로봇을 위한 인공 두뇌 개발

Artificial Brain for Robots

이 규 빈[1] · 권 동 수[2]

Lee Kyoobin[1] · Kwon Dong-Soo[2]

Abstract This paper introduces the research progress on the artificial brain in the Telerobotics and Control Laboratory at KAIST. This series of studies is based on the assumption that it will be possible to develop an artificial intelligence by copying the mechanisms of the animal brain. Two important brain mechanisms are considered: spike-timing dependent plasticity and dopaminergic plasticity. Each mechanism is implemented in two coding paradigms: spike-codes and rate-codes. Spike-timing dependent plasticity is essential for self-organization in the brain. Dopamine neurons deliver reward signals and modify the synaptic efficacies in order to maximize the predicted reward. This paper addresses how artificial intelligence can emerge by the synergy between self-organization and reinforcement learning. For implementation issues, the rate codes of the brain mechanisms are developed to calculate the neuron dynamics efficiently.

Keywords: Artificial Brain, Spiking Neural Network, Spike-timing Dependent Plasticity, Dopamine Model

1. 서 론

로봇 기술의 발전은 더욱 빠르고 정확한 제품 생산을 가능케 함으로써 인류의 삶을 더욱 풍요롭게 만들었다. 센서 기술의 발전과 액추에이터의 소형화로 인해 로봇은 산업 현장의 울타리를 벗어나 점차적으로 인간의 생활 공간으로 투입되어 보다 복잡하고 섬세한 작업을 수행하는 단계에 접어들고 있다. 국가기술지도총론에 의하면 일본의 동경대에서는 서비스 로봇 산업의 규모가 2020년경에는 자동차 산업의 규모를 앞지를 것으로 예상하고 있다. [1]

그러나 서비스 로봇이 지금의 퍼스널 컴퓨터와 같이 각 가정에서 사용되려면 극복해야 할 기술적 문제들이 많이 있다. 물체 인식, 위치 추정, 얼굴 인식, 표정 인식, 음성 인식, 매니퓰레이터 제어, 비주얼 서보잉, 온톨로지 DB의 구축과 활용 등의 기술들이 유기적으로 결합되어 해결되어야만 실생활에서 사용 가능한 지능적인 로봇이 탄생할 수 있다. 이러한 이유로 현재 상용화 된 서비스 로봇은 공공 기관에서의 안내 로봇이나 가정용 청소 로봇, 그리고 애완용 장난감 로봇 정도로 한정되어 있는 실정이다.

산업용 로봇과 서비스 로봇은 작동 환경과 요구 사항에서 큰 차이가 있다. 산업용 로봇은 정형화된 작업장 내에서 미리 계획된 동작만 반복하는 반면, 서비스 로봇은 변화가 심한 환경에서 상황에 따라 작업의 내용이 수시로 변화하게 된다. 이런 차이점을 극복하기 위해 지능 로봇 연구자들은 로봇이 복잡한 환경을 인식하고 학습할 수 있는 여러 알고리즘들을 개발하고 있지만, 대부분의 알고리즘들은 주어진 작업을 성공적으로 수행하기 위해서 학습 파라메터들을 조정하는 등 사용자의 개입을 필요로 하기 때문에 비전공자가 일상생활에서 사용하기엔 아직 무리가 있다. 산업용 로봇의 경우 그 목적이 효율적인 제품 생산이므로 사용자가 개입하는 것에 대해 문제가 크게 없으나, 서비스 로봇은 그 목적이 인간을 대신하여 스스로 작업을 수행하는 것이므로 사용자의 개입이 최소화되어야 한다. 예를 들어 자동차 조립 라인의 로봇의 경우, 새로운 파트를 조립해야 할 때 관리자가 새로운 파트의 모양과 작업의 궤적(trajectory) 등을 지정해 주어도 문제가 없지만, 서비스 로봇의 경우, 불특정의 물체를 집기 위해 일일이 궤적을 지정해 줄 수는 없으므로 일상 생활 속에서 실시간으로 학습할 수 있는 능력이 필수적이다. 이러한 학습 방식은 인간을 포함한 실제 동물들이 환경과 상호작용

※ 본 연구는 21 세기 프론티어 R&D 사업 중 하나인 지능로봇 사업단과 학술진흥재단의 신진연구장려금지원사업의 지원으로 수행되었음.

[1] 한국과학기술원 기계공학과 박사과정(leekb@robot.kaist.ac.kr)

[2] 한국과학기술원 기계공학과 교수(kwonds@kaist.ac.kr)

을 통해 학습하는 패턴과 유사하다. 동물의 학습 방식 중에서 어떤 특성들이 중요하며 어떻게 구현할 수 있는지를 파악한다면 새로운 로봇 지능 알고리즘을 개발할 수 있을 것이다.

