



(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0080087  
(43) 공개일자 2017년07월10일

(51) 국제특허분류(Int. Cl.)  
G06F 1/32 (2006.01) G06F 1/04 (2006.01)  
G06N 3/04 (2006.01) G06N 3/06 (2006.01)  
(52) CPC특허분류  
G06F 1/3237 (2013.01)  
G06F 1/04 (2013.01)  
(21) 출원번호 10-2015-0191287  
(22) 출원일자 2015년12월31일  
심사청구일자 2015년12월31일  
기술이전 희망 : 기술양도

(71) 출원인  
인천대학교 산학협력단  
인천광역시 연수구 아카데미로 119 (송도동)  
(72) 발명자  
정재용  
인천광역시 연수구 해송로 70, 211동 1303호(웰카운티아파트2단지)  
(74) 대리인  
특허법인충정

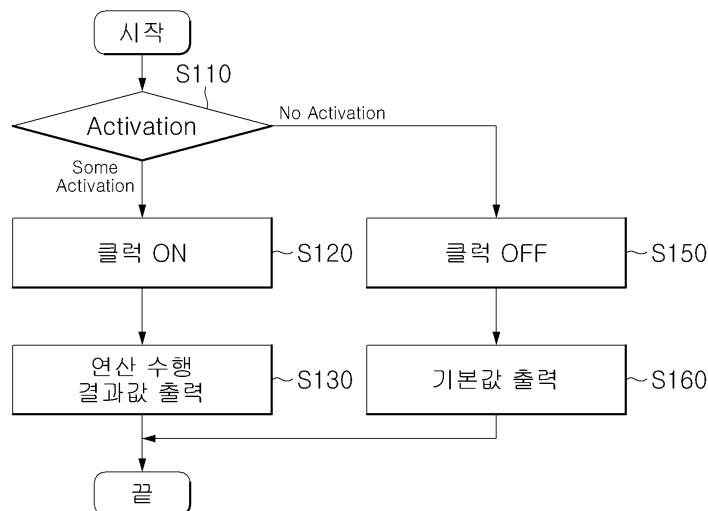
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치 및 방법**

**(57) 요약**

본 발명은 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치 및 방법으로서, 인공 신경망에서 뉴런의 활동을 감지하고 활동 정도에 따른 회소 활동에 대하여 처리 요소의 클럭을 차단함으로써 동기 회로에서 불필요하게 소비되는 동적 전력을 절감할 수 있는 장치와 방법을 제시한다.

**대표도** - 도6



(52) CPC특허분류

*G06F 1/3287* (2013.01)

*G06N 3/04* (2013.01)

*G06N 3/06* (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1345232925

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 이공학개인지초연구지원

연구과제명 사물 인터넷을 위한 범용 임베디드 뉴로모픽 프로세서 연구

기 여 율 1/1

주관기관 인천대학교 산학협력단

연구기간 2015.05.01 ~ 2016.04.30

---

## 명세서

### 청구범위

#### 청구항 1

인공 신경망(neural networks)에서 뉴론(neuron)의 활동을 감지하는 활동 감지부; 및  
상기 활동 감지에 기초하여 상기 인공 신경망의 처리 요소(processing element)에 대한 클럭 스위칭을 수행하는 클럭 게이팅 회로부를 포함하는 것을 특징으로 하는 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치.

#### 청구항 2

제 1 항에 있어서,  
상기 활동 감지부는,  
상기 뉴론의 활동에 따른 상기 뉴론의 출력이 제로(zero)가 되는 상기 뉴론의 회소(sparse) 활동을 감지하며,  
상기 클럭 게이팅 회로부는 상기 뉴론의 회소 활동 감지에 따라 상기 처리 요소의 클럭을 차단하는 것을 특징으로 하는 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치.

#### 청구항 3

제 1 항에 있어서,  
상기 클럭 게이팅 회로부는,  
뉴로모픽 컴퓨팅 시스템의 단위 셀에 포함된 레지스터의 클럭 핀에 입력되는 클럭을 스위칭하는 것을 특징으로 하는 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치.

#### 청구항 4

제 1 항에 있어서,  
상기 처리 요소는,  
세미시스토픽 멀리플라이어(semi-systolic multiplier)을 구성하는 단위 셀인 것을 특징으로 하는 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치.

#### 청구항 5

인공 신경망(neural networks)에서 뉴론(neuron)의 활동을 감지하는 뉴론 활동 감지 단계; 및  
상기 활동 감지에 기초하여 상기 인공 신경망의 처리 요소(processing element)에 대한 클럭 게이팅을 수행하는 클럭 게이팅 단계를 포함하는 것을 특징으로 하는 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 방법.

#### 청구항 6

제 5 항에 있어서,  
상기 뉴론 활동 감지 단계는,  
뉴론의 활동에 따른 상기 뉴론의 출력이 제로(zero)가 되는 상기 뉴론의 회소(sparse) 활동을 감지하며,  
상기 클럭 게이팅 단계는,  
상기 뉴론의 회소 활동 감지시에 상기 처리 요소에 대한 클럭을 차단하는 것을 특징으로 하는 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 방법.

## 청구항 7

제 6 항에 있어서,

상기 클럭 게이팅 단계의 수행에 따른 상기 처리 요소에 대한 클럭 차단으로 상기 처리 요소가 기본값을 출력하는 단계를 더 포함하는 것을 특징으로 하는 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 방법.

## 발명의 설명

### 기술 분야

[0001] 본 발명은 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치 및 방법에 대한 것으로서, 보다 상세하게는 인공 신경망의 뉴런 활동을 감지하고 활동 정도에 따라 회소 활동에 대한 처리 요소의 클럭을 차단함으로써 동기회로에서 불필요하게 소비되는 동적 전력을 절감할 수 있는 장치와 방법에 관한 것이다.

### 배경 기술

[0003] 딥 러닝(Deep learning)으로 알려진 뉴럴 네트워크의 최근 개발은 단지 인간만이 해결할 수 있는 문제들을 성공적으로 해결하고 있다. 딥 러닝(Deep learning) 기술은 이미지 이해, 음성 인식 검색, 자연어 처리, 무인 차량 운행 등과 같은 다양한 인지 어플리케이션을 위한 실질적인 도구가 되고 있다. 실제 현실 어플리케이션에 적용되기 위한 딥 러닝으로 거대 뉴럴 네트워크가 적용되며, 이는 수백만 요소를 포함한다.

[0004] 실시간으로 뉴럴 네트워크에서의 계산량은 방대하며 이로 인해 많은 전력 또한 소비하므로, 배터리 전력 시스템에서 심각한 문제가 된다. CPU는 실시간 계산 수행을 전달할 수 없고 모바일 CPU 및 FPGA 기반 엑셀러레이터는 10W 이상을 소비하는데, 이는 휴대폰 및 웨어러블과 같은 모바일 디바이스들에 적합하지 않다. 또한 항상 온 상태인 디바이스들의 경우, 더 엄격한 파워 소비가 요구된다. 그러므로 폰 노이만(Von Neumann) 구조 기반 시스템과는 근본적으로 다른 연산 플랫폼을 찾기 위한 개발 연구가 되고 있다.

[0005] 1980년대 시작된 뉴로모픽 엔지니어링(Neuromorphic engineering)은 아날로그 회로를 사용하는 신경 시스템 상의 신경생물학적 구조(neuro-biological architectures)를 모방하는 개념이지만, "뉴로모픽(neuromorphic)" 어원은 뉴런 시스템의 적용 모델인 회로와 시스템으로 폭넓게 사용되고 있으며, 뉴로모픽 컴퓨팅 시스템은 종래 컴퓨팅 플랫폼들의 전력과 여러 수행 문제들의 해결 대안이 되고 있다.

[0006] 그러나 최근 브레인스케일스(BrainScaleS)와 뉴로그리드(Neurogrid)와 같은 대규모 뉴로모픽 시스템이 브레인 시뮬레이션을 가속화시키고 인간 수준까지 확대시키기 위한 다른 목적으로 개발되었다. 이들은 주로 뇌과학 연구(neuroscience research)에 적용된다. 이와 같은 두가지 시스템과는 다르게, IBM에서 제안된 디지털 뉴로모픽 시스템인 트루노쓰(TrueNorth)는 브레인 시뮬레이션뿐만 아니라 실제 현실 어플리케이션에 적용하는 것을 목적으로 한다. 트루노쓰(TrueNorth)는 스파이킹 뉴런 모델(spiking neuron model)을 적용하고 뉴런이 활동하지 않을 때 적은 전력은 소비하도록 동기-비동기 디자인(synchronous-asynchronous design design)을 혼합하여 적용한다.

