(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년01월15일
(11) 등록번호 10-2203746
(24) 등록일자 2021년01월11일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
(21) 출원번호 10-2018-7015434
(22) 출원일자(국제) 2016년04월01일
심사청구일자 2018년05월30일
(85) 번역문제출일자 2018년05월30일
(65) 공개번호 10-2018-0102059
(43) 공개일자 2018년09월14일
(86) 국제출원번호 PCT/CN2016/078281
(87) 국제공개번호 WO 2017/124642
국제공개일자 2017년07월27일
(30) 우선권주장
201610037645.1 2016년01월20일 중국(CN)

(56) 선행기술조사문헌
Chen, Tianshi, et al. A high-throughput neural network accelerator. IEEE Micro 35.3. 2015.*
Domingos, Pedro O., Fernando M. Silva, and Horácio C. Neto. An efficient and scalable architecture for neural networks with backpropagation learning. IEEE. 2005.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
캠브리콘 테크놀로지스 코퍼레이션 리미티드
중국 베이징 100191, 하이뎬 디스트릭트, 지춘 로드, 넘버 7, 즈쩐 빌딩, 블록 디, 16/에프, 룸 1601
(72) 발명자
리우, 샤오리
중국 피.알., 베이징 100190, 하이디엔 디스트릭트, 커쉐위안 사우스 로드, 넘버 6, 사이언티픽 리서치빌딩, 스위트 644
궈, 치
중국 피.알., 베이징 100190, 하이디엔 디스트릭트, 커쉐위안 사우스 로드, 넘버 6, 사이언티픽 리서치빌딩, 스위트 644
(뒷면에 계속)
(74) 대리인
이정현

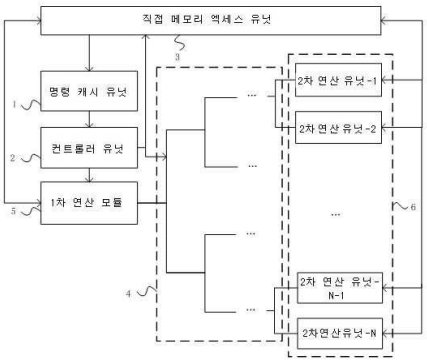
전체 청구항 수 : 총 12 항 심사관 : 서광훈

(54) 발명의 명칭 **인공 신경망 정방향 연산 실행용 장치와 방법**

(57) 요 약

본 발명은 인공 신경망 정방향 연산 실행용 장치에 관한 것으로서, 명령 캐시 유닛, 컨트롤러 유닛, 직접 메모리 엑세스 유닛, H 트리 모듈, 1차 연산 모듈, 및 복수개의 2차 연산 모듈을 포함한다. 상기 장치를 사용하여 다층 인공 신경망 정방향 연산을 구현할 수 있다. 각 한 층에 있어서, 먼저 입력 뉴런 벡터에 대하여 가중 합산을 진행하여 본 층의 중간 결과 벡터를 계산한다. 상기 중간 결과 벡터에 오프셋을 추가하고 활성화하여 출력 뉴런 벡터를 얻는다. 출력 뉴런 벡터는 다음 층의 입력 뉴런 벡터로 삼는다.

대 표 도 - 도1



(72) 발명자

천, 윈지

중국 피.알., 베이징 100190, 하이디엔 디스트릭트, 커쉐위안 사우스 로드, 넘버 6, 사이언티픽 리서치빌딩, 스위트 644

천, 티엔스

중국 피.알., 베이징 100190, 하이디엔 디스트릭트, 커쉐위안 사우스 로드, 넘버 6, 사이언티픽 리서치빌딩, 스위트 644

명 세 서

청구범위

청구항 1

H 트리 모듈을 통해 연결되는 복수개의 연산 모듈들; 및

명령을 한 개 이상의 마이크로 명령들로 디코딩하는 컨트롤러 유닛을 포함하고,

상기 복수개의 연산 모듈들은 상기 마이크로 명령들 중에서 상응하는 마이크로 명령들을 실행하도록 구비되고,

명령 캐시 유닛, 직접 메모리 엑세스 유닛, H 트리 모듈을 더 포함하되;

상기 명령 캐시 유닛은 상기 직접 메모리 엑세스 유닛을 통하여 명령을 판독하고 캐싱하는데 사용되고;

상기 직접 메모리 엑세스 유닛은 외부 주소 공간에서 제1 연산 모듈과 각 제2 연산 모듈에 상응하는 데이터 캐시 유닛 중 데이터를 작성하거나 상기 데이터 캐시 유닛에서 외부 주소 공간으로 데이터를 판독하는데 사용되고; 상기 제1 연산 모듈은 H 트리 모듈을 통해 모든 제2 연산 모듈로 본 층의 입력 뉴런 벡터를 전송하고, 제2 연산 모듈의 계산 과정이 종료된 후 H트리 모듈은 단계적으로 각 2차 연산 모듈의 출력 뉴런 값을 중간 결과 벡터로 합산하며; 상기 컨트롤러 유닛은 상기 명령 캐시 유닛에서 명령을 판독하고, 상기 명령을 H 트리 모듈, 제1연산 모듈, 및 제2 연산 모듈 행위를 제어하는 마이크로 명령으로 디코딩하는데 사용되도록 구비되며,

상기 제1 연산 모듈은, 연산 유닛, 데이터 종속성 판단 유닛 및 뉴런 캐시 유닛을 포함하고; 상기 뉴런 캐시 유닛은 제1 연산 모듈이 계산 과정에서 사용한 입력 데이터와 출력 데이터를 캐싱하는데 사용되고; 상기 연산 유닛은 상기 H 트리 모듈에서 유래하는 출력 기울기 벡터를 수신하고, 제1 연산 모듈의 각종 연산 기능을 완료하는데 사용되고; 상기 데이터 종속성 판단 유닛은 연산 유닛이 뉴런 캐시 유닛을 판독하여 기록하는 포트이며, 뉴런 캐시 유닛 중 데이터의 판독 기록에 일치성 충돌이 존재하지 않도록 보장하고, 뉴런 캐시 유닛에서 입력 뉴런 벡터를 판독하여 H 트리 모듈을 통하여 제2 연산 모듈에 발송하는 것을 책임지는데 사용하도록 구비되고,

각 제2 연산 모듈은 연산 유닛, 데이터 종속성 판단 유닛, 뉴런 캐시 유닛, 및 가중치 캐시 유닛을 포함하고; 상기 연산 유닛은 상기 컨트롤러 유닛에서 발송하는 마이크로 명령을 수신하고 산술 논리 연산을 진행하는데 사용되고; 상기 데이터 종속성 판단 유닛은 계산 과정 중에 있어서, 뉴런 캐시 유닛과 가중치 캐시 유닛의 판독 기록 조작을 책임지며, 뉴런 캐시 유닛과 가중치 캐시 유닛의 판독 기록에 일치성 충돌이 존재하지 않도록 보장하는데 사용되고; 상기 뉴런 캐시 유닛은 입력 뉴런 벡터 데이터 및 상기 제2 연산 모듈에서 계산하여 얻은 출력 뉴런 값을 캐싱하는데 사용되고; 및 상기 가중치 캐시 유닛은 상기 제2 연산 모듈이 계산 과정에서 필요하는 가중치 벡터를 캐싱하는데 사용되는 것을 특징으로 하는 다층 인공 신경망 정방향 연산 실행용 장치.

