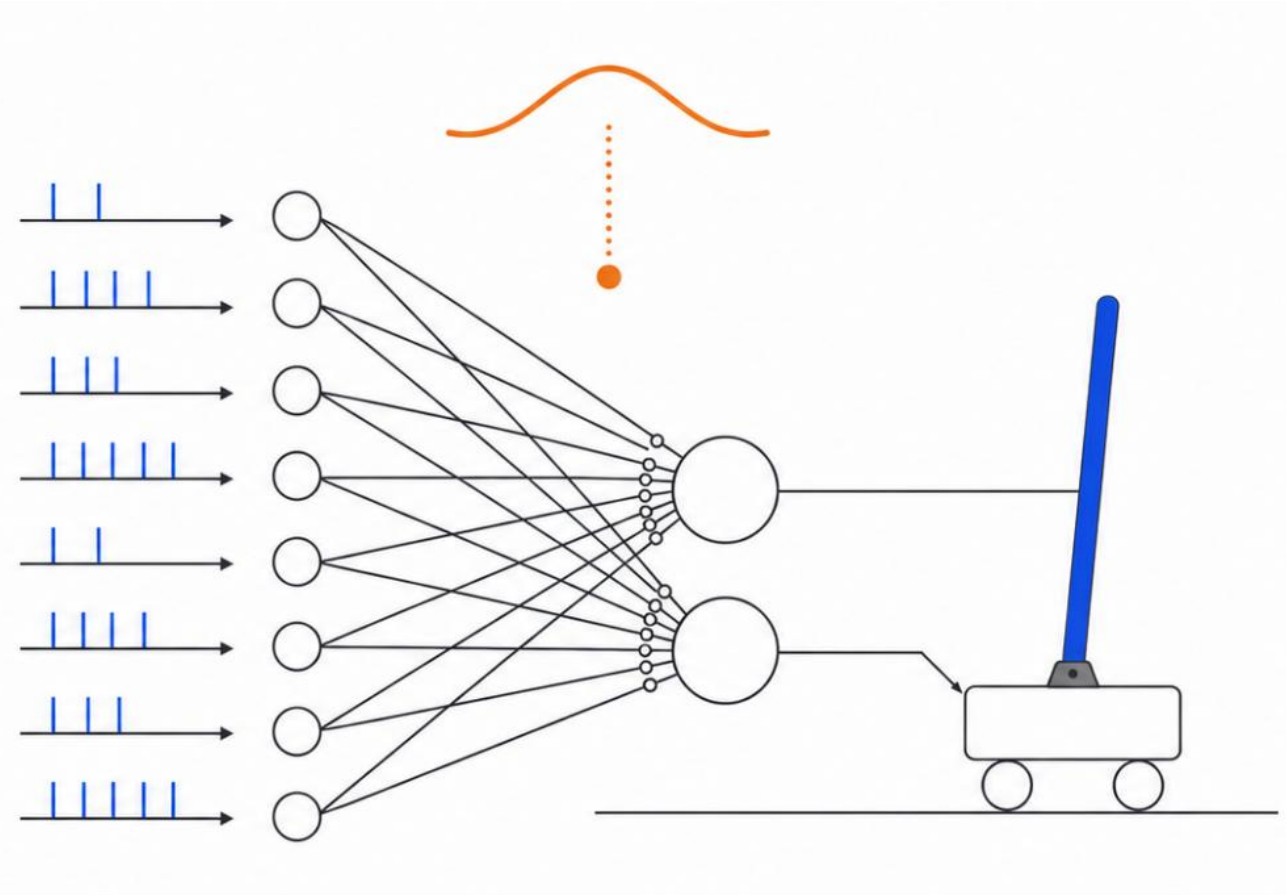
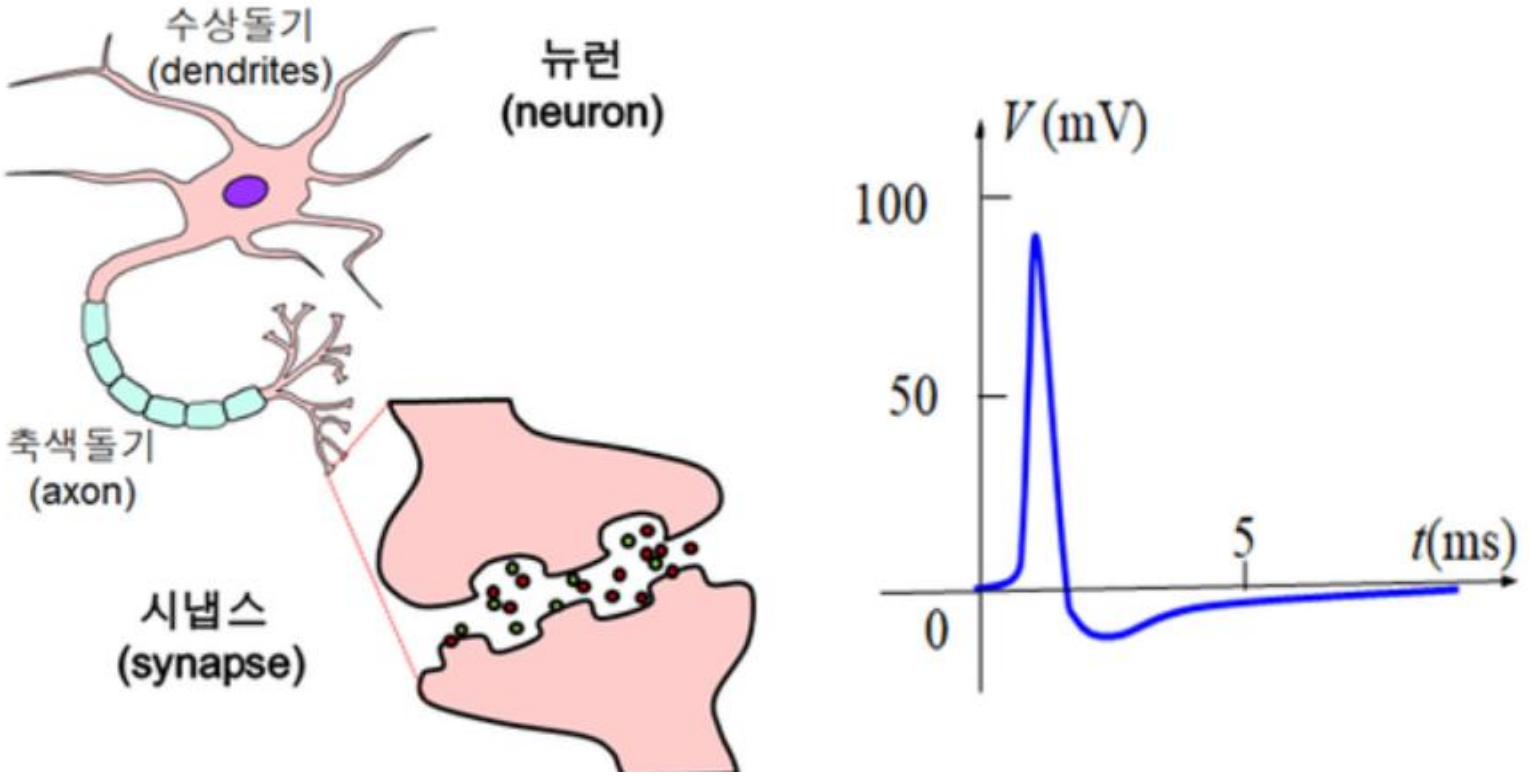
도파민 유사 학습 신호 기반 STDP-SNN 강화학습 방법 분석



2026-07-21

SNN, Spiking Neural Network란?

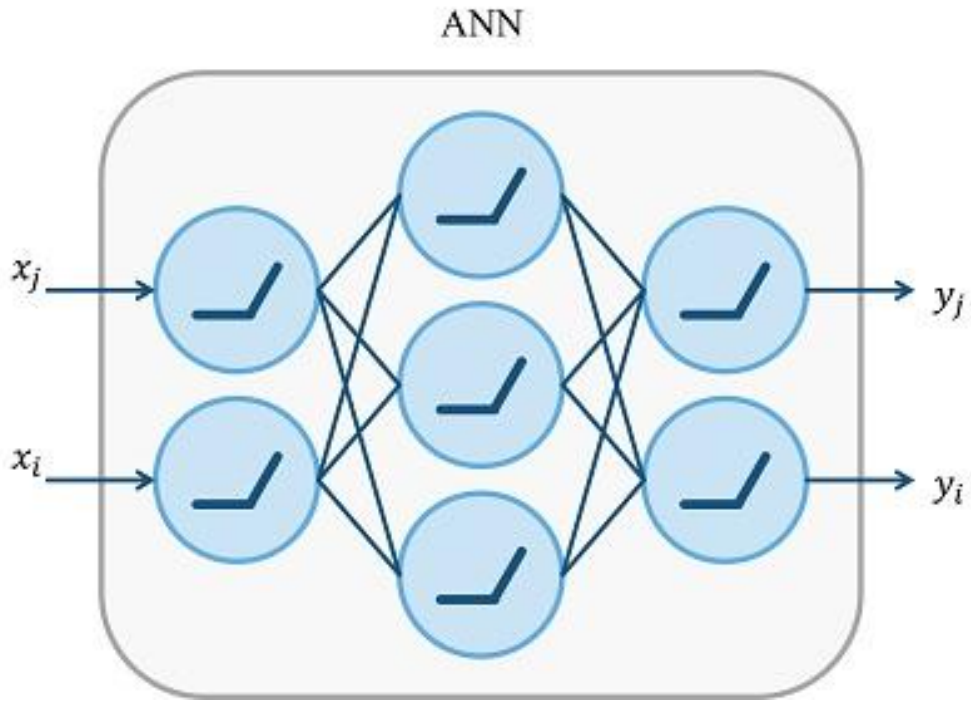
SNN은 **생물의 신경망을 모방한 모델**로 뉴런이 연속적인 실수값을 출력하는 것이 아니라,
특정 순간에 발생하는 **스파이크(spike)**를 통해 정보를 전달하는 신경망



생체 뉴런의 입출력

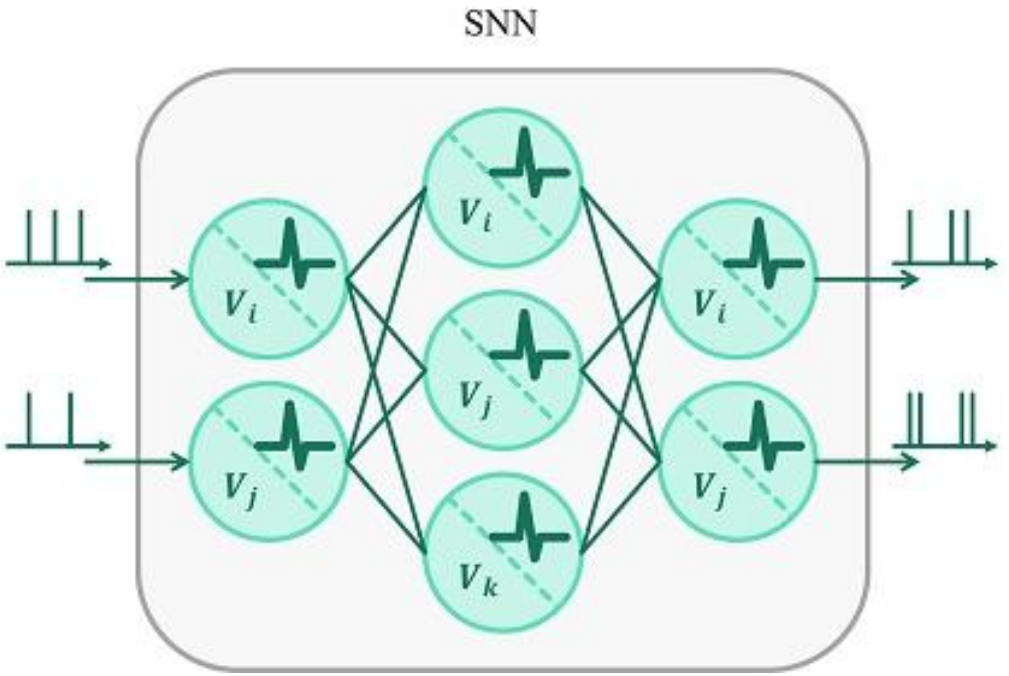
SNN, Spiking Neural Network란?

