creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

다중 멤리스티브 시냅스 소자를 활용한 뉴로모픽 시스템의 온라인 학습 향상 방법 분석

2021년 8월

서울대학교 대학원

재료공학부 재료공학전공

조영운

다중 멤리스티브 시냅스 소자를 활용한 뉴로모픽 시스템의 온라인 학습 향상 방법 분석

지도교수 김상범

이 논문을 공학석사 학위논문으로 제출함
2021년 8월

서울대학교 대학원
재료공학부 재료공학전공
조영운

조영운의 석사 학위논문을 인준함
2021년 8월

위 원 장 이윤석

부위원장 김상범

위 원 김세영

초 록

차세대 비휘발성 메모리 소자를 활용한 크로스 포인트 어레이(cross point array) 구조는 가중치 합(weighted sum)과 가중치 갱신(weight update) 등의 학습에 필요한 병렬 연산 기능을 온 칩 구현하기 위해 제안되어왔다. 그러나 아날로그 시냅스 소자가 가지는 비이상적인 특성들, 가중치 갱신과정의 비선형성 및 비대칭성, 컨덕턴스 단계의 제한적인 개수, 장치 편차 등이 잠재적으로 온라인 학습의 정확성을 저해시키고 있어 큰 문제로 대두되고 있다. 본 연구에서는 이러한 실제 소자가 가지는 비이상적인 특성들이 학습 정확도에 끼치는 영향을 살펴보고 다중 시냅스 구조를 활용한 훈련 최적화 알고리즘 적용과 선형 및 대칭 스케일링 전략(Linear and symmetric scaling strategy)을 제안하였다. 그리고 시냅스 소자의 병렬 연결형태를 갖는 다중 시냅스 구조의 사례 연구를 심층 분석하였다.

먼저 실제 소자의 아날로그 시냅스 특성의 측정 결과를 바탕으로 가상의 다중 시냅스 소자가 실제 시냅스 거동(realistic synaptic behavior)을 보이도록 교정하였다. 그 후 인공 신경망(Artificial Neural Network, ANN)기반의 다중 레이어 퍼셉트론 시뮬레이터(Multi-Layer Perceptron Simulator, MLP simulator)를 실험에 바탕이 되는 뉴럴 네트워크 구조로 하여 MNIST 손글씨 데이터를 통해 온라인 학습의 정확도를 평가하였다. 또한, 선행 연구에서 지적하지 못하였던 다중 시냅스 구조가 가지는 큰 장점인 비이상적인 특성을 완화 시키는 현상을 면밀히 분석하였으며 결과적으로 선형성 개선 및 컨덕턴스 상태의 개수를 증가함을 확인하였다. 이러한 결과를 바탕으로 전통적인 훈련 알고리즘인 확률적 경사 하강법(stochastic gradient decent, 이하 SGD)이 아닌 모멘텀(Momentum), RMSprop, Adam(Adaptive Estimation method)등의 다양한 훈련 최적화 기법(Training optimizer)을 사용하여 다중 시냅스 구조의 성능을 평가하였다. 그 결과 다중 시냅스 구조와 훈련 최적화 기

법은 융합적으로 사용할 수 있으며, 특히 Adam 옵티마이저와 사용하였을 경우 비이상적인 특성으로 30%의 낮은 학습 정확도를 가지는 소자의 경우에 89%까지 높일 수 있었다. 그러나 다중 시냅스 구조는 전체 뉴로모픽 회로의 크기가 커지고 복잡해지는 문제점이 있었으며 훈련 최적화 기법 또한 하드웨어 구현 문제가 남아있었다.

이러한 문제점을 해결하기 위해 우리는 다중 시냅스 구조가 컨덕턴스 상태의 개수를 증가시키는 현상에 착안하여 선형성 및 비대칭성 스케일링 전략을 제안하였고 결과적으로 비이상적인 특성이 심한 소자의 온 칩 러닝의 검증 정확도를 30%에서 94%까지 높일 수 있었다.

주요어: 차세대 비휘발성 메모리 소자 (Emerging nonvolatile memory device, eNVM), **다중 시냅스 구조**(multi-memristive synapse), **비이상적 특성**(non-ideal properties), **비선형성**(non-linearity), **비대칭성**(asymmetry), **컨덕턴스 상태의 수**(number of conductance states), **장치 편차**(device variation), **학습 정확도**(learning accuracy)

학 번 : 2019-24121

목 차

Chapter 1 서론 ····· 1

Chapter 2 문헌 연구 ····· 4
2.1 다중 퍼셉트론 시뮬레이터 ····· 4
2.2 다중 멤리스티브 시냅스 소자 연구 ····· 19

Chapter 3 실험 방법 ····· 23
3.1 다중 시냅스 구조의 컨덕턴스 반응 분석 ····· 23
3.2 가중치 변화량 분석 ····· 28
3.3 유효 컨덕턴스 상태 수 ····· 32
3.4 훈련 최적화 기법에 따른 학습 성능 ····· 34
3.5 선형 및 대칭 스케일링 방법 ····· 36

Chapter 4 결과 및 고찰 ····· 23
4.1 다중 시냅스 구조 검증 성능 결과 ····· 42
4.2 선형 및 대칭 스케일링 방법 학습 결과 ····· 55

Chapter 5 요약 및 결론 ····· 59

참고문헌 ····· 61
Abstract ····· 66

표 목 차

Table 1.1 시냅스 소자의 특성과 학습 정확도 분석 ····· 18
Table 4.1 실제 소자의 비이상적인 특성 평가 및 다중 시냅스 구조와 Adam 옵티마이저의 성능 비교 ····· 54

그 림 목 차

Figure 2.1 MLP 시뮬레이터의 동작 흐름도 ····· 5
Figure 2.2 아날로그 eNVM 소자의 거동 모델 ····· 8
Figure 2.3 아날로그 eNVM 시냅틱 코어 및 크로스바 어레이 구조 ····· 11
Figure 2.4 2중 층 MLP 뉴럴 네트워크의 구조 ····· 13
Figure 2.5 2중 층 MLP 뉴럴 네트워크의 하드웨어 회로 구현 모식도 ····· 14
Figure 2.6 비이상적인 특성 값에 따른 온라인 학습의 정확도 분석 ····· 17
Figure 2.7 다중 시냅스 구조의 가중치 합 과정 ····· 21
Figure 2.8 소자 선택 카운터를 통한 가중치 갱신 중재 모식도 ····· 22
Figure 3.1 선형적인 소자의 경우, 다중 시냅스 구조의 컨덕턴스 반응 분석 ····· 26
Figure 3.2 비선형적인 소자의 경우, 다중 시냅스 구조의 컨덕턴스 반응 분석 ····· 27

Figure 3.3 NL(LTP/LTD)=1.0/-1.0인 경우, 가중치 변화량 비율 분포 분석 ····· 30
Figure 3.4 NL(LTP/LTD)=3.0/-3.0인 경우, 가중치 변화량 비율 분포 분석 ····· 30
Figure 3.5 NL(LTP/LTD)=5.0/-5.0인 경우, 가중치 변화량 비율 분포 분석 ····· 31
Figure 3.6 NL(LTP/LTD)=2.4/-4.88인 경우, 가중치 변화량 비율 분포 분석 ····· 31
Figure 3.7 N의 개수에 따른 학습에 요구되는 최소 컨덕턴스 상태의 개수 ····· 33
Figure 3.8 아날로그 시냅스 소자의 선형 및 대칭 구간을 설계한 모식도 ····· 39
Figure 3.9 선형 및 대칭 스케일링 방법의 소프트웨어 흐름도 ····· 41
Figure 4.1 NL(LTP/LTD)=0.0/0.0인 경우 다중 시냅스 구조의 정확도 결과 ····· 43
Figure 4.2 NL(LTP/LTD)=2.0/-2.0인 경우 다중 시냅스 구조의 정확도 결과 ····· 43
Figure 4.3 NL(LTP/LTD)=3.0/-3.0인 경우 다중 시냅스 구조의 정확도 결과 ····· 44
Figure 4.4 NL(LTP/LTD)=5.0/-5.0인 경우 다중 시냅스 구조의 정확도 결과 ····· 44
Figure 4.5 모멘텀 옵티마이저와 다중 시냅스 구조를 활용한 정확도 결과 ····· 45
Figure 4.6 RMSProp 옵티마이저와 다중 시냅스 구조를 활용한 정확도 결과 ····· 46

Figure 4.7 Adam 옵티마이저와 다중 시냅스 구조를 활용한 정확도 결과 ····· 47
Figure 4.8 AlO_x/HfO_2 소자, 다중 시냅스 구조 및 Adam 옵티마이저를 통한 학습 정확도 ····· 48
Figure 4.9 HZO-FeFET 소자, 다중 시냅스 구조 및 Adam 옵티마이저를 통한 학습 정확도 ····· 49
Figure 4.10 PCMO 소자, 다중 시냅스 구조 및 Adam 옵티마이저를 통한 학습 정확도 ····· 50
Figure 4.11 epiRAM 소자, 다중 시냅스 구조 및 Adam 옵티마이저를 통한 학습 정확도 ····· 51
Figure 4.12 TaO_x/HfO_x 소자, 다중 시냅스 구조 및 Adam 옵티마이저를 통한 학습 정확도 ····· 52
Figure 4.13 Ag:a-Si 소자, 다중 시냅스 구조 및 Adam 옵티마이저를 통한 학습 정확도 ····· 53
Figure 4.14 NL(LTP/LTD)=1.0/-1.0인 경우, 선형 및 대칭 스케일링을 통한 학습 정확도 ····· 56
Figure 4.15 NL(LTP/LTD)=3.0/-3.0인 경우, 선형 및 대칭 스케일링을 통한 학습 정확도 ····· 57
Figure 4.16 NL(LTP/LTD)=5.0/-5.0인 경우, 선형 및 대칭 스케일링을 통한 학습 정확도 ····· 58

List of Abbreviations

eNVM	emerging nonvolatile device
MLP	multi layer perceptron network
SGD	stochastic gradient descent
LTP	long term potentiation
LTD	long term depression
G	conductance
W	synaptic weight
NL	Nonlinearity of synaptic device
L	the number of conductance states
N	the number of device per 1-synapse
N-G	multi-synapse architecture, N synapse devices
C2C var	cycle to cycle variation
accuracy(%)	learning accuracy

Chapter 1. 서 론

뉴로모픽 컴퓨팅(Neuromorphic computing)은 인간의 두뇌의 주요한 기능을 모방한 초저전력으로 작동하여, 기존의 CMOS 기반의 불리안 컴퓨팅(Boolean Computing)이 한계에 도달하면서 상대적으로 많은 이점을 가지고 있다고 보고되었다[1].

특히나, 전통적인 연산 구조인 폰 노이만 구조(von Neumann architecture) 방식 기반 연산의 한계를 극복할 수 있는 기술로 떠오르고 있다. 기존의 폰 노이만 연산은 컴퓨팅 유닛(computing unit)과 메모리 사이에 데이터 버스 통로(data-bus path)가 존재하였고 이에 따라 많은 연산이 요구되는 실시간 훈련 및 분류 과정에서 데이터 버스를 이용한 데이터 이동(data communication)에 병목 현상(bottle neck)이 일어나면서 흔히 “von Neumann-bottleneck”이라 불리는 문제점이 제기되고 있었다[2].

이에 따라 신경망을 구조로 하는 인공 지능 연산에서 대규모의 병렬 연산 결과인 막대한 양의 정보를 다루기 위하여 빠른 연산 속도와 높은 효율, 저전력 소모 등의 특징을 가지는 작은 면적의 시냅스 장치가 필요해졌다.

그리고 대규모의 뇌-영감 컴퓨팅(neuro-inspired computing)기반의 하드웨어 플랫폼, FACETs[3], Caviar[4], SpiNNaker[5], True North[6] 등등이 최근 몇 년 동안 발전해왔으나, 막대한 수의 시냅스 소자를 포함하는 뉴로모픽 시스템을 설립하기 위해서는 CMOS 기반의 메모리 소자의 큰 크기가 하드웨어 측면에서 과도한 비용으로 요구되었다[7].

따라서 뉴로모픽 컴퓨팅을 구현하기 위해 초고밀도 메모리 배열, 초고속 연산 속도, 낮은 프로그래밍 전압 등의 특성을 갖는 비휘발성 차세대 메모리 소자(emerging nonvolatile memory, eNVM)들이 연구되었다[8-13].

비휘발성 차세대 메모리 소자는 ON, OFF의 바이너리 상태(binary states)를 표현하는 디지털 시냅스로 사용될 수 있는데 단일 바이너리

소자가 모여 다단계(multi level)상태를 나타낼 수 있다. 반면 차세대 메모리 소자를 아날로그 시냅스 형태로 사용하게 되면 컨덕턴스의 값을 연속적으로 조절할 수 있어 다단계 컨덕턴스 상태(multi-level conductance states)를 표현할 수 있다[14]. 아날로그 소자의 다단계 컨덕턴스 상태는 온 칩 상황의 학습 알고리즘에서 뉴럴 네트워크의 가중치 요소(weight element)로 대변될 수 있으며 크로스 포인트 배열(cross-point array)구조는 가중치 행렬(weight matrix)을 대변하여 가중치 합(weighted sum)과 가중치 갱신(weight update)의 병렬 연산을 수행할 수 있다[15-17].

그러나 비휘발성 차세대 메모리 소자를 아날로그 시냅스로 사용하게 되면 디지털 시냅스로 사용한 경우나 기존의 CMOS 기반 디지털 회로에 비교하여 신뢰성이 떨어지는데 이는, 비휘발성 차세대 메모리 소자가 가지는 비이상적인 특징들인 제한된 컨덕턴스 ON/OFF 비율, 비선형성 및 비대칭성, 장치의 편차(variation) 및 노이즈(noise) 그리고 제한된 컨덕턴스 상태의 개수 등으로부터 기인한다[18]. 이와 같은 아날로그 시냅스 소자의 비이상적인 특징들은 온 칩 학습의 정확도를 저해시키는 주요한 문제점으로 제기된다.

따라서 뉴로모픽 컴퓨팅을 통한 실시간 훈련 및 분류 과정을 궁극적으로 실현하기 위하여 차세대 메모리 소자의 비이상적인 특성을 완화 시킬 수 있는 회로 및 알고리즘 단계의 기술을 개발하는 연구가 진행되었다.

그 중, 차세대 메모리 소자인 상변화 메모리(phase change memory)를 활용한 연구에서는 상변화 메모리가 가지는 근본적인 특성인 비대칭적인 컨덕턴스 반응을 조절하기 위해 두 개의 상변화 메모리 소자가 differential pair로 구성하여 컨덕턴스의 차이를 가중치 요소로 표현하였다[19].

반면 다중 멤리스티브 시냅스(multi-memristive synapse) 구조를 제안한 연구에서는 상변화 메모리를 여러 개로 병렬연결한 뒤 컨덕턴스의 합을 하나의 가중치 요소로 나타내어 non diffrential 구조에서도 학습을 성공적으로 진행하였다[20].

우리는 다중 멤리스티브 시냅스 구조가 상변화 메모리라는 특수한 경우

뿐만 아니라 다양한 아날로그 시냅스 소자에서도 일반적으로 사용될 수 있는 방법임에 착안하였으며 비이상적인 특성들을 완화 시킬 수 있을 것으로 예상하였다.

따라서 본 연구에서는, 아날로그 소자의 특성 데이터를 추출 및 검정하여 실제 시냅스 소자처럼 컨덕턴스 반응을 보이는 가상의 시냅스 회로를 설계하고 이를 다층 레이어 퍼셉트론 시뮬레이터인 MLP Simulator Neurosim[21]에 다중 시냅스 소자 구조를 추가하여 온 칩 러닝을 시뮬레이션하였으며 특히 아날로그 시냅스 소자의 다양한 비이상적인 특징 중 비선형성 및 비대칭성, 제한된 컨덕턴스 상태의 개수, 장치 편차 등에 초점을 맞추어 다중 시냅스 구조를 평가하였다.

그리고 기존에 사용되는 훈련 최적화 알고리즘인 확률적 경사 하강(Stochastic Gradient Descent, SGD) 방식이 아닌 학습 방향과 학습률을 훈련 과정에서 적절하게 조절할 수 있는 모멘텀(Momentum)[22], Adagrad(Adaptive Gradient)[23], RMSProp, Adam(Adaptive Momentum Estimation)[24] 등의 방식을 알고리즘 단계에서 다중 시냅스 회로에 적용하여 온 칩 학습 성능이 더욱 개선됨을 확인하였다.

반면, 위의 훈련 최적화 방식들은 뉴로모픽 시스템의 하드웨어로 구현할 기술이 현재까지 제안되지 않았으며 다중 시냅스 구조 또한 병렬 연결된 시냅스 소자의 개수가 많아지면 회로의 복잡성 및 크기가 커지는 문제점을 가진다고 판단하였다.

이러한 문제점들을 고려하여 비이상적인 특성에서 기인한 학습 성능이 저해되는 문제를 해결하기 위해 회로 및 알고리즘 단계에서 새로운 훈련 모델이 필요하였다.

본 연구에서는 다중 시냅스 구조가 가지는 컨덕턴스 단계 수 증가 효과를 활용한 선형 및 대칭 스케일링 전략(Linear and symmetric scaling strategy)을 제시함으로써 학습 성능을 개선할 수 있었다.

Chapter 2. 문헌 연구

2.1 다층 퍼셉트론 시뮬레이터

2.1.1 뉴로심의 구조

선행 연구에서 시뮬레이터 플랫폼으로 PRIME[25]. Harmonica[26], ISSAC[27]등이 효과적으로 시스템 단계 설계의 유연성 및 가능성을 검증하였지만 eNVM 소자의 비이상적인 특징을 고려하지 못하였다.

우리는 뇌-영감 구조를 회로 단계에서 설계하며 아날로그 시냅스 소자의 학습 성능을 평가하는데 보조할 수 있는 시뮬레이터가 필요하였으며 오픈 소스로 공유되는 ANN 시뮬레이터인 뉴로심(NeuroSim)을 커스터마이징하였다. 뉴로심은 뉴럴 네트워크의 머신 러닝, 딥러닝을 보조해주는 시뮬레이션 프레임워크이며, 시뮬레이터에 입력 값으로 메모리 타입, 비이상적인 소자 매개 변수, 트랜지스터 테크놀로지 노드(transistor technology node), 네트워크 형상(network topology), 훈련 데이터 세트(training dataset) 등을 조정 할 수 있다. 이후 시뮬레이터의 출력 값은 회로 단계의 전반적인 성능인 면적, 레이턴시, 동적 에너지, 누설 전력 등이 있으며 알고리즘 단계에서는 학습 정확도를 계산한다(Fig 2.1).

뉴로심은 크게 세 가지 단계로 분류되는데, 먼저 소자 단계(device level)에서 시냅스 소자의 특성을 고려하여 뉴로모픽 시스템을 설계할 수 있다. 또한, 회로 단계(circuit level)에서는 시냅스 소자의 배열(synaptic array)과 주변 회로(periphery circuit)등이 구성되어 시냅스 코어의 디자인 옵션을 바꾸어 줄 수 있다. 마지막으로 알고리즘 단계(algorithm level)에서는 학습 알고리즘을 조정할 수 있으며 기본적으로 제공하는 뉴럴 네트워크는 2중 층 퍼셉트론 구조이다[28]. 이는 하나의 층에 속한 각각의 뉴런 노드가 다음 층의 모든 뉴런 노드와 연결된 완전 연결 네트워크 (fully connected network)이다.