[0007] 최근 딥러닝 기반 어플리케이션을 위해 특별히 디자인된 뉴로모픽 컴퓨팅 시스템으로 인사이트(INSight)가 제안되었다. 이 시스템은 완전 동기 디자인을 채용함으로써, 현재의 디자인 자동화 툴들에 의해 구현이 제공될 수 있다. 그러나 인사이트(INSight)와 같은 동기 뉴로모픽 시스템은 클럭으로 인해 많은 양의 동적 전력 소비가 요구된다.

[0008] 본 발명에서는 이와 같은 디지털 동기 뉴로모픽 시스템을 위한 저전력 소비 방안을 제시하는데, 본 발명은 근본적으로 뉴로모픽 시스템을 위해 개발되었으나, 뉴럴 네트워크와 연관된 다양한 동기 프로세서들에도 적용될 수 있다.

## 발명의 내용

## 해결하려는 과제

- [0010] 본 발명은 상술한 바와 같은 종래 기술의 문제점을 해결하고자 하는 것으로서, 디지털 동기 뉴로모픽 시스템을 위한 저전력 소비 방안을 제시하는 것을 주된 목적으로 한다.
- [0011] 특히, 스파이크 방식의 뉴럴 네트워크 외의 다양한 동기 뉴로모픽 회로에서도 전력을 절감할 수 있는 방안을 제시하고자 하며, 동기 회로에서 클럭으로 인해 유용한 작업을 수행하지 않음에도 불구하고 많은 전력을 소비하는 문제를 해결하고자 한다.

## 과제의 해결 수단

- [0013] 상기 기술적 과제를 달성하고자 본 발명에 따른 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치는, 인공 신경망(neural networks)에서 뉴론(neuron)의 활동을 감지하는 활동 감지부; 및 상기 활동 감지에 기초하여 상기 인공 신경망의 처리 요소(processing element)에 대한 클럭 스위칭을 수행하는 클럭 게이팅 회로부를 포함할 수 있다.
- [0014] 바람직하게는 상기 활동 감지부는, 상기 뉴론의 활동에 따른 상기 뉴론의 출력이 제로(zero)가 되는 상기 뉴론의 회소(sparse) 활동을 감지하며, 상기 클럭 게이팅 회로부는 상기 뉴론의 회소 활동 감지에 따라 상기 처리 요소의 클럭을 차단할 수 있다.
- [0015] 나아가서 상기 클럭 게이팅 회로부는, 뉴로모픽 컴퓨팅 시스템의 단위 셀에 포함된 레지스터의 클럭 핀에 입력되는 클럭을 스위칭할 수 있다.
- [0016] 본 발명의 일실시예로서, 상기 처리 요소는, 세미시스토릭 멀리플라이어(semi-systolic multiplier)를 구성하는 단위 셀이 될 수 있다.
- [0017] 또한 본 발명에 따른 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 방법은, 인공 신경망(neural networks)에서 뉴론(neuron)의 활동을 감지하는 뉴론 활동 감지 단계; 및 상기 활동 감지에 기초하여 상기 인공 신경망의 처리 요소(processing element)에 대한 클럭 게이팅을 수행하는 클럭 게이팅 단계를 포함할 수 있다.
- [0018] 바람직하게는 상기 뉴론 활동 감지 단계는, 뉴론의 활동에 따른 상기 뉴론의 출력이 제로(zero)가 되는 상기 뉴론의 회소(sparse) 활동을 감지하며, 상기 클럭 게이팅 단계는, 상기 뉴론의 회소 활동 감지시에 상기 처리 요소에 대한 클럭을 차단할 수 있다.
- [0019] 나아가서 상기 클럭 게이팅 단계의 수행에 따른 상기 처리 요소에 대한 클럭 차단으로 상기 처리 요소가 기본값을 출력하는 단계를 더 포함할 수 있다.

## 발명의 효과

- [0021] 이와 같은 본 발명에 의하면, 인공 신경망의 뉴런 활동을 감지하고 활동 정도에 따른 회소 활동에 대한 처리 요소의 클럭을 차단함으로써 동기회로에서 불필요하게 소비되는 동적 전력을 절감할 수 있게 된다.
- [0022] 특히 동기 회로에서도 스파이크 기반의 비동기 회로와 같이 활동량에 따라 동적 전력을 절감하는 것이 가능하게 되어, 동기 뉴로모픽 컴퓨팅 시스템에 대한 저전력 소비 효과를 이룰 수 있게 된다.

## 도면의 간단한 설명

- [0024] 도 1은 레이어의 피드-포워드 뉴럴 네트워크에 대한 DAG(Directed Acyclic Graph)의 개념도를 나타낸다.
- 도 2는 알렉스넷(AlexNet)이라 불리는 실용적인 CNN(Convolutional Neural Networks)을 나타낸다.
- 도 3은 입력 이미지에 대한 알렉스넷(AlexNet)의 각 나선형 레이어의 활동들을 도시한다.
- 도 4는 CNN에서 각 레이어의 활동에 대한 평균 회소치를 나타낸다.

도 5는 본 발명에 따른 전력 절감 장치의 실시예에 대한 구성도를 도시한다.

도 6은 본 발명에 따른 전력 절감 방법의 일실시예에 대한 흐름도를 도시한다.

도 7은 세미 시스토크 멀티플라이어(semi-systolic multiplier)에 기반한 시넵스에 본 발명을 적용하는 실시예를 나타낸다.

도 8은 본 발명에 따른 전력 절감 방법의 구현예에 대한 흐름도를 나타낸다.

도 9는 4비트 세미 시스토크 멀티플라이어의 부분적 구성으로 생성되어 추가되는 개념을 도시한다.

도 10은 본 발명에서의 시그널 형태를 나타낸다.

도 11은 회소성이 변화할 때 기본 디자인과 본 발명에 따른 디자인에 대한 동적 파워와 전체 파워를 대비한 결과이다.

### 발명을 실시하기 위한 구체적인 내용

- [0025] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시예에 의하여 달성되는 목적을 설명하기 위하여 이하에서는 본 발명의 바람직한 실시예를 예시하고 이를 참조하여 살펴본다.
- [0026] 먼저, 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니며, 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다. 또한 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0027] 본 발명을 설명함에 있어서, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다.
- [0029] 본 발명은 인공 신경망의 회소 활동을 활용하는 동기 직접 회로의 소비 전력 절감 장치 및 방법으로서, 인공 신경망에서 뉴런의 활동을 감지하고 활동 정도에 따라 회소 활동에 대한 처리 요소의 클럭을 차단함으로써 동기 회로에서 불필요하게 소비되는 동적 전력을 절감할 수 있는 장치와 방법을 제시한다.
- [0031] 피드-포워드 뉴럴 네트워크(Feed-forward neural networks)는 실제 현실에 적용되는 뉴럴 네트워크의 가장 일반적인 형태이며, 위상적 순서(topological order)를 이용하는 유닛들이 구분되는 레이어들로 구성된다. 첫번째 레이어는 입력 레이어이고, 마지막 레이어는 출력 레이어이며, 다른 레이어들은 히든 레이어들이다.
- [0032] 피드-포워드 네트워크가 하나 이상의 히든 레이어를 갖는 경우, 이를 딥 뉴럴 네트워크라 말한다. 대부분 공통 레이어 형태는 완전히 연결된 레이어로서, 인접한 두 개의 레이어들 간의 뉴런들은 쌍으로 연결된다. m개 뉴런들과 완전히 연결된 n개 뉴런의 레이어는 비선형 변환을 수행한다.
- [0033] 입력 레이어에서 뉴런들의 활동과 상기 레이어의 출력을 각각  $\mathbf{x} \in \mathbb{R}^m$  와  $\mathbf{y} \in \mathbb{R}^n$  이라하면, 비선형 변환은 하기 [식 1]로 표현될 수 있다.
- [0034]  $\mathbf{y} = \mathbf{f}(\mathbf{W}\mathbf{x} + \mathbf{b})$  [식 1]
- [0035] 여기서,  $\mathbf{W} \in \mathbb{R}^{n \times m}$  는 가중치 행렬(weight matrix)이고,  $\mathbf{b} \in \mathbb{R}^n$  은 바이어스(bias)이며, f는 활동 함수이다.
- [0036] 도 1은 3개 레이어의 피드-포워드 뉴럴 네트워크에 대한 DAG(Directed Acyclic Graph)를 개념도를 나타낸다. 상기 도 1에서 원은 뉴론들을 나타내고 라인들은 연결을 나타내는데, 히든 레이어와 출력 레이어들은 완전히 연결된다.
- [0038] 나선형 뉴럴 네트워크(Convolutional Neural Networks; CNNs)는 멀티-레이어 뉴럴 네트워크로서, 일반적으로

시각 인식 태스크로 이용되며 입력으로서 이미지를 획득한다. CNNs는 다양한 이미지 인식 태스크들에서 획기적인 발전을 이루었고 딥 러닝에 대한 관심을 불러일으켰다.

