ETRI

2018
Electronics and
Telecommunications
Trends

최신 반도체, 하드웨어 기술 동향 특집

차세대 뉴로모픽 하드웨어 기술 동향

Next-Generation Neuromorphic Hardware Technology

문승언 (S.E. Moon, semoon@etri.re.kr) ICT 소재연구그룹 책임연구원
임종필 (J.P. Im, jpim@etri.re.kr) ICT 소재연구그룹 선임연구원
김정훈 (J.H. Kim, JeongHun@etri.re.kr) ICT 소재연구그룹 선임연구원
이재우 (J. Lee, jaewoo@etri.re.kr) ICT 소재연구그룹 책임연구원
이미영 (M.Y. Lee, sharav@etri.re.kr) 프로세서연구그룹 책임연구원
이주현 (J.H. Lee, juehyun@etri.re.kr) 프로세서연구그룹 책임연구원
강승열 (S.Y. Kang, kang2476@etri.re.kr) 유연소자연구그룹 책임연구원
황치선 (C.S. Hwang, hwang-cs@etri.re.kr) 실감디스플레이연구그룹, 책임연구원
윤성민 (S-M. Yoon, sungmin@khu.ac.kr) 경희대학교 정보전자신소재공학과 교수
김대환 (D.H. Kim, drlife@kookmin.ac.kr) 국민대학교 전자공학부 교수
민경식 (K.S. Min, mks@kookmin.ac.kr) 국민대학교 전자공학부 교수
박배호 (B.H. Park, baehpark@konkuk.ac.kr) 건국대학교 물리학과 교수

A neuromorphic hardware that mimics biological perceptions and has a path toward human-level artificial intelligence (AI) was developed. In contrast with software-based AI using a conventional Von Neumann computer architecture, neuromorphic hardware-based AI has a power-efficient operation with simultaneous memorization and calculation, which is the operation method of the human brain. For an ideal neuromorphic device similar to the human brain, many technical huddles should be overcome; for example, new materials and structures for the synapses and neurons, an ultra-high density integration process, and neuromorphic modeling should be developed, and a better biological understanding of learning, memory, and cognition of the brain should be achieved. In this paper, studies attempting to overcome the limitations of next-generation neuromorphic hardware technologies are reviewed.

* DOI: 10.22648/ETRI.2018.J.330607
* 본 연구는 과학기술정보통신부의 ETRI 충연금 사업의 지원을 받아 수행하였음.

Ⅰ. 서론

최근, 디지털 헬스케어나 스마트 도시, 자율주행 자동차 등의 다양한 서비스가 본격적으로 제공되면서 수많은 데이터를 한꺼번에 처리해야 하는 수요가 급증하게 되었다. 이를 위해 전 세계 각 지역에 수 천대의 서버를 갖춘 데이터 센터가 건립 중이며, 이로 인한 엄청난 규모의 컴퓨터와 막대한 전력을 필요로 하고 있다. 또한, 클라우드 서버에 모이는 방대한 양의 데이터 중에는 인간에게는 익숙하지만, 기계는 쉽게 인식하기 어려운 비정형적인 문자·이미지·음성·영상 등이 혼재해 있는데, 이의 정보처리 과정에서 막대한 전력이 소모되는 기존 컴퓨팅 시스템의 문제를 해결해야 한다. 따라서, 적은 전력을 사용하고도 많은 정보를 처리하는 사람의 뇌를 모방한 뉴로모픽 칩에 대한 관심이 높아지고 있다.

특히, 2016년 3월에 Google DeepMind 사는 자체 개발한 인공지능 컴퓨터인 알파고(AlphaGo)에게 바둑을 학습시켜 한국의 이세돌 9단과 대국을 진행하여 화제가 되었다. 이 대국에서 알파고는 이세돌 9단을 4:1로 압도적으로 이김으로써 사람들을 더욱 놀라게 하였는데, 주목해야 할 점은 알파고가 바둑을 둘 수 있도록 미리 프로그램된 것이 아니라, 바둑을 학습하며 스스로 진화하였다는 점이다.

알파고에서 구글이 선보인 학습을 통한 컴퓨팅은 기존의 컴퓨팅 방식과 다르다. 인간의 두뇌가 지식을 학습하고 응용하는 것과 매우 유사한, 혁신적인 소프트웨어 알고리즘을 통해 슈퍼컴퓨터인 알파고가 바둑을 학습하도록 하였다. 이를 위해 알파고는 1,202개의 중앙처리장치(CPU: Central Processing Unit)와 176개의 그래픽처리장치(GPU: Graphic Processing Unit)를 사용한 것으로 알려져 있다. 이 시스템의 전력을 단순 계산하면 170kW 정도가 필요한데, 20W 정도로 알려진 인간 두뇌의 소모전력과 비교하면 약 8,500배 정도 많은 에너지를 소모하는 것이다. 연산하는 과정을 뇌의 구조에서 영감을 얻어 소프트웨어적으로 모사하는 구조였을 뿐 모든 단계의 연산들은 기존 폰 노이만 구조의 컴퓨팅에 적합한 하드웨어 상에서 이루어졌기 때문에 방대한 시스템과 비효율적인 에너지 소모가 수반된다. 여기에서 구글이 구현한 소프트웨어 기반 인공지능의 한계점이 드러난다.

〈표 1〉 CMOS 기반 뉴로로픽 시스템과 인체 뇌와의 성능 비교

Ref.	인체 뇌	뉴로모픽 시스템 (CMOS)
속도	1msec	1nsec
크기	1~10mm	10~100nm
구동 전압	~0.1V	V_{dd}~1.0V
뉴런 밀도	$10^5/mm^2$	$5\times10^3/mm^2$
신뢰도	80%	〉99.999%
인식 에러율	75%	~0%
Fan in-out	1,000~10,000	3~4
시냅스 동작 에너지	~10fJ	~10pJ
전체 소모전력	20W	〉〉10^3W
Noise effect	Stochastic resonance	Bad

기존 폰 노이만 컴퓨팅 아키텍쳐는 인지 처리 기능을 수행 함에 있어서 인간의 뇌에 비해 효율성이 낮고, 기억, 연산, 추론, 학습 등을 동시에 수행할 수 없었다[〈표 1〉 참조], [1], [2]. 따라서 인간 뇌의 하드웨어인 뉴런과 시냅스를 모사하는 뉴로모픽 소자 기반의 컴퓨팅 아키텍처와 인공 신경망을 구현함으로써 인간의 뇌가 학습하고 연산하는 과정을 효율적으로 직접 모방하고자 하는 연구가 활발히 진행 중이다.

