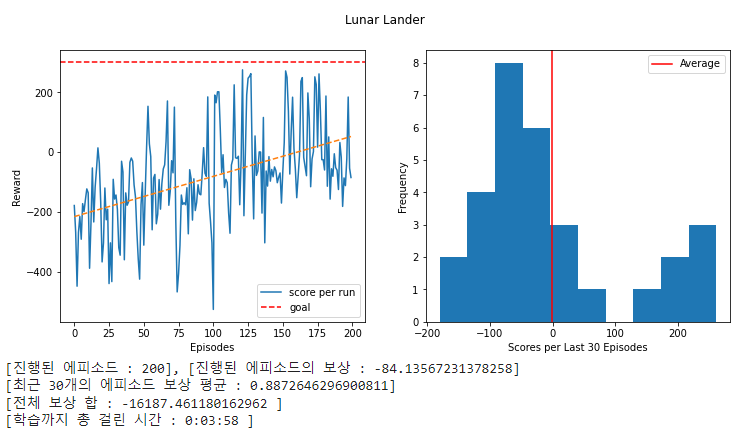
최근 30개의 에피소드 보상 평균 및 전체 보상 합이 제일 높으며

학습까지 총 걸린 시간 또한 적당하다고 판단되었기 때문이다.

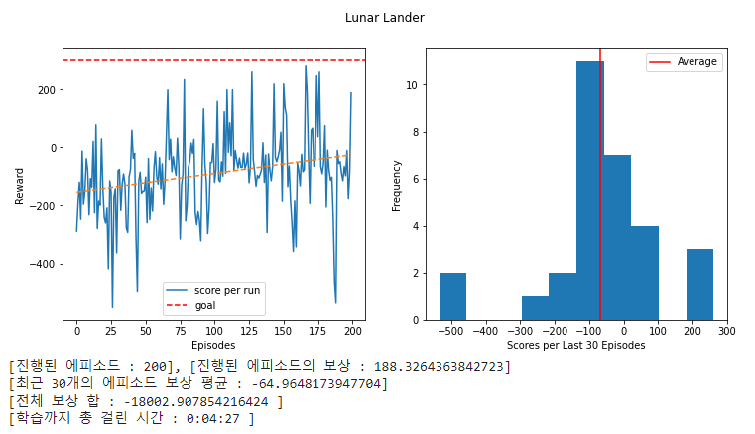
그 외에 학습까지 총 걸린 시간이 젤 낮은 LR 0.1은 LR 0.005와 비교해서

전체 보상 합과 최근 30개의 에피소드 보상 평균이 지나치게 낮아서 채택하지 않았다.

