강화학습 이론의 신경과학적 고찰

한국과학기술원 ▌이상완

1. 서 론

인공지능 분야의 기술적 진보에 힘입어 다양한 실제 문제에 적용 가능한 강화학습 알고리즘들이 개발되어 왔다. 그러나, 최신 강화학습 알고리즘으로도 완벽하게 해결하지 못하는 근본적인 이슈들이 산재해 있다. 최근 계산 뇌과학 연구에서는 어떻게 인간의 뇌가 단일 유기체로서 이러한 문제들을 쉽게 풀어내는지를 밝혀내기 시작했다. 본 논문에서는 강화학습의 신경과학과 알고리즘을 병치하는 논의 방식을 통해, 강화학습에 대한 신경과학적 연구를 이용한 강화학습 알고리즘 연구의 발전 가능성에 대해 논의한다. 첫째, 강화학습의 계산적 핵심 요소에 대한 최신 신경과학 연구 결과들을 소개함으로써 뇌 안에서 일어나는 강화학습 과정에 대해 논의한다. 둘째, 뇌의 강화학습 관련 정보처리 과정을 기능적으로 분리하여 알고리즘에 구현되어 있지 않은 뇌의 고유한 강화학습 능력에 대해 살펴본다. 더 나아가 이러한 강화학습의 신경과학과 알고리즘에 대한 논의를 종합하여, 프로세스의 복잡도와 학습 속도 간의 균형점을 찾아가는 뇌기반 강화학습 이론에 대해 소개한다. 끝으로 이러한 이론을 바탕으로 한 공학적 응용 가능성에 대해 논의한다.

2. 강화학습의 기술적 발전

강화학습 이론은 전통적인 최적제어 문제에서 출발하여 다이나믹 프로그래밍과 근사화 알고리즘 연구[1]를 거쳐 오늘날의 심층 강화학습 알고리즘에 이르기 까지 꾸준한 발전이 이루어져 왔다. 강화학습은 정확한 에러 계산 없이도 작업 목적에 맞도록 입력 시퀀스를 최적화 할 수 있다는 점에서 로보틱스, 게임, 대형 프로세스 관리 등 적용 범위가 매우 넓다 (그림1). 그러나, 거의 같은 시간대에서 입력-출력 계산이 이루어지는 일반적인 신경망의 동작 특성과는 달리, 강화학습 문제는 일반적으로 입력과 출력 사이의 시간 차이가 매우 크며, 시간 순서에 의존적인 동적 시스템을 제어해야 한다는 제약 조건이 있다. 이러한 이유로 차원이나 복잡도가 큰 문제에 적용하는데 기술적 어려움을 겪어왔다.

최근 딥러닝과 같은 인공지능 기술의 발전에 힘입어 강화학습의 적용분야가 크게 넓어지기 시작했다. 대표적인 예로, 구글의 딥마인드에서는 딥러닝과 강화학습을 결합한 심층 강화학습 알고리즘을 이용하여 현실 세계 문제의 복잡도에 근접하는 다양한 ATARI 게임 환경에서 유용성을 보인 바 있다[2]. 더 나아가 문제의 복잡도가 더욱 높은 바둑 환경에서 강화학습과 전문가 지식(기보)을 결합하여 프로기사 이상의 성능을 보였으며[3], 최근에는 같은 바둑 환경에서 기보와 같은 전문가 지식 없이 순수한 심층 강화학습 알고리즘만으로도 성능을 더욱 향상시킬 수 있음을

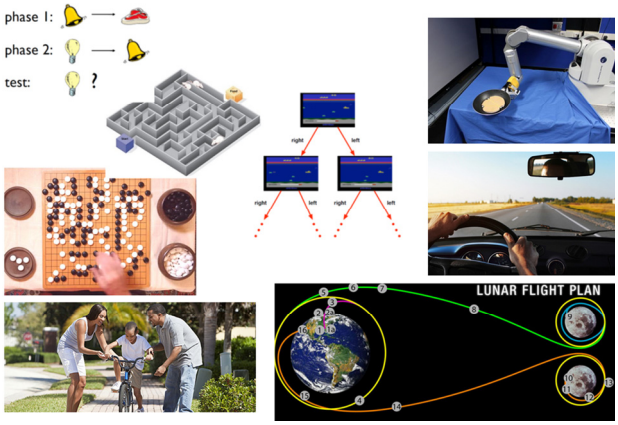


그림 1 강화학습이 적용될 수 있는 다양한 최적제어 문제들

† 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). 본 연구는 과학기술정통부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [2016-0-00563, 자율지능 동반자를 위한 적응형 기계학습 기술 연구개발] 이 논문은 KAIST의 지원을 받아 수행되었음 (과제번호 G04150045).

보였다[4]. 이러한 추세에 힘입어 계산과 구조의 복잡도를 줄이면서 더 나은 전략 탐색 성능을 가진 다양한 알고리즘들이 발표되고 있다.

3. 인간의 강화학습

이러한 발전에도 불구하고 아직 강화학습 알고리즘은 인간의 작업수행 능력과 관련된 몇가지 근본적인 문제에 대한 해답을 내놓지 못하고 있다. 한가지 예는 학습과 행동의 유연성(behavioral flexibility)이다[1]. 알고리즘은 주어진 작업 특성에 맞춰서 개별적인 최적화 과정이 이루어져야 하지만, 인간은 하나의 생물학적 개체로서 다양한 작업을 유연하게 학습하고 수행할 수 있다. 인간의 유연한 학습의 비밀은 목적 지향적(goal-directed)인 탑다운 제어(top-down control)과정에 있다. 기존에 학습된 정보를 새로운 컨텍스트에 맞춰 빠르게 변형하는 적응능력은 인간의 시각, 청각 피질에서 쉽게 발견할 수 있으며, 행동 제어 측면에서는 전두엽이 관여한다고 알려져 있다.

또 하나의 예는 추론 학습 속도 문제다. 이는 환경 변화에 대한 적응력과 직결되는 중요한 문제다. 일반적인 알고리즘의 경우 새로운 환경에 놓였거나 환경의 컨텍스트가 변화하는 경우 많은 학습 샘플을 필요로 한다. 관측 가능한 샘플로부터 직접 관측이 불가능한 변수의 상태를 추정하는 추론(inference) 문제에서도 증거(evidence)가 가진 정보량과 추론 정확도는 비례한다. 이에 비해 인간은 한 두번의 경험만으로 결론을 도출하기도 한다.