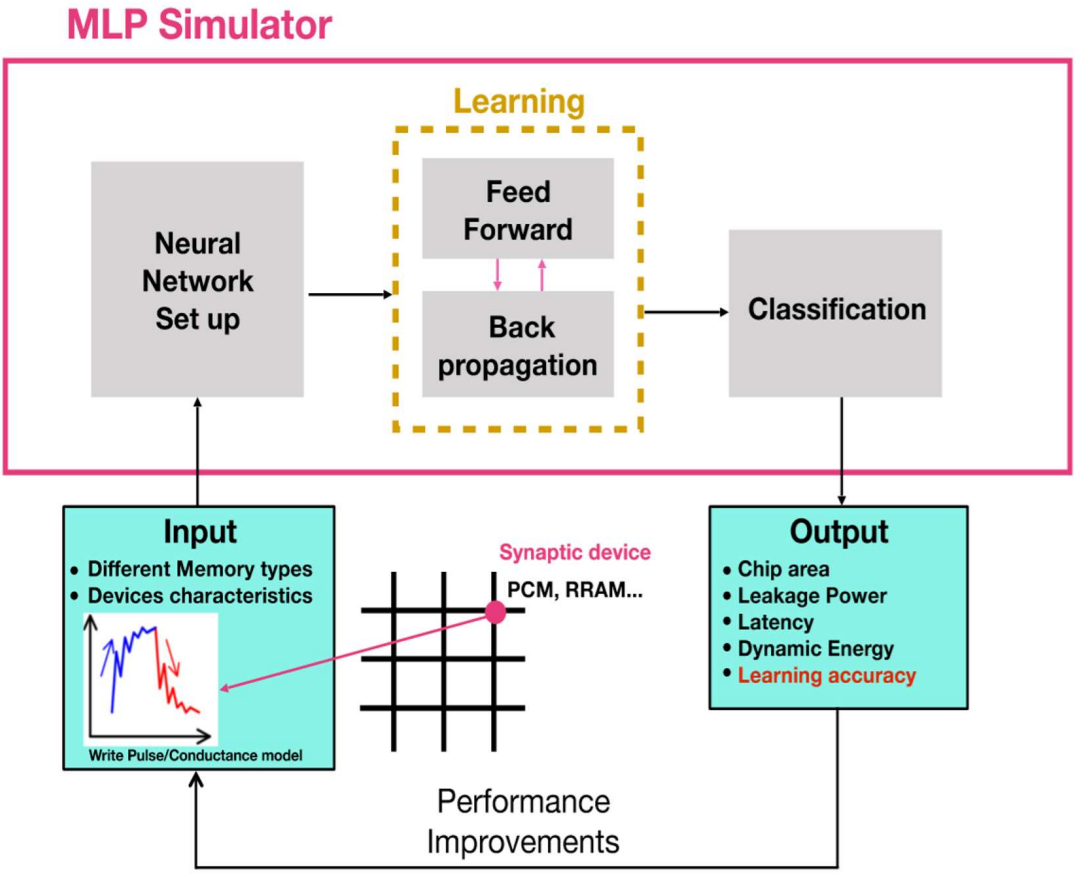

Fig 2.1. MLP Simulator의 작동 흐름도. 사용자가 아날로그 시냅스 특성을 입력해주면 뉴럴 네트워크가 형성되고 훈련 및 분류 과정을 거쳐 전반적인 회로 성능을 평가한다.

2.1.2 소자 단계: 시냅스 소자의 특성

실제 아날로그 시냅스 소자가 가지는 비이상적인 특징을 소개하고 이에 대한 분석 방법을 살펴볼 것이다. 먼저, 이상적인 뉴로모픽 소자는 가중치가 증가하는 장기 기억(long term potentiation, LTP)과 가중치가 감소하는 장기 억제(long term depression, LTD) 과정에서 입력한 쓰기 펄스(write pulse)의 개수에 비례하여 가중치가 선형적으로 변화한다. 그러나 실제 아날로그 소자는 이러한 이상적인 궤도를 따르지 않으며 LTP와 LTD 초기 단계에서 급격하게 변화하는 것이 특징적이다[29-31].

2.1.2.1 Nonlinear weight update

선행 연구에서는 실제 아날로그 소자의 비선형적인 가중치 갱신 거동을 포착하기 위해 펄스의 개수(P)에 관한 지수 함수로 표현하였으며 아래 수식과 같다[32].

Conductance for LTP

$$G_{LTP} = B(1 - e^{-\frac{P}{A}}) + G_{\min} \tag{2.1}$$

Conductance for LTD

$$G_{LTP} = -B(1 - e^{\frac{P - P_{\max}}{A}}) + G_{\max} \tag{2.2}$$

B is function of A within range of G

$$B = (G_{\max} - G_{\min}) / (1 - e^{-\frac{P_{\max}}{A}}) \tag{2.3}$$

(여기서 Gmax, Gmin Pmax는 각각 최대 컨덕턴스, 최소 컨덕턴스 그리고 최소 컨덕턴스 상태에서 최대 컨덕턴스 상태로 바꾸는데 필요한 최대 펄스 개수를 의미한다. Gmax, Gmin Pmax는 소자의 측정 데이터에서 추출하여 분석할 수 있다.)

따라서 수식 (2.1)-(2.3)을 통해, 가상의 시냅스 소자의 가중치 증가(파란색)와 가중치 감소(빨간색)의 거동을 Fig2.1에 나타내었다. 수식 (2.1)-(2.3)에 해당하는 A 값이 이 그림에 표기된 -6~6의 숫자에 해당하며 A가 변화함에 따라 비선형적인 곡선이 조절됨을 확인하였다.
본 연구에서는 A 값이 비선형성을 결정하는 정도라고 판단하였으며 이후 실험에서 가상의 시냅스 소자를 설계 단계에 이 A 값을 조정해주며 실험을 진행하였다.

다음으로 비선형성 이외에 다른 비이상적인 특성들인 컨덕턴스 상태의 개수(Precision), 장치 사이 편차(Device-to-device variation), 사이클 사이 편차(Cycle-to-cycle variation), 동적 범위(Dynamic range)에 대한 분석 방법을 아래에 설명하였다.

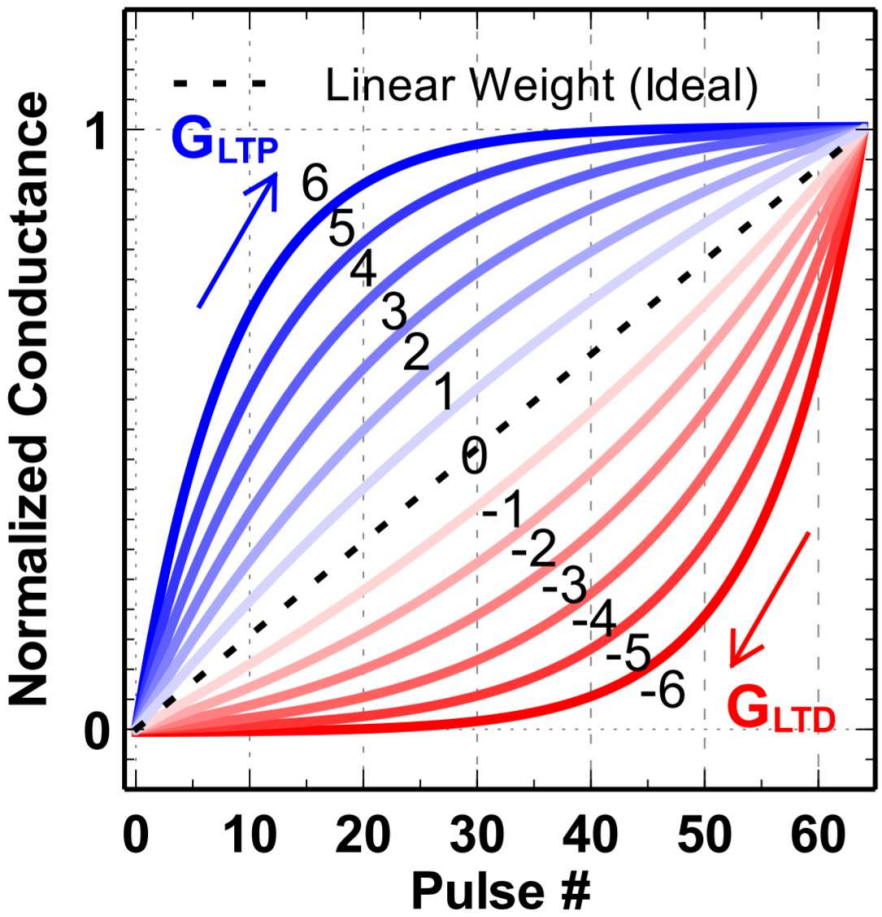

Fig 2.2 아날로그 eNVM 소자의 거동 모델, 가중치 갱신과정에서 비선형성 정도를 -6과 6 사이 값으로 표기하였다[33]. (Adopted, with permission, from P. Chen, et al, NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. © 2017, IEEE.)

2.1.2.2 Precision

Precision은 아날로그 eNVM 소자의 컨덕턴스 상태의 개수에 결정되며 수식 (2.1)-(2.3)에서 P_{max}에 해당한다.

2.1.2.3 Device-to-device variation

Device-to-device variation은 뉴로모픽 소자와 다른 소자 사이의 편차이며 Fig 2.1에서 분석한 비선형성의 기준에 따라 장치 사이의 표준 편차를 계산하였다.

2.1.2.4 Cycle-to-cycle variation

사이클과 사이클 사이 편차는 매번의 프로그래밍 펄스를 통해 변화하는 컨덕턴스의 편차로, 이 값은 전체 컨덕턴스 범위의 백분율로 표현한다.

2.1.2.5 Dynamic range

동점 범위는 최대 컨덕턴스와 최소 컨덕턴스의 비율(ON/OFF ratio)로 계산한다. 이상적으로 eNVM 소자의 시냅스 가중치 값의 범위는 0부터 1이 되어야 하지만, ON/OFF ratio가 충분히 크면 최소 컨덕턴스에 해당하는 가중치 값을 0으로 볼 수 있다.

2.1.3 회로 단계: 시냅스 코어와 어레이 구조

시냅틱 코어(Synaptic core)는 뉴로모픽 컴퓨팅의 핵심 기술인 가중치 합과 가중치 갱신과정의 병렬 연산을 효율적으로 수행하기 위해 설계된 연산 단위 (computation unit)이다. 시냅틱 코어는 시냅스 어레이(synaptic array)와 주변 회로 배열(array periphery)로 구성되며 시냅스 어레이 종류와 크기, 시냅스의 개수 등을 변수로 고려하여 병렬 연산 기능을 단독으로 수행한다[34].

뉴로심은 세 가지 시냅틱 코어를 소개하였는데 이는 SRAM, 디지털 eNVM, 아날로그 eNVM 등이 있다. 그중에서 본 연구에서 다룰 아날로그 시냅스 배열구조를 살펴보면, 아날로그 eNVM 소자들을 크로스바 어레이(cross-bar array) 구조[35]로 형상하여 가중치 행렬을 나타내도록 하였다. 이 구조는 각각의 eNVM 소자가 Fig2.3의 가로 전선에 해당하는 워드 라인(WL)과 세로 전선에 배치된 비트 라인(BL)의 교차점(cross point)에 위치하게 된다. 만약, 인풋 벡터가 읽기 신호로부터 인코딩되면 이 신호가 워드 라인을 통해 각각의 시냅스 소자에 전달되고 비트 라인을 통해 흘러나간 전체 전류를 읽어내어 시냅스 가중치 합 연산(matrix-vector multiplication)을 병렬적으로 수행할 수 있다.

본 연구에서는 Fig 2.3 아날로그 eNVM 소자를 크로스바 배열을 한 시냅틱 코어를 골조로 하여 이후 소개할 다중 시냅스 구조, 즉 하나의 시냅스에 여러 개의 단일 소자가 병렬 연결된 구조를 설계하였다.

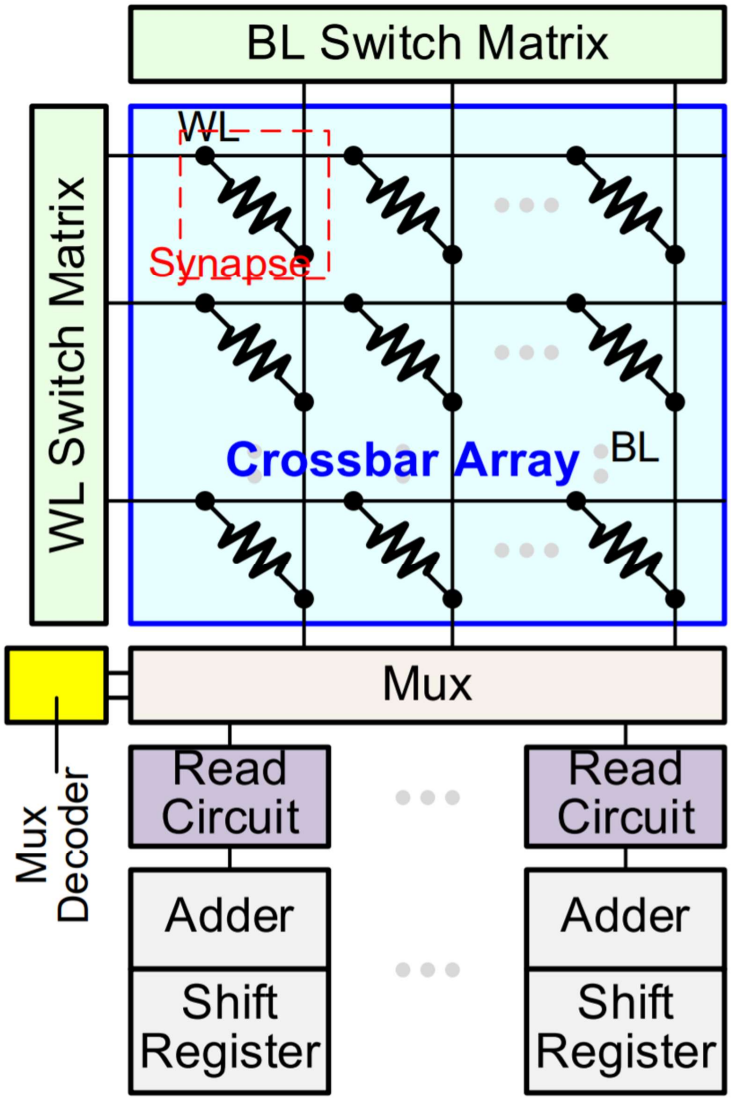

Fig 2.3. 크로스바 어레이 구조의 병렬화 기능을 수행하기 위한 아날로그 eNVM 시냅틱 코어[36]. (Adopted, with permission, from P. Chen et al, NeuroSim: A Circuit-Level Macro Model for Benchmarking Neuro-Inspired Architectures in Online Learning. © 2018, IEEE.)

2.1.4 알고리즘 단계 : 다중 퍼셉트론 레이어 뉴럴 네트워크

알고리즘 단계(algorithm level)에서는 400개의 MNIST 손글씨 데이터 입력층(Input layer), 100개의 은닉층(hidden layer) 그리고 10개의 출력층(Output layer)로 400-100-10의 2중 층 신경망 구조를 사용하였다.

완전 연결 계층 구조인 MLP는 두 개의 뉴런 노드 사이 연결에 해당하는 시냅스가 존재하며, 시냅스의 강도(strength)가 가중치 값을 나타낸다. 그리고 회로 설계의 복잡성을 해소하고 전체 연산량을 줄이기 위해 입력 데이터인 MNIST 이미지를 20 × 20으로 잘라낸 뒤 흰색과 검은색의 1-bit 데이터로 변환하였다(Fig 2.4).

훈련의 핵심적인 두 가지 동작은 순전파(feed forward)와 역전파(back propagation)이며, 순전파는 입력층에 들어온 입력 데이터가 정방향으로 이동하여 일련의 가중치 합 동작과 뉴런 활성화 함수(neuron activation function)을 거치며 최종적으로 출력층에 도달하게 된다. 출력층에 도달한 순전파의 결과는 출력층에 표기된 정답과 비교하여 예측 오차(prediction error)를 계산한다. 반면, 역전파는 계산된 오차가 출력층으로부터 역방향으로 전파되어 각 층의 가중치 값이 오차를 최소화하는 방향으로 조절된다.

그리고 검증(test)과정 혹은 오프라인 분류(offline classification)에서는 순전파 동작만 수행하며 검증 결과를 도출하기 때문에 시냅스 어레이의 가중치 값이 변화하지 않는다. 이 과정에서는 Heaviside step function과 같은 1-bit low-precision 활성화 함수로 가중치 합 연산이 가능하다. 그러나 훈련 과정에서는 가중치 갱신과정의 오차를 줄이기 위해 high precision 활성화 함수가 필요하다(Fig 2.5).

또한, 본 연구에서는 시냅스 어레이의 양의 가중치 값을 양과 음의 값을 모두 가지는 시냅스 가중치로 변환하였으며 다음 수식과 같다.

$$W_A V = 2(W_H - 0.5J)V = 2W_H V - JV \qquad (2.4)$$

(여기서 V는 입력 벡터, J는 WA와 WH의 크기가 같고 값이 1인 행렬이다. WHV는 시냅스 코어로부터 출력된 가중치 합이다. 따라서 WH(0~1)의 범위를 WA(-1~1)로 변환하였다.)

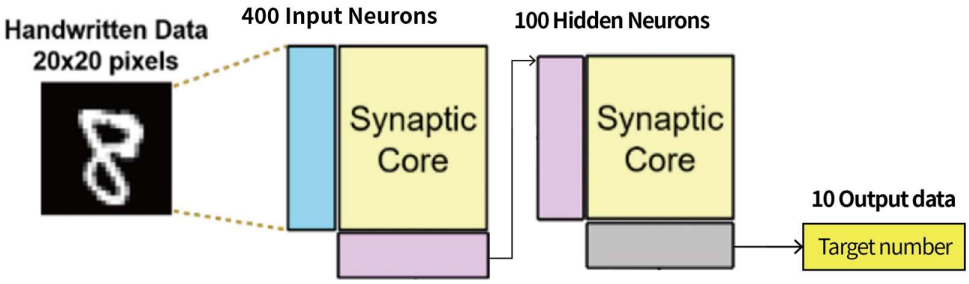

Fig 2.4. 2중 층 MLP 뉴럴 네트워크로 20 × 20 MNIST 훈련 데이터를 입력층에 전달 후 병렬 연산 기능을 수행하는 시냅틱 코어를 거친 뒤 10개의 타겟 숫자 중 하나를 결정한다.

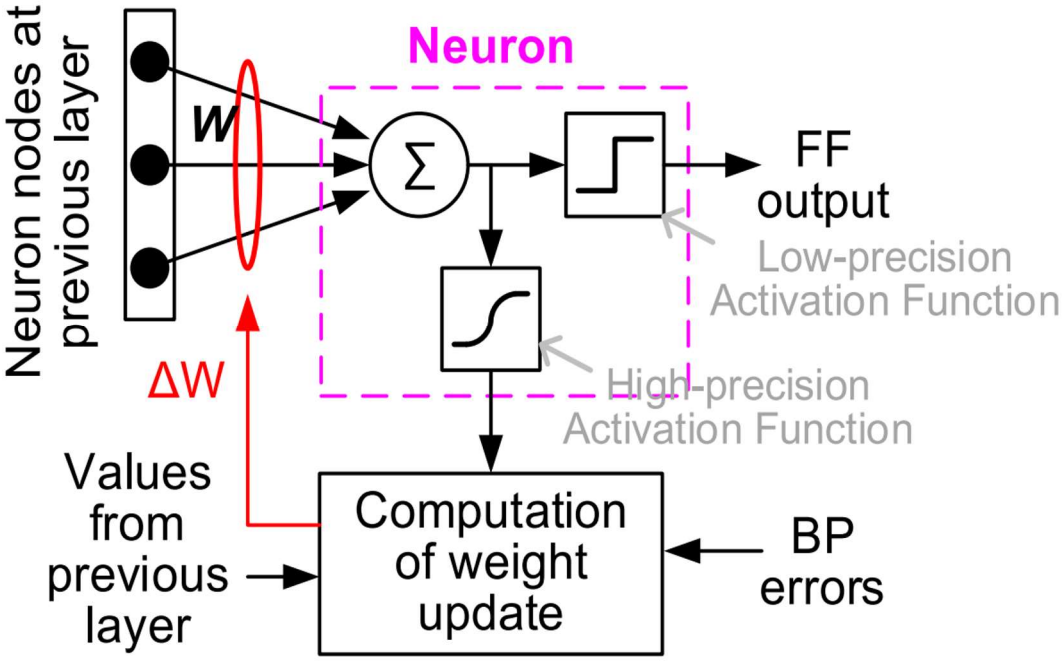

Fig 2.5. 2중 층 MLP 뉴럴 네트워크의 하드웨어 회로 구현 모식도[37] (Adopted, with permission, from P. Chen et al, NeuroSim: A Circuit-Level Macro Model for Benchmarking Neuro-Inspired Architectures in Online Learning. © 2018, IEEE.)

