J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr
Copyright ⓒ 2026 KSII

# 도파민 유사 학습 신호 기반 STDP-SNN 강화학습 방법 분석☆

# Analysis of Dopamine-like Learning Signal-Based STDP-SNN Reinforcement Learning Methods

문 종 인[1] 노 기 섭[1*]
Jongin Mun Giseop Noh

요 약

본 논문은 Spike-Timing-Dependent Plasticity(STDP) 기반 Spiking Neural Network(SNN)에 강화학습을 적용할 때, 시냅스 가중치 갱신을 조절하는 도파민 유사 학습 신호의 설계가 학습 성능과 안정성에 미치는 영향을 비교 분석한다. 이를 위해 CartPole-v1 환경에서 SNN이 직접 행동을 선택하는 actor 구조를 구성하고, 동일한 spike encoding 방식과 SNN actor 구조를 사용하는 Actor-Critic TD-STDP와 Proximal Policy Optimization(PPO) TD-STDP를 비교하였다. Actor-Critic TD-STDP는 매 환경 step에서 계산한 Temporal-Difference error (TD-error)를 도파민 유사 학습 신호로 사용하며, PPO TD-STDP는 episode 단위 Generalized Advantage Estimation(GAE)를 통해 추정한 advantage의 정규화 및 제한과 policy ratio 기반 갱신 제한을 적용하였다. 실험 결과, Actor-Critic TD-STDP는 학습 초반에 빠른 성능 향상을 보였으나 학습 후반에는 성능 변동성이 증가하였다. 반면 PPO TD-STDP는 초기 학습 속도는 상대적으로 느렸지만, 급격한 시냅스 가중치 변화와 정책 변화를 완화하여 비교적 안정적인 학습 경향을 보였다. 이러한 결과는 STDP-SNN 강화학습에서 도파민 유사 학습 신호의 시간적 반영 범위와 정책 변화 제한 방식을 학습 설계 시 함께 고려할 필요가 있음을 시사한다. 본 연구는 동일한 SNN 구조에서 도파민 유사 학습 신호와 이에 결합된 정책 갱신 방식의 영향을 비교함으로써, 향후 생물학적으로 타당하고 안정적인 SNN 강화학습 알고리즘을 설계하고 평가하기 위한 기초 자료와 설계 방향을 제공한다.

☞ 주제어 : SNN, STDP, 강화학습, 도파민 유사 학습 신호, CartPole, Actor-Critic

ABSTRACT

This paper analyzes how the design of dopamine-like learning signals affects the learning performance and stability of reinforcement learning in Spiking Neural Networks(SNNs) based on Spike-Timing-Dependent Plasticity(STDP). An SNN actor that directly selects actions was implemented in the CartPole-v1 environment. Actor-Critic TD-STDP and Proximal Policy Optimization(PPO) TD-STDP were compared using the same spike encoding method and SNN actor architecture. Actor-Critic TD-STDP uses the Temporal-Difference error(TD-error) calculated at each environment step as a dopamine-like learning signal. In contrast, PPO TD-STDP uses normalized and clipped episode-level advantages estimated through Generalized Advantage Estimation(GAE), together with policy-ratio-based update clipping. The results showed that Actor-Critic TD-STDP improved rapidly during early training but exhibited greater performance variability in later training. PPO TD-STDP learned more slowly initially but maintained relatively stable performance by mitigating abrupt changes in synaptic weights and the policy. These findings suggest that the temporal range of dopamine-like learning signals and the approach used to constrain policy changes should be considered together when designing STDP-SNN reinforcement learning algorithms. By comparing learning-signal and policy-update designs under an identical SNN architecture, this study provides a foundation and design guidance for developing and evaluating biologically plausible and stable SNN reinforcement learning algorithms.

☞ keyword : Spiking Neural Networks; Spike-Timing-Dependent Plasticity; Reinforcement Learning; Dopamine-like Learning Signal; CartPole; Actor-Critic

## 1. 서 론

Spiking Neural Network(SNN)는 생물학적 신경망의 정보처리 방식을 모방한 인공신경망으로, 뉴런 간 정보를 연속적인 실수값이 아니라 이산적인 전위 차이 신호인 스파이크(spike) event를 통해 전달한다.

SNN의 학습 중 하나인 Spike-Timing-Dependent Plasticity

[1] Dept. of Software and Communications Engineering, Hongik University, Sejong Special City, 2639, Korea
* Corresponding author (kafa46@hongik.ac.kr)
[Received 19 June 2026, Reviewed 11 August 2026, Accepted 14 August 2026]
☆ 본 연구는 과학기술정보통신부/정보통신산업진흥원의 세종 SW융합클러스터2.0 사업의 지원으로 수행된 연구입니다 (과제번호: PJTS080126041007000100030200).

(STDP)는 시냅스 뉴런 사이의 spike 발생 시점 차이에 따라 시냅스 가중치를 강화하거나 약화하는 SNN 학습 방식이다. 이 방식은 각 시냅스가 국소적인 스파이크 시간 정보에 기반하여 업데이트된다는 점에서 생체 신경망의 시냅스 가소성을 설명하는 대표적인 학습 방식으로 볼 수 있다[1,2]. 그러나 기본적인 STDP는 외부 보상이나 장기적인 목표를 직접 반영하지 못하기 때문에, 보상 또는 critic 기반 강화학습 신호를 결합한 STDP 방식이 필요하다[3,4].