SNN은 **생물의 신경망을 모방한 모델**로 뉴런이 연속적인 실수값을 출력하는 것이 아니라,
특정 순간에 발생하는 **스파이크(spike)**를 통해 정보를 전달하는 신경망



ANN 뉴런의 입출력

[0.13, 0.72, -0.21, 1.35 ...]

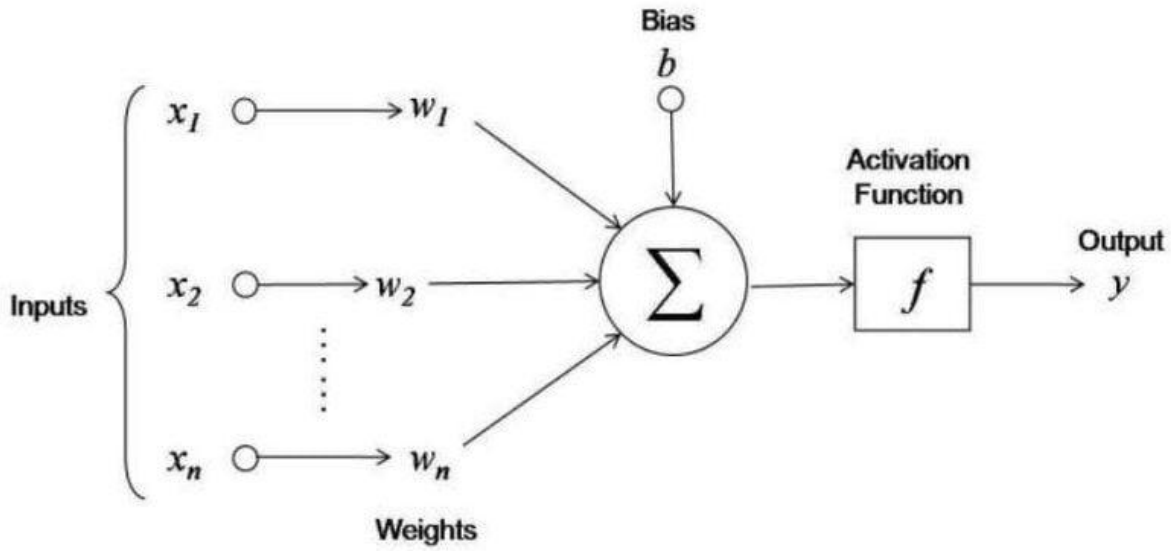


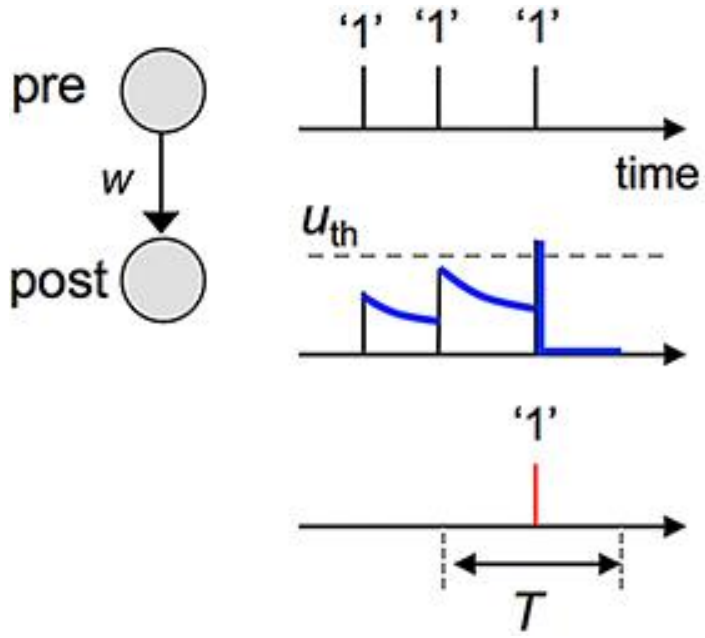
SNN 뉴런의 입출력

[0, 0, 1, 0, 0, 1, 0 ...]

SNN, Spiking Neural Network란?

SNN은 **생물의 신경망을 모방한 모델**로 뉴런이 연속적인 실수값을 출력하는 것이 아니라,
특정 순간에 발생하는 **스파이크(spike)**를 통해 정보를 전달하는 신경망





ANN 뉴런의 입출력

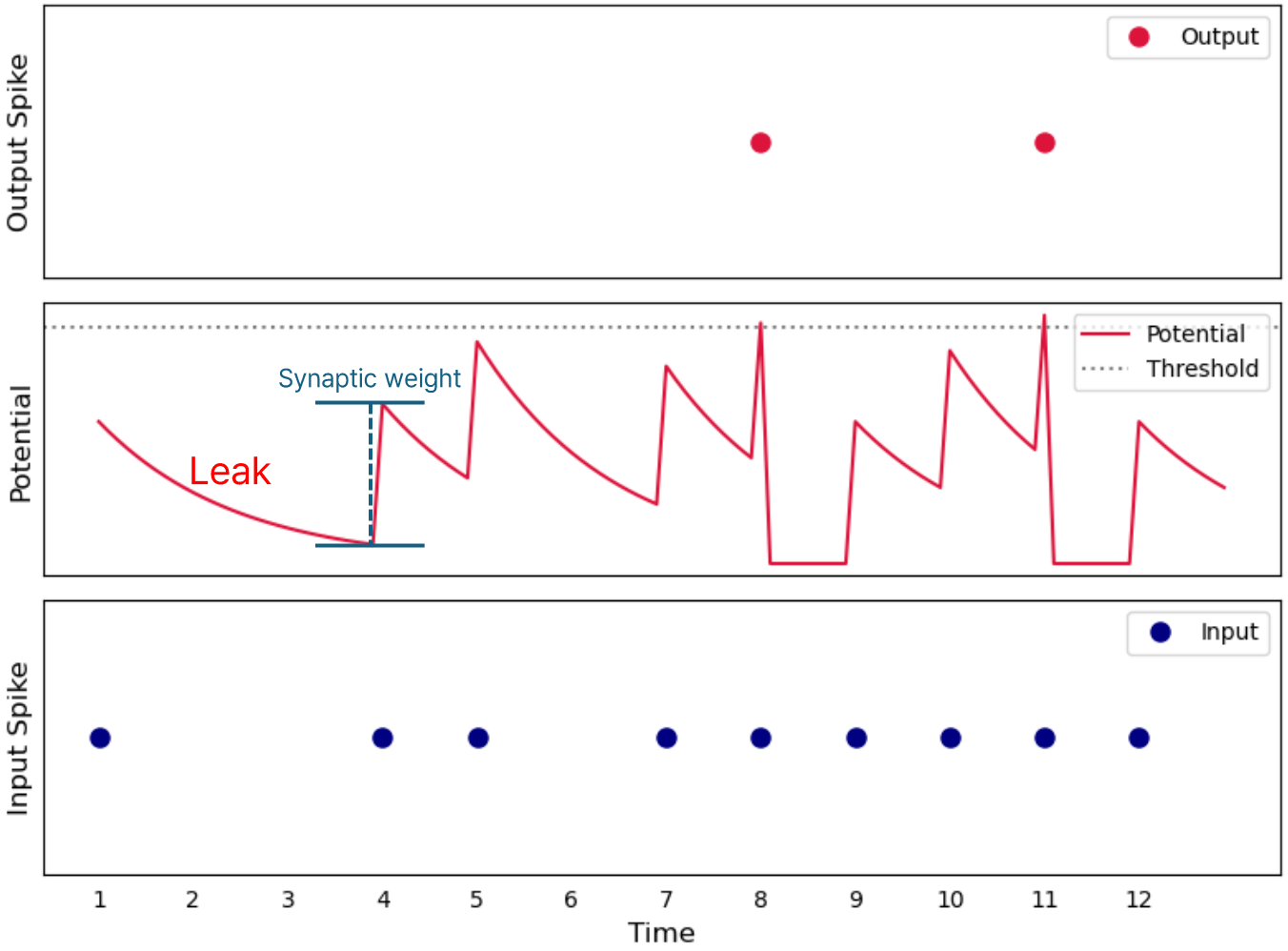
[0.13, 0.72, -0.21, 1.35 ...]

SNN 뉴런의 입출력

[0, 0, 1, 0, 0, 1, 0 ...]

SNN 뉴런의 동작 과정

뉴런의 **막전위(potential)**가 **임계값(Threshold)**을 넘어야 **발화(Firing)**하여 다음 뉴런으로 전달
(SNN에서 가장 기본적으로 사용되는 뉴런 모델 **LIF, Leaky Integrate-and-Fire 모델**)