이러한 뉴로모픽 하드웨어는 과도기인 3~7년간 자동차 첨단운전자보조시스템(ADAS: Advanced Driver Assistance Systems), 실시간 얼굴 및 물체 인식, 실시간 문자 번역, IoT(Internet of Things) 센서 등에서 활용되고, 10년 이후부터는 지능형 로봇, 무인기, 자율주행 자동차, AI(Artificial Intelligence) 비서 등에서 폭넓게 활용될 것으로 예측된다(가트너, 테크놀로지 하이프 싸이클, 2017). 뉴로모픽 소자 세계시장 규모는 2016년

약 12억 달러 규모에서 2022년 약 48억 달러 규모로 연평균 26.3% 성장할 것으로 전망되었다(Markets and Markets, Neuromorphic Chip Market – Global Forecast to 2022, 2015).

본문에서는 차세대 뉴로모픽 시스템과 그에 필요한 기술에 대한 간략한 소개를 담도록 한다.

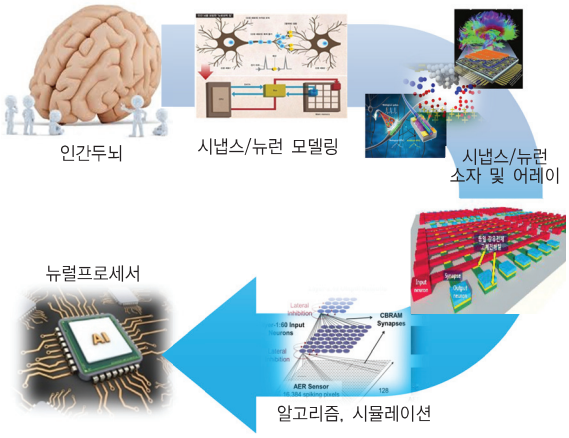
II. 기술 개요

뉴로모픽 공학이라는 개념은 1980년대 후반 Carver Mead (1990)[3] 논문에서 처음으로 제안되었다. 이 당시 두뇌에 존재하는 뉴런에서의 신호 전달 방식인 action potential이 알려지기 시작하였는데, 이를 당시의 반도체 기술로 회로를 구성하여 재현하는 것이 뉴로모픽 컴퓨팅의 시작이었다.

뇌의 동작에 대한 공학적 이해에 의하면 뉴런은 정보를 처리하는 각각의 코어이며 뉴런들 사이는 시냅스로 연결되어 있어서 이를 통해 뉴런 간에 스파이크 신호(활동전위)를 주고받아 정보를 처리한다. 이때 시냅스는 화학적, 전기적 반응을 통해 뉴런에서 발생하는 스파이크 신호를 다른 뉴런으로 전달해주는 역할을 하는데, 시냅스의 흥분과 억제를 통해 단·장기 기억 강화 또는 약화 기능을 수행하며 인접한 두 뉴런 사이에 스파이크 신호의 타이밍에 따른 가소성(STDP: Spike Timing Dependent Plasticity)을 가지고 있다.

평균적으로 2L 정도로 알려진 성인의 뇌에는 약 10^{11}개의 뉴런이 존재하며 하나의 뉴런은 평균적으로 10^{3}~10^{4}개의 시냅스를 통해 다른 뉴런들과 연결이 되어 있다. 뇌는 인지, 학습, 판단 등의 고차원적인 기능들을 동시에 병렬적으로 처리하는데 약 20W의 에너지를 소비하는 것으로 알려져 있으며 시냅스의 활동이 평균 수 Hz라고 생각할 때 시냅스 연산 한 번에 소비되는 에너지는 10fJ 내외로 추측할 수 있다.

현재 뉴로모픽 시스템을 하드웨어로 구현하는 기술은



(그림 1) 뉴로모픽 공학 연구 개념도

구현하는 소자의 관점에서, 기존의 실리콘 트랜지스터만으로 하드웨어를 구현하는 방법과 멤리스터 등의 차세대 뉴로모픽 소자와 실리콘 트랜지스터의 혼합 등을 통해 하드웨어를 구현하는 방법으로 나눌 수 있다. 뉴런과 시냅스 회로로 구성된 뉴로모픽 하드웨어가 구현되면, 다양한 뉴럴넷 알고리듬을 뉴로모픽 하드웨어에 이식하여 인간 두뇌의 다양한 인식(cognition) 기능을 모방할 수 있게 된다. 뉴널넷 알고리즘은 단순한 퍼셉트론 알고리즘, back-propagation 학습에 기반한 인공 신경망(ANN: Artificial Neural Network) 알고리즘, 스파이크 신호 기반의 스파이킹 신경망(SNN: Spiking Neural Network)등의 다양한 알고리즘이 있고 현재에도 많은 새로운 두뇌 모방의 신경망 알고리즘이 계속해서 개발되고 있다. 이와 같이 뉴로모픽 공학은 인간의 뇌처럼 에너지를 적게 소비하면서 단순 사칙연산보다는 고차원적인 인지와 지능적인 기능을 수행할 수 있는 전자기기를 구현하고자 하는 연구분야이다. 이를 위해 현재 다양한 관점에서 연구가 진행 중인데, 뉴런이나 시냅스의 하드웨어적인 구현 연구, 이들을 하나의 시스템상에서 동작하도록 하는 시스템 레벨에서의 연구, 이러한 하드웨어 상에서 인지, 판단 등의 기능을 수행하기 위한 다양한 신경망 기반 알고리즘에 대한 연구 등이 진행 중이다[(그림 1) 참조], [4].