딥러닝 기반 학습에는 역전파를 통해 학습하는 방식이 주로 사용된다. 그러나 이러한 학습 방식은 미분 불가능한 스파이크 신호와 국소적 시냅스 업데이트 특성을 가지는 SNN에 직접 적용하기 어렵다. 따라서 SNN 기반 강화학습에서는 temporal-difference error(TD-error)와 같은 강화학습 신호를 선택된 행동에 관여한 시냅스의 가중치를 조절하는 도파민 유사 학습 신호로 활용할 수 있다[3-5]. 즉, SNN actor가 spike activity를 통해 행동을 선택하면, 보상 또는 critic 기반 신호가 eligibility trace와 결합되어 시냅스 가중치를 갱신한다[3,4,6].

최근에는 SNN을 강화학습 및 제어 문제에 적용하려는 연구가 활발히 진행되고 있다. 학습 신호 측면에서는 critic 기반 학습 신호를 결합한 Actor-Critic 구조와 policy 기반 학습 방법이 제안되고 있다[6,7]. 구조적 측면에서는 SNN actor의 출력 표현과 내부 연결 구조를 개선하거나, 환경 상태를 spike 형태로 표현하기 위한 인코딩 방법을 강화하는 연구가 진행되고 있다[8-11]. 이러한 연구들은 SNN 기반 강화학습에서 네트워크 구조뿐 아니라 시냅스 가중치를 조절하는 학습 신호의 정의가 성능과 안정성에 영향을 줄 수 있음을 시사한다.

그러나 최근 연구의 상당수는 surrogate gradient와 역전파 기반 강화학습을 이용하여 SNN actor를 학습하는 데 초점을 두고 있으며, 동일한 STDP 기반 SNN 구조에서 도파민 유사 학습 신호의 계산 방식에 따른 학습 특성을 비교한 연구는 제한적이다. 본 논문은 STDP 기반 SNN actor를 사용한 강화학습에서 도파민 학습 신호의 설계가 학습 성능과 안정성에 미치는 영향을 비교 분석하였으며, 향후 생물학적으로 타당하고 안정적인 SNN 강화학습 알고리즘을 설계하고 평가하기 위한 기초 자료와 설계 방향을 제공하는 것을 목표로 한다. 이를 위해 CartPole-v1 환경을 대상으로 두 가지 STDP 기반 SNN 강화학습 방법을 비교하였고, 비교 대상은 1-step TD-error를 도파민 학습 신호로 사용하는 Actor-Critic TD-STDP, Proximal Policy Optimization(PPO)의 policy ratio 기반 갱신 제한 원리와 Generalized Advantage Estimation(GAE)을 STDP 기반 도파민 학습 신호에 결합한 PPO TD-STDP이다. 두 방법은 모두 동일한 SNN 구조를 사용하며, 차이는 Actor의 행동 선택 방식과 시냅스 업데이트에 사용되는 도파민 학습 신호의 정의에 있다.

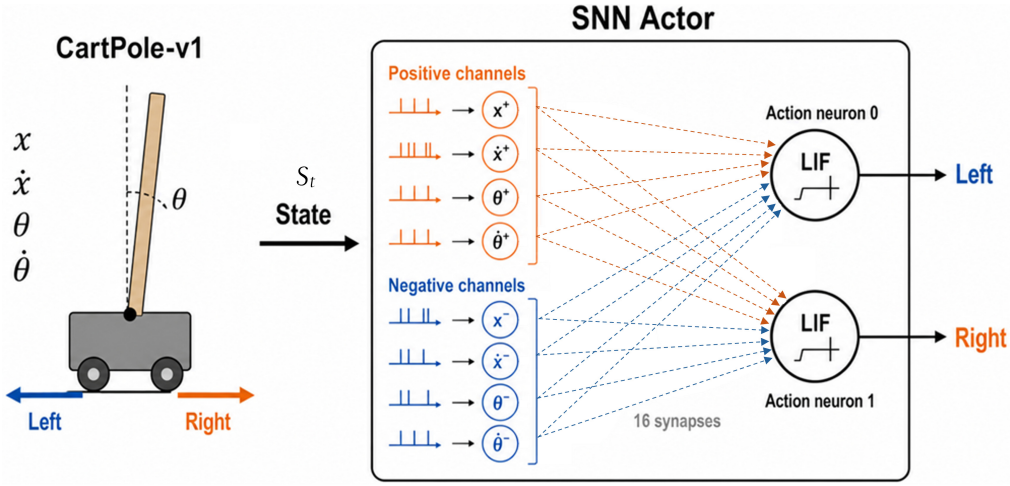
## 2. 실험 환경 및 SNN 구성

### 2.1 환경 구성

본 연구에서는 실험 환경으로 CartPole-v1을 사용하였다. CartPole-v1은 강화학습 분야에서 널리 사용되는 대표적인 제어 문제이며, 최근 SNN 기반 deep reinforcement learning 연구에서도 CartPole과 같은 제어 환경이 SNN 구조의 성능과 한계를 분석하기 위한 benchmark로 활용되고 있다[8]. Agent는 매 환경 step마다 cart를 왼쪽 또는 오른쪽으로 이동시키는 행동을 선택하며, pole의 각도가 수직축으로부터 $\pm 12°$를 초과하거나 cart의 위치가 $\pm 2.4$ 범위를 벗어나면 실패로 종료되며, 최대 500 environment step 동안 균형을 유지하면 성공으로 종료된다. CartPole-v1의 상태는 cart와 pole의 동역학을 나타내는 네 개의 연속 변수로 구성된다. 시간 $t$에서의 상태 $s_t$는 $s_t = [x_t, \dot{x}_t, \theta_t, \dot{\theta}_t]$라고 표현할 수 있으며, $x_t$는 cart의 수평 위치, $\dot{x}_t$는 cart의 수평 속도, $\theta_t$ 는 pole이 수직축으로부터 기울어진 각도, $\dot{\theta}_t$는 pole의 각속도를 의미한다. CartPole-v1의 행동은 시간 $t$에서의 행동 $a_t$로 정의할 수 있다. $a_t = 0$은 cart를 왼쪽으로 이동, $a_t = 1$ 은 오른쪽으로 이동을 의미한다.