발화 (Firing)
뉴런이 스파이크를 발생시키는 과정

막전위 (potential)
뉴런 내부에 누적된 전기적 상태값

누설(Leak)
시간이 지나면서 막전위가 자연스럽게 감소

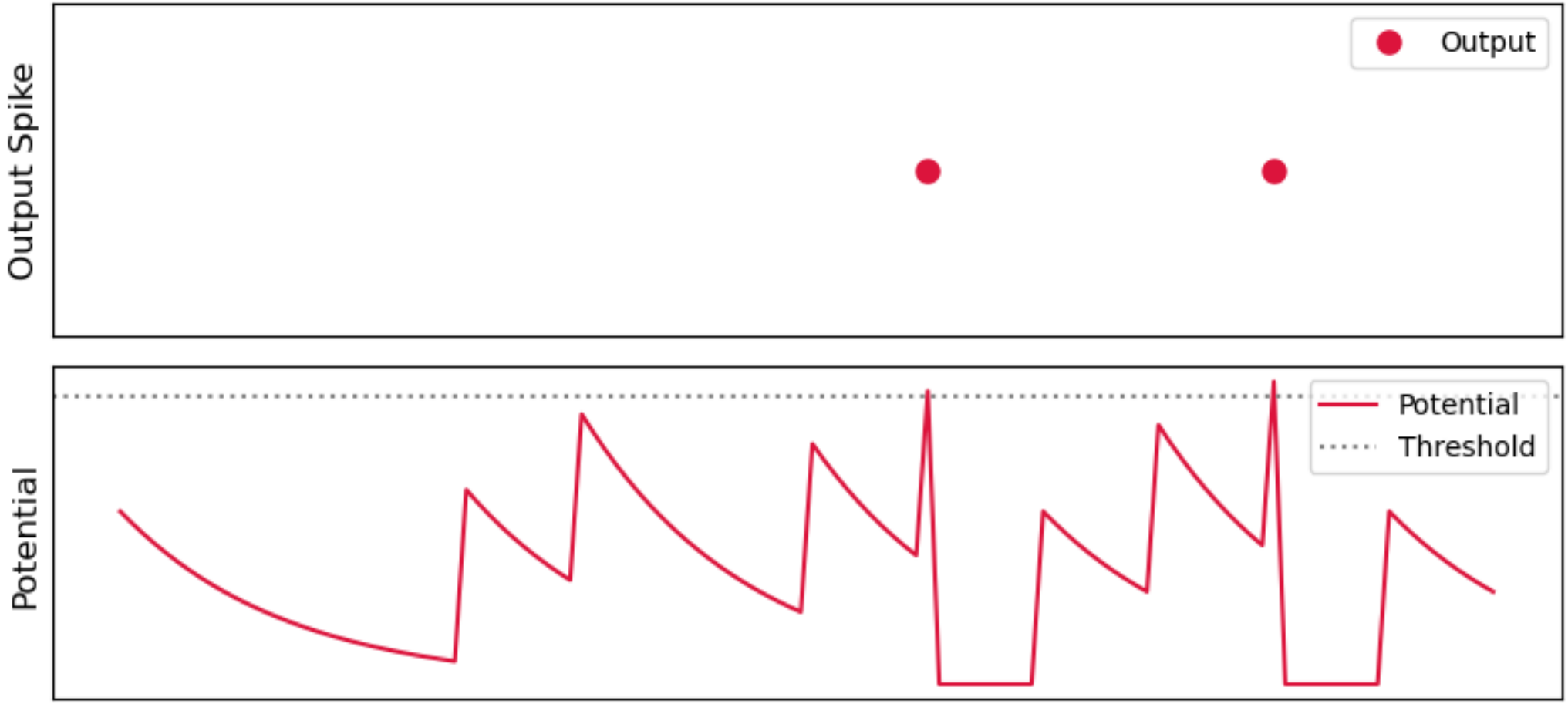
시냅스 입력(Synaptic input)
이전 스파이크가 막전위에 미치는 영향

초기화(Reset)
발화 후 막전위가 초기화 됨 (이후 잠시동안 불활성화)

시냅스 가중치(Synaptic weight)
시냅스 입력의 정도

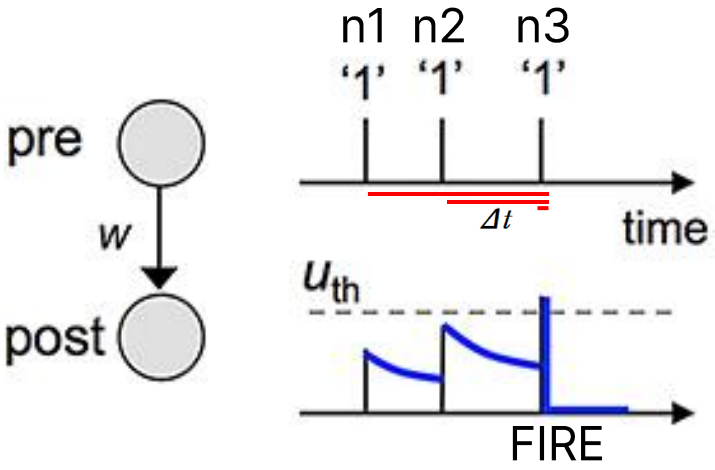
학습이 어려운 SNN 모델

SNN의 **출력이 연속적인 값이 아니라,** 임계값을 넘는 순간 발생하는 **불연속적인 스파이크**이다.
각 뉴런은 **시냅스 입력**과 **발화 시 급격하게 막전위가 변화**한다.
(때문에 미분이 불가능하여 일반적인 **역전파 (Backpropagation)를 적용하기 어렵다.**)



역전파 방식을 대체한 SNN 학습 방법

STDP(Spike-Timing-Dependent Plasticity)는 역전파처럼 전체 오차를 미분해서 전달하는 방식이 아니라,
pre-neuron과 post-neuron의 스파이크 발생 시간 차이를 이용해 시냅스 가중치를 조정하는 SNN 학습 방법



$$\Delta t = t_{post} - t_{pre}$$

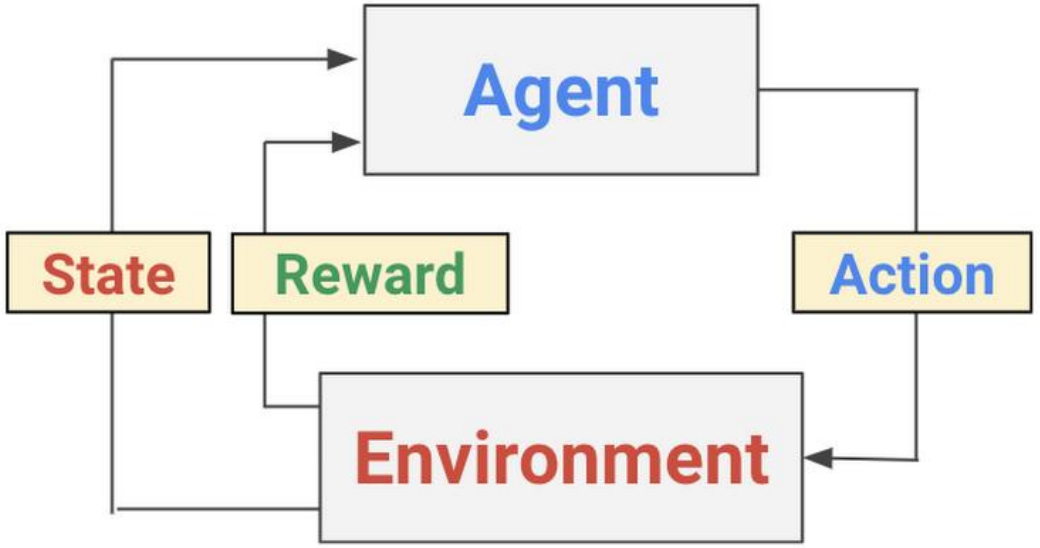
$$\Delta w = \begin{cases} A_+ e^{-\frac{\Delta t}{\tau_+}}, & \Delta t > 0 \quad (LTP) \\ -A_- e^{\frac{\Delta t}{\tau_-}}, & \Delta t < 0 \quad (LTD) \end{cases}$$

post 발화 직전에 가까운 pre spike일수록 **(n3 > n2 > n1)**
post 발화에 더 크게 기여한 것으로 판단

A_+,A_-: 가중치 변화 크기
τ_+,τ_-: 시간 차이에 따른 감쇠 상수

STDP를 강화학습으로 확장

강화학습의 상태/보상 개념과 **STDP** 방식을 함께 사용하여
R-STDP(Reward-modulated Spike-Timing-Dependent Plasticity) 구현



강화학습 프로세스

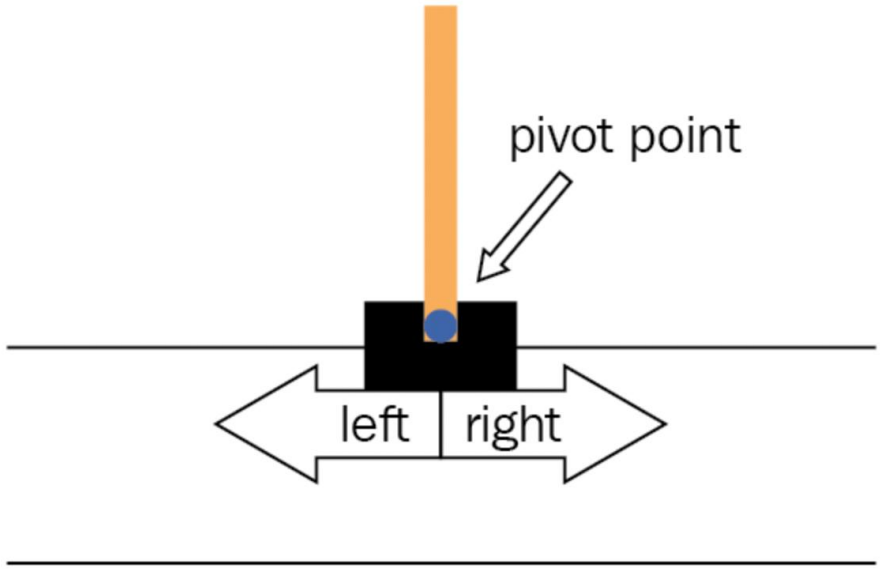
$$\Delta t = t_{post} - t_{pre}$$

$$\Delta w = \begin{cases} A_+ \cdot e^{-\frac{\Delta t}{\tau_+}} & , \Delta t > 0 \quad (LTP) \\ -A_- \cdot e^{\frac{\Delta t}{\tau_-}} & , \Delta t < 0 \quad (LTD) \end{cases}$$

STDP 메커니즘

실험 환경: CartPole-v1

SNN actor는 네 개의 연속적인 상태값을 입력받아 **왼쪽 또는 오른쪽 행동을 선택**



실험 환경: CartPole-v1

$$s_t = [x_t, \dot{x}_t, \theta_t, \dot{\theta}_t] \qquad a_t \in \{0,1\}$$

상태(s_t)	의미
x_t	수레의 수평 위치
$\dot{x}_t$	수레의 수평 속도
θ_t	막대가 수직축에서 기울어진 각도
$\dot{\theta}_t$	막대의 각속도

행동(a_t)	의미
0	수레를 왼쪽으로 이동
1	수레를 오른쪽으로 이동

상태 및 행동 정의

실험 환경: CartPole-v1

막대의 균형을 유지하는 **매 step마다 +1의 보상**을 받으며 **종료 조건**에서 에피소드 종료

$$r_t = 1, \quad R = \sum_{t=1}^{T} r_t = T$$

$$|x_t| > 2.4 \quad \text{or} \quad |\theta_t| > 0.2095 \quad \text{or} \quad T \geq 500$$

기호	의미
t	현재 환경 step
r_t	시간 t에서 환경으로부터 받은 보상
T	Episode가 지속된 전체 step 수
R	한 episode에서 누적된 전체 보상