2.1.5 시냅스 소자의 특성이 정확도에 끼치는 영향

앞서 2.1.2절에서 언급한 아날로그 소자의 비이상적인 특징들이 실제 학습 성능에 끼치는 영향을 온라인 학습 및 오프라인 분류를 수행하여 분석한 선행 연구 결과이다(Fig 2.6).

2.1.4.1 Requirement of weight precision

아날로그 시냅스 소자는 다단계의 컨덕턴스 상태를 가지는데 학습에 요구되는 최소 컨덕턴스 상태의 개수를 예측하였다. 결과적으로 MNIST 데이터를 통한 온라인 학습에서는 6-bit weight이 요구되며, 오프라인 분류 과정에서는 2-bit weight만으로도 충분히 수행하였다.

2.1.4.2 Limited Conductance On/off ratio

온라인 학습을 원활히 진행하기 위해서는 ON/OFF ratio> 10, 오프라인 분류를 진행하기 위해서는 ON/OFF ratio> 50이 요구되었다.

2.1.4.3 비선형성 및 비대칭성

비선형성이 심하고 대칭적인 소자는 학습 정확도가 소폭 감소한 것에 비해 비대칭적인 소자는 비선형성이 증가할수록 급격하게 학습 성능이 저하되었다. 따라서 비대칭성이 학습 정확도를 결정하는 주요 원인이다.

2.1.4.4 Cycle-to-cycle variation

낮은 variation(<2%)에 대해서는 학습 정확도가 크게 저하되지 않았지만, 비교적 높은 variation(>2%)은 학습 성능을 악화시키는 결정적인 요인으로 작용하였다.

2.1.4.5 실제 시냅스 소자의 학습 정확도 평가

다양한 시냅스 소자에 대해 특성을 분류하였는데 특히, 학습 정확도에 큰 영향을 끼치는 비선형성 및 비대칭성, 컨덕턴스 상태의 개수, cycle to cycle variation에 따른 학습 정확도를 분석하였다(Table 1). 분류한 아날로그 eNVM 소자는 총 6가지며 먼저, Ag:a-Si[38] 소자는 시냅스 가중치 증가 혹은 감소 과정에서 높은 비선형성 및 비대칭성을 가지며 73%의 학습 정확도를 보여주었다. 그리고 TaO_x/TiO_2[39]소자는 cycle to cycle variation이 학습 정확도를 저해하는 주요 원인으로 작용하였다. 다음으로 PCMO[40]는 비선형성 및 비대칭성이 이들 중 가장 심한 소자로 낮은 학습 정확도(~30%)를 가졌다. EpiRAM[41]은 이들 중 이상적인 소자에 매우 가까웠으며 높은 학습 성능(92%)을 보여주었다. 또한, HZO-FeFET[42]은 낮은 컨덕턴스 상태의 수가 AlO_x/HfO_2[43] 소자는 높은 cyce-to-cycle variation으로 인해 준수한 선형성에도 불구하고 낮은 정확도를 보여주었다.

결과적으로, 뉴로모픽 회로에서 병렬 연산을 수행하는 과정 중 이러한 비이상적인 특성들이 가중치 갱신과정에서 큰 영향을 주어 학습 성능을 떨어뜨렸다고 유추할 수 있었다. 따라서 우리는 회로 및 알고리즘 단계에서 비이상적인 특성들이 학습에 끼치는 영향을 완화하는 연구가 필요하다고 판단하였다.

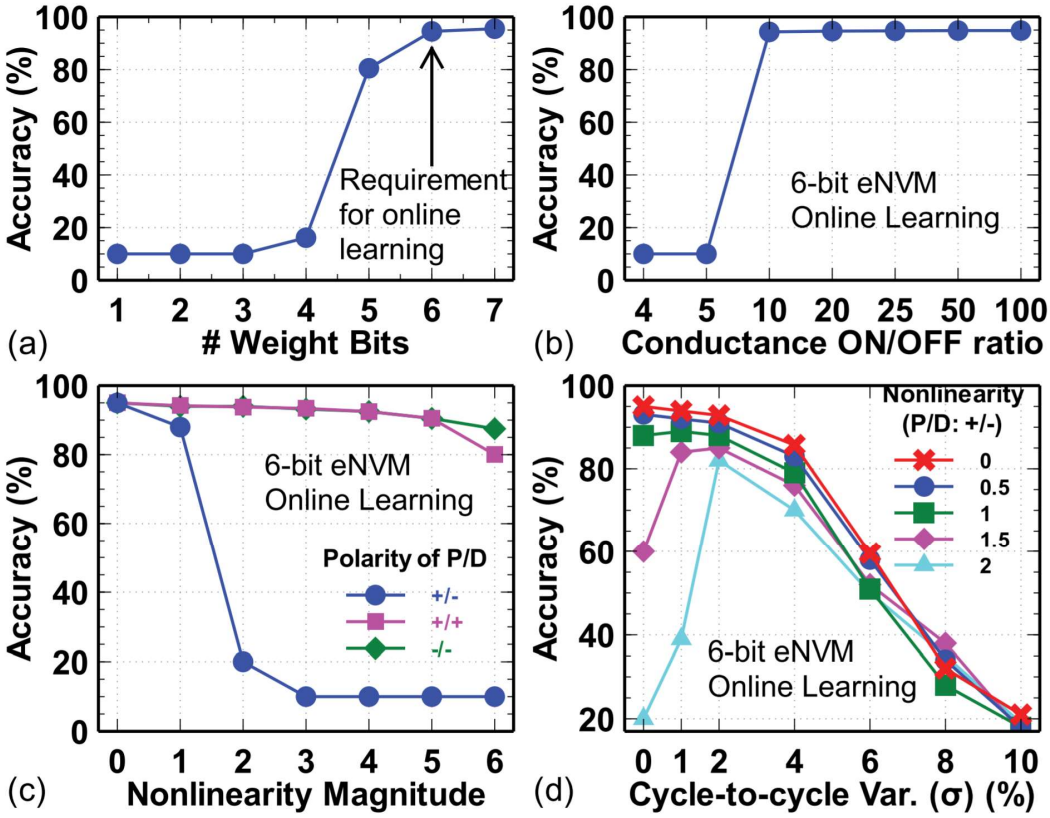

Fig 2.6. 온라인 학습에 끼치는 영향 (a) weight precision, (b) conductance ON/OFF ratio, (c) weight update asymmetry/nonlinearity, and (d) weight update cycle-to-cycle variation[44]. (Adopted, with permission, from P. Chen et al, Technological Benchmark of Analog Synaptic Devices for Neuroinspired Architectures," in IEEE Design & Test, © 2019, IEEE.).

소자 타입	**Ag:a-Si[38]**	**TaOx/HfOx[39]**	**PCMO[40]**	**EpiRAM[41]**	**HZO FeFET[42]**	**AlOx/HfO2[43]**
컨덕턴스 상태의 개수 (# of conductance states)	97	128	50	64	32	40
비선형성 및 비대칭성(LTP/LTD)	2.4/-4.88	0.04/-0.63	3.68/-6.76	0.5/-0.5	1.75/-1.46	1.94/-0.61
cycle to cycle variation	3.5%	3.7%	<1%	2%	<0.5%	5%
온라인 학습 정확도	72%	80%	30%	92%	88%	20%

Table 1.1. 시냅스 소자의 특성과 학습 정확도 분석표

2.2 다중 멤리스티브 시냅스 소자 연구

본 연구에서는 다중 시냅스 구조를 핵심 모델로 벤치마킹하여 아날로그 소자가 가지는 비이상적인 특성들 완화 시키는 방법을 구현하였다. 선행 연구인 다중 시냅스 구조는 하나의 시냅스에 단일 소자가 병렬로 연결된 방식을 취하는데 그러므로 시냅스가 가지는 가중치는 병렬 연결된 소자들의 컨덕턴스가 합해진 형태로 표현할 수 있다. 이를 수식으로 표현하면 다음과 같다.

$$W_{ij} \propto \left(\sum_{n=1}^{N} G_n \right) \tag{2.5}$$

여기서 Wij는 크로스바 어레이에서 i열, j행에 위치한 시냅스 가중치 값을 의미하며, N은 병렬로 연결된 시냅스 소자의 총 개수이다. 이를 통해 시냅스의 전반적인 동적 범위(dynamic range)와 분해능(resolution)이 향상된다.

그리고 가중치 합을 계산하기 위해 이전 층의 시냅틱 코어에서 뉴럴 활성화 함수를 거쳐 나온 전압이 다음 층의 시냅틱 코어의 N 개의 소자들이 구성하는 다중 시냅스에 입력된다. 그 후 각각의 소자에서 흘러나온 전류의 합을 통해 시냅스의 출력 값을 계산한다(Fig 2.7).

반면, 가중치 갱신과정은 특수하게 진행되는데 한 번의 프로그래밍 단계에서 N개의 소자 중 한 개를 선택한다. 이것은 소자 선택 카운터(device selection counter)를 통해 1과 N 사이 범위를 단계적으로 이동하여 이와 대응되는 하나의 소자를 선택한다.

따라서 가중치 갱신과정에서 가중치 변화량은 다음과 같은 수식을 따른다.

$$W_{ij} \propto \left(\sum_{n=1}^{N} G_n \right) + \triangle G \tag{2.6}$$

특히, 소자 선택 카운터는 고정된 비율로 증가하기 때문에 충분히 많은 가중치 갱신 동작이 이뤄지는 훈련 과정에서 N개의 소자가 균일하게 선택되는 것을 보장한다. 또한, 주의할 점은 한 시냅스 내 단일 소자가 반복적으로 가중치 갱신되는 것을 방지하기 위해서 소자 선택 카운터의 길이(Counter Length)는 다중 시냅스 소자의 개수 N과 서로소가 되어야 한다(Fig 2.8).

이외에도 독립적인 카운터로 potentiation counter 및 depression counter를 설계하여 LTP와 LTD의 주기를 조절할 수 있다. 위와 같은 방식은 상변화 메모리[45]와 비슷한 비대칭성이 심한 소자에 효과적인데 부분적인 SET 동작에서는 시냅스 가중치 값이 선형적으로 증가하는 반면 RESET 동작에서 급격하게 떨어지기 때문에 depression counter를 통해 주기를 조절하여 가중치 갱신과정을 보조할 수 있다.

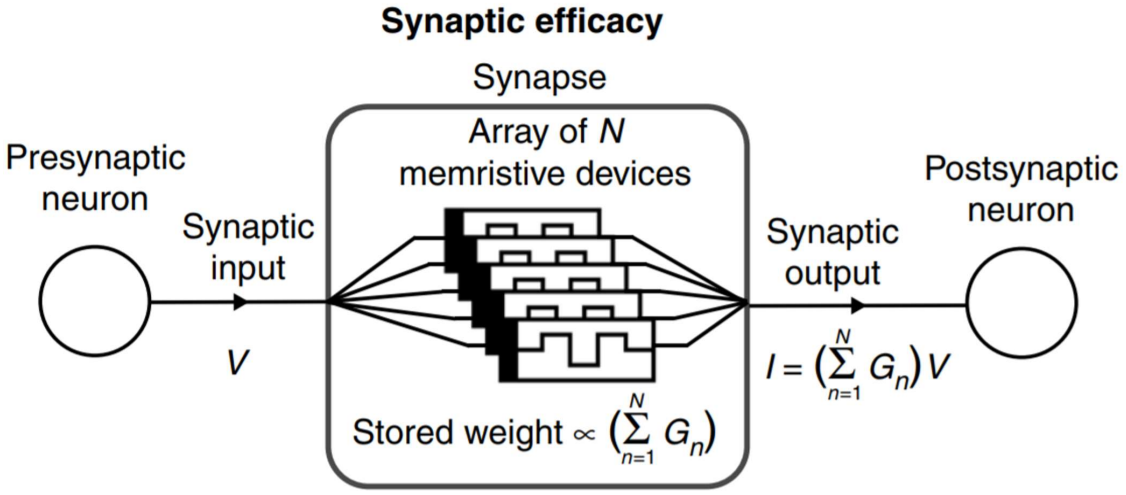

Fig 2.7. 전체 시냅스 가중치는 다중 시냅스 소자의 컨덕턴스를 결합한 값으로 표현한다. 입력 전압 신호 V가 전체 소자에 가해지면 각 소자의 흘러나온 전류를 합한 출력 신호가 다음 뉴런으로 전파된다[46]. (Adopted, with permission, from Irem Boybat et al, Neuromorphic computing with multi-memristive synapses, © 2018, Springer Nature.)

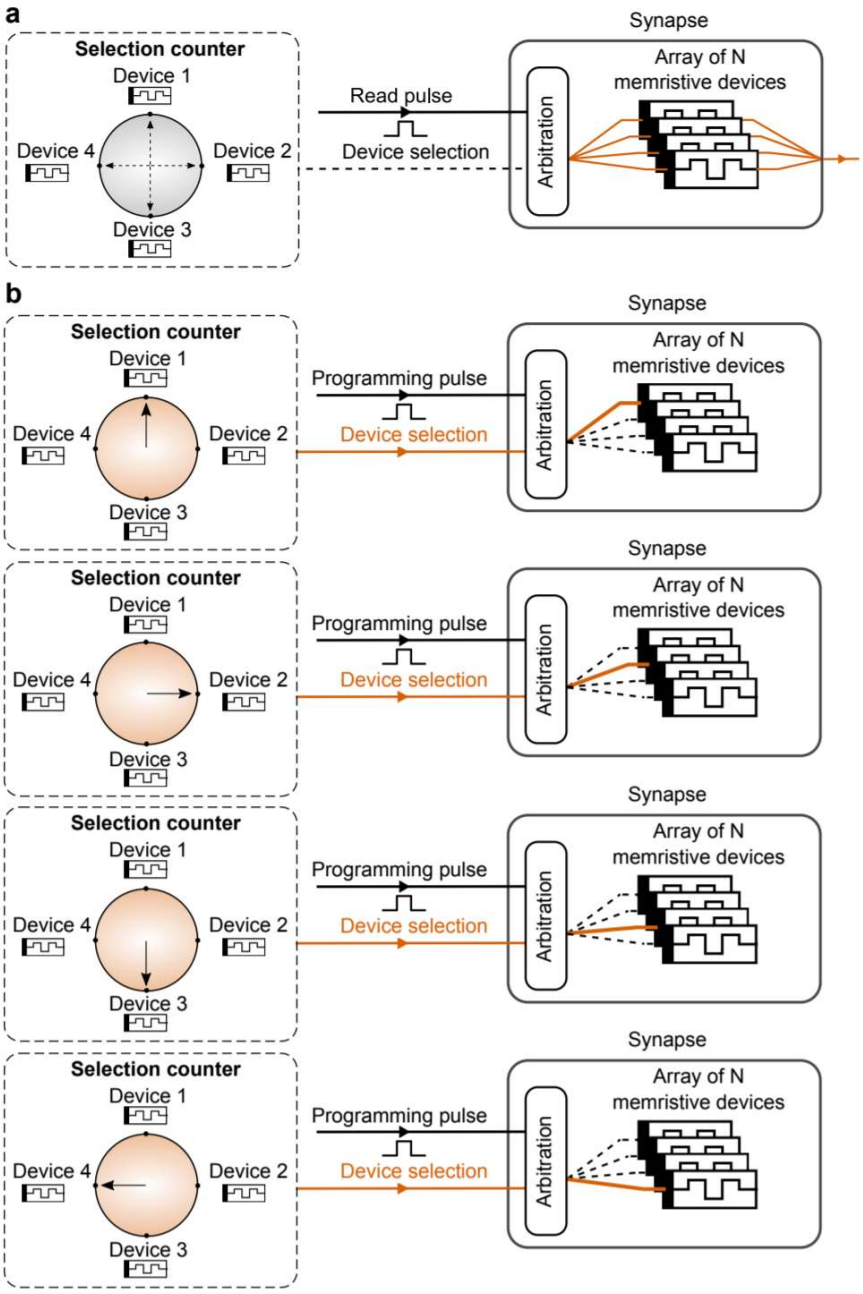

Fig 2.8. 소자 선택 카운터를 이용한 가중치 갱신 중재 모식도 [47](arbitration scheme). (Adopted, with permission, from Irem Boybat et al, Neuromorphic computing with multi-memristive synapses, © 2018, Springer Nature.)

Chapter 3. 실험 방법

3.1 다중 시냅스 구조의 컨덕턴스 반응 분석

3.1.1 시뮬레이션 모델 구성

훈련을 진행하기에 앞서 입력 신호에 따른 컨덕턴스 반응 시뮬레이션을 수행하였는데, 다중 시냅스 구조는 2.2절에서 언급한 소자 선택 카운터를 통해 전체 시냅스의 가중치 갱신이 이루어진다. 그러므로 시냅스 가중치 값을 측정하더라도 시냅스를 구성하는 각각의 소자에 대한 컨덕턴스 정보를 알 수가 없다. 특히, 단일 시냅스 구조의 선형성은 2.1.2.1절에서 실험 데이터를 통해 쉽게 판단할 수 있으나 다중 시냅스 구조는 여러 개의 소자가 통합되어 하나의 시냅스로 나타내기 때문에 선형성을 판단하기 힘든 문제가 따른다. 그 이유는 소자 선택이 이루어졌을 때 가중치 변화가 균일하게 이루어지는 것이 아니라 입력된 가중치 변화량에 따라 무작위적으로 증가 혹은 감소하여 어떤 소자가 어느 컨덕턴스 상태에 있는지 확정할 수 없기 때문이다. 따라서 학습에 막대한 영향을 끼치는 선형성을 판단하기 위해 다중 시냅스 소자의 컨덕턴스 반응이 단일 시냅스 소자와 비교하여 어떠한 차이점이 있는지 살펴보았다.

3.1.1.1 시뮬레이션 변수 및 환경

먼저, 소자 단계에서 두 가지 가상의 시냅스 소자를 설계하였으며 수식(2.1-2.3)과 Fig 2.1을 바탕으로 비선형성 크기를 나타내는 변수 NL(LTP/LTD)를 조정하였다. 그리고 NL(LTP/LTD)의 크기가 0.0/0.0인 이상적인 소자와 5.0/-5.0인 비선형성 및 비대칭성이 심한 소자를 기반으로 하여 시냅스 어레이를 구성하였다. 이렇게 구성된 시냅스 어레이에 속하는 소자의 개수 N을 독립 변수(independent variable)로 N=1, 2, 4,

8, 16 와 같이 점진적으로 늘려가며 컨덕턴스 반응의 변화를 정량화하였다. 아울러 실험에 영향을 끼치는 변수인 컨덕턴스 상태의 개수(L)는 64개, On/OFF ratio는 100, variation은 0으로 변인 통제하였다.

다른 한편, 시뮬레이션의 네트워크는 2.1.4절에서 설계한 구조와 상동인 2중 층 구조를 골조로 하여 입력층, 은닉층, 출력층 순으로 뉴런 개수가 400-100-10이며 총 41,000개의 시냅스를 가진다. 또한, 강조할 점은 앞으로 분석할 실험들에 대해 이와 일치하는 네트워크 구조를 사용하였다.

3.1.1.2 시냅스 가중치 갱신 방법

우리는 가중치 값이 증가하는 쓰기 펄스만 주어 단일 방향(unidirectional)으로 이루어지도록 하였다. 그 후 한 번의 가중치 갱신에 하나의 쓰기 펄스 신호(write pulse)를 주어 단계적으로 컨덕턴스 상태를 높일 수 있도록 하였다.