### 2.2 Spike Encoding

CartPole-v1의 상태는 네 개의 연속 변수로 구성되지만, SNN actor는 연속값을 직접 입력으로 사용하지 않고 spike 형태의 이산적인 입력 신호를 사용한다. 본 연구에서는 연속적인 환경 상태를 별도의 encoder 학습 없이 spike 발생 빈도로 표현하여 STDP 기반 학습에 활용하기 위해 Poisson 기반 spike encoding을 적용하였다. SNN 기반 강화학습에서는 연속적인 환경 상태를 spike 기반 입력으로 변환하거나, SNN 내부에서 spike 표현을 형성하여 행동 선택에 활용하는 방식이 사용된다[8-11].




(그림 1) CartPole-v1 환경 및 SNN Actor 구조
(Figure 1) CartPole-v1 Environment and SNN Actor Architecture

본 연구에서는 상태 변수 $[x_t, \dot{x_t}, \theta_t, \dot{\theta_t}]$을 SNN 입력으로 변환하기 위해 시뮬레이션의 허용 수치를 고려하여 $[2.4, 3.0, 0.2095, 3.5]$을 기준으로 $[-1,1]$ 범위로 정규화하였다. 이후 변수의 크기뿐만 아니라 방향성도 SNN actor에 전달하기 위해 positive channel과 negative channel로 분리되어 최종적으로 8차원 입력 벡터가 된다. 이를 통해 위치, 방향, 기울기 방향, 각속도 방향을 다른 입력 channel의 spike pattern으로 구분할 수 있으며 정규화된 값에 비례한 확률로 입력 spike가 발생한다.

## 2.3 SNN 구조

본 연구에서 비교하는 두 강화학습 방법은 동일한 SNN 구조를 사용한다. 그림 1은 본 연구에서 공통적으로 사용한 SNN의 전체 구조를 나타내며, 은닉층이 없는 단순한 feedforward 구조로 설계하여, 8개의 input spike channel과 2개의 Leaky Integrate-and-Fire(LIF) action neuron으로 구성된다. SNN은 연속적인 출력값이나 일반적인 확률 분포를 직접 출력하는 대신, 하나의 환경 상태에 대해 SNN actor를 $K$번 동안 실행하고, 그동안 발생한 action neuron의 spike count를 기반으로 행동을 선택한다. 이때의 각 과정을 SNN simulation step이라 하며 이러한 구조는 SNN이 행동을 직접 생성하는 정책으로 작동한다는 점에서 일반적인 역전파 기반 인공신경망 actor와 구분된다. 또한 두 비교 알고리즘은 동일한 SNN 구조를 공유하므로, 알고리즘에 따른 성능의 차이는 시냅스 가중치 조절에 사용되는 도파민 학습 신호, 행동 선택 및 정책 갱신 방식의 차이에서 발생한다.

## 2.4 STDP 기반 시냅스 가소성

SNN에서는 뉴런 사이의 spike 발생 시점에 따라 시냅스 가중치를 조절하는 STDP가 대표적인 학습 규칙으로 사용된다. 일반적으로 pre-synaptic spike가 post-synaptic spike보다 먼저 발생하면 시냅스 가중치가 강화되는 Long-Term Potentiation(LTP)이 발생하고, 반대 순서로 발생하면 가중치가 약화되는 Long-Term Depression(LTD)이 발생한다[1,2].

본 연구에서는 $k$번째 SNN simulation step에서 두 trace를 수식 1과 같이 갱신하며, $x_i^{(k)}$와 $y_a^{(k)}$는 각각 입력 spike channel $i$와 action neuron $a$의 spike 발생 여부이며, $P_i^{(k)}$와 $Q_a^{(k)}$는 각각 최근 입력 및 출력 spike의 흔적이며, $\beta_{tr}$는 과거 spike 정보가 시간에 따라 감소하는 정도를 조절하는 trace decay 계수로 1에 가까울수록 과거 spike 정보가 더 오래 유지된다.

$$P_i^{(k)} = \beta_{tr} P_i^{(k-1)} + x_i^{(k)} \tag{1}$$

$$Q_a^{(k)} = \beta_{tr} Q_a^{(k-1)} + y_a^{(k)}$$

# 3. 연구 방법

## 3.1 Dopamine-modulated STDP Update

본 연구에서는 2.4절에서 정의한 두 trace를 이용하여, 하나의 환경 step에서 시냅스별 eligibility $e_{ia,t}$를 수식 2와 같이 계산한다.

$$e_{ia,t} = \frac{1}{K}\sum_{k=1}^{K}\left[A_{+}P_i^{(k)}y_a^{(k)} - A_{-}x_i^{(k)}Q_a^{(k)}\right] \tag{2}$$