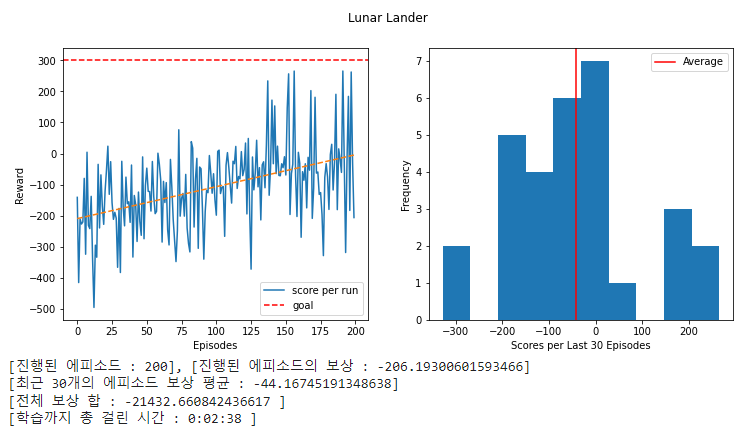
**학습 과정 ( 엡실론 )**

* 엡실론(EPS) 변경

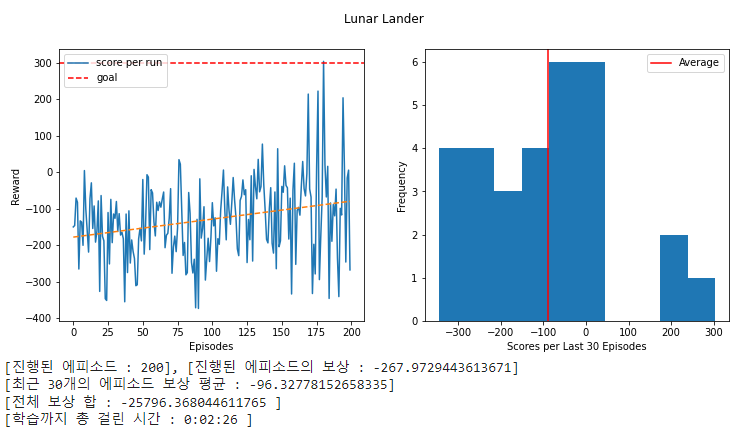
|  |  |
| --- | --- |
| EPS : 0.2 | |
| 최근 30개의 에피소드 보상 평균 | 0.8 |
| 전체 보상 합 | -16187.4 |
| 학습까지 총 걸린 시간 | 0:03:58 |



|  |  |
| --- | --- |
| EPS : 0.4 | |
| 최근 30개의 에피소드 보상 평균 | -64.9 |
| 전체 보상 합 | -18002.9 |
| 학습까지 총 걸린 시간 | 0:04:27 |



|  |  |
| --- | --- |
| EPS : 0.6 | |
| 최근 30개의 에피소드 보상 평균 | -44.1 |
| 전체 보상 합 | -21432.6 |
| 학습까지 총 걸린 시간 | 00:02:38 |



|  |  |
| --- | --- |
| EPS : 0.8 | |
| 최근 30개의 에피소드 보상 평균 | -96.3 |
| 전체 보상 합 | -25796.3 |
| 학습까지 총 걸린 시간 | 00:02:26 |

* 엡실론(EPS) 변경 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | EPS 0.2 | EPS 0.4 | EPS 0.6 | EPS 0.8 |
| 최근 30개의 에피소드 보상 평균 | 0.8 | -64.9 | -44.1 | -96.3 |
| 전체 보상 합 | -16187.4 | -18002.9 | -21432.6 | -25796.3 |
| 학습까지 총 걸린 시간 | 0:03:58 | 0:04:27 | 00:02:38 | 00:02:26 |

엡실론 변경 결과로 EPS 0.6을 채택하기로 했다.

EPS 0.2, EPS 0.4에 비해 전체 보상 합이 낮지만 학습까지 총 걸린 시간이 EPS 0.2, EPS 0.4에 비해 빠르며

히스토그램 그래프도 EPS 0.2에 견줄 수 있을 정도로 높은 값이 관찰되었기 때문이다.

그 외에 학습까지 총 걸린 시간이 제일 낮은 EPS 0.8은 EPS 0.6에 비해 히스토그램 그래프의 높은 값이

낮에 관찰되기 때문에 채택하지 않았다.

여기서 Q-Learning의 MDPs(Markov Decision Processes)의 특징이 보인다.

EPS의 값이 높으면 높을수록 학습시간이 짧아지는 것이 관찰되는데

EPS가 높으면 학습의 불확실성이 높아 초반에 우주선이 금방 추락해 학습이 짧아지지만

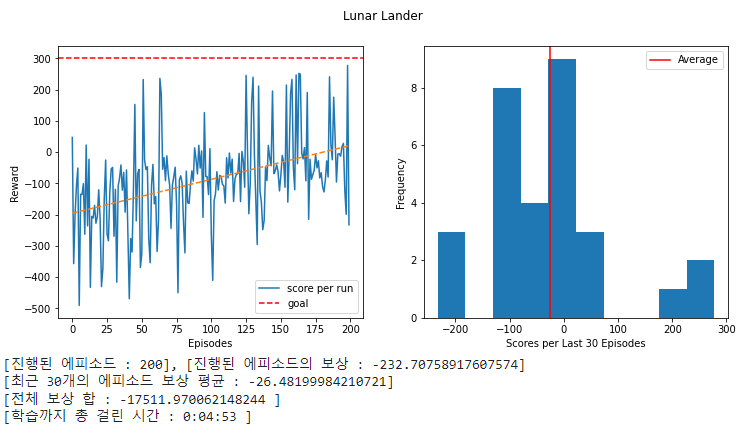
엡실론 디케이에 의해 학습이 점점 안정을 가져 후반에 그래프가 진동한다.

