스파이킹 신경망 기반 뉴로모픽 기술 동향

박성모 한국전자통신연구원 초경량지능형반도체연구실 책임연구원
최병건 한국전자통신연구원 초경량지능형반도체연구실 책임연구원
이재진 한국전자통신연구원 초경량지능형반도체연구실 책임연구원
여순일 지능형 반도체 PG(417) 의장, 한국전자통신연구원 책임연구원
박경환 한국전자통신연구원 초경량지능형반도체연구실 실장

1. 머리말

뉴로모픽(Neuromorphic)이란 직렬로 동작하는 기존 컴퓨터에서 병렬로 동작하는 뇌를 모방하여 처리하는 반도체를 기반으로 설계하는 기술이다. 직관적으로 인식하기 어려운 비정형적인 문자, 이미지, 음성 등을 효율적으로 처리할 수 있는 장점이 있다. 뇌는 약 1,000억 개의 뉴런이 시냅스로 연결되어 있고, 각 뉴런에서 스파이크 전기 자극을 만들어 다른 뉴런으로 신호가 전달된다. 이때, 신호의 자극 세기가 임계값 보다 높으면 신호가 전송되고, 낮으면 전송되지 않는다. 뇌는 이러한 복잡도의 연산을 약 20W의 낮은 에너지로 연산, 저장 및 학습을 동시에 수행한다[1].

한편, 최근 들어 주목받고 있는 스파이킹 신경망(SNN, Spiking Neural Network)은 폰 노이만기반 컴퓨터 구조의 단점을 해결하기 위해 생물학적 신경 네트워크 구조를 유사하게 모방하였다. 이런 특성으로 인해 스파이킹 신경망 기반의 뉴로모픽 칩은 적은 전기신호로도 동작이 가능하고, 심층 신경망(DNN, Deep Neural Network) 및 합성곱 신경망(CNN, Convolutional Neural Network)보다 소모 전력이 적어 에너지 효율이 좋다는 장점이 있다. 뉴로모픽 기술은 드론, 인텔리전스 엣지, IoT 디바이스, 웨어러블 디바이스, 자율주행 자동차, 인지로봇 및 모바일 단말 분야를 중심으로 인공지능 반도체 시장이 급속도로 성장하는 등 산업전반에 걸쳐 파급 효과를 클 것으로 예상된다[1].

2. 스파이킹 뉴럴 네트워크 비교 및 현황

2.1 심층 신경망(DNN)과 스파이킹 신경망(SNN)

심층 신경망(DNN)은 입력층(input layer)과 출력층(output layer) 사이에 여러 개의 은닉층(hidden layer)들로 이루어진 인공 신경망(ANN, Artificial Neural Network)이다[2][3]. 심층 신경망은 기존의 인공 신경망과 같이 복잡한 연산으로 학습될 경우 많은 문제가 발생할 수 있다. 기계 학습에서 학습 데이터를 과하게 학습하여 오차가 증가하거나 연산에 시간이 많이 걸리는 문제점이 대표적이다[4].