전통적으로 학습 알고리즘들은 동물이 학습하는 과정을 관찰하거나 연구자 자신이 학습하는 과정을 논리적으로 분석하고 수식화하여 개발되었다. 전자의 대표적인 예는 인공신경망 (artificial neural network) 또는 유전 알고리즘 (genetic algorithm) 등을 들 수 있고, 후자의 예로는 기호논리학으로부터 출발한 게임 트리 (game tree), 인공지능 언어 등을 들 수 있다. 본 연구는 전자의 범주에 속한다. 동물의 학습 메커니즘을 모방하여 학습 알고리즘을 만드는 가장 큰 이유는 동물의 학습 방법은 이미 구현되어진, 실재하는 지능이라는 점이다. 즉, 모방하는 대상과 방법이 옳다면 인공 지능을 구현할 수 있을 가능성이 있다는 것이다. 다시 말해, 인간 수준의 인공 지능을 구현하기 위해서는 다양한 기호 논리학 방법론을 채택하여 구현할 수도 있지만, 있는 그대로의 인간의 두뇌를 모방하는 방법도 있을 수 있다는 것이다. 물론 이러한 모방 과정에서 가장 중요하고 어려운 측면들을 살펴보면, 첫째로, 동물 두뇌에 대한 신비가 아직 다 밝혀져 있지 않다는 것이다. 동물 두뇌가 가지고 있는 기능들 중에서 지능을 발현하게끔 하는 기능이 아직 밝혀져 있지 않다면 모방할 대상도 없는 것이다. 본 연구에서는 이미 밝혀진 뇌의 기능으로부터 인공 지능을 구현할 수 있을 것이라는 가설을 토대로 연구가 진행중이며 자세한 기능에 대한 설명은 뒤에서 밝히도록 한다. 둘째로 동물 두뇌의 여러 가지 기능들이 밝혀졌다고 했을 때 인공 지능을 구현하기 위해 사용되어야 할 기능은 무엇이고, 사용되지 않아도 될 기능은 무엇인지 판별해 내는 것이다. 예를 들어 동물의 두뇌는 생명체이므로 세포가 죽지 않고 유지될 수 있도록 해 주는 영양 효과 (Tropic effects) 와 신경 성장 인자 (nerve growth factor) 의 기능을 갖고 있으나, 컴퓨터로 구현하는 지능 알고리즘에서는 이러한 기능들은 배제해도 좋을 것이다. 셋째로 현존하는 컴퓨터의 성능으로 모방해 낼 수 있느냐 하는 문제가 있을 수 있다. 인간의 두뇌의 경우 신경세포의 수는 1조 개가 넘고 하나의 신경 세포는 1만개 이상의 신경 세포와 연접 (synapse)을 이룰 수 있다. 만약 각각의 신경 세포와 신경 연접을 하나의 비트 (bit) 만 가지고 표현한다 해도 대략 1,000,000 기가바이트 (gigabyte) 의 메모리가 필요하고, 이들을 시뮬레이션 하기 위한 CPU 의 속도도 현존하는 가장 빠른 컴퓨터라고 해도 턱없이 부족한 현실이다. 그러므로 모방하려는 뇌의 기능과 실제 적용하려는 어플리케이션 사이의 적절한 타협이 필요하다.

이러한 어려움에도 불구하고 낙관적인 시각으로, 장기적인 연구 개발의 토대를 구축하는 의미로서 본 연구가 시작되었다. 연구의 방향성을 고려했을 때 본 연구는 기존의 인공 신경망의 재해석을 통한 새로운 인공 신경망 알고리즘을 고안하는 것으로도 볼 수 있을 것이다. 그러나 본 연구는 기존의 인공 신경망이 고안되고 연구되어온 것 보다 더욱 생물학적 발견들을 토대로 이루어지고 있으며, 2000년 이후의 가장 최근의 생물학적 발견들도 포함되어 있다.

본 연구는 로봇의 지능을 구현하는데 있어서 특정한 작업 수행의 성능에 초점을 두기 보다, 변화하는 환경 속에서 새로운 상황을 인식하고 자신의 행동을 평가하여 수정해 나갈 수 있는 능력을 가지는 학습 알고리즘을 개발하고 로봇에 이식하는 데 목적이 있다. 각 알고리즘의 세부적 사항들은 추후 발표되는 논문들에서 다루기로 하고 이 논문에서는 전체적인 연구의 흐름과 결과를 요약하는 데 중점을 두기로 한다.

2. 자가 조직화와 강화학습

로봇을 위한 인공 두뇌 알고리즘을 개발하기 위하여 먼저 동물이 학습하는 메커니즘을 분석했다. 동물이 학습하는 과정은 크게 [인식 → 판단 → 행동 → 평가 → 수정] 의 단계로 단순화할 수 있다. 즉, 동물은 새로운 자극이 들어왔을 때 그 자극이 다른 자극과 구분됨을 알게 되고 그 자극에 대한 반응을 결정하여 행동한 후, 그 행동의 결과에 따라 학습해 나가는 일련의 과정을 지속적으로 거치게 된다. 이와 같은 과정을 동물의 두뇌에서 발생하는 현상과 비교하여 단순화 하는 과정을 거치면 자가 조직화(self-organization)와 강화 학습 (reinforcement learning)으로 재분류할 수 있다. 자가 조직화는 주어진 자극으로부터 신경세포(neuron)의 발화 패턴이 점차적으로 특정한 패턴으로 수렴해 가는 과정으로 해석할 수 있다. 실제로 동물 두뇌에서는 헵의 학습 (Hebbian learning) 이용하여 자가 조직화가 이루어진다는 연구가 1950년대부터 발표되었고,[9] 최근에는 헵의 학습보다 보다 발전된 형태의 STDP (spike-timing dependent plasticity)가 네이처지 등에서 발표되고 있다. [2,7,13] STDP란 신경 세포가 발화하는 순서에 따라 신경 연접

이 강화되거나 약화되는 현상을 말한다. STDP에 관한 더 자세한 설명은 3장에서 논하도록 한다.

익숙해진 자극에 대해 동물은 스스로 판단하여 어떤 행동을 보이게 되는데, 그 행동이 주변 환경에 영향을 끼쳐 동물의 신체에 도움이 되는 결과를 가져오게 되면 그 자극과 행동 사이의 상관관계는 더욱 강화되고, 반대로 그 행동이 신체에 악영향을 끼친다면 상관관계가 약화된다.

일상 생활 속에서 실시간으로 학습한다는 것은 미리 짜여진 틀 속에서 학습이 이루어지는 것이 아닌, 환경과의 상호작용을 통해서 학습이 이루어지는 점에서 강화 학습 (reinforcement learning) 의 범주에 속한다고 볼 수 있다. 강화 학습은 동적프로그래밍 (dynamic programming) 기법에 기반하여 창안되기는 했지만[14], 최근 동물이 강화 학습을 하는데 중요한 역할을 하는 신경전달물질 (neurotransmitter)인 도파민(dopamine) 시스템을 이 기법을 이용하여 모델링할 수 있다는 연구 결과가 발표되고 있다.[5,6,16]