앞서 소개한 개별적인 문제 해결 방식은 상황에 따라 장단점이 있으므로, 결국 컨텍스트 의존적인 제어 문제로 귀결된다. 예를 들어 유연한 학습 능력의 경우 적응력이 뛰어나지만 외부 노이즈에 강건(robust)하지 못하다. 고속 추론 학습 속도의 경우 환경에 대한 빠른 모델링을 가능케 하므로 직접적 샘플링 과정 없이 결과 예측이 가능하나, 잘못된 결론을 도출할 확률이 높으며 이러한 오류가 학습 전체에 전파될 수 있다. 종합해 보면 인간은 이러한 학습과 추론 과정을 상위 레벨에서 균형적으로 제어하는 능력을 가지고 있을 것이라는 가설을 세워 볼 수 있다.

의사결정 신경과학 분야의 화두는 앞서 소개한 학습, 추론, 그리고 고등 제어 기능에 대한 뇌과학적인 이해이다. 다음 장에서는 이러한 연구의 바탕이 되는 강화학습의 이론적 배경에 대해 소개하고, 이어 강화학습의 신경과학에 대해 보다 깊이 있는 논의를 이어가기로 한다.

4. 강화학습의 기본 개념

4.1 Markov decision process를 이용한 최적제어 문제의 형식화

강화학습은 보통 시간-공간적인 내부 상태 의존성이 존재하는 상황에서 '보상'(reward)'이라는 형태의 단순한 피드백만을 이용하여 문제 해결을 위한 전략을 학습하는 방법을 일컫는다. 보상을 최대화 하기 위한 전략 학습 문제는 비용(cost)을 최소화하는 최적제어(optimal control) 문제로 볼 수 있으며, 동적 프로그래밍(dynamic programming)과 같은 방식을 통해 이론적, 기술적인 발전을 거듭하여 왔다[6].

강화학습 연구에서는 다음과 같은 요소를 도입하여 전략 학습을 위한 문제 형성을 하였다. 전략 학습의 주체가 되는 것은 에이전트이고 에이전트가 전략을 수립해야 하는 대상은 환경이다. 에이전트는 현재 가지고 있는 전략을 바탕으로 행동을 하게 되고, 이는 환경의 입력이 된다. 환경은 에이전트의 행동에 대해 보상이라는 형태로 반응하게 되는데, 여기서 보상은 에이전트의 관측 변수에 포함된다. 다양한 상황을 고려하기 위해 일반적으로 환경은 불확실성을 내재한 동적 시스템으로 설정된다. 이를 수학적으로 정의한 것이 마르코프 의사결정 프로세스(Markov decision process; MDP)[7]이다.

4.2 Principle of optimality와 Bellman equation를 이용한 해법

강화학습 문제는 에이전트가 받을 미래의 보상(reward)에 대한 기대치를 최대화 (또는 비용을 최소화)하는 행동 전략을 찾는 과정으로 표현할 수 있다. 여기서 행동 전략은 매 상황(state) 또는 상황에서의 행동(action)에 대한 가치값(value)으로 표현되며, value는 일반적으로 에이전트가 받을 미래의 보상의 총 합에 대한 기대치로 정의된다. MDP 문제에서 보상의 기대치는 에이전트와 환경이 상호작용하는 경험으로부터 얻어지는 샘플을 이용하게 된다. 이 문제가 어려운 이유는 매 상황에서의 입력(에이전트의 행동)-출력(환경으로부터의 피드백)에 대한 보상이 간헐적으로 주어진다는 점이다(sparseness). 보상신호는 이러한 입력-출력이 반복되는 인터렉션 중간, 또는 가장 마지막에 주어지게 된다. 더구나 보상은 보상을 받는 시점의 상태에만 의존적이 아니라, 과거에 일어난 일련의 입출력 집합(episode)에 의존적이다.

MDP 세팅에서는 매 상황의 입출력은 바로 이전 상황의 입출력 쌍에 의존적인 형태로 표현된다. 만일 모든 상황(S)에 대한 피드백/보상치(R)의 정보와 상황들

간의 확률적인 관계를 나타내는 상태천이 행렬(P)이 주어진다면 우리는 아래와 같이 최적의 전략을 상태의 가치값 행렬로 표현할 수 있다.

$$v = R + \gamma P v$$
$$\Rightarrow v = (I - \gamma P)^{-1} R.$$

그러나 위와 같은 해법은 다음과 같은 이유로 일반적인 상황에 적용할 수 없는 경우가 많다. 우선적으로 생각해 볼 수 있는 문제는 환경에 대한 완벽한 정보인 P가 주어지지 않는다는 것이다. 환경 안에서 무한한 시간동안 탐험하면서 sampling을 한다면 P를 추정할 수 있으나, 위 식의 적용은 작은 규모의 이산적 공간에 국한되므로, 복잡한 문제에 대해서는 전체 R을 얻기도 어렵다. 앞서 언급한 여러 가지 일반적인 최적 제어의 문제에서는 고려해야 할 상황의 개수가 너무 많고, 이에 비해 sampling할 수 있는 기회는 상대적으로 적기 때문에 현실적으로 위와 같은 행렬 자체를 구하기가 어렵다.

Principle of optimality는 이러한 문제 해결을 위한 이론적 기틀을 제공한다[6]. 최초 상황(S_0) 에서 최종 보상을 받는 상황(S_n)을 이어주는 최적의 전략 M*가 존재한다고 가정하였을 때, 상황(S_{n-1})에서 S_n을 이어주는 최적의 해 M_{n-1}*는 M*의 부분집합이 된다. M_{i-1}*는 M_i*의 부분집합이 되며, 이를 풀어서 설명하면 보상을 획득한 시점에서의 부분전략을 재귀적(recursive)으로 확장해 나가면 임의의 상황(S_i)에서부터 시작되는 전체 전략과 같게 된다고 할 수 있다. 결국 최종적으로 얻은 보상에 대한 정보는 그 보상을 얻기까지 이루어진 과거의 episode상에 놓인 상황-행동-피드백 세트들로 역전파가 가능하다고 할 수 있다. 물론 이러한 이상적인 결론은 특정한 가정 하에 보장되는 것이다.