이 부분에 대해서는 디케이 값을 조정해 다음 페이지에서 보완할 예정이다.

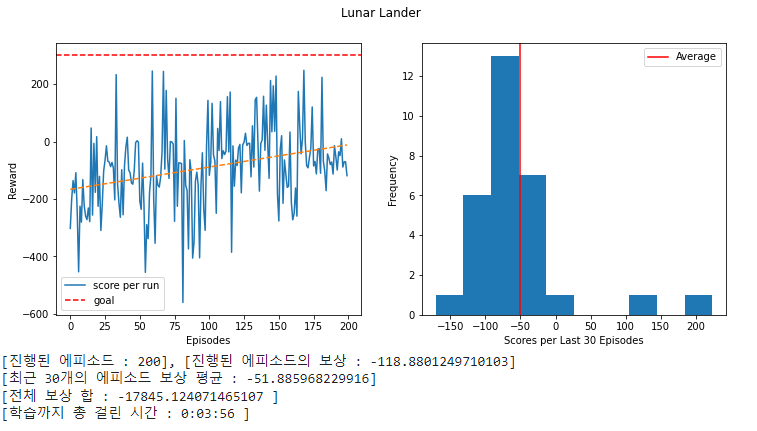
**학습 과정 ( 엡실론 디케이 )**

* 엡실론 디케이(E-DECAY) 변경

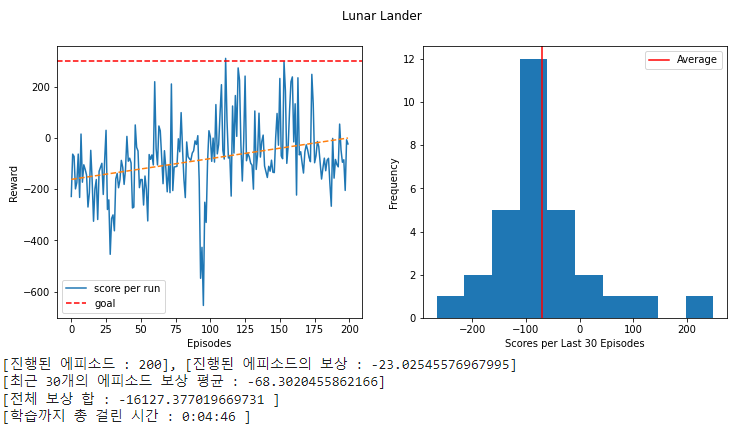
|  |  |
| --- | --- |
| E-DECAY : 0.96 | |
| 최근 30개의 에피소드 보상 평균 | -26.4 |
| 전체 보상 합 | -17511.9 |
| 학습까지 총 걸린 시간 | 0:04:53 |

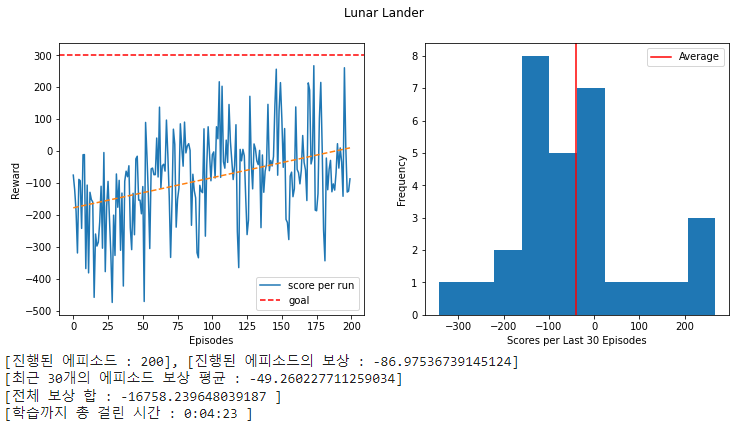


|  |  |
| --- | --- |
| E-DECAY : 0.97 | |
| 최근 30개의 에피소드 보상 평균 | -51.8 |
| 전체 보상 합 | -17845.1 |
| 학습까지 총 걸린 시간 | 0:03:56 |