- [0039] CNNs에서 뉴런들은 2차원 공간으로 구성된다. 여기서 차원은 차원 정보로서 폭과 높이 및 깊이이다. 예를 들어, 입력 이미지의 각 채널의 픽셀들에 의해 활동되는 입력 뉴런들은 픽셀들 및 채널들의 수직 및 수평 위치에 따라 배치될 수 있다. CNN의 각 레이어는 3차원 볼륨으로 표현되는 입력 활동을 3차원 출력 볼륨으로 변환시킨다.
- [0040] CNN은 두가지 주요 형태의 레이어들로 이루어질 수 있는데, 나선형 레이어와 완전히 연결된 레이어이다. 나선형 레이어에서 뉴런들은 공간적인 지역 연결을 가지며, 동일 깊이의 뉴런들은 지역 연결을 위한 가중치(weights)를 공유하므로, 레이어는 컨벌루션(convolution)을 수행한다. 상기 지역 연결의 공간적 규모는 수용장(receptive field)이라 불린다. 나선형 레이어는 풀링 레이어(pooling layer)와 LRN(Local Response Normalization) 레이어를 포함할 수 있다.
- [0041] 도 2는 알렉스넷(AlexNet)이라 불리는 실용적인 CNN을 나타낸다.
- [0042] 상기 알렉스넷(AlexNet)은 한계를 뛰어넘는 획기적인 도전으로 많은 관심을 얻었다. 상기 알렉스넷은 입력으로 224\*224 칼라 이미지를 획득하므로 입력 레이어에서 뉴런들은 224\*224\*3의 볼륨으로 표현되고, 획득된 이미지를 1000개 카테고리 분류한다. 상기 알렉스넷은 3개의 완전 연결된 레이어들에 의해 5개의 나선형 레이어들을 갖고 60M(메가) 파라미터들과 650K(킬로) 뉴런들을 포함한다. 첫번째 나선형 레이어는 11\*11 수용장(receptive field)을 갖고, 풀링 레이어(pooling layer)와 LRN 레이어를 포함한다. 상기 네트워크는 1.2M(메가) 트레이닝으로 분리되고 50K(킬로) 유효 이미지인 ILSVRC-2012 데이터셋(dataset)에서 18.2%의 상위 5위 에러 레이트(top-5 error rate)를 달성한다.
- [0043] IBM에서 제안된 디지털 뉴로모픽 시스템인 트루노쓰(TruNorth)는 동기-비동기가 혼합된 뉴로모픽 시스템으로 뉴런들(neurons)과 시냅스들(synapses)로 구성된다. 디지털, 스파킹(spiking) 뉴론 모델이 이용되며, 코딩율(rate coding)이 적용될 때 뉴론의 활동이 시간 스텝, 즉 시간 윈도우의 특별한 숫자인 1(spikes)의 숫자로 표현된다. 뉴런들은 AER(Address Event Representation)을 통해 소통하며, 여기서 스파이크들(spikes)이 발생하는 시간들이 전송된다. 트루노쓰(TruNorth)는 이벤트 기반이기 때문에, 코어(core)에 대한 AER 패킷이 수신되지 않는 경우, 비동기 시스템으로서 코어의 시냅스들과 뉴런들에서 아무 활동이 발생되지 않는다. 그러므로 뉴런들이 스파이크되지 않는, 즉 뉴런들이 활동되지 않을 때 회로의 스위칭 활동은 매우 낮다.
- [0044] 이와 같이 비동기 회로는 단지 작동시에만 동적 전력을 소비하지만, 반면에 동기 회로는 클럭으로 인해 유용한 작업을 수행하지 않음에도 불구하고 많은 전력을 소비하는 과약된다. 그러나 클럭 게이팅이 고려되는 경우에는 그렇지 않을 수 있다. 클럭 게이팅은 클럭의 전력 문제를 해결하는 저전력 회로 디자인 기술이며, 최근 디자인 자동화 툴들에 의해 제공된다.
- [0045] 이러한 툴들은 플롭(flop)이 존재하는 경우, 각 플롭에 대한 '인에이블(enable)' 신호를 자동적으로 추출하고, 동일한 '인에이블' 신호를 공유하는 플롭들의 그룹을 위해 기존 클럭(original clock)을 '인에이블' 신호에 의해 게이팅된 새로운 클럭으로 대체한다. 이는 클럭 트리들(clock trees)에서 많은 양의 클럭 부하를 보호하고 클럭 핀들이 불필요하게 스위칭하는 것을 제거한다. 이와 같은 변환은 매우 효과적이며, 심지어 10 플롭들 이하의 미세 레벨(fine-grained level)에서도 수행된다. 디지털 신호처리 회로에서, '인에이블' 신호들은 보통 '데이터 유효(data valid)' 신호들이 된다.
- [0046] 그럼에도 불구하고, 회소 활동으로 인해 뉴로모픽 컴퓨팅 시스템에 클럭 게이팅된 회로를 적용하는 경우보다 비동기 회로가 여전히 더 전력이 효율적이다. 정류된 선형 유닛(Rectified Linear Unit; ReLU)  $f(x)=\max(0,x)$ 는 최단 딥 뉴럴 네트워크에서 활동 함수(activation function)를 위한 가장 일반적인 선택이다. 이는  $x$ 가 0보다 작을 때 정확히 제로(zero)가 생성되기 때문에 레이어의 활동은 보통 많은 제로(zero)들을 갖는데, 즉 이것이 회소(sparse)이다. 또한 출력 레이어로 갈수록 활동들이 더 회소되는 경향이 있다.
- [0047] 도 3은 입력 이미지에 대한 알렉스넷(AlexNet)의 각 나선형 레이어의 활동들을 도시한다.
- [0048] 상기 도 3의 (a)는 입력 이미지를 나타내며, 상기 도 3의 (b) 내지 (f)는 알렉스넷(AlexNet)의 각 나선형 레이어에서 뉴런들의 활동들을 시각적으로 나타낸다. 각 레이어의 출력 3차원 볼륨은 깊이 차원에 따라 분리되며, 각 분리 조각이 타일링(tile)된다. 첫번째와 두번째 나선형 레이어는 LRN에 의해 뒤따르는 풀링(pooling)을 포함하며, 다섯번째 레이어는 단지 풀링(pooling)만을 포함한다.
- [0049] 상기 도 3에서 제로 활동은 검정으로 나타나는데, 이들 활동 맵은 대부분의 활동들이 제로나 작기 때문에 검게

나타나며, 이미지는 출력 레이어로 갈수록 더 검게 된다.