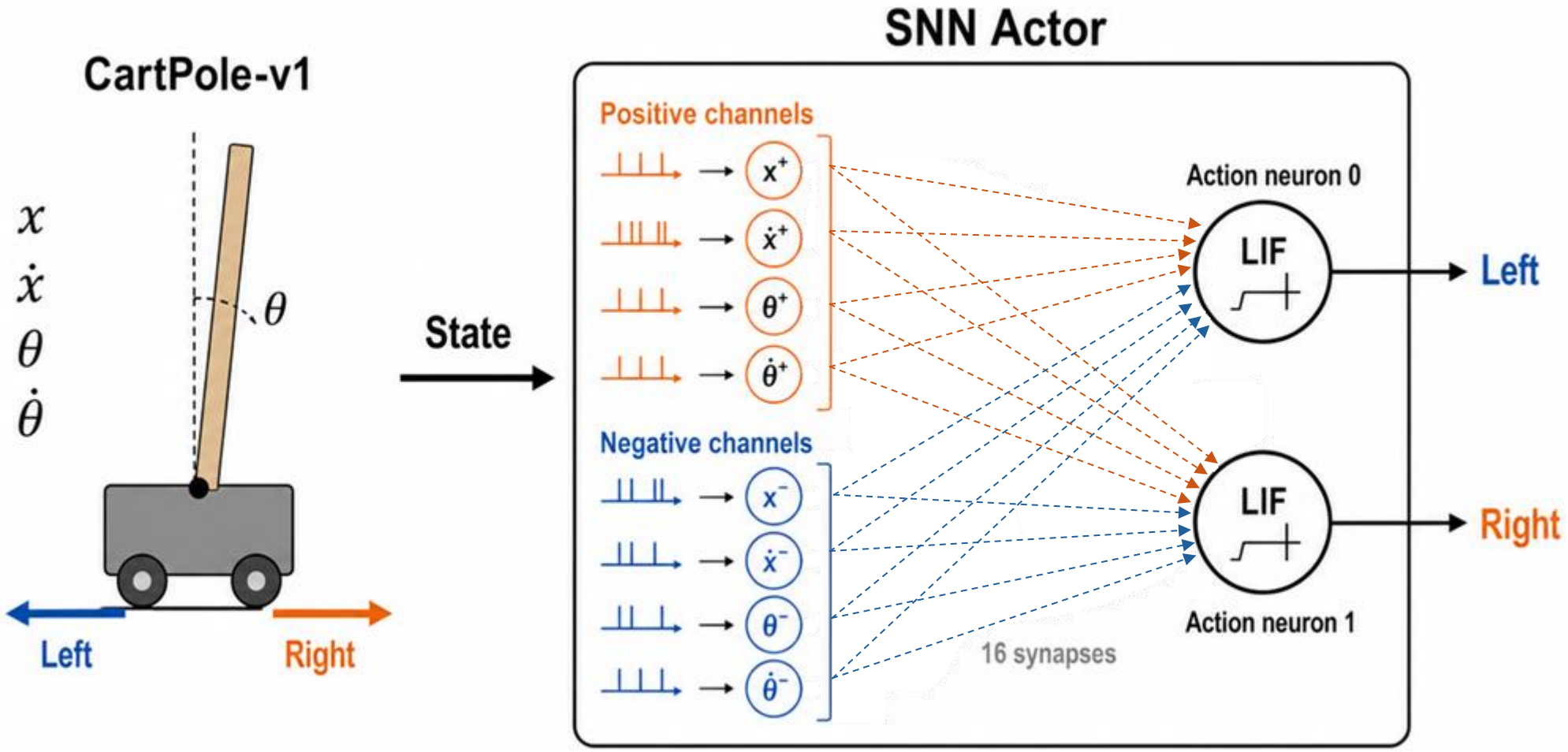
보상구조

기호	의미
$\|x_t\| > 2.4$	수레의 위치가 허용 범위를 벗어남 (위치 최대 구역 설정)
$\|\theta_t\| > 0.2095$	막대의 각도가 약 12°를 초과 (허용되는 최대 막대 각도 약 12°를 radian으로 표현)
$T \geq 500$	최대 500 step에 도달

종료 조건

SNN Actor 뉴런 구조

CartPole의 **4개 상태값**을 **양·음 방향**에 따라 **8개의 입력** spike channel로 변환하고,
16개 시냅스로 연결된 LIF 행동 뉴런의 발화 결과를 통해 **좌·우 행동을 결정**



SNN Actor의 인코딩

상태값 → 정규화 → 채널 분리 → spike 발생 확률 계산 → 내부 step에서 spike train 생성

$$s_t = [x_t, \dot{x}_t, \theta_t, \dot{\theta}_t]$$

정규화

$\widehat{s_{t,i}} \in [-1,1]$

각 상태값을 해당 기준으로 정규화

$$c = [2.4, 3.0, 0.2095, 3.5]$$

정규화 예시

$$\widehat{s_t} = \left[\frac{0.6}{2.4}, \frac{-0.9}{3.0}, \frac{0.10475}{0.2095}, \frac{-0.7}{3.5}\right] = [0.25, -0.30, 0.50, -0.20]$$

채널 분리

$$u_t = [r_t^+, r_t^-]$$

채널 분리 예시

$$u_t = [0.25,0,0.50,0,0,0.30,0,0.20]$$

spike 발생 확률 계산

$$p_t = p_{\max} u_t$$

(R-STDP)

spike 발생 확률 계산 예시

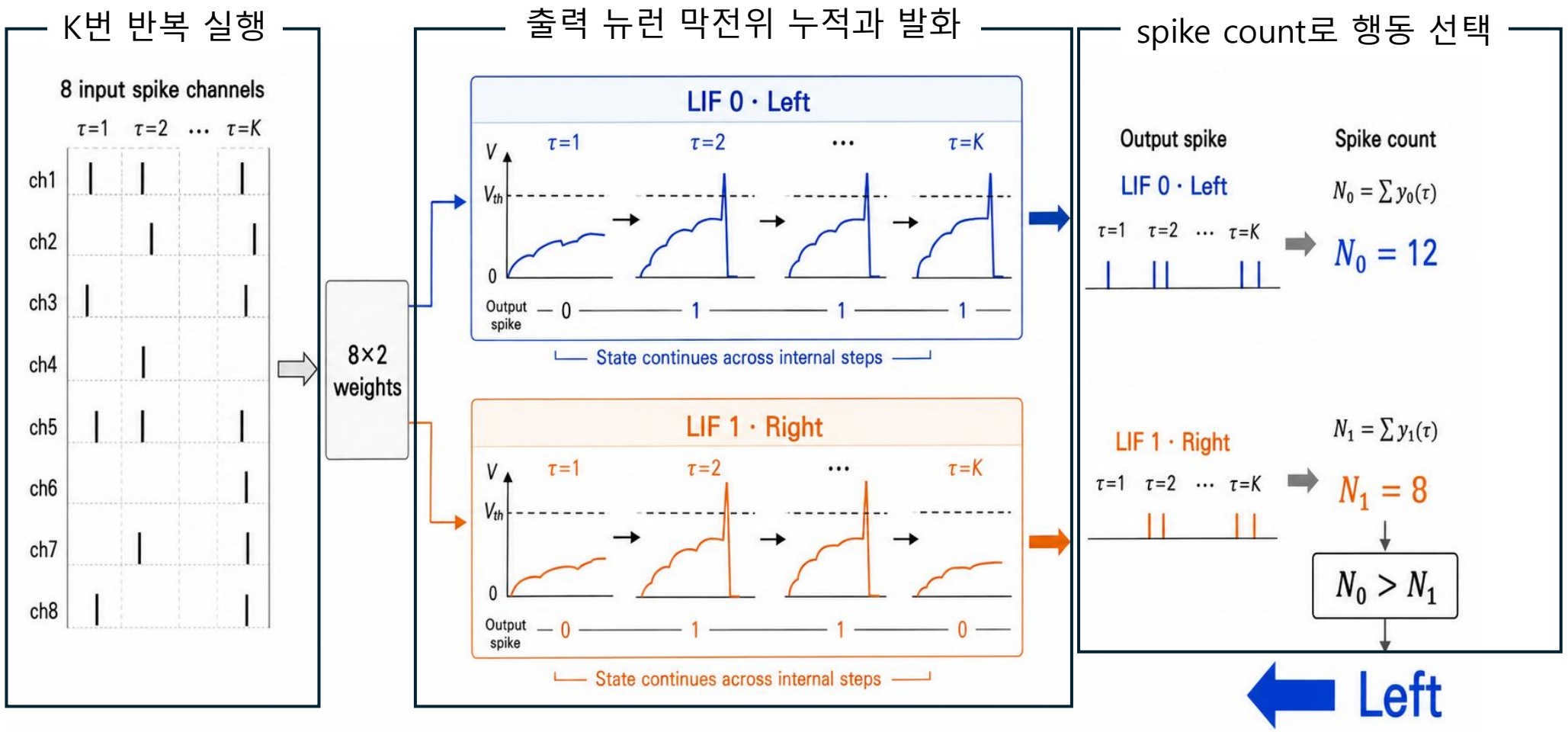
$$p_t = [0.15,0,0.30,0,0,0.18,0,0.12]$$

상태(s_t)	의미
x_t	수레의 수평 위치
$\dot{x}_t$	수레의 수평 속도
θ_t	막대가 수직축에서 기울어진 각도
$\dot{\theta}_t$	막대의 각속도

알고리즘	$p_{\max}$
R-STDP	0.60
Episodic R-STDP	0.60
TD-STDP Actor-Critic	0.50
PPO-modulated TD-STDP	0.50

SNN Actor의 디코딩

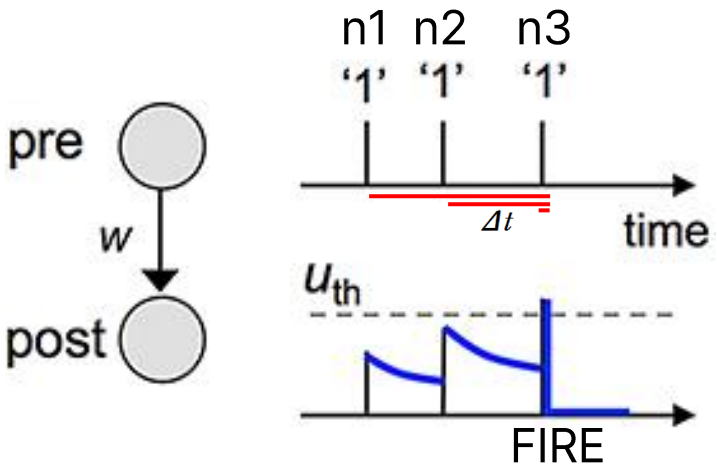
CartPole 환경은 **시간 t에서 하나의 상태값을 전달**하지만 SNN은 이 상태를 **한 번만 처리하지 않고** 하나의 **환경 step t 안에서** 동일한 입력 확률을 사용해 SNN을 **K 번 반복 실행 후 행동 결정**



R-STDP: 보상 조절 STDP

입력·출력 spike의 시간적 상관관계로 STDP eligibility를 형성하고,
행동 결과로 계산한 도파민 유사 신호를 결합하여 선택된 행동 **뉴런의 시냅스를 매 환경 step마다 갱신**

상태 입력→SNN 실행→행동 선택→**eligibility 형성**→결과 확인→도파민 신호 계산→가중치 갱신



post 발화 직전에 가까운 pre spike일수록 **(n3 > n2 > n1)**
post 발화에 더 크게 기여한 것으로 판단

$$e_{ij,t} = \frac{1}{K}\sum_{\tau=1}^{K}\left(A_{+}\mathrm{LTP}_{ij}^{(\tau)} - A_{-}\mathrm{LTD}_{ij}^{(\tau)}\right)$$

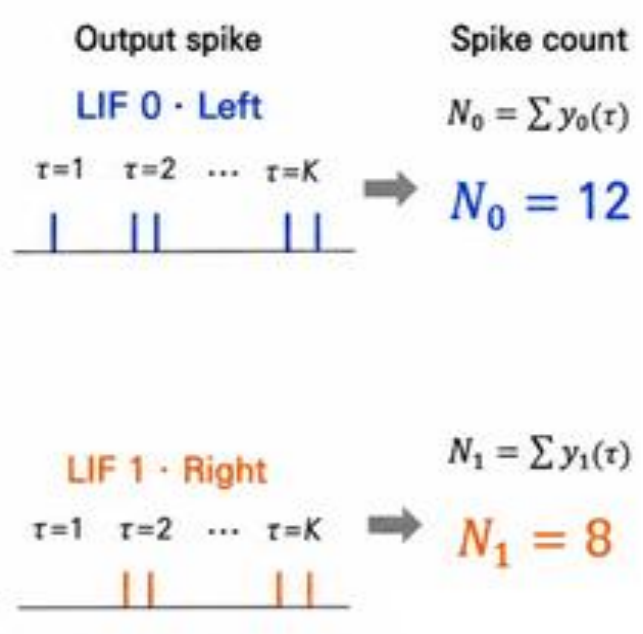
기호	의미
K	spike 발생 관찰 수
i	스파이크 채널 종류
j	출력(행동) 종류

