Reinforcement Learning

나는 강화학습으로 축구한다

Google Research와 Manchester City F.C.의
인공지능 축구 프로젝트를 대한민국에서
재조명하고 직접 구현해보는 캠프

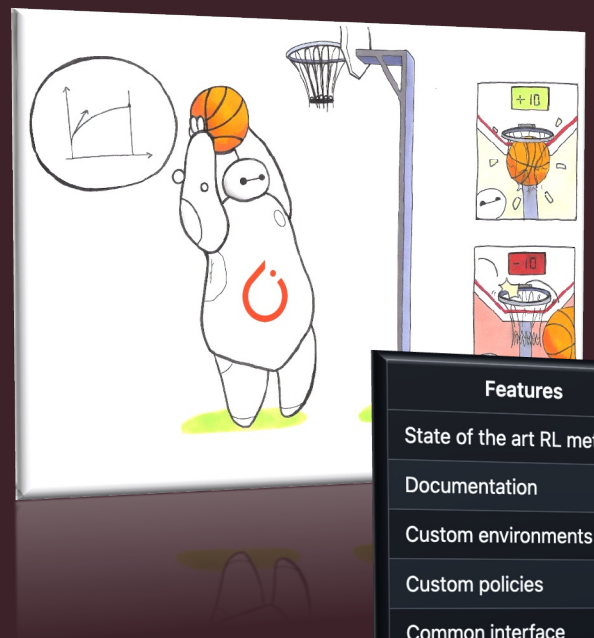
SESSION

11

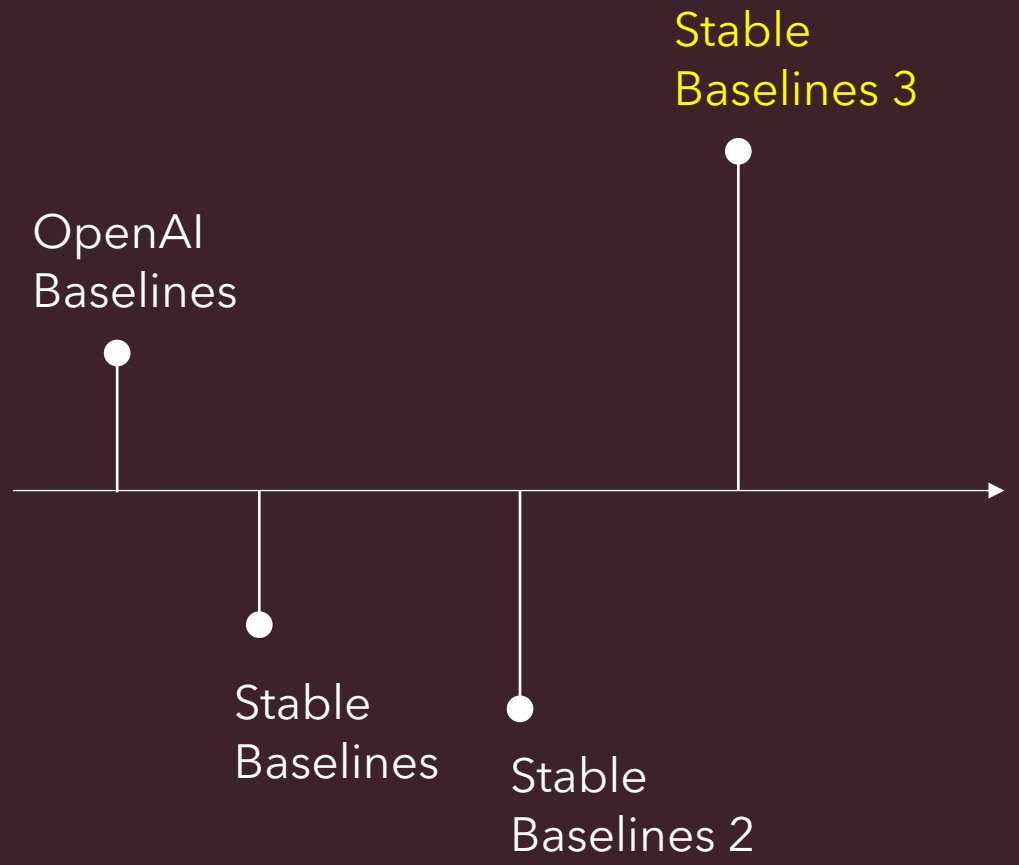
한동대학교 빅데이터사업단
Winning Twelve / 전재영

Copyright Notice

Stable Baselines 3 (SB3)



