뉴로모픽 소자의 현재와 미래

박종길

1. 서론 (뉴로모픽 공학의 대두)

인간처럼 생각하고 학습하는 기계를 만들겠다는 과학자들의 노력은 수십년에 걸쳐서 다양한 형태의 노력으로 진행되어 왔다. 인간의 뇌가 학습하고 연산하는 과정을 모방하여 컴퓨터의 연산 구조 형태를 인공 신경망 형태로 모방하여 인간의 학습능력을 모사하고자 하는 연구가 진행되기도 했다. 하지만 이는 연산하는 과정을 뇌의 구조에서 영감을 얻은 구조였을 뿐 모든 단계의 연산들은 기존 폰 노이만 구조의 컴퓨터 상에서 이루어졌다. 이러한 가운데 1980년대 후반 캘리포니아 공대의 Carver Mead 교수는 뉴로모픽 공학이라는 개념을 제안한다[1,2]. 뉴로모픽 공학이란 트랜지스터의 특정 동작 바이어스 영역(subthreshold 영역)에서의 물리적 현상 및 동작 특성이 생물학적 시냅스의 동작 특성과 유사함에 착안하여 생물학적 뇌의 구조 및 기능적 특성을 전자 회로로 구현 할 수 있다는 아이디어이다. 이를 통해 연산하는 과정의 알고리즘 구조 자체만 뇌의 그것을 모방한 것이 아니라 알고리즘을 계산하고 처리하는 컴퓨터 자체도 뇌의 구조와 기능을 모사한 전자기기로 만들 수 있을 것이라는 것이다.

뇌의 동작 원리에 대한 가장 일반적인 공학적 이해는 뉴런이 정보를 처리하는 각각의 코어이며 뉴런들 사이는 시냅스로 연결되어 있어서 이를 통해 뉴런 간에 스파이크 신호(활동전위)를 주고 받아 정보를 처리한다는 것이다. 이때 각각의 시냅스의 강도는 뉴런에 전달하고자 하는 정보에 따라 세기가 정해진다. 시냅스가 가지는 가소성을 이용하여 학습이라는 과정을 통해 시냅스의 강도를 새롭게 전달되는 정보에 맞춰 조절해 나가기도 한다. 평균적으로 1.3~1.4kg 정도로 알려진 성인의 뇌에는 약 10^{11} 개의 뉴런이 존재하며 하나의 뉴런은 평균적으로 $10^3 \sim 10^4$ 개의 시냅스를 통해 다른 뉴런들과 연결이 되어 있다. 뇌는 인지, 학습, 판단 등의 고차원적인 기능들을 동시에 병렬적으로 처리하는데 이를 위해 약 20W의 에너지를 소비하는 것으로 알려져 있으며 이는 시냅스의 활동이 평균 2Hz라고 생각할 때 시냅스 연산 한번에 소비되는 에너지는 10fJ에 불과함을 알 수 있다.

이와 같이 뉴로모픽 공학은 인간의 뇌처럼 에너지를 적게 소비하면서 단순 사칙연산 보다는 고차원적 기능을 수행할 수 있는 전자기기를 구현하고자 하는 연구분야이다. 이를 위해 현재 다양한 관점에서 연구가 진행중이다. 뉴런의 하드웨어적인 구현과 이들을 하나의 시스템상에서 동작하도록 하는 시스템 레벨에서의 연구가 진행중이며, 이러한 하드웨어 상에서 인지, 판단 등의 기능을 수행하기 위한 스파이킹 신경망 기반 알고리즘에 대한 연구가 함께 진행중이다. 또한 기존의 CMOS 기반의 하드웨어 구현에서 벗어나 시냅스의 동작 특성을 모방한 새로운

〈저자 약력〉

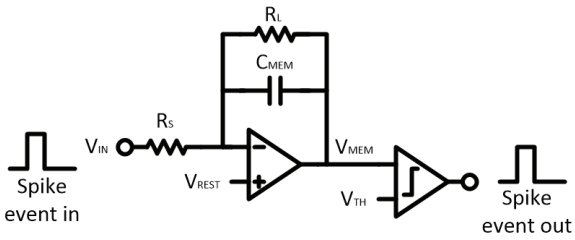
박종길 박사는 2014년 미국 University of California, San Diego에서 전자공학 박사학위를 받았으며, 2014년부터 현재까지 한국전자통신연구원에 재직중이다. 뉴로모픽 시스템 설계 및 스파이크 기반 온라인 학습법을 이용한 신경망 학습 구현등을 활발히 연구 중이다. (jkpark@etri.re.kr)

소자를 개발 하고자 하는 연구도 진행중이다. 본문에서는 뉴로모픽 시스템이 발전되어 온 방향을 되돌아 보며 향후에 필요로 하는 뉴로모픽 시스템과 그에 필요한 시냅스 소자의 특성이 무엇인지 간략한 소개를 담도록 한다.