앞서 기술한 [인식 → 판단 → 행동 → 평가 → 수정] 의 단계를 자가 조직화와 강화 학습으로 명확하게 구분할 수는 없고 각각의 단계에서 자가 조직화와 강화 학습이 동시에 기여한다고 보는 것이 옳다. 예를 들어 설명하기 위해 쥐가 실험 우리 (cage) 안에서 특정한 색깔의 전등이 켜졌을 때 레버를 누르면 치즈라는 보상 (reward)을 받게 되고 전등이 켜지지 않았을 때 레버를 누르면 전기 충격이라는 처벌 (punishment)을 받게 되는 상황을 가정해 보자. 실험의 초기 단계에서 레버와 전등은 쥐에게 아무 의미가 없고 그들에 대한 지식 (knowledge)이 없다고 볼 수 있다. 시간이 지남에 따라 신경 세포의 자가조직화에 의해 레버와 전등을 볼 때 신경 세포의 발화 패턴이 일정하게 수렴된다. 동시에 우연한 행동의 결과로서 보상이나 처벌을 받으면 그 우연한 행동은 또다시 자가 조직화가 이루어지면서 전등을 보았을 때 발화하는 신경 세포의 발화 패턴과 우연한 행동이 발생할 때 발화하는 신경 세포의 발화 패턴이 강화 학습에 의해 상관관계가 강화 되거나 약화된다. 이러한 관점에서 볼 때 자가 조직화와 강화 학습은 동물이 학습하는 각 과정에서 항상 동시에 작용할 것으로 보인다.

'과연 자가 조직화와 강화 학습만으로 지능이 발현될 수 있을까'에 대해서는 아직 명확한 해답은 없다. 그러나 자가 조직화와 강화 학습이 뇌를 모방하여 구현할 인공 지능 알고리즘의 필수 요소라는 것은 분명하다.

3. STDP (Spike-Timing Dependent Plasticity)

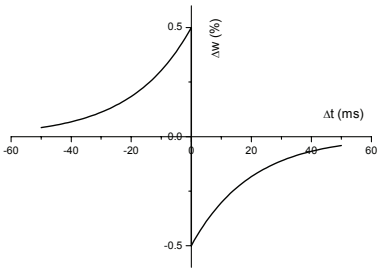
헵의 학습 (Hebbian learning)은 기존의 인공 신경망에서도 주요하게 다루고 있는 학습 방법이다. 헵의 학습은 가장 단순한 형태의 학습 법칙으로 수식적으로도 간단하며 선형 시스템으로 이루어진 인공 신경망의 경우 주성분 분석 (principal component analysis)로 잘 체계화되어 있기도 하다. 헵의 학습은 간단히 '신경 세포 A가 신경 세포 B를 자극시키는 데 충분하고, 지속적인 원인으로 작용한다면, 신경 세포 B에 대한 신경 세포 A의 작용력이 내부적으로 변화를 일으켜 증가하게 된다.' 라고 정의될 수 있고 이는 1949년에 헵 (Hebb)에 의해 처음 발표되었다.[9] 그러나 계측 기술이 발달함에 따라 실제 동물 두뇌에서는 좀 더 복잡한 메커니즘 (STDP)이 존재한다는 사실이 밝혀졌다. '신경 세포 A가 신경 세포 B에 영향을 주는 연접 (synapse)에서 짧은 시간 창문 (time window)에서 볼 때 A가 먼저 발화하고 B가 나중에 발화하면 연접 효율 (synaptic efficacy)는 증가하고, 반대로 B가 먼저 발화하고 A가 나중에 발화하면 연접 효율은 감소한다.'라는 것이다.[2]

신경생리학적 실험을 통해 얻어진 신경 세포 발화의 시간차와 연접 효율의 변화량 사이의 관계는 식 (1)과 같은 불연속 지수함수로 표현되고 그림 1과 같이 도식화할 수 있다.[13]

$$\Delta w = \begin{cases} A_+ \exp(\Delta t / \tau_+) & \text{if } \Delta t < 0 \\ -A_+ \exp(-\Delta t / \tau_+) & \text{if } \Delta t < 0 \end{cases} \qquad (1)$$

A_+, A_-, τ_+, τ_- 는 양의 상수

이는 두뇌가 가지고 있는 신경망이 기존의 인공 신경망에서 쓰이는 학습 규칙과는 달리 시간적 역학 (temporal dynamics)에 영향을 크게 받는다는 것을 말한다. 기존의 인공 신경망도 시간에 따른 동역학적 변화를 고려하기 위해 재귀망(recurrent network) 이라는 특수한 형태로 발전되기도 했으나, 이 경우에는 신경 세포 하나 하나의 다이나믹스(dynamics)를 고려해야만 STDP를 구현할 수 있게 된다. 최근의 신경생리학적 연구 결과에 의하면, STDP는 신경 세포막 내의 칼슘 이온의 농도에 의해 결정되게 되는데 몇몇 연구에서는 매우 복잡한 세포내 이온 유동식을 이용하여 칼슘 이온의 농도를



[그림 1] Spike-timing dependent plasticity

구하고 그에 따르는 STDP에 의한 연접 효율 변화를 구하는 공식도 발표되었다.[3,4,17]

그러나 본 연구는 생물학적 정확성을 목표로 하는 것이 아니라 동물 두뇌가 가지고 있는 중요한 기능은 구현하되 최대한 간단한 공식으로 단순화하여 보다 많은 신경 세포를 계산하는데 있기 때문에 앞서 설명한 복잡한 공식을 사용하지는 않는다. 또한 신경 세포의 펄스 형태의 발화 (spike)를 모델링하는 것은 많은 계산량을 요구하기 때문에 신경 세포의 시간당 발화수 (firing rate)로 단순화한 rate-code 형태의 데이터를 처리할 수 있는 모델을 구축하였다. 본 연구에서 개발된 VTDP(variation-timing dependent plasticity)는 비교적 간단한 아날로그 값의 계산을 통해 펄스 모델인 STDP를 구현할 수 있다.