스파이킹 신경망은 신경세포 각각에 대해서 학습을 수행한다면, 심층 신경망은 전체를 학습하는 방식이다. 심층 신경망에는 이미지 인식율이 높은 합성

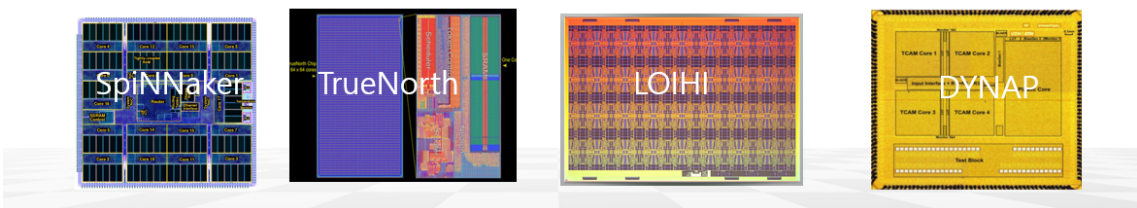
곱 신경망, 시계열적인 데이터에 적합한 순환 신경망(RNN, Recurrent Neural Network), 게임 등에 활용되는 강화학습(Reinforcement Learning)이 있다. 스파이킹 신경망의 학습방법은 스파이크타이밍 종속성(Spike-timing-dependent plasticity)으로 뇌의 뉴런 간의 연결 강도를 조절하는 생물학적 과정을 통해 수행한다. 이는 특정 신경 세포의 출력과 입력 활동 전위의 상대적인 타이밍을 기반으로 연결 강도를 조정하며, 시냅스 전 스파이크와 시냅스 후 스파이크의 시간 차이를 통해 시냅스 가중치를 활용해 학습한다. 심층 신경망은 현재 인공지능 분야에서 가장 널리 활용되고 있으며, 성능 면에서 스파이킹 신경망보다 뛰어나다. 특히 이미지 분류 분야는 GPU의 성능 향상에 따라 인간의 판단력보다 빠르다. 반면 스파이킹 신경망은 현재 초기 연구단계이며 향후 뉴로모픽 칩의 성장으로 심층 신경망의 문제점을 근본적으로 개선 할 수 있는 연구 분야이다.

2.2 뉴로모픽 칩의 현황 및 연구 동향

인텔은 [그림 1]과 같이 스파이킹 신경망 기반의 학습이 가능한 로이히(LOIHI)라는 뉴로모픽 칩을 구현하였다. 특히 기존 방식보다 1,000배의 에너지 효율을 내고 13만 개의 뉴런과 13억 개의 시냅스를 내장하고 있으며, 14nm 공정을 이용하여 설계하였다. 또한 유럽의 취리히 대학에서는 아날로그와 디지털 방식을 혼합한 구조의 스파이킹 신경망 기반의 뉴로모픽 칩을 개발하였는데, 스파이크를 입력받아 동작하는 다이내믹 비전 센서를 이용하여 빠른 동작이 요구되는 응용분야에 적합한 것으로 알려져 있다. 영국의 맨체스터 대학에서는 암코어를 여러 개 연결하여 뉴로모픽을 구현하였으며, IBM에서는 트루노스(TrueNorth)라는 칩을 개발했다. 이는 사람의 얼굴과 동작 인식 등에 응용하여 활용 할 수 있을 것으로 보인다.

뉴로모픽 칩은 인공지능의 기술발달로 인해 다양한 분야에서 반도체 수요가 증가하고 있기에 전망은 낙관적이다. 또한 뉴로모픽 칩은 제4차 산업혁명의 핵심기술인 반도체와 스마트폰에 핵심부품으로 사용될 것으로 보이며 향후 세계를 선도하는 기술로 보인다.

제품명	회사	특징	전력	문제점
SpiNNaker	맨체스터 대학	18-ARM9 코어/칩 SoC (102mm^2) 18K 뉴런/칩 130nm	50KW @500M Neurons	고비용
TrueNorth	IBM	45pJ/스파이크 1M 뉴런, 256M 시냅스	73mW	단순한 모델
LOIHI	INTEL	14nm, SNN기반 STDP 학습, 130,000 뉴런,130M 시냅스	고 에너지 효율	
Dynap	취리히 대학	28nm, SNN-기반 온칩 학습	13배 에너지 효율	특정 응용분야