2. 뉴로모픽 시스템의 발전 방향

초창기 뉴로모픽 공학은 CMOS 반도체 공정을 이용하여 생물학적 뉴런의 다양한 역학모델들의 구현 가능성을 보여주는 형태로 연구가 진행되었다. 뉴런의 하드웨어 구현은 뉴로모픽 공학의 핵심 중 하나인데 생물학적 뉴런 행동을 더 자세하게 묘사하는 것에 초점을 맞추는 방법과 뉴런이 네트워크 상에서 수행하는 공학적인 기능에 더 초점을 맞추는 방법이 존재한다. Hodgkin–Huxley 뉴런 모델은 뉴런의 막전위 (membrane voltage)에 따라 변화하는 이온 채널을 전기적 컨덕턴스로 표현하여 여러 이온 채널들의 상호작용으로 생기는 동적 변화를 미분 방적식으로 표현하였다. 모델의 복잡성으로 인하여 네트워크 레벨에서 시뮬레이션 하기 어려운 점이 있으며 회로로 구성할 때 구성 요소들의 복잡도가 높아 여러 개의 뉴런을 하나의 칩에 집적 시키는 것에 어려움이 존재한다. 하지만 생물학적 뉴런의 실제 동작 특성과 비슷한 뉴런의 다양한 활동전위 형태 및 패턴을 모사할 수 있다는 장점이 존재한다. Izhikevich 뉴런 모델은 생물학적 뉴런의 실제 동작 특성을 모사 하면서도 큰 규모의 신경망 모델을 컴퓨터 상에서 효과적으로 시뮬레이션 할 수 있는 수학적 모델에 초점을 맞추었다. 이 또한 수식 자체는 미분방정식의 형태를 띄지만 이온 채널의 현상 등 생물학적인 특성에 근거를 두지않고 미분방적식의 답이 뉴런의 활동 전위의 패턴과 유사해 질 수 있게 하는 매개 변수를 찾아서 만든 형태이다. 회로로 이를 구성하기 위해서는 미분 방정식의 상태 변수를 저장할 수 있는 여분의 커패시터가 추가로 필요한 점과 미분방정식의 매개 변수의 변화의 따른 결과의 민감성이 너무 높아서 트랜지스터 부정합 문제로 인해 구현하기 어렵다는 점 때문에 큰 규모로 뉴런을 집적하기에는 어려움이 따른다.

위에 언급한 두개의 모델과는 다르게 뉴런의 가장 기본적인 공학적 원리를 압축한 뉴런 모델로 Integrate–and–fire 뉴런 모델이 존재한다. 이는 뉴런이 시냅스로 받는 신호에 따라 전하를 적분 (integrate)하고 이로 인해



$$C_{MEM}\frac{dV_{MEM}}{dt} = G_S(V_{IN} - V_{REST}) - G_L(V_{MEM} - V_{REST})$$

[Fig. 1] Integrate-and-fire 뉴런 모델 블록 다이어그램과 수학적 모델

막전위 값이 특정 문턱 전압을 넘는 순간 활동전위를 발현 (fire) 하는 동작을 모사하고 있다. 이 뉴런 모델은 다양한 뉴런의 활동 전위 형태와 패턴을 모사하지는 못한다. 하지만 신경망 구조의 연산 등에 필요한 공학적인 모델로는 충분한 정도의 뉴런 역학 모델로 받아들여지고 있다. 이로 인해 큰 규모로 뉴런을 하나의 시스템에 집적하기 위한 모델로 많이 사용된다. 그림1은 Integrate–and–fire 뉴런을 구현하는 하나의 블록 다이어그램과 수식으로 표현된 동적 모델을 보여준다.