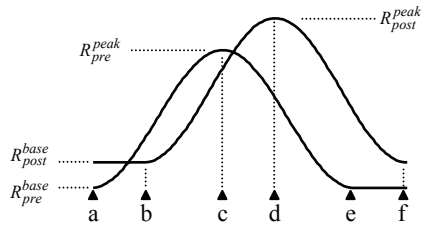
$$\frac{d}{dt}w(t) = R_{pre}(t)\cdot\frac{d}{dt}R_{post}(t) - \frac{d}{dt}R_{pre}(t)\cdot R_{post}(t) \quad (2)$$

여기서 $\frac{d}{dt}w(t)$ 는 신경 연접 효율의 시간당 변화량을, $R_{pre}(t)$ 는 연접 이전 신경 세포의 활성화 정도 (firing rate)를, $R_{post}(t)$ 는 연접 이후 신경 세포의 활성화 정도를 나타낸다.

식 (2)가 작동하는 원리를 보다 직관적으로 이해하기 위해 기하학적 해석 방법을 제안하였다.

그림 2와 같이 연접전 신경 세포와 연접후 신경 세포가 짧은 시간차를 갖고 활성화되고, 이 때 각각의 신경 세포의 활성화는 단조 증가, 단조 감소한다고 가정하자. 다음으로 연접 효율 변화량을 구하기 위해 식 (2)를 적분하면 식 (3)이 된다.

$$\int_{t_1}^{t_2}\frac{dw}{dt}dt = \int_{t_1}^{t_2}\alpha\cdot\left(R_{pre}(t)\cdot\frac{dR_{post}(t)}{dt} - \frac{dR_{pre}(t)}{dt}\cdot R_{post}(t)\right)dt \quad (3)$$



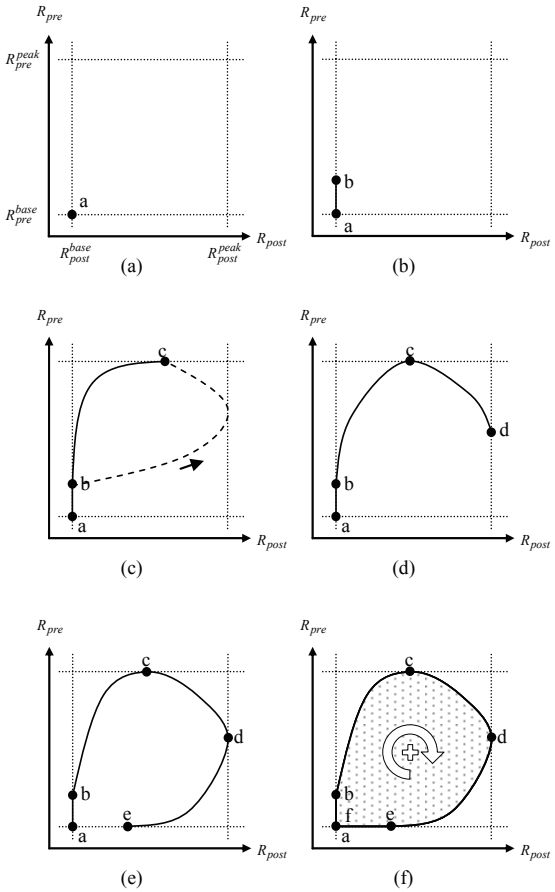
[그림 2] 연접전 신경 세포가 연접후 신경 세포보다 먼저 활성화되는 예

이 때 R_{pre} 와 R_{post} 를 서로간의 함수로 나타낼 수 있다고 가정하면, 치환 적분법을 적용하여 식 (4)와 같이 표현 가능하다.

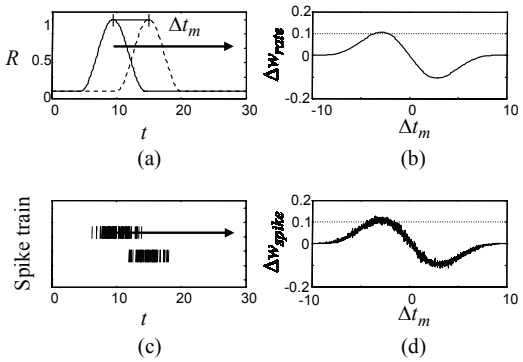
$$\int_{t_1}^{t_2}\frac{dw}{dt}dt = \alpha\int_{t_1}^{t_2}R_{pre}\left(R_{post}(t)\right)\frac{dR_{post}(t)}{dt}dt - \alpha\int_{t_1}^{t_2}R_{post}\left(R_{pre}(t)\right)\frac{dR_{pre}}{dt}dt \quad (4)$$
$$\int_{t_1}^{t_2}\frac{dw}{dt}dt = \alpha\int_{C_{post}}R_{pre}\left(R_{post}\right)dR_{post} - \alpha\int_{C_{pre}}R_{post}\left(R_{pre}\right)dR_{pre}$$

식 (4)의 첫째항 $\int_{C_{post}}R_{pre}\left(R_{post}\right)dR_{post}$ 을 도식화 하면 그림 3과 같이 그릴 수 있다. 처음에 연접전 신경세포가 먼저 활성화된다고 가정하였으므로 $\left(R_{post}, R_{pre}\right)$ 로 나타나는 곡선은 반드시 위쪽 경계선을 먼저 만난 후에 오른쪽 경계선을 만나게 된다. 그러므로 연접전 신경세포가 먼저 활성화 될 경우, $\left(R_{post}, R_{pre}\right)$ 가 만드는 폐곡선에 의해 얻어지는 식 (4)의 첫째항의 적분값은 반드시 양수가 된다. 마찬가지 방법으로 식 (4)의 두번째 항 $\int_{C_{post}}R_{post}\left(R_{pre}\right)dR_{pre}$ 의 적분값은 항상 음이 되고, 마이너스 부호에 의해 식 (4) 전체의 적분값은 항상 양의 값을 갖게 된다. 반대로 연접후 신경세포가 먼저 활성화되면 마찬가지 방법으로 연접 효율 변화량이 항상 음의 값을 가짐을 보일 수 있다.