Ⅲ. 기술 동향

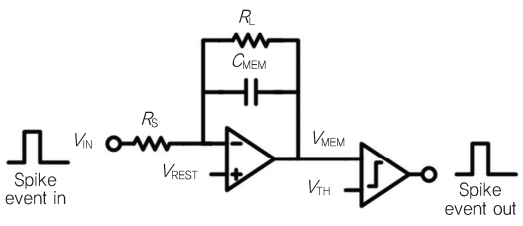
이러한 뉴로모픽 하드웨어 원천 기술을 확보하기 위해 전 세계적으로 정부 차원의 지원 하에 많은 연구가 경쟁적으로 진행되고 있는데, EU에서는 2013년도부터 향후 10년간 총 10억 유로 규모가 지원되는 미래기술 주력 사업 프로그램을 통해 인간 뇌의 수리적 모델 구축 및 컴퓨터 모사를 하는 Human Brain Project를 진행 중이다. 미국에서는 국가프로젝트로 인공지능 사업 기획 및 신개념 반도체 소자 연구를 진행 중인데, DARPA에서는 SyNAPSE(Systems of Neuromorphic Adaptive Plastic Scalable Electronics) 프로젝트를 통해 지능형반도체 관련 인간 두뇌모사형 뉴로시냅틱 칩 'TrueNorth'를 개발하여 보행자 및 차량에 부분 적용하고 있으며, Brain Initiative 프로그램을 통해 2013년부터 10년간 총 30억 달러의 예산으로 뇌의 종합적인 이해와 뇌질환 극복을 위한 뇌지도 구축 프로젝트를 진행 중이다.

뉴로모픽 하드웨어 기술 관련하여 인간의 신경회로가 어떻게 데이터를 처리하는지 회로 및 시스템 수준에서의 이해가 여전히 부족하며, 현재는 간단한 수준에서 단일 신경세포들을 모사하는 수준이다. 그래서 실제 인간의 뇌와 같이 복잡하고 집적도 높은 구조를 가지기 위해서는 기술적 측면에서 해결해야 할 문제가 많은 상황이다. 특히 현재 많은 관심을 받고 있는 기술적 이슈로는 시냅스나 뉴런용 신소재 및 소자 구조에 대한 이슈, 뉴로모픽 칩 안에서 가장 기본이 되는 단위인 '멤리스터(memristor)'를 고밀도로 집적하기 위한 반도체 집적 기술 이슈, 뉴로모픽 모델링 및 학습인지 프레임워크 기술 이슈 등이 있다.

1. 뉴런

초창기 뉴로모픽 공학은 CMOS(Complementary Metal-Oxide-Semiconductor) 반도체 공정을 이용하여 생물학적 뉴런의 다양한 역학모델들의 구현 가능성을 보여주는 형태로 연구가 진행되었다. 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 시뮬레이션과 집적화에 어려움이 있지만, 생물학적 뉴런의 실제 동작처럼 다양한 활동전위 형태를 모사할 수 있는 장점이 있다. Izhikevich 뉴런 모델은 생물학적 뉴런의 실제 동작 특성을 모사하면서도 큰 규모의 신경망 모델을 컴퓨터 상에서 효과적으로 시뮬레이션할 수 있는 수학적 모델에 초점을 맞추었다. 회로화할 때 여분의 커패시터가 추가로 필요한 점과 트랜지스터 부정합 문제로 큰 규모로 뉴런을 집적하기에는 어려움이 따른다.

위에 언급한 두 개의 모델과는 다르게 뉴런의 가장 기본적인 공학적 원리를 압축한 뉴런 모델로 Lapicque et al.(1907)[5]의 논문에서 제안된 Integrate-and-fire 뉴런 모델이 존재한다. 이는 뉴런이 시냅스로부터 받는 신호에 따라 전하를 적분(integrate)하고 이로 인해 막전위 값이 특정 문턱 전압을 넘는 순간 활동전위를 발현(fire)하는 동작을 모사하고 있다. 이 뉴런 모델은 다양한 뉴런의 활동 전위 형태와 패턴을 모사하지는 못하지만, 신경망 구조의 연산 등에 필요한 공학적인 모델로는 충분한 정도의 뉴런 역학 모델로 받아들여지고 있다. (그림 2)는 Integrate-and-fire 뉴런을 구현하는 하나의 블록 다이어그램과 수식으로 표현된 동적 모델을 보



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

(그림 2) Integrate-and-fire 뉴런 모델 블록 다이어그램과 수학적 모델

여준다.

앞서 뉴런 소자의 구현은 대부분 CMOS 기반 소자로 제작되어 왔다. 이 경우, 기존의 반도체 기술 기반으로 적용이 용이하지만, 뉴런 동작 구현을 위한 회로 구조가 매우 복잡하여 집적화에 불리한 측면도 있다. 최근 뉴런 동작을 멤리스터를 이용하여 더욱 간단한 회로 구성으로 구현하려는 시도가 이루어지고 있다.

Leon Chua에 의해 이론상 예상되었던 멤리스터를 산화물에서 최초로 실현하였던 HP의 Stanley Williams 그룹의 Pickett et al.(2013)[6], Jin et al.(2016)[7]과 Stoliar et al.(2017)[8] 논문에서는 Mott 멤리스터를 이용한 뉴런 소자의 제작 가능성을 보여주었고, 이 경우 집적도를 높인 뉴런(neurisotr) 소자가 가능함을 보고하였다. 또한, IBM Zurich 연구소의 Tuma et al.(2016)[9] 논문에서는 상변화 물질(phase change material)을 이용한 stochastic phase-change neuron 기술을 발표하였고, 주변 회로를 CMOS로 구성한 집적도가 높은 뉴런 소자를 구현하였다. Zhang et al.(2018)[10]과 Jaiswal et al.(2017)[11] 논문에서는 Ag/SiO_2/Au threshold switching 멤리스터나 강자성체의 magnetoelectric switching 특성을 기반으로 integration-and-fire 기능의 인공 뉴런 소자를 구현하였다.

한편 Wang et al.(2018)[12] 논문에서는 이러한 단위 뉴런 소자 연구에서 벗어나, 최근의 연구 방향은 시냅스 어레이와 뉴런 모두를 멤리스터 기반으로 모로리틱하게 집적한 뉴로모픽 소자 연구 등으로 기존 CMOS 기반의 소자 한계를 뛰어넘으려 하고 있는데 ETRI에서도 건국대 등과 공동 연구를 통해 유사 연구를 진행 중이다.