|  |  |
| --- | --- |
| E-DECAY : 0.98 | |
| 최근 30개의 에피소드 보상 평균 | -68.3 |
| 전체 보상 합 | -16127.3 |
| 학습까지 총 걸린 시간 | 00:04:46 |





|  |  |
| --- | --- |
| E-DECAY : 0.99 | |
| 최근 30개의 에피소드 보상 평균 | -49.2 |
| 전체 보상 합 | -16758.2 |
| 학습까지 총 걸린 시간 | 00:04:23 |

* 엡실론 디케이(E-DECAY) 변경 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | E-DECAY 0.96 | E-DECAY 0.97 | E-DECAY 0.98 | E-DECAY 0.99 |
| 최근 30개의 에피소드 보상 평균 | -26.4 | -51.8 | -68.3 | -49.2 |
| 전체 보상 합 | -17511.9 | -17845.1 | -16127.3 | -16758.2 |
| 학습까지 총 걸린 시간 | 0:04:53 | 0:03:56 | 00:04:46 | 00:04:23 |

엡실론 디케이 변경 결과로 E-DECAY 0.97을 채택하기로 했다.

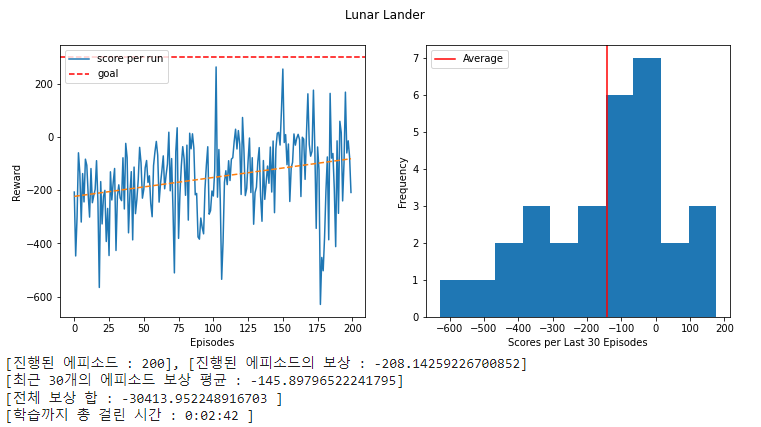
각각의 그래프 개형이 비슷하게 관찰돼 학습까지 총 걸린 시간을 기준을 두었다.

E-DECAY값은 낮추면 낮출수록 Q값이 금방 수렴하기에 크게 낮출 수 없다.

