(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0104406
(43) 공개일자 2019년09월09일

(51) 국제특허분류(Int. Cl.)
G06N 3/02 (2019.01)
(52) CPC특허분류
G06N 3/02 (2019.01)
(21) 출원번호 10-2019-7023878
(22) 출원일자(국제) 2018년07월13일
심사청구일자 2019년08월14일
(85) 번역문제출일자 2019년08월14일
(86) 국제출원번호 PCT/CN2018/095548
(87) 국제공개번호 WO 2019/076095
국제공개일자 2019년04월25일
(30) 우선권주장
201710989575.4 2017년10월20일 중국(CN)
(뒷면에 계속)

(71) 출원인
상하이 캠브리콘 인포메이션 테크놀로지 컴퍼니 리미티드
중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6
(72) 발명자
리우 사울리
중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6
조우 슈다
중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6
(뒷면에 계속)
(74) 대리인
제일특허법인(유)

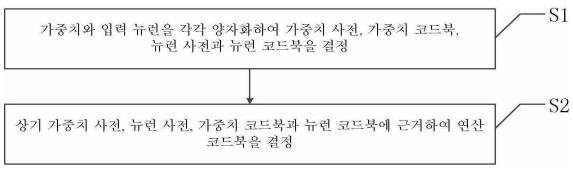
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 **처리방법 및 장치**

(57) 요 약

본원 발명은 가중치와 입력 뉴런을 각각 양자화하여 가중치 사전, 가중치 코드북, 뉴런 사전과 뉴런 코드북을 결정하는 단계; 및 상기 가중치 코드북과 뉴런 코드북에 근거하여 연산 코드북을 결정하는 단계를 포함하는 처리방법 및 장치를 제공한다. 이와 동시에 본원 발명은 양자화한 후의 데이터에 근거하여 연산 코드북을 결정하고 두 가지 양자화한 후의 데이터를 결합하여 데이터 처리에 용이하도록 한다.

대 표 도 - 도1a



(72) 발명자

두 지동

중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6

리우 다오푸

중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6

장 레이

중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6

첸 티안시

중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6

후 수아이

중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6

웨이 지에

중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6

멩 샤오푸

중국 상하이 201306 푸동 뉴 아레아 동후이 로드 넘버 168 블럭 비 플로어 6

(30) 우선권주장

201711004974.7 2017년10월24일 중국(CN)
201711061069.5 2017년10월24일 중국(CN)
201711029543.6 2017년10월29일 중국(CN)
201711118938.3 2017년10월29일 중국(CN)
201711289667.8 2017년12월07일 중국(CN)

명 세 서

청구범위

청구항 1

가중치와 입력 뉴런을 각각 양자화하여 가중치 사전, 가중치 코드북, 뉴런 사전과 뉴런 코드북을 결정하는 단계; 및

상기 가중치 코드북과 뉴런 코드북에 근거하여 연산 코드북을 결정하는 단계를 포함하는 것을 특징으로 하는 처리방법.

청구항 2

제1항에 있어서,

가중치를 양자화하는 상기 단계는,

상기 가중치를 그루핑하되, 매 그룹의 가중치에 클러스터링 알고리즘을 이용하여 클러스터링 작업을 진행하고 상기 매 그룹의 가중치를 m타입로 나누며 m는 자연수이고 매 타입의 가중치는 하나의 가중치 인덱스와 대응하여 상기 가중치 사전을 결정하는데 여기서 상기 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 상기 가중치 위치는 가중치가 신경망 구조에서의 위치를 가리키는 단계; 및

매 타입의 모든 가중치를 하나의 중심 가중치로 대체하여 상기 가중치 코드북을 결정하되, 여기서 상기 가중치 코드북은 가중치 인덱스와 중심 가중치를 포함하는 단계를 포함하는 것을 특징으로 하는 처리방법.

청구항 3

제1항 또는 제2항에 있어서,

입력 뉴런을 양자화하는 상기 단계는,

상기 입력 뉴런을 p세그먼트로 나누고 매 세그먼트의 입력 뉴런은 하나의 뉴런 범위 및 하나의 뉴런 인덱스와 대응하여 상기 뉴런 사전을 결정하되, 여기서 p는 자연수인 단계; 및

상기 입력 뉴런을 코딩하고 매 세그먼트의 모든 입력 뉴런을 하나의 중심 뉴런으로 대체하여 상기 뉴런 코드북을 결정하는 단계를 포함하는 것을 특징으로 하는 처리방법.

청구항 4

제3항에 있어서,

상기 연산 코드북을 결정하는 단계는 구체적으로,

상기 가중치에 근거하여 상기 가중치 코드북에서의 대응되는 가중치 인덱스를 결정한 다음 가중치 인덱스를 통해 상기 가중치와 대응되는 중심 가중치를 결정하는 단계;

상기 입력 뉴런에 근거하여 상기 뉴런 코드북에서의 대응되는 뉴런 인덱스를 결정한 다음 뉴런 인덱스를 통해 상기 입력 뉴런과 대응되는 중심 뉴런을 결정하는 단계; 및

상기 중심 가중치와 중심 뉴런에 대해 연산작업을 진행하여 연산 결과를 얻고 상기 연산 결과를 매트릭스로 구성하여 상기 연산 코드북을 결정하는 단계를 포함하는 것을 특징으로 하는 처리방법.

청구항 5

제4항에 있어서,

상기 연산작업은 덧셈, 곱셈과 풀링에서의 적어도 하나를 포함하되, 여기서 상기 풀링은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함하는 것을 특징으로 하는 처리방법.