STDP eligibility 계산

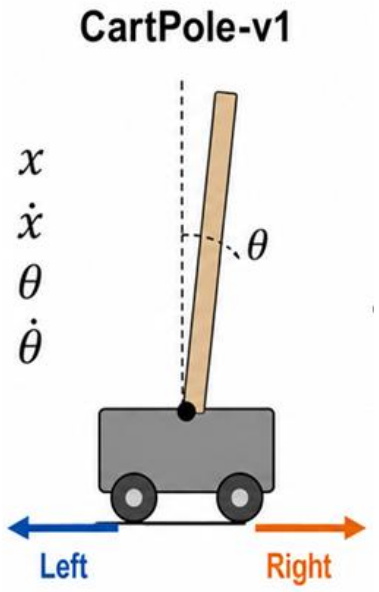
R-STDP: 보상 조절 STDP

입력·출력 spike의 시간적 상관관계로 STDP eligibility를 형성하고,
행동 결과로 계산한 도파민 유사 신호를 결합하여 선택된 행동 뉴런의 시냅스를 매 환경 step마다 갱신

상태 입력→SNN 실행→행동 선택→eligibility 형성→**행동 결과 확인**→도파민 신호 계산→가중치 갱신



행동 선택



행동 적용 및 결과 확인

$$J(s) = \omega_x \frac{|x|}{c_x} + \omega_{\dot{x}} \frac{|\dot{x}|}{c_{\dot{x}}} + \omega_\theta \frac{|\theta|}{c_\theta} + \omega_{\dot{\theta}} \frac{|\dot{\theta}|}{c_{\dot{\theta}}}$$

기호	의미
$J(s)$	불안정도
ω	상태 수치의 중요도
c	상태값 정규화 기준

상태의 불안정도 계산

R-STDP: 보상 조절 STDP

입력·출력 spike의 시간적 상관관계로 STDP eligibility를 형성하고,
행동 결과로 계산한 도파민 유사 신호를 결합하여 선택된 행동 뉴런의 시냅스를 매 환경 step마다 갱신

상태 입력→SNN 실행→행동 선택→eligibility 형성→<u>행동 결과 확인</u>→**도파민 신호 계산→가중치 갱신**

$$J(s) = \omega_x \frac{|x|}{c_x} + \omega_{\dot{x}} \frac{|\dot{x}|}{c_{\dot{x}}} + \omega_\theta \frac{|\theta|}{c_\theta} + \omega_{\dot{\theta}} \frac{|\dot{\theta}|}{c_{\dot{\theta}}}$$

$$d_t = \begin{cases} +1, & T = 500 \\ -1, & \text{실패 종료} \\ J(s_t) - J(s_{t+1}), & \text{일반 step} \end{cases}$$

$J(s_t) < J(s_{t+1})$, (행동 후 더 불안정) $\quad J(s_t) > J(s_{t+1})$ (행동 후 더 안정)

도파민 신호 계산

기호	의미
$J(s)$	불안정도
ω	상태 수치의 중요도
c	상태값 정규화 기준

최종 가중치 계산 / 학습률 / 어떤 연결인가? / 결과가 좋았는가?

$$\underbrace{\Delta w_{ij,t}}_{\text{Weight update}} = \eta \times \underbrace{e_{ij,t}}_{\text{Which synapse}} \times \underbrace{d_t}_{\text{Good or bad}}$$

가중치 계산

가중치 계산 예시

$j = 0$ (왼쪽 선택) $\quad \eta = 0.001 \quad e_{i0,t} = 0.4$

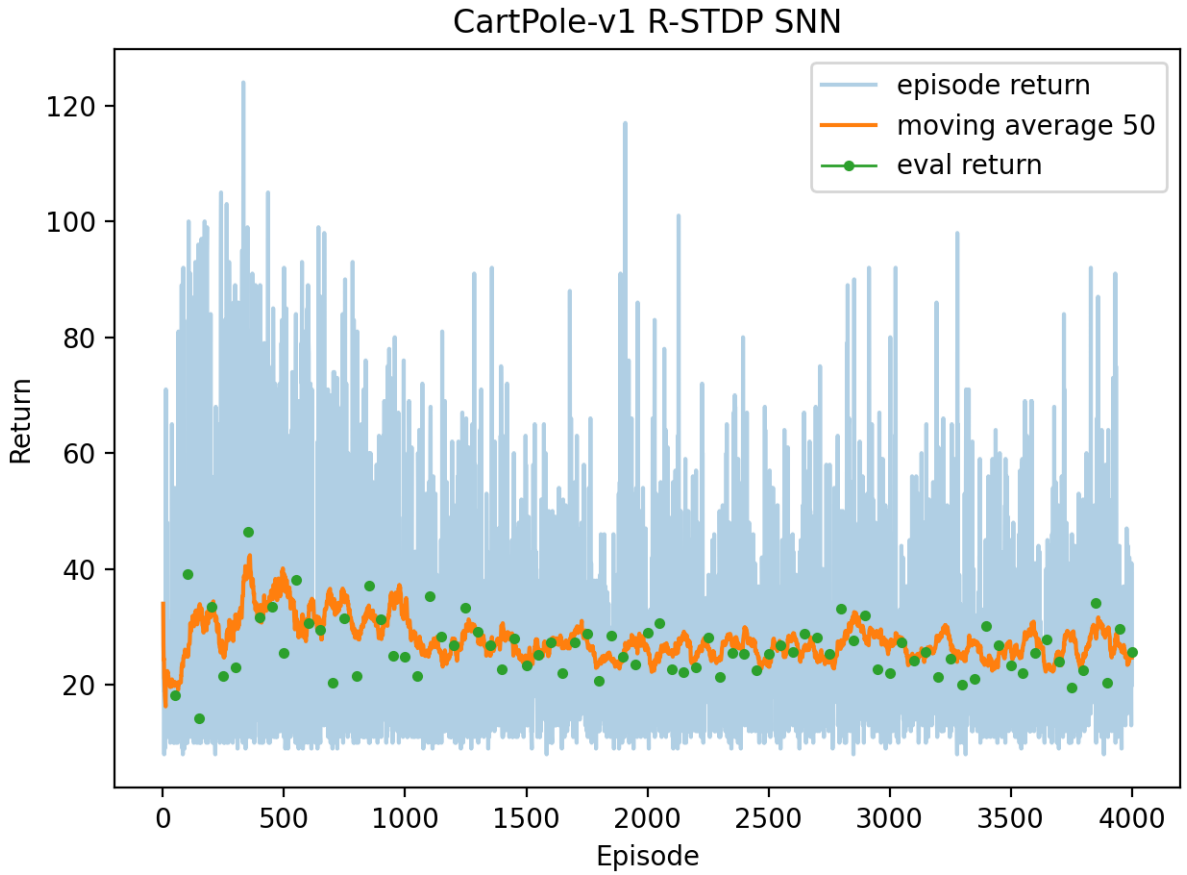
$J(s_t) = 0.60, \quad J(s_{t+1}) = 0.55$

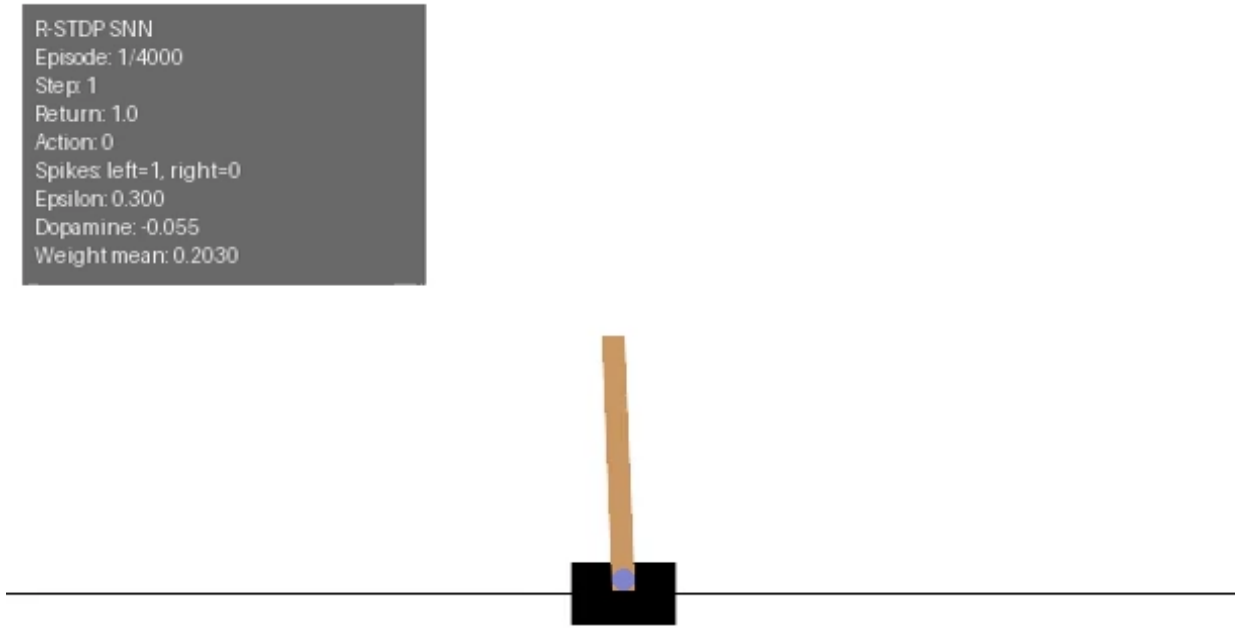
$$d_t = J(s_t) - J(s_{t+1}) = 0.05$$

$$\boxed{\Delta w_{i0,t} = 0.001 \times 0.05 \times 0.4 = 0.00002}$$

R-STDP: 보상 조절 STDP

입력·출력 spike의 시간적 상관관계로 STDP eligibility를 형성하고,
행동 결과로 계산한 도파민 유사 신호를 결합하여 선택된 행동 뉴런의 시냅스를 매 환경 step마다 갱신





Episodic R-STDP: 에피소드별 R-STDP

R-STDP는 매 환경 step의 상태 변화를 즉시 평가하지만, Episodic R-STDP는 한 에피소드가 끝날 때까지 eligibility를 저장한 후 전체 성과가 baseline보다 좋았는지를 기준으로 학습

$$e_{ij,t} = \frac{1}{K}\sum_{\tau=1}^{K}\left(A_{+}\mathrm{LTP}_{ij}^{(\tau)} - A_{-}\mathrm{LTD}_{ij}^{(\tau)}\right)$$