Bellman equation은 위와 같은 원리를 이용하여 전략으로부터 샘플링되는 상황 또는 행동의 가치값을 미래에 받을 보상의 총합의 기대치로 표현한다[1]. Bellman optimality equation은 최적 전략에 대한 가치와 기대값의 관계를 의미한다. 기대치는 MDP의 특성을 이용하여 recursive한 형태로 풀어쓸 수 있으며, 식을 정리해 보면 다음에 이어질 상황(s')에 대한 행동(a')의 가치값(Q(s',a'))을 현재 상황(S)에 대한 행동 가치값(Q(s,a))에 반영하는 단순한 형태의 업데이트 식으로 표현할 수 있다.

$$Q^*(s,a) = E_{(s,a,s')}\left[R + \gamma \max_{a'} Q^*(s',a')\right].$$

여기서 두 가지 특징에 주목해 볼 필요가 있다. 하나는 다음 상황(s')에서의 행동 가치값 중 최대치(max)를 선택하여 업데이트에 반영한다. 이는 에이전트 본인의 전략이 최적이라고 가정하는 것을 뜻한다(optimistic). 또 하나는 기대값을 추정하기 위해서는 환경에 대한 모델, 즉 (s,a,s')에 대한 확률분포가 필요하다는 점이다.

4.3 TD learning 알고리즘

앞 절에서 소개한 아이디어를 알고리즘의 형태로 표현한 것을 시간차 학습 (Temporal difference learning; TD learning)이라 부른다. 이에 비해, 전체 에피소드에 대하여 업데이트 하는 방식을 Monte Carlo learning (MC)이라 부른다. 여기서 전략을 매 "상황"에 대한 가치값으로 표현하는 방식을 TD evaluation, MC evaluation과 같이 부르며, 전략을 매 상황에서의 "행동"에 대한 가치값으로 표현하는 것을 TD control, MC control 등으로 부른다. 본 글에서는 인간의 실제 학습과 가장 가깝다고 알려진 TD control (이하 TD learning이라 표기)에 대해 논의하기로 한다.

TD learning은 앞서 소개한 Bellman equation에 기반하며, 환경에 대한 확률분포에 대한 추정 없이 현재 전략에 의존한 sampling 방식을 이용해 가치값을 업데이트한다는 점에서, model-free RL로 분류된다.

$$Q(s,a) \leftarrow^{\alpha} R + \gamma \max_{a'} Q(s',a') \ : \text{off-policy}$$
$$Q(s,a) \leftarrow^{\alpha} R + \gamma Q(s',a') - Q(s,a) \ : \text{on-policy}$$

이와 같이 현재 에이전트의 전략에서 고려되는 다음 상황(s') 에서의 가치 추정 정보를 현재 상황(s)의 업데이트에 이용한다(alpha는 learning rate를 의미). 업데이트 식은 다음 상황에서 가치 추정치를 높이는 방향으로 실제로 움직이면서 업데이트 하는지의 여부에 따라 on-policy와 off-policy로 분류된다. Off-policy와 on-policy TD learning 업데이트 식에서 아래 각각의 term은 현재의 가치값이 근접해야 할 목표라는 의미에서 TD target이라 불린다.

$$R + \gamma \max_{a'} Q(s',a') : \text{TD target(off-policy)}$$
$$R + \gamma Q(s',a') : \text{TD target(on-policy)}$$

딥러닝과 같이 파라미터화 된 신경망을 이용해 가치값(가치망)을 학습시키는 최근의 심층 강화학습 문제 역시 TD target에 기반한다. 여기서 TD target은 임시 목적함수로 설정되어 TD target에 대한 가치망의 예측 손실(loss)인 TD error를 최소화 시키는 방식(gradient descent)으로

가치망의 파라미터를 업데이트 한다.

$$\theta^* = \arg\min_\theta \sum_t \left\| Q_\theta(s_t, a_t) - \widehat{TQ}_t \right\|^2,$$
$$\widehat{TQ}_t = r_t + \gamma \max_{a'} Q_{\theta^-}(s', a').$$

여기서 한 batch의 episode를 이용한 학습은 Bellman backup operator를 가치망에 적용하는 것으로 간주된다.

지금까지 설명한 심층 강화학습 알고리즘 학습의 핵심은 TD error를 최소화하는데 있다. 딥러닝을 이용하여 이산적인 action 공간에서 가치망을 학습시키는 방식을 일반적으로 value iteration이라 하며, action 자체에 대한 정보를 직접 출력할 수 있는 정책망을 학습시키는 방식을 policy gradient (PG)라 한다. 특히, (stochastic) gradient descent 방식의 학습과정에서 환경에 대한 확률분포에 대한 별도의 미분항이 필요하지 않으므로[8,9], TD learning과 마찬가지로 model-free RL로 분류된다.

4. 강화학습의 신경과학

강화학습의 알고리즘과 평행하게 최근 계산신경과학 연구에서는 강화학습의 계산적 핵심 요소들이 뇌에서 어떻게 구현되어 있는지에 대한 연구가 활발히 진행되어 왔다. 신경과학에서의 강화학습 연구 앞서 소개한 강화학습의 기본 이론에서 다루고 있는 핵심 정보들이 뇌의 신경 활성도에 반영되어 있는지를 확인하는데서 시작한다.