동시에, 소자 선택 카운터를 이용하여 N개의 소자 중 하나를 선택하여 가중치 갱신을 진행하였다. 그리고 훈련 과정과 유사한 상황을 재현하기 위해 소자 선택 카운터가 하나씩 증가하는 것이 아닌 무작위로 하나의 소자를 선택하도록 설계하였다. 그러므로 선택된 소자의 컨덕턴스 상태가 최대에 도달하지 않은 중간 단계(intermediate state)에서는 컨덕턴스의 단계가 하나씩 증가할 수 있다. 반면, 최대 컨덕턴스값에 도달한 포화 단계(saturation state)에 쓰기 펄스가 입력되면 컨덕턴스 단계가 더 증가할 수 없다.

3.1.2 시뮬레이션 모델 유효성 평가

시뮬레이션의 결과 선형적인 소자와 비선형적인 소자의 경우 모두 다중 시냅스 구조를 사용하였을 때 비선형성이 증가하는 경향을 보인다. 이를 살펴보면, 첫 번째로 선형적인 소자(Fig 3.1)는 N=1(검은색)일 때 입력된 신호에 비례하여 컨덕턴스값이 선형적으로 증가한다. 그러나 N=2(빨간색)일 때 비선형성 정도가 급격하게 커지며 N의 개수를 더욱 증가시키면 비선형성이 점차 감소하는 특이한 반응을 보인다. 마찬가지로 비선형적인 소자(Fig. 3.2)에서도 다중 시냅스 구조를 사용하면 선형성이 감소하는 현상을 보인다. 이러한 경향성이 나타난 원인은 소자 선택 카운터가 무작위로 하나의 소자를 선택하기 때문이며 최대 컨덕턴스값에 도달한 소자는 입력 신호에 반응하지 않아 선형성을 떨어뜨리게 된다

그렇지만 본 실험에서 진행한 가중치 갱신 방법은 실제 온라인 학습 과정에서 이루어지는 가중치 갱신과 근본적인 차이점이 존재한다. 그 이유는 본 실험은 가중치 값이 증가하기만 하는 단방향 소자를 가정하였는데, 실제 온라인 학습에서 아날로그 소자의 컨덕턴스값이 증가 혹은 감소하는 양방향 소자를 사용하기 때문이다. 게다가 실제 훈련에서는 소자 카운터가 무작위로 선택하는 것이 아닌 단계적으로 움직이고 가중치 변화량이 무작위 값으로 입력되기 때문에 이것을 모두 고려할 필요가 있다. 결과적으로 훈련 과정 중 다중 시냅스 구조가 갖는 선형성이 달라질 것으로 예상할 수 있으며 이것을 정량적으로 알아보기 위해 가중치 변화량 산포를 측정하였다.

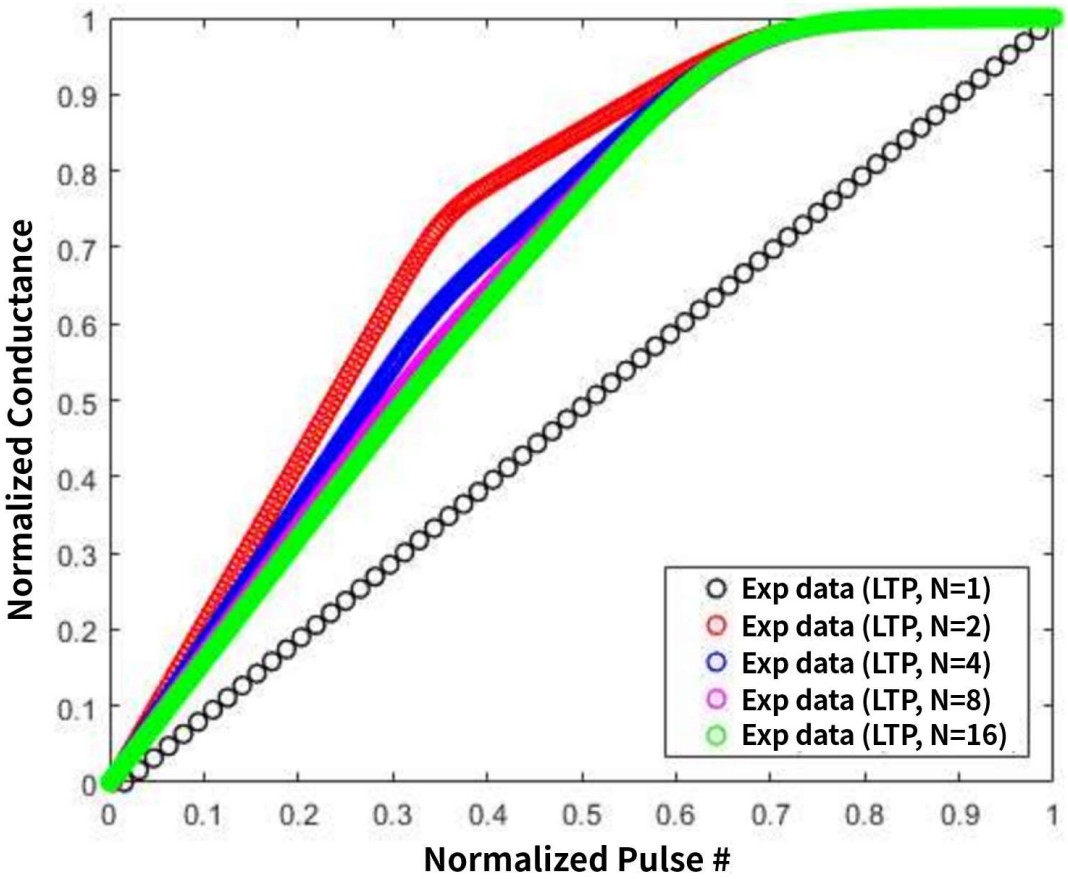

Fig 3.1 NL(LTP)=0(선형적인 소자)의 경우, 가상의 시냅스 소자를 이용한 Normalized pulse 개수 vs Normalized conductance 결과. N의 개수를 1, 2, 4, 8, 16으로 변화시켜가며 측정

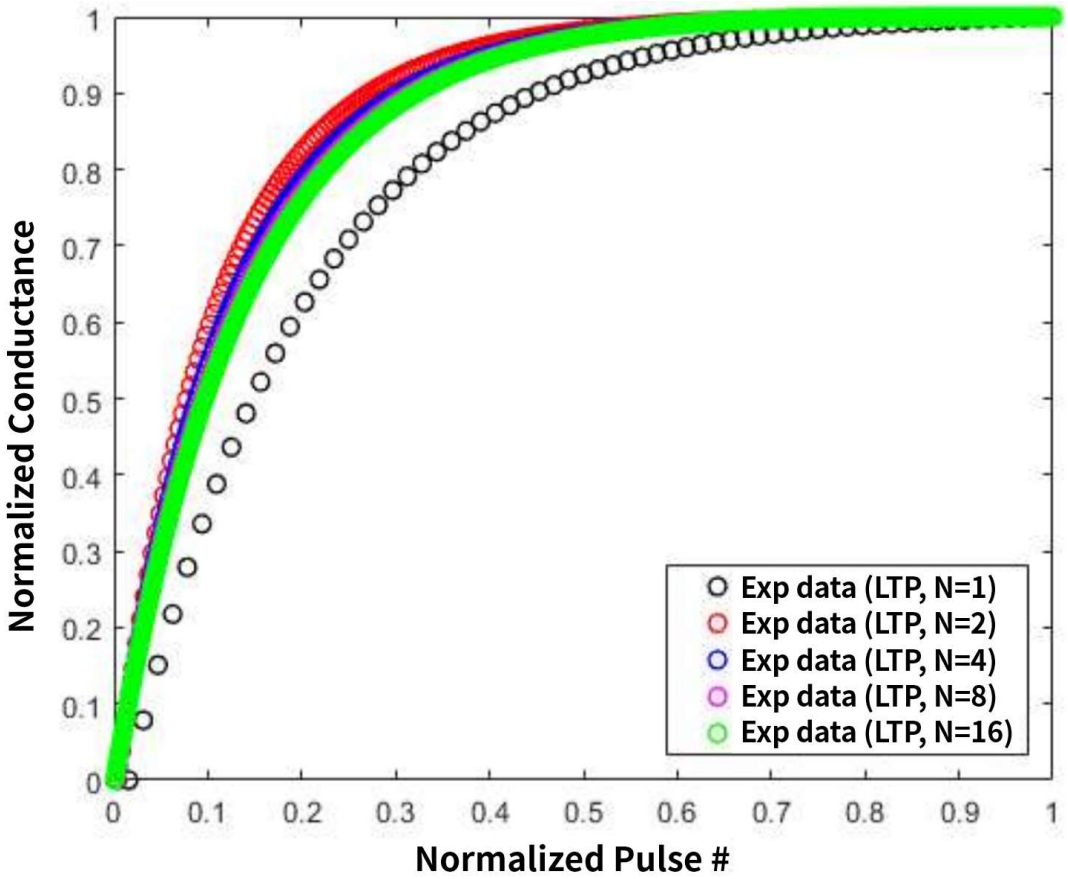

Fig 3.2 NL(LTP)=5.0(비선형적인 소자)의 경우, 가상의 시냅스 소자를 이용한 Normalized pulse 개수 vs Normalized conductance 결과. N의 개수를 1, 2, 4, 8, 16으로 변화시켜가며 측정

3.2 가중치 변화량 산포 분석

3.2.1 아날로그 소자 특성 및 훈련 방법 설계

앞서 3.1절에서 진행한 컨덕턴스 반응 시뮬레이션은 다중 시냅스 구조가 선형성을 떨어뜨린다고 판단하였는데, 실제 온라인 학습 과정에서 다중 시냅스 구조의 선형성을 분석하기 위해 다음과 같이 실험을 설계하였다.

첫 번째로 비선형성이 다른 4개의 가상의 소자, NL(LTP/LTD)의 크기 1.0/-1.0, 3.0/-3.0, 5.0/-5.0인 비대칭적인 소자와 2.4/-4.88인 실제 Ag:a-Si 소자를 디자인하였다. 그리고 시냅스 소자의 개수가 선형성에 끼치는 영향을 분석하기 위해 여타 비이상적인 특성들을 통제 변인 하여 3.1.1.1과 동일한 조건으로 진행하였다. 동시에, 선형성을 판단하기 위한 종속 변수로 시냅스에 입력된 가중치 변화량과 실제로 변화된 가중치 변화량의 비율(R)을 연산과정에서 계산하였다.

$$R = \frac{\triangle W_{output}}{\triangle W_{input}} \tag{3.1}$$

그러고 나서 설계한 아날로그 시냅스 소자를 뉴럴 네트워크에 배치하여 MNIST 손글씨 데이터를 학습하였다. 훈련 과정은 8,000개의 Training images를 1-epoch으로 10-epoch까지 진행하였으며 총 80,000번의 가중치 갱신에서 R 값을 산정하였다. 추가적인 작업으로 시냅스 가중치(weight = -1~1)의 영역을 10등분 한 뒤 각 영역에서 해당하는 시냅스 소자의 R 값을 평균하여 측정하였다.

3.2.2 가중치 변화량 산포 분석 평가

Fig 3.3-3.6은 순서대로 NL(LTP/LTD)의 값이 1.0/-1.0, 3.0/-3.0, 5.0/-5.0, 2.4/-4.8인 경우이다. 그림에서 세로축에 해당하는 R 값은 선형성을 나타내는 지표로 R=1인 경우는 이상적인 소자에 해당한다. 반대로 R의 값이 1에 멀어질수록 비선형적인 소자에 가깝다고 볼 수 있다. 그리고 가로축에 해당하는 것은 시냅스 가중치(weight)로 -1과 1 사이의 값을 가지며 10등분 한 각 영역에 해당하는 시냅스 소자의 R 값을 평균하여 계산한다.

주목할 점은, 모든 소자 상황에 대해 N의 개수가 1에서 4, 9, 16으로 증가함에 따라 R=1인 영역으로 분포가 이동하며 결과적으로 시냅스의 선형성이 개선된다고 판단할 수 있다.

구체적으로, Fig 3.1의 선형적인 경우는 R 값의 감소량이 평균적으로 0.2이었고 Fig 3.5의 비선형적인 경우는 이 감소량이 2에 가까웠다. 결과적으로 소자의 비선형적인 특성이 심할수록 다중 시냅스의 선형성 개선 영향이 더욱 효과적임을 알 수 있었다.

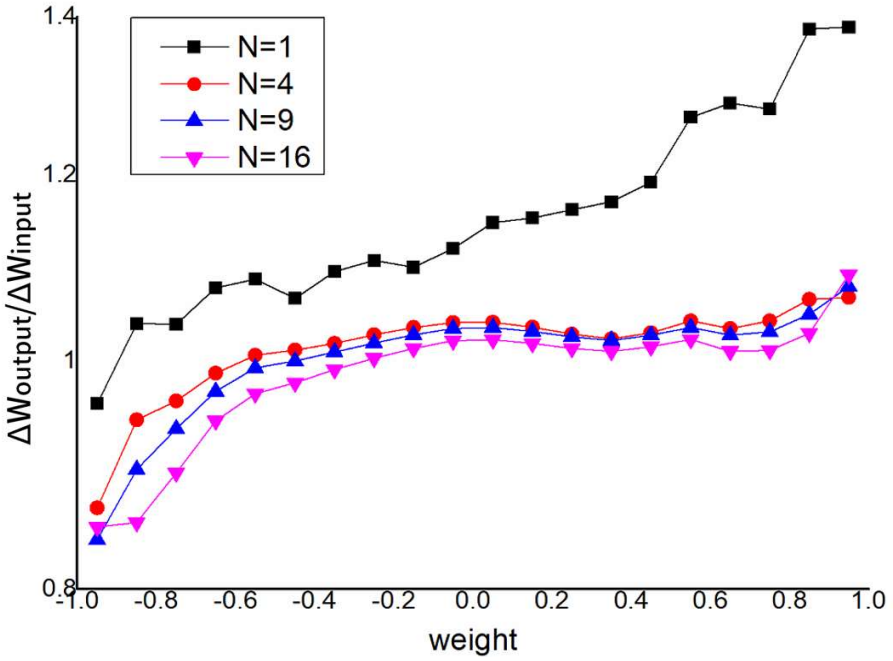

Fig 3.3. NL(LTP/LTD)=1.0/-1.0인 경우, 시냅스 가중치 값에 따른 가중치 변화량 비율 분포

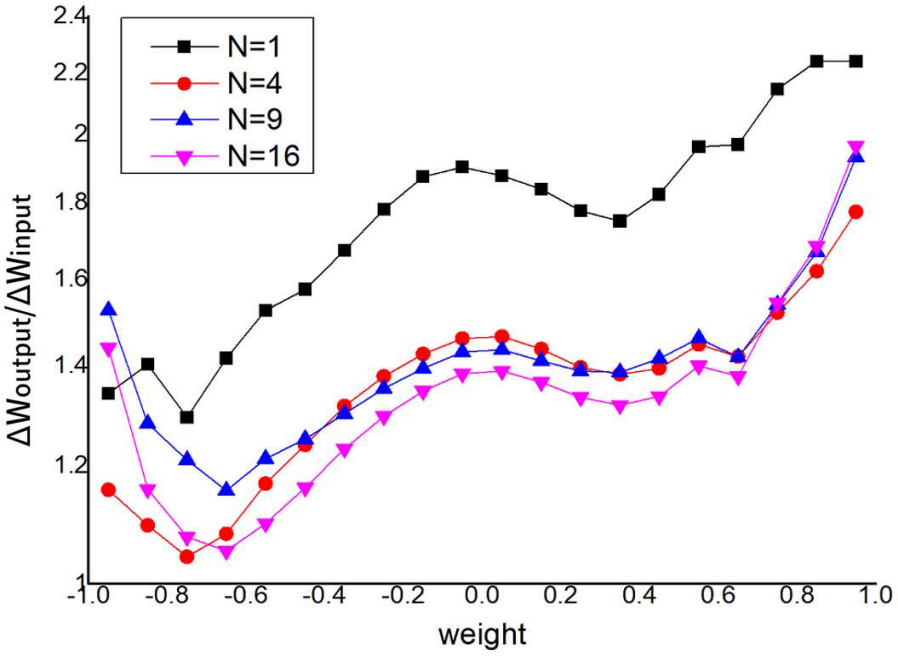

Fig 3.4. NL(LTP/LTD)=3.0/-3.0인 경우, 시냅스 가중치 값에 따른 가중치 변화량 비율 분포

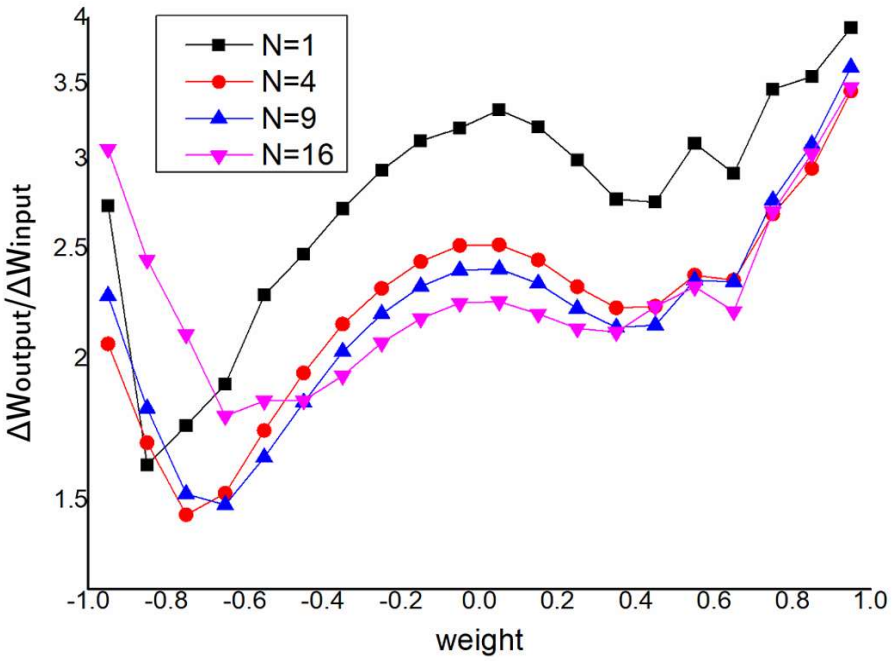

Fig 3.5. NL(LTP/LTD)=5.0/-5.0인 경우, 시냅스 가중치 값에 따른 가중치 변화량 비율 분포

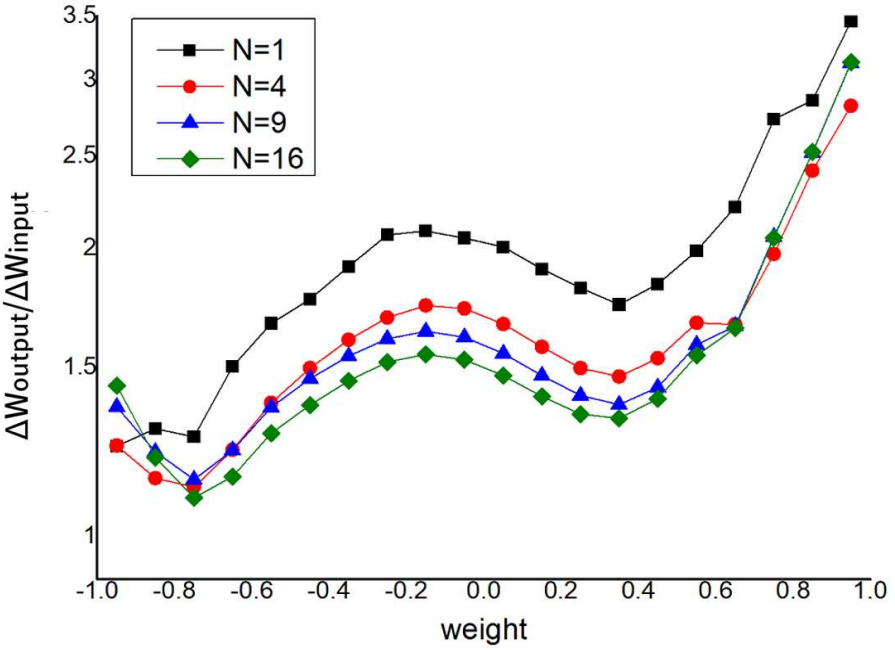

Fig 3.6. NL(LTP/LTD)=2.4/-4.88인 경우, 시냅스 가중치 값에 따른 가중치 변화량 비율 분포

3.3 유효 컨덕턴스 상태 수

다중 시냅스 구조가 가지는 또 다른 이점을 살펴볼 것인데, 이는 다중 시냅스 구조를 활용하면 단일 시냅스 소자가 표현하는 컨덕턴스 상태가 확장된다는 것이다. 이것을 유효 컨덕턴스 상태 수(effective number of conductance states)로 두어 다중 시냅스 소자가 가지는 컨덕턴스 상태를 정량적으로 분석하였다.