여기서 $A_{+}$와 $A_{-}$는 각각 LTP와 LTD의 크기를 조절하는 계수이고, 첫 번째 항은 입력 spike의 흔적이 남아 있는 상태에서 action neuron이 발화한 경우의 LTP 성분, 두 번째 항은 action neuron의 발화 흔적이 남아 있는 상태에서 입력 spike가 발생한 경우의 LTD 성분이다. 이후 선택된 행동 $a_t$에 대응하는 $e_{ia_t,t}$는 도파민 학습 신호 $D_t$와 결합되어 선택된 시냅스 가중치 $w_{ia_t}$를 갱신한다.

$$\Delta w_{ia_t} = \eta_a D_t\, e_{ia_t,t} \tag{3}$$

$$w_{ia_t} \leftarrow clip\left(w_{ia_t} + \Delta w_{ia_t},\ w_{\min},\ w_{\max}\right)$$

여기서 $w_{ia_t}$ 는 입력 spike channel $i$와 선택된 action neuron $a_t$ 사이의 시냅스 가중치, $\eta_a$는 학습률이다.

## 3.2 Actor-Critic TD-STDP

Actor-Critic TD-STDP는 SNN을 actor로 사용하고, 별도의 critic이 상태 가치 $V(s_t)$를 추정하는 강화학습 방법이다. Actor는 상태 $s_t$에 대해 $K$개의 SNN simulation step을 실행하여 spike count를 비교하여 더 높은 count의 행동 $a_t$를 선택한다.

학습 중에는 $\epsilon$-greedy 탐험을 적용하여 탐험 발생 시 actor가 임의의 행동을 선택할 수 있다. 에피소드 번호 $e$에서의 탐험 확률 $\epsilon_e$는 초기 탐험 확률 $\epsilon_{start}$, 탐험 확률의 최솟값 $\epsilon_{end}$, 최솟값까지 감소하는 데 필요한 episode 수 $E_{decay}$에 따라서 선형적으로 감소하며 수식 4와 같다.

$$\epsilon_e = \max\left(\epsilon_{end},\ \epsilon_{start} - e\frac{\epsilon_{start}-\epsilon_{end}}{E_{decay}}\right) \tag{4}$$

Critic은 우선 연속적인 $s_t$를 각 상태 변수별로 이산화하고, 상태의 가치를 테이블에 $V[k_x,k_{\dot{x}},k_\theta,k_{\dot{\theta}}]$ 형태로 저장한다. 이후 critic은 수식 5와 같이 행동 $a_t$를 수행한 후 보상 $r_t$, 할인율 $\gamma$, critic 학습률 $\eta_c$, 마지막 환경 step $T_e$를 통해 TD-error $\delta_t$를 계산하고 $V(s_t)$를 갱신한다.

$$\delta_t = \begin{cases} r_t - V(s_t), & t = T_e \\ r_t + \gamma V(s_{t+1}) - V(s_t), & t < T_e \end{cases} \tag{5}$$

$$V(s_t) \leftarrow V(s_t) + \eta_c\delta_t$$

$$r_t = \begin{cases} 10, & (t = T_e,\ T_e = 500) \\ 1, & (t < T_e) \\ -10, & (t = T_e,\ T_e < 500) \end{cases}$$

Actor 학습에는 $\delta_t$를 $[c_{\min}^{TD}, c_{\max}^{TD}]$ 범위로 제한하여 급격한 시냅스 변화를 방지하고, $\epsilon$-greedy 탐험으로 선택된 행동은 현재 SNN 정책이 직접 선택한 행동이 아니므로 탐험 보정 계수 $m_t$를 적용한다. 이를 통해 Actor-Critic TD-STDP의 도파민 학습 신호 $D_t^{TD}$는 수식 6과 같이 정의하며, 환경 step마다 $D_t^{TD}$를 통해 해당 시냅스를 업데이트 한다.

$$D_t^{TD} = m_t\overline{\delta}_t \tag{6}$$

$$\overline{\delta}_t = \begin{cases} c_{\min}^{TD}, & \delta_t < c_{\min}^{TD} \\ \delta_t, & c_{\min}^{TD} \le \delta_t \le c_{\max}^{TD} \\ c_{\max}^{TD}, & \delta_t > c_{\max}^{TD} \end{cases}$$

$$m_t = \begin{cases} 0.2, & exploratory\ action \\ 1.0, & policy\ action \end{cases}$$

## 3.3 PPO TD-STDP

PPO TD-STDP는 GAE와 policy-ratio 기반 안정화 원리를 STDP 시냅스 갱신을 조절하는 도파민 학습 신호로 변환한다[12,13]. Actor는 현재 상태 $s_t$에서 $K$번의 SNN simulation step을 실행하고, 각 action neuron의 spike count $n_{a,t}$를 softmax 함수에 적용하여 행동 확률 $\pi_t^{SNN}(a \mid s_t)$에 따라 행동 $a_t$를 선택한다.