개발된 VTDP가 실제로 STDP와 비슷한 거동을 보이는지 확인하기 위해 포아송 과정 (Poisson Process)으로 표현된 신경세포 발화 패턴 (spike train)을 이용하여 두 알고리즘을 비교하였다. 그림 4a에서 실선은 연접전 신경 세포의 활성화 정도를, 점선은 연접후 신경 세포의 활성화 정도를 나타낸다. 이 때 연접전 신경 세포가 발화하는 시점을 변화시켜가면서 식 (2)에 의해 얻어지는 연접 효율 변화량을 구하면 그림 4b와 같이 얻어진다. 그림 4c는 같은 신경 세포 활성화 정도를 이용하여 신경 세포가 발화하는 패턴을 포아송 과정으로 표현한 것이고, 이를 식 (1)을 이용하여 연접 효율 변화량을 구하



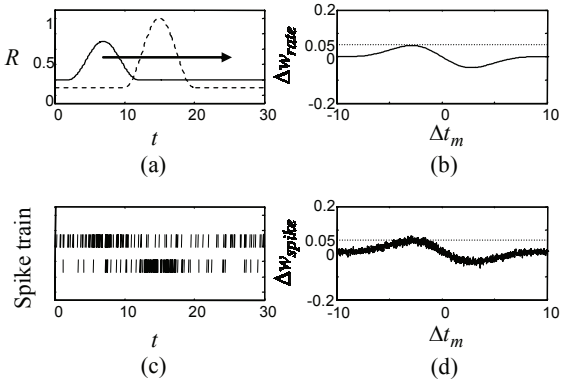
[그림 3] (R_{post}, R_{pre})에 의한 연접 효율 변화량



[그림 4] STDP와 VTDP의 비교 1

면 그림 4d를 얻을 수 있다. 그림 5는 마찬가지 방법으로 신경 세포 활성화 정도를 변화시켜 얻은 그림이다.

이 두 결과를 살펴 보면, 제안한 VTDP 알고리즘과 STDP가 유사한 결과를 보임을 알 수 있다. VTDP 알고리즘은 rate-code의 속도 면에서의 장점과, 실제 두뇌에서 일어나는 시간적 요소를 포함한 학습 방법 (STDP)을 동시에 구현 가능하다는 점에서 큰 장점이 있다.



[그림 5] STDP와 VTDP의 비교 2

4. 펄스 형태 인공신경망을 위한 강화 학습법

앞서서 동물의 뇌에서는 자가 조직화와 강화 학습이 동시에 발생한다고 설명했다. 강화 학습을 위해 동물의 뇌는 도파민이라는 신경전달물질을 이용한다. 도파민은 흑색질(substantia nigra), 배쪽 피개부(ventral tegmental area)의 신경 세포가 발화할 때 그 신경 세포가 연접하고 있는 기저핵(basal ganglia), 줄무늬핵(striatum), 그리고 전두피질(frontal cortex)의 뇌의 광범위한 영역으로 전파된다. 이렇게 전파된 도파민은 보상 신호로 작용하여 해당 신경 세포들 사이의 연접 효율을 변화시키는 역할을 한다.

도파민에 의해 연접 효율을 변화시킬 때는 도파민이 분비가 된 영역의 모든 연접 효율을 변화시키는 것이 아니라 어떤 특정한 연접만을 변화시켜야 한다. 본 연구에서는 어떤 신경 연접들을 도파민 보상 신호에 의해 변화시켜야 하고 어떤 방식으로 변화할 지에 대한 모델을 제시하고 있다. 본 연구에서 제안된 알고리즘에서는 보상 신호에 의해 변화하는 연접을 적격 연접 (eligible synapse)로 정의 하고 적격 연접을 다음과 같은 방법을 이용하여 정의하였다.

"하나의 연접에 대해 연접 이전 신경 세포가 발화한 후 연접 이후 신경 세포가 발화할 경우 그 연접은 짧은 시간 동안 적격 연접으로 분류되고 그렇지 않은 경우, 예를 들어, 연접 이후 신경 세포가 먼저 발화한 이후 연접 이전 신경 세포가 발화하거나 할 경우엔 그 연접은 적격 연접으로 분류되지 않는다."

본 연구에서는 이러한 알고리즘을 펄스 형태 신경망 (spiking neural network) 구조의 수식으로 표현하고 시뮬레이션을 통하여 입증한 바 있다.

$$\frac{dPSI(t)}{dt} = \alpha \cdot (1 - PSI(t)) \cdot \delta_{pre}(t) - \frac{PSI(t)}{\tau_{PSI}} \quad (5)$$

$$\frac{dPPSC(t)}{dt} = \beta \cdot PSI(t) \cdot (1 - PPSC(t)) \cdot \delta_{post}(t) - \frac{PPSC(t)}{\tau_{PPSC}} \quad (6)$$

where

$\delta_{pre}(t) = \sum_{i=1}^{\infty} \delta(t - (\mathrm{i_{th}}\ \text{presynaptic spike time}))$,

$\delta_{post}(t) = \sum_{i=1}^{\infty} \delta(t - (\mathrm{i_{th}}\ \text{postsynaptic spike time}))$,

$\delta(t)$ is Dirac delta function,

α and β are constants,

τ_{PSI} and τ_{PPSC} are time constants.

PSI 는 presynaptic spike indicator의 약자로서 연접 이전 신경 세포가 발화하면 증가하고 시간이 지남에 따라 단조 감소하는 변수이고, PPSC는 pre- and postsynaptic spike correlator의 약자로서 연접 이전 신경 세포가 발화한 후 일정한 시간 이내에 연접 이후 신경 세포가 발화할 때 증가하고 시간이 지남에 따라 단조 감소하는 변수이다.

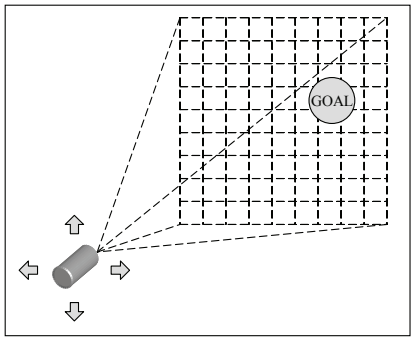
이 알고리즘은 논리적 인과관계로부터 도출되었으나, 최근의 도파민 관련 신경생리학 연구에서 이를 뒷받침할 수 있는 결과가 발표되었다. 연접 내의 칼슘 이온의 농도는 연접 이전 신경 세포가 발화하고 짧은 시간 안에 연접 이후의 신경 세포가 발화할 때 높아지게 되는데 이 칼슘 이온의 농도가 도파민이 어떤 신경 연접 효율을 변화하는데 필요 조건이 된다. [11,15]

이렇게 얻어진 PPSC와 주어진 보상 신호 D와의 상호작용을 통해 최종적으로 연접 효율 w의 증가분이 된다.