Features	Stable-Baselines	OpenAI Baselines
State of the art RL methods	✓ (1)	✓
Documentation	✓	✗
Custom environments	✓	✓
Custom policies	✓	– (2)
Common interface	✓	– (3)
Tensorboard support	✓	– (4)
Ipython / Notebook friendly	✓	✗
PEP8 code style	✓	✓ (5)
Custom callback	✓	– (6)



Proximal Policy Optimizatino (PPO)

```python
import gfootball.env as football_env
from stable_baselines3 import PPO

env = football_env.create_environment(
    env_name="academy_empty_goal_close",
    representation="simple115v2"
)

env.reset()

model = PPO("MlpPolicy", env, verbose=1)

model.learn(total_timesteps=10_000)

model.save("model-ppo.zip")

env.close()
```

Maximize $J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}[G_0]$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\theta}(\boldsymbol{a}_t|\boldsymbol{s}_t) G_t\right] = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\theta}(\boldsymbol{a}_t^i|\boldsymbol{s}_t^i) G_t^i$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\theta}(\boldsymbol{a}_t^i|\boldsymbol{s}_t^i) G_t^i$$

$$\boldsymbol{a} \sim \pi_{\theta}(\cdot\,|\boldsymbol{s}_t)$$

DQN, A2C, PPO

```
import gfootball.env as football_env
from stable_baselines3 import DQN

env = football_env.create_environment(...)

env.reset()

model = DQN("MlpPolicy", env, verbose=1)

model.learn(...)

model.save("...")

env.close()
```

```
import gfootball.env as football_env
from stable_baselines3 import A2C

env = football_env.create_environment(...)

env.reset()

model = A2C("MlpPolicy", env, verbose=1)

model.learn(...)

model.save("...")

env.close()
```

```
import gfootball.env as football_env
from stable_baselines3 import PPO

env = football_env.create_environment(...)

env.reset()

model = PPO("MlpPolicy", env, verbose=1)

model.learn(...)

model.save("...")

env.close()
```