청구항 2

제1항에 있어서, 상기 복수개의 연산 모듈들은,

상기 H 트리 모듈을 통해 수신되는 입력 뉴런 벡터에 기초하여 복수개의 출력 뉴런 값들을 병렬적으로 계산하는 복수개의 제2 연산 모듈들; 및

상기 복수개의 제2 연산 모듈들에 의해 계산되는 상기 복수개의 출력 뉴런 값들에 기초하여 상기 H 트리 모듈에 의해 병합되는 한 개 이상의 중간 결과 벡터들에 기초하여 최종 출력 뉴런 벡터를 생성하는 제1 연산 모듈을 포함하는 다층 인공 신경망 정방향 연산 실행용 장치.

청구항 3

제2항에 있어서, 상기 제1 연산 모듈은,

상기 중간 결과 벡터에 대한 오프셋 추가 연산;

활성 함수를 사용한 상기 중간 결과 벡터에 대한 활성화 연산; 및

상기 중간 결과 벡터에 대한 풀링(pooling) 연산;

을 포함하는 그룹으로부터 선택된 연산을 수행하는 다층 인공 신경망 정방향 연산 실행용 장치.

청구항 4

삭제

청구항 5

제2항에 있어서, 상기 H 트리 모듈은 각각의 단계가 복수개의 노드들을 갖는 복수개의 단계들을 포함하는 2진 트리형의 구조를 갖고,

각 단계의 상기 노드들 각각은 다운스트림의 두 개의 노드들과 연결되고,

상기 노드들 각각은 동일한 데이터를 상기 다운스트림의 두 개의 노드들에 전송하고, 상기 다운스트림의 두 개의 노드들로부터 수신되는 데이터를 병합하는 다층 인공 신경망 정방향 연산 실행용 장치.

청구항 6

삭제

청구항 7

삭제

청구항 8

제1항에 있어서, 상기 데이터 종속성 판단 유닛은,

수신되는 마이크로 명령들을 내부의 명령 큐에 저장하고,

상기 명령 큐에 저장된 판독 명령에 따른 데이터의 범위가 상기 명령 큐에서 상기 판독 명령보다 선행하는 기록 명령에 따른 데이터의 범위와 충돌하는 경우, 상기 판독 명령은 상기 기록 명령이 실행된 이후에 실행되도록 상기 뉴런 캐시 유닛에 대한 판독 및 기록 동작을 제어하는 다층 인공 신경망 정방향 연산 실행용 장치.

청구항 9

제1항에 있어서, 상기 명령은,

각층의 인공 신경망의 계산 시작 전에 현재 층의 계산에 필요한 상수들을 배치하기 위한 CONFIG 명령,

각층의 인공 신경망의 산술 논리 연산을 완료하기 위한 COMPUTE 명령, 및

외부 주소 공간으로부터 계산에 필요한 입력 데이터를 판독하여 입력하고, 계산 완료 후 데이터를 상기 외부 주소 공간에 저장하기 위한 IO 명령을 포함하는 그룹으로부터 선택되는 다층 인공 신경망 정방향 연산 실행용 장치.

청구항 10

삭제

청구항 11

컨트롤러 유닛이 명령을 수신하는 단계;

상기 컨트롤러 유닛이 상기 명령을 한 개 이상의 마이크로 명령들로 디코딩하는 단계; 및

상기 컨트롤러 유닛이 상기 한 개 이상의 마이크로 명령들을 복수개의 연산 모듈들에 각각 할당하는 단계를 포함하고,

명령 캐시 유닛이 직접 메모리 엑세스 유닛을 통하여 명령을 판독하고 캐싱하는데 사용하는 단계;

직접 메모리 엑세스 유닛이 외부 주소 공간에서 제1 연산 모듈과 각 제2 연산 모듈에 상응하는 데이터 캐시 유닛에 데이터를 작성하거나 상기 데이터 캐시 유닛에서 외부 주소 공간으로 작성된 데이터를 판독하는데 사용하

는 단계;

복수개의 연산 모듈 중 제1 연산 모듈이 H 트리 모듈을 통해 모든 제2 연산 모듈로 본 층의 입력 뉴런 벡터를 전송하고, 제2 연산 모듈의 계산 과정이 종료된 후 H 트리 모듈이 단계적으로 각 2차 연산 모듈의 출력 뉴런 값을 중간 결과 벡터로 합산하는 단계;

상기 컨트롤러 유닛이 상기 명령 캐시 유닛에서 명령을 판독하고, 상기 컨트롤러 유닛이 상기 명령을 H 트리 모듈, 제1연산 모듈, 및 제2 연산 모듈의 행위를 제어하는 마이크로 명령으로 디코딩하는데 사용하는 단계를 포함하며,

상기 제1 연산 모듈의 뉴런 캐시 유닛이 제1 연산 모듈의 계산 과정에서 사용한 입력 데이터와 출력 데이터를 캐싱하는데 사용하는 단계; 상기 제1 연산 모듈의 연산 유닛이 상기 H 트리 모듈에서 유래하는 출력 기울기 벡터를 수신하고 각종 연산 기능을 완료하는데 사용하는 단계; 상기 제1 연산 모듈의 데이터 종속성 판단 유닛이 뉴런 캐시 유닛을 판독하여 기록하는 포트이며, 상기 제1 연산 모듈의 데이터 종속성 판단 유닛이 뉴런 캐시 유닛 중 데이터의 판독 기록에 일치성 충돌이 존재하지 않도록 보장하는데 사용하고, 상기 제1 연산 모듈의 데이터 종속성 판단 유닛이 뉴런 캐시 유닛에서 입력 뉴런 벡터를 판독하여 H 트리 모듈을 통하여 제2 연산 모듈에 발송하는 것을 책임지는데 사용하는 단계를 포함하며,

각 제2 연산 모듈의 연산 유닛이 상기 컨트롤러 유닛에서 발송된 마이크로 명령을 수신하고, 각 제2 연산 모듈의 연산 유닛이 마이크로 명령에 따라 산술 논리 연산을 진행하는데 사용하는 단계; 상기 데이터 종속성 판단 유닛이 계산 과정 중 뉴런 캐시 유닛과 가중치 캐시 유닛의 판독 기록 조작을 책임지는데 사용되고, 상기 데이터 종속성 판단 유닛이 뉴런 캐시 유닛과 가중치 캐시 유닛의 판독 기록에 일치성 충돌이 존재하지 않도록 보장하는데 사용하는 단계; 상기 뉴런 캐시 유닛이 입력 뉴런 벡터 데이터 및 상기 제2 연산 모듈에서 계산하여 얻은 출력 뉴런 값을 캐싱하는데 사용하는 단계; 및 상기 가중치 캐시 유닛이 상기 제2 연산 모듈이 계산 과정에서 필요하는 가중치 벡터를 캐싱하는데 사용하는 단계를 포함하는 것을 특징으로 하는 다층 인공 신경망 정방향 연산을 실행하는 방법.