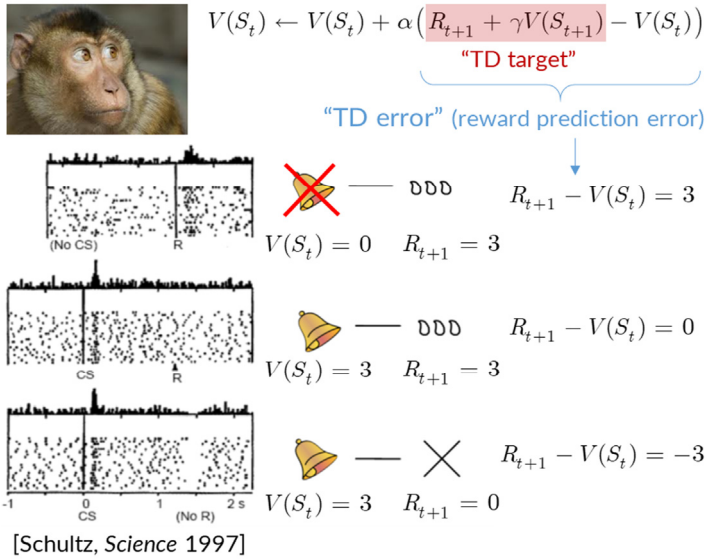
4.1 기저핵의 TD learning

신경과학 분야에서는 뇌의 쾌락 중추로 잘 알려진 도파민 시스템에 대한 연구가 활발히 이루어져왔다. 강화학습의 관점에서 도파민 시스템을 이해하는 연구는 1990년대 후반 Wolfram Schultz의 발견에서 출발한다[10]. 이후 수많은 후속 연구가 이루어져 왔고, 이를 통해 앞 절에서 설명한 on-policy 방식과 같은 TD learning은 인간을 포함한 실제 동물의 강화학습 방식과 유사하다고 알려져 있다.

Schultz의 대표적인 도파민 연구[10]에서는 원숭이에게 큐(cue)가 주어지고 일정 시간이 지난 이후 주스와 같은 보상이 주어지는 실험을 사용하였다. 이러한 경험이 반복됨에 따라 원숭이는 시간적으로 먼저 발생하는 사건인 큐를 이용해 이후 보상을 예측하게 된다. 연구자들은 학습과정에서 일어나는 도파민 뉴런의 신경활성 패턴을 측정하였는데, 이 패턴이 TD error 와 높은 정확도로 일치한다는 것을 발견하였다(그림 2).

큐가 보상을 미처 예측하지 못하는 학습 초기에는 보상이 주어지는 시점에서 도파민 뉴런이 활성화 된다 (TD error > 0; 그림의 첫 번째 경우). 학습이 진행됨에 따라 큐가 주어지는 시점에서 도파민 뉴런이 활성화되기 시작하며 (value > 0; 그림의 두 번째 경우) 보상이 주어지는 시점에서는 도파민 뉴런의 활성도가 정상 레벨도 돌아온다(TD error = 0; 그림의 두 번째 경우). 이후 갑자기 보상을 주지 않게 되면(devaluation) 보상이 예측된 시점에서 도파민 뉴런이 비활성화 된다 (TD error < 0; 그림의 세 번째 경우). 이는 도파민 뉴런이 큐와 같은 외부자극에 대하여 보상 예측 에러를 바탕으로 가치값을 학습함을 의미한다.

이 실험 결과는 동물의 뇌가 TD learning과 같은 방



[Schultz, *Science* 1997]

그림 2 강화학습의 신경생리학적 증거: 가치값과 보상 예측 에러 (reward prediction error)

식으로 강화학습을 한다는 직접적인 신경과학적 증거이다. 이후 도파민 시스템이 보상의 불확실성과 같은 상위 레벨의 정보도 반영한다는 발견[11]등과 같은 다양한 연구로 이어지면서, 강화학습이 뇌 안에서 어떻게 구현되어 있는지에 대한 수많은 연구가 활발하게 이루어져 왔다.

4.2 Bellman equation으로 본 두 가지 형태의 강화학습

앞서 3절에서는 강화학습은 Bellman equation으로 요약될 수 있으며, TD learning과 같은 방식은 - 환경에 대한 모델 없이 학습한다 - 는 점에서 model-free RL로 분류된다는 것을 논의한 바 있다. 이렇게 model-free로 학습된 전략은 매 상황에서 가치값을 즉시 추출하여 행동으로 변환하므로, 습관적인 행동패턴(habitual behavior)을 설명할 수 있다[12]. 습관적인 행동 패턴은 devaluation과 같이 환경이 변하는 상황에서 반복적인 경험을 통해 서서히 전략을 수정하게 되므로 행동의 유연성이 떨어지는 양상을 보인다[5].

4.1절에 소개한 바와 같이 TD learning과 같은 model-free RL이 뇌 안에서 어떻게 구현되어 있는지에 대한 이해가 깊어지면서, model-free가 설명하지 못하는 목적 지향적인 행동패턴 (goal-directed behavior)의 계산적/신경과학적 메커니즘에 대한 질문이 대두되었다.

2005년 발표된 Daw의 논문[13]에서는 이 두 가지 행동패턴을 만들어 내는 동물의 강화학습 전략에 대한 다음과 같은 가설을 제시하였다. 동물의 뇌에서는 습관적인 행동 전략을 만들어내는 model-free RL 뿐만 아니라, 목적 지향적인 행동 전략을 만들어내는 model-based RL 역시 구현되어 있을 것이며, 이 두 가지 형태의 강화학습을 제어하는 상위 레벨의 메커니즘 역시 뇌 안에 구현되어 있을 것이라는 가설이다.

그림 3에서 보이고 있듯이 model-based 강화학습은 환경에 대한 모델 (4.2 절 Bellman optimality equation의 state-action-state transition 확률분포 부분)을 추정하고, 이를 가치값 업데이트에 반영한다. 이러한 점에서 확률분포를 uniform 으로 가정하고 sampling을 통해 가치값을 업데이트하는 model-free와 구분된다. 성능면에서는 두 방법 모두 같은 해로 수렴하지만, model-based의 경우 환경 모델을 이용하므로 일반적으로 수렴 속도가 더욱 빠르다. 극단적으로 동적인 환경이나 잡음이 너무 큰 환경에서는 부정확한 모델의 추정이 오히려 학습에 방해되므로 model-based 보다는 model-free를 사용하는 것이