- [0050] 도 4의 (a)는 ILSVRC 2012 유효 데이터 셋에서 첫번째 256 입력 이미지들에 대한 알렉스넷(AlexNet)에서 각 레이어의 활동에 대한 평균 회소치를 나타낸다.
- [0051] 상기 도 4의 (a)는 알렉스넷(AlexNet)에서 각 레이어의 출력 활동에 대한 회소치를 256 이미지들에 대하여 측정 한 것으로서, 출력 레이어로 갈수록 회소치는 증가하며, 출력 레이어 근방에서 0.8까지 높아진다.
- [0052] 도 4의 (b)는 ILSVRC에 대한 기술 수준 결과로 생성되어 19 레이어와 160M(메가) 파라미터를 갖는 가장 최근의 CNN으로서 VGG-19과 동일한 경향을 나타낸다.
- [0053] 활동에 대한 회소성은 스파이크 기반 비동기 회로에서 활용될 수 있다. "무활동(No activation)"은 보통 제로 스파이크(zero spike)로 인코딩되므로, 전혀 동적 전력을 소비하지 않는다. 이를 클럭 게이팅으로 적용함으로써 동기 회로에서도 활용할 수 있다.
- [0054] 도 5는 본 발명에 따른 전력 절감 장치의 실시예에 대한 구성도를 도시한다.
- [0055] 본 발명에서는 뉴론들의 활동에 따라 처리 요소(processing element)를 위한 클럭을 게이트하는데, 처리 요소(processing element)에 대한 클럭은 "무활동" 감지에 의해 게이트될 수 있다. 이와 같은 접근에서, 오버헤드(overhead)가 이득을 추월하지 않도록 난제로(nonzero) 활동 감지를 작게 유지하는 것이 중요하다.
- [0056] 이를 위해 본 발명에 따른 전력 절감 장치(200)는, 개략적으로 처리 요소(processing element)(100)에 대한 활동을 감지하는 활동 감지부(210)와 활동 감지부(210)에서 감지한 회소 활동을 기초로 처리 요소(100)에 대한 클럭을 스위칭하는 클럭 게이팅 회로부(250)를 포함할 수 있다.
- [0057] 도 6은 본 발명에 따른 전력 절감 방법의 일실시예에 대한 흐름도를 도시한다.
- [0058] 본 발명에서는 상기 도 5와 같은 전력 절감 장치를 적용하여 뉴론들의 활동에 따라 처리 요소를 위한 클럭 게이팅을 통해 전력을 절감하는데, 활동 감지부(210)가 뉴론들의 활동을 감지(S110)하여, 뉴론의 활동이 감지되는 경우 클럭 게이팅 회로부(250)가 처리 요소(100)로 클럭을 제공(S120)하여 연산 수행에 따른 결과값이 출력(S130)되지만, 만약 활동 감지부(210)가 뉴론들이 무활동인 것으로 판단하는 경우에는 클럭 게이팅 회로부(250)가 처리 요소(100)에 대한 클럭을 차단(S150)하여 단순히 기본값만 출력(S160)되게 된다.
- [0059] 본 발명은 이와 같은 구성을 인사이트(INsight) 뉴로모픽 컴퓨팅 시스템에 어떻게 실현시킬지를 제안한다. 인사이트(INsight)는 시냅스들이 주된 처리 요소(processing element)인 디지털 동기 뉴로모픽 시스템이다.
- [0060] 도 7은 세미 시스토크 멀티플라이어(semi-systolic multiplier)에 기반한 시냅스에 본 발명을 적용하는 실시예를 나타낸다.
- [0061] 본 발명에서는 단위 셀들(110a, 110b, 110c, 110d)에 있는 레지스터들의 클럭 핀들에 동적 전력 소비를 줄이는 간단한 회로로서 본 발명에 따른 전력 절감 장치(200)를 추가함으로써 입력 활동의 제로들에 의해 스위칭되지 않을 수 있다.
- [0062]  $n$  비트 바이너리로 표현되는 뉴론의 활동과 시냅스의 가중치(weight)를 각각  $x$ 와  $y$ 라 설정하고,  $x$  및  $y$ 의  $i$ 번째 비트를  $x_i$ 와  $y_i$ 로 나타낸다. 시냅스의 입력으로  $x$ 의 각 비트는 연속적인 최소 유효 비트(LSB)가 된다.
- [0063] 세미 시스토크 멀티플라이어(semi-systolic multiplier)는 두 개의 레지스터와 두 개의 입력 AND 게이트 및 전가산기(full adder)로 구성된 단위 셀(primitive cell)에 기반하며,  $n$  비트 세미 시스토크 멀티플라이어는  $n$ 개의 단위 셀들로 구성된다.
- [0064]  $PC_0$ 부터  $PC_{n-1}$ 로 단위 셀들을 표현하면, 시냅스의 가중치(weight)는 병렬의 시냅스로부터 제공되어  $PC_i$ 는  $y_i$ 를 갖는다. 멀티플라이어는 입력을 위한 공통 와이어(global wire)가 요구되기 때문에 "세미 시스토크(semi-systolic)"으로 칭해지며, 로컬 와이어들을 통해 이웃한 처리 요소(processing element)들이 상호 연결되는 시스토크 시스템과는 다르다. 공통 와이어는 긴 와이어이기에 고주파 동작에 적절하지는 않을지라도 뉴로모픽 시스템에 맞게 더욱 조밀한 구성을 가능하게 한다.
- [0065] 상기 도 7과 같은 본 발명에 따른 실시예에 본 발명에 따른 전력 절감 방법을 적용하는 경우, 도 8에 따른 본 발명에 따른 전력 절감 방법의 구현예에 대한 흐름도가 적용될 수 있다.
- [0066] 뉴론 활동  $x$ 를 감지(S210)하여,  $x$ 가 0이 아닌 경우에는 각 단위 셀들(110a, 110b, 110c, 110d)에 클럭을 인가

하여 각 단위 셀들(110a, 110b, 110c, 110d)에서 뉴런 활동  $x$ 와 가중치(weight)를 고려한 결과값  $y$ 가 출력(S230)되게 된다.

[0067] 그러나 뉴런 활동  $x$ 가 0인 경우에는 각 단위 셀들(110a, 110b, 110c, 110d)에 클럭을 차단(S250)하여 각 단위 셀들(110a, 110b, 110c, 110d)에서 기본값인 0이 출력(S260)된다.

[0068] 도 9는 4비트 세미 시스토크 멀티플라이어의 부분적 구성으로 생성되어 추가되는 개념을 도시한다.

[0069] 세미 시스토크 멀티플라이어에서 각 단위 셀은 부분 구성과의 연결된 동작들에 기여한다. 각 단위 셀은 부분 구성으로 적용되고, 동일 주기로 점들의 대각선 상에 디지털(digit)들로 생성된다. 멀티플라이어는 구성의 각 비트를 개별적으로 생성하여 부분 구성들을 점차적으로 추가한다.

[0070] 상기 도 9에서 A는 부분 구성이 어떻게 추가되는지를 나타낸다.  $PC_i$ 의 전가산기(full adder)는 하나의 입력에 따라 현재 주기(current cycle)의 1 비트 구성을 갖는다. 나머지 두 입력은 플립을 통한 이전 주기로부터 얻는데, 하나는 동일한 단위 셀  $PC_i$ 로부터 얻고 다른 하나는  $PC_{i+1}$ 로부터 얻는다.

[0071] 부호 곱셈(signed multiplication)을 위해, 피승수  $x$ 는 마지막 4 주기 동안 부호 확장(signed-extended)이 요구되고,  $PC_{n-1}$ 은  $y$ 의 신호 비트(sign bit)가  $-2^{N-1}$ 의 가중치(weight)를 갖기 때문에 수정이 요구된다.

[0072] 그러므로, 보상 레지스터(Cout register)를 1로 초기화하고, AND 게이트를 NAND 게이트로 대체한다. 수정된 PC는 PC'로 나타내며, 마지막 부분 구성은 제외한다.

[0073] 세미 시스토크 멀티플라이어에서 레지스터들은 입력으로부터 첫번째 '1'이 나타날 때까지 초기값을 유지한다. 첫번째 '1'이전에 '0'이 나타나는 경우, 전가산기(full adder)가 캐리(carry)를 생성하는 것을 유지하기 때문에,  $PC'_{n-1}$ 에서 보상 레지스터의 값은 '1'이 유지된다. 그러므로 '0'들이 나타나는 경우, 레지스터들에 새로운 값을 부여할 필요가 없고, 클럭 게이팅이 가능하다.

[0074] 상기 도 7에서 활동 감지부(210)는 이와 같은 데이터 구동 클럭 게이팅을 위해 추가되는 로직인데, 단지 두 개의 추가 결합 게이트들과 하나의 추가 레지스터가 클럭 게이팅 상태를 감지할 수 있다.