2. 시냅스

시냅스의 특성을 모방한 소자는 비휘발성 메모리 소자이면서 여러 단계의 시냅스 강도를 표현할 수 있어야 하고 시냅스 학습을 구현하기에 용이 하여야 한다. 대표적인 시냅스 모방 소자는 멤리스터인데, 이는 메모리(memory)와 레지스터(resistor)의 합성어이다. 저항의 특성을 띄는 소자이면서 저항 값이 일정하지 않고 양단에 인가되는 특정 전압 펄스에 따라 저항 값이 변화하며 일정 시간 이를 저장하는 메모리 역할을 한다고 하여 붙여진 이름이다. 1971년 Chua 교수가 물리학의 이론적 모델을 기반으로 전하(Charge)와 자기 선속(Magntic flux)과의 비선형적 관계를 표현하는 제4의 소자(레지스터, 인덕터, 커패시터 이외의 소자)의 존재를 예측하는 것에서 시작되었다.

최근 멤리스터 특성을 정성적으로 보여주는 소자들이 다양한 물질과 구조 들을 활용한 형태로 구현되어 발표되는데, 주요 요소로 분류하면 금속 이온 이동 기반 소자, 산소정공 이동 기반 소자, 상 변화 기반 소자, 스핀 기반 소자, 전계효과 트랜지스터(FET: Field Effect Transistor) 기반 소자, 강유전체 분극 반전 기반 소자 등으로 분류할 수 있다.

Ohno et al.(2011)[13], Wang et al.(2017)[14], Hu et al.(2017)[15] 논문에서는 고체 전해질 내부로 이동하는 금속이온의 반복적인 산화·환원을 통한 전도성 경로 형성과 파괴를 통한 저항 변화 현상을 보고하였다. 반복된 자극(펄스)의 조건에 의존하는 시냅스의 생물학적 특성인 단기 강화 장기 강화 전환 및 단기 기억 장기 기억 전환 특성 등을 확인했었지만 선택소자 필요성, 낮은 On/Off 비 등을 개선해야 한다.

Choi et al.(2017)[16], Chang et al.(2011)[17], Tan et al.(2017)[18] 논문에서는 산화물 박막 내부에 존재하는 산소 정공의 이동을 통해 형성 및 파괴되는 전도성 경로로 인한 저항 변화 현상을 보고하였다. 이를 이용한 array 소자 제작과 반복된 자극(펄스)에 의한 시냅스의 생물학적 특성인 시냅스 강화 및 시냅스 약화 그리고 반복된 자극(펄스) 횟수에 의존하는 장기기억 특성 등을 확인한 바 있으나 sneak path, abrupt set와 낮은

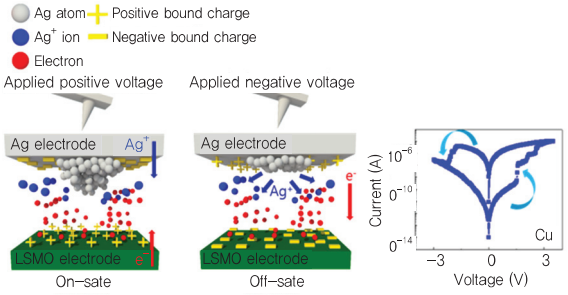
On/Off 비 등을 개선해야 한다.

Kuzum et al.(2012)[19]와 Wright et al.(2011)[20] 논문에서는 외부에서 인가된 전압에 의해 상 변화 물질($Ge_2Sb_2Te_5$) 내부에 국소적으로 발생하는 열로 결정성의 제어를 통해 저항 변화 현상(비정질 off-state, 결정질 on-state)을 보고하였다. 반복된 자극(펄스)에 의해 유도되는 점진적인 컨덕턴스 변화를 이용한 시냅스 가소성 모방 등을 확인한 바 있지만 큰 소모전력 등이 극복해야 할 과제이다.

Wang et al.(2009)[21]와 Srinivasan et al.(2016)[22] 논문에서는 스핀 토크에 의한 자기장 스위칭에 기반한 MTJ(Magnetic Tunneling Junction) 기반의 스핀트로닉 멤리스터 등을 보고하였지만 낮은 On/Off 비 특성과 외부 자기장이 필요하여 집적화에 어려움이 예상된다.

Kim et al.(2017)[23], Burgt et al.(2017)[24], Yang et al.(2017)[25]와 Shi et al.(2013)[26] 논문에서는 CNT(Carbon Nano Tube), 폴리머, 2D MoO_3, $SmNiO_3$ 등을 채널로 이용하여 전계효과 트랜지스터를 제작하고, 3-terminal 구조에서의 저항 변화 현상을 보고하였다. 제작된 소자의 게이트에 인가된 펄스의 조건에 의해 채널 컨덕턴스를 제어하여, 서로 다른 dynamic range (on/off ratio or variation margin)의 특성을 확보하고 이를 패턴 인지 시뮬레이션에 활용함으로써 패턴 인지 소자 응용 가능성 등을 확인한 바 있지만, FET 구조 특성상 공정 과정이 복잡하고 소자의 소형화에 어려움이 있다.

Kim et al.(2017)[27], Shao et al.(2016)[28], Wan et al.(2014)[29]과 Yoon et al.[30]-[32]의 논문에서는 트랜지스터의 게이트 절연막으로 각각 모바일 이온의 이동이 가능한 소재나 강유전체 박막을 적용하였다. 전자에서는 인가 전압신호의 크기와 폭 제어를 통해 트랜지스터의 전류 특성을 점진적으로 변화시키는 시냅스 모방 기능을 확인한 바 있고, 후자에서는 강유전체 박막이



(그림 3) 강유전체 분극 반전과 금속 이온 이동을 동시에 이용하는 멤리스터 소자의 동작 원리와 측정 결과

가지는 부분분극 스위칭 특성을 이용하여 외부 인가전압(자극)의 변화에 따라 FET의 드레인 전류(반응)가 점진적으로 증감하는 시냅스의 특성을 확인한 바 있지만, 계면 상태 문제와 retention 등에서 개선이 필요하다.