뉴런의 개별 모델을 구현 하는 단계를 넘어서 신경망 모델을 구현할 수 있는 하드웨어를 제작하기 위한 연구들로 발전하게 된다. 뉴런을 어레이 형태로 하나의 칩에 집적하고 뉴런들 간의 주고받는 스파이크 정보를 처리할 수 있는 하드웨어를 만든다. 이를 이용하여 다양한 스파이크 기반 신경망 (Spiking neural network) 알고리즘의 동작을 보여줌으로써 하드웨어를 검증한다. 이러한 하드웨어 구조상에서 개별 시냅스 강도는 외부 메모리에 값이 저장되어 있는 형태로 구현된다. 뉴런의 스파이크 발현에 의해 시냅스가 다른 뉴런으로 전달 되어야 할 때는 외부 메모리에 저장되어 있는 시냅스 강도가 이벤트 형태로 연결되어야 하는 뉴런에 전달 된다. 이 값이 뉴런의 막전위 값을 증가 시키고 특정 문턱 값을 넘어서면 새로운 스파이크를 발현하게 된다. 시냅스 학습은 뉴런들이 발현하는 스파이크의 시간적 상관관계에 따라 시냅스 강도를 강화 시키거나 약화 시키는 방향으로 진행되게 된다. 이를 하드웨어 상에서 구현하기 위해서는 추가적인 회로 구성이 요구 된다. 따라서 소규모의 뉴런 어레이 수준에서 학습의 발현 원리 등을 하드웨어 적으로 구현 가능함을 검증하는 수준에서 시냅스 학습에 대한 연구가 이루어졌다.

인간의 뇌에 있는 뉴런의 개수에 맞추려는 생물학적 동

기도 존재하지만 특정 기능을 수행하기 위해 제안된 뉴로모픽 알고리즘을 동작시키기 위해서 시스템에 집적해야 하는 뉴런의 개수가 늘어나게 되었다. 이를 위해 칩에 집적하는 뉴런의 개수를 늘리기 위한 회로 설계 측면에서의 연구가 진행되었다. 하지만 물리적 공간의 한계로 한 개의 칩에 집적시키는 뉴런의 개수는 제한적일 수 밖에 없고 이러한 칩 여러 개를 하나의 시스템에 집적하고 개별 칩에 존재하는 뉴런 끼리 스파이크를 주고 받을 수 있는 시스템 제작 관점에서의 연구로 발전하게 된다[3]. 최근에는 하나의 시스템에 뉴런을 백만개 단위로 집적한 시스템들이 발표되었다. 이러한 시스템은 스파이크 기반 신경망 알고리즘을 실시간으로 동작 시킬 수 있는 능력은 구현하였다는 것에 의미가 있다. 하지만 신경망의 학습은 GPU등의 기존 컴퓨터를 이용하여 오프라인 상에서 선행되어야 한다는 문제점이 존재한다. 이는 실시간 학습을 구현하기 위해 필요한 추가적인 회로 구성부의 복잡성과 학습 규칙에 대한 이해의 부족에서 문제점이 기인한다. 이로 인해 가장 최근에는 시스템 설계 관점에서 온라인 학습이 가능한 칩을 설계하는 방향으로 연구가 진행되고 있다.

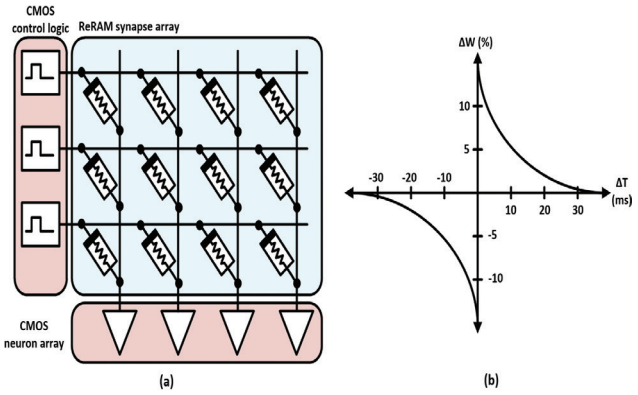
3. 뉴로모픽 소자의 대두

시냅스를 학습 시키는 기능을 하드웨어로 모사하는 것은 매우 중요한 일이다. 하지만 앞서 말한 바와 같이 시스템의 규모가 커질 수록 이를 하드웨어에 구현하는데 구조적인 어려움이 따른다. 또한 시스템의 규모가 커지면서 시냅스의 강도를 표현하기 위해 필요한 메모리의 용량도 많아지면서 발생하는 문제점도 있다. 시스템 구성의 구조적인 변화에서 이들을 해결하고자 하는 노력도 수반되고 있지만 시냅스를 단일 소자로 구현하여 문제점들을 해결하고자 하는 노력들이 함께 진행되고 있다.