[0076] 이하에서는 본 발명을 적용하는 경우에 대한 실험적 결과를 살펴보기로 한다.

[0077] n비트 세미 시스토크 멀티플라이어의 기준 방식과 베릴로그(Verilog)로 제안된 방식을 적용하였으며, Synopsys Design Compiler를 사용하여 합성하고 0.95V, 125도 및 SS 코너(corner)에 특화된 높은 Vt 라이브러리를 사용하는 32nm 기술에 맵핑하였다.

[0078] 클럭 게이팅은 합성(synthesis)동안 자동적으로 수행된다. 합성의 결과를 하기 표 1로 요약하였다.

[0079] [표 1]

| Metric                  | Word size |          |          |          |
|-------------------------|-----------|----------|----------|----------|
|                         | 8-bit     |          | 16-bit   |          |
|                         | Baseline  | Proposed | Baseline | Proposed |
| Critical path delay     | 0.73ns    | 0.73ns   | 0.78ns   | 0.78ns   |
| Cell area ( $\mu m^2$ ) | 210.18    | 218.06   | 405.36   | 413.24   |
| Comb. cell count        | 39        | 39       | 72       | 72       |
| Seq. cell count         | 17        | 18       | 33       | 34       |
| Gated reg.              | 15        | 16       | 31       | 32       |

[0080]

[0081] 기준 방식은 단위 셀들을 위한  $2n-1$  레지스터들과 추가로 2개의 레지스터들을 가지며,  $PC_0$ 에 대한 합계 레지스터를 제거하였다. 직렬 입력의 부호 확장(sign extension)을 위해 하나의 추가 레지스터를 적용하였고, 클럭 게이팅 셀에 래치(latch)를 적용하였다.

[0082]  $2n-1$  레지스터들은 정확하게 클럭 게이팅되며, 부호 확장(sign extension)을 위한 레지스터는 나머지들과 다른 인에이블 상태를 갖기 때문에 클럭 게이팅되지 않는다. 결합 게이트 카운트는 동기 리셋을 갖는 플립 플롭들이 AND 게이트와 더 좋은 타이밍(better timing)을 위한 간단한 D 플립 플롭으로 맵핑된다. 그러므로 게이트 레벨로 설계된다면 더 영역을 줄이는 것이 가능하다. 본 발명에서는 기본 방식에 단지 레지스터를 하나 더 추가하

로 셀 영역에서의 차이가 미미하다.

[0083] 도 10은 본 발명에서의 시그널 형태를 나타낸다.

[0084] 내부 레지스터에 대한 클럭 시그널은 입력 활동에 따라 스위칭되는데, 시그널 형태에서 'last in ph0'은 부호 확장(sign extension)이 수행될 필요가 있는 경우를 나타내고, 'last'는 활동의 마지막 비트를 의미한다.

[0085] '유효 데이터(data in valid)'와 'last'가 나타남에 따라 입력이 계속적으로 나타나지만, 연결된 뉴런이 활동하지 않는 경우 내부 레지스터들의 클럭 시그널은 스위칭하지 않으므로 동적 전력이 절감된다.

[0086] 본 발명을 적용하여 임의적으로 1000벡터들의 20개 세트를 생성하고 상기 벡터 세트의 희소성이 0.0부터 1.0까지 0.05 간격으로 변화도록 설정하고, 각 합성된 넷리스트(netlist)를 위해, Synopsys VCS를 이용하고 SAIF (Switching Activity Interchange Format) 파일을 생성하여 이들 벡터 세트를 적용한 게이트 레벨 시뮬레이션을 수행하였다.

[0087] 각 SAIF 파일에서의 스위칭 동작을 넷리스트(netlist)에 기록(back-annotated)하고, 동적 전력 분석을 각 벡터 세트에 대한 파워 리포트를 도출하는 Design Compiler에서 수행하였다.

[0088] 하기 [표 2]는 희소성이 0.2, 0.6 및 0.8인 결과를 나타낸다.

[0089] [표 2]

| Sparsity | Power ( $\mu$ W) | Baseline (8-bit) |      |      |                    | Proposed (8-bit) |      |      |                    | Red  | Baseline (16-bit) |      |      |                    | Proposed (16-bit) |      |      |                    | Red  |
|----------|------------------|------------------|------|------|--------------------|------------------|------|------|--------------------|------|-------------------|------|------|--------------------|-------------------|------|------|--------------------|------|
|          |                  | clk              | seq  | comb | total <sup>a</sup> | clk              | seq  | comb | total <sup>b</sup> |      | clk               | seq  | comb | total <sup>a</sup> | clk               | seq  | comb | total <sup>b</sup> |      |
| 0.2      | Internal         | 1.1              | 16.3 | 1.5  | 19.1               | 1.1              | 12.8 | 1.5  | 15.5               | 0.81 | 1.3               | 33.1 | 3.2  | 37.7               | 1.2               | 25.4 | 3.2  | 29.8               | 0.79 |
|          | Switching        | 1.5              | 0.2  | 0.5  | 2.3                | 1.2              | 0.2  | 0.5  | 1.9                | 0.83 | 3.2               | 0.5  | 0.9  | 4.7                | 2.4               | 0.5  | 0.9  | 3.8                | 0.82 |
|          | Leakage          | 0.1              | 2.3  | 1.5  | 3.9                | 0.1              | 2.4  | 1.5  | 4.0                | 1.04 | 0.1               | 4.6  | 2.9  | 7.7                | 0.1               | 4.8  | 2.9  | 7.8                | 1.02 |
|          | Dynamic          | 2.7              | 16.5 | 2.0  | 21.4               | 2.3              | 13.1 | 2.1  | 17.4               | 0.81 | 4.6               | 33.6 | 4.1  | 42.3               | 3.5               | 25.9 | 4.1  | 33.6               | 0.79 |
|          | Total            | 2.8              | 18.8 | 3.5  | 25.3               | 2.4              | 15.5 | 3.5  | 21.5               | 0.85 | 4.7               | 38.3 | 7.1  | 50.0               | 3.7               | 30.7 | 7.0  | 41.5               | 0.83 |
| 0.6      | Internal         | 1.3              | 15.5 | 1.0  | 17.8               | 0.9              | 7.9  | 1.0  | 9.8                | 0.55 | 1.3               | 30.8 | 1.5  | 33.6               | 0.9               | 12.3 | 1.5  | 14.7               | 0.44 |
|          | Switching        | 1.5              | 0.1  | 0.4  | 2.1                | 0.7              | 0.1  | 0.3  | 1.2                | 0.56 | 3.2               | 0.2  | 0.5  | 3.9                | 1.2               | 0.2  | 0.4  | 1.8                | 0.46 |
|          | Leakage          | 0.1              | 2.2  | 1.5  | 3.8                | 0.1              | 2.4  | 1.5  | 4.0                | 1.04 | 0.1               | 4.4  | 2.9  | 7.4                | 0.1               | 4.6  | 2.9  | 7.6                | 1.02 |
|          | Dynamic          | 2.9              | 15.7 | 1.4  | 19.9               | 1.6              | 8.0  | 1.4  | 11.0               | 0.55 | 4.6               | 31.0 | 2.0  | 37.6               | 2.1               | 12.5 | 1.9  | 16.5               | 0.44 |
|          | Total            | 3.0              | 17.9 | 2.8  | 23.7               | 1.8              | 10.4 | 2.8  | 14.9               | 0.63 | 4.7               | 35.5 | 4.8  | 45.0               | 2.2               | 17.2 | 4.8  | 24.1               | 0.54 |
| 0.8      | Internal         | 1.3              | 15.0 | 0.6  | 16.8               | 0.8              | 4.5  | 0.6  | 5.8                | 0.35 | 1.3               | 29.4 | 0.5  | 31.3               | 0.8               | 5.2  | 0.5  | 6.5                | 0.21 |
|          | Switching        | 1.5              | 0.1  | 0.3  | 1.9                | 0.4              | 0.1  | 0.2  | 0.6                | 0.33 | 3.2               | 0.1  | 0.2  | 3.5                | 0.4               | 0.1  | 0.2  | 0.7                | 0.19 |
|          | Leakage          | 0.1              | 2.2  | 1.4  | 3.7                | 0.1              | 2.4  | 1.4  | 3.9                | 1.05 | 0.1               | 4.3  | 2.8  | 7.2                | 0.1               | 4.5  | 2.8  | 7.5                | 1.03 |
|          | Dynamic          | 2.9              | 15.0 | 0.8  | 18.7               | 1.2              | 4.5  | 0.8  | 6.5                | 0.35 | 4.6               | 29.5 | 0.8  | 34.8               | 1.2               | 5.3  | 0.7  | 7.1                | 0.21 |
|          | Total            | 3.0              | 17.2 | 2.3  | 22.4               | 1.3              | 6.9  | 2.2  | 10.4               | 0.46 | 4.7               | 33.8 | 3.6  | 42.1               | 1.3               | 9.8  | 3.5  | 14.6               | 0.35 |

[0090]

[0091] 상기 도 4에서와 같이, 알렉스넷(AlexNet)의 두번째 레이어와 7번째 레이어의 희소성은 대략 0.2와 0.8이고, 이를 통해 각 레이어에 대한 결과가 도출될 수 있다.