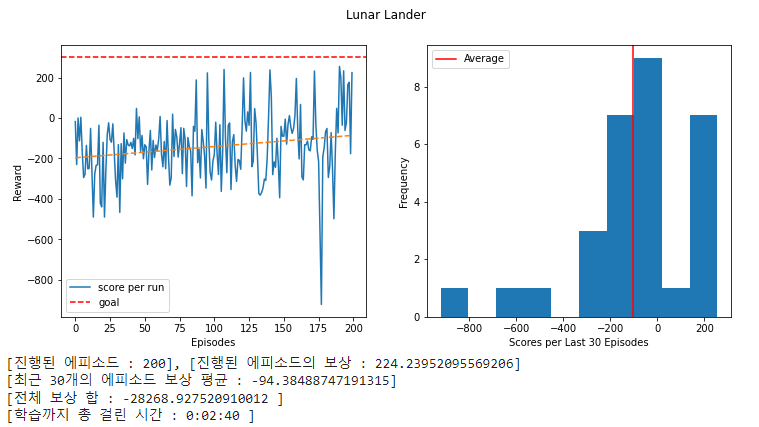
**학습 과정 ( 감마 )**

* 감마(G) 변경

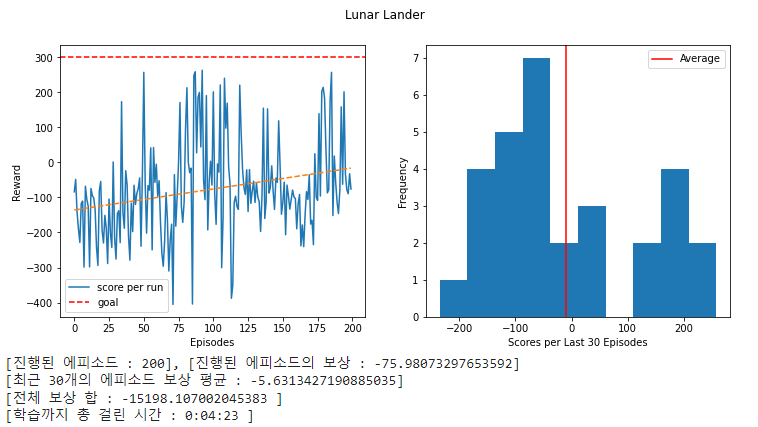
|  |  |
| --- | --- |
| G : 0.8 | |
| 최근 30개의 에피소드 보상 평균 | -145.8 |
| 전체 보상 합 | -30413.9 |
| 학습까지 총 걸린 시간 | 0:02:42 |

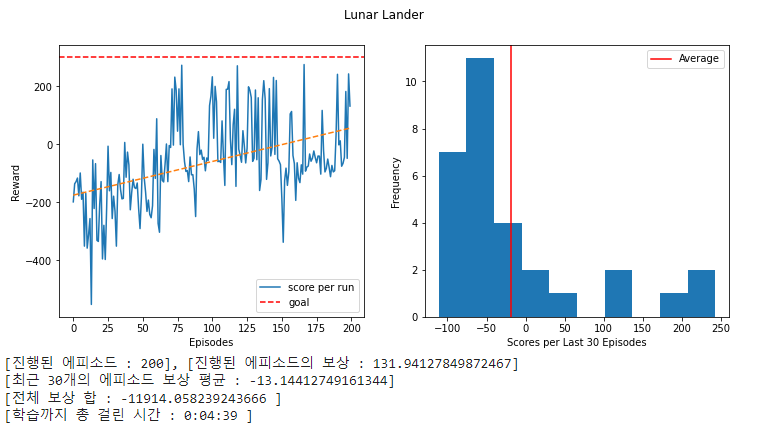


|  |  |
| --- | --- |
| G : 0.85 | |
| 최근 30개의 에피소드 보상 평균 | -94.3 |
| 전체 보상 합 | -28268.9 |
| 학습까지 총 걸린 시간 | 0:02:40 |



|  |  |
| --- | --- |
| G : 0.9 | |
| 최근 30개의 에피소드 보상 평균 | -75.9 |
| 전체 보상 합 | -15198.1 |
| 학습까지 총 걸린 시간 | 00:04:23 |





|  |  |
| --- | --- |
| G : 0.95 | |
| 최근 30개의 에피소드 보상 평균 | -13.1 |
| 전체 보상 합 | -11914.05 |
| 학습까지 총 걸린 시간 | 00:04:39 |

* 감마(G) 변경 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | G 0.8 | G 0.85 | G 0.9 | G 0.95 |
| 최근 30개의 에피소드 보상 평균 | -145.8 | -94.3 | -75.9 | -13.1 |
| 전체 보상 합 | -30413.9 | -28268.9 | -15198.1 | -11914.05 |
| 학습까지 총 걸린 시간 | 0:02:42 | 0:02:40 | 00:04:23 | 00:04:39 |

감마 변경 결과로 G 0.95를 채택하기로 했다.

최소 제곱 그래프(노란색 실선 그래프)의 기울기가 제일 높으며

학습까지 총 걸린 시간이 G 0.9값과 큰 차이를 보이지 않는다.

이 결과를 토대로 감마 값의 특성을 유추할 수 있다.