먼저, 실험 조건은 대칭적이며 비선형성인 NL의 크기가 각각 1.0, 2.0, 3.0, 4.0, 5.0인 5개의 소자를 준비하고 N의 개수를 1, 2, 4, 8, 16으로 증가시키며 총 25개의 실험 데이터 세트의 학습 성능을 측정하였다. Fig 2.6에서 N=1인 단일 소자가 온라인 학습에 요구되는 weight bits 수가 6임을 보고하였고 이에 따라 컨덕턴스 상태의 개수 L=64일 때 학습 정확도를 기준으로 삼았다. 그 후 25개의 실험 데이터 세트에 해당하는 L의 값을 64에서 1까지 하나씩 감소시키면서 학습에 필요한 최소 컨덕턴스 상태의 개수를 평가하였는데 이 개수는 기준 학습 정확도와 오차 5% 이내인 지점으로 판단하였다.

실험 결과 Fig 3.7.에서 N의 개수가 증가하면 학습에 요구되는 스텝 레벨(Step level), 즉 학습에 요구되는 컨덕턴스 상태의 수가 감소하게 됨을 확인하였으며 이는 다음과 같은 수식을 따른다.

$$L_{eff} = \frac{N \times L}{\alpha} \tag{3.2}$$

(여기서 Leff는 유효 컨덕턴스 상태의 수(effective number of

conductance states), N은 병렬 연결된 소자의 수, L은 단일 소자의 컨덕턴스 상태 수, α는 비선형성의 크기와 관련된 유효 계수(effective coefficeint)이다.)

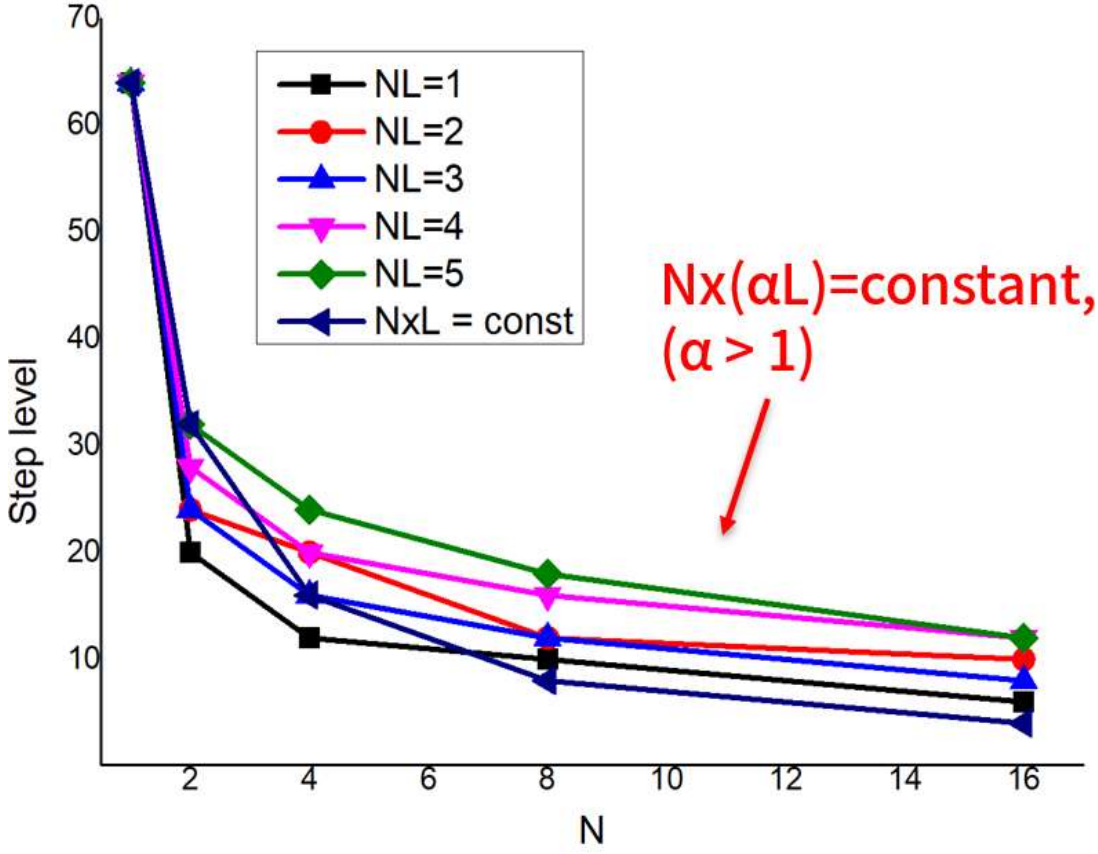

Fig 3.7. N의 개수=1, 2, 4, 8, 16에 따라 학습에 필요한 최소한의 스텝 레벨의 개수. Magnitude of Nonlinearity, NL=1.0/1.0(black), 2.0/2.0(red), 3.0/3.0(blue), 4.0/4.0(margenta), 5.0/5.0(green), ideal(deep blue)

3.4 훈련 최적화 기법에 따른 학습 성능 분석

선행 연구에서는 상변화 메모리 소자라는 특수한 경우에 다중 시냅스 구조를 통해 학습 성능을 개선하였는데, 그렇기에 다양한 소자의 특성들에 대한 실험이 부족하였고 비이상적인 특성들이 끼치는 영향을 정확하게 분석하지 못하였다.

3.4.1. 가상의 다중 시냅스 구조 설계

다중 시냅스 구조를 활용한 실험 3.2, 3.3에서 선형성 증가 및 유효 컨덕턴스 상태의 개수가 증가함을 확인했다. 우리는 이를 효과적으로 검증하기 위해 비선형성 및 비대칭성이 다양한 상황을 선정하였는데 비선형성의 크기 NL(LTP/LTD)= 0.0/0.0, 2.0/-2.0, 3.0/-3.0, 5.0/-5.0 등의 4가지 경우가 대부분의 아날로그 시냅스 소자의 특성을 표현할 수 있다고 판단하였다. 따라서 위의 4가지 경우에 컨덕턴스 상태의 개수 L= 4, 8, 16, 32, 64, 128인 경우를 고려하여 총 24가지 가상의 아날로그 소자를 만들었다. 무엇보다도 다중 시냅스 구조의 영향을 살펴보기 위해 병렬 연결된 시냅스 소자의 개수 N을 1, 2, 4, 9, 16을 변화 주어 총 120가지의 상황에서 학습 정확도를 계산하였다.

그리고 훈련 최적화 알고리즘에 따른 학습 성능을 판단하기 위해 모멘텀, RMSprop, Adam 등의 3가지 방식을 추가하여 위와 일치하는 120가지 실험 세트를 구성하였다. 이를 통해 훈련 최적화 기법이 비이상적인 특성에 끼치는 영향을 알 수 있으며 다중 시냅스 소자와 통합적으로 사용하여 학습 성능 향상을 기대하였다.

3.4.2. 실제 eNVM 소자의 학습 성능 검증

우리는 Table 1에 분석한 AlO_x/HfO_2, HZO-FeFET, PCMO, EpiRAM, TaO_x/HfO_x 그리고 Ag:a-Si 등의 6가지 아날로그 소자에 대해 다중 시냅스 구조를 적용하고 학습 성능을 분석하였다.

각각의 소자는 비선형성 및 비대칭성, 제한된 컨덕턴스 상태의 개수, cycle-to-cycle variation 등이 다양하게 문제 될 수 있으며 예를 들어, PCMO 소자는 비선형성 및 비대칭성과 적은 컨덕턴스 상태의 수로 인해 학습 정확도가 30% 이하로 나타나며, AlO_x/HfO_2 소자는 높은 variation 때문에 20% 이하의 학습 정확도를 가진다. 따라서 본 실험을 통해 다양한 소자 상황에서 다중 시냅스 구조가 가지는 이점인 선형성과 유효 컨덕턴스 상태 수 증가를 평가하는 것이 목표이며 훈련 최적화 기술 중 하나인 Adam 옵티마이저와 비교 분석하여 학습 성능에 끼치는 영향을 정량화하였다.

3.5 선형 및 대칭 스케일링 방법

비이상적인 아날로그 시냅스 소자는 비선형성과 비대칭성을 가지고 있는데 이러한 특징을 최소화시킬 수 있는 선형 및 대칭 스케일링 전략을 고안하였다.

3.5.1 선형 및 대칭 구간 설계

먼저, 아날로그 소자의 선형 및 대칭 구간은 다음과 같은 수식을 통해 설정할 수 있다.

수식 (3.1)에서 선형성을 나타내는 매개 변수인 R에 체인 룰을 적용하면 다음과 같은 수식을 얻는다.

$$\frac{dW_{output}}{dW_{input}} = \frac{dp}{dW_{input}} \times \frac{dG_{ltp}}{dp} \times \frac{dW_{output}}{dp} \tag{3.3}$$

그리고 수식 (2.1)을 양변 미분하면 다음과 같은 수식을 얻는다.

$$\frac{dG_{ltp}}{dp} = \frac{Be^{-P/A}}{A} \tag{3.4}$$

수식 (2.6)에서 가중치 값과 컨덕턴스 사이의 관계를 이용하면 다음과 같은 수식을 얻는다.

$$\frac{dW_{output}}{dG_{ltp}} = \frac{W_{\max} - W_{\min}}{G_{\max} - G_{\min}} \tag{3.5}$$

$$\therefore \frac{\triangle W_{output}}{\triangle W_{input}} = \frac{P_{\max} e^{-P/A}}{A(1-e^{-P_{\max}/A})} \tag{3.6}$$

수식 (3.6)을 통해 시냅스 소자의 선형 지점(Linear point)와 대칭 지점(Symmetric point)를 구할 수 있다.

Linear point

$$\frac{\triangle W_{output}}{\triangle W_{input}} = \frac{P_{\max} e^{-P/A}}{A(1-e^{-P_{\max}/A})} = 1 \tag{3.7}$$

$$P_{linear} = Aln(\frac{P_{\max}}{A(1-e^{-P_{\max}/A})}) \tag{3.8}$$

수식 (3.7)에서 완벽하게 선형적일 때 R=1이므로 선형 지점을 구하면 수식 (3.8)과 같다.

Symmetric point

$$(\frac{\triangle W_{output}}{\triangle W_{input}})_{ltp} = (\frac{\triangle W_{output}}{\triangle W_{input}})_{ltd} \tag{3.9}$$

대칭 지점은 LTP와 LTD에서의 R 값이 같으므로 수식 (3.9)를 통해 계산할 수 있다.

수식 (3.8)-(3.9)를 통해 선형 지점과 대칭 지점을 구할 수 있으며 이를 바탕으로 선형 구간 및 대칭 구간을 설계하였다. (Fig 3.8) 그리고 온 칩 훈련에서 충분한 학습 성능을 얻기 위해 아날로그의 weight bits 수가 6-bit 이상이 되어야 하는 것을 고려하여 선형 및 대칭 구간에 해당하는 weight bits 수가 4에서 5 사이가 되도록 설정하였다. 그리고 수식 (3.2)에서 다중 시냅스 구조가 N의 개수에 비례하여 유효 컨덕턴스 상태 수를 가지는 효과를 통해 스케일링 방법으로 부족해진 컨덕턴스 상태의 개수를 보충할 수 있다.

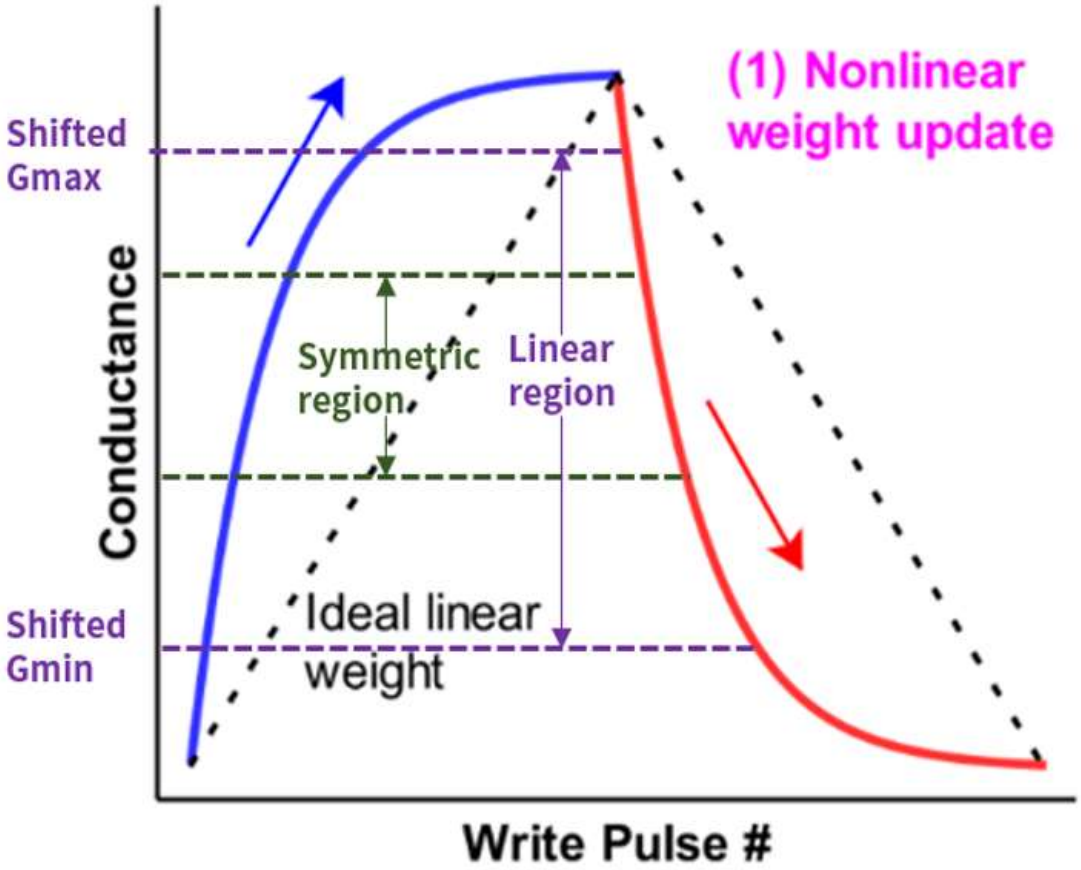

Fig 3.8. 아날로그 시냅스 소자에서 선형 구간(Linear region)과 대칭 구간(Symmetric region)을 설명한 그림.

3.5.1 훈련 방법 및 모델 설계

소프트웨어상에서 진행되는 훈련 알고리즘은 Fig 3.9와 같다. 첫 번째 단계로 인풋 뉴런을 통해 MNIST 데이터가 입력이 되고 배치된 가상의 시냅스 소자에 좁혀진 컨덕턴스 구간(scaled conductance range)을 통해 가중치 합 연산을 진행한다. 그 과정을 살펴보면, 기존의 최대 컨덕턴스(Gmax) 혹은 최소 컨덕턴스(Gmin)에 해당하는 시냅스 가중치 1과 -1을 좁혀진 컨덕턴스 범위 안의 G'max~ G'min 으로 이동시키는 작업을 진행한다. 아울러 좁혀진 컨덕턴스 범위 밖의 상태들에 해당하는 시냅스 가중치는 -1과 1로 고정이 되는데, G'min 보다 작은 영역에서는 -1로 G'max보다 큰 영역은 1로 설정하였다. 결과적으로 Gmin~Gmax 에서 증가 혹은 감소하던 컨덕턴스 반응이 G'min~G'max 영역으로 옮겨지고 역전파에서 일어나는 가중치 갱신이 이 구간을 통해 일어나기 때문에 선형 및 대칭적인 궤도를 따를 수 있다.

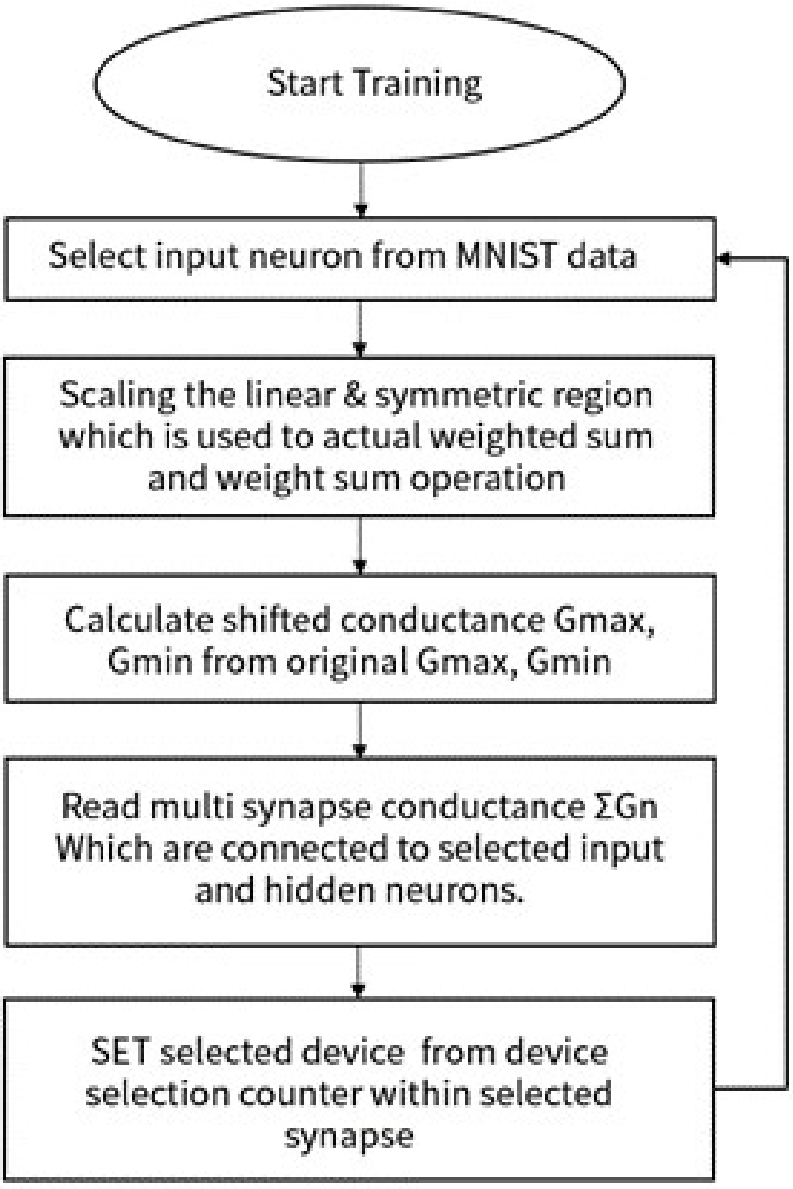

Figure 3.9. Software execution of Linear & symmetry scaling with multi-synapses architecture.

Chapter 4. 결과 및 고찰

4.1 다중 시냅스 구조 검증 성능 결과

3.4.1절에서 소개했던 다중 시냅스 구조가 소자의 비이상적인 특성에 끼치는 영향을 분석한 실험의 결과이다. Fig 4.1-4.4는 컨덕턴스 상태의 개수와 병렬 연결된 소자의 개수 N에 따른 학습 정확도를 나타낸다.