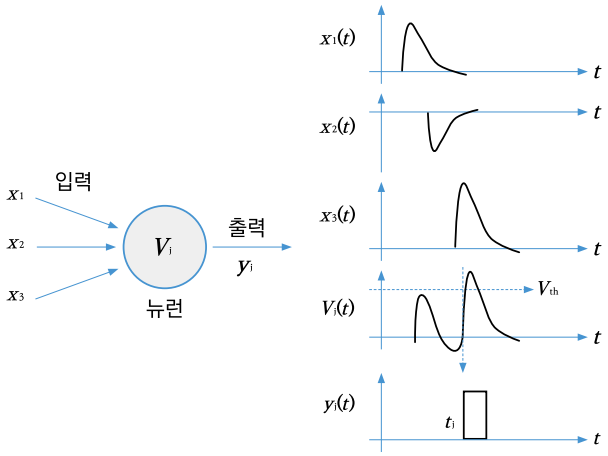


[그림 1] 뉴로모픽 칩 현황

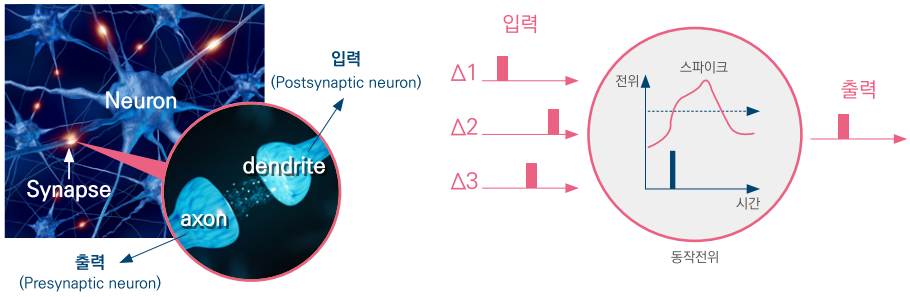
3. 스파이킹 뉴럴 네트워크 동작 및 세대별 신경망

펄스 신경망에서 뉴런의 상태는 막전위 및 활성화 임계값에 의해 결정되고, 뉴런의 막전위는 이전 층의 뉴런으로부터 온 시냅스 후 전위에 의해 결정된다. 흥분성 연접후 전위(EPSP, Excitatory Post Synaptic Potential)는 뉴런의 막전위를 증가시키고, 억제성 연접후 전위(IPSP, Inhibitory Post Synaptic Potential)는 뉴런의 막전위를 감소시킨다. 뉴런의 막전위가 활성화 임계값 이상으로 상승하면 뉴런은 스파이크를 생성한다. 펄스는 [그림 2]와 같이 뉴런의 축색을 통해 다음 뉴런으로 전달되는데, 시냅스를 따라 전달되는 펄스 과정에는 특정 시간이 필요하며, 이 시간을 시냅스 지연이라고 한다.

생물학적 뉴런과 인공 스파이킹 뉴런 동작의 연관성은 [그림 3]과 같다. 뇌는 뉴런으로 구성되어 있고, 각 뉴런은 다른 시냅스와 연결된다. 수상돌기(dendrite)를 통해서 정보가 입력되면, 다른 뉴런으로 향하는 전기적 신호인 스파이크를 발생시키고, 생성된 신경 신호는 시냅스를 통해 인접한 뉴런으로 전달된다[6]. 뉴로모픽 칩은 이와 같은 뇌의 생물학적 뉴런의 동작을 모방한 것으로, 스파이크가 입력되면 이 스파이크의 시간적 간격과 형태에 따라 전위가 증가하고 이 전위가 일정 수준 이상의 전위에 도달하면 스파이크가 발생하여 출력한다. 이처럼 스파이킹 신경망은 시냅스를 통해 정보를 전송하는 뇌의 신경정보 처리를 모방한 것이다.