청구항 12

제11항에 있어서,

상기 복수개의 연산 모듈들 중의 복수개의 제2 연산 모듈들이 H 트리 모듈을 통해 수신되는 입력 뉴런 벡터에 기초하여 복수개의 출력 뉴런 값들을 병렬적으로 계산하는 단계;

상기 H 트리 모듈이 상기 복수개의 출력 뉴런 값들을 병합하여 중간 결과 벡터를 생성하는 단계; 및

상기 복수개의 연산 모듈들 중의 제1 연산 모듈이 상기 중간 결과 벡터에 기초하여 최종 출력 뉴런 벡터를 생성하는 단계를 더 포함하는 다층 인공 신경망 정방향 연산을 실행하는 방법.

청구항 13

제12항에 있어서,

상기 제1 연산 모듈이,

상기 중간 결과 벡터에 대한 오프셋 추가 연산;

활성 함수를 사용한 상기 중간 결과 벡터에 대한 활성화 연산; 및

상기 중간 결과 벡터에 대한 풀링(pooling) 연산;

을 포함하는 그룹으로부터 선택된 연산을 수행하는 단계를 더 포함하는 다층 인공 신경망 정방향 연산을 실행하는 방법.

청구항 14

삭제

청구항 15

제12항에 있어서, 상기 H 트리 모듈은 각각의 단계가 복수개의 노드들을 갖는 복수개의 단계들을 포함하는 2진

트리형의 구조를 갖고,

각 단계의 상기 노드들 각각은 다운스트림의 두 개의 노드들과 연결되고,

상기 노드들 각각은 동일한 데이터를 상기 다운스트림의 두 개의 노드들에 전송하고, 상기 다운스트림의 두 개의 노드들로부터 수신되는 데이터를 병합하는 다층 인공 신경망 정방향 연산을 실행하는 방법.

청구항 16

삭제

청구항 17

삭제

청구항 18

제11항에 있어서,

상기 데이터 종속성 판단 유닛이 수신되는 마이크로 명령들을 내부의 명령 큐에 저장하는 단계; 및

상기 데이터 종속성 판단 유닛이 상기 명령 큐에 저장된 판독 명령에 따른 데이터의 범위가 상기 명령 큐에서 상기 판독 명령보다 선행하는 기록 명령에 따른 데이터의 범위와 충돌하는 경우, 상기 판독 명령은 상기 기록 명령이 실행된 이후에 실행되도록 상기 뉴런 캐시 유닛에 대한 판독 및 기록 동작을 제어하는 단계를 더 포함하는 다층 인공 신경망 정방향 연산을 실행하는 방법.

청구항 19

제11항에 있어서, 상기 명령은,

각층의 인공 신경망의 계산 시작 전에 현재 층의 계산에 필요한 상수들을 배치하기 위한 CONFIG 명령,

각층의 인공 신경망의 산술 논리 연산을 완료하기 위한 COMPUTE 명령, 및

외부 주소 공간으로부터 계산에 필요한 입력 데이터를 판독하여 입력하고, 계산 완료 후 데이터를 상기 외부 주소 공간에 저장하기 위한 IO 명령을 포함하는 그룹으로부터 선택되는 다층 인공 신경망 정방향 연산을 실행하는 방법.

청구항 20

삭제

청구항 21

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 인공 신경망 분야에 관한 것으로서, 더욱 상세하게는 인공 신경망 정방향 트레이닝 실행용 장치와 방법에 관한 것이다.

배 경 기 술

[0002] 다층 인공 신경망은 패턴 인식, 이미지 처리, 함수 근사, 최적화 계산 등 분야에서 광범위하게 응용되며, 다층 인공망은 최근 몇 년 동안 비교적 높은 식별 정확도와 비교적 우수한 병렬 가능성으로 인해 학술계와 산업계에서 갈수록 많은 주목을 받고 있다.

[0003] 다층 인공 신경망 정방향 트레이닝을 지원하는 종래의 방법은 범용 프로세서를 사용하는 것이다. 상기 방법은 범용 레지스터 파일과 범용 기능 부품을 통하여 범용 명령을 실행하여 상기 알고리즘을 지원한다. 상기 방법의 단점 중 하나는 단일 범용 프로세서의 연산 성능이 비교적 낮아 통상적인 다층 인공 신경망 연상의 성능 수요를

충족시킬 수 없다는 것이다. 복수 개의 범용 프로세서를 병렬 실행할 경우, 범용 프로세서 간 상호 통신도 성능 병목 현상을 만든다. 또한 범용 프로세서는 다층 인공 신경망 정방향 연산을 긴 열 연산 및 메모리 엑세스 명령 시퀀스로 디코딩하기 때문에 프로세스 전단 디코딩에 비교적 큰 전력이 소모된다.

[0004] 다층 인공 신경망 정방향 트레이닝을 지원하는 또 다른 종래의 방법은 그래픽 처리 유닛(GPU)을 사용하는 것이다. 상기 방법은 범용 레지스터 파일과 범용 스트림 프로세서를 통하여 범용 SIMD 명령을 실행함으로써 상기 알고리즘을 지원한다. GPU는 그래픽 이미지 연산 및 과학 계산을 전문적으로 실행하는 설비이기 때문에, 다층 인공 신경망 연산을 전문적으로 지원할 수 없어 여전히 대량의 전단 디코딩 작업이 있어야 다층 인공 신경망 연산을 실행할 수 있으므로 대량의 추가적 비용이 든다. 또한 GPU는 비교적 작은 온칩(on-chip) 캐시만 있기 때문에 다층 인공 신경망의 모델 데이터(가중치)를 반복적으로 칩 외부에서 운반해야 하므로 오프칩(off-chip) 대역폭이 주요 성능의 병목 현상을 일으키며, 동시에 엄청난 전력이 소모된다.

발명의 내용

[0005] 본 발명은 한편으로 인공 신경망 정방향 트레이닝 실행용 장치에 관한 것으로서, 명령 캐시 유닛, 컨트롤러 유닛, 직접 메모리 엑세스 유닛, H 트리 모듈, 1차 연산 모듈, 및 복수 개의 2차 연산 모듈을 포함한다. 여기에서, 상기 명령 캐시 유닛은 상기 직접 메모리 엑세스 유닛을 통하여 명령을 판독하고 캐싱하는 데 사용된다. 상기 컨트롤러 유닛은 명령 캐시 유닛에서 명령을 판독하고, 상기 명령은 H 트리 모듈, 1차 연산 모듈, 및 2차 연산 모듈 동작을 제어하는 마이크로 명령으로 디코딩하는 데 사용된다. 직접 메모리 엑세스 유닛은 외부 주소 공간에서 1차 연산 모듈과 각 2차 연산 모듈에 상응하는 데이터 캐시 유닛 중 데이터를 작성하거나 상기 데이터 캐시 유닛에서 외부 주소 공간으로 데이터를 판독하는 데 사용된다. H 트리 모듈은 각층 신경망 정방향 트레이닝 계산 시작의 단계에서 1차 연산 모듈이 H 트리 모듈을 통해 모든 2차 연산 모듈로 본 층의 입력 뉴런 벡터를 전송하는 데에 사용되고, 2차 연산 모듈의 계산 과정이 종료된 후 H 트리 모듈은 단계적으로 각 2차 연산 모듈의 출력 뉴런 값을 중간 결과 벡터로 합산하며; 1차 연산 모듈은 중간 결과 벡터를 이용하여 후속 계산을 완성한다.

