뉴로모픽 하드웨어 기술 동향

DOI: 10.3938/PhiT.28.013

정 두 석

Recent Progress in Neuromorphic Hardware Development

Doo Seok JEONG

Neuromorphic hardware realizes spiking neural networks by using dedicated hardware far more efficiently than general-purpose hardware such as central processing units and graphics processing units. Neuromorphic engineering originally aimed to reverse engineer the brain by using analog very large-scale integrated (VLSI) circuits, which dates back to the 1980s. Nowadays, it includes various implementation methods (fully digital, analog/digital mixed circuits) and application domains. In this article, I first address the key functionality of the brain for neuromorphic engineering and move on to its difference from deep learning. The recent trend in neuromorphic hardware development is followed; particularly, important attributes that are pursued are addressed.

들어가는 글

사람의 두뇌는 수많은 스파이킹 뉴런(spiking neuron)과 더 많은 수의 시냅스(synapse)로 구성된 매우 복잡한 스파이킹 신경망(spiking neural network)으로 구성되어 있다. 특히 두뇌의 신피질(neocortex)은 진화의 역사상 가장 최근에 발달된 영역으로 감각으로부터 생성된 기억을 토대로 인지/학습 등의 고차원적인 기능을 수행하는 영역이다. 이는 현재 대부분의 인공지능기능이 구현하고자 하는 핵심 기능이다.

생물학적 스파이킹 신경망을 전용 하드웨어¹⁾를 이용하여 구현하고 이를 이용하여 스파이킹 신경망의 인지/학습 등의 기능을 구현하고자 하는 기술을 뉴로모픽 기술이라고 하며 이에

저자약력

정두석 교수는 독일 RWTH Aachen 공학박사(2008)로 한국과학기술연구원 연구원/선임/책임연구원(2008.9-2018.2)를 거쳐 현재 한양대학교 신소재공학부 부교수(2018.3-현재)로 재직 중이다.(dooseokj@hanyang.ac.kr)

사용된 전용 하드웨어를 뉴로모픽 하드웨어¹⁾라 한다. 뉴로모픽 하드웨어의 역사는 1980년대로 거슬러 올라간다. 1980년대 미국 캘리포니아 공과대학의 Carver Mead 교수는 시신경 경로 하단부²⁾의 신경회로를 아날로그³⁾ 대규모집적(VLSI) 회로를 이용하여 모방하는 기술을 제안하였으며 이를 뉴로모픽 기술로 명명하였다.^[2] 따라서 초기 뉴로모픽 기술의 정의는 아날로그 VLSI 회로를 이용하여 두뇌를 역설계(reverse engineering)하는 기술이다. 현재 뉴로모픽 기술은 초기에 정의된 기술 뿐 아니라 다양한 새로운 방법론 및 목적을 포함하고 있으며 특히 시각정보 인식, 음성정보 인식 등 감각기관 말단에서부터 최종 인지/학습을 관장하는 신피질에 이르는 경로를 기능주의(functionalism) 관점에서 모사하는 기술을 주요기술로 포함하고 있다. 초기 뉴로모픽 기술이 신경경로를 신경세포의 전기생리학(electrophysiology)적 거동을 자세히 모사했다면 최근 기술은 전체 경로의 수준별 기능을 모사하지만 전기생리학적 기전과의 거리감이 커져가는 추세이다.

이러한 기술개발 동향은 최근 인공지능기술의 핵심인 딥러닝(deep learning)의 하드웨어형 대안기술로서 뉴로모픽 하드웨어를 바라보는 관점에 기인한다. 즉, 과거 뉴로모픽 기술이 두뇌 모사향 기술이었다면 현재는 딥러닝향 기술이 주류를 이루며 뉴로모픽 하드웨어는 딥러닝을 매우 효율적으로 구현할 수 있는 차세대 딥러닝 가속기⁴⁾ 또는 딥러닝이 구현할 수 없는 그 이상의 가치를 창출할 수 있는 전용 하드웨어로 간주되고 있다.

REFERENCES

- [1] D. S. Jeong, J. Appl. Phys. **124**, 152002 (2018).
- [2] C. Mead, *Analog VLSI and Neural Systems* (Adison-Wesley, Reading, MA, 1989).

- 1) 중앙처리장치(CPU), 그래픽 처리장치(GPU) 등의 범용 하드웨어와 반대되는 하드웨어로 특수목적을 위하여 설계된 하드웨어를 지칭함.
- 2) 망막의 광수용체(photoreceptor)로부터 망막 신경절세포(retinal ganglion cell)에 이르는 경로
- 3) 아날로그 회로는 회로 상 노드의 연속적인 전압 값을 입출력으로 취하는 회로를 의미함
- 4) 딥러닝 가속기(deep learning accelerator)는 기존 소프트웨어 기반 딥러닝 기술을 전용 하드웨어를 이용하여 구현하여 기존 범용 하드웨어 기술 기반 딥러닝 대비 연산속도의 개선을 목표로 한다.