여기까지 기존 R-STDP와 동일하나
Episodic R-STDP는 이 eligibility로
가중치를 바로 갱신하지 않음

STDP eligibility 계산

$$G_n = \sum_{t=1}^{T_n} r_t$$

예를 들어
80 step 생존 : $G_n = 80$
300 step 생존 : $G_n = 300$

CartPole은 생존한 매 step마다
+1의 보상을 제공

Episode의 생존 return 계산

$$A_n^{\mathrm{epi}} = G_n - b_n$$

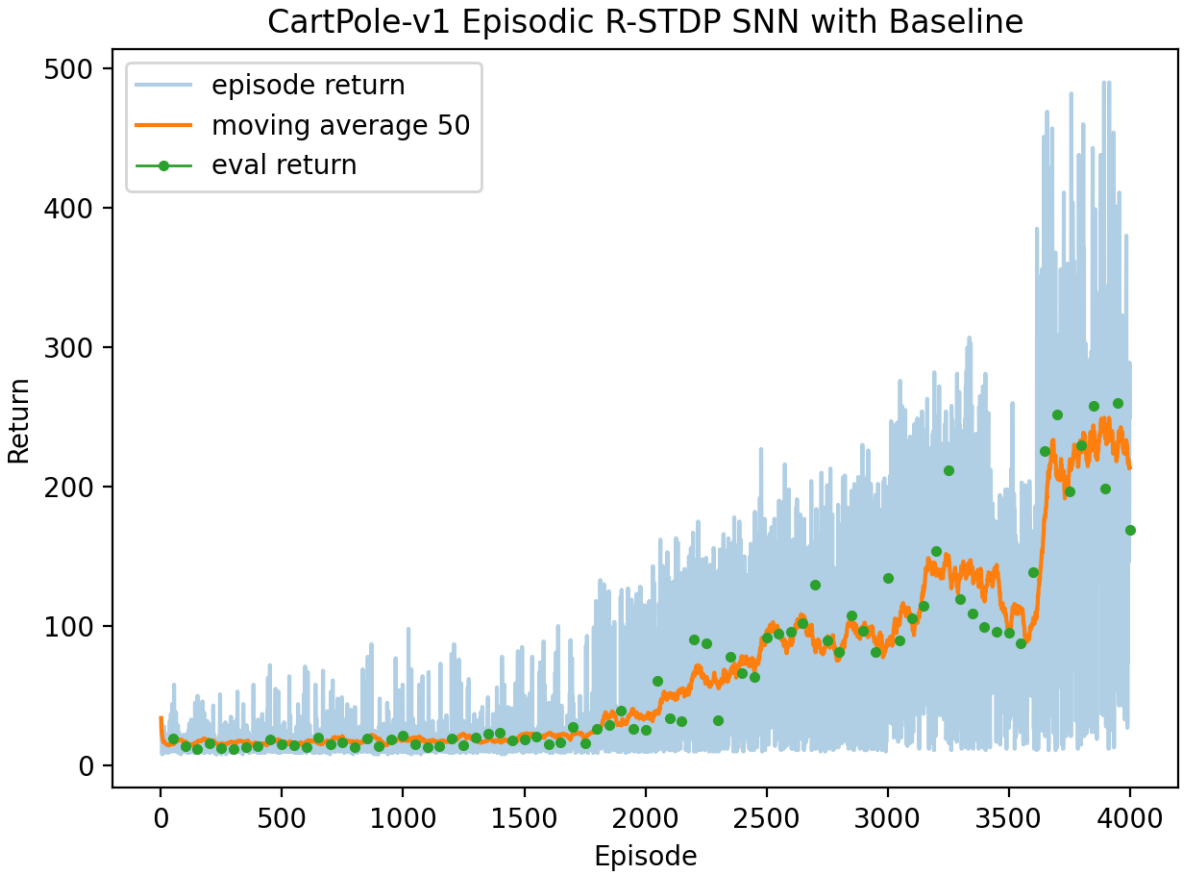
$G_n > b_n$: 평소보다 좋은 episode
$G_n < b_n$: 평소보다 나쁜 episode
$G_n = b_n$: 평소보다 비슷한 episode

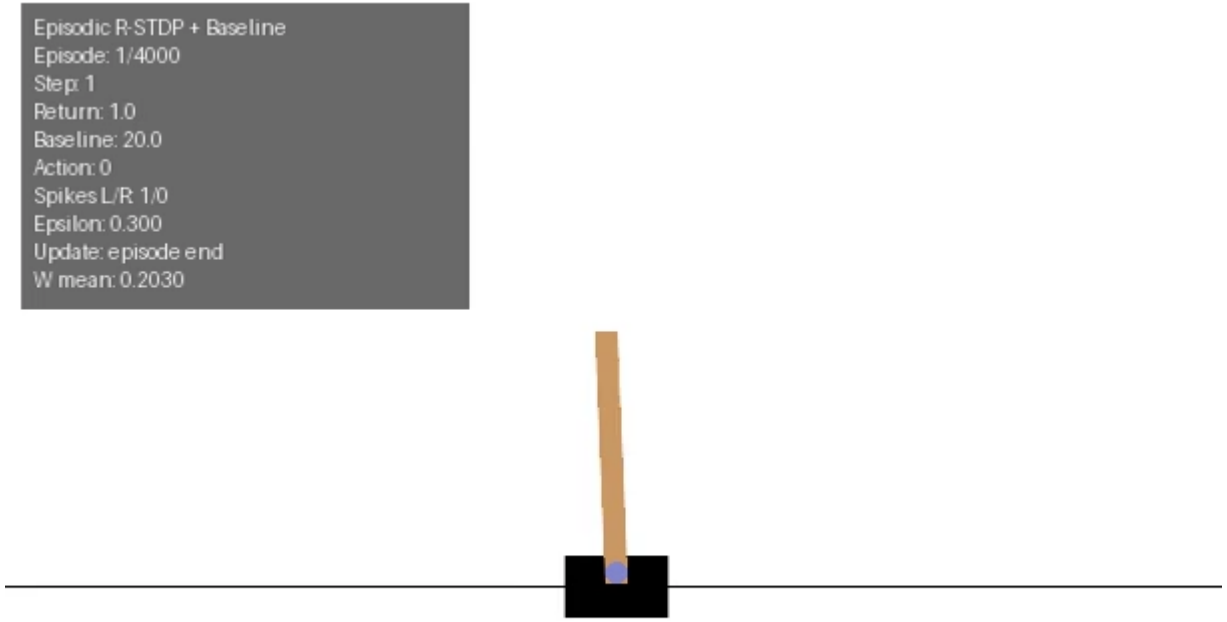
episode return만 사용하면 항상 양수 (강화 판단이 어려움)
때문에 최근의 평균적 성능을 나타내는 baseline 과 비교
(R-STDP의 d_t 역할)

Running baseline과 비교

Episodic R-STDP: 에피소드별 R-STDP

R-STDP는 매 환경 step의 상태 변화를 즉시 평가하지만, Episodic R-STDP는 한 에피소드가 끝날 때까지 eligibility를 저장한 후 전체 성과가 baseline보다 좋았는지를 기준으로 학습





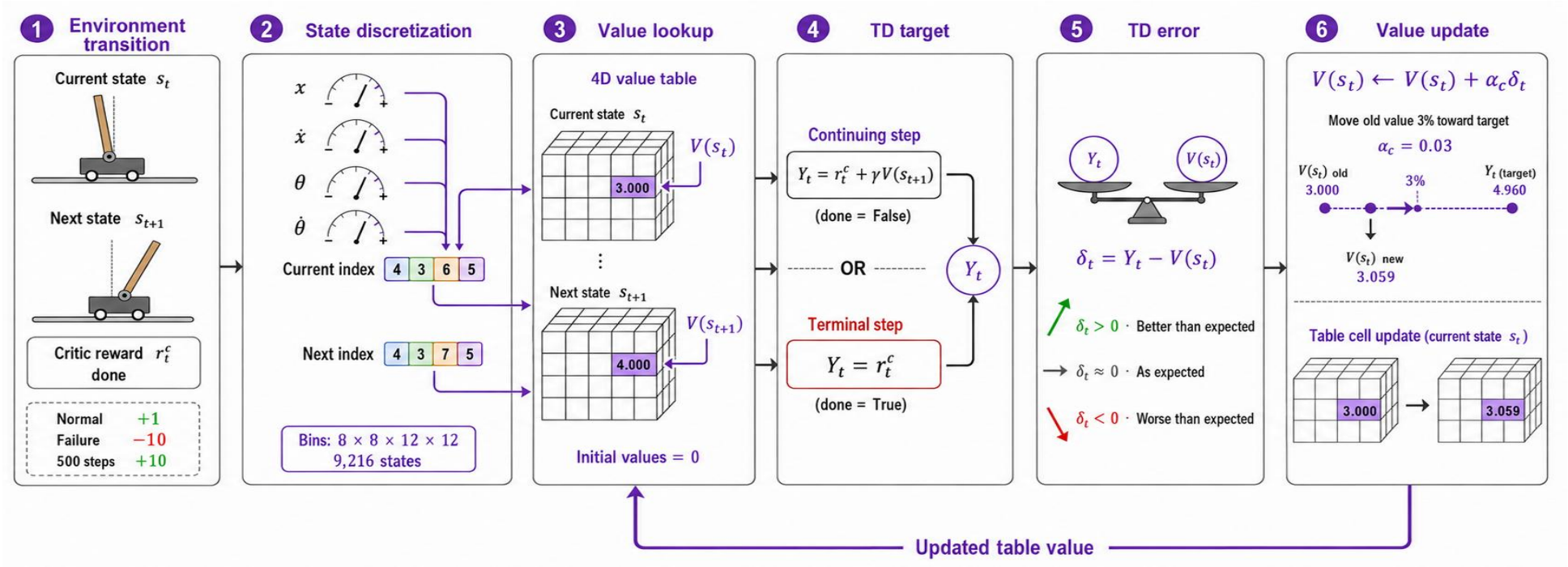
TD-STDP Actor-Critic: TD error 조절 STDP

SNN Actor가 행동을 선택하고, Critic이 행동 전후 상태의 가치 차이로 TD error를 계산하며,
이 TD error가 STDP eligibility를 조절하는 도파민 신호로 사용

$V(s) \approx$ 상태 s에서 앞으로 받을 누적 보상의 기대값

막대가 수직에 가깝고 움직임이 안정적: 높은 $V(s)$
막대가 크게 기울고 빠르게 넘어지는 중: 낮은 $V(s)$
아직 경험하지 못한 상태: 초기값 $V(s) = 0$

Critic : 상태가치 $V(s)$를 학습



상태 이산화 → table 조회 → TD target 계산 → table 갱신

TD-STDP Actor-Critic: TD error 조절 STDP

SNN Actor가 행동을 선택하고, Critic이 행동 전후 상태의 가치 차이로 TD error를 계산하며,
이 TD error가 STDP eligibility를 조절하는 도파민 신호로 사용

$$[x, \dot{x}, \theta, \dot{\theta}] \rightarrow [k_x, k_{\dot{x}}, k_\theta, k_{\dot{\theta}}]$$

Critic은 table 형태이기 때문에 연속 상태를 그대로 인덱스로 사용할 수 없음 때문에 각 상태축을 여러 구간으로 나눔

상태	범위	Bin 수
x_t	([-2.4,2.4])	8
$\dot{x}_t$	([-3.0,3.0])	8
θ_t	([-0.2095,0.2095])	12
$\dot{\theta}_t$	([-3.5,3.5])	12

$$\phi(s_t) = [4,3,6,5] \quad \phi(s_{t+1}) = [4,3,7,5]$$

상태 이산화

$$V(s_t) = V[4,3,6,5] \quad V(s_{t+1}) = V[4,3,7,5]$$

현재 상태와 다음 상태에 대응하는 값을 table에서 읽음
(초기에는 모두 0으로 시작)

table 조회

$$Y_t = r_t^{\mathrm{C}} + \gamma V(s_{t+1})$$

$$r_t^{\mathrm{C}} = \begin{cases} +10, & T = 500 \\ -10, & \text{실패 종료} \\ +1, & \text{일반 step} \end{cases}$$

$$\delta_t = Y_t - V(s_t)$$

기호	의미
r_t^{C}	지금 받은 보상
$\gamma V(s_{t+1})$	다음 상태부터 기대되는 미래 가치
γ	미래 할인률

TD error	Critic의 판단
$\delta_t > 0$	결과가 예상보다 좋음
$\delta_t < 0$	결과가 예상보다 나쁨
$\delta_t \approx 0$	예상과 결과가 비슷함