[0006] 본 발명은 다른 한편으로 상기 장치를 사용하여 단층 인공 신경망 정방향 트레이닝을 실행하는 방법을 제공한다.

[0007] 본 발명은 또 다른 한편으로 상기 장치를 사용하여 다층 인공 신경망 정방향 트레이닝을 실행하는 방법을 제공한다.

도면의 간단한 설명

[0008] 이하에서는, 본 발명의 도면을 통해 발명 및 본 발명의 장점을 더욱 상세하게 설명한다.

도 1은 본 발명 실시예에 따른 인공 신경망 정방향 트레이닝 실행용 장치의 전체 구조에 대한 블록 다이어그램이고;

도 2는 본 발명 실시예에 따른 인공 신경망 정방향 트레이닝 실행용 장치 중 H 트리 모듈의 구조이고;

도 3은 본 발명 실시예에 따른 인공 신경망 정방향 트레이닝 실행용 장치 중 1차 연산 모듈 구조의 블록 다이어그램이고;

도 4는 본 발명 실시예에 따른 인공 신경망 정방향 트레이닝 실행용 장치 중 2차 연산 모듈 구조의 블록 다이어그램이고;

도 5는 본 발명 실시예에 따른 신경망 정방향 트레이닝 과정의 블록 다이어그램이고; 및

도 6은 본 발명 실시예에 따른 단층 인공 신경망 연산의 흐름도이다.

상기 도면에서 동일한 장치, 부품, 유닛 등은 동일한 부호로 표시하였다.

발명을 실시하기 위한 구체적인 내용

[0009] 이하에서는, 본 발명의 예시적인 실시형태들을 도면을 통해 보다 상세히 설명한다. 본 발명의 기타 측면, 장점 및 특징은 본 발명이 속한 기술분야의 당업자가 쉽게 이해할 수 있다.

[0010] 본 발명에 있어서, 전문용어 “포함”과 “함유” 및 그 파생어의 뜻은 포괄하며 제한적이지 않다는 것이고, 전

문용어 “또는”은 “및/또는”의 뜻으로서 포함성을 가진다.

[0011] 본 발명의 설명에 있어서, 이하의 내용은 본 발명 원리의 각종 실시예를 설명하기 위한 것에 불과하므로 어떠한 방식으로든 본 발명의 보호범위를 제한하지 않는다. 이하의 첨부 도면을 참조하여 특허청구범위 및 그와 동등한 물건으로 한정되는 본 발명의 예시적 실시예를 전면적으로 이해하도록 한다. 이하의 설명에는 이해를 돕기 위하여 다양한 구체적인 세부사항을 포함하나, 이러한 세부사항은 예시적인 것에 불과하다. 따라서 본 발명이 속한 기술분야의 당업자는 본 발명의 범위와 정신에서 위배되지 않는 상황에서 본 발명에서 설명한 실시예에 대하여 변경과 수식을 진행할 수 있다는 것을 이해하여야 한다. 그 외, 명확함과 간결함을 위하여 공지된 기능과 구조에 대한 설명은 생략하였다. 또한 첨부 도면에서 일관되게 동일한 참고 숫자는 유사한 기능과 조작에 사용하였다.

[0012] 본 발명 실시예에 따른 다층 인공 신경망의 정방향 연산은 2층 이상의 복수 개 뉴런을 포함한다. 각 층에 있어서, 입력 뉴런 벡터는 먼저 가중치 벡터와 내적 연산(dot product operation)을 진행하고, 결과는 활성 함수를 거쳐 출력 뉴런을 얻는다. 여기에서, 활성 함수는 sigmoid 함수, tanh, relu, softmax 함수 등일 수 있다.

[0013] 도 1은 본 발명 실시예에 따른 인공 신경망 정방향 트레이닝 실행용 장치의 전체 구조에 대한 블록 다이어그램을 도시한 것이다. 도 1에서 도시하는 바와 같이, 상기 장치는 명령 캐시 유닛(1), 컨트롤러 유닛(2), 직접 메모리 엑세스 유닛(3), H 트리 모듈(4), 1차 연산 모듈(5), 및 복수 개의 2차 연산 모듈(6)을 포함한다. 명령 캐시 유닛(1), 컨트롤러 유닛(2), 직접 메모리 엑세스 유닛(3), H 트리 모듈(4), 1차 연산 모듈(5), 및 2차 연산 모듈(6)은 모두 하드웨어 회로(예를 들어 전용 집적회로 ASIC)를 통하여 구현할 수 있다.

[0014] 명령 캐시 유닛(1)은 직접 메모리 엑세스 유닛(3)을 통하여 명령을 판독하여 입력하고, 판독하여 입력한 명령을 캐싱하는 데 사용된다.

[0015] 컨트롤러 유닛(2)은 명령 캐시 유닛(1)에서 명령을 판독하고, 명령을 기타 모듈 동작을 제어하는 마이크로 명령으로 디코딩하고, 상기 기타 모듈에는 예를 들어 직접 메모리 엑세스 유닛(3), 1차 연산 모듈(5), 및 2차 연산 모듈(6) 등이 있다.

[0016] 직접 메모리 엑세스 유닛(3)은 외부 주소 공간을 메모리 엑세스하고 장치 내부의 각 캐시 유닛에 직접 데이터를 판독하여 기록함으로써 데이터의 로딩과 저장을 완료할 수 있다.

[0017] 도 2는 H 트리 모듈(4)의 구조를 도시한 것이다. H 트리 모듈(4)은 1차 연산 모듈(5)과 복수 개의 2차 연산 모듈(6) 사이의 데이터 통로를 구성하고, H 트리형의 구조를 가진다. H 트리는 복수 개의 노드로 구성된 2진 트리 통로이고, 각 노드는 업스트림의 데이터를 마찬가지로 다운스트림의 2개 노드로 발급하고, 다운스트림의 2개 노드에서 반환하는 데이터를 병합하여 업스트림의 노드에 반환한다. 예를 들어 각 층의 인공 신경망 계산 시작 단계에서 1차 연산 모듈(5) 내의 뉴런 데이터는 H 트리 모듈(4)을 통하여 각 2차 연산 모듈(6)에 발송한다. 2차 연산 모듈(6)의 계산 과정이 완료된 후, 각 2차 연산 모듈이 출력하는 뉴런의 값은 H 트리에서 단계적으로 뉴런으로 구성되는 하나의 완전한 벡터로 합쳐져 중간 결과 벡터로 삼을 수 있다. 신경망 완전 연결층(MLP)으로 설명할 경우, 장치 중에 총 N개의 2차 연산 모듈이 있다고 가정하면 중간 결과 벡터는 N 구간으로 나누고, 각 구간에는 N개 원소가 있고, 제i번째 2차 연산 모듈은 각 구간 중의 제i번째 원소를 계산한다. N개 원소는 H 트리 모듈을 거쳐 길이가 N인 벡터로 합쳐지고 1차 연산 모듈로 반환된다. 따라서 만약 망에 N개 출력 뉴런밖에 없다면 각 2차 연산 모듈은 단일 뉴런의 값만 출력하면 된다. 만약 망에 m*N개 출력 뉴런이 있다면 각 2차 연산 모듈은 m개 뉴런 값을 출력해야 한다.