뇌의 핵심기능

딥러닝 또는 뉴로모픽 컴퓨팅⁵⁾을 통해 구현하고자 하는 두뇌의 핵심 기능은 크게 인지, 학습, 예측 기능이며 이 기능을 가능케 하는 우리 두뇌의 핵심은 패턴인식(pattern recognition) 기능이다. 우리 두뇌는 감각기관을 통해 접수된 무질서한 대용량 데이터를 입력으로 취한다. 예를 들어 우리는 아침에 눈을 뜨는 순간부터 우리의 안구를 통해 입력되는 다양한 파장을 가지는 광자(photon) 데이터를 접수한다. 이와 같은 광자 데이터를 하위수준(low level) 데이터라 한다. 하위수준 데이터는 데이터의 양이 방대하며 시시각각 데이터의 형태가 변화하며 겉보기에 매우 무질서해 보인다. 하지만 우리 두뇌는 방대하며 무질서해 보이는 하위수준 데이터의 숨겨진 질서(pattern)를 매우 짧은 시간에 찾아내어 입력된 하위수준 데이터를 “추상화”한다. 추상화된 데이터는 상위수준(high level) 데이터로 하위수준 데이터가 포함하고 있는 핵심질서를 포함하는 매우 간결한 데이터이다.

10년 만에 만나는 친구의 모습을 생각해 보라. 지난 10년 동안 친구의 생김새는 바뀌었을 것이다. 즉, 하위수준 데이터의 변화가 매우 클 것이다. 하지만 여전히 친구를 인식할 수 있고 이는 하위수준 데이터의 변화에도 불구하고 친구 모습의 핵심패턴이 변화하지 않았고 우리의 두뇌는 그 패턴을 찾아낼 수 있기 때문에 가능한 일이다.

시각정보 인식의 경우 우리 안구에 2차원적으로 투사되는 광자의 공간배열로 구성된 하위수준 데이터의 숨겨진 패턴을 우리 두뇌는 인식하는 것이며 음성인식의 경우 고막을 통해 입력되는 음파(하위수준 데이터)의 시간 상 숨겨진 패턴을 인식할 수 있으며 현재 이후에 이어질 시간 패턴을 예측할 수 있다. 따라서 우리 두뇌는 하위수준의 시공간 데이터로부터 숨겨진 시공간 패턴을 인식할 수 있다. 이는 딥러닝 또는 뉴로모픽 컴퓨팅을 통해 모방하고자 하는 우리 두뇌의 핵심 기능이다.

현재 뉴로모픽 기술의 범위

초기 뉴로모픽 기술은 아날로그 VLSI 회로를 이용하여 생리학적인 신경망의 역설계를 목표로 했다. 하지만 지난 30년 동안 뉴로모픽 기술의 범위는 크게 확장되었다. 생리학적 신경망의 역설계 방법론으로 아날로그 VLSI 회로뿐 아니라 디지털 VLSI 회로 또는 아날로그/디지털 혼성회로 등의 기술을 포함한다.^[3-7] 최근 비휘발성 메모리 기반의 방법론^[8] 또한 등장하여 외연은 더욱 넓어지고 있는 추세다. 그럼에도 불구하고 뉴로모픽 기술은 앞서 언급한 스파이킹 신경망을 기반으로 하고 뉴로모픽 하드웨어는 스파이킹 신경망을 전용 하드웨어로 구현

한다는 점은 뉴로모픽 기술의 핵심이다.

뉴로모픽 기술의 목표 역시 다변화 중이다. 초기 기술이 우리 두뇌의 역설계를 목표로 하였다면 최근에는 기능주의 관점에서 딥러닝향 기술이 각광을 받고 있다.

딥러닝과의 차별성

뉴로모픽 컴퓨팅과 딥러닝은 사용하는 신경망 모델에서 크게 차별화 된다. 앞서 언급한 바와 같이 뉴로모픽 컴퓨팅에서는 스파이킹 신경망을 신경망 모델로 취하며 딥러닝은 인공 신경망(artificial neural network, ANN)을 모델로 취한다. 스파이킹 신경망과 인공신경망은 공통적으로 노드(뉴런)와 엷지(시냅스)로 구성되어 있는 수학적 그래프 모델이나 노드의 동작원리가 상이하다.^[1] 스파이킹 신경망의 노드는 스파이킹 뉴런(spiking neuron)으로 뉴런 간 통신의 기본 언어는 활동전위 혹은 스파이크로 명명된 전압 펄스다. 시냅스전 뉴런(presynaptic neuron)은 시냅스후 뉴런(postsynaptic neuron)에 엷지(시냅스)를 통해 스파이크를 전달하고 시냅스후 뉴런은 입력 스파이크를 시간 상 적분하여 적분 값이 문턱 값을 초과 시 스파이크를 생성하고 해당 시냅스후 뉴런으로 스파이크를 전달한다. 즉, 스파이킹 신경망은 시간 의존적(time-dependent) 모델이다.