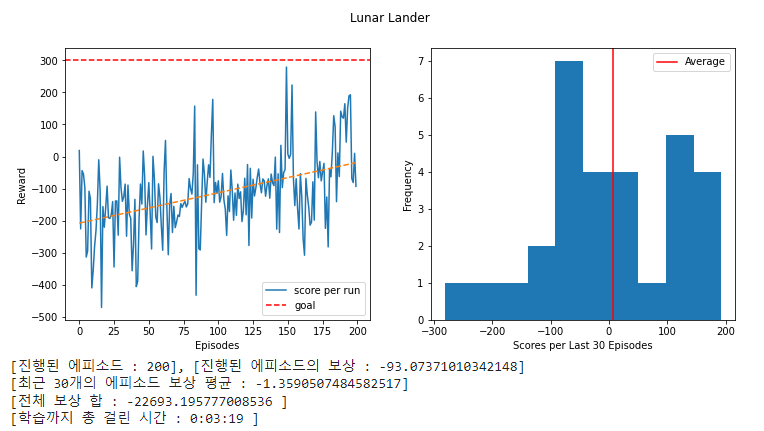
감마 값이 높을수록 최소 제곱 그래프의 기울기가 상승한다는 것을 관찰할 수 있다.

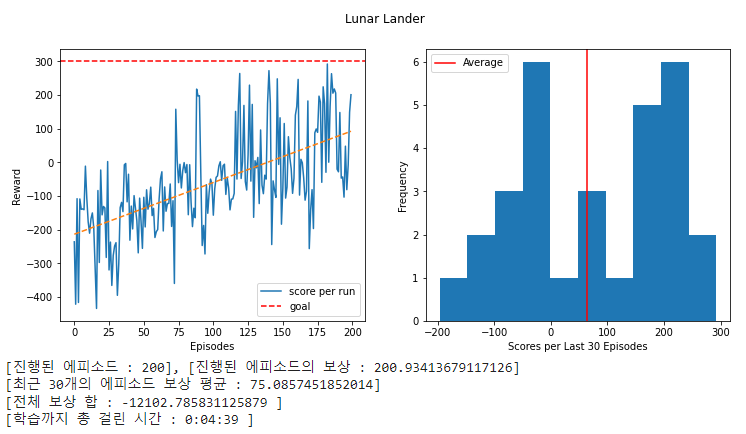
따라서 에피소드의 진척과 보상은 양의 상관관계를 가진다는 결론을 도출해낼 수 있다.

**학습 과정 ( 뉴럴 네트워크 )**

* 뉴럴 네트워크(NN) 변경

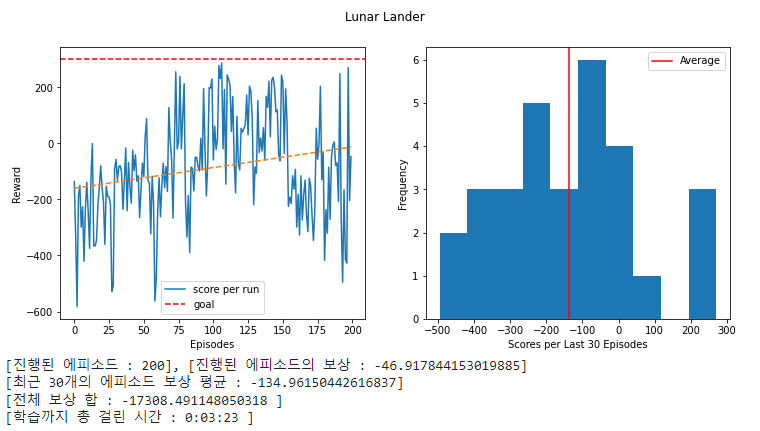
|  |  |
| --- | --- |
| NN : [16, 16] | |
| 최근 30개의 에피소드 보상 평균 | -1.3 |
| 전체 보상 합 | -22693.1 |
| 학습까지 총 걸린 시간 | 0:03:19 |