[0018] 도 3은 본 발명 실시예에 따른 인공 신경망 정방향 연산 실행용 장치 중 1차 연산 모듈(5) 구조의 블록 다이어그램이다. 도 3에서 도시하는 바와 같이, 1차 연산 모듈(5)은 연산 유닛(51), 데이터 종속성 판단 유닛(52) 및 뉴런 캐시 유닛(53)을 포함한다.

[0019] 뉴런 캐시 유닛(53)은 1차 연산 모듈(5)이 계산 과정에서 사용한 입력 데이터와 출력 데이터를 캐싱하는 데 사용된다. 연산 유닛(51)은 1차 연산 모듈의 각종 연산 기능을 완료한다. 데이터 종속성 판단 유닛(52)은 연산 유닛(51)이 뉴런 캐시 유닛(53)을 판독하여 기록하는 포트이며, 동시에 뉴런 캐시 유닛 중 데이터의 판독 기록에 일치성을 보장할 수 있다. 또한, 데이터 종속성 판단 유닛(52)은 판독한 데이터를 H 트리 모듈(4)을 통해 계산 모듈로 발송하는 것도 책임지며, 2차 연산 모듈(6)의 출력 데이터는 H 트리 모듈(4)을 통해 곧바로 연산 유닛(51)으로 이송된다. 컨트롤러 유닛(2)이 출력하는 명령은 계산 유닛(51)과 데이터 종속성 판단 유닛(52)에 발송하여 그 동작을 제어한다.

[0020] 도 4는 본 발명 실시예에 따른 인공 신경망 정방향 연산 실행용 장치 중 2차 연산 모듈(6) 구조의 블록 다이어

그램이다. 도 4에서 도시하는 바와 같이, 각 2차 연산 모듈(6)은 연산 유닛(61), 데이터 종속성 판단 유닛(62), 뉴런 캐시 유닛(63), 및 가중치 캐시 유닛(64)을 포함한다.

[0021] 연산 유닛(61)은 컨트롤러 유닛(2)에서 발송하는 마이크로 명령을 수신하고 산술 논리 연산을 진행한다.

[0022] 데이터 종속성 판단 유닛(62)은 계산 과정 중 뉴런 캐시 유닛의 판독 및 기록 조작을 책임진다. 데이터 종속성 판단 유닛(62)은 판독 기록 조작 실행 전에 먼저 명령 간에 사용하는 데이터에 판독 기록의 일치성 충돌이 존재하지 않도록 보장한다. 예를 들어, 데이터 종속성 판단 유닛(62)으로 발송되는 모든 마이크로 명령은 데이터 종속성 판단 유닛(62) 내부의 명령 큐(instruction queue) 내에 저장될 수 있고, 상기 큐에 있어서, 판독 명령 판독 데이터의 범위는 만약 큐 위치에서 앞쪽에 가까운 기록 명령 기록 데이터의 범위와 충돌이 일어나는 경우, 상기 명령은 반드시 종속되는 기록 명령이 실행된 후에만 실행될 수 있다.

[0023] 뉴런 캐시 유닛(63)은 상기 2차 연산 모듈(6)의 입력 뉴런 벡터 데이터와 출력 뉴런 값 데이터를 캐싱한다.

[0024] 가중치 캐시 유닛(64)은 상기 2차 연산 모듈(6)이 계산 과정에서 필요한 가중치 데이터를 캐싱한다. 각 2차 연산 모듈(6)은 전체 입력 뉴런과 부분 출력 뉴런 사이의 가중치만 저장할 수 있다. 완전 연결층을 예를 들면, 출력 뉴런은 2차 연산 모듈의 개수 N에 따라 구간을 나누며, 각 구간의 제n번째 출력 뉴런에 대응하는 가중치는 제n번째 2차 연산 모듈 내에 저장한다.

[0025] 2차 연산 모듈(6)은 각 층 인공 신경망 정방향 연산 과정 중 병렬 가능한 전반 부분을 구현한다. 인공 신경망의 완전 연결층(MLP)을 예를 들면, 과정은 y=f(wx+b)이고, 여기에서 가중치 행렬 w와 입력 뉴럭 벡터 x의 곱셈은 상관없는 병렬 연산 서브태스크로 나눌 수 있으며, out과 in은 열 벡터이고, 각 2차 연산 모듈(6)은 in 중 상응하는 부분 스칼라(scalar) 원소와 가중치 행렬 w에 대응하는 열의 곱만 계산하고, 수득한 각 출력 벡터는 모두 최종 결과의 하나의 누적 대기 부분합이고, 이러한 부분합은 H 트리 모듈(4)에서 단계적으로 2개씩 서로 더하여 최후의 결과를 얻는다. 상기 계산 과정은 부분합을 병렬 연산하는 과정과 후속의 누적하는 과정으로 된다. 각 2차 연산 모듈(6)은 출력 뉴런 값을 계산하고, 모든 출력 뉴런 값은 H 트리 모듈(4)에서 합하여 중간 결과 벡터를 얻는다. 각 2차 연산 모듈(6)은 중간 결과 벡터 y 중의 본 모듈에 대응하는 출력 뉴런 값만 계산하면 된다. H 트리 모듈(4)은 모든 2차 연산 모듈(6)에 대하여 출력하는 뉴런 값을 더하여 최종적인 중간 결과 벡터 y를 얻는다. 1차 연산 모듈(5)은 중간 결과 벡터 y를 기반으로 후속 계산을 진행하는데, 예를 들어 오프셋 추가, 풀링(예를 들어 최대값 풀링(MAXPOOLING) 또는 평균값 풀링(AVGPOOLING 등), 활성화 및 샘플링 등이 있다.

[0026] 본 발명 실시예에 의거하여 전술한 장치에서 인공 신경망 정방향 연산을 실행하는 명령 집합을 더 제공한다. 명령 집합 내에는 CONFIG 명령, COMPUTE 명령, IO 명령, NOP 명령, JUMP 명령, 및 MOVE 명령이 포함된다.

[0027] 여기에서, CONFIG 명령은 각층 인공 신경망 계산 시작 전에 현재 층 계산에 필요한 각종 상수를 배치한다.

[0028] COMPUTE 명령은 각층 인공 신경망의 산술 논리 연산을 완료한다.