인공신경망의 노드는 실수 입력 값을 엷지를 통해 받아들이며 이를 더하여 출력 값을 결정한다. 즉, 인공신경망의 노드는 스파이킹 뉴런과 달리 시간 상 적분과정을 포함하고 있지 않으므로 인공신경망은 시간 비의존적(time-independent) 모델이다. 따라서 인공신경망은 정적데이터(static data) 인식에 적합하다. 이에 반해 스파이킹 신경망은 정적/동적데이터 인식에 공통적으로 사용할 수 있으므로 적용 가능한 문제의 범위가 넓다. 물론 RNN(recurrent neural network) 또는 LSTM(long short-term memory)^[9] 등의 신경망을 이용하여 데이터 순서의 패턴을 파악하고 다음 순서를 예측하는 기능을 구현할 수 있으나 이는 단순한 데이터의 순서로 실제 물리적인 시간 상 정

REFERENCES

- [3] P. A. Merolla *et al.*, Science **345**, 668 (2014).
- [4] M. Davies *et al.*, IEEE Micro **38**, 82 (2018).
- [5] A. Necker *et al.*, Proc. IEEE **107**, 144 (2018).
- [6] B. V. Benjamin *et al.*, Proc. IEEE **102**, 699 (2014).
- [7] N. Q. S. Moradi *et al.*, IEEE Trans. Biomed. Circuits Syst. **12**, 106 (2018).
- [8] D. S. Jeong and C. S. Hwang, Adv. Mater. 1704729 (2018).
- [9] S. Hochreiter and Schmidhuber, Neural Comput. **9**, 1735 (1997).

5) 뉴로모픽 하드웨어를 이용한 연산.

6) 아날로그 회로는 회로상 노드의 연속적인 전압 값을 입출력으로 취하나 디지털 회로는 노드의 전압 값을 ‘0’과 ‘1’로 변환하여 디지털 연산을 함.

Table 1. Comparison between state-of-the-art neuromorphic chips.

Prototype	TrueNorth ^[3]	Loihi ^[4]	Neurogrid ^[6]	DyNAP ^[7]
Technology	28 nm	14 nm	180 nm	180 nm
Die size	4.3 cm ²	60 mm ²	1.7 cm ²	43.8 mm ²
Neuron type	Digital	Digital	Mixed	Mixed
# neurons Fan-in/out	256/256	-	-	64/4k
Embedded learning	N	Y	N	N
Real-time learning	N	N	N	N

의된 데이터가 아니다. 따라서 실제 물리적인 시간 상 정의된 데이터의 시간 상 데이터 패턴을 인식할 수 있는 모델은 스파이킹 신경망이 유일하다. 따라서 기능적인 측면에서 답러닝은 뉴로모픽 컴퓨팅의 부분집합⁷⁾으로 간주할 수 있다.

뉴로모픽 기술의 현재 동향

과거 아날로그 VLSI 회로 기반 뉴로모픽 하드웨어가 주류를 이루었다면 현재는 디지털 VLSI 회로 기반 기술 또는 디지털과 아날로그 혼성신호 회로 기반 기술이 주류를 이루고 있다. 아날로그 회로는 회로 제작 공정 상 필연적으로 발생하는 트랜지스터 크기의 부조화(size mismatch), 도핑 농도의 불균일 등의 이유로 실제 동작 특성과 예상특성 사이의 차이가 보이며 이 차이는 트랜지스터 크기가 줄어들수록 더욱 두드러진다. 따라서 아날로그 회로 기반 뉴로모픽 하드웨어는 사용 트랜지스터의 소형화(고집적화)에 따른 회로 제작의 어려움을 내포하고 있다.

반면 디지털 회로 기반 뉴로모픽 하드웨어는 아날로그 회로와 달리 트랜지스터의 고집적화에 유리하며 높은 동작의 신뢰도를 보이는 장점이 있다. 따라서 미국 IBM, Intel 등에서 선보인 TrueNorth,^[3] Loihi^[4] 등의 뉴로모픽 하드웨어는 디지털 회로를 기반으로 하고 있다. 하지만 디지털 뉴로모픽 하드웨어는 아날로그 뉴로모픽 하드웨어 대비 단위 기능 구현에 더 많은 트랜지스터가 필요하므로 제작비용이 큰 단점이 있다.