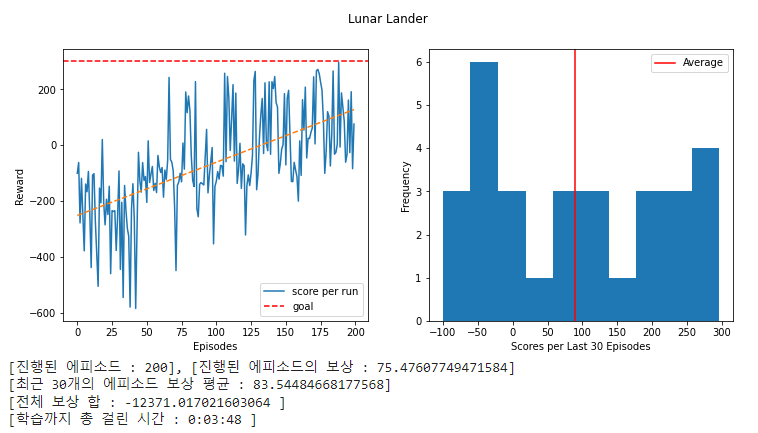




|  |  |
| --- | --- |
| NN : [32, 32] | |
| 최근 30개의 에피소드 보상 평균 | 75.08 |
| 전체 보상 합 | -12102.7 |
| 학습까지 총 걸린 시간 | 0:04:39 |

|  |  |
| --- | --- |
| NN : [64, 64] | |
| 최근 30개의 에피소드 보상 평균 | -134.9 |
| 전체 보상 합 | -17308.4 |
| 학습까지 총 걸린 시간 | 00:03:23 |





|  |  |
| --- | --- |
| NN : [128, 128] | |
| 최근 30개의 에피소드 보상 평균 | 83.5 |
| 전체 보상 합 | -12371.01 |
| 학습까지 총 걸린 시간 | 00:03:48 |

* 뉴럴 네트워크(NN) 변경 결과

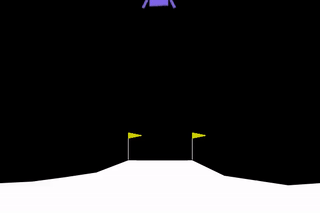
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NN [16, 16] | NN [32, 32] | NN [64, 64] | NN [128, 128] |
| 최근 30개의 에피소드 보상 평균 | -1.3 | 75.08 | -134.9 | 83.5 |
| 전체 보상 합 | -22693.1 | -12102.7 | -17308.4 | -12371.01 |
| 학습까지 총 걸린 시간 | 0:03:19 | 0:04:39 | 00:03:23 | 00:03:48 |

뉴럴 네트워크 변경 결과로 NN [128, 128]을 채택하기로 했다.

4개의 결과중에 최소제곱 그래프의 기울기 크기나

최근 30개의 에피소드 보상 평균이 가장 높게 측정되었고

학습까지 총 걸린 시간 또한 4분 이내로 본 학습의 목적에 있어 유의미한 결과를 도출해낼 수 있다.



이 경우는 첨부한 영상과 마찬가지로 시간제한에 다다를 때까지

우주선이 공중에 떠있는 결과를 관찰할 수 있었다.

따라서 NN [32, 32]는 본 학습 목적에 적합하지 않다고 판단할 수 있다.