[0029] IO 명령은 외부 주소 공간에서 계산에 필요한 입력 데이터를 판독하여 입력하고, 및 계산 완료 후 데이터를 외부 공간에 저장한다.

[0030] NOP 명령은 현재 내부 모든 마이크로 명령 캐시 큐(queue) 내에 담긴 마이크로 명령을 정리하여 NOP 명령 전의 모든 명령이 전부 완료되도록 보장한다. NOP 명령 자체에는 어떠한 조작도 포함되지 않는다.

[0031] JUMP 명령은 컨트롤러가 명령 캐시 유닛에서 판독한 다음 명령 주소로 건너뛰는 것을 책임져 제어 흐름의 점프를 구현하는 데 사용된다.

[0032] MOVE 명령은 장치 내부 주소 공간의 특정 주소 데이터를 장치 내부 주소 공간의 다른 주소로 옮기는 것을 책임지며, 상기 과정은 연산 유닛에 독립적이고 실행 과정 중 연산 유닛의 자원을 점용하지 않는다.

[0033] 도 5는 본 발명 실시예에 따른 신경망 정방향 연산 과정의 블록 다이어그램이다. 다른 2차 연산 모듈(6)에서 입력 뉴런 벡터는 각각 상기 2차 연산 모듈(6)의 가중치 벡터와 내적 연산을 진행하여 대응하는 출력 뉴런 값을 얻고, 모든 이러한 출력 뉴런 값은 중간 결과 벡터를 구성하며, 상기 중간 경과 벡터는 오프셋 추가 벡터 및 활성 연산을 거쳐 상기 층 신경망의 최종 출력 뉴런 벡터를 얻는다. 공식은 out=f(w*in+b)이고, 여기에서 out은 출력 뉴런 벡터이고, in은 입력 뉴런 벡터이며, b는 오프셋 벡터이고, w는 가중치 행렬이며, f는 활성 함수이다. 각 2차 연산 모듈(6)의 가중치 벡터는 가중치 행렬 중의 상기 2차 연산 모듈(6)에 대응하는 열 벡터이다. H 트리 모듈은 입력 뉴런 벡터[in0,…,inN]를 모든 2차 연산 모듈에 발송하고, 뉴런 캐시 유닛 내에 일시 저장한다. 제i번째 2차 연산 모듈에 있어서, 그 상응하는 가중치 벡터[w_i0,…,w_iN]와 입력 뉴런 벡터의 내적

을 계산한다. 2차 연산 모듈이 출력하는 결과는 H 트리 모듈을 거쳐 완전한 출력 벡터로 합쳐져 1차 연산 유닛으로 반환하고, 1차 연산 유닛에서 활성 연산을 진행하여 마지막의 출력 뉴런 벡터[out0,out1,out2,…,outN]를 얻는다.

[0034] 도 6은 본 발명 실시예에 따른 단층 인공 신경망 정방향 연산 흐름도를 도시한 것이다. 상기 흐름도는 본 발명의 장치와 명령 집합을 이용하여 도 4에서 도시하는 단층 신경망 정방향 연산을 구현하는 과정에 관한 것이다.

[0035] 단계 S1: 명령 캐시 유닛(1)의 첫 주소에 하나의 IO 명령을 사전 저장한다.

[0036] 단계 S2: 연산을 시작하며, 컨트롤러 유닛(2)은 명령 캐시 유닛(1)의 첫 주소에서 상기 IO 명령을 판독하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 엑세스 유닛(3)은 외부 주소 공간에서 상응하는 모든 인공 신경망 연산 명령을 판독하며, 이를 명령 캐시 유닛(1) 내에 캐싱한다.

[0037] 단계 S3: 컨트롤러 유닛(2)은 이어서 명령 캐시 유닛에서 다음 IO 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 엑세스 유닛(3)이 외부 주소 공간에서 1차 연산 모듈(5)이 필요하는 모든 데이터(예를 들어 입력 뉴런 벡터, 보간표, 상수표, 오프셋 등을 포함)를 1차 연산 모듈(5)의 뉴런 캐시 유닛(53)까지 판독한다.

[0038] 단계 S4: 컨트롤러 유닛(2)은 이어서 명령 캐시 유닛에서 다음 IO 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 엑세스 유닛(3)이 외부 주소 공간에서 2차 연산 모듈(6)에 필요하는 가중치 행렬 데이터를 판독한다.

[0039] 단계 S5: 컨트롤러 유닛(2)은 이어서 명령 캐시 유닛에서 다음 CONFIG 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 장치는 상기 층 신경망 계산에 필요하는 각종 상수를 배치한다. 예를 들어 연산 유닛(51, 61)은 마이크로 명령 내의 파라미터에 의거하여 유닛 내부 레지스터의 값을 배치하고, 상기 파라미터에는 예를 들어 본 층 계산의 정밀도 설정, 활성 함수의 데이터(예를 들어 본 층 계싼의 정밀도 비트, Lrn층 알고리즘의 rang 파라미터, AveragePooling층 알고리즘 윈도 사이즈의 역수 등)가 포함된다.

[0040] 단계 S6: 컨트롤러 유닛(2)은 이어서 명령 캐시 유닛에서 다음 COMPUTE 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 1차 연산 모듈(5)은 먼저 H 트리 모듈(4)을 통하여 뉴런 벡터를 각 2차 연산 모듈(6)로 전송하고, 2차 연산 모듈(6)의 뉴런 캐시 유닛(63)에 저장한다.

[0041] 단계 S7: COMPUTE 명령이 디코딩한 마이크로 명령에 의거하여, 2차 연산 모듈(6)의 연산 유닛(61)은 가중치 캐시 유닛(64)에서 가중치 벡터(가중치 행렬 중 상기 2차 연산 모듈(6)에 대응하는 열 벡터)를 판독하고, 뉴런 캐시 유닛에서 입력 뉴런 벡터를 판독하여 가중치 벡터와 입력 뉴런 벡터의 내적 연산을 완료하고 중간 결과를 H 트리를 통해 반환한다.

[0042] 단계 S8: H 트리 모듈(4)에 있어서, 각 2차 연산 모듈(6)에서 반환되는 중간 결과는 단계적으로 완전한 중간 결과 벡터로 합산된다.

[0043] 단계 S9: 1차 연산 모듈(5)은 H 트리 모듈(5)의 반환 값을 획득하고, COMPUTE 명령이 디코딩한 마이크로 명령에 의거하여 뉴런 캐시 유닛(53)에서 오프셋 벡터를 판독하고, H 트리 모듈(4)에서 반환되는 벡터와 서로 더한 후 더한 결과를 다시 활성화하고, 최종의 출력 뉴런 벡터를 뉴런 캐시 유닛(53)에 기록한다.

[0044] 단계 S10: 컨트롤러 유닛(2)은 이어서 명령 캐시 유닛에서 다음 IO 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 엑세스 유닛(3)은 뉴런 캐시 유닛(53) 중의 출력 뉴런 벡터를 외부 주소 공간에 지정되는 주소에 저장하고, 연산을 종료한다.