TD target 계산

TD-STDP Actor-Critic: TD error 조절 STDP

SNN Actor가 행동을 선택하고, Critic이 행동 전후 상태의 가치 차이로 TD error를 계산하며,
이 TD error가 STDP eligibility를 조절하는 도파민 신호로 사용

TD target 계산 예시

$$r_t^{\mathrm{C}} = 1, \qquad \gamma = 0.99,$$
$$V(s_t) = 3, \quad V(s_{t+1}) = 4$$

$$Y_t = 1 + 0.99 \times 4 = 4.96$$

현재 상태에서 +1의 보상을 받고, 가치가 4인 다음 상태로
이동했으므로 현재 상태의 가치는 약 4.96이 되는 것이 적절

$$\delta_t = 1.96 > 0$$

현재 상태의 가치를 3으로 저장해 두었지만,
이번 경험을 보면 4.96에 더 가까워야 한다
(예상보다 결과가 좋다)

$$V(s_t) \leftarrow V(s_t) + \eta_{\mathrm{C}} \delta_t$$

η_C: Critic 학습률 (예시는 0.03)

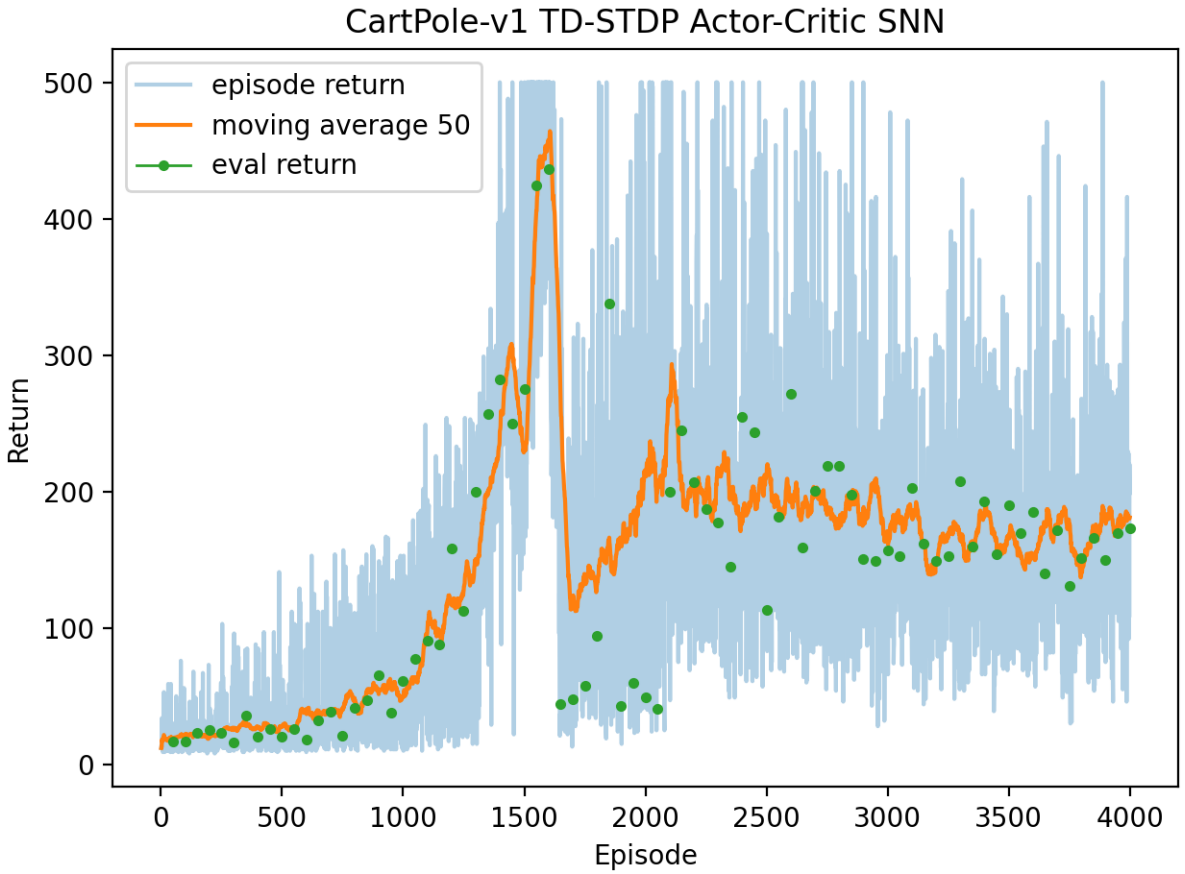
$$V(s_t) \leftarrow 3 + 0.03 \times 1.96 = 3.0588$$

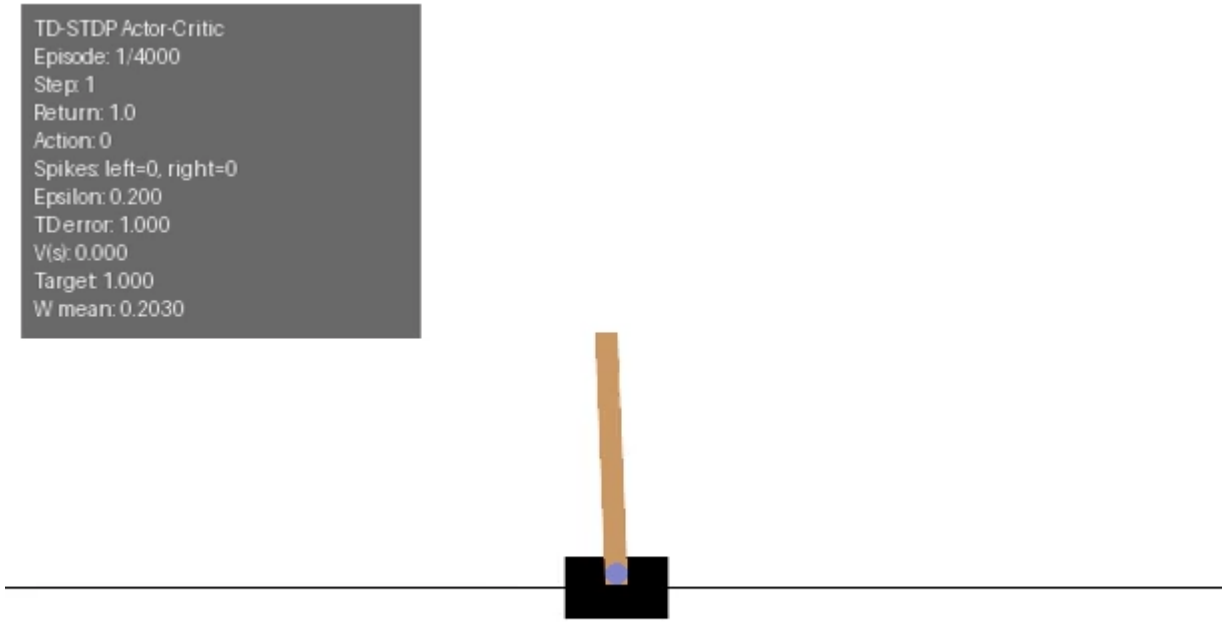
$$V[4,3,6,5]: 3.000 \rightarrow 3.0588$$

table 갱신

TD-STDP Actor-Critic: TD error 조절 STDP

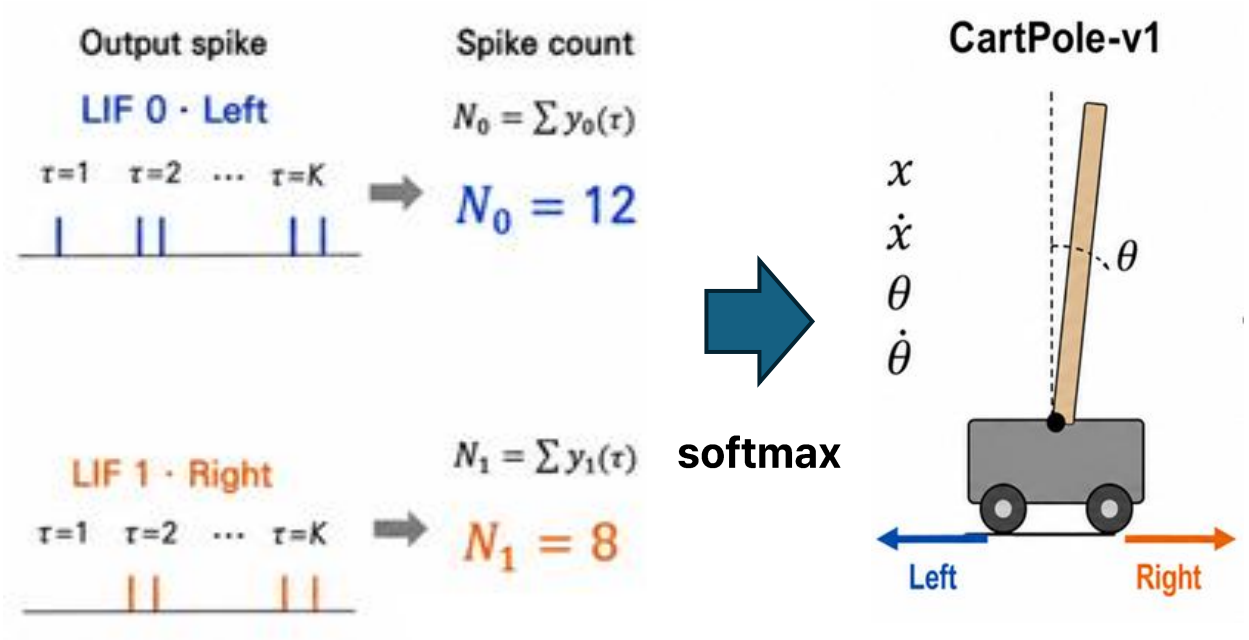
SNN Actor가 행동을 선택하고, Critic이 행동 전후 상태의 가치 차이로 TD error를 계산하며,
이 TD error가 STDP eligibility를 조절하는 도파민 신호로 사용





PPO-modulated TD-STDP: 과하지 않은 TD error 조절 STDP

TD-STDP처럼 Critic의 가치 추정을 사용하지만, 단일 TD error를 즉시 Actor에 적용하지 않고,
episode 전체에서 계산한 advantage와 PPO probability ratio를 이용해 STDP 학습 신호를 제한



확률적 행동 적용

PPO-modulated TD-STDP: 과하지 않은 TD error 조절 STDP

TD-STDP처럼 Critic의 가치 추정을 사용하지만, 단일 TD error를 즉시 Actor에 적용하지 않고, episode 전체에서 계산한 advantage와 PPO probability ratio를 이용해 STDP 학습 신호를 제한

최종 가중치 계산 / 학습률 / 학습을 허용할 것인가? / 어떤 시냅스가 관여했는가

$$\Delta w_{i,a_t,t} = \eta \times m_t^{\mathrm{PPO}} \times e_{i,a_t,t}$$

행동이 좋았는가 (도파민 신호)

$$m_t^{\mathrm{PPO}} = \begin{cases} 0, & \bar{A}_t \geq 0 \wedge \rho_t > 1.2 \\ 0, & \bar{A}_t < 0 \wedge \rho_t < 0.8 \\ \min\left(\rho_t \bar{A}_t, \rho_t^{\mathrm{clip}} \bar{A}_t\right), & \bar{A}_t \geq 0 \\ \max\left(\rho_t \bar{A}_t, \rho_t^{\mathrm{clip}} \bar{A}_t\right), & \bar{A}_t < 0 \end{cases}$$