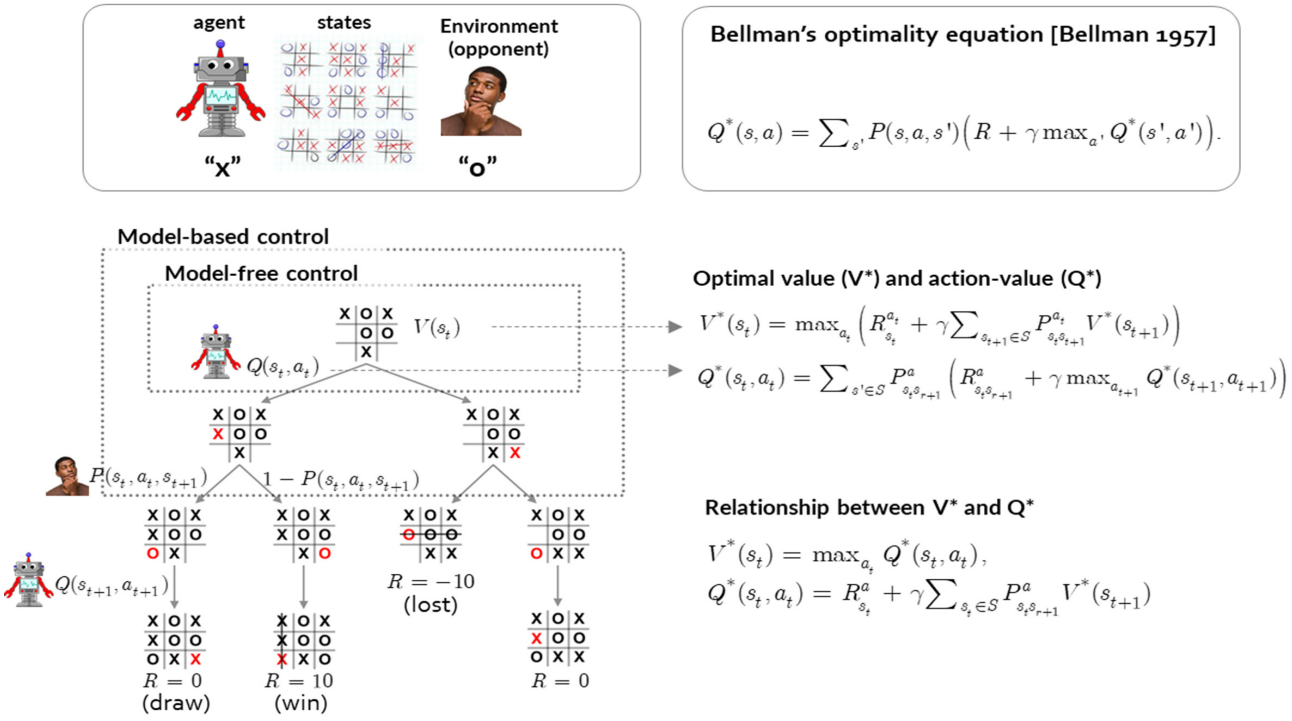


그림 3 Model-free와 model-based 강화학습 개념. Model-free control 방식으로 강화학습하는 에이전트는 매 상황(state)에서 가치값(Q)을 최대화 시키는 방향으로 행동한다. 이 모델은 상황변화(state-transition; P(s,a,s'))을 별도로 가정하지 않고 경험적(sampling)인 방식으로 가치값을 업데이트한다. 이에 비해 Model-based control 방식의 강화학습은 매 상황에서 행동한 이후의 상황변화에 대한 추정(P(s,a,s')) 이용하여 가치값을 업데이트한다. 예시로 주어진 게임 상황에서 P(s,a,s'))는 상대방의 전략 모델로 볼 수 있다.

optimal한 전략이라 할 수 있다[14].

성능면에서는 model-based가 model-free보다 우수하지만, model-based 전략을 수행하기 위해서는 추가적인 계산 자원이 필요하다. 이는 신경과학 관점에서 일반적으로 cognitive load[15]라고 불리우는 요소이자, model-based와 model-free를 구분하는 또 하나의 중요한 지표이다.

요약하면 model-based RL은 빠른 가치값 수렴속도와 목적의 수정에 따른 행동의 유연성이라는 장점을 가지고 있으나, 추가적인 계산으로 인해 실제 에이전트의 반응속도(reaction time)가 느리고 cognitive load가 많다는 단점이 있다. 이에 비해 model-free는 수렴속도가 느리고 목적이 변경될 경우 반복적인 에피소드를 통해 점진적으로 학습해야 한다는 단점이 있으나, 현재 policy를 바로 행동으로 전환하는 형태로써 반응속도가 빠르고 cognitive load가 적다는 장점이 있다. 이와 같이 model-based RL와 model-free 강화학습 전략은 상호보완적인 관계로서 상황과 작업 조건에 따라 최적 제어하는 메커니즘이 요구된다.

4.3 전전두피질–기저핵의 메타 강화학습 메커니즘

Model-based와 model-free 강화학습 메커니즘이 동시에 존재할 수 있다는 신경과학적 증거는 2011년에야 비로소 제시되었다[16]. 이는 fMRI 데이터에 기반한 연구로, 인간은 학습 초기에 model-based 전략으로 시작하지만, 안정적인 학습이 이루어짐에 따라 점차 model-free 전략을 선호하는 행동패턴을 보임을 발견하였다. 강화학습 전략의 선호도가 model-based에서 model-free로 이동한다는 현상은 인간의 뇌에서 메타 강화학습이 실제로 일어날 수 있다는 주장을 뒷받침한다. 또 다른 연구에서는 model-based와 model-free와 관련된 정보가 기저핵에서 처리되고 있으며, 개개인에게 내재된 강화학습 전략의 선호도가 기저핵 정보처리 과정에 영향을 끼친다는 것을 발견하였다[17].

상황변화에 민감한 강화학습 제어 과정을 규명하기 위해서는 model-based와 model-free 강화학습의 행동이 분리될 수 있는 시나리오가 필요하다. 앞서 설명한 바와 같이 환경이 안정적인 경우 두 프로세스의 수렴성에 의해 그 차이가 사라지게 된다. 2014년 연구에서는 model-based에 핵심인 목적(goal)과 state-action-state transition 확률이라는 두 가지 환경 변수를 조절함으로서 두 프로세스가 각각 다른 예측을 하도록 유도할 수 있는 시나리오 (Markov decision task)를 만들었다[18]. 이 시나리오에서는 두 모델의 TD error와 가치값 신호가 잘 분리된다.

