Journal of Korean Institute of Intelligent Systems
Vol. 28, No. 5, October 2018, pp. 481–486
http://dx.doi.org/10.5391/JKIIS.2018.28.5.481

ISSN(Print) 1976–9172
ISSN(Online) 2288–2324

JKIIS

확률적 시뮬레이션을 이용한 모델 기반 강화학습

Model-based reinforcement learning using probabilistic simulation

주하람 · 김준오 · 이상완†
Haram Joo, Juno Kim and Sang Wan Lee†

카이스트 바이오및뇌공학과
Department of Bio&Brain Engineering, KAIST

요 약

본 논문은 상태천이에 불확실성이 있는 동적 환경에서도 안정적인 학습이 가능한 model-based 강화학습 전략을 제안한다. 기존의 강화학습 알고리즘은 보상의 기대치 최대화에 초점을 둔 model-free 방식으로 환경의 불확실성을 경험적으로 습득하므로 적응 속도가 느리다. 이에 비해 환경 모델을 학습하는 model- based 방식은 아직 경험하지 못한 상황에 대한 시뮬레이션 결과를 보상의 기대치 학습에 적용함으로써 환경변화에 빠른 적응이 가능하다. 본 연구에서는 환경의 상태천이에 대한 확률 모델을 온라인 학습하고, 학습된 모델을 이용하여 확률적으로 시나리오를 시뮬레이션하며, 이를 바탕으로 보상의 기대치를 최대화하는 전략을 찾아내는 model-based 강화학습 방식을 구현하였다. OpenAI의 FrozenLake 시뮬레이터를 이용하여 불확실성을 내포한 동적 환경을 구현하였으며, 제안한 모델과 기존 방법의 성능을 다양한 측면에서 비교하였다. 제안된 모델은 상태천이의 불확실성과 환경변화의 불안정성이 모두 존재하는 극한 상황 속에서도 변화에 강인한 전략 탐색의 기틀을 제공한다.

키워드 : 상태천이 불확실성, 모델기반 강화학습, FrozenLake.

Abstract

This paper proposes a model-based reinforcement learning strategy that enables stable learning even in a dynamic environment containing state transition uncertainty. The existing reinforcement learning algorithm is a model-free method that focuses on maximizing the expectation of the reward, and the adaptation speed is slow because it empirically learns the uncertainty of the environment. In contrast, the model-based method that learns the environmental model can adapt quickly to changes in the environment by applying the simulation results to the expectation reward. In this paper, we propose a model-based reinforcement learning method that finds a strategy that maximizes the expectation of reward based on the on-line learning of the probability transition model of the environment, simulates the scenario probabilistically using the learned model. We implemented the dynamic environment containing uncertainty using FrozenLake simulator of OpenAI and compared the performance of the proposed model with the existing method in various aspects. The proposed model provides a framework for strategy exploration even in extreme situations where both uncertainty of state transition and instability of environmental change exist.

Key Words : State Transition Uncertainty, Model-based Reinforcement Learning, FrozenLake.

Received: Apr. 3, 2018
Revised: Aug. 13, 2018
Accepted: Aug. 13, 2018
†Corresponding authors
sangwan@kaist.ac.kr

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신 · 방송 연구개발사업의 일환으로 수행하였음.[2016-0-00563, 자율지능 동반자를 위한 적응형 기계학습기술 연구개발]
이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1C 1B 2008972).
이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No. 2017-0- 00451, 딥러닝을 이용하여 사람의 의도를 인지하는 BCI 기반 뇌인지컴퓨팅 기술 개발).

1. 서 론

강화학습은 개체가 환경과의 상호 작용을 통해 경험한 보상을 바탕으로 목표를 달성하는 과정을 배우는 방법이다. 이러한 강화학습 기술은 즉각적인 결과와 지연된 결과를 모두 고려할 수 있는 최적의 장기 행동 선택을 위한 해법 조합이다. 가장 보편적인 강화학습 방법은 시간차(Temporal difference, TD) 학습 방식이다.