[0092] 기본 방식과 본 발명에 따른 방식에서, 워드 크기와 상관없이 대부분의 전력은 단락 전류로 인해 순차 논리 회로(sequential logic)에서 소비되었다.

[0093] 입력 희소가 0.2인 경우의 8비트 기본 디자인에서, 동적 파워는 전체 파워의 85%를 차지하고 정적 파워는 15%를 차지하였다. 희소성이 0.2일 때, 본 발명에 따른 8비트 디자인은 4% 정적 파워가 증가하나 19% 동적 파워를 절감하여 15%의 전체 파워 절감을 얻었다. 16 비트의 경우도 유사한 향상이 달성되었으며, 희소성이 0.8로 증가할 때, 더욱 현저한 향상이 나타났다. 본 발명에 따른 8비트 방식은 기본 방식보다 65%의 동적 파워를 절감하여 54%의 전체 파워를 절감했다.

[0094] 16비트로 워드 크기가 증가함에 따라 클럭 게이팅에 의해 클럭 핀들이 더 숨겨지기 때문에 더욱 현저하게 향상된다. 본 발명에 따른 방식은 기본 방식보다 79% 동적 파워를 덜 소비한다.

[0095] 도 11은 희소성이 변화할 때 기본 디자인과 본 발명에 따른 디자인에 대한 동적 파워와 전체 파워를 대비한 결과이다.

[0096] 기본 방식의 경우에도 희소성이 증가함에 따라 스위칭 활동이 감소하며, 결국 파워 소비가 줄어든다. 본 발명의 경우에는 스파이크 기반 비동기 회로일 때 희소성이 1.0으로 도달함에 따라 동적 파워 소비는 더 작아진다.

[0097] 이와 같이 본 발명은 딥 뉴럴 네트워크에서 희소 활동을 감지하여 동기 뉴로모픽 컴퓨팅 시스템에 대한 저전력 소비 효과를 이룰 수 있게 된다. 나아가서 동기 회로에서도 스파이크 기반의 비동기 회로와 같이 활동량에 따라 동적 전력을 절감하는 것이 가능하게 된다.

[0098] 상기에서는 본 발명을 세미 시스토픽 멀티플라이어(semi-systolic multiplier)에 기반한 디지털 시냅스 방식으로 설명하였으나 다양하게 제시될 수 있는 다른 형태의 디지털 회로에도 본 발명을 적용할 수 있다.

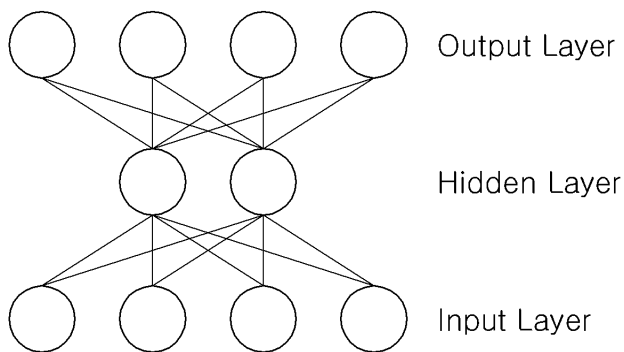
[0100] 이상의 설명은 본 발명의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서 본 발명에 기재된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 기술 사상이 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

### 부호의 설명

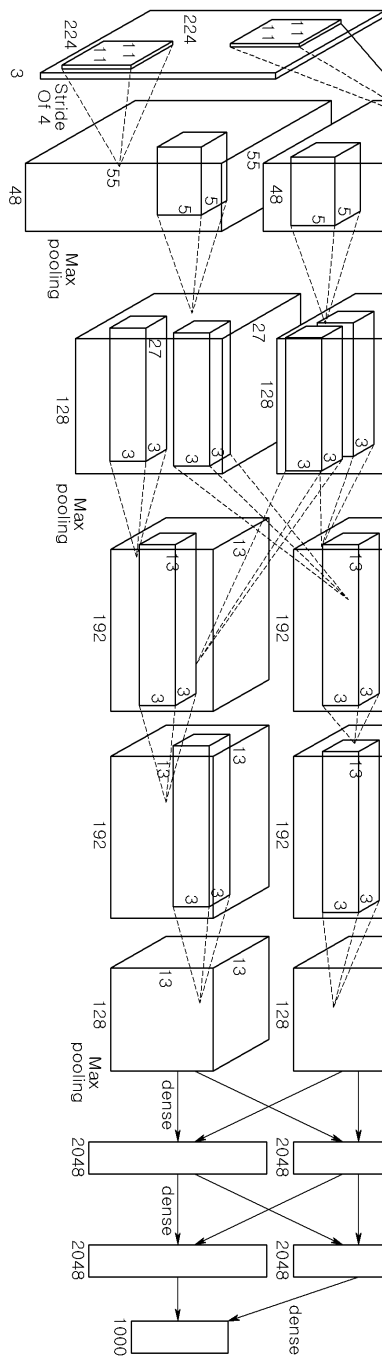
[0102] 100 : 처리 요소,  
200 : 전력 절감 장치,  
210 : 활동 감지부,  
250 : 클럭 게이팅 회로부.