학습 중에는 특정 행동으로 정책이 조기에 고정되는

것을 완화하기 위해 수식 7과 같이 무작위 정책을 혼합한 behavior policy $\pi_t^B(a \mid s_t)$를 사용한다. 여기서 $\tau$는 행동 확률의 민감도를 조절하는 policy temperature, $\bar{a}$은 $a$와 반대되는 action neuron, $\epsilon_e$는 해당 에피소드에서의 탐험 확률, $|A|$는 전체 행동의 수이다.

$$\pi_t^{SNN}(a \mid s_t) = \frac{1}{1+\exp((n_{\bar{a},t} - n_{a,t})/\tau)} \quad (7)$$

$$\pi_t^B(a \mid s_t) = (1-\epsilon_e)\pi_t^{SNN}(a \mid s_t) + \epsilon_e \frac{1}{|A|}$$

PPO TD-STDP는 환경 step마다 상태 전이 정보 $z_t$를 $T_e$까지 rollout $H_e$으로 저장한다. Episode가 종료되면 수식 8과 같이 $H_e$의 TD-error $\delta_t$를 계산하고 $T_e$부터 역순으로 GAE $A_t^{GAE}$을 계산한다. 여기서 $\lambda_G$는 이후 TD-error의 반영 범위를 조절하는 계수, $\pi_{old}^B(a_t \mid s_t)$는 rollout 수집 당시 저장한 행동 $a_t$의 선택 확률이다.

$$z_t = (s_t, a_t, r_t, s_{t+1}, e_{ia_t,t}, \pi_{old}^B(a_t \mid s_t)) \quad (8)$$

$$H_e = (z_1, z_2, ..., z_{T_e})$$

$$\delta_t = \begin{cases} r_t - V(s_t), & t = T_e \\ r_t + \gamma V(s_{t+1}) - V(s_t), & t < T_e \end{cases}$$

$$A_t^{GAE} = \begin{cases} \delta_t, & t = T_e \\ \delta_t + \gamma\lambda_G A_{t+1}^{GAE}, & t < T_e \end{cases}$$

Critic 학습에는 GAE advantage를 이용하여 상태 가치 테이블을 갱신하며, 이는 수식 9와 같다. Actor-Critic TD-STDP가 1-step TD-error로 상태 가치를 즉시 갱신하는 것과 달리, PPO TD-STDP는 episode 종료마다 GAE advantage와 critic 학습률 $\eta_c$을 이용하여 critic을 갱신한다.

$$V(s_t) \leftarrow V(s_t) + \eta_c A_t^{GAE} \quad (9)$$

Actor는 GAE advantage와 행동 정책의 변화량을 이용하여 학습한다. 특정 step의 큰 GAE advantage가 시냅스 갱신을 과도하게 지배하지 않도록 수식 10과 같이 rollout 내 GAE advantage를 정규화 및 $[c_{\min}^{GAE}, c_{\max}^{GAE}]$ 범위로 제한한다. 여기서 $\mu_A$와 $\sigma_A$는 각각 GAE advantage의 평균과 표준편차, $\varepsilon_n$는 0으로 나누는 것을 방지하는 상수이다.

$$\widetilde{A}_t = \frac{A_t^{GAE} - \mu_A}{\sigma_A + \varepsilon_n} \quad (10)$$

$$\overline{A}_t = \begin{cases} c_{\min}^{GAE}, & \widetilde{A}_t < c_{\min}^{GAE} \\ \widetilde{A}_t, & c_{\min}^{GAE} \le \widetilde{A}_t \le c_{\max}^{GAE} \\ c_{\max}^{GAE}, & \widetilde{A}_t > c_{\max}^{GAE} \end{cases}$$

이후 $H_e$에 저장된 $z_t$의 인덱스를 무작위로 배열하고, 각 전이의 정책의 행동 선택 확률 $\pi_{\theta_{old}}^B(a_t \mid s_t)$현재 actor의 정책의 행동 선택 확률 $\pi_{\theta_{new}}^B(a_t \mid s_t)$제한 범위 $[c_{\min}^{PPO}, c_{\max}^{PPO}]$를 통해 PPO TD-STDP의 도파민 학습 신호 $D_t^{PPO}$는 수식 11과 같이 계산된다. 계산된 $D_t^{PPO}$는 해당 전이의 $e_{ia_t,t}$와 결합되어 시냅스 가중치에 즉시 반영되고 다음 전이의 $\pi_{\theta_{new}}^B(a_t \mid s_t)$계산에 적용된다. PPO TD-STDP의 actor는 critic 학습과 같이 episode 종료마다 $D_t^{PPO}$를 통해 $T_e$회의 가중치 갱신을 시도한다.

$$\rho_t = \frac{\pi_{\theta_{new}}^B(a_t \mid s_t)}{\pi_{\theta_{old}}^B(a_t \mid s_t)} \quad (11)$$

$$D_t^{PPO} = \begin{cases} 0, & (\overline{A}_t \ge 0, \rho_t > c_{\max}^{PPO}) \\ 0, & (\overline{A}_t < 0, \rho_t < c_{\min}^{PPO}) \\ \rho_t \overline{A}_t, & otherwise \end{cases}$$

## 4. 실험 결과 및 분석

### 4.1 학습 설정 및 평가 지표

본 연구에서는 Actor-Critic TD-STDP와 PPO TD-STDP를 동일한 조건에서 비교하기 위해 spike encoding 방식, SNN 구조 및 전체 학습 episode 수를 동일하게 설정하였다. 두 알고리즘의 공통 실험 조건은 표 1과 같으며, 알고리즘 간 성능 차이가 공통 SNN 구조가 아니라 도파민 학습 신호, 행동 선택 및 정책 갱신 방식의 차이에서 나타나도록 하였다. 이러한 알고리즘의 세부 학습 설정은 각각 표 2, 표 3과 같다.