시간차 학습 방식은 실제 경험한 보상과 이전 단계에서 예측한 보상의 차이를 학습에 이용하는 증분학습방법이다[1]. 시간차 학습 방식은 시간적인 연속성 하에서 예측의 오차에 의해 유도되며, 학습은 예측에 변화가 있을 때마다 이루어지게 된다. 특히 시간차 학습 방식은 점진적인 특성이 있어 계산이 용이하고 경험을 효율적으로 활용할 수 있다. 가장 최초로 시간차 학습 방식을 활용한 연구는 Samuel[2]에 의해 이루어졌으며 그 후 이와 유사한 방법의 연구들이 Holland[3], Moore, et al.[4] 등에 의해 진행되었다. 국내에서 진행된 연구들은 다소 제한적인 측면이 있지만, 확률론적 강화학습 방식에 관한 연구들이 진행되었다[5,6].

본 연구에서는 시간차 학습 방식의 한 종류인 Q- Learning 알고리즘을 사용했다. 어떤 유한 마르코프 결정 프로세스(MDP)에 대해서도 Q-Learning 알고리즘은 궁극적으로 모든 연속 단계에 대한 보상의 기대치를 최대화할 수 있다[7,8]. Q-Learning 알고리즘은 주로 model-free 강화학습 방식에 사용된다.

Model-free 강화학습 방식과 model-based 강화학습 방식 모두 주된 목표는 보상의 기대치를 최대화하기 위해서 행동 선택 정책을 개선하는 것이다. Model-free 강화학습 방식은 환경 모델이나 행동의 결과에 대한 명시적 지식을 피하고 시행착오적 학습을 통해 좋은 행동의 정도를 평가한다. Model-free 강화학습 방식은 특정 자극에 직면했을 때 반사적으로 표출되는 습관적 및 파블로브(Pavlovian)의 조건부 반응에 기초하며 광범위한 경험이 요구된다[9].

그러나 model-based 강화학습 방식은 agent가 환경을 학습하려고 시도한다는 점에서 model-free 강화학습 방식과는 차이가 있다. 이러한 환경 모델의 학습을 통해 미래의 결과들을 미리 예측하여 가상 경험하고, 나아가 적절한 행동 sequence를 얻을 수 있다. 따라서 model-based 강화학습 방식은 과거 경험에 기초한 학습뿐만 아니라 미래의 행동 선택에 따른 결과 예측도 포함한다[10].

Q-Learning 알고리즘을 기반으로 하여 환경 모델에 대한 학습이 가능하도록 한 것이 dyna-Q 알고리즘이다. 다시 말해, dyna-Q 알고리즘은 model-based 강화학습 방식에 속한다.

본 연구에서는 확률적인 상태천이를 가지는 동적 환경에서 model-free 강화학습 방식과 model-based 강화학습 방식의 성능을 비교해 보고자 한다. 나아가, 상태천이확률이 변하는 불안정성을 가지는 환경에 대해서도 확장해 본다. 이를 통해 상태천이의 불확실성과 환경변화의 불안정성이 모두 존재하는 상황에서 효율적인 학습 방식의 기틀을 제공한다.

2. 확률적인 상태천이를 가진 동적 환경

본 연구에서는 확률적인 상태천이를 가지는 동적 환경으로 OpenAI의 FrozenLake를 이용하였다. FrozenLake 환경은 2차원 격자 형태로, 각각의 칸은 시작점(S), 빙판(F), 구멍(H), 목적지(G)의 네 가지 중 하나의 상태를 가진다.

빙판(F)은 안전한, 다시 말해 목적지로의 여정을 계속 진행할 수 있는 칸을 의미한다. 반면에, 구멍(H)에 빠지게 되면 해당 게임은 종료되고, 보상도 받지 못한다. 각각의 칸에서 취할 수 있는 행동은 상하좌우 중 하나의 방향으로 움직이고자 하는 것이다. 하지만, 빙판인 만큼 항상 우리가 원하는 방향으로 움직이지는 않는다. 다시 말해, 미끄러짐이 발생할 수 있다. 미끄러짐은 원래 움직이고자 했던 방향과 인접한 방향으로만 발생하고, 완벽하게 대응되는 방향으로 미끄러지지는 않는다. 예를 들어 동(오른)쪽으로 움직이고자 했을 때, 북(위)쪽이나 남(아래)쪽으로 미끄러짐이 발생할 수도 있지만, 서(왼)쪽으로 미끄러지는 경우는 없다.

이러한 확률적인 상태천이를 가지는 환경에서 시작점에서 출발하여 목적지에 도달하는 것이 목표이다. 그림 1은 FrozenLake 환경의 예이다.

