환경 = [현재 State, 다음 State, Action, Action 확률, reward ···]

↳ 기본값.

Optimizer = Adam (요즘 경사도가 가장 빠름)

Activation Function = ReLU

Gamma = 0.009

buffer_Size = 5000

learning_rate = 0.0001.

batch_size = 64.

State → Agent → Action

↑ reward

next State ← Env ←

buffer를 사용하여 S_t, S_{t+1}, Action, reward, Episode 등을 메모리 저장.

↳ Sample 추출