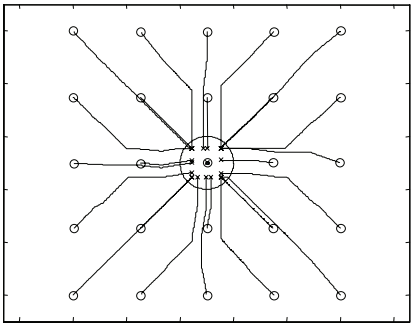
$$\frac{dw}{dt} = \gamma \cdot D \cdot PPSC(t) \quad (7)$$

본 연구에서는 제안된 알고리즘을 이용하여 간단한 시뮬레이션을 통하여 강화 학습이 이루어지는지 살펴보았다. 시뮬레이션 환경은 다음과 같다.

그림 6과 같이 하나의 비전 센서와 2자유도의 움직임을 갖는 로봇이 있을 때 학습 되지 않은 로봇이 무작위적으로 움직여서 로봇이 목적지에 다다르게 되면 보상 신호가 주어지고, 목적지로부터 멀어지면 처벌 신호가 주어지도록 하여 앞서 설명한 알고리즘으로 강화 학습을 시키면, 그림 7의 행동 패턴이 나타난다.



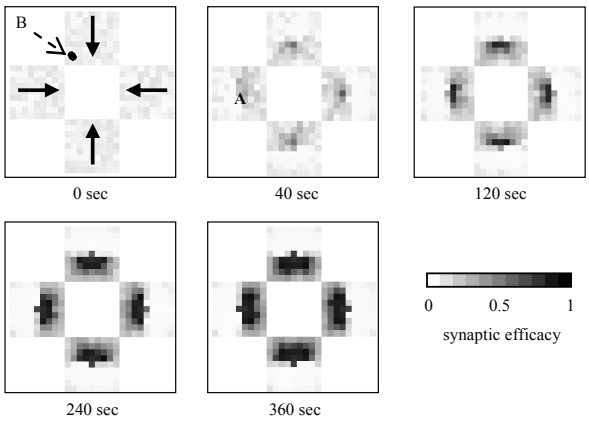
[그림 6] 강화 학습 알고리즘을 위한 시뮬레이션 환경



[그림 7] 개발된 강화 학습 알고리즘의 시뮬레이션 결과

여기서 가운데의 큰 원은 목적지이고 평면에 고루 분포되어 있는 작은 원은 로봇의 출발지이다. 로봇은 작업 영역 내의 어디에서 출발해도 목적지를 찾아 움직이는 행동을 학습한 결과를 보여주고 있다.

그림 8은 시간에 따른 로봇의 신경 세포들 간의 연접 효율의 변화를 보이고 있다. 각 사각형 안의 네 개의 사각형 형태의 패턴은 각각 비전 센서에 연결된 신경 세포와 운동 신경 세포 사이의 연접 효율을 나타낸다. 예를 들어, 점 B의 밝기는 (8,2) 위치의 비전 센서에 연결된 신경 세포가 아래쪽으로 움직이는 운동 신경 세포와 연결된 연접 효율을 나타낸다. 검은색에 가까울수록 연접 효율은 높다. 강화 학습이 진행될수록 연접 효율은 목적지를 찾는데 더욱 효율적이 될 수 있도록 변화하고 있다. 예를 들어, A 영역의 연접 효율은 오른쪽으로 움직이는 운동 신경 세포와 연결된 연접 효율을 뜻하므로 오른쪽 부분이 어두워져야 하고, 학습이 진행될수록 오른쪽 부분이 점점 어두워지는 것을 확인할 수 있다. 그러므로 위의 시뮬레이션은 제안된 알고리즘에 의해 펄스 형태의 신경망 (spiking neural network)에서 강화학습이 잘 구현됨을 보여준다.



[그림 8] 강화 학습 시뮬레이션을 진행하는 동안의 연접 효율의 변화

현재 본 연구에서는 구현된 펄스 형태의 신경망용 (spiking neural network) 강화 학습 알고리즘을 개량하여 rate-code 기반의 VTDP (variation-timing dependent plasticity) 와 호환될 수 있는 알고리즘을 식 (8)과 같이 구축한 바 있다. 여기서 E 는 식 (6)의 PPSC의 역할을 하게 된다.

$$\frac{dE(t)}{dt} = \left\{ \frac{dw_{VTDP}}{dt} \cdot \mathrm{sgn}(w_{VTDP}) \right\} - \frac{E(t)}{\tau_E}$$
$$\frac{dw}{dt} = \alpha \frac{dw_{VTDP}}{dt} + \gamma \cdot E(t) \cdot D(t) \qquad (8)$$

where

$$\{x\} = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

$$\frac{dw_{VTDP}(t)}{dt} = R_{pre}(t) \cdot \frac{dR_{post}(t)}{dt} - \frac{dR_{pre}(t)}{dt} \cdot R_{post}(t)$$