(표 1) 공통 실험 조건
(Table 1) Common Experimental Conditions

| 설정 항목 | 기호 | 공통 설정값 |
|---|---|---|
| 학습 episode 수 | - | 4,000 |
| 기준 random seed | - | 0 |
| 입력 spike 확률 범위 | - | [0.0, 0.5] |
| LIF threshold | - | 0.85 |
| Episode당 최대 step | $T_{\max}$ | 500 |
| SNN simulation step | $K$ | 40 |
| LTP, LTD 가중치 | $A_+$, $A_-$ | (1.0, 0.5) |
| 시냅스 가중치 범위 | $[w_{\min}, w_{\max}]$ | [0.0, 1.5] |
| Actor 학습률 | $\eta_a$ | 0.0003 |
| Spike trace decay | $\beta_{tr}$ | 0.90 |
| 상태 이산화 구간 수 | $[k_x, k_{\dot{x}}, k_\theta, k_{\dot{\theta}}]$ | [8, 8, 12, 12] |
| 할인율 | $\gamma$ | 0.99 |
| Critic 학습률 | $\eta_c$ | 0.03 |

(표 2) Actor-Critic TD-STDP 설정
(Table 2) Actor-Critic TD-STDP Settings

| 설정 항목 | 기호 | 설정값 |
|---|---|---|
| 행동 선택 기준 | - | Spike count |
| Actor, Critic 갱신 시점 | - | 환경 step |
| 초기 탐험 확률 | $\epsilon_{start}$ | 0.20 |
| 최종 탐험 확률 | $\epsilon_{end}$ | 0.01 |
| 탐험 감소 기간 | $E_{decay}$ | 800 episodes |
| TD-error 범위 | $[c_{\min}^{TD}, c_{\max}^{TD}]$ | [-0.7, 0.7] |
| 탐험 보정 계수 | $m_t$ | 0.2 |

(표 3) PPO TD-STDP 설정
(Table 3) PPO TD-STDP Settings

| 설정 항목 | 기호 | 설정값 |
|---|---|---|
| 행동 선택 기준 | - | 확률적 정책 |
| Actor, Critic 갱신 시점 | - | 에피소드 종료 |
| 초기 탐험 확률 | $\epsilon_{start}$ | 0.05 |
| 최종 탐험 확률 | $\epsilon_{end}$ | 0.00 |
| 탐험 감소 기간 | $E_{decay}$ | 800 episodes |
| Policy temperature | $\tau$ | 1.5 |
| GAE 계수 | $\lambda_G$ | 0.95 |
| GAE 정규화 상수 | $\varepsilon_n$ | $1.0\times10^{-8}$ |
| GAE clipping | $[c_{\min}^{GAE}, c_{\max}^{GAE}]$ | [-2.0, 2.0] |
| Policy ratio 범위 | $[c_{\min}^{PPO}, c_{\max}^{PPO}]$ | [0.8, 1.2] |

학습 성능은 episode return, 최근 50 episode return의 이동평균 및 evaluation return을 이용하여 평가하였다. Episode return은 도파민 학습 신호 계산에 사용된 보상 $r_t$의 합이 아니라, CartPole-v1에서 획득한 원래 환경 보상의 누적값으로 정의하였다. 따라서 episode return은 pole의 균형을 유지한 environment step 수와 동일하며 최대값은 500이다. Moving average 50은 episode return의 50 에피소드 간의 평균으로 전반적인 학습 추세를 확인하기 위해 사용하였다. Evaluation return은 50 episode 간격으로 탐험과 시냅스 가중치 갱신을 비활성화한 상태에서 현재 정책을 평가하여 얻은 평균 return으로 학습 중 정책의 실제 성능을 확인하기 위한 지표로 사용하였다.

## 4.2 도파민 학습 신호에 따른 성능 해석

그림 2와 표 4는 두 가지 STDP-SNN 강화학습 방법의 CartPole-v1 학습 결과를 나타낸다. 두 방법은 동일한 SNN actor 구조를 사용하므로, 성능 차이는 도파민 유사 학습 신호의 계산 방식과 이에 결합된 행동 선택 및 정책 갱신 방식의 차이에서 나타난 것으로 해석할 수 있다.

Actor-Critic TD-STDP는 critic이 계산한 1-step TD-error를 도파민 학습 신호로 사용하므로, 선택한 행동의 결과가 현재 가치 예측보다 좋은지 또는 나쁜지를 매 환경 step마다 시냅스 갱신에 직접 반영한다[6]. 이에 따라 학습 초반에는 빠른 성능 향상이 나타났으나, 1-step TD-error가 순간적인 가치 변화에 민감하게 반응하고 우수한 정책이 형성된 이후에도 가중치 갱신이 지속되면서 학습 후반의 성능 변동성이 증가한 것으로 해석할 수 있다.

PPO TD-STDP는 이러한 불안정성을 완화하기 위해 episode rollout에서 계산한 GAE advantage를 정규화하고 제한한 뒤, policy ratio를 이용하여 기준 정책으로부터 지나치게 벗어나는 방향의 시냅스 갱신을 억제한다. Advantage는 선택한 행동이 현재 가치 기준에서 상대적으로 얼마나 유리한지를 나타내며, policy-ratio 기반 제한은 특정 전이로 정책과 시냅스 가중치가 급격하게 변화하는 것을 방지한다. 따라서 Actor-Critic TD-STDP보다 초기 학습 속도는 느리지만, 학습 후반까지 상대적으로 안정적인 성능을 유지한 것으로 해석할 수 있다.