이렇게 분리된 각각의 신호와 실제 각각의 뇌 부위 신호(fMRI)와의 상관성을 계산하고(statistical map), 그 중 상관도가 확률적으로 유의미한 부위를 찾아낼 수가 있다. 이러한 분석 방법은 계산 모델을 이용해 뇌 신호를 분석한다는 의미에서 model-based fMRI 분석이라 불린다.

찾아낸 model-free와 model-based RL의 보상/상태 예측 에러(state/reward prediction error) 신호는 그림4의 좌측 상단과 같다(보상/상태 예측 에러 정보). Model-based 강화학습의 상태 예측 에러는 측전두엽 앞부분(lateral prefrontal cortex)과 전두엽 뒷부분(intraparietal sulcus)에서 처리되고 있으며, model-free 강화학습의 보상 예측 에러(TD error)는 기저핵의내부의 ventral striatum에서 처리된다.

Model-based 강화학습 에이전트가 보여주고 있는 목적 기반 행동은 devaluation과 같은 실험 연구에서의 행동 패턴 분석을 통해서도 드러나지만, planning과 같이 보다 직접적으로 드러나는 경우도 있다. 예를 들면, 갑자기 작업 목적이 변경된 경우를 생각해 볼 수 있다. 에이전트는 새로운 목적을 성취하기 위해 행동/경로수정을 하게 되며, 이를 위해서는 예전에 가지고 있던 가치 정보를 새 목적에 맞게 한꺼번에 업데이트하는 과정이 필요하다. 이러한 개념에 기반하여 model-based 강화학습 에이전트의 planning과정을 가치 정보의 업데이트 크기로 정량화할 수 있다. 기억을 담당하는 해마(hippocampus)와 전두엽의 여러 부위가 해당 정보 처리에 관여하는 것으로 보고된 바 있다 (그림 4의 목적 기반 계획 정보 참조).

다음으로는 강화학습의 정책의 기반이 되는 가치값 신호인데, 이는 대체로 도파민 중추에 해당되는 기저핵(basal ganglia)과 신호를 주고받는 부위들에서 나타난다. 먼저 model-based 강화학습의 가치 신호는 후내방 전전두피질(dorsomedial prefrontal cotex)과 운동 피질 영역에서 관찰되며, model-free 강화학습의 가치 신호는 기저핵 내부의 후부 피간 (posterior putamen)에서 주로 관찰된다 [19,20].

에이전트가 강화학습을 통해 가치를 실제 행동으로 변환하려면 이 두 가지 가치 신호가 통합되어야 한다는 논리가 필요하다. 실제 선택과 연관된 가치 신호는 선택한 옵션의 가치값(chosen value)과 선택하지 않은 옵션(unchosen value)의 가치값의 차이로 정의되며, 복내측 전전두피질(ventromedial prefrontal cortex)을 중심으로 신호가 나타난다 (그림4의 가치값 정보). 이는 해당 뇌 부위가 강화학습을 통해 추정된 가치값들을 실제 행동으로 변환하는 과정에 관여한다는 것을 의

미한다. 이 결과는 의사결정 신경과학과 신경 경제학 분야의 연구에서 여러차례 재현되어 왔기 때문에 [21,22,23] 이제는 가설이 아니라 새로운 연구을 위한 가설을 세우는데 활용되기도 한다. (그림4의 가치값 정보 관련 논문 참조)

Model-based와 model-free 강화학습 관련 정보들이 통합되는 계산 과정을 이해하기 위한 연구도 진행되어 오고 있다. 이는 앞 절에서 설명한 2005년 Daw의 불확실성 기반 제어 가설[13]에서 출발한다. 이 가설을 뒷받침하는 직접적인 증거는 약 10년 뒤인 2014년에야 비로소 제시되었다[18]. 이 연구에서는 – model-based와 model-free 강화학습 프로세스가 뇌 수준에서 관측하는 신호인 "보상/상태 예측 에러의 불확실성"에 기반하여 학습 전략의 선호도가 결정된다 – 는 계산적 메타 강화학습 가설을 검증하였다. 관련된 신호들은 복외측 전전두피질(ventrolateral prefrontal cortex)과 전두엽피질 부위(frontopolar cortex)에서 처리되고 있다. 이러한 발견에 기초하여 이루어진 기저핵-전전두피질 뇌 네트워크 연구에서는, (1) model-free 강화학습 가치값 정보 처리의 핵심 부위인 기저핵과 (2) 가치값이 통합되는 복내측 전전두피질 사이의 정보 전달이 model-based 강화학습 전략을 사용하는 동안에 약해진다는 것이 밝혀졌다. 이러한 결과는 목적기반 행동과 습관적 행동 전략을 비교하는 연구의 결과와도 일치한다[20].

앞서 소개된 연구 결과들을 종합하면 (1) 우리의 뇌는 예측 에러 (prediction error)를 기반으로 강화학습 전략을 선택하며, (2) 습관적 행동패턴 형성에 관여하는 model-free 강화학습 전략을 사용하는 동안에는 도파민 시스템이 우리의 의사결정에 많은 영향을 끼치고, 반대로 (3) 목적기반 행동패턴 형성에 관여하는 model-based 강화학습 전략을 사용하는 동안에는 우리의 의사결정이 도파민 시스템보다는 전두엽에 의존적하고 있다고 요약할 수 있다.