5. 도파민 보상 예측 모델

앞서 설명한 바와 같이 도파민은 동물 두뇌에서의 강화 학습의 보상 신호로 작용하는 신경 전달 물질이다. 도파민을 분비하는 신경 세포군이 갖는 또 다른 중요한 기능은 미래에 얻게 될 보상을 미리 예측하는 기능이다. 실제 동물 실험에서 도파민 시스템이 미래의 보상 신호를 예측할 뿐만 아니라 보상이 주어질 시간도 예측할 수 있다는 연구 결과가 보고되었다.[8,10,12] 미래의 보상 신호를 예측하는 능력은 지능을 발현하는 데 매우 중요한 역할을 하는 것으로 보인다. 당장의 보상에만 반응하기만 한다면 지능이 높다고 보기 힘들다. 예를 들어, 단기적인 처벌 (punishment)을 거쳐 보다 많은 보상을 얻게 된다면 강화 학습의 결과는 보다 많은 보상을 얻기 위해서 단기적 처벌을 감수하는 방향으로 가야 하기 때문이다. 이런 능력을 갖기 위해서는 미래에 받을 보상을 예측할 수 있는 기능이 필수적이다. 이 기능은 TD learning이라는 강화학습법도 갖고 있다. [14] 기계 학습 (machine learning)과 뇌연구를 병행하는 연구자들은 TD learning을 이용하여 도파민의 보상 예측 모델을 세우기도 했다. [12,16]

$$\frac{dMP_d}{dt} = -\frac{MP_d}{\tau_d} + C + F - S$$
$$\ddot{F} + \xi\omega_F \dot{F} + \omega_F^2 F = \sum_{j=i,R} \delta_j(t)$$
$$\ddot{S} + \xi\omega_S \dot{S} + \omega_S^2 S = \sum_{j=i,R} \delta_j(t)$$
$$\ddot{N}_i + \xi\omega_N \dot{N}_i + \omega_N^2 N_i = \delta_i(t) \qquad (9)$$
$$\frac{dw_{di}}{dt} = \left(N_i - \frac{dN_i}{dt} \right) \cdot (F - S)$$
$$\xi = 1$$
$$\omega_F > \omega_S$$

본 연구에서는 spiking neural network과 호환되는 도파민 보상 예측 모델을 개발하였다. 두 종류의 신경 세포 이온 채널이 서로 다른 적응 시간을 가지고 있을 때 빠른 적응 시간을 갖는 이온 채널에서 +이온이, 느린 적응 시간을 갖는 이온 채널에서 –이온이 유입되어 신경 세포의 막전압 (membrane potential)이 변화되는 모델을 개발하여 식 (9)와 같은 도파민 보상 예측 모델을 구축하였다. MP_d 는 도파민 뉴런의 활성도를, F 는 빠른 적응 시간을 갖는 이온 채널의 활성도를, S 는 느린 적응 시간을 갖는 이온 채널의 활성도를 나타낸다. 이 알고리즘의 자세한 유도 과정은 생략하기로 한다. 위의 도파민 보상 예측 모델을 이용한 시뮬레이션 결과를 동물 실험에서 얻어진 결과와 비교하여 유사성을 확인하였다.

그림 9는 도파민 뉴런이 보상을 예측하는 과정을 나타내는 그림이다. 먼저 그림 9의 좌측 그림은 Schultz가 원숭이의 뇌에 전극을 꽂아 자극, 보상, 그리고 도파민 뉴런의 상관관계를 얻은 실험 결과이다 [12]. 상단의 그래프는 보상을 예측하지 못한 상태에서의 도파민 뉴런

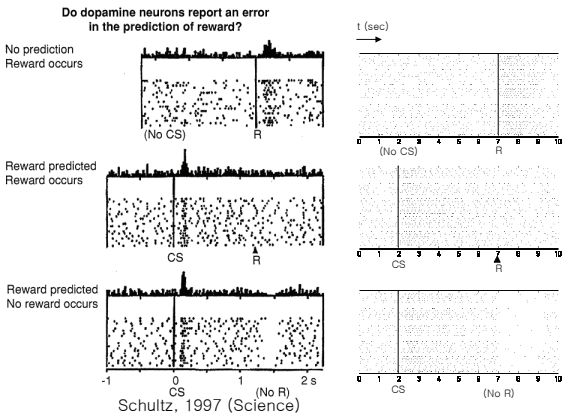
의 발화 패턴을 나타낸 것인데, 보상이 주어진 시점에서 도파민 뉴런이 발화하는 것을 볼 수 있다. 중단의 그래프는 시각 자극과 보상을 반복적으로 학습시킨 후의 도파민 뉴런의 발화 패턴을 나타낸 것인데, 도파민 시스템이 학습된 후에는 시각 자극이 주어질 때 이미 보상을 예측하기 때문에 보상이 주어지는 시간에서는 도파민 뉴런의 발화에 변화가 없게 되고, 대신 시각 자극이 주어지는 시간에 도파민 뉴런의 발화가 증가하는 것을 보여준다. 하단의 그래프는 시각 자극은 주어지나 보상이 주어지지 않으면 예상했던 보상이 주어지지 않아서 오히려 도파민 뉴런의 발화가 감소하는 것을 보여준다. 제안한 알고리즘을 이용하여 Schultz의 원숭이 실험을 시뮬레이션 한 결과 그림 9의 우측 그래프와 같이 거의 유사한 도파민 뉴런의 발화 패턴을 얻을 수 있었다.

도파민 뉴런은 보상을 예측할 뿐만 아니라 보상이 주어질 시간도 예측한다는 연구 결과가 있다. 그림 10의 좌측의 그래프는 Hollerman이 수행한 원숭이 실험 결과이다. Hollerman은 앞서 Schultz가 실험한 것과 마찬가지로 조건 자극 (conditional stimulus) 를 부여한 후 보상을 주는 상황을 반복적으로 원숭이에게 학습 시킨 후, 보상이 주어지는 시간을 좀 더 이르게, 그리고 좀 더 나중에 주는 변화를 일으켰다. 결과적으로, 보상이 좀 더 일찍 주어지면, 그 보상은 예측하지 못한 보상으로 여겨지게 되고 그 시간에 도파민 뉴런의 발화가 증가하고, 보상이 좀 더 나중에 주어지면, 원래 보상이 주어지던 시간에서는 도파민 뉴런의 발화가 감소했다가 나중에 보상이 주어지는 시간에 발화가 증가하게 된다. 이는 도파민 뉴런이 보상이 주어지는 시간도 예측한다는 결정적 증거로 여겨지고 있다 [10]. 제안한 알고리즘을 이용하여 Hollerman의 실험을 시뮬레이션한 결과 그림 10의 우측 그래프와 같이 유사한 발화 패턴을 얻을 수 있었다.