[그림 2] 펄스 뉴런 활성화[2]



[그림 3] 생물학적 뉴런과 인공 스파이킹 뉴런과의 연관성[3]

〈표 1〉 세대별 인공지능 신경망

	입출력	특징	예제
1세대	이진, 이진	• 장점: 디지털 계산을 위한 범용성, 스파이크와 같은 매우 직관적인 해석, 최적화해야 하는 매개 변수 수가 적음 • 단점: 이진 출력으로 제한됨, 아날로그 세계에 제한됨, 시간 개념을 포함하지 않음	다층퍼셉트론 신경망
2세대	실수, 실수	• 장점: 디지털 계산을 위한 범용성, 1세대보다 생물학적으로 접근이 향상, 대규모 경사 기반 최적화 기법, • 단점: 스파이크로 해석은 직관적이지 않음, 파라미터 값에 민감 동작함, 최적화해야 하는 많은 파라미터 필요	합성곱 신경망, 순환 신경망
3세대	실수, 실수	• 장점: 디지털 계산을 위한 범용성, 2세대보다 생물학적으로 접근, 계산이 강력, VLSI 구현에 성공적, 연산의 대규모 병렬화 • 단점: 매개 변수 값에 민감함, 최적화해야 하는 많은 매개 변수, 아직 확실한 이론이 없음, 다양한 시공간 데이터에 대해 알려지지 않음, 생물학적으로 강력한 이론을 공식화하기에 충분하지 않음, 다양한 모델과 코딩 체계에 대한 일관된 프레임 워크 부족	스파이킹 신경망

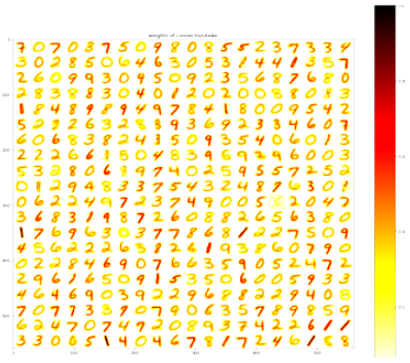
〈표 2〉 스파이킹 신경망 시뮬레이터[7]

	뉴런모델			시냅스모델				시냅틱가소성					입력		툴			통합모델			프론트엔드			백엔드					플랫폼		
	Leaky integrate-and-fire (LIF)	Izhikevich 4-param	Hodgkin-Huxley	Current-based (CUBA)	Conductance-based (COBA)	AMPA, NMDA, GABA	Neuromodulation	Short-term plasticity (STP)	E-STDP	I-STDP	DA-STDP	Synaptic scaling / homeostasis	Current injection	Spike injection	Prameter tuning	Analysis and visualisation	Regression suite	Forward / exponential Euler	Exact integration	Runge-Kutta	Python / PyNN	C / C++	Java	Single-threaded	Multi-threaded	distributed	Single GPU	Multi-GPU	Linux	Mac OS X	Windows
CARLsim		×		×	×	×	/	×	×	×	×	×	×	×	×	×	×	×				×		×			×	/	×	×	×
Brian	×	×	×	×	×	×	/	×	×	×	/	×	×	×	/	/	×	×	×	×	×	×		×	×	/			×	×	×
GENN		×	×	×	×	×	/	/	×			/	×	/				×				×		×			×		×	×	×
NCS	×	×	×	×	×	/		×	×				×	×		/		×			×	×		×	×	×	×	×	×		
NEMO		×			×		/	×	×	×			×	×		×	×	×			×	×		×	×		×		×	×	×
Nengo	×	×	×	×	×	×	×	×	/			×	×		×	×	×			×	×		×	×	×		×		×	×	×
NEST	×	×	×	×	×	×	/	×	×		×	×	×	×			×	×	×	×	×			×	×	×			×	×	
PCSIM	×	×	×	×	×	×	/	×	×		×	×	×	×		/	×	×			×	/	×	×	×	×			×	×	/