시냅스의 특성을 모방한 소자는 비휘발성 단일 소자이면서 여러 단계의 시냅스 강도를 표현할 수 있어야 하고 시냅스 학습을 구현하기에 용이 하여야 한다. 대표적으로 최근에 활발히 연구되는 소자로는 멤리스터(memristor) 속성을 띠는 소자들이 있다. 멤리스터는 메모리 (memory)와 레지스터 (resistor)의 합성어로 저항의 특성을 띄는 소자가 저항 값이 일정하지 않고 양단에 인가되는 특정 전압 펄스에 따라 저항 값이 변화하며 일정 시간 이를 저장하는 메모리 역할을 한다고 하여 붙여진 이름이다. 멤리스터는 1971년 Chua교수에 의해 물리학의 이론적 모델로 전하(Charge)와 자기 유동(Magntic flux)과의 관계가 비선형적 관계를 보이는 제4의 소자(레지스터, 인덕터, 커패시터 이외의 소자)가 존재할 것으로 예측하는 것에서 시작되었다. 이론적인 소자의 구현은 아직 이루어 지지 않았지만 최근 들어 멤리스터 특성을 정성적으로 보여주는 소자들이 다양한 물질들을 활용한 형태로 구현되어 발표되고 있다. 대표적으로 멤리스터 특성을 보이는 소자로는 ReRAM (Resistive RAM), PCRAM (Phase Change RAM), FeRAM (Ferroelectric RAM)등이 있다[4].

이들은 외부에서 인가하는 전압 펄스에 따라 저항 값이 변하는 특성을 이용하여 시냅스의 학습을 구현한다. ReRAM은 아날로그 저항 변화 특성을 이용하는 메모리 형태를 통칭한다. PCRAM은 물질의 결정질의 상태간 변화에 따른 저항 특성의 변화를 이용한 비휘발성 메모리 소자이며, FeRAM은 강유전막의 계면 상태의 변화에 따라 생기는 저항 변화를 이용한다. 대부분의 소자는 2단자 소자로 수직 크로스바 형태의 시냅스 어레이 구조로 구성할 수 있다. 크로스바 형태의 시냅스 어레이 구조는 시냅스의 집적도를 높일 수는 있지만 Sneak-path를 통한 누설 전류의 문제가 생길 수 있는 단점이 존재한다. 이 때문에 스위치를 추가한 3단자 형태로 만들어 층을 쌓기도 하는데 스위치를 추가 함으로써 소자의 크기가 커지는 단점이 있다.

멤리스터 특성으로 기대되는 점들과 다르게 실제 시스템상에서 사용하기까지는 멤리스터 특성 소자가 해결되어야 할 기술적 문제들이 존재한다. 대부분의 멤리스터 특성 소자의 경우 필라멘트 형성이 나타나는데 필라멘트가 형성되는 위치나 길이들을 제어하는 방법이 개발되지 않아 멤리스터의 특성을 일정하게 조절하는데 어려움이 있다. 소자의 면적을 작게 만들면 시냅스 어레이의 집적도도 높일 수 있고 필라멘트 형성 가능성이 작아지는 장점이 있지만 특성의 부정합 (mismatch)문제가 커지는 단점을 가진다. 또한 멤리스터 특성 소자는 정체성(retention) 문제를 가지고 있다. 아직 저항 값을 긴 시간 유지 하는 능력이 부족하다. 이러한 문제는 시냅스의 단기적응효과(short-term plasticity) 구현은 가능할 수 있지만 장기적응효과(long-term plasticity) 구현에는 어



[Fig. 1] (a) 하이브리드 뉴로모픽 시스템 블록 다이어그램. CMOS 기반 뉴런 회로 위에 적층된 ReRAM 시냅스 어레이. (b) 시냅스 어레이를 통해 구현 하고자 하는 시냅스 학습 방법의 도식적 표현. Spike timing-dependent plasticity의 동작 특성을 보여준다.