S	F	F	F
F	H	F	H
F	F	F	H
H	F	F	G

그림 1. FrozenLake 환경 예시

Fig. 1. Example of FrozenLake environment

3. Model-free 강화학습

본 연구의 model-free 강화학습 방식은 Q-Learning 알고리즘을 기반으로 한다[11].

$\forall\ (s,a)$에 대해 $Q(s,a) \leftarrow 0$

반복:

(a) $s \leftarrow$ 현재 상태

(b) $a \leftarrow \epsilon\text{-}greedy(s,Q)$

(c) a를 실행하고, 보상 r과 결과 상태 s'을 얻는다.

(d) $Q(s,a)$를 다음과 같이 갱신한다:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha[r+\gamma \max_{a'} Q(s',a')]$$

과정 (b)에서 행동을 선택할 때 ϵ-greedy 방법을 사용한다. Explore 시에는 random한 행동을 선택하며, Exploit 시에는 $Q(s,a)$값을 최대로 만드는 행동 a를 선택한다.

Model-free 강화학습 방식의 주요 파라미터 초기 값은 표 1과 같다.

표 1. Model-free 강화학습 방식 주요 파라미터 초기 값

Table 1. Initial value of model-free RL parameters

Parameter	Meaning	Initial Value
ALPHA	learning rate of Q-table	0.1
GAMMA	reward discount	0.99
EPSILON	related to ϵ-greedy algorithm	0.2
TURN_LIMIT	if it exceeds turn_limit, episode ends and don't get any reward	100

4. Model-based 강화학습

본 연구의 model-based 강화학습 방식은 (a)~(f) 과정으로 이루어지는 dyna-Q 알고리즘[10]을 기반으로 한다.

$\forall\ (s,a)$에 대해 $Q(s,a) \leftarrow 0$
반복:
(a) $s \leftarrow$ 현재 상태
(b) $a \leftarrow \epsilon\text{-}greedy(s,Q)$
(c) a를 실행하고, 보상 r과 결과 상태 s'을 얻는다.
(d) $Q(s,a)$를 다음과 같이 갱신한다:
$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha[r+\gamma \max_{a'} Q(s',a')]$
(e) s, s', a를 이용하여 $T(s,a,a')$ 갱신
(f) 반복 10회:
$s_2 \leftarrow$ 기존에 방문한 상태 중 랜덤 선택
$a \leftarrow s_2$에서 실행된 행동 중 랜덤 선택
예측 보상 r과 예측 상태 s_2'을 모델로부터 얻는다.
$Q(s_2,a)$를 갱신한다:
$Q(s_2,a) \leftarrow (1-\alpha)Q(s_2,a)+\alpha[r+\gamma \max_{a'} Q(s_2',a')]$

(a)~(d)까지의 과정은 model-free 강화학습 방식과 유사하다. 다만, ϵ-greedy(과정 (b))의 Exploit 시에 $Q(s,a)$값을 최대로 만드는 행동 a를 선택하는 model- free 강화학습 방식과 달리, model-based 강화학습 방식에서는 상태천이행렬(state_transition-table, T)을 활용하여 행동 a를 선택한다. 다시 말해, 다음의 값을 최대로 만드는 행동 a를 선택한다. (0부터 3까지의 값은 행동의 방향을 의미한다. 순서대로 서남동북)

$$\sum_{i=0}^{3} Q(s,i)\,T(s,a,i)$$

$T(s,a,i)$는 상태 s에서 행동 a를 선택하였을 때, 행동 i가 실행될 확률의 예측(학습) 값이다. 행렬 T는 본 연구의 model-based 강화학습 방식에서 환경 모델에 대한 학습 정보를 가지고 있는 유일한 변수이다. 환경 모델에 대한 학습, 다시 말해 상태천이행렬 값의 갱신은 과정 (e)를 통해 일어난다. 먼저 상태천이행렬의 초기 값은 다음과 같다.

$$\forall\ (s,a,i),\ T(s,a,i) = \begin{cases} 0.5 & \text{if } a = i \\ 0.25 & \text{if } abs(a-i) = 1 \\ 0 & \text{if } abs(a-i) = 2 \end{cases}$$

상태 s는 각각의 칸에 대응되는 0부터 N×N-1에 해당하는 숫자를 가질 수 있다. 행동 a와 행동 i는 서남동북에 대응되는 0부터 3까지의 숫자를 가질 수 있다. 행동 a와 행동 i가 같은 경우, 즉 미끄러짐 없이 우리가 원하는 방향으로 진행할 확률의 초기 값으로 0.5를 두었다. 또한 미끄러짐이 발생해도 원하는 방향과 완벽하게 대응되는 방향으로 미끄러지지는 않으므로, 이 경우들에 대한 예측 값은 0이다. (abs() 함수는 절댓값을 의미한다.)