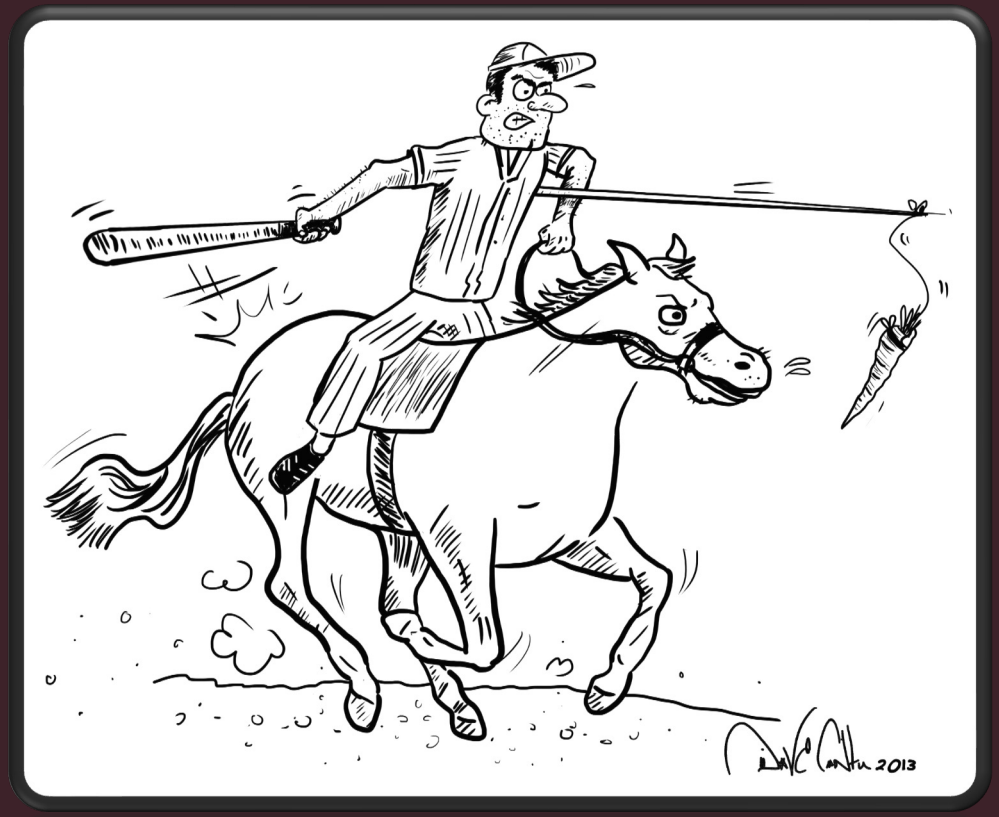
Hands-On Exercise

PPO로 학습시켜보기

LAB 06

Live

당근과 채찍질



Reward Design

Decides what the agent ultimately optimizes

+1

득점

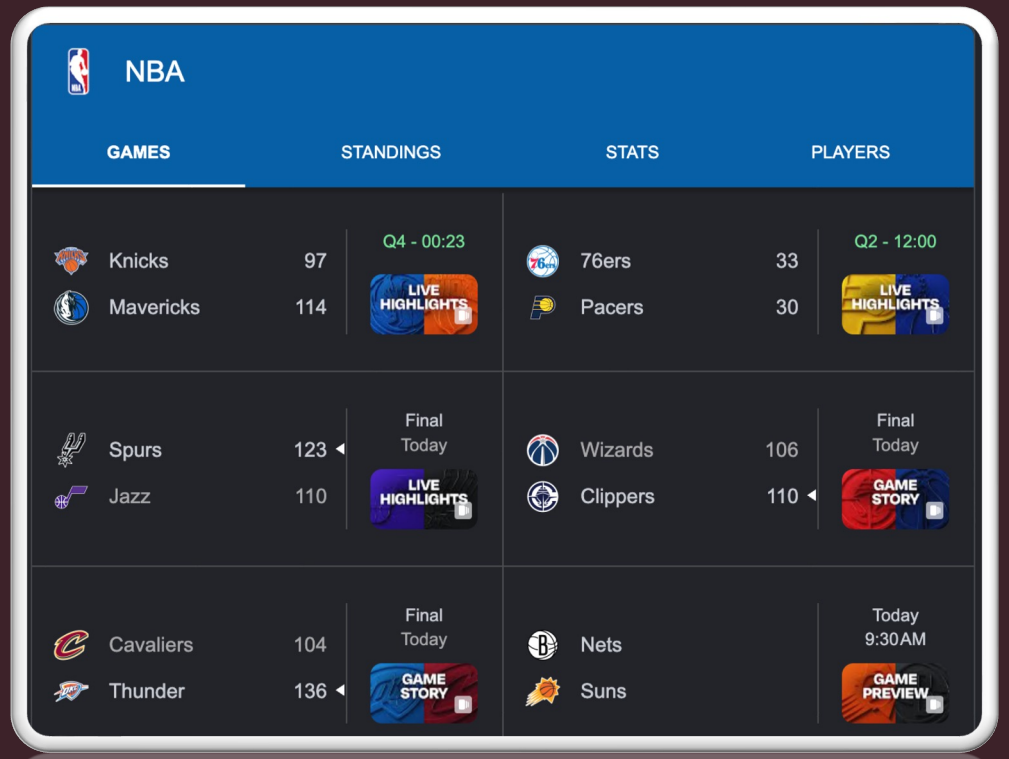
-1

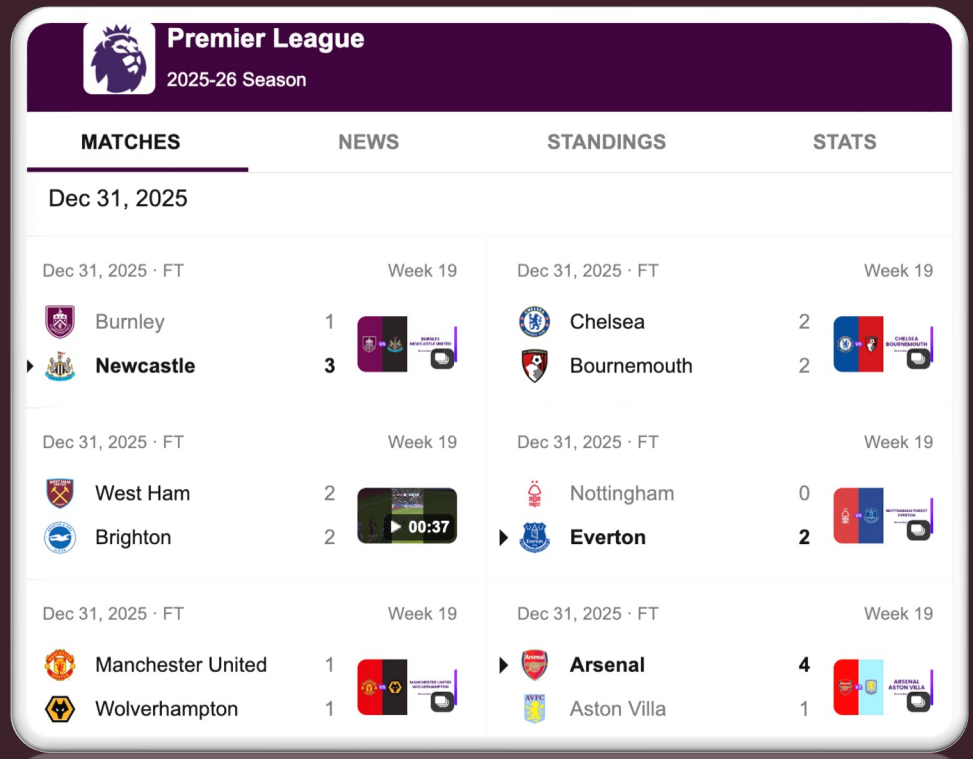
실점

Environment 만들 때 정한
궁극적으로 Agent가 성취하려는 목적

Sparse Rewards

골이 너무 적게 나온다...





Reward Design vs. Reward Shaping