려움이 있다. 또한 양단에 인가하는 일정 전압 패턴에 따른 저항의 변화량이 선형적이어야 시스템 상에서 사용이 용이한데 선형성 (lineary)을 가지지 못하고 일정 구간에서는 비선형적인 특성을 가지게 됨으로 사용에 어려움이 따른다.

멤리스터 특성 소자는 해결해야 하는 기술적인 장벽들이 존재 한다. 하지만 문제가 해결되어 실제 구현되면 가질 수 있는 뉴로모픽 공학 전반에 의미가 크기 때문에 기존 뉴로모픽 시스템의 발전과 함께 필요성이 요구되는 분야이다. 시냅스를 모방한 소자를 이용하면 시냅스의 강도를 표현하기 위해 사용하였던 메모리의 물리적 공간을 줄일 수 있다는 장점이 있다. 여러 단계의 비트 정보를 단일 소자의 아날로그 값으로 나타낼 수 있기 때문에 단일 비트를 표현하기 위해 커패시터를 사용하는 DRAM이나 여러 개의 트랜지스터를 사용하는 SRAM에 비해 집적도를 획기적으로 높일 수 있다. 또한 양단에 인가하는 전압 펄스 만을 이용하여 저장된 값을 변화 시킬 수 있기 때문에 시냅스 학습을 위한 회로 구현이 단순해 지는 장점이 있다. 기존 CMOS 트랜지스터로 구현한 뉴런 어레이 위에 멤리스터 소자의 시냅스 어레이를 개별 층으로 쌓는 구조의 하이브리드 뉴로모픽 시스템 형태로 구현 가능할 것으로 예상한다 [5]. 그림2는 이러한 하이브리드 뉴로모픽 시스템의 블록 다이어그램과 구현하고자 하는 학습의 형태를 보여주고 있다. 기존의 integrate-and-fire 뉴런 및 컨트롤 로직은 CMOS로 설계하고 시냅스를 구성하는 레지스터를 멤리스터 소자의 시냅스 어레이로 대체한다. 그림1에 있는 뉴런의 기본 형태를 사용하면서 저항(R_S)를 시냅스 어레이로 대체하는 것이다. 그림 2(b)는 양단에 인가하는 펄스의 시간의 차에 따른 시냅스 학습 발현량을 도식화 한 것이다. 이는 시냅스 학습 방법 중에 하나인 spike timing-dependent plasticity에 따른 학습 발현과 유사함을 가짐을 볼 수 있다.

4. 뉴로모픽 공학의 미래

뉴로모픽 공학은 다양한 분야의 학문이 함께 연구 되어야 하는 분야이다. 생물학에서 연구되는 뇌의 학습, 기억, 그리고 인지 기능 등의 발현에 대한 이해의 노력도 필요하고 뉴로 사이언스에서 연구되는 계산 과학 분야의 이해도 접목시켜야 할 때도 있다. 또한 이를 공학적으로 구현하기 위한 뉴로모픽 시스템, 알고리즘, 소자 등 다양한 공학분야에서의 지식의 발전이 필요하다. 이러한 융합 연구를 통해 향후 인간의 뇌의 동작원리를 가깝게 모사한 인지 기능을 하는 저전력 고집적 뉴로모픽 시스템의 등장을 기대해 본다.

References

[1] Carver Mead, "Neuromorphic electronic systems," Proceedings of the IEEE, vol. 78, no. 10, pp.1629-1636, 1990
[2] Indiveri et al., "Neuromorphic silicon neuron circuits," Frontiers in Neuroscience, vol. 5, no. 73, 2011
[3] Jongkil Park et al.,"Hierarchical address event routing for reconfigurable large-scale neuromorphic systems,", IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2408-2422, 2017
[4] S. G. Hu et al.,"Review of nanostructured resistive switching memristor and its applications," Nanoscience and Nanotechnology Letters, vol. 6, pp. 729-757, 2014
[5] Carlos Zamarreno-Ramos et al.,"On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex," Frontiers in Neuroscience 5:26. doi: 10.3389/fnins.2011.00026