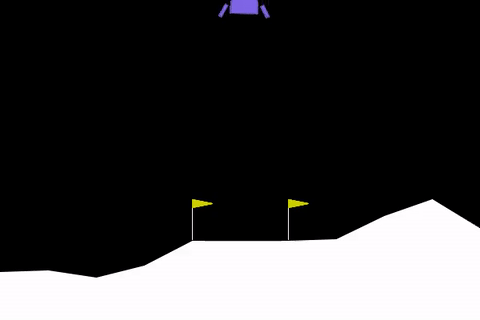
**학습 결과 ( 영상 비교 )**

* 하이퍼 파라미터 최종 세팅

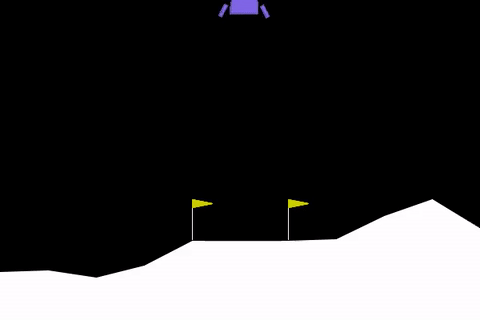
|  |  |
| --- | --- |
| 하이퍼 파라미터 | |
| 러닝 레이트(LR) | 0.005 |
| 엡실론(E) | 0.6 |
| 엡실론 디케이(E-DECAY) | 0.97 |
| 감마(G) | 0.95 |
| 뉴럴 네트워크(NN) | [128, 128] |

* 에피소드(EP) 횟수 변경 별 최종 영상

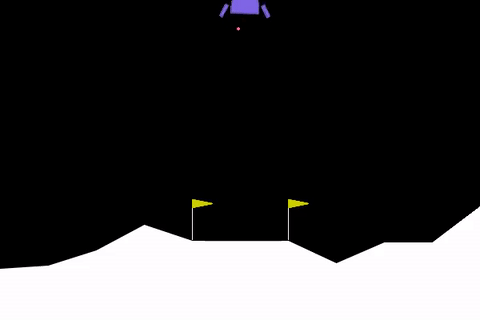
|  |  |
| --- | --- |
| EP : 50 | |
| 총 재생시간 | 1초 |
| 총 얻은 보상 | 0 |
| 성공 여부 | X |



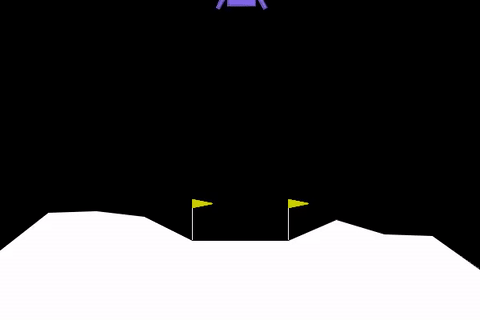
|  |  |
| --- | --- |
| EP : 100 | |
| 총 재생시간 | 1초 |
| 총 얻은 보상 | 0 |
| 성공 여부 | X |



|  |  |
| --- | --- |
| EP : 150 | |
| 총 재생시간 | 4초 |
| 총 얻은 보상 | 218 |
| 성공 여부 | X |



|  |  |
| --- | --- |
| EP : 200 | |
| 총 재생시간 | 6초 |
| 총 얻은 보상 | 239 |
| 성공 여부 | O |



* 에피소드(EP) 횟수 변경 별 최종 영상 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | EP : 50 | EP : 100 | EP : 150 | EP : 200 |
| 총 재생시간 | 1초 | 1초 | 4초 | 6초 |
| 총 얻은 보상 | 0 | 0 | 218 | 239 |
| 성공 여부 | X | X | X | O |

예상된 결과대로 횟수를 늘리면 늘릴수록 성공 확률이 높아진다.

EP 150부터 우주선이 착륙에 성공하지만 깃발 범위에 벗어나는 경우가 많으며

EP 200부터는 우주선이 깃발 안에 안착하는 경우가 많아졌다.

**학습 후기**

* 이번 실험을 통해 깨달은 것

Lunar Lander는 CartPole처럼 무한히 버티는 것이 아닌 문제를 푸는 것을 중점으로 두고 있다.

간단한 문제는 Experience Replay, Double DQN같은 복잡한 기법보단 DQN으로 빠르게 푸는 것도

효율적인 시간관리 측면에서는 더 나은 판단이라고 할 수 있다.

하지만 이게 성공적으로 깃발 사이에 우주선이 안착해 연속적으로 진행하는 식의 문제였다면

DQN은 오래 버티지 못하고 계속 멤돌다 Q값이 수렴해

앞서 말한 기법보다 효율이 떨어졌을지도 모른다.

또한, 최적화 방식이나 손실 계산 방식을 여러가지 방식으로 테스트 했다면

좀 더 좋은 결과를 보여줄 것이라 예상된다.