먼저, Fig 4.1은 이상적인 소자의 경우로 가로축은 컨덕턴스 상태의 개수 세로축은 학습 정확도에 해당한다. 시냅스 소자의 개수 N=1(black)일 때 낮은 컨덕턴스 상태의 개수(L<16)에서 30% 이하의 학습 정확도를 보이는데 N의 개수가 2, 4, 9, 16으로 증가 할수록 이러한 조건에서도 높은 학습 정확도를 확보할 수 있다. 이는 실험 3.3에서 유효 컨덕턴스 상태의 개수 증가 현상으로 설명할 수 있다.

다음으로, 비선형성 및 비대칭성이 문제가 되는 상황인 Fig 4.2~4.4은 각각 NL(LTP/LTD) = 2.0/-2.0, 3.0/-3.0, 5.0/-5.0일 때의 분석 결과이다. N=1일 때 결과를 살펴보면 비선형성 및 비대칭성이 커질수록 요구되는 weight bits의 수가 6-bit 이상으로 많아짐을 확인할 수 있다. 그리고 선형적일 경우와 마찬가지로 유효 컨덕턴스 상태의 개수가 증가함을 확인할 수 있었다. 구체적으로 살펴보면 N=1, L<16으로 낮은 컨덕턴스 상태의 개수를 가질 때 세 가지 상황에서 모두 20% 이하의 학습 정확도를 보이는데 N의 개수를 증가시킬수록 학습 정확도가 대폭 상승하는 것을 확인할 수 있다. 이는 실험 3.3에서 분석한 유효 컨덕턴스 증가와 일치하는 결과이다.

한편, 3.2절에서 보여준 선형성 증가 현상은 Fig 4.4에서 두드러지게 나타나는데 컨덕턴스 증가 효과를 통제하기 위해 충분한 컨덕턴스 상태 수(L=128)에서 N의 개수를 늘렸을 때 살펴본 결과 학습 정확도가 30%에서 84%까지 증가하였다. 이를 통해 다중 시냅스 구조는 선형성을 개선한다고 판단할 수 있다.

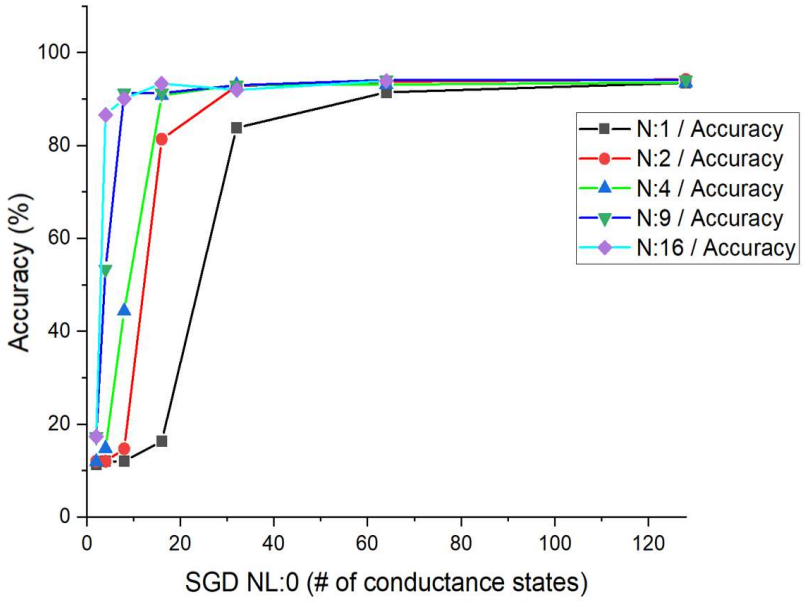

Fig 4.1. NL(LTP/LTD)=0.0/0.0, N의 개수=1(black), 2(red), 4(blue), 9(green), 16(purple)에 따른 conductance states의 개수=4, 8, 16, 32, 64, 128인 소자의 학습 정확도

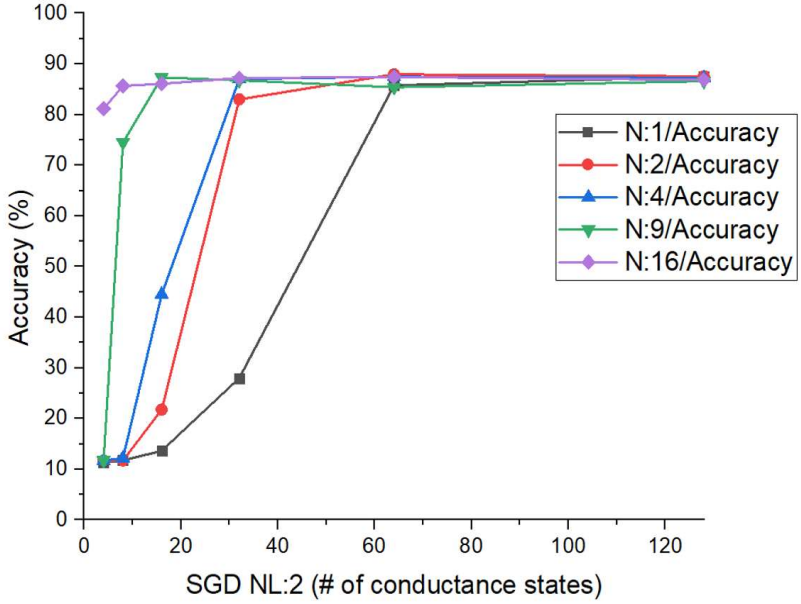

Fig 4.2. NL(LTP/LTD)=2.0/-2.0, N의 개수=1(black), 2(red), 4(blue), 9(green), 16(purple)에 따른 conductance states의 개수=4, 8, 16, 32, 64, 128인 소자의 학습 정확도

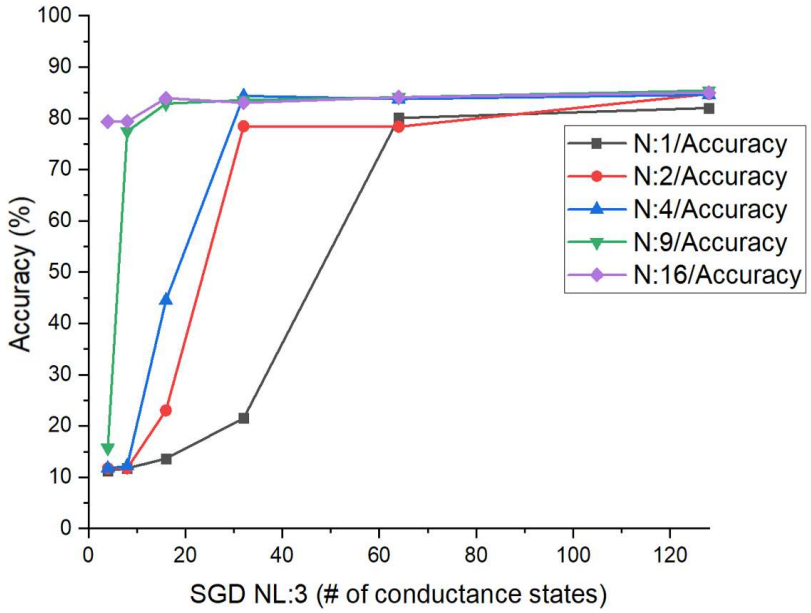

Fig 4.3. NL(LTP/LTD)=3.0/-3.0, N의 개수=1(black), 2(red), 4(blue), 9(green), 16(purple)에 따른 conductance states의 개수=4, 8, 16, 32, 64, 128인 소자의 학습 정확도

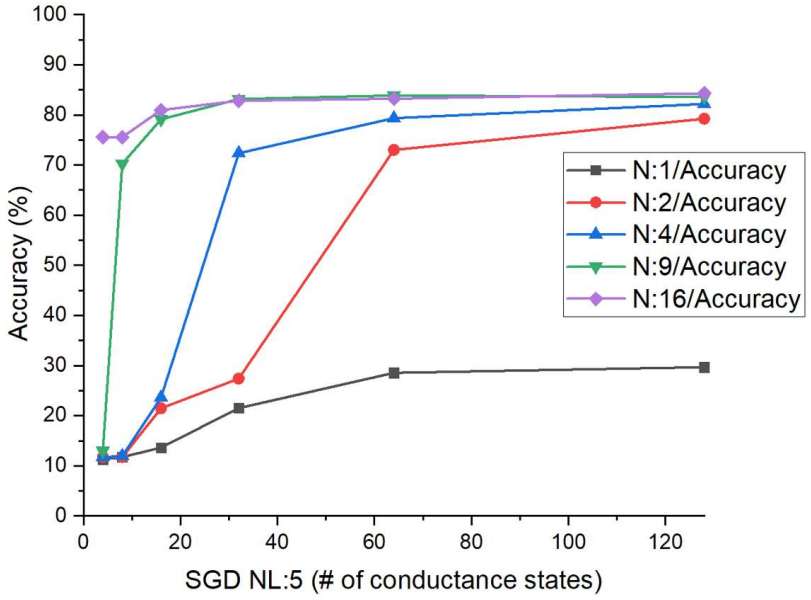

Fig 4.4 NL(LTP/LTD)=5.0/-5.0, N의 개수=1(black), 2(red), 4(blue), 9(green), 16(purple)에 따른 conductance states의 개수=4, 8, 16, 32, 64, 128인 소자의 학습 정확도

다음으로 Fig 4.5~4.7은 훈련 최적화 알고리즘 기법을 모멘텀, RMSprop, Adam으로 조정해가며 측정한 결과이다. 이러한 방식들이 학습 성능에 끼치는 영향을 효과적으로 알아보기 위하여 비선형성 및 비대칭성이 심한 소자 NL(LTP/LTD)= 5.0/-5.0인 소자를 설계하였다.

첫 번째로 Fig 4.5는 모멘텀 방식을 통해 얻은 결과이다. 기존의 SGD 방식에서는 N=1인 경우에 L=128일 때 30% 이하의 학습 정확도를 보였는데 모멘텀 방식을 통해 78%까지 개선할 수 있었다. 그리고 다중 시냅스 구조를 사용하여 N=16일 때 85%의 정확도를 보였다. 같은 조건에서 SGD 방식(84%)과 비교하였을 때 미세한 차이를 보였다.

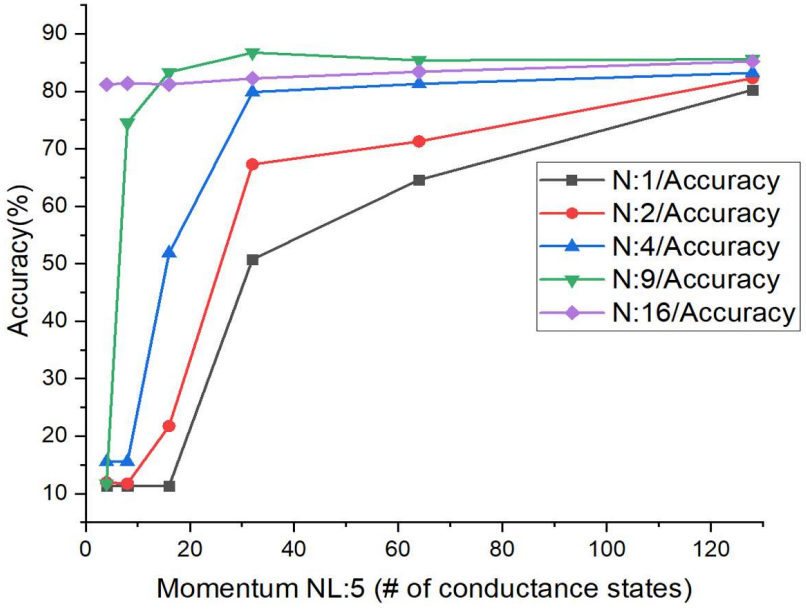

Fig 4.5. Momentum, NL(LTP/LTD)=5.0/-5.0, N의 개수=1(black), 2(red), 4(blue), 9(green), 16(purple)에 따른 conductance states의 개수=4, 8, 16, 32, 64, 128인 소자의 학습 정확도

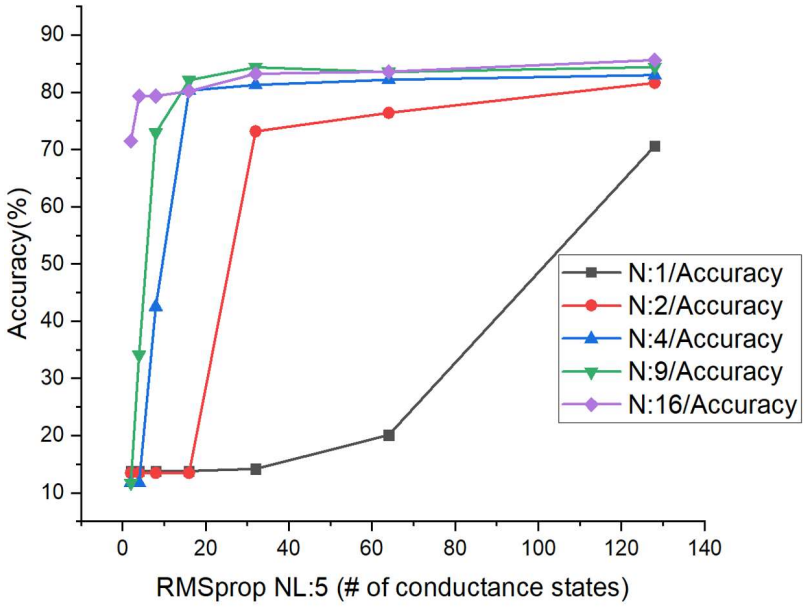

Fig 4.6. RMSprop, NL(LTP/LTD)=5.0/-5.0, N의 개수 =1(black), 2(red), 4(blue), 9(green), 16(purple)에 따른 conductance states의 개수=4, 8, 16, 32, 64, 128인 소자의 학습 정확도

두 번째로 Fig 4.6은 RMSprop 옵티마이저를 사용하여 얻은 학습 정확도이다. 특징적으로 SGD 및 모멘텀 방식과 비교하였을 때 컨덕턴스 상태의 수가 증가함에 따라 학습 정확도가 빠르게 회복되었다. 이는 더 낮은 컨덕턴스 상태의 개수에서도 학습이 원활하게 진행됨을 의미한다. 또한, N의 개수를 늘렸을 때 최대 87%의 학습 정확도를 기록하였다.

마지막으로 Fig 4.7은 Adam 방식을 통해 분석한 결과이며 훈련 옵티마이저 중 가장 높은 성능을 보여주었다. 특히, 다중 시냅스 구조를 사용하였을 때 낮은 컨덕턴스 상태 수에서도 높은 정확도를 기록 할 수 있었으며 N=16일 때 2-weight bits 수로 80% 이상의 정확도를 나타내었다. 그리고 N=1일 때 충분한 컨덕턴스 상태 수가 있으면 84%의 정확도를 보여주었으며 이는 SGD 방식에서 N=16일 때의 성능과 견줄 수 있다. 더욱이 다중 시냅스 구조와 통합적으로 사용하였을 때 89%의 가장 높은 정확도를 보여주었다.

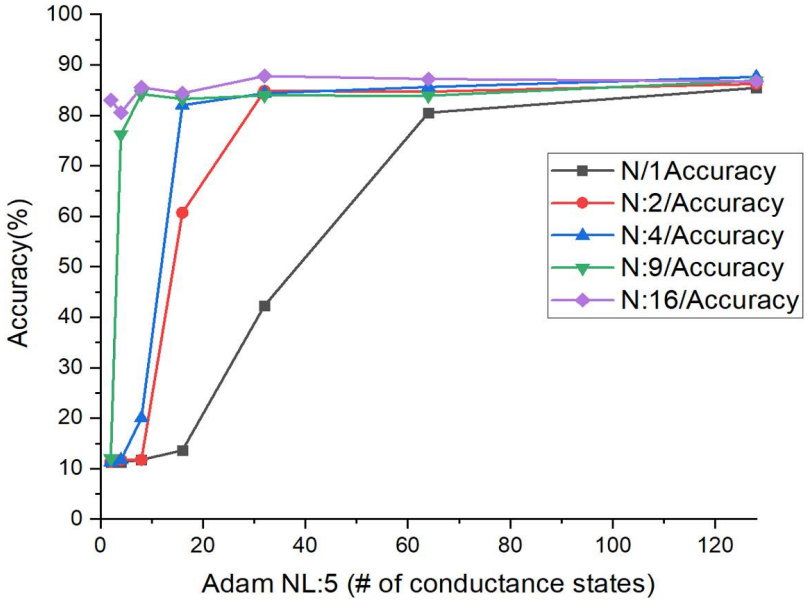

Fig 4.7. Adam, NL(LTP/LTD)=5.0/-5.0, N의 개수=1(black), 2(red), 4(blue), 9(green), 16(purple)에 따른 conductance states의 개수=4, 8, 16, 32, 64, 128인 소자의 학습 정확도

다음으로 실제 eNVM 소자 특성을 이용하여 가장 효과적인 훈련 최적화 기법이었던 Adam 옵티마이저와 다중 시냅스 구조를 비교 분석하고 통합적으로 사용했을 때의 학습 정확도를 측정하였다.

Figure 4.8은 AlO_x/HfO_2 소자를 다중 시냅스 방법과 Adam을 이용한 정확도 결과이다. 먼저, SGD 방식에서 N의 개수를 1, 2, 5로 늘렸을 때 (그림 상의 2G와 5G에 해당) 학습 정확도가 약 60%, 74%, 86%로 증가하였다. 그리고 Adam을 사용한 경우 N=5인 경우와 유사한 결과를 보였으며, Adam과 N=2인 다중 시냅스 구조를 사용하였을 때 가장 높은 90%의 정확도를 보였다.

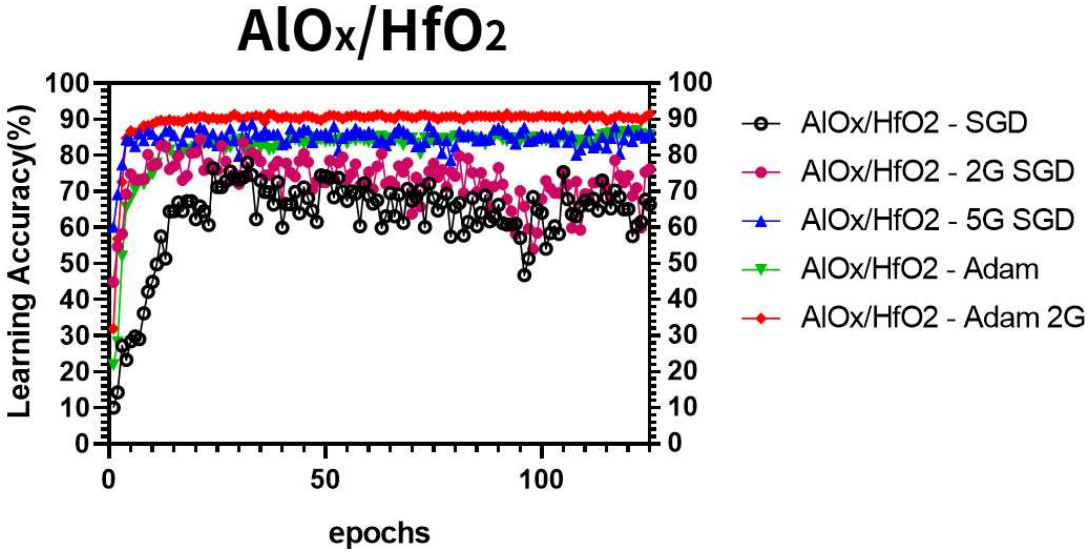

Fig 4.8. AlO_x/HfO_2 소자, 비선형성 NL(LTP/LTD)=1.94/-0.61, 컨덕턴스 상태의 수 L=40, C2C var=5%, SGD(검은색), 2G와 SGD(자홍색), 5G와 SGD(파란색), Adam(초록색), 2G와 Adam(빨간색)

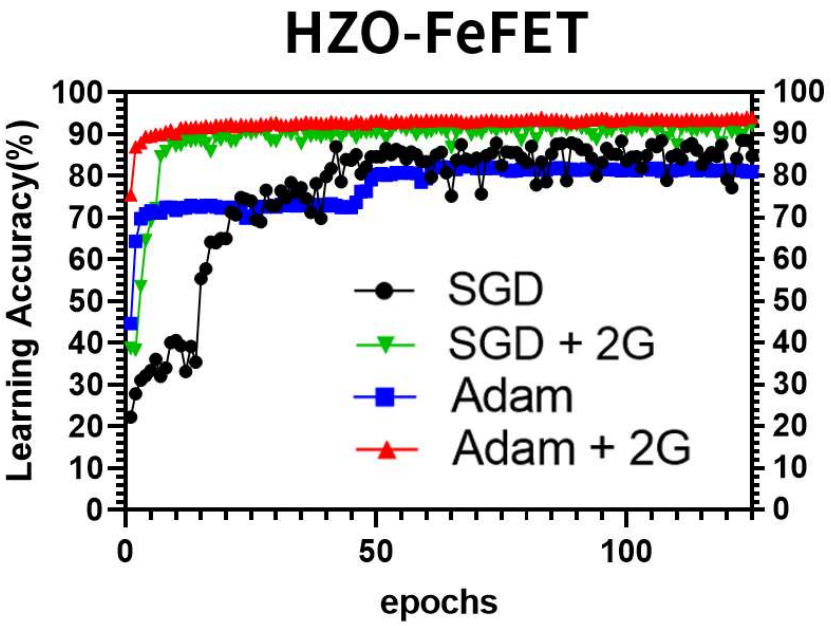

Fig 4.9. HZO-FeFET 소자, 비선형성 NL(LTP/LTD)= 1.75/1.46, 컨덕턴스 상태의 수 L=32, C2C var<0.5%, SGD(검은색), 2G와 SGD(초록색), Adam(파란색), 2G와 Adam(빨간색)

그리고 Figure 4.9를 살펴보면 낮은 컨덕턴스 수가 문제가 되는 HZO-FeFET 소자에서는 Adam이 오히려 SGD 보다 낮은 정확도를 보였으며 다중 시냅스 구조의 N=2인 경우에 단일 소자에 비해 약 3%의 정확도가 향상되었다. 그리고 Adam과 다중 시냅스 구조를 함께 사용하였을 때 92%의 정확도가 측정되었다.