(표 4) Actor-Critic TD-STDP와 PPO TD-STDP의 주요 성능 비교
(Table 4) Comparison of Key Performance Metrics for Actor-Critic TD-STDP and PPO TD-STDP

| 평가 지표 | Actor-Critic TD-STDP | PPO TD-STDP |
|---|---|---|
| Max episode return 최초 도달 episode | 1,399 | 2,334 |
| 최고 evaluation return | 436.60 | 384.70 |
| Episode return의 후반 500 episode 평균 | 167.73 | 404.97 |
| Evaluation return 후반 500 episode 평균 | 160.68 | 358.82 |




(그림 2) Actor-Critic TD-STDP와 PPO TD-STDP 방식의 CartPole-v1 학습 결과
(Figure 2) CartPole-v1 Learning Results for Actor-Critic TD-STDP and PPO TD-STDP Methods

### 4.3 기존 연구와의 비교 및 논의

기존 SNN 강화학습 연구는 전역적인 강화학습 신호를 국소적 시냅스 가소성과 결합하는 방법과 SNN의 구조를 개선하는 방법으로 발전해 왔다. Frémaux 등은 spiking Actor-Critic 구조에서 critic이 계산한 보상 예측 오차(reward prediction error)를 actor의 시냅스 학습에 적용하였으며[6], Yang 등은 전역 policy gradient를 보상 신호로 STDP를 조절하는 Reward-modulated STDP(R-STDP) 규칙으로 변환하였다[7]. 이러한 연구들은 학습 신호로 행동 선택에 관여한 시냅스의 가중치 변화를 조절한다는 점에서 본 연구와 공통점을 가진다.

한편 최근 연구에서는 PPO 기반 SNN의 계층 및 뉴런 구성에 따른 성능 분석이 수행되었으며[8], Chen 등은 뉴런의 막전위(membrane potential)를 행동 출력에 활용하고 출력 뉴런 사이의 내부 연결을 추가한 fully spiking actor를 제안하였다[9]. 또한 뉴런 자체의 발화 특성을 적응적으로 조절하는 intrinsic plasticity coding을 적용한 방법[10]과, 하나의 뉴런을 여러 내부 구획으로 구성하는 multi-compartment neuron 및 여러 뉴런의 발화 분포로 상태를 표현하는 population encoding을 결합한 방법도 제안되었다[11]. 이들 연구는 SNN의 구조, spike encoding, 행동 출력 방식이 강화학습 성능에 영향을 미칠 수 있음을 보여준다. 하지만 이러한 연구들은 네트워크 구조와 학습 알고리즘이 함께 변경되는 경우가 많으므로, 시냅스 갱신을 조절하는 학습 신호 자체의 영향을 독립적으로 비교하기에는 한계가 있다.

본 연구는 동일한 spike encoding 방식과 SNN 구조를 유지한 상태에서 1-step TD-error와 episode 단위 GAE advantage를 각각 도파민 유사 학습 신호로 적용하여 학습 방식의 차이를 비교하였다는 점에서 기존 연구와 구분된다. 특히 PPO TD-STDP는 일반적인 PPO 목적함수를 역전파로 최적화하는 대신, GAE advantage의 정규화 및 제한과 policy ratio 기반 갱신 제한을 STDP eligibility에 결합되는 학습 신호로 변환하였다[12,13]. 따라서 본 연구는 SNN 구조의 개선보다 도파민 유사 학습 신호의 시간적 반영 범위와 정책 변화 제한 방식이 STDP-SNN의 학습 특성에 미치는 영향을 분리하여 분석하였다는 의의를 가진다.

## 5. 결론 및 향후 연구과제

본 연구에서는 CartPole-v1 환경에서 동일한 STDP 기반 SNN을 사용하고, 시냅스 갱신을 조절하는 도파민 유사 학습 신호의 설계에 따라 Actor-Critic TD-STDP와 PPO TD-STDP의 학습 특성을 비교하였다. Actor-Critic TD-STDP는 매 환경 step에서 계산한 1-step TD-error를 도파민 유사 학습 신호로 사용하였으며, PPO TD-STDP는 episode 단위로 계산한 GAE advantage의 정규화 및 제한과 policy ratio 기반 갱신 제한을 적용하였다. 두 방법 모두 일반적인 인공신경망 actor의 역전파를 사용하지 않고, STDP 기반 국소적 시냅스 가소성과 critic 기반 전역 학습 신호를 결합하여 행동 정책을 학습하도록 구성하였다.

실험 결과, Actor-Critic TD-STDP는 PPO TD-STDP보다 episode return 500점에 더 이르게 도달하고 더 높은 최고 evaluation return을 기록하여, 학습 초반의 빠른 정책 개선에 강점을 보였다. 하지만 후반에는 지속적인 TD-error 기반 시냅스 갱신으로 인해 evaluation return이 감소하고 성능 변동성이 증가하였다. 반면 PPO TD-STDP는 초기 학습 속도는 상대적으로 느렸지만, GAE advantage와 policy ratio를 이용하여 과도한 시냅스 갱신과 급격한 정책 변화를 제한함으로써 학습 후반까지 높은 성능을 유지하였다. 이러한 결과는 본 단일 random seed 사례에서 STDP-SNN 강화학습의 학습 양상이 SNN 구조뿐 아니라 도파민 유사 학습 신호의 시간적 반영 범위와 정책 변화 제한 방식에 따라서도 달라질 수 있음을 시사한다.

본 연구의 학술적 의의는 동일한 SNN 구조를 유지한 상태에서 도파민 유사 학습 신호와 이에 결합된 행동 선택 및 정책 안정화 방식을 달리하여 학습 속도와 안정성의 차이를 비교하였다는 점에 있다. 특히 1-step TD-error 기반 신호는 빠른 성능 향상에 유리한 반면, episode 단위 GAE advantage와 policy ratio 기반 제한이 학습 후반의 정책 안정성에 유리한 경향을 보임을 확인하였다. 이는 향후 SNN 강화학습 알고리즘에서 전역 학습 신호의 계산 범위와 시냅스 갱신 크기를 함께 고려해야 한다는 설계 방향을 제시한다.