- Scoring: +1, -1
  - Learning becomes very slow
  - 당근도 안 주고 채찍질도 안 하는...
- Reward Shaping
  - Instead of giving the agent a reward only at the end (e.g. scoring a goal), you provide additional intermediate rewards that encourage progress toward the goal.
  - Agent receives feedback earlier.
  - Less exploration burden.

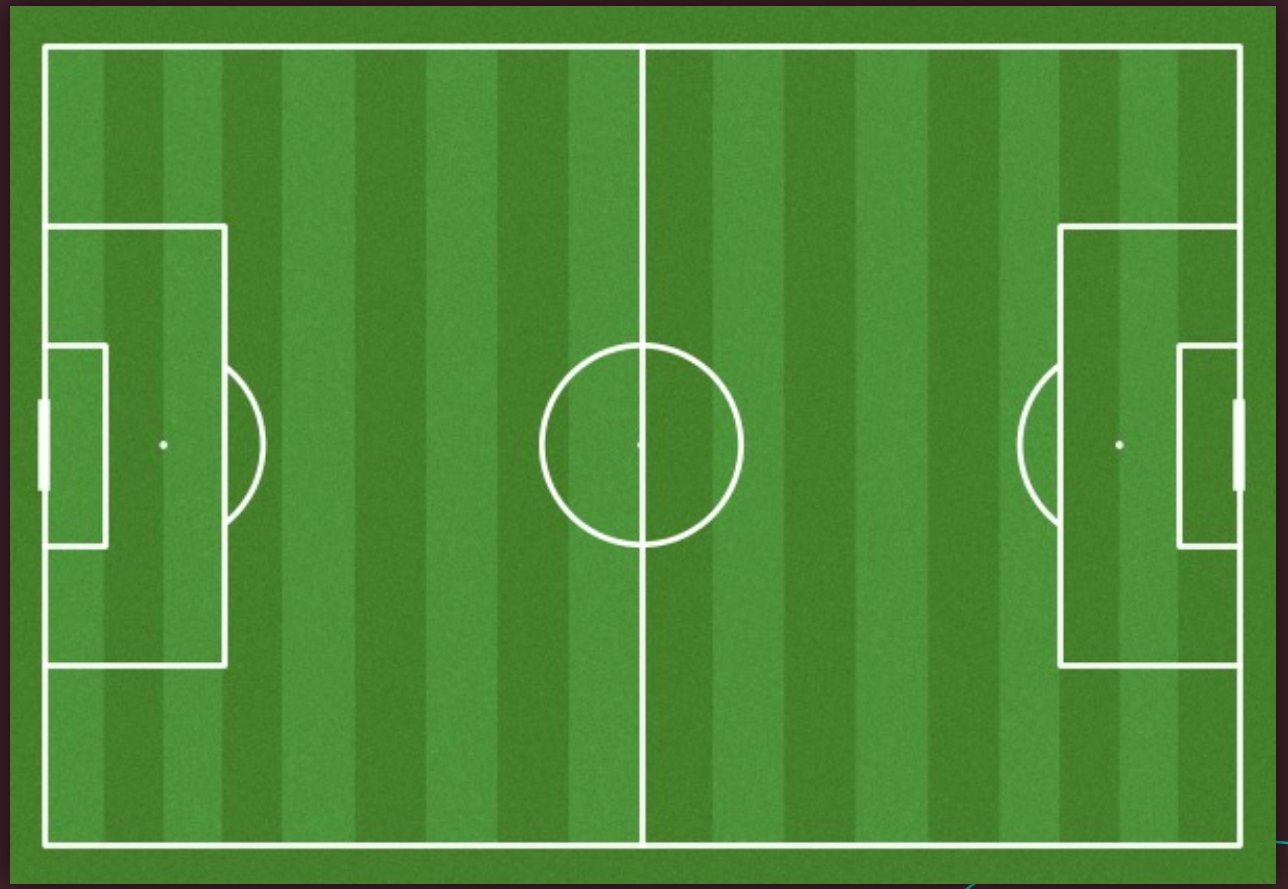
힌트주기? 빵가루 떨어뜨려놓기?