정책이 얼마나 변했는가

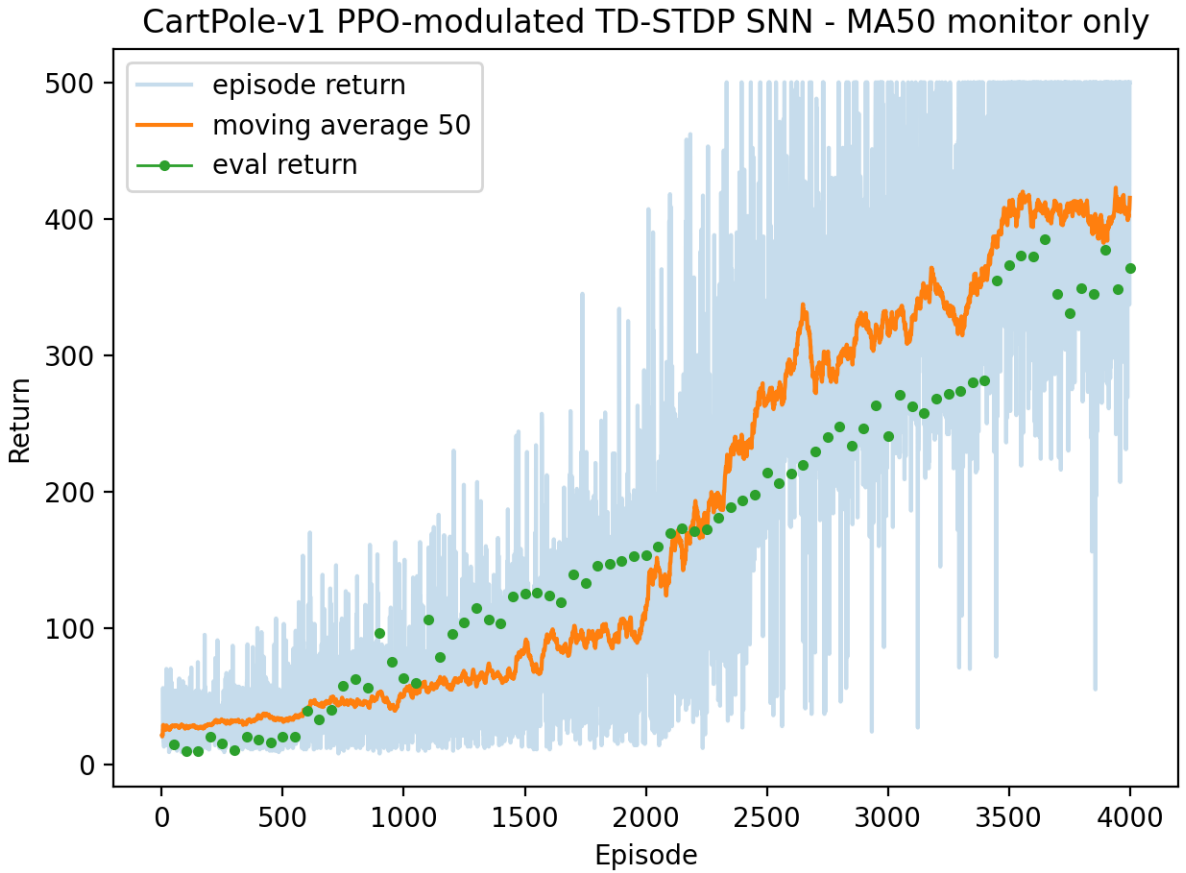
$$\rho_t = \frac{p_t^{\mathrm{new}}}{p_t^{\mathrm{old}}} = \frac{\pi_{\mathrm{new}}(a_t \mid s_t)}{\pi_{\mathrm{old}}(a_t \mid s_t)}$$

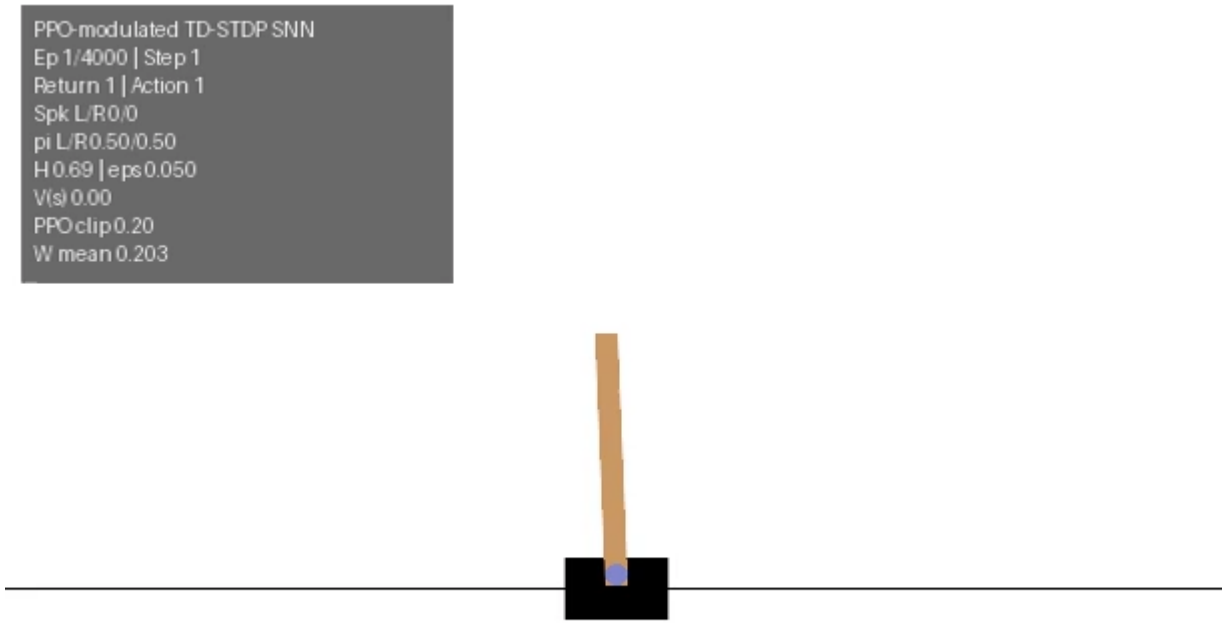
$$\rho_t^{\mathrm{clip}} = \mathrm{clip}(\rho_t, 0.8, 1.2)$$

행동 평가	정책 변화	처리
좋은 행동 $\bar{A}_t \geq 0$	확률 증가가 20% 이내	강화
좋은 행동 $\bar{A}_t \geq 0 \wedge \rho_t > 1.2$	확률이 20% 초과 증가	추가 강화 차단
나쁜 행동 $\bar{A}_t < 0$	확률 감소가 20% 이내	약화
나쁜 행동 $\bar{A}_t < 0 \wedge \rho_t < 0.8$	확률이 20% 초과 감소	추가 약화 차단
정책이 잘못된 방향으로 변화	좋은 행동 감소 또는 나쁜 행동 증가	반대 방향으로 수정

PPO-modulated TD-STDP: 과하지 않은 TD error 조절 STDP

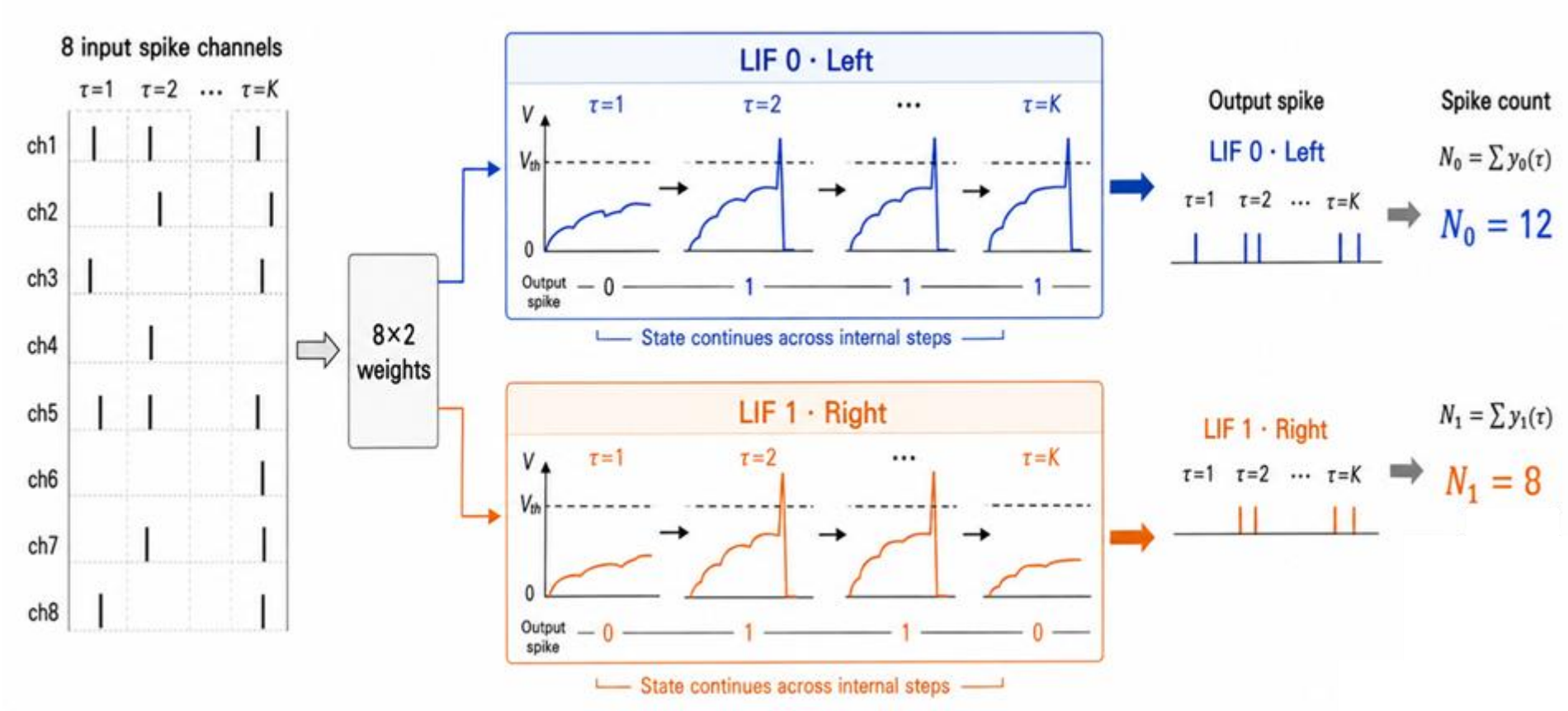
TD-STDP처럼 Critic의 가치 추정을 사용하지만, 단일 TD error를 즉시 Actor에 적용하지 않고,
episode 전체에서 계산한 advantage와 PPO probability ratio를 이용해 STDP 학습 신호를 제한





종합 결론

강화학습의 **상태/보상 개념**과 **SNN의 STDP 방식을 함께 사용**하여 **R-STDP 구현 및 다양한 강화학습 알고리즘을 적용**시키는데 성공하였다.



종합 결론

하지만 정책 붕괴 등의 강화학습적 문제점들이 발견되었으며, 단순한 입력·출력 spike의 시간적 상관관계를 통해 갱신하는 R-STDP 같은 policy 계열보다 Actor-Critic 계열의 모델이 전체적으로 성능이 높았다.