본 연구는 단일 random seed, 은닉층이 없는 소규모 SNN actor, 그리고 CartPole-v1 단일 환경을 사용했다는 한계가 있다. 후속 연구에서는 다양한 random seed와 하이퍼파라미터 설정 및 보다 복잡한 제어 환경을 적용하여 연구 결과의 일반성을 검증할 필요가 있다. 또한 별도의 critic과 상태 가치 추정 없이 SNN actor만으로 정책을 학습하는 강화학습 구조를 설계하고, 이를 critic 기반 방법과 비교하여 학습 성능, 안정성 및 계산 복잡도를 분석할 필요가 있다. 이러한 연구는 생물학적 학습 원리에 더욱 부합하는 SNN 강화학습 구조를 설계하는 데 기여할 수 있을 것으로 기대된다.

## 참고문헌(Reference)


[1] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, “A neuronal learning rule for sub-millisecond temporal coding,” Nature, Vol. 383, No. 6595, pp. 76-78, 1996. https://doi.org/10.1038/383076a0

[2] G.-Q. Bi and M.-M. Poo, “Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type,” Journal of Neuroscience, Vol. 18, No. 24, pp. 10464-10472, 1998. https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998

[3] R. Legenstein, D. Pecevski, and W. Maass, “A Learning Theory for Reward-Modulated Spike-Timing-Dependent Plasticity with Application to Biofeedback,” PLOS Computational Biology, Vol. 4, No. 10, e1000180, 2008. https://doi.org/10.1371/journal.pcbi.1000180

[4] N. Frémaux, H. Sprekeler, and W. Gerstner, “Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity,” Journal of Neuroscience, Vol. 30, No. 40, pp. 13326-13337, 2010. https://doi.org/10.1523/JNEUROSCI.6249-09.2010

[5] W. Schultz, P. Dayan, and P. R. Montague, “A Neural Substrate of Prediction and Reward,” Science, Vol. 275, No. 5306, pp. 1593-1599, 1997. https://doi.org/10.1126/science.275.5306.1593

[6] N. Frémaux, H. Sprekeler, and W. Gerstner, “Reinforcement Learning Using a Continuous Time Actor-Critic Framework with Spiking Neurons,” PLOS Computational Biology, Vol. 9, No. 4, e1003024, 2013. https://doi.org/10.1371/journal.pcbi.1003024

[7] Z. Yang, S. Guo, Y. Fang, Z. Yu, and J. K. Liu, “Spiking Variational Policy Gradient for Brain Inspired Reinforcement Learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 47, No. 3, pp. 1975-1990, 2025. https://doi.org/10.1109/TPAMI.2024.3511936

[8] L. Zanatta, F. Barchi, S. Manoni, S. Tolu, A. Bartolini, and A. Acquaviva, “Exploring spiking neural networks for deep reinforcement learning in robotic tasks,” Scientific Reports, Vol. 14, Article No. 30648, 2024. https://doi.org/10.1038/s41598-024-77779-8

[9] D. Chen, P. Peng, T. Huang, and Y. Tian, “Fully Spiking Actor Network With Intralayer Connections for Reinforcement Learning,” IEEE Transactions on Neural Networks and Learning Systems, Vol. 36, No. 2, pp. 2881-2893, 2025. https://doi.org/10.1109/TNNLS.2024.3352653

[10] X. Liang, Q. Wu, W. Liu, Y. Zhou, C. Tan, H. Yin, and C. Sun, “Intrinsic plasticity coding improved spiking actor network for reinforcement learning,” Neural Networks, Vol. 184, Article No. 107054, 2025. https://doi.org/10.1016/j.neunet.2024.107054

[11] Y. Sun, F. Zhao, Z. Zhao, and Y. Zeng, “Multi-compartment neuron and population encoding powered spiking neural network for deep distributional reinforcement learning,” Neural Networks, Vol. 182, Article No. 106898, 2025. https://doi.org/10.1016/j.neunet.2024.106898

[12] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, “High-Dimensional Continuous Control Using Generalized Advantage Estimation,” in Proc. of International Conference on Learning Representations, 2016. https://doi.org/10.48550/arXiv.1506.02438

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” arXiv preprint arXiv:1707.06347, 2017. https://doi.org/10.48550/arXiv.1707.06347

## ◐ 저 자 소 개 ◑



**문 종 인** (Jongin Mun)
2025년 홍익대학교 소프트웨어융합학과(학사)
2025년～현재 홍익대학교 소프트웨어융합학과(석사)
관심분야 : 딥러닝, 데이터분석
E-mail : whddls5206@naver.com



**노 기 섭**(Giseop Noh)
1998년 공군사관학교 산업공학과(공학사)
2009년 콜로라도대학교 대학원 컴퓨터학과(공학석사)
2014년 서울대학교 대학원 컴퓨터공학과(공학박사)
2016년 12월～2018년 02월 공군사관학교 전산학과 조교수
2018년 03월～2025년 02월 청주대학교 인공지능SW학과 조교수
2025년 03월～현재 홍익대학교 SW융합학과 조교수
관심분야 : 자연어처리, 신경망 이론, 강화학습
E-mail : kafa46@hongik.ac.kr