### 도면

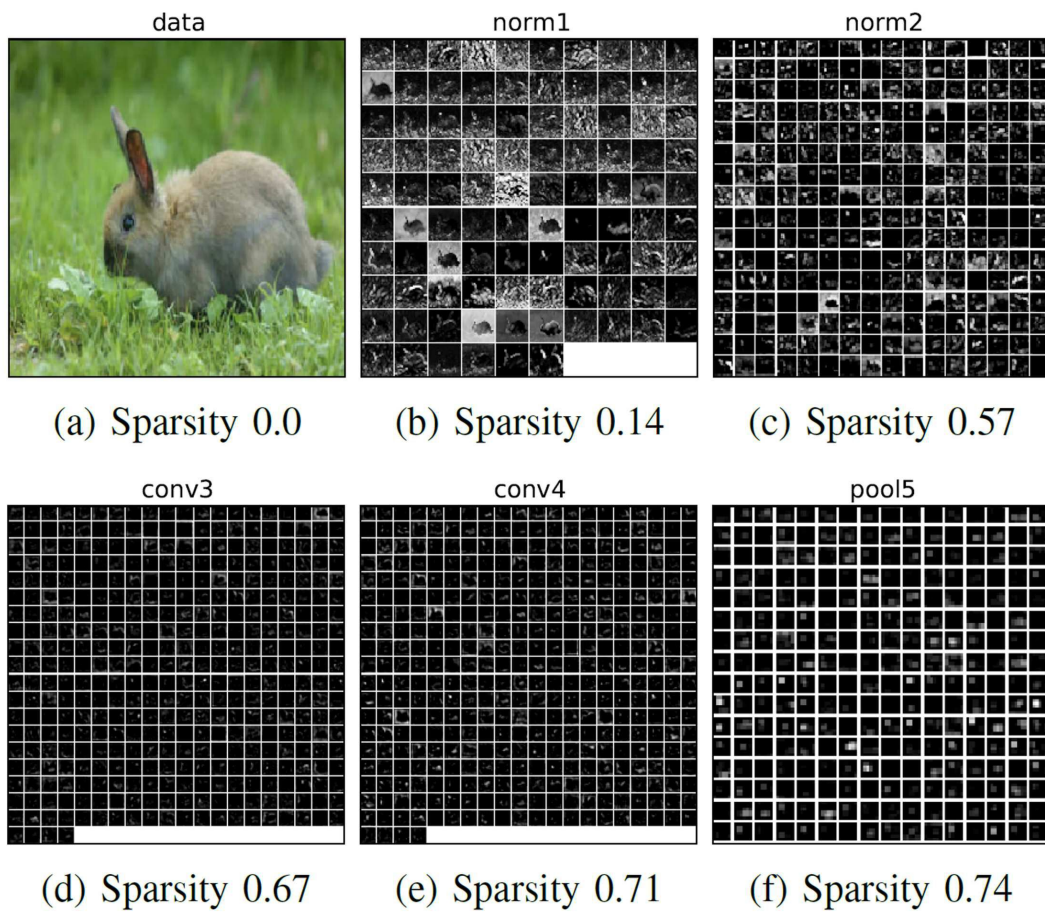
#### 도면1



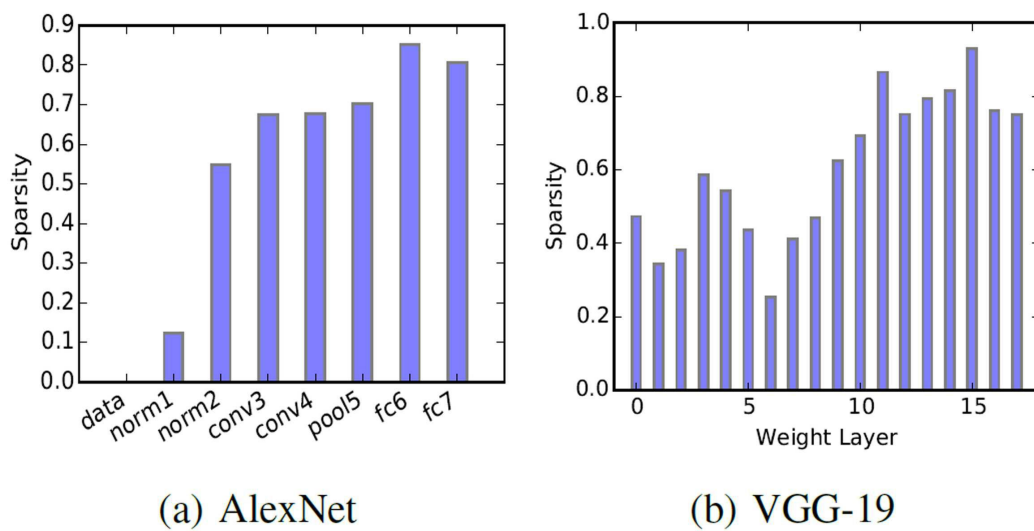
도면2



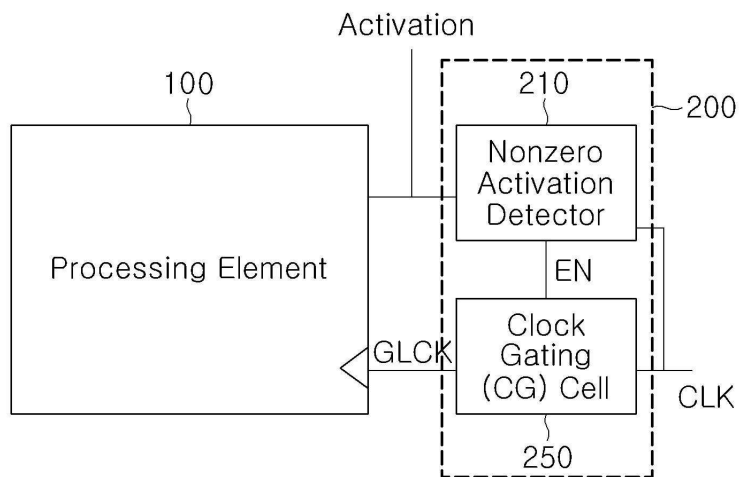
도면3



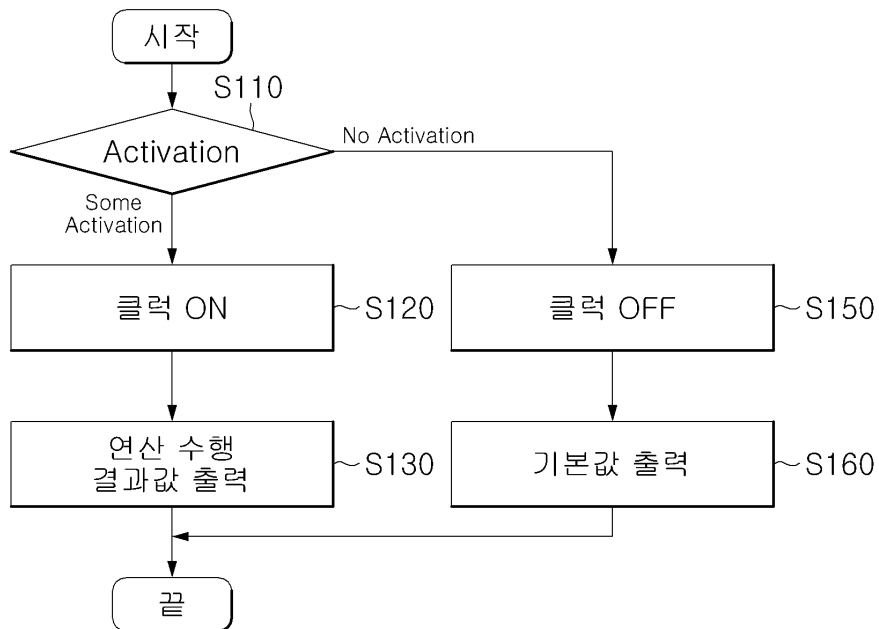
도면4



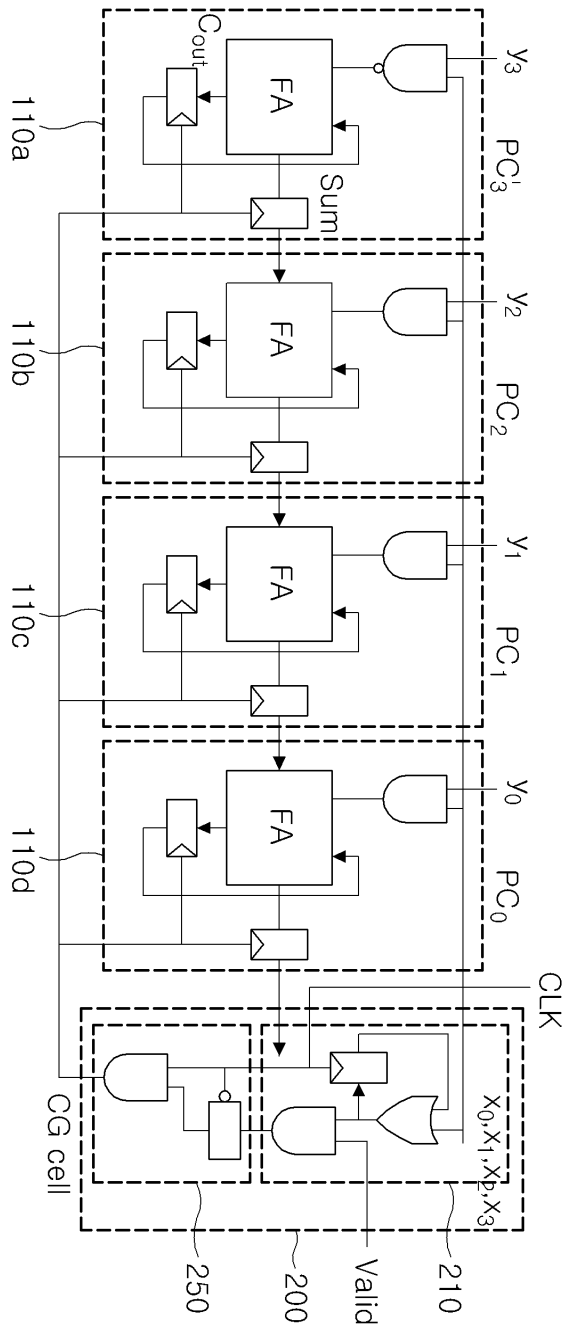
도면5



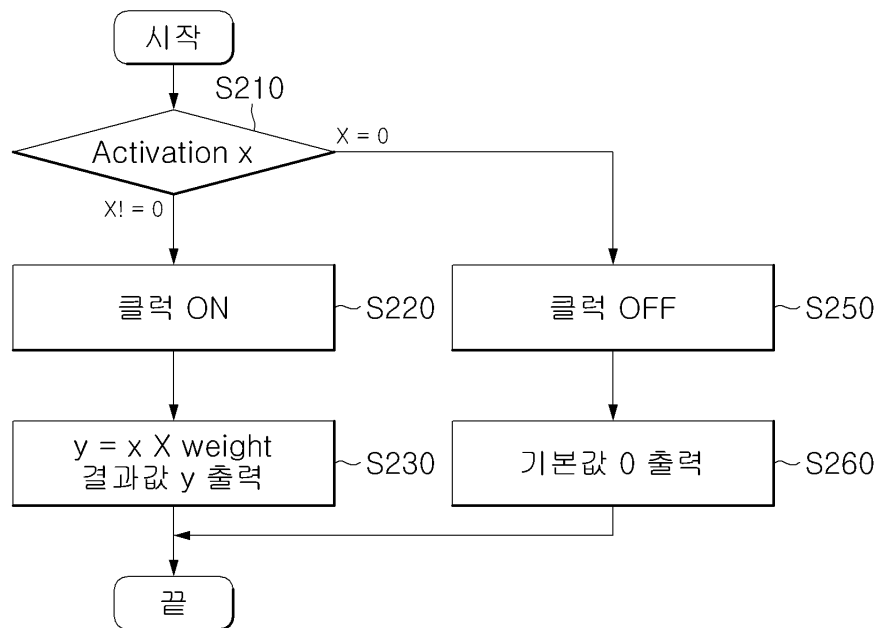
도면6