5. 결론: 유연한 강화학습 연구를 위한 기계학습-신경과학 융합적 접근

지금까지 강화학습의 기본 개념과 알고리즘, 그리고 이러한 아이디어가 어떻게 신경과학 연구에 적용되어 왔는지를 살펴보았다. 앞 절에서 설명한 인간의

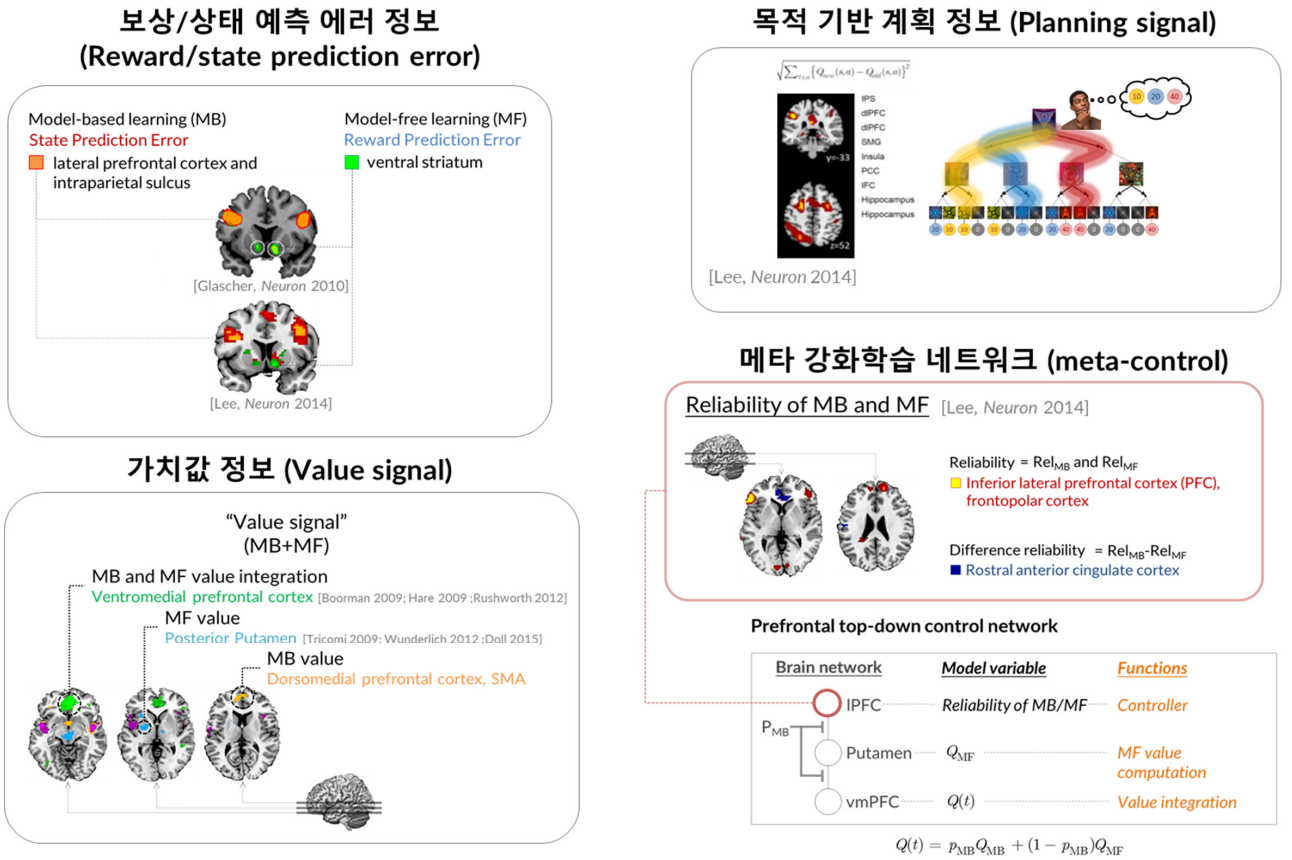


그림 4 Model−free 강화학습, model−based 강화학습, 그리고 메타 강화학습과 관련된 정보가 인간의 뇌에서 처리되고 있다는 것을 보여주는 신경과학적 연구결과들

유연한 강화학습은 문제해결의 복잡성과 문제 해결의 속도 측면에서 다음과 같이 요약해 볼 수 있다.

새로운 환경에 노출되어서 빠른 적응이 필요하거나 목적이 유동적인 경우에는 보다 적극적으로 환경 모델을 학습하게 되며, 이는 model-based 강화학습에 해당된다. 이러한 전략을 사용할 경우 환경 모델에 대한 제어력(controllability)를 확보하게 되므로, 게임과 같은 상황에서 상대방의 전략을 고려하거나 상대방의 모델을 역이용하여 상대방을 특정 상태로 유도하는 등 보다 높은 차원의 작업 수행이 가능해진다. 그러나, 메모리 등 인지적 부담(cognitive load) 크고 많은 계산량이 요구되므로 에너지 효율 측면에서는 좋지 않은 전략이다. 따라서 일정 수준이상의 전략이 형성된 이후에는 계산량이 적어 효율이 좋고 반응속도가 빠른 model-free 강화학습을 사용하는 것이 더욱 효과적이다. 이렇게 인간은 두 가지 강화학습 전략을 유연하게 제어함으로써 성능, 에너지, 시간 효율사이의 trade-off를 균형적으로 추구한다.

Model-based 강화학습을 위해서는 환경과 상호 교류하는 sampling 과정이 필요한데, 인간은 환경의 불확실성이 높은 부분에 집중하여 빠르게 정보를 업데이트(예: one-shot inference)하는 방식으로 환경에 대한 전체적인 불확실성을 줄여나간다. 그러나 적은 observation기반의 고속 학습은 에러가 전체 지식구조에 전파되어 잘못된 환경 모델 또는 준 최적의 전략으로 수렴할 위험성이 있으므로, 인간은 느린 학습과 빠른 학습을 적절히 조합함으로써 학습 효율과 일반화 문제를 모두 해결한다고 할 수 있다.

본 글에서 자세히 다루지는 않았으나 인간의 강화학습은 컨텍스트에 민감하다. 어려운 종류의 작업에서는 스스로의 성능에 대한 기대치를 낮춰 예측에러에 덜 민감해지며, 반대로 쉬운 작업에서는 예측에러에 대한 민감도를 높여 보다 높은 성능을 추구한다. 이는 강화학습 알고리즘에서 다뤄지고 있는 보상 예측 에러의 variance문제를 인간은 이미 효율적으로 풀고 있음을 의미한다.