상태천이행렬의 갱신은 다음과 같은 방법을 통해 이루어진다.

def $A(s_p, s_c, i_p)$:
/* s_p(이전 상태), s_c(현재 상태) 및 i_p(선택된 행동)을 바탕으로 값을 리턴.
만약, s_p에서 i_p를 통해 미끄러짐 없이 s_c에 도달할 수 있다면 1을 리턴.
그렇지 않으면, 0을 리턴. */

(e) s, s', a를 이용하여 $T(s,a,a')$ 갱신
$0 \le \forall\ i \le 3$에 대해,
$T(s,a,i) \leftarrow (1-\beta)T(s,a,i)+\beta A(s,s',i)$

Model-based 강화학습 방식의 주요 파라미터 초기 값을 정리하면 표 2와 같다.

표 2. Model-based 강화학습 방식 주요 파라미터 초기 값
Table 2. Initial value of model-based RL parameters

Parameter	Meaning	Initial Value
ALPHA	learning rate of Q-table	0.1
BETA	learning rate of state_transition table	0.05
GAMMA	reward discount	0.99
EPSILON	related to ϵ-greedy algorithm	0.2
TURN_LIMIT	if it exceeds turn_limit, episode ends and don't get any reward	100

표 1, 표 2에서 알 수 있듯이 model-free 강화학습 방식과 model-based 강화학습 방식의 파라미터 대부분은 초기 값이 동일하다. 이는 상태천이행렬을 통한 환경 모델 학습 이외의 변수들을 줄이기 위함이다.

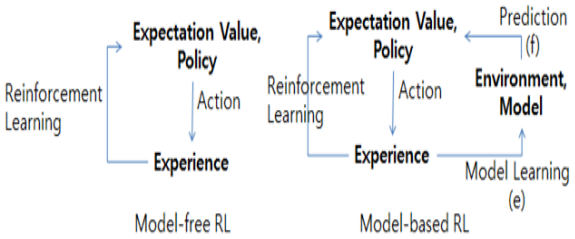

그림 2. Model-free와 Model-based 강화학습 방식
Fig. 2. Model-free RL vs Model-based RL

본 연구의 model-free 강화학습 방식과 model-based 강화학습 방식을 정리하면 그림 2[10]와 같다. 오른쪽의 model-based 강화학습 방식에서 (e)와 (f)는 dyna-Q 알고리즘의 해당 과정을 의미한다.

5. 시뮬레이션 결과 및 분석

표 3에서 "# of episodes for first_reward"는 처음 보상을 얻게 되는 episode의 번호를 의미한다. "Average reward after 1000(5000) episodes"는 각각 1000(5000)개의 episodes를 학습한 후의 누적평균보상을 의미한다. 누적평균보상은 지금까지 얻은 보상의 총합을 경험한 episodes의 수로 나눈 값이다. "Average reward for test"는 5000개의 episodes로 학습이 완료된 이후 1000번의 테스트를 통해 나오는 평균보상이다. 테스트 시에는 환경 모델의 추가학습이나 새로운 행동(경로)을 탐험하지 않고, 기존의 table값을 바탕으로 최적의 행동을 선택한다(Explore 없이 Exploit만 한다).

표 3. 주요 성능 비교 지표(괄호 안은 표준편차)

Table 3. Major performance indicator
(Number in parentheses means standard deviation)

	Model-free	Model-based
# of episodes for first_reward	32 (16)	6 (3)
Average reward after 1000 episodes	0.23 (0.02)	0.41 (0.01)
Average reward after 5000 episodes	0.31 (0.01)	0.58 (0.01)
Average reward for test	0.73 (0.03)	0.74 (0.01)

표 3에서 볼 수 있듯이 주요 성능 비교 지표들에서 model-based 강화학습 방식이 model-free 강화학습 방식보다 좋은 성능을 보였다. 이는 상태천이행렬을 활용한 환경 모델의 학습을 통해 보다 정확한 보상의 기대치를 구할 수 있었기 때문으로 생각된다. 그림 3은 학습 episode 수에 따른 누적평균보상을 나타낸다.