Chanthbouala et al.(2012)[33], Boyn et al.(2017)[34], Yoon et al.(2017)[35] 논문에서는 외부에서 인가된 반복된 자극(펄스)에 의한 강유전체 초박막 $BaTiO_3$, $BiFeO_3$ 도메인 변화와 그에 따른 터널링 전류 변화 현상을 보고하였다. 고집적 응용 가능성을 보였을 뿐만 아니라, 극단적으로 짧은 자극(펄스) 시간에 의한 효율적인 에너지 소비 특성, 반복된 자극(펄스)에 의한 시냅스 특성(시냅스 감소와 시냅스 강화) 모사 등을 확인한 바 있다.

ETRI에서는 건국대 등과 같이 강유전체 분극 반전과 금속 이온 이동을 동시에 이용하는 소자를 이용하여, progressive set/reset 구현과 선택소자로 작동하는 성상세포를 모사한 기능까지 구현하는 소자를 연구 중이고[(그림 3) 참조], 경희대 등과 같이 원내 실리콘 팹을 활용하여 CMOS 소자와의 확장성이 용이한, 강유전체 박막을 게이트 절연막으로 사용하는 3단자 소자인 FET 소자를 연구 중이다.

3. 뉴로모픽 반도체 집적 기술

반도체 집적 기술 이슈에는 고집적 뉴런-시냅스 어레이 기술, 인체의 뉴런간 연결하는 Axon에 해당하는 저

저항 연결 기술, 집적화 공정 기술 등이 포함된다.

CMOS 뉴로모픽 시스템에서 뉴런의 네트워크 연결도 향상과 메모리 운용의 한계를 극복하기 위해서 Intel의 3D XPoint memory와 유사한 구조가 제안되고 있다. CMOS 뉴런과 나노선이 crossbar 네트워크 구조로 배열되고 나노선이 겹쳐지는 부분에 시냅스 기능을 가지는 CrossNets 구조의 나노소자가 배치되는 하이브리드 반도체/나노선/분자(CMOL) 집적회로의 한 종류로 Likharev et al.(2003)[36], [37]과 Strukov et al.(2005)[38]에 의해 제안되었고, Lin et al.(2014)[39] 논문에서는 CMOS 공정으로 제작된 소자위에 나노임프린트 등의 차세대 반도체 공정 등을 활용하여 CMOL 아키텍쳐를 구현한 바 있다.

Jo et al.(2010)[40] 논문에서는 Al/PCMO 기반 cross point 구조의 ReRAM(Resistive Random Access Memory)과 계면에서의 self-formed Schottky barrier를 선택소자로 활용하여 $4F^2$ 크기의 소자로 구성된 어레이 구조를 보고하였다.

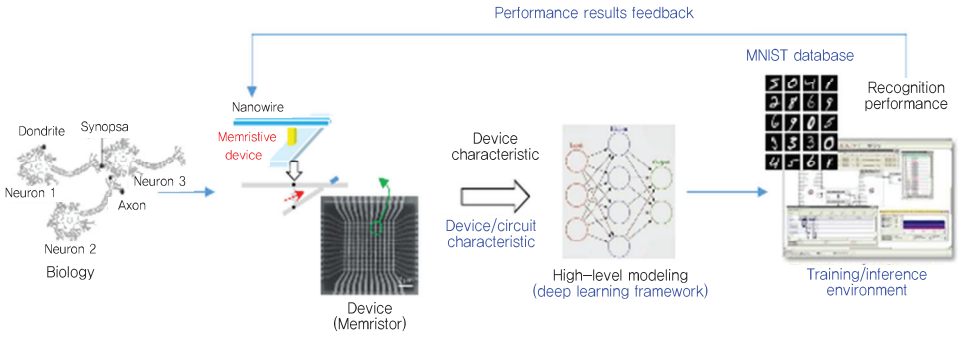
저저항 연결 기술의 대안으로는 광 뉴로모픽 소자가 연구 중이며, Lee et al.(2017)[41]와 Qin et al.(2017)[42] 논문에서는 비정질의 IGZO 기반 2-terminal 소자나 CNT-Graphene 하이브리드 소재를 채널로 이용한 전계효과 트랜지스터를 제작하고 제작된 소자에 자극에 해당하는 광 펄스를 가했을 때, 광에 의한 산소 공공의 이온화나 생성된 전자 등에 의한 저항 변화 현상을 보고하였지만, 속도가 느리며 집적화에 어려움이 예상된다.

파나소닉 (2013)[43]에서는 CMOS 디지털 회로에 강유전체 멤리스터를 적용하여 이미지를 아날로그 방식으로 처리할 수 있는 영상시스템의 개발을 발표하였는데, 소모 전력이 기존 디지털시스템에 비해 1/10 정도이고, 이 시스템을 구현하는 데 있어서 강유전체 성장기술과 디지털-아날로그 혼성 신호처리기술이 핵심이다. Moon et al.(2015)[44]와 Burr et al.(2015)[45] 논문에서는 산화물이나 상변화 박막 기반 시냅스와 뉴런으로 구성된 10k 비트 수준의 cross point array를 제작하여 필기체 인식 성능을 확인한 바 있다. ETRI에서는 CMOS 호환 반도체 공정을 이용하여 강유전체 초박막 기반 고집적 시냅스 어레이 제작 공정 기술을 개발하고 있으며, 이를 기반으로 필기체나 이미지 인식을 위한 뉴로모픽 시스템에 적용할 수 있는 초소형 뉴로모픽 하드웨어 등을 연구 중이다.

4. 뉴로모픽 모델링 및 학습인지 프레임워크

뉴로모픽 소자가 실제 인식 테스트를 통해 어느 정도의 인식 성능을 보이는지 측정하여, 적용 가능성을 판단하고 필요한 소자 변수의 설정을 위한 뉴로모픽 상위 수준 모델링 및 인지 성능 평가 기술이 필요하다. 소자의 응용분야를 선정하고 이에 적합한 학습 데이터 세트를 이용하여, 소자 특성을 반영한 상위 수준 모델링을 통한 학습과 인지 과정을 미리 수행하여 소자 연구 개발의 초기 단계에서 요구되는 사양을 정의할 필요가 있는데, ETRI에서는 2018년도부터 인공지능 소자의 학습 및 인지 성능의 평가를 위한 인공지능 소자 모델링 기반 기계학습 프레임워크를 연구 개발 중이다[(그림 4)참조].