Examples of Reward Shaping

- Acquiring Possession
- Moving Ball Closer to Opponent Goal
- Losing Possession
- Staying Longer in the Opponent's Side
- …

Scoring

```
env = football_env.create_environment(
    env_name="academy_run_to_score",
    rewards="scoring"
)
```



-1

실점

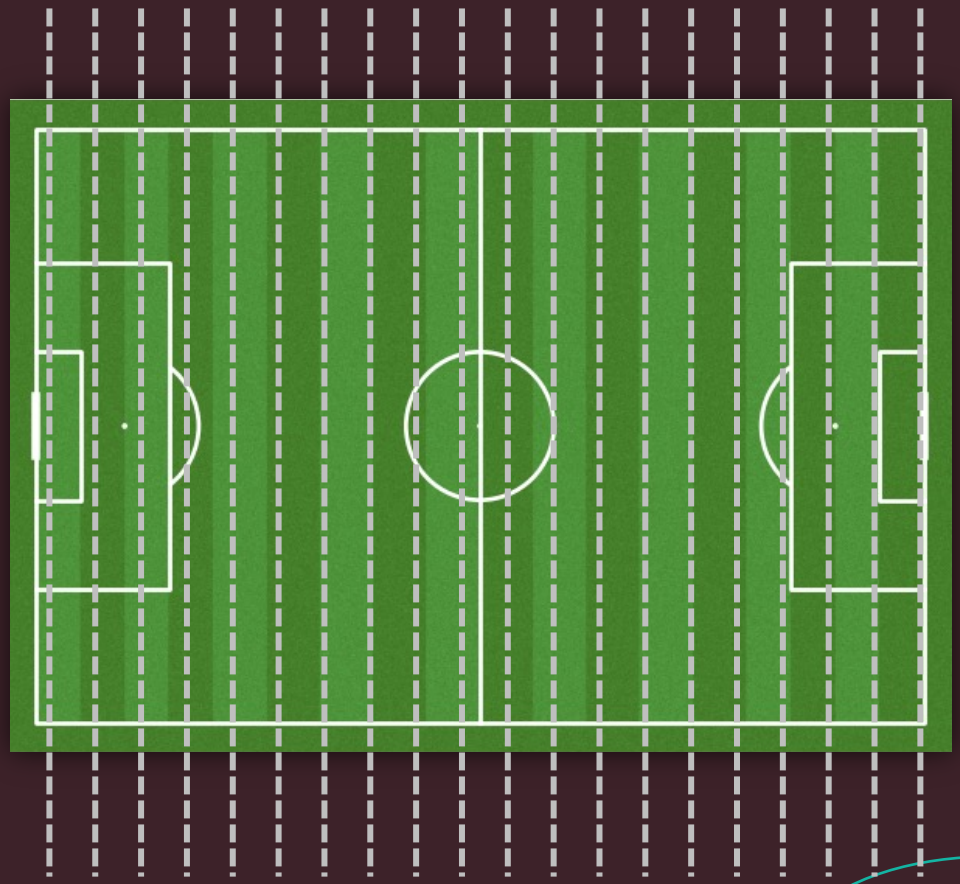
+1

득점

Checkpoints

(GRF Built-in Reward Shaping)

```
env = football_env.create_environment(
    env_name="academy_run_to_score",
    rewards="scoring,checkpoints"
)
```



-0.1

+0.1

(the lines are arbitrarily drawn for the illustrative purpose)

Other Issues

- Reward Hacking
  - Cheating
- Overfitting
  - Fails to generalize
- Conflicting Incentives
  - Dribble vs. Shot

Applying SB3's PPO

academy_run_to_score_with_keeper