나아가, map의 구성(시작점, 목적지, 구멍의 위치 등)은 일정하지만, 환경의 불안정성이 있는 상황에서 누적평균보상을 구해 보았다. 본 연구에서 "환경의 불안정성"이란 상태천이확률이 변함을 의미한다.

그림 4-A는 매 100의 배수 번째 episode가 끝나면 상태천이확률이 변하는 상황에서 누적평균보상을 구한 것이다. 그림 4-B는 매 1000의 배수 번째 episode가 끝나면 상태천이확률이 변하는 상황에서 누적평균보상을 구한 것이다. 그림 4에 나타나듯이 model-based 강화학습 방식의 순간 보상 감소폭은 더 크지만, 누적평균보상은 여전히 model-free 강화학습 방식보다 크다.

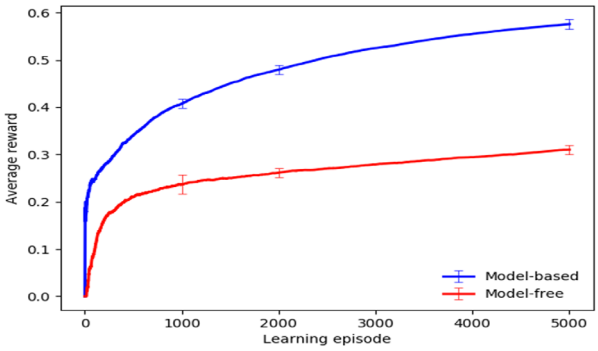

그림 3. 학습 episode 수에 따른 누적평균보상

Fig. 3. Average reward based on number of learning episodes

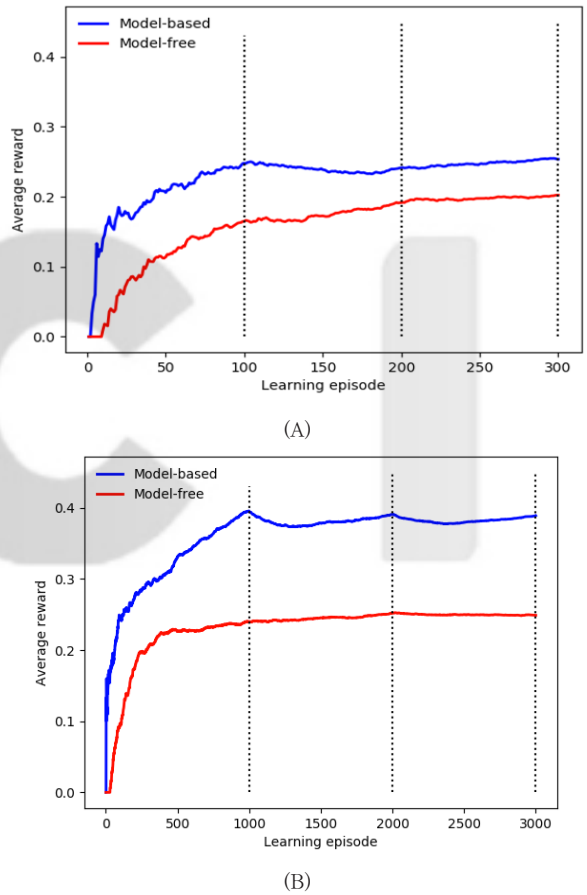

그림 4. 환경(상태천이확률)이 변화하는 상황에서 episode 수에 따른 누적평균보상

Fig. 4. In a situation where the environment(state_transition probability) changes, average reward based on episodes

6. 결론 및 향후 연구과제

상태천이의 불확실성이 있는 동적 환경에서 model- free 강화학습

방식보다 환경 모델을 학습하고 이를 바탕으로 보상의 기대치를 최대화하는 model-based 강화학습 방식이 더 나은 성능을 보임을 확인할 수 있었다. 또한 상태천이확률이 변하는 불안정한 환경에서도 model- based 강화학습 방식의 누적평균보상이 model-free 강화학습 방식의 누적평균보상보다 크다.