다음으로 Figure 4.10은 비선형성 및 컨덕턴스 상태의 개수가 문제인 PCMO 소자의 결과이다. Table 1에서 PCMO는 30%의 낮은 정확도를 보였는데 Adam을 사용하여도 이를 개선하기 어려웠다. 그런데 Adam을 단독으로 사용하며 소자의 컨덕턴스 상태의 개수를 200개로 인위적으로 바꾸었을 때 90%의 높은 정확도를 나타내었다. 이것은 Adam이 컨덕턴스 상태의 개수에 끼치는 영향이 미미한 사실을 반증한다. 반면, 다중 시냅스 소자의 개수 N이 5인 경우 정확도가 약 80%까지 상승하였는데 Adam과 비교하였을 때 비교적 낮은 수치이다. 즉, PCMO가 가지는 극심한 비선형성 및 비대칭성으로 인해 나타난 다중 시냅스 구조의 한계라고 여겨진다.

.

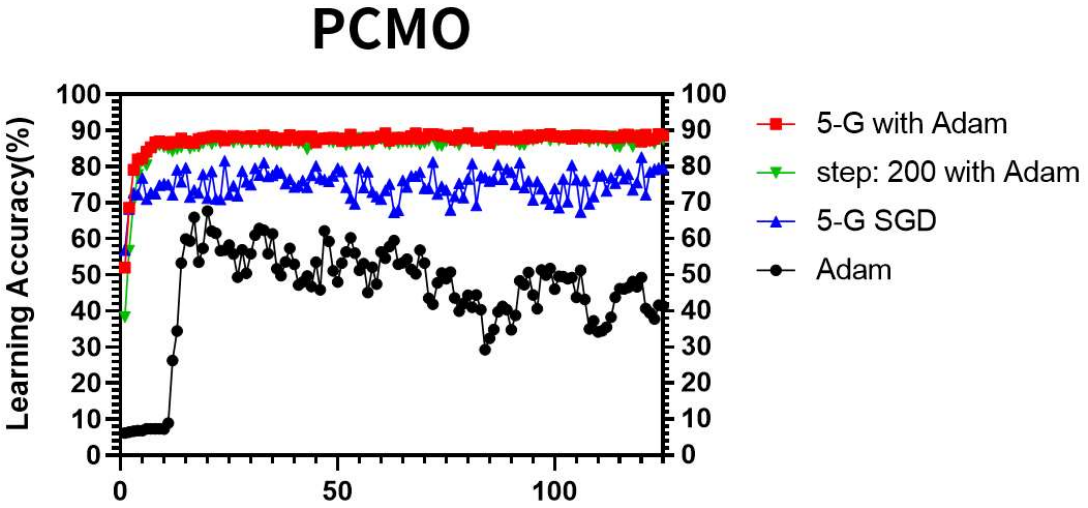

Fig 4.10. PCMO 소자, 비선형성 NL(LTP/LTD)=3.68/-6.76, 컨덕턴스 상태의 수 L=50, C2C var<1% Adam(검은색), 5G와 SGD(파란색), 5G와 Adam(빨간색), L=200인 소자에 Adam을 사용한 결과(초록색)

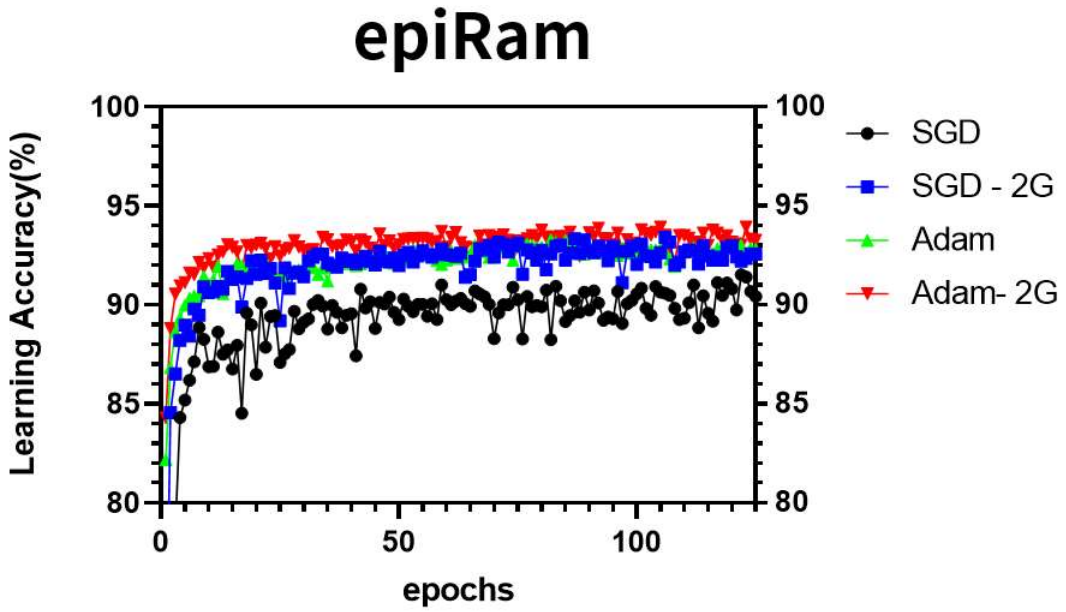

Fig 4.11. epiRam 소자, 비선형성 NL(LTP/LTD)=0.5/-0.5, 컨덕턴스 상태의 수 L=64, C2C var=2% SGD(검은색), 2G와 SGD(파란색), Adam(초록색), 2G와 Adam(빨간색)

Figure 4.11은 이상적인 특성에 가장 가까운 EpiRAM 소자의 결과이며, 다른 방법들의 도움 없이 92% 학습 정확도를 가진다. 그리고 Adam 및 다중 시냅스 구조를 각각 사용하였을 때 학습 정확도가 93.5%로 소폭 상승하였으며, 통합적으로 사용하게 되면 94%까지 증가함을 관찰하였다. Figure 4.12 결과도 마찬가지로 선형적이고 컨덕턴스 상태의 수가 충분한 경우인데, 소자의 variation이 학습을 저해시키는 주요한 원인으로 작용하고 있다. 따라서 variation을 무시한 결과(빨간색)가 가장 높은 정확도를 보여주었으며 N=2인 다중 시냅스 구조 보다 Adam의 경우가 2% 정도 높은 정확도를 보였다.

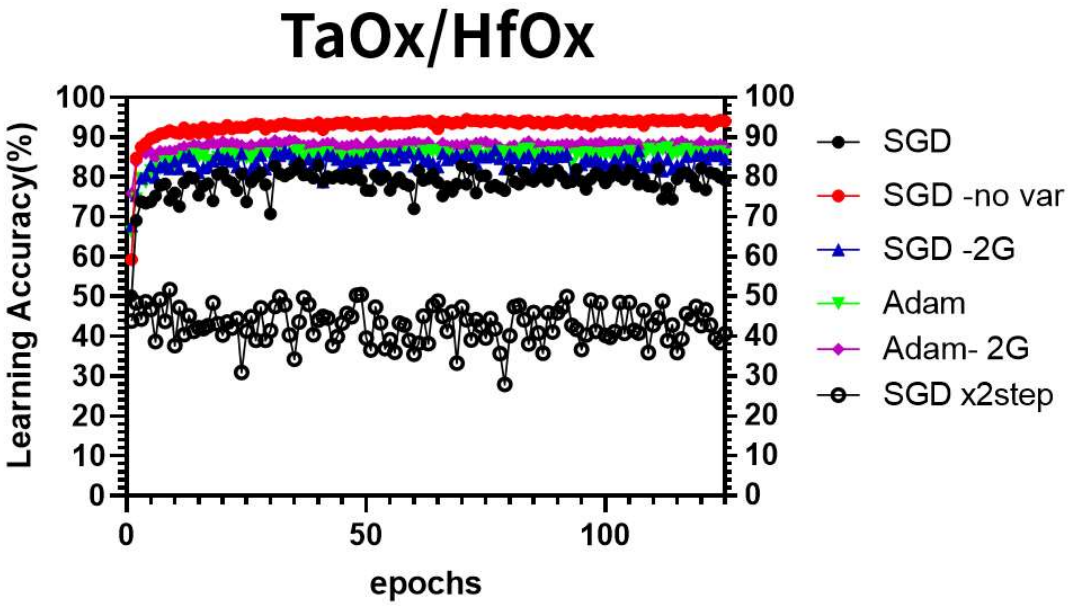

Fig 4.12. TaOx/HfOx 소자, 비선형성 NL(LTP/LTD)=0.04/-0.63, 컨덕턴스 태수 L=128, C2C var=3.7%, SGD(검은색), 2G와 SGD(파란색), Adam(초록색), 2G와 Adam(자홍색), variation이 없는 경우의 SGD(빨간색)

마지막으로 Figure 4.13은 비선형성 및 비대칭성과 소자의 variation 이 큰 문제점인 Ag:a-Si 소자의 결과이다. 컨덕턴스 상태의 개수는 충분하다고 여겨지며 N=2인 다중 시냅스 구조의 결과(보라색)를 살펴보면 86%의 정확도를 보인다. 그리고 N=5인 경우에는 그 값보다 소폭 상승한 88%의 정확도를 가진다. 반면, Adam을 사용했을 때 90%의 높은 정확도를 가졌으며 Adam과 다중 시냅스 구조를 융합한 방식 역시 90%에 가깝게 측정되었다.

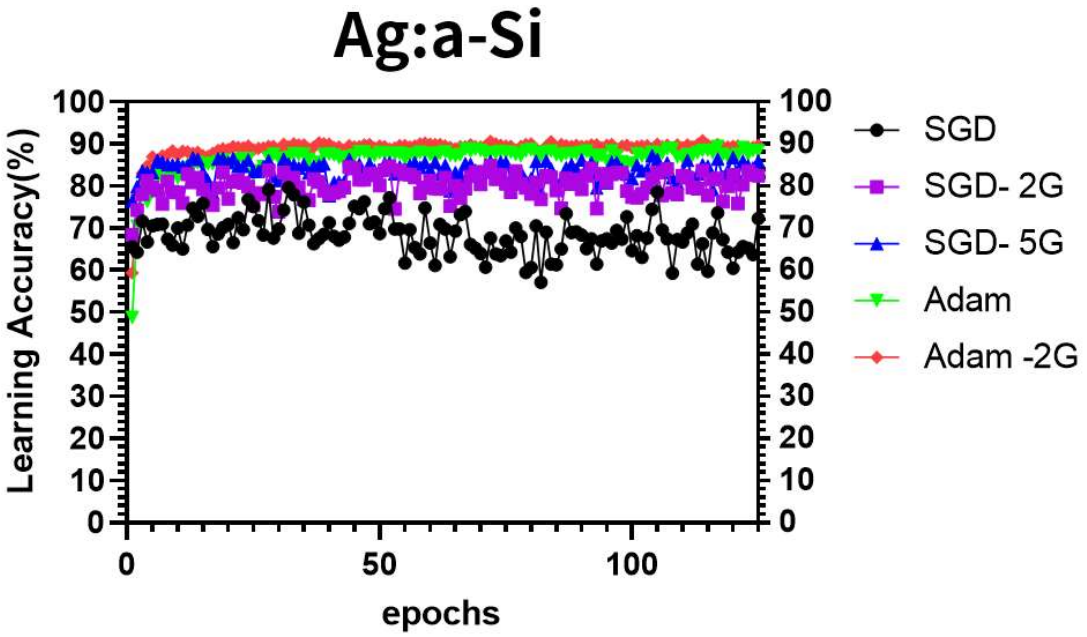

Fig 4.13.Ag:a-Si 소자, 비선형성 NL(LTP/LTD)=2.44/-4.88, 컨덕턴스 상태의 수 L=97, C2C var=3.5%, SGD(검은색), 2G와 SGD(보라색), 5G와 SGD(파란색), Adam(초록색), 2G와 Adam(빨간색)

위의 결과를 요약한 Table 4.1은 6개의 시냅스 소자에 대해 선형성, 컨덕턴스 상태의 개수 그리고 cycle-to-cycle variation 등을 분류하고 이를 완화 시키는 효과적인 방법을 채택하여 Top contributor로 선정하였다. 이 표에서 주목할 점은 다중 시냅스 구조와 Adam 옵티마이저가 아날로그 소자의 특성에 영향을 끼치는 주요 인자가 다르다는 점인데, 다중 시냅스 구조(N-G)는 전형적으로 컨덕턴스 상태의 개수가 낮은 소자에 학습 성능을 효과적으로 개선하는 모습을 보였다. 반면, Adam 옵티마이저는 높은 variation 때문에 학습 성능이 저해되던 소자를 효율적으로 개선할 수 있다. 그리고 모든 경우에 Adam 옵티마이저와 다중 시냅스 구조를 통합적으로 사용하였을 때 가장 높은 학습 성능을 보여주었다.

Modification	Linearity	# of conductance states	C2C variation	Top contributor	Best method
AlOx/HfO2	moderate	bad	bad	N-G	N-G with Adam
HZO-FeFET	good	bad	good	N-G	N-G with Adam
PCMO	bad	bad	good	N-G	N-G with Adam
EpiRAM	good	**moderate**	**moderate**	Adam	N-G with Adam
TaOx/HfOx	good	good	bad	Adam	N-G with Adam
Ag:a-Si	bad	good	bad	Adam	N-G with Adam

Table 4.1. 실제 소자의 Non-ideal properties 평가 및 Adam과 다중 시냅스 소자의 학습 성능을 Top contributor와 Best method로 비교

4.2 선형 및 대칭 스케일링 방법의 학습 결과

다중 시냅스 구조는 회로의 복잡성 및 크기가 커지는 문제점을 가지고 있으며 훈련 최적화 기법 또한, 뉴로모픽 회로로 구현하는 기술의 한계를 가지고 있다. 따라서 우리는 이 두 가지 문제점을 해소하기 위해 선형 및 대칭 스케일링 방법을 제안하였다. 먼저, 아날로그 소자의 선형 및 대칭 지점을 수식 (3.8),(3.9)로 예측할 수 있으며 이를 바탕으로 선형 및 대칭 구간을 설계하는데 이 구간에 속하는 컨덕턴스 단계의 수는 4~5 weight bits를 가지도록 하였다. 이렇게 좁혀진 컨덕턴스 범위에서만 실제 훈련 과정의 가중치 갱신이 일어날 수 있도록 소프트웨어상에 설계하여 실험을 진행하였다.

Fig 4.13은 비교적 선형적인 경우로 NL(LTP/LTD)=1.0/-1.0, 컨덕턴스 상태의 개수가 128개인 소자를 고려하였다. 그리고 스케일링 방법을 사용하지 않은 N=16인 경우(초록색)와 스케일링 방법을 사용하였을 때 N=1, 4, 9, 16을 비교하였다. 이때 선형 및 대칭 스케일링은 학습에 미미한 영향(약 2% 정확도 향상)을 주었는데 그 이유는 이미 선형성과 대칭성이 확보된 소자이며 좁혀진 컨덕턴스 범위로 인해 낮아진 weight bits 수가 학습 성능을 저해시키기 때문이다.

반면 Fig 4.14~4.15는 비선형성 및 비대칭성이 문제가 되는 소자의 경우로 스케일링을 사용하지 않은 다중 시냅스 구조와 비교하여 10% 정도의 학습 정확도가 향상되었다. 특히, Figure 4.14 결과를 살펴보면 선형적인 소자와 유사한 94%의 정확도가 나타났으며 Fig 4.15 결과에서는 Adam 옵티마이저와 다중 시냅스를 사용하였을 때 얻은 최대 정확도인 89%보다 더 높은 92%의 학습 정확도를 보였을 뿐만 아니라 더 적은 시냅스 소자의 개수로도 정확도를 개선하여 회로의 복잡성을 완화 시킬 수 있었다.

.