학습 데이터 세트는 학습을 위한 샘플 데이터 DB와 정답 정보로 구성되며, 대표적인 이미지 기반 데이터 세트로 MNIST(Modified National Institute of Science and Technology database), CIFAR-10(Canadian Institute For Advanced Research database), ILSVRC (ImageNet Large Scale Visual Recognition Competition database), PASCAL VOC(Pascal Visual Object Classes database), MS COCO(Microsoft Common Objects in Context database) 등이 있다. 다른 이미지 기반 데이터 세트에 비해 작은 규모의 신경망의 성능 테스트에 적합한 MNIST 데이터 세트가 뉴로모픽 소자 성능 테스트에 많이 활용되고 있다. MNIST 데이터 세트



(그림 4) 뉴로모픽 모델링 및 인식 성능 평가 기술

는 뉴욕대의 LeCun 교수가 최초로 CNN(Convolutional Neural Networks)을 제안하면서 사용한 데이터 세트로, 28×28(784픽셀)의 해상도를 갖는 그레이 이미지인 필기체 숫자(0~9, 총 10종)들로 구성되어 있으며, 6만 개의 학습 데이터, 1만 개의 테스트 데이터로 구성된다.

인지 성능 평가 기술을 위한 기계학습 프레임 워크로는 대표적으로 Caffe[46], Tensorflow[47], Theano[48], Torch[49] 등이 널리 활용되고 있다. C++, Python, Lua 등 다양한 언어를 기반으로 개발된 기계학습 프레임워크들로 사용자가 원하는 신경망을 구성하고, 학습하며, 인지 테스트를 진행하는 환경을 제공한다. 하지만 기존의 기계학습 프레임워크는 소자 등의 모델링을 위한 기능 제공이 전무하므로 인공지능 소자를 위한 기계학습 프레임워크를 개발하기 위해서는 아날로그 특성의 인공지능 소자의 동작 연산 과정의 소프트웨어적인 모델링이 필요하다.

기계학습 프레임워크를 통한 학습 과정은 학습 대상 신경망에 학습 데이터를 입력하고 신경망의 인식 패스를 수행하는 단계, 인식 결과와 정답과의 오류를 신경망 연결 역 순서로 역 전파하여 레이어 파라미터 업데이트 양을 계산하는 단계, 레이어 파라미터를 업데이트하는 단계로 구성된다. 또한, 대부분의 기계학습 프레임워크들은 학습 과정에서 발생하는 방대한 양의 연산을 빠르게 수행하기 위해 GPU를 통한 가속화 라이브러리를 연동하여 동작하게 된다. 이러한 기존의 기계학습 연산과정에 아날로그 소자를 적용하기 위해서는 여러 가지 소자 파라미터의 특성을 GPU 등을 이용하는 기계학습 프레임워크에 어떻게 적절히 모델링하여 옮겨 놓을 수 있는지가 관건이다.

뉴로모픽 소자의 인지 성능 평가 프레임워크로 사용하기 위해서는 부동 소수점(floating point) 연산 기반으로 구성된 기존의 기계학습 프레임워크를 고정 소수점(fixed point) 연산 기반으로 변경하는 기술을 필요로 한다. 고정 소수점 연산을 위한 신경망 파라미터 양자화에 관한 연구로는 Gysel et al.(2015)[50], Han et al.(2016[51], 2015[52]) 논문이 있다.

Gysel et al.(2015)[50] 논문에서 'Dynamic range fixed point quantization' 방법을 신경망 양자화에 적용하여, 프레임워크의 고정 소수점 연산 모델의 가능성을 제시했다. 신경망의 파라미터 및 데이터의 값 분포가 레이어 별로 많은 차이를 보이는 것을 착안하여, 레이어 별로 파라미터 및 데이터 분포를 분석하고 레벨이 다른 양자화를 수행하며 재학습하는 방법으로 오류가 적은 신경망 양자화를 보였다. Han et al.(2016[51], 2015[52]) 논문에서는 신경망 압축 방법에 관한 다양한 방법을 제시했는데, 그중 파라미터 그룹화 방식의 양자화 방식을 소개한 바 있다.

뉴로모픽 소자 특성을 반영한 상위 수준 모델링에서는 뉴로모픽 소자의 특성을 프레임워크의 레이어 연산

모델에 적용한다. 뉴로모픽 소자의 동작 모델링, 시냅스 소자의 시냅스 비트 할당에 따른 인식 성능 평가를 위해 시냅스 비트 레벨 정확도 모델, 뉴로모픽 소자의 비선형성 특성 모델 등을 포함한다.

뉴로모픽 소자 특성을 반영한 상위 수준 모델링 관련 기술로 Du at el.(2017)[53] 연구에서는 멤리스터 소자의 특성 변이(variation)를 기계학습 과정으로 학습하여 MNIST 데이터에 대한 분류 성능을 보였다. 멤리스터 소자의 비선형성, 셀에 따른 동작 변이, 멤리스터의 단기 기억 메모리 특성을 반영하여 학습하기 위해, 88개의 멤리스터 소자의 출력 단의 데이터를 학습 입력으로 분류기 신경망을 학습하여 일정 수준 이상의 분류 성능을 얻을 수 있었다. 이는 기계학습 과정으로 멤리스터 특성 학습이 가능함을 시사한다.

Ⅳ. 결론

뉴로모픽 소자 연구는 생물학에서 연구되는 뇌의 학습, 기억, 그리고 인지 기능 등의 발현에 대한 이해의 노력도 필요하고 뉴로 사이언스에서 연구되는 계산 과학 분야의 이해와 더불어 이를 공학적으로 구현하기 위한 뉴로모픽 시스템, 알고리즘, 소자 등 다양한 공학분야에서의 지식의 발전이 필요하다. 이를 통해 뉴로모픽 반도체 칩 및 컴퓨터뿐만 아니라, 뉴로모픽 인공두뇌 모델을 응용하여 의학/생물학 데이터를 집적한 인공 장기 개념의 생물학적 인공 뇌 개발을 통해 신경계 질환 모델 구축 및 신경계 질환 약물 스크리닝에도 응용할 수 있다. 따라서 IT 및 BT 분야에의 다양한 신산업 창출이 가능한 차세대 연구 분야로서, 뉴로모픽 인공두뇌 모델링 기술 개발 분야에 대한 중요성이 국제적으로 부각되고 있다.