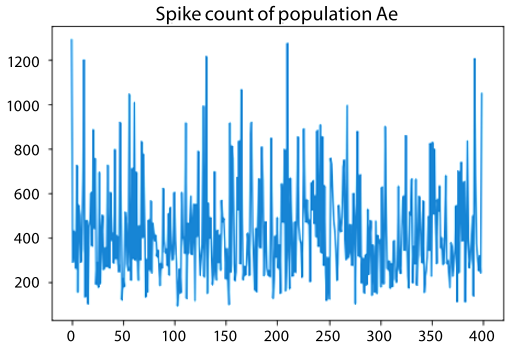
세대별 인공지능 신경망은 〈표 1〉과 같다. 1세대는 다층퍼셉트론 신경망으로 스파이크가 매우 직관적이며 최적화해야 하는 변수가 단순하다는 특징이 있다. 2세대는 합성곱 신경망과 순환 신경망이 해당하며, 1세대 보다 생물학적 접근이 향상되었고 최적화해야 하는 파라미터가 더 많이 필요하다. 3세대 인공지능 신경망은 스파이킹 신경망으로 매개 변수값에 민감하며 최적화해야 하는 매개 변수가 많고, 다양한 시공간 데이터에 대해 알려지지 않고 있다. 생물학적으로 강력한 이론을 공식화하기에 충분하지 않으며, 다양한 모델과 코딩 체계에 대한 일관된 프레임 워크가 부족하나, 2세대 보다 생물학적인 접근이 용이하여 칩으로서 성공 가능성이 높고, 연산의 대규모 병렬화가 가능하다.

4. 스파이킹 뉴럴 네트워크 시뮬레이터

스파이킹 신경망 시뮬레이터는 〈표 2〉와 같이 시

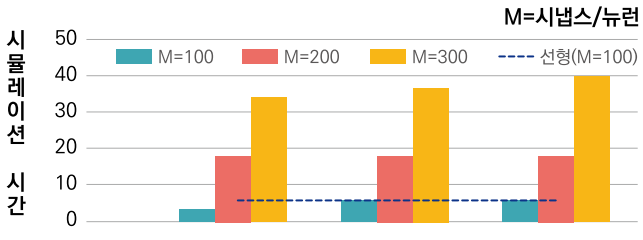


〈시냅스 웨이트값〉



〈각 뉴런 수에 따른 스파이크 수〉

[그림 4] 스파이킹 신경망 시뮬레이션[8]



	10	50	90
M=100	3.45	5.59	5.62
M=200	17.71	17.95	18.1
M=300	32.99	35.9	40.27

뉴런 수(x1000)

[그림 5] 칼심(CARLsim-4) 환경에서 뉴런과 시냅스 최적화 시뮬레이션

뮬레이터의 환경을 제공하고 있다. 제공된 환경은 뉴런모델, 시냅스모델, 시냅틱 가소성, 프론트엔드, 백엔드, 지원하는 플랫폼의 종류를 나타내고 있다. 이를 이용하여 신경회로망의 상위수준 모델링, 뉴로모픽 칩을 개발하기 위한 검증 도구로 사용 가능하다. 제공되는 시뮬레이터는 신경과학, 실시간 응용 로봇, 뉴로모픽 분야 등에서 활용되고 있다.

브라이언(Brian)은 스파이킹 신경망을 시뮬레이션 할 수 있는 오픈 소스이다. 파이선 기반의 브라이언2를 이용하였으며 [그림 4]와 같이 뉴런과 시냅스의 모델링을 시뮬레이션 한다. 설계 환경은 윈도우 환경에서 아나콘다 4.4.0이며, 엠니스트(MNIST) 데이터셋을 입력해 스파이크로 변환하여 학습을 수행하였다. 이때 뉴런의 수는 400개를 사용하였다. 뉴런 네트워크 구조를 하드웨어로 구현하기 위해서는 강화뉴런과 억제뉴런이 일대일로 복잡한 구조를 가지고 있다. 브라이언2 기반 강화뉴런과 억제뉴런의 배열 및 네트워크 구조와 뉴런 개수를 최적화하였다. 시뮬레이션 환경은 파이선 기반 브라이언2를 이용하여 시뮬레이션을 수행했으며, 결과 정확도는 91.4% 였다[8].

칼심(CARLsim-4)은 스파이킹 신경망 시뮬레이터로, GPU를 이용한 가속의 라이브러리를 제공한다. 칼심을 이용하여 [그림 5]와 같이 뉴런과 시냅스의 최적 조건을 시뮬레이션했다. 시뮬레이션 결과 시

냅스와 뉴런의 비가 300일 때 최적화 조건을 알았다.