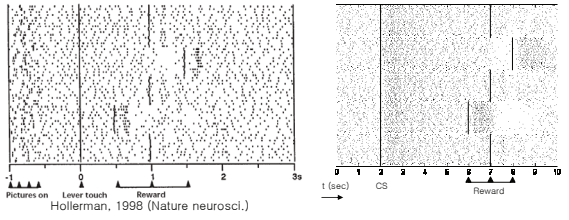
현재 본 연구에서는 spiking neural network를 위해 개발된 상기 알고리즘을 VTDP에 호환되는 모델인 DRPM (dopamine reward prediction model)을 다음과 같이 구축한 바 있다.

$$\frac{d}{dt}w_{CD}(t) = R_C(t)\cdot R_D(t) - \frac{d}{dt}R_C(t)\cdot R_D(t)$$
$$R_D(t) = \frac{d}{dt}\left(\sum_i \left(w_{C_iD}(t)\cdot R_C(t)\right) + REW(t)\right) \quad (10)$$

여기서 w_{CD} 는 일반 신경 세포가 도파민 신경 세포



[그림 9] 도파민 보상 예측 모델의 시뮬레이션 결과와 Schultz의 동물 실험 결과와의 비교



[그림 10] 도파민 보상 시간 예측 알고리즘의 시뮬레이션 결과와 Hollerman의 동물 실험 결과와의 비교

에 연결된 연접 효율이고 R_C 는 일반 신경 세포의 활성화 정도, R_D 는 도파민 신경 세포의 활성화 정도를 나타낸다. $REW(t)$ 는 보상 신호로서 도파민의 예측된 보상 신호와는 구분되는 실제의 보상을 나타낸다.

6. 결 론

본 논문은 현재 한국과학기술원 텔레로보틱스 및 제어 연구실에서 진행중인 로봇을 위한 인공 두뇌 연구에 관한, 동물 두뇌의 생물학적 기능을 분석하고 이해하여 새로운 알고리즘을 개발하는 연구의 전반적 진척 상황에 대해 기술한 논문이다. 동물 두뇌의 기능 중에서 자가 조직화와 강화 학습이 지능의 발현의 중요한 요소가 된다는 가설 아래, 동물 두뇌가 사용하는 펄스 형태의 신경망 (spiking neural network)에 호환되는 강화 학습 방법과 도파민의 보상 예측 모델을 구축하였다. 이 모델을 로봇 두뇌로 구현하기 위해서는 계산이 보다 효율적이어야 하므로 자가 조직화 모델, 강화 학습 모델, 그리고 도파민 보상 예측 모델을 각각 서로 호환되는 아날

로그 모델로 구현하였다. 향후 연구로 개발된 모델을 실제 로봇에 구현하여 알고리즘을 검사하고 개선하려 한다.

참고문헌

[1] 과학기술부, 국가기술지도작성. 2002.

[2] G.-q. Bi, M.-m. Poo, Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type, The Journal of Neuroscience, 18 10464-10472, 1998.

[3] R. J. Cormier, A. C. Greenwood, J. A. Conner, Bidirectional Synaptic Plasticity Correlated With the Magnitude of Dendritic Calcium Transients Above a Threshold, Journal of Neurophysiology, 85 399-406, 2001.

[4] J. A. Cummings, R. M. Mulkey, R. A. Nicoll, R. C. Malenka, Ca2+ Signaling Requirements for Long-Term Depression in the Hippocampus, Neuron, 16 825-833, 1996.

[5] N. D. Daw, D. S. Touretzky, Long-Term Reward Prediction in TD Models of the Dopamine System, Neural Computation, 14 2567-2583, 2002.

[6] K. Doya, Metalearning and Neuromodulation, Neural Networks, 15 495 - 506, 2002.

[7] R. C. Froemke, Y. Dan, Spike-timing-dependent synaptic modification induced by natural spike trains, Nature, 416 433-438, 2002.

[8] O. K. Hassani, H. C. Cromwell, W. Schultz, Influence of Expectation of Different Rewards on Behavior-Related Neuronal Activity in the Striatum, Journal of Neurophysiology, 85 2477-2489, 2001.

[9] D. O. Hebb, The organization of behavior. New York: Wiley. 1949.

[10] J. R. Hollerman, W. Schultz, Dopamine neurons report an error in the temporal prediction of reward during learning, Nature Neuroscience, 1 304-309, 1998.

[11] T. M. Jay, Dopamine: a potential substrate for synaptic plasticity and memory mechanisms, Progress in Neurobiology, 69 375-390, 2003.

[12] W. Schultz, P. Dayan, P. R. Montague, A Neural Substrate of Prediction and Reward, Science, 275 1593-1599, 1997.

[13] S. Song, K. D. Miller, L. F. Abbott, Competitive Hebbian learning through spike-timing-dependent synaptic plasticity, Nature Neuroscience, 3 919-926, 2000.

[14] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press. 1998.

[15] T. Suzuki, M. Miura, K.-y. Nishimura, T. Aosaki, Dopamine-Dependent Synaptic Plasticity in the Striatal Cholinergic Interneurons, The Journal of Neuroscience, 21 6492-6501, 2001.

[16] F. Wörgötter, B. Porr, Temporal Sequence Learning, Prediction, and Control - A Review of different models and their relation to biological mechanisms, Neural Computation, 17 245-319, 2005.

[17] S.-N. Yang, Y.-G. Tang, R. S. Zucker, Selective induction of LTP and LTD by postsynaptic [Ca2+]i elevation., Journal of Neurophysiology, 81 781-787, 1999.

이 규 빈

1998 한국과학기술원 기계공학과 학사

2000 한국과학기술원 기계공학과 석사

2000~현재 한국과학기술원 기계공학과 박사과정

관심분야 : Artificial Intelligence, Neuroscience, Robotics

권 동 수

1980 서울대학교 기계공학과 학사

1982 한국과학기술원 기계공학과 석사

1991 미국 Georgia Institute of Technology 기계공학과 박사

1991~1995 미국 Oak Ridge 국립 연구소 선임 연구원

1995~현재 한국과학기술원 기계공학과 조교수, 부교수, 정교수

2003~현재 인간-로봇 상호작용 핵심연구센터 소장

관심분야 : HRI, Haptics, Medical Robots