하지만 국내에서는 소수의 연구팀만이 개별적으로 관련 연구가 진행되고 있어, 국제적 연구 동향에 비하여 매우 미흡한 실정이다. 또한 융합 연구를 통해 향후 인간의 뇌의 동작원리를 가깝게 모사한 인지 기능을 하는 저전력 고집적 뉴로모픽 시스템을 개발하는 것은 메모리와 디스플레이 위주인 국내 반도체 산업에서의 경쟁력을 지속적으로 유지하기 위해 필요하므로, 국내/해외 연구 개발 현황을 분석을 바탕으로 국가적 대응 전략 수립이 시기적으로 매우 중요함을 알 수 있다.

약어 정리

ADAS	Advanced Driver Assistance Systems
AI	Artificial Intelligence
ANN	Artificial Neural Network
CIFAR-10	Canadian Institute For Advanced Research database
CMOS	Complementary Metal-Oxide-Semiconductor
CNN	Convolutional Neural Networks
CNT	Carbon Nano Tube
CPU	Central Processing Unit
FET	Field Effect Transistor
GPU	Graphic Processing Unit
ILSVRC	ImageNet Large Scale Visual Recognition Competition database
IoT	Internet of Things
MNIST	Modified National Institute of Science and Technology database
MS COCO	Microsoft Common Objects in Context database
MTJ	Magnetic Tunneling Junction
PASCAL VOC	Pascal Visual Object Classes database
ReRAM	Resistive Random Access Memory
SNN	Spiking Neural Network
SyNAPSE	Systems of Neuromorphic Adaptive Plastic Scalable Electronics

참고문헌

[1] I.K. Schuller et al., “Neuromorphic Computing: From Materials to Systems Architecture,” Report of a Roundtable Convened to Consider Neuromorphic Computing Basic Research Needs, 2015, pp. 1-37.

[2] G. Indiveri et al., “Neuromorphic silicon neuron circuits,”

Frontiers Neurosci., vol. 5, May 2011, pp. 1–23.

[3] C. Mead, "Neuromorphic Electronic Systems," *Proc. IEEE*, vol. 78, no. 10, Oct. 1990, pp. 1629–1636.

[4] J. Park et al., "Hierarchical Address Event Routing for Reconfigurable Large-Scale Neuromorphic Systems," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 28, no. 10, Oct. 2017, pp. 2408–2422.

[5] L. Lapicque, "Recherches Quantitatives sur L'excitation E'lectrique des Nerfs Traite´e Comme une Polarization," *J. Physiol. Pathol. Gen.*, vol. 9, 1907, pp. 620–635.

[6] M.D. Pickett et al., "A Scalable Neuristor Built with Mott Memristors," *Nat. Mater.*, vol. 12, Dec. 2013, pp. 144–147.

[7] J. Lin et al., "Low-Voltage Artificial Neuron Using Feedback Engineered Insulator-to-Metal-Transition Devices," in *Int. Electon. Dev. Meeting*, San Francisco, CA, USA, Dec. 3–7, 2016, p. 34.5.1–34.5.4.

[8] P. Stoliar et al., "A Leaky-Interate-and-Fire Neuron Analog Realized with a Mott Insulator," *Adv. Funct. Mater.*, vol. 27, no. 11, Mar. 2017, pp. 1–7.

[9] T. Tuma et al., "Strochastic Phase-Change Neurons," *Nat. Nanotech.*, vol. 11, 2016, pp. 693–670.

[10] X. Zhang et al., "An Artificial Neuron Based on a Threshold Switching Memristor," *IEEE Electron Dev. Lett.*, vol. 39, no. 2, Feb. 2018, pp. 308–311.

[11] A. Jaiswal et al., "Proposal for a Leaky-Integrate-Fire Spiking Neuron Based on Magnetoelectric Switching of Ferromagnets," *IEEE Trans. Electr. Dev.*, vol. 64, no. 4, Apr. 2017, pp. 1818–1824.

[12] Z. Wang et al., "Fully Memristive Neural Networks for Pattern Classification with Unsupervised Learning," *Nat. Electron.*, vol. 11, 2018, pp. 137–145.

[13] T. Ohno et al., "Short-Term Plasticity And Long-Term Potentiation Mimicked In Single Inorganic Synapses," *Nature Mat.*, vol. 10, 2011, pp. 591–595.

[14] Z. Wang et al., "Memristors with Diusive Dynamics as Synaptic Emulators for Neuromorphic Computing," *Nature Mat.*, vol. 16, 2017, pp. 101–110.

[15] L. Hu, "Ultrasensitive Memristive Synapses Based on Lightly Oxidized Sulfide Films," *Adv. Mater.*, vol. 29, no. 24, 2017, pp. 1–11.

[16] S. Choi et al., "Experimental Demonstration of Feature Extraction and Dimensionality Reduction Using Memristor Networks," *Nano Lett.*, vol. 17, no. 5, 2017, pp. 3113–3118.

[17] T. Chang et al., "Short-Term Memory to Long-Term Memory Transition in a Nanoscale Memristor," *ACS Nano*, vol. 5, no. 9, 2011, pp. 7669–7676.

[18] Z.H. Tan et al., "Synaptic Metaplasticity Realized in Oxide Memristive Devices," *Adv. Mater.*, vol. 28, no. 2, Jan. 2016, pp. 377–384.

[19] D. Kuzum et al., "Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing," *Nano Lett.*, vol. 12, no. 5, 2012, pp. 2179–2186.

[20] C.D. Wright et al., "Arithmetic and Biologically-Inspired Computing Using Phase-Change Materials," *Adv. Mater.*, vol. 23, no. 30, Aug. 2011, pp. 3408–3413.

[21] X. Wang et al., "Spintronic Memristor Through Spin-Torque-Induced Magnetization Motion," *IEEE Electr. Dev. Lett.*, vol. 30, no. 3, Mar. 2009, pp. 294–297.

[22] G. Srinivasan et al., "Magnetic Tunnel Junction Based Long-Term Short-Term Stochastic Synapse for a Spiking Neural Network with On-Chip STDP Learning," *Sci. Rep.*, vol. 6, 2016, Article no. 29545.

[23] S. Kim et al., "Pattern Recognition Using Carbon Nanotube Synaptic Transistors with an Adjustable Weight Update Protocol, " *ACS Nano*, vol. 11, no. 3, 2017, pp. 2814–2822.