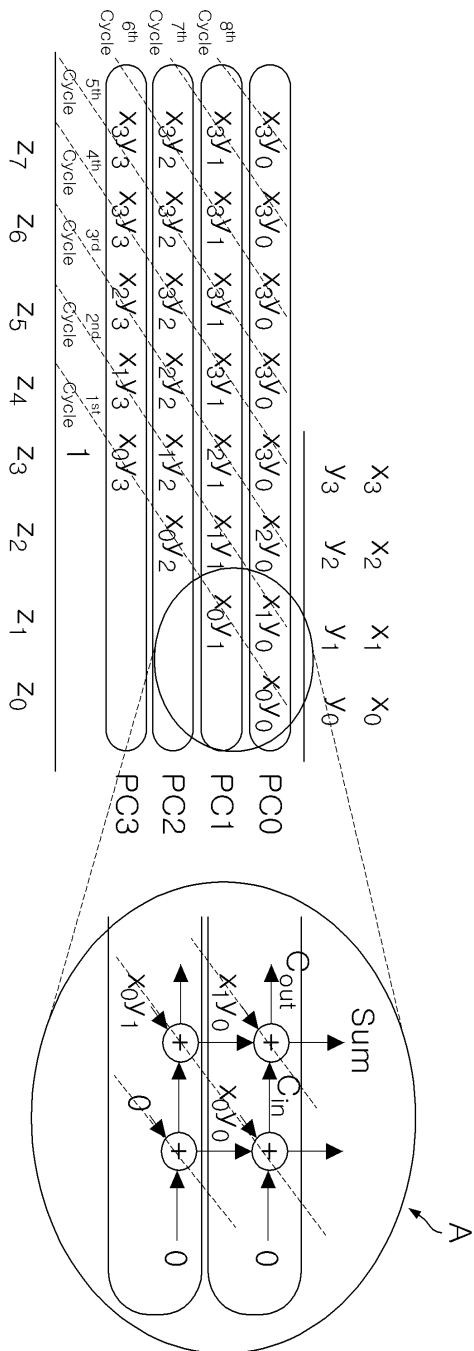
도면7



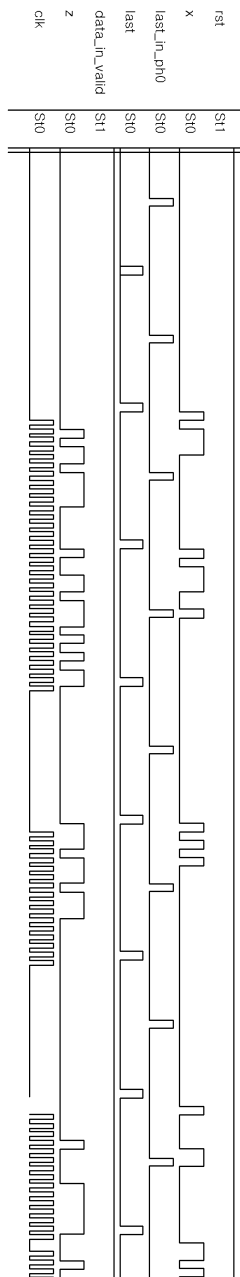
도면8



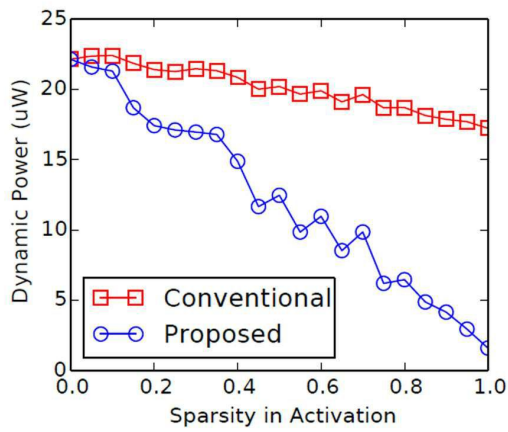
도면9



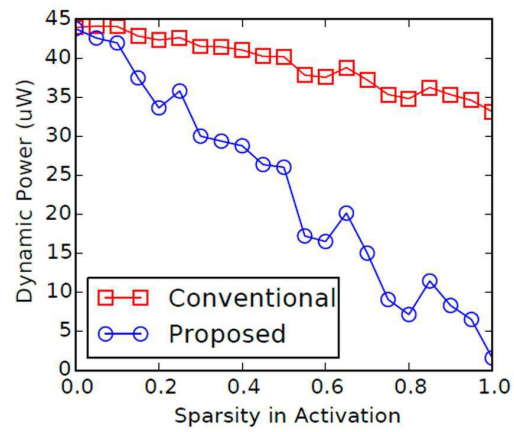
도면10



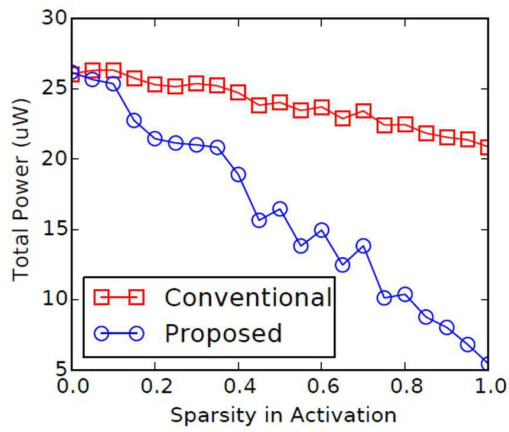
도면11



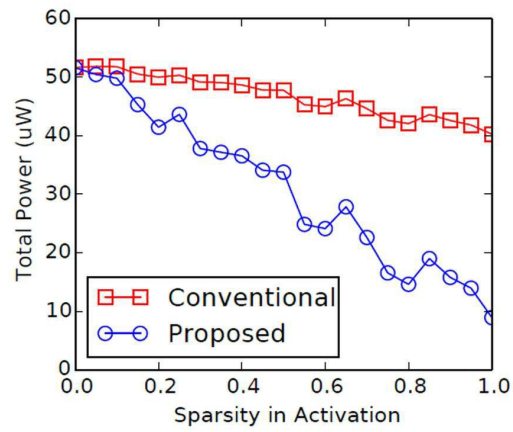
(a) 8-bit



(b) 16-bit



(c) 8-bit



(d) 16-bit