이러한 인간의 강화학습 능력은 여러 가지 측면에서 심층 강화학습 알고리즘 보다 유연하다고 할 수 있다. 가장 먼저 생각해 볼 수 있는 응용은 이러한 메타 제어 원리를 알고리즘 개발에 적용하는 것이다. 그러나 보다 중요한 이슈는 기계학습의 관점에서 보았을 때 준 최적(suboptimal)인 인간의 강화학습 프로세스 자체를 모델링하는 것이다. 이러한 모델에 기반하여 동작하는 컴퓨터는 인간-컴퓨터/로봇 상호작용 환경에서 인간의 만족도를 더욱 높일 수 있을 것이다. 또 하나의 공학적 응용은 인간의 학습 과정을 최적의 상태로 유지하는 시스템 개발이다. 교육, 의료정보, 법률과 같이 다수의 컨텍스트에서 발생하는 복잡한 정보를 학습하고 추론해야 하는 직업군에서 이러한 시스템은 인간과 지식 기반 시스템 사이의 효율적 매개체 역할을 할 수 있다. 또 하나의 예는 메타 제어 개념을 이용한 Brain-Computer Interface (BCI) 시스템이다. 앞 절에서 소개하였든 인간이 문제해결과정에서 유연하게 다양한 전략을 사용하고, 각각의 전략에 관여하는 뇌 부위가 다른 경우가 많으므로, 이러한 예측을 바탕으로 매 시간 뇌에서 정보를 읽어내는 위치와 특징을 유동적으로 조절하는 BCI 방식을 이용한다면 성능 향상을 기대할 수 있을 것이다.

마지막 응용 가능성은 정신질환의 정밀진단과 예후예측이다. 계산정신의학이라 불리우는 연구 분야는 뇌의 계산적, 시스템적 이해를 바탕으로 복잡한 정신질환을 이해하고자 하는 새로운 접근 방법으로, 최근 관심이 높아지고 있는 추세이다.

이와 같이 강화학습의 신경과학적 연구, 그리고 신경과학적 이해를 강화학습 알고리즘에 적용하는 연구는 무한한 가능성을 가지고 있다. 이러한 연구를 위해서는 두 분야의 접근 방식에 대한 깊은 이해를 바탕으로 하는 융합적인 연구가 요구된다.

참고문헌

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning. MIT press, 1998.

[2] V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529-533, Feb. 2015.

[3] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, pp. 484-489, Jan. 2016.

[4] D. Silver et al., “Mastering the game of Go without human knowledge,” Nature, vol. 550, no. 7676, pp. 354-359, Oct. 2017.

[5] J. P. O’Doherty, S. W. Lee, and D. McNamee, “The structure of reinforcement-learning mechanisms in the human brain,” Curr. Opin. Behav. Sci., vol. 1, pp. 94-100, Oct. 2014.

[6] D. P. Bertsekas, Dynamic programming and optimal control. Athena Scientific, 2005.

[7] M. L. Puterman, Markov decision processes : discrete

stochastic dynamic programming. Wiley-Interscience, 2005.

[8] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, “Policy Gradient Methods for Reinforcement Learning with Function Approximation.” pp. 1057-1063, 2000.

[9] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. JMLR.org, p. I-387, 2014.

[10] W. Schultz, P. Dayan, and P. R. Montague, “A neural substrate of prediction and reward,” Science (80-.)., vol. 275, pp. 1593-1599, 1997.

[11] C. D. Fiorillo, P. N. Tobler, and W. Schultz, “Discrete coding of reward probability and uncertainty by dopamine neurons.,” Science, vol. 299, no. 5614, pp. 1898-902, Mar. 2003.

[12] B. W. Balleine and J. P. O’Doherty, “Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action.,” Neuropsychopharmacology, vol. 35, no. 1, pp. 48-69, Jan. 2010.

[13] N. D. Daw, Y. Niv, and P. Dayan, “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control,” Nat. Neurosci., vol. 8, pp. 1704-1711, 2005.

[14] P. D. Máté Lengyel, “Hippocampal Contributions to Control: The Third Way,” in Advances in Neural Information Processing Systems (NIPS), 2008, pp. 889-896.

[15] S. a Sheth et al., “Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation.,” Nature, pp. 3-7, Jun. 2012.

[16] J. Gläscher, N. Daw, P. Dayan, and J. P. O’Doherty, “States versus Rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning,” Neuron, vol. 66, no. 4, pp. 585-95, May 2010.

[17] N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan, “Model-based influences on humans’ choices and striatal prediction errors.,” Neuron, vol. 69, no. 6, pp. 1204-15, Mar. 2011.

[18] S. W. Lee, S. Shimojo, and J. P. O’Doherty, “Neural Computations Underlying Arbitration between Model-Based and Model-free Learning,” Neuron, vol. 81, no. 3, pp. 687-699, Feb. 2014.

[19] E. Tricomi, B. W. Balleine, and J. P. O’Doherty, “A specific role for posterior dorsolateral striatum in human habit learning,” Eur. J. Neurosci., vol. 29, pp. 2225-2232, 2009.

[20] K. Wunderlich, P. Dayan, and R. J. Dolan, “Mapping value based planning and extensively trained choices in the human brain,” Nat. Neurosci., vol. 15, pp. 786-791, 2012.

[21] E. D. Boorman, T. E. Behrens, M. W. Woolrich, and M. F. S. Rushworth, “How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action,” Neuron, vol. 62, pp. 733-743, 2009.

[22] T. a Hare, C. F. Camerer, and A. Rangel, “Self-control in decision-making involves modulation of the vmPFC valuation system,” Science (80-.)., vol. 324, pp. 646-648, 2009.

[23] M. F. S. Rushworth, M. P. Noonan, E. D. Boorman, M. E. Walton, and T. E. Behrens, “Frontal Cortex and Reward-Guided Learning and Decision-Making,” Neuron, vol. 70, pp. 1054-1069, 2011.

약 력

이 상 완

2003 연세대학교 전기전자공학과 졸업 (학사)
2005 KAIST 전자전산학과 졸업 (석사)
2009 KAIST 전자전산학과 졸업 (박사)
2010~2011 MIT 박사후연수연구원
2011~2015 CALTECH 박사후연수연구원
2015~현재 KAIST 바이오및뇌공학과 조교수
관심분야: 뇌기반 인공지능, 계산신경과학
Email: sangwan@kaist.ac.kr