[24] Y. Burgt et al., "A Non-Volatile Organic Electrochemical Device as a Low-Voltage Artificial Synapse for Neuromorphic Computing," *Nature Mat.*, vol. 16, 2017, pp. 414–419.

[25] C.S. Yang et al., "A Synaptic Transistor based on Quasi-2D Molybdenum Oxide," *Adv. Mater.*, vol. 29, no. 27, July 2017, Article no. 1700906.

[26] J. Shi et al., "A correlated nickelate synaptic transistor," *Nature Commun.*, vol. 4, 2013, Article no. 2676.

[27] Y.M. Kim et al., "Short-Term and Long-Term Memory Operations of Synapse Thin-Film Transistors using an In-Ga-Zn-O Active Channel and a Poly(4-vinylphenol)-Sodium β-Alumina Electrolytic Gate Insulator," *RSC Adv.*, vol. 6, 2017, pp. 52913–52919.

[28] F. Shao at al., "Oxide-Based Synaptic Transistors Gated by Sol-Gel Silica Electrolytes," *ACS Appl.* Mater. Interfaces, vol. 8, 2016, pp. 3050–3055.

[29] C. Wan et al., "Classical Conditioning Mimicked in Junctionless IZO Electric-Double-Layer Thin-Film Transistors," *IEEE Electr. Dev. Lett.*, vol. 35, no. 3, Mar. 2014, pp. 414–416.

[30] S.M. Yoon et al., "An Electrically Modifiable Synapse Array Composed of Metal-Ferroelectric-Semiconductor (MFS)

FET's using SrBi2Ta2O9 Thin Films," *IEEE Electron. Dev. Lett.*, vol. 20, no. 5, May 1999, pp. 229-231.

[31] S.M. Yoon et al., "Adpative-Learning Neuron Integrated Circuits Using Metal-Ferroelectric (SrBi2Ta2O9)-Semiconductor (MFS) FET's," *IEEE Electron. Dev. Lett.*, vol. 20, no. 5, May 1999, pp. 526-528.

[32] S.M. Yoon et al., "Ferroelectric Neuron Integrated Circuits Using SrBi2Ta2O9-Gate FET's and CMOS Schmitt-Trigger Oscillators," *IEEE Trans. Electron. Dev.*, vol. 47, no. 8, Aug. 2000, pp. 1630-1635.

[33] A. Chanthbouala et al., "A Ferroelectric Memristor," *Nature Mater.*, vol. 11, Sept. 2012, pp. 860-864.

[34] S. Boyn et al., "Learning Through Ferroelectric Domain Dynamics in Solid-State Synapses," *Nature Commun.*, vol. 8, Apr. 2017, Article no. 14736.

[35] C. Yoon et al., "Synaptic Plasticity Selectively Activated by Polarization-Dependent Energy-Efficient Ion Migration in an Ultrathin Ferroelectric Tunnel Junction," *Nano Lett.*, vol. 17, no. 3, 2017, pp. 1949-1955.

[36] K. Likharev et al., "CrossNets: High-Performance Neuromorphic Architectures for CMOL Circuits," *Annu. NY Acad. Sci.*, vol. 1006, no. 1, Dec. 2003, pp. 146-163.

[37] K. Likharev, "CrossNets: Neur Omorphic Hybrid CMOS/Nanoelectronic Networks," *Sci. Adv. Mater.*, vol. 3, no. 3, June 2011, pp. 322-331.

[38] D.B. Strukov et al., "CMOL FPGA: a Reconfigurable Architecture for Hybrid Digital Circuits with Two-Terminal Nanodevices," *Nanotechnol.*, vol. 16, no. 6, Apr. 2005, pp. 888.

[39] P. Lin et al., "3D Integration of Planar Crossbar Memristive Devices with CMOS Substrate," *Nanotechnol.*, vol. 25, no. 40, 2014, Article no. 405202.

[40] M. Jo et al., "Novel Cross-Point Resistive Switching Memory with Self-Formed Schottky Barrier," *Symp. VLSI Technol.*, Honolulu, HI, USA, June 15-17, 2010, pp. 53-54.

[41] M. Lee et al., "Brain-Inspired Photonic Neuromorphic Devices using Photodynamic Amorphous Oxide Semiconductors and their Persistent Photoconductivity," *Adv. Mater.*, vol. 29, no. 28, July 2017, Article no. 1700951.

[42] S. Qin et al., "A Light-Stimulated Synaptic Device Based on Graphene Hybrid Phototransistor," *2D Mater.*, vol. 4, no. 3, Aug. 2017, Article no. 035022.

[43] http://www.panasonic.co.jp/corp/news/official.data/data.dir/2013/06/en130610-3/en130610-3.html

[44] Moon et al., "High Density Neuromorphic System with Mo/Pr0.7Ca0.3MnO3 Synapse and NbO2 IMT Oscillator Neuron," *IEEE Int. Electr. Dev. Meeting*, Washington, DC, USA, Dec. 7-9, 2015, pp. 17.6.1-17.6.4.

[45] G.W. Burr et al., "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element," *IEEE Trans. Electr. Dev.*, vol. 62, no. 11, Nov. 2015, pp. 3498-3507.

[46] http://caffe.berkeleyvision.org/

[47] https://www.tensorflow.org/

[48] http://deeplearning.net/software/theano/

[49] http://torch.ch/

[50] P. Gysel et al., "Hardware-Oriented Approximation of Convolutional Neural Networks," *Int. Conf. Learn. Representations*, San Juan, Puerto Rico, May 2-4, 2016, pp. 1-8.

[51] S. Han et al., "Learning Both Weights and Connections for Efficient Neural Network." In *Proc. Int. Conf. Neural Inform. Process. Syst.*, Montreal, Canada, Dec. 7-12, 2015, 2016, pp. 1135-1143.

[52] S. Han et al., "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," *Proc. Int. Conf. Neural Inform. Process. Syst.*, Montreal, Canada, Dec. 7-12, 2015, pp. 1-9.

[53] C. Du et al., "Reservoir Computing Using Dynamic Memristors for Temporal Information Processing," *Nature Commun.*, vol. 8, 2017, Article no. 2204.