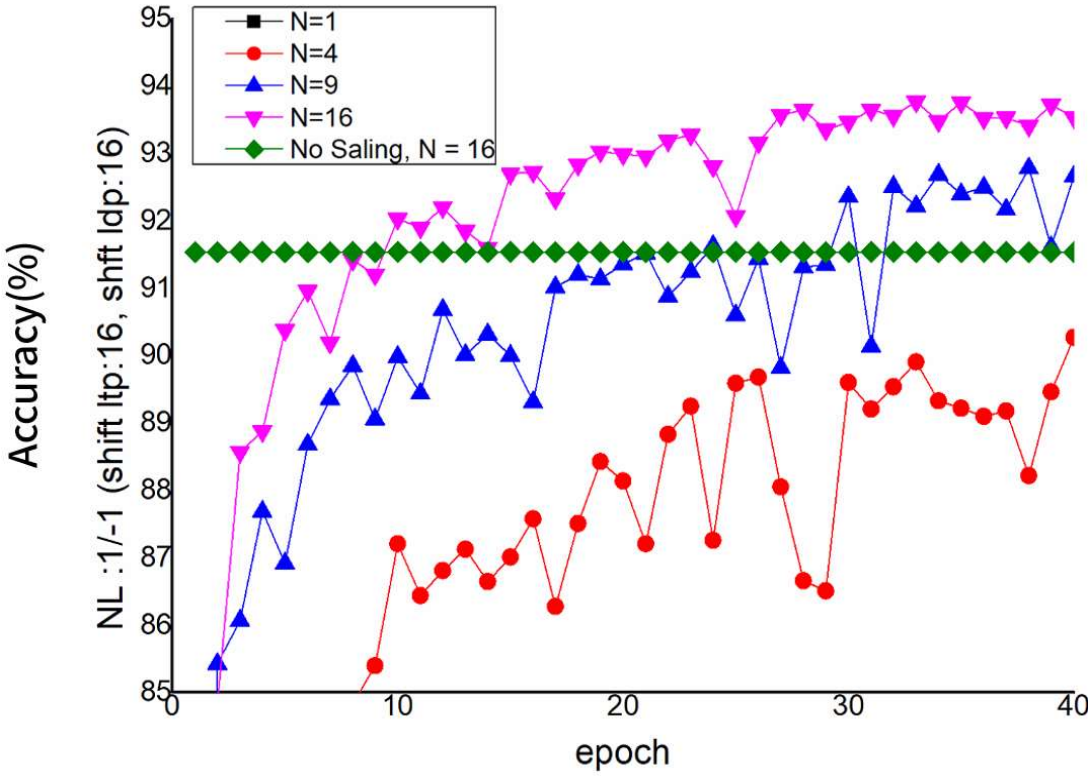

Fig 4.14. 비선형성 NL(LTP/LTD)=1.0/−1.0 컨덕턴스 상태의 수 L=128, scaled conductance range=16

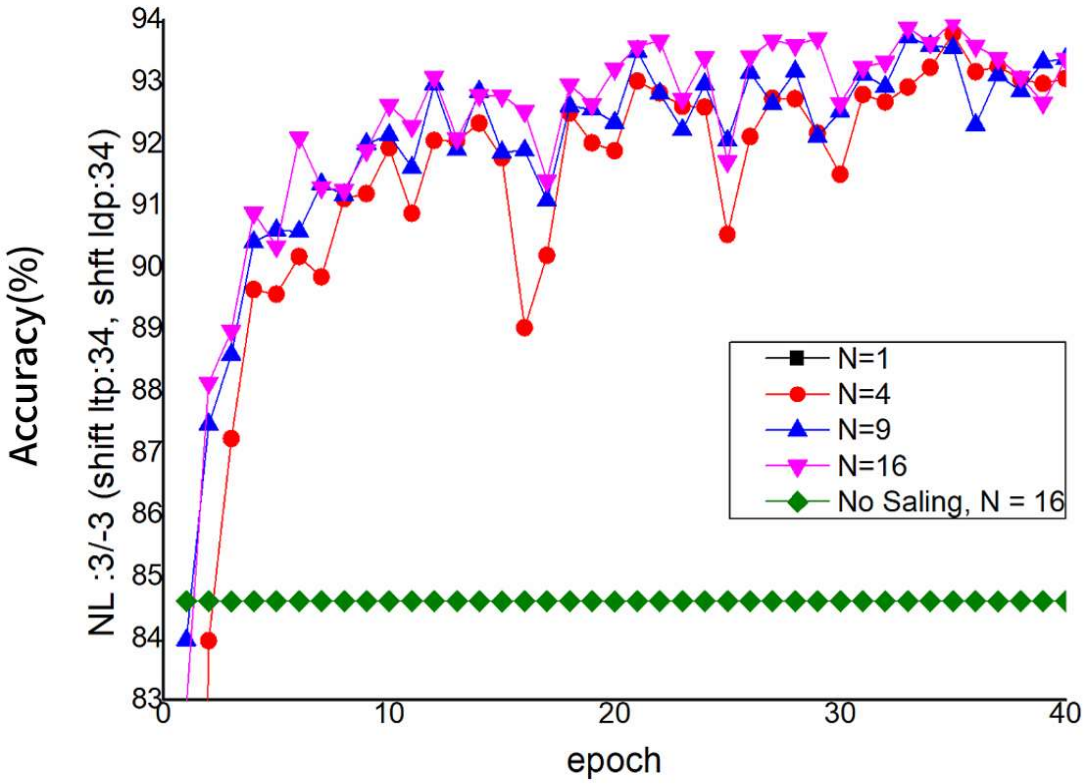

Fig 4.15. 비선형성 NL(LTP/LTD)=3.0/-3.0 컨덕턴스 상태의 수 L=128, scaled conductance range=34

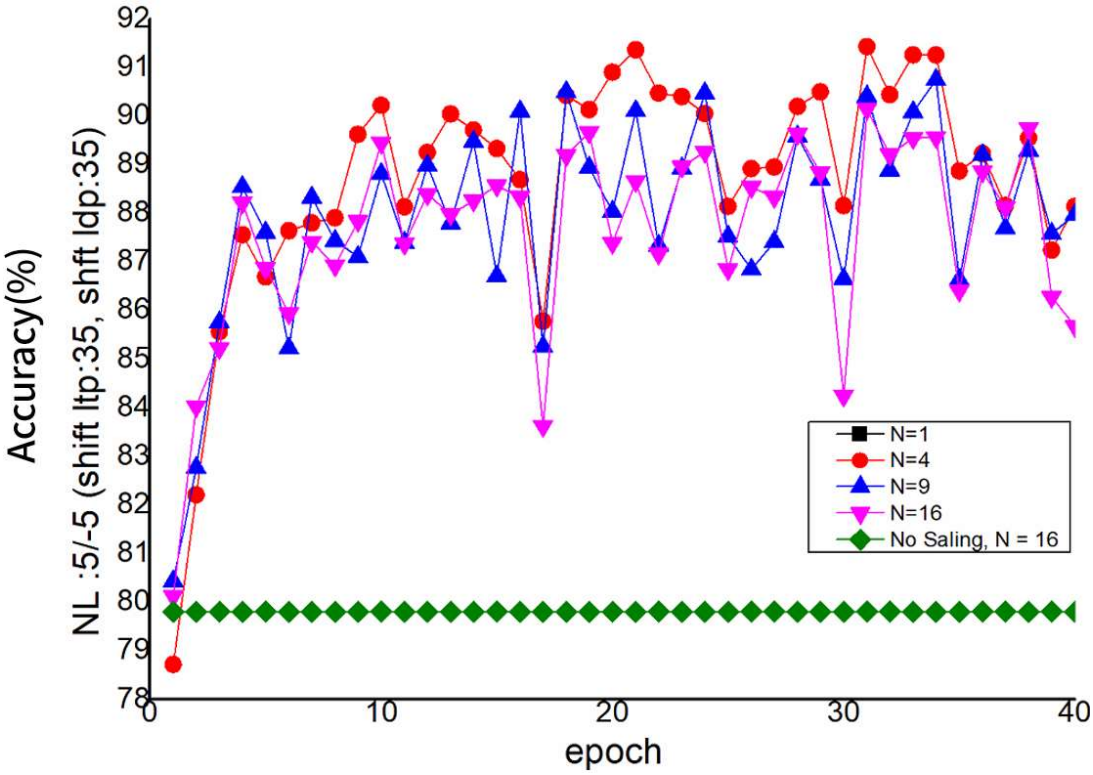

Fig 4.16. 비선형성 NL(LTP/LTD)=5.0/-5.0 컨덕턴스 상태의 수 L=128, scaled conductance range=35

Chapter 5. 요약 및 결론

선행 연구인 다중 멤리스티브 시냅스 구조를 제안한 연구에서는, 다중 시냅스 구조를 통해 비대칭성이 심한 상변화 메모리 소자를 이용하여 높은 학습 성능을 보여주었다. 그러나 상변화 메모리라는 특수한 소자를 사용하였기 때문에 비선형성 및 비대칭성의 개선되는 정도를 정량적으로 분석하지 못하였고 컨덕턴스 단계의 개수가 증가한다는 점 또한 지적하지 못하였다.

본 연구에서는 다중 시냅스 구조를 활용하였을 때 훈련 과정에서 비선형성 및 비대칭성이 크기를 가중치 산포 분석을 통해 통계적으로 관찰하였다. 그 결과 다중 시냅스 구조가 소자의 비선형성 및 비대칭성을 효과적으로 개선할 수 있음을 포착하였으며 뿐만 아니라 유효 컨덕턴스 단계의 개수를 통계적으로 분석한 뒤 시냅스에 병렬 연결된 소자의 개수 N에 비례하여 증가한다고 판단하였다.

또한, 다중 시냅스 구조와 Adam 옵티마이저를 비교 분석하여 각각이 가지는 이점을 확인할 수 있었는데 다중 시냅스 구조는 컨덕턴스 단계의 개수가 낮은 소자에 대해 학습 성능을 효율적으로 개선하였다. 반면에 Adam 옵티마이저는 cycle to cycle variation이 큰 소자에 대해 성능을 효과적으로 증진 시켰다. 결과적으로 다중 시냅스 구조에 훈련 최적화 알고리즘을 통합적으로 사용하여 학습하였을 때 가장 효과적이었으며 비선형성 및 비대칭성이 심한 소자도 기존의 30%의 낮은 검증 정확도에서 89%까지 높일 수 있었다.

그러나 다중 시냅스 구조에서 병렬 연결된 소자의 개수를 늘릴수록 회로의 복잡성이 커지게 되는 문제점이 있으며 훈련 최적화 기법은 뉴로모픽 회로로 구현할 기술이 제안되지 않았다.

그리하여 이 두 가지 문제점을 해결할 수 있는 선형 및 대칭 스케일링 전략을 소개하였고 소자의 컨덕턴스 범위를 선형 및 대칭 구간으로 좁혀주고 다중 시냅스 구조의 이점인 컨덕턴스 상태 수 증가를 통해 비선형성 및 비대칭성이 극심한 소자도 92% 이상의 높은 학습 정확도를 확보

할 수 있었다.

결론적으로 본 연구에서는 다중 시냅스 구조의 비이상적인 소자에 끼치는 영향을 정량적으로 살펴보고 그 중 컨덕턴스 단계의 개수를 증가시키는 것을 포착하여 선형 및 대칭 구간을 설계한 뒤 이 구간에서 학습이 일어날 수 있도록 소프트웨어상의 보조적인 기술을 개발하였다. 하지만 이 연구는 뉴로모픽 컴퓨팅의 핵심 기술인 병렬 연산에 대한 점검이 이루어져 있지 않아 향후 연구가 필요할 것으로 예상한다.

참 고 문 헌

[1] J. Hasler and B. Marr, “Finding a roadmap to achieve large neuromorphic hardware systems,” Front. Neurosci., vol. 7, no. 118, pp. 1 - 29, 2013.
[2] J. Von Neumann, “The principles of large-scale computing machines,” Ann. History Comput., vol. 3, no. 3, pp. 263 - 270, 1981.
[3] J. Schemmel, et al., “A wafer-scale neuromorphic hardware system for large-scale neural modeling,” IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1947-1950, 2010.
[4] R. Serrano-Gotarredona, et al., “CAVIAR: A 45k neuron, 5M synapse, 12G connects/s AER hardware sensory - processing - learning - actuating system for high-speed visual object recognition and tracking,” IEEE Transactions on Neural Networks, vol. 20, no. 9, pp. 1417-1438, 2009.
[5] X. Jin, et al., “Modeling spiking neural networks on SpiNNaker,” Computing in Science & Engineering, vol. 12, no. 5, pp. 89-94, 2010.
[6] P. A. Merolla, et al., “A million spiking-neuron integrated circuit with a scalable communication network and interface,” Science, vol. 345, no. 6197, pp. 668-673, 2014.
[7] J. Partzsch, et al., “Analyzing the scaling of connectivity in neuromorphic hardware and in models of neural networks,” IEEE Transactions on Neural Networks, vol. 22, no. 6, pp. 919-940, 2011.
[8] S. Yu, et al., “Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect,” IEEE International Electron Devices Meeting (IEDM), pp. 451-454, 2015.
[9] L. Gao et al., “Fully parallel write/read in resistive synaptic array for accelerating on-chip learning,” Nanotechnology, vol. 26, no. 45, 2015, Art. no. 455204.

[10] S. Park et al., “Neuromorphic speech systems using advanced ReRAM based synapse,” in Proc. IEEE Int. Electron Device Meeting (IEDM), Washington, DC, USA, 2013, pp. 621 – 629
[11] J. Woo et al., “Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems,” IEEE Electron Device Lett., vol. 37, no. 8, pp. 992 – 998, Aug. 2016.
[12] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H.-S. P. Wong, “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing,” Nano Lett., vol. 12, no. 5, pp. 2176 – 2190, 2011.
[13] M. Suri et al., “Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction,” in Proc. IEEE Int. Electron Devices Meeting (IEDM), Washington, DC, USA, 2011, pp. 76 – 85.
[14] S. Yu, Ed., Neuro-Inspired Computing Using Resistive Synaptic Devices Cham, Switzerland: Springer, 2017.
[15] P. Gu, et al.,“Technological exploration of rram crossbar array for matrix-vector multiplication,” Asia and South Pacific Design Automation Conference (ASP-DAC), pp. 106-111, 2015.
[16] M. Hu, et al., “Memristor crossbar based hardware realization of BSB recall function,” International Joint Conference on Neural Networks (IJCNN), pp. 1-5, 2012.
[17] P. Gu, et al., “Technological exploration of rram crossbar array for matrix-vector multiplication,” Asia and South Pacific Design Automation Conference (ASP-DAC), pp. 106-113, 2015.
[18] P. Chen et al., "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," 2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2015, pp. 194-199,

[19] Geoffrey W Burr, Robert M Shelby, Severin Sidler, Carmelo Di Nolfo, Junwoo Jang, Irem Boybat, Rohit S Shenoy, Pritish Narayanan, Kumar Virwani, Emanuele U Giacometti 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 Transactions on Electron Devices 62(11), pp. 3498 - 3503, 2015.
[20] Boybat, I., Le Gallo, M., Nandakumar, S.R. et al. Neuromorphic computing with multi-memristive synapses. Nat Commun 9, 2514 , pp 1-5, 2018.
[21] https://github.com/neurosim/MLP_NeuroSim_V3.0
[22] Qian, N. (1999). On the momentum term in gradient descent learning algorithms. Neural Networks : The Official Journal of the International Neural Network Society, 12(1), pp. 145 - 151, 1999
[23] Duchi, J., Hazan, E., & Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research, 12, pp. 2121 - 2159, 2011.
[24] Kingma, D. P., & Ba, J. L. Adam: a Method for Stochastic Optimization. International Conference on Learning Representations, pp. 1 - 13, 2015.
[25] P. Chi et al., “PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory,” in Proc. ACM/IEEE Int. Symp. Comput. Archit. (ISCA), Seoul, South Korea, 2016, pp. 24 - 33.
[26] X. Liu et al., “Harmonica: A framework of heterogeneous computing systems with memristor-based neuromorphic computing accelerators,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 63, no. 5, pp. 614 - 626, May 2016.
[27] A. Shafiee et al., “ISAAC: A convolutional neural network

accelerator with in-situ analog arithmetic in crossbars," in Proc. Int. Symp. Comput. Archit. (ISCA), Seoul, South Korea, 2016, pp. 14 – 26.
[28] P. Chen, X. Peng and S. Yu, "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," 2017 IEEE International Electron Devices Meeting (IEDM), pp. 6.1.1-6.1.4, 2017.
[29] S. H. Jo et al., "Nanoscale memristor device as synapse in neuromorphic systems," Nano Lett., vol. 10, no. 4, pp. 1297 – 1301, 2010.
[30] L. Gao et al., "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," Nanotechnology, vol. 26, no. 45, 2015, Art. no. 455204.
[31] S. Park et al., "Neuromorphic speech systems using advanced ReRAM based synapse," in Proc. IEEE Int. Electron Device Meeting (IEDM), Washington, DC, USA, 2013, pp. 625 – 628
[32] Chen, Pai Yu ; Peng, Xiaochen ; Yu, Shimeng. / NeuroSim+ : An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. 2017 IEEE International Electron Devices Meeting, IEDM 2017. Institute of Electrical and Electronics Engineers Inc., 2018. pp. 6.1.1-6.1.4
[33] Ibid., pp. 6.13.
[34] P. Chen, X. Peng and S. Yu, "NeuroSim: A Circuit-Level Macro Model for Benchmarking Neuro-Inspired Architectures in Online Learning," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 37, no. 12, pp. 3067-3080, Dec. 2018.
[35] M. Hu, H. Li, Q. Wu, and G. S. Rose, "Hardware realization of BSB recall function using memristor crossbar arrays," in Proc. Design Autom. Conf. (DAC), San Francisco, CA, USA, 2012, pp. 498 – 503.

[36] P. Chen, "NeuroSim: A Circuit-Level Macro
Model for Benchmarking Neuro-Inspired Architectures in Online
Learning,", pp. 3071-3080.
[37] Ibid., pp. 3076.
[38] S. H. Jo et al., “Nanoscale memristor device as synapse in
neuromorphic systems,” pp. 1299-1301.
[39] L. Gao et al., “Fully parallel write/read in resistive
synaptic array for accelerating on-chip learning,” pp. 455204.
[40] S. Park et al., “Neuromorphic speech systems using
advanced ReRAM-based synapse,” pp. 625 - 628.
[41]Choi, Shinhyun & Tan, Scott & Li, Zefan & Kim, Yunjo & Choi,
Chanyeol & Chen, Pai-Yu & Yeon, Hanwool & Yu, Shimeng & Kim,
Jeehwan. (2018). SiGe epitaxial memory for neuromorphic computing
with reproducible high performance based on engineered dislocations.
Nature Materials. 17.
[42] M. Jerry et al., “Ferroelectric FET analog synapse for
acceleration of deep neural network training,” in Proc.
IEEE Int. Elect. Dev. Meet., 2017, pp. 139 - 142
[43] J. Woo et al., “Improved synaptic behavior under identical pulses
using AlOx /HfO2 bilayer RRAM array for neuromorphic systems,”
IEEE Electr. Dev. Lett., vol. 37, no. 8, pp. 994 - 997, 2016.
[44] P. Chen and S. Yu, "Technological Benchmark of Analog
Synaptic Devices for Neuroinspired Architectures," in IEEE Design &
Test, vol. 36, no. 3, pp. 31-38, June 2019.
[45] Sebastian, A., Le Gallo, M. & Krebs, D. Crystal growth within a
phase change memory cell. Nat. Commun. 5, 4314 (2014).
[46] Boybat, I., Le Gallo, M., Nandakumar, S.R. et al. Neuromorphic
computing with multi-memristive synapses. pp. 3-4.
[47] Boybat, I., Le Gallo, M., Nandakumar, S.R. et al. Neuromorphic
computing with multi-memristive synapses. supplementary note.

Abstract

Analysis to improve online learning of neuromorphic system with multi-memristive synapses

Young Woon Cho

Department of Material science and Engineering

The Graduate School

Seoul National University

The cross point array architecture with eNVM synaptic device has been proposed for on-chip implementation of weighted sum and weight update in the training operation. However, the non-ideal properties of the synaptic devices, such as the nonlinearity and asymmetry in weight increase or decrease, the limited number of conductance states and, device variations, hinder the learning accuracy. To alleviate this issue, the proposed strategies include 1) multi-synapses architecture to increase the number of conductance states, 2) training optimizer with a smart algorithm for mitigating device variation, 3) linear and symmetry scaling to achieving linear weight update.

Firstly, we have built a device behavioral model to capture the nonlinear weight update behavior of multi-memristive synapses based on experimental data including characteristics of eNVM devices. And then, by using 2-layer Multi-layer perceptron simulator, we performed

online learning and offline classification with MNIST handwritten digits as the training and testing data-set. Furthermore, we pointed out the phenomena, such as increasing the number of conductance states, improving linearity and asymmetry during the training process with multi-synapses architecture. Based on these results, various training optimizers including momentum, RMSprop, and Adam combined with multi-synapses architecture were performed for additional verification. As a result, training optimizer and multi-synapses architecture can be integrated appropriately, especially the highest learning accuracy(89%) which is much higher than the case(30%) single synapse with SGD was obtained by Adam optimizer with multi-synapses in the condition of nonlinear and asymmetric synaptic devices.

Even so, multi-synapses architecture faced difficulties with complexity and a large volume of neuromorphic circuits as the number of devices per synapse increases. Moreover, there has not been yet found a technique to implement training optimizer in hardware. To overcome these issues, we suggested linear and symmetric scaling with multi-synapses architecture. Consequently, the highest learning accuracy(94%), which is near to the learning accuracy with software training, was obtained from the devices that have bad linearity and symmetry.

keywords : emerging nonvolatile memory, multi-memristive synapses, non-ideal properties, non-linearity, asymmetry, number of conductance states, device variation, learning accuracy

Student Number **: 2019-24121**