[0045] 다층 인공 신경망은 그 구현 과정이 단층 신경망과 유사하며, 이전 층 인공 신경망 실행을 완료한 후, 다음 층의 연산 명령은 1차 연산 모듈 중에 저장되는 이전 층의 출력 뉴런 주소를 본 층의 입력 뉴런 주소로 삼는다. 마찬가지로, 명령 중의 가중치 주소와 오프셋 주소도 본 층에서 대응하는 주소로 변경될 수 있다.

[0046] 인공 신경망 정방향 트레이닝 실행용 장치와 명령 집합을 채택하여 CPU와 GPU 연산 성능 부족과 전단 디코딩 비용이 큰 문제를 해결하였다. 또한 다층 인공 신경망 정방향 트레이닝에 대한 지원을 효과적으로 향상시킨다.

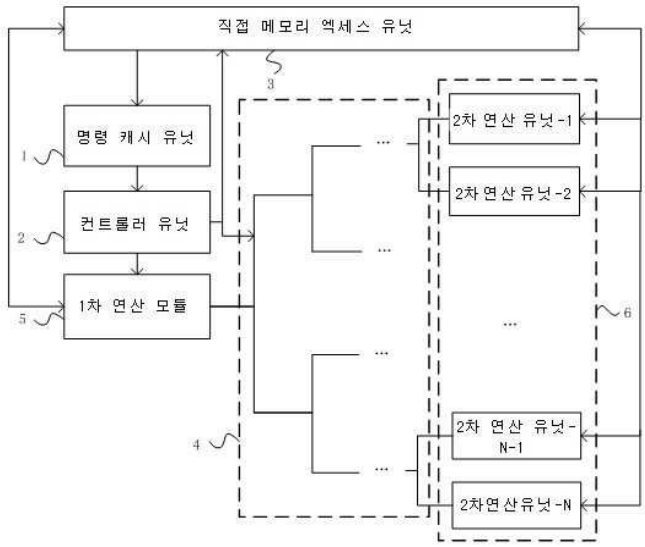
[0047] 다층 인공 신경망 정방향 트레이닝 전용 온칩 캐시를 채택하여 입력 뉴런과 가중치 데이터의 재사용성을 충분히 발굴하며, 반복적으로 메모리가 이러한 데이터를 판독하는 것을 방지하고 메모리 엑세스 대역폭을 낮추며 메모리 대역폭이 다층 인공 신경망 정방향 트레이닝 성능 병목이 되는 현상을 방지한다.

[0048] 앞서 도면에서 기재한 진행과정 또는 방법에는 하드웨어(예를 들어 회로, 전용 논리 등), 펌웨어, 소프트웨어(예를 들어 구체화된 비일시적 컴퓨터 판독 가능 매체)를 포함할 수 있으며, 또는 양자 조합의 처리 논리(processing logic)로 실행할 수 있다. 비록 상기 내용이 특정 순서 조작에 따라 진행과정 또는 방법을 설명하기는 하나, 상기에서 설명한 특정 조작은 다른 순서로 실행할 수 있다. 그 외 병렬하여 비(非)순서적으로 일부 조작을 실행할 수 있다.

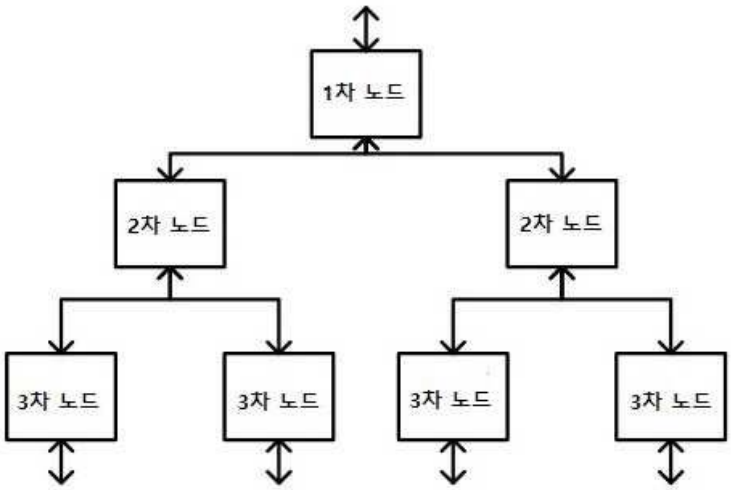
[0049] 상기 발명의 설명에서는 특정한 예시적 실시예를 참조하여 본 발명의 각 실시예를 설명하였다. 각 실시예에 대하여 진행할 수 있는 각종 수식은 상기 첨부한 특허청구범위에 있어서 본 발명의 더욱 광범위한 정신과 범위에 위배되지 않는다. 이에 상응하여, 발명의 설명과 첨부 도면은 설명을 위한 것이므로 본 발명을 제한하지 않는다.