상태천이확률이 변하는 환경에서 model-based 강화학습 방식의 순간 감소폭을 줄이기 위한 방법에 대한 연구가 필요할 것으로 보인다. Model-based 강화학습의 순간 보상 감소폭이 큰 현상은 환경 구조에 대한 model-based 강화학습의 강한 가정으로 설명할 수 있다. 더 나아가 model-based 강화학습의 적응능력을 넘어선 지극히 불안정한 동적환경 시나리오에서는 오히려 uniform 상태천이확률분포를 가정하는 model-free 강화학습이 더욱 안정적일 것으로 예상된다. 향후 연구에서는 환경의 동적 변화의 안정성을 추정하여 model-free와 model-based 강화학습을 유동적으로 제어할 수 있는 모델이 필요하다.

References

[1] R. S. Sutton, "Learning to predict by the methods of temporal differences." *Machine Learning*, 3, pp. 9-44, 1988.

[2] A. Samuel, "Some studies in machine learning using the game of checkers." *IBM Journal of Research and Development*, 3(3), pp. 210-229, 1959.

[3] J. H. Holland, "Escaping brittleness: The possibilities of general purpose learning algorithms applied to parallel rule-based systems." *Machine learning: An artificial intelligence approach*, 2, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. Los Altos, CA: Morgan Kaufmann, pp. 593-623, 1986.

[4] J. W. Moore, J. E. Desmond, N. E. Berthier, D. E. Blazis, R. S. Sutton, and A. G. Barto, "Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: response topography, neuronal firing, and interstimulus intervals." *Behavioral Brain Research*, 21(2), pp. 143-154, 1986.

[5] J. Y. Park, S. H. Ji, K. H. Sung, S. M. Heo, and K. W. Park, "Investigations on data-driven stochastic optimal control and approximate-inference-based reinforcement learning methods." *Journal of The Korean Institute of Intelligent Systems*, 25(4), pp. 319-326, 2015.

[6] J. G. Kim and H. S. Lee, "Multi-agent Reinforcement Learning based Evacuation Framework Considering Both Evacuation Time and Crowdedness." *Journal of The Korean Institute of Intelligent Systems*, 27(4), pp. 334-341, 2017.

[7] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms." *Neural Computation*, 6(6), pp. 1185-1201, 1994.

[8] C. Ribeiro and C. Szepesvári, "Q-learning combined with spreading: Convergence and results." *Proceedings of the ISRF-IEE International Conference: Intelligent and Cognitive Systems (Neural Networks Symposium)*, pp. 32-36, 1996.

[9] Q. J. M. Huys, A. Cruickshank, and P. Seriès, "Reward-Based Learning, Model-Based and Model-Free." *Encyclopedia of Computational Neuroscience*, D. Jaeger and R. Jung, Eds. Springer, 2014.

[10] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction." Cambridge, MA: MIT Press, pp. 227-241, 1998.

[11] T. M. Mitchell, "Machine Learning." Redmond, WA: McGraw-Hill Science/Engineering/Math, pp. 373-383, 1997.

저 자 소 개

주하람(Haram Joo)

2018년 : KAIST 전산학부(바이오및뇌공학 복수전공) 학사

2018년~현재 : KAIST 바이오및뇌공학과 석사과정

관심분야 : Brain-inspired AI, Computational Neuroscience

Phone : +82-42-350-4334

E-mail : haramjoo@kaist.ac.kr

김준오(Juno Kim)

2017년 : KAIST 바이오및뇌공학과 학사

2017년~현재 : KAIST 바이오및뇌공학과 석박사통합과정

관심분야 : Brain-inspired AI, Computational Neuroscience

Phone : +82-42-350-4334

E-mail : bium@kaist.ac.kr

이상완(Sang Wan Lee)

2003년 : 연세대학교 전기전자 학사
2005년 : KAIST 전자전산학과 석사
2009년 : KAIST 전자전산학과 박사
2010년~2011년 : MIT 박사후연수연구원
2011년~2015년 : CALTECH 박사후연수연구원
2015년~현재 : KAIST 바이오및뇌공학과 조교수

관심분야 : Brain-inspired AI, Computational Neuroscience
Phone : +82-42-350-4334
E-mail : sangwan@kaist.ac.kr