5. 맺음말

스파이킹 신경망(SNN) 기반의 뉴로모픽 칩은 향후 인공지능의 발달과 함께 기존의 심층 신경망(DNN) 기반의 뉴로모픽 칩에서 해결하지 못한 문제점을 해결하는 솔루션으로 각광받을 연구 분야이다. 하지만 지금까지의 연구결과에 의하면 스파이킹 신경망이 가지는 장점은 많으나 학습의 정확도나 성능 면에서는 해결해야 할 점이 많다. 생물학적 동작과 유사한 뇌를 하드웨어로 구현하려면 알고리즘, 구조, 회로설계, 소자 및 플랫폼의 종합적인 기술 개발이 필요하다. 스파이킹 신경망 기반 뉴로모픽 칩은 아주 적은 전기신호로 동작하고 에너지 효율이 좋다는 장점으로 인해, 향후 인공지능 반도체 시장이 급성장하는데 기여하는 등 산업전반에 걸쳐 파급 효과가 클 것으로 예상된다. TTA

※ 본 연구는 MSIT / IITP의 ICT R & D 프로그램에 의해 수행됨[2018-0-00197, 경량 RISC-V 기반 초저전력 인텔리전트 엣지 지능형반도체 기술 개발].

참고문헌

[1] '사람 뇌 닮은 반도체칩, 뉴로모픽', The Science Times, June 10, 2019.

[2] Y. Bengio, A. Courville, and P. Vincent., 'Representation Learning: A Review and New Perspectives,' IEEE Trans. PAMI, special issue Learning Deep Architectures, 2013.

[3] J. Schmidhuber, 'Deep Learning in Neural Networks: An Overview', 2014.

[4] James Mnatzaganian & Ross Reinhardt, Memristors and Neuromorphic Computing, 2012 위키백과.

[5] Gerstner W, Kistler W M. Spiking neuron models: Single neurons, populations, plasticity[M]. New York: Cambridge University Press, 2002, 20-25.

[6] W. Gerstner and W. M. Kistler, Spiking Neruon Models, Cambridge University Press, 2002.

[7] Cognitive Anteater Robotics Laboratory, 'CARLsim: a GPU-Accelerated SNN Simulator,' Mar. 2017.

[8] Peter U. Diehl, Matthew Cook, 'Unsupervised Learning of Digit Recognition Using Spike-Timing-Dependent Plasticity', IEEE TRANSACTIONS IN NEURAL NETWORKS AND LEARNING SYSTEMS, vol. 1, pp1-6, 2014.

주요 용어 풀이

- EPSP(excitatory post synaptic potential): 이온통로가 열리면서 연접 후부로 양이온이 흘러들어가는 것으로 인해 연접후부의 막전위가 일시적으로 탈분극 현상을 일으켜 나타내는 전위
- IPSP(Inhibitory Post Synaptic Potential): 향후 활동전위가 연접 후 뉴런 또는 알파운동뉴런에서 일어날 확률을 감소시키는 연접 후 전위
- Synapse: 뉴런 간 연결을 하며 화학적·전기적인 반응을 통하여 뉴런에서 발생하는 스파이크 신호를 다른 뉴런으로 전달해주는 역할
- STDP(Spike Timing Dependent Plasticity): 스파이크 신호의 타이밍에 따른 가소성
- SNN(Spiking Neural Netwok): 인공 신경망 모델로서 자연 신경망을 보다 가깝게 모방한 네트워크로, 신경 세포와 시냅스 상태뿐만 아니라 시간 개념 모델을 기반으로 함
- MNIST(Modified National Institute of Standards and Technology database): 손으로 쓴 숫자들로 이루어진 대형 데이터베이스이며, 다양한 영상을 처리·학습하기 위해 일반적으로 사용