도면

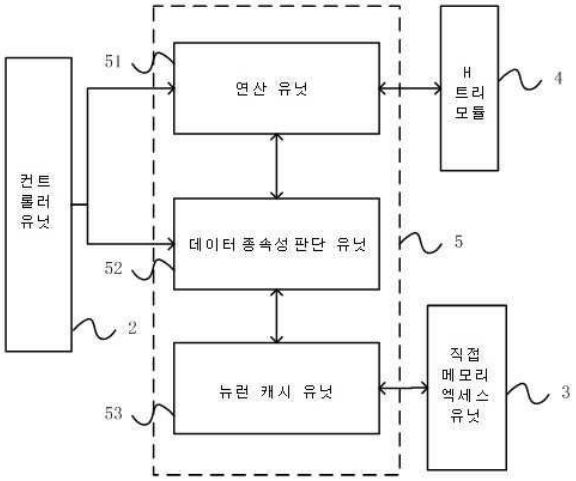
도면1



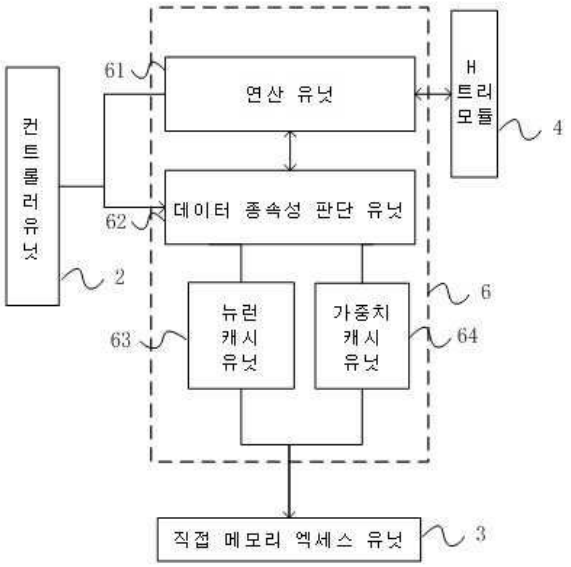
도면2



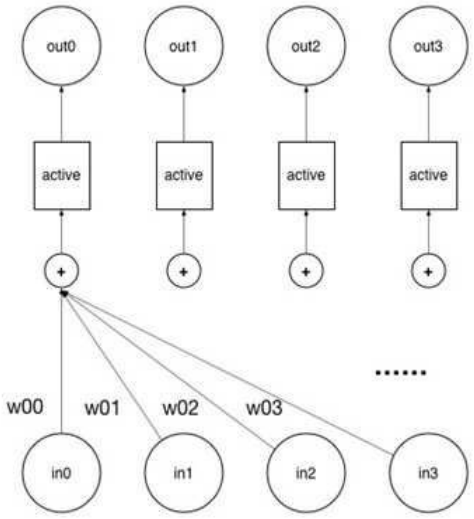
도면3


51
연산 유닛
H 트리 모듈
4
컨트롤러 유닛
데이터 종속성 판단 유닛
5
52
2
직접 메모리 엑세스 유닛
3
뉴런 캐시 유닛
53

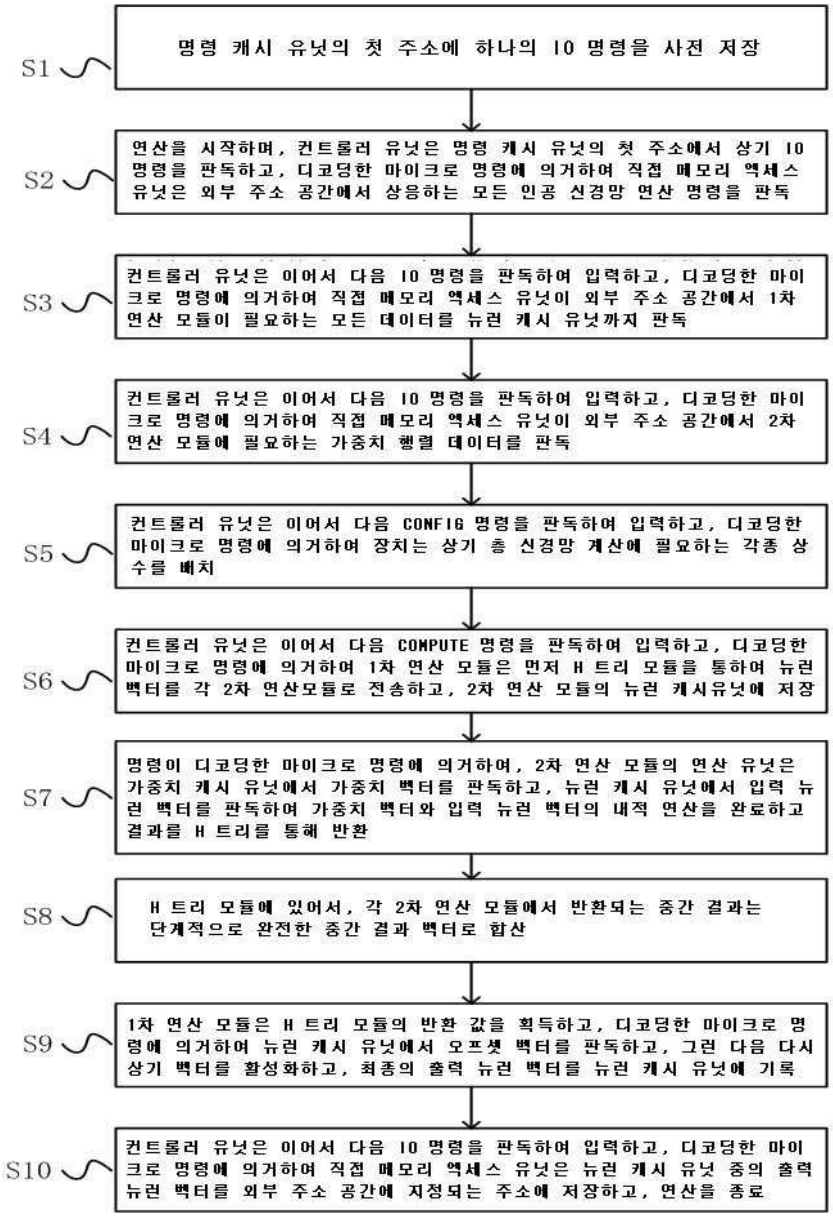
도면4


61
연산 유닛
H 트리 모듈
4
컨트롤러 유닛
데이터 종속성 판단 유닛
62
2
6
뉴런 캐시 유닛
가중치 캐시 유닛
63
64
직접 메모리 엑세스 유닛
3

도면5


out0
out1
out2
out3
active
active
active
active
+
+
+
+
.....
w00
w01
w02
w03
in0
in1
in2
in3

도면6


S1
명령 캐시 유닛의 첫 주소에 하나의 IO 명령을 사전 저장
S2
연산을 시작하며, 컨트롤러 유닛은 명령 캐시 유닛의 첫 주소에서 상기 IO 명령을 판독하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 액세스 유닛은 외부 주소 공간에서 상응하는 모든 인공 신경망 연산 명령을 판독
S3
컨트롤러 유닛은 이어서 다음 IO 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 액세스 유닛이 외부 주소 공간에서 1차 연산 모듈이 필요하는 모든 데이터를 뉴런 캐시 유닛까지 판독
S4
컨트롤러 유닛은 이어서 다음 IO 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 액세스 유닛이 외부 주소 공간에서 2차 연산 모듈에 필요하는 가중치 행렬 데이터를 판독
S5
컨트롤러 유닛은 이어서 다음 CONFIG 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 장치는 상기 층 신경망 계산에 필요하는 각종 상수를 배치
S6
컨트롤러 유닛은 이어서 다음 COMPUTE 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 1차 연산 모듈은 먼저 H 트리 모듈을 통하여 뉴런 벡터를 각 2차 연산모듈로 전송하고, 2차 연산 모듈의 뉴런 캐시유닛에 저장
S7
명령이 디코딩한 마이크로 명령에 의거하여, 2차 연산 모듈의 연산 유닛은 가중치 캐시 유닛에서 가중치 벡터를 판독하고, 뉴런 캐시 유닛에서 입력 뉴런 벡터를 판독하여 가중치 벡터와 입력 뉴런 벡터의 내적 연산을 완료하고 결과를 H 트리를 통해 반환
S8
H 트리 모듈에 있어서, 각 2차 연산 모듈에서 반환되는 중간 결과는 단계적으로 완전한 중간 결과 벡터로 합산
S9
1차 연산 모듈은 H 트리 모듈의 반환 값을 획득하고, 디코딩한 마이크로 명령에 의거하여 뉴런 캐시 유닛에서 오프셋 벡터를 판독하고, 그런 다음 다시 상기 벡터를 활성화하고, 최종의 출력 뉴런 벡터를 뉴런 캐시 유닛에 기록
S10
컨트롤러 유닛은 이어서 다음 IO 명령을 판독하여 입력하고, 디코딩한 마이크로 명령에 의거하여 직접 메모리 액세스 유닛은 뉴런 캐시 유닛 중의 출력 뉴런 벡터를 외부 주소 공간에 지정되는 주소에 저장하고, 연산을 종료