디지털과 아날로그 회로의 장점을 동시에 취한 디지털/아날로그 혼성신호 회로 기반 뉴로모픽 하드웨어는 또 다른 접근 방법으로 스탠포드대학에서 발표한 Braindrop,^[5] Neurogrid^[6]는 그 대표적인 예이다. 표 1은 대표적인 뉴로모픽 하드웨어 프로토타입을 요약하고 있다.

임베디드 학습(embedded learning) 기능은 최근 뉴로모픽 하드웨어에서 구현하고자 하는 핵심 기능이다. 임베디드 학습은 자가학습으로도 불리며 외부기기의 도움 없이 뉴로모픽 하드웨어가 주어진 학습알고리즘을 이용해 스스로 학습하는 기능을 의

미한다. Intel의 Loihi 뉴로모픽 하드웨어는 임베디드 학습기능을 포함하고 있다. 임베디드 학습을 위해 적절한 하드웨어용 학습 알고리즘이 매우 중요하다. 하드웨어용 학습 알고리즘은 소프트웨어 학습 알고리즘과 매우 상이하다. 하드웨어용 학습 알고리즘은 소프트웨어 알고리즘과 달리 학습용 데이터의 사용에 있어서 회로적인 제약조건이 까다로우며 시공간상 로컬 데이터만을 사용한 학습을 목표로 한다. 반면 소프트웨어용 학습 알고리즘은 학습에 필요한 데이터의 사용에 크게 제약조건이 없다. 따라서 성공적인 임베디드 학습기능 구현을 위해 시공간 로컬 데이터만을 이용하는 학습 알고리즘의 개발이 필수적이며 현재 전 세계적으로 알고리즘 연구가 활발히 진행되고 있다.

뉴로모픽 하드웨어 개발에 있어서 중요한 항목은 높은 재구성도(reconfigurability)이다. 사용자별로 구현하고자 하는 스파이킹 신경망의 구성이 상이하다. 예를 들어 사용자마다 선호하는 스파이킹 신경망의 뉴런, 시냅스 개수, 그리고 다층의 신경망의 은닉층(hidden layer) 개수 등이 다르므로 뉴로모픽 하드웨어가 범용적인 스파이킹 신경망 플랫폼 기술로 성공하기 위해서 사용자가 쉽게 스파이킹 신경망을 재구성 가능해야 한다. 최근 발표된 순람표(lookup table) 기반의 매우 높은 재구성도를 가지며 임베디드 학습에 용이한 뉴로모픽 하드웨어 아키텍처^[10]는 뉴로모픽 기술개발의 동향을 보여주는 대표적인 예이다.

나오는 글

뉴로모픽 하드웨어는 인공지능기술의 플랫폼 하드웨어로 앞으로 인공지능 기술발전에 박차를 가할 것으로 예상되는 기술이다. 뉴로모픽 하드웨어를 포함한 뉴로모픽 기술의 발전을 위해 다학제 접근법이 반드시 필요하다. 뇌과학, 컴퓨터과학, 전자공학, 물리학, 수학 등의 기초, 응용학문의 종합적인 이해가 필요한 대표적인 융합기술이다. 향후 거대 시장이 열릴 것으로 예상되는 뉴로모픽 하드웨어의 원천기술확보를 통한 기술선점을 위해 융합사고에 기반한 실증적 연구개발이 필요하다.

REFERENCES

[10] V. Kornijcuk *et al.*, Adv. Mater. Technol. 1800345 (2018).

- 7) 시간 비의존적 모델은 시간 의존적 모델의 특수한 경우에 해당하므로 시간 의존적 모델은 시간 비의존적 모델을 포함함.
- 8) 스파이킹 신경망에서 학습은 뉴런을 연결하는 시냅스의 연결정도를 최적화하는 과정을 의미하며 이 과정에서 많은 양의 데이터가 사용된다. 시냅스는 시냅스전, 시냅스후 뉴런 사이에 위치하므로 이 두 뉴런의 정보가 연결정도 최적화에 중요한 데이터로 사용된다. 시간상 로컬 데이터는 과거가 아닌 현재의 데이터를 의미한다. 공간상 로컬 데이터는 한 시냅스의 연결정도를 최적화하기 위해 공간상 이 시냅스와 직접적으로 관련이 있는 시냅스전, 시냅스후 뉴런의 정보를 의미한다. 따라서 시공간상 로컬 데이터를 이용한 학습 알고리즘은 주어진 시냅스의 시냅스전, 시냅스후 뉴런의 현재 데이터만을 이용한 학습 알고리즘을 의미함.