청구항 6

제1항 내지 제5항 중 임의의 한 항에 있어서,

상기 가중치와 입력 뉴런에 대해 리트레이닝을 진행하되, 리트레이닝 할 경우 상기 가중치 코드북과 뉴런 코드북만을 트레이닝하고 상기 가중치 사전과 뉴런 사전에서의 콘텐츠는 변하지 않으며 상기 리트레이닝은 오차역전파법을 사용하는 단계를 더 포함하는 것을 특징으로 하는 처리방법.

청구항 7

제2항에 있어서,

상기 가중치를 그루핑하는 상기 단계는,

신경망에서의 모든 가중치를 한 그룹으로 하는 한 그룹으로의 그루핑;

상기 신경망에서의 모든 합성곱 층의 가중치, 모든 풀 연결층의 가중치와 모든 장단기 메모리 네트워크층의 가중치를 각각 한 그룹으로 구획하는 레이어 타입 그루핑;

상기 신경망에서의 하나 또는 다수의 합성곱 층의 가중치, 하나 또는 다수의 풀 연결층의 가중치와 하나 또는 다수의 장단기 메모리 네트워크층의 가중치를 각각 한 그룹으로 구획하는 층간 그루핑; 및

상기 신경망의 한 층 내의 가중치를 분할하고 분할한 후의 매 한 부분을 한 그룹으로 그루핑하는 층내 그루핑을 포함하는 것을 특징으로 하는 처리방법.

청구항 8

제2항에 있어서,

상기 클러스터링 알고리즘은 K-means, K-medoids, Clara 및/또는 Clarans를 포함하는 것을 특징으로 하는 처리방법.

청구항 9

제2항 내지 제8항 중 임의의 한 항에 있어서,

매 타입의 대응하는 중심 가중치의 선택방법은 비용 함수 $J(w, w_0)$의 값이 제일 작도록 할 경우의 W0의 값을 결정하되, 이때 W0의 값은 상기 중심 가중치인 단계를 포함하는데,

여기서 $J(w, w_0) = \sum_{i=1}^{n}(w_i - w_0)^2$, $J(w, w_0)$는 비용 함수이고 W는 이러한 타입에서의 모든 가중치이며 W0는 중심 가중치이고 n은 이러한 타입에서의 모든 가중치의 수량이며 Wi는 이러한 타입에서의 i번째 가중치이고 $1 \leq i \leq n$이며 i는 자연수인 것을 특징로 하는 처리방법.

청구항 10

작동 명령을 저장하기 위한 메모리;

메모리에서의 작동 명령을 실행하고 상기 작동 명령을 실행할 경우 청구항 1 내지 청구항 9에서의 임의의 한 항의 처리방법에 따라 작동하는 프로세서를 포함하는 것을 특징으로 하는 처리장치.

청구항 11

제10항에 있어서,

상기 작동 명령은 이진수로서 옵 코드와 주소 코드를 포함하되, 옵 코드는 프로세서가 진행하게 될 작동을 지시하고 주소 코드는 프로세서로 하여금 메모리의 주소에서 작동에 참여한 데이터를 판독하도록 지시하는 것을 특징으로 하는 처리장치.

발명의 설명

기 술 분 야

[0001] 본원 발명은 데이터 처리분야에 관한 것으로 특히 처리방법 및 장치, 연산방법 및 장치에 관한 것이다.

[0002] 본원 발명은 데이터 처리분야에 관한 것으로 특히 처리방법 및 장치, 연산방법 및 장치에 관한 것이다.

배 경 기 술

[0003] 본원 발명의 목적은 처리방법 및 장치, 연산방법 및 장치를 제공하여 상술한 적어도 하나의 기술적 과제를 해결하는 것이다.

[0004] 본원 발명의 한 양태는,

[0005] 가중치와 입력 뉴런을 각각 양자화하여 가중치 사전, 가중치 코드북, 뉴런 사전과 뉴런 코드북을 결정하는 단계; 및

[0006] 상기 가중치 코드북과 뉴런 코드북에 근거하여 연산 코드북을 결정하는 단계를 포함하는 처리방법을 제공한다.

[0007] 본원 발명의 하나의 가능한 실시예에서 가중치를 양자화하는 단계는,

[0008] 상기 가중치를 그루핑하되, 매 그룹의 가중치에 클러스터링 알고리즘을 이용하여 클러스터링 작업을 진행하고 상기 매 그룹의 가중치를 m타입으로 나누며 m는 자연수이고 매 타입의 가중치는 하나의 가중치 인덱스와 대응하여 상기 가중치 사전을 결정하는데 여기서 상기 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 상기 가중치 위치는 가중치가 신경망 구조에서의 위치를 가리키는 단계; 및

[0009] 매 타입의 모든 가중치를 하나의 중심 가중치로 대체하여 상기 가중치 코드북을 결정하되, 여기서 상기 가중치 코드북은 가중치 인덱스와 중심 가중치를 포함하는 단계를 포함한다.

[0010] 본원 발명의 하나의 가능한 실시예에서 입력 뉴런을 양자화하는 단계는,

[0011] 상기 입력 뉴런을 p세그먼트로 나누고 매 세그먼트의 입력 뉴런은 하나의 뉴런 범위 및 하나의 뉴런 인덱스와 대응하여 상기 뉴런 사전을 결정하되, 여기서 p는 자연수인 단계; 및

[0012] 상기 입력 뉴런을 코딩하고 매 세그먼트의 모든 입력 뉴런을 하나의 중심 뉴런으로 대체하여 상기 뉴런 코드북을 결정하는 단계를 포함한다.

[0013] 본원 발명의 하나의 가능한 실시예에서 상기 연산 코드북을 결정하는 단계는 구체적으로,

[0014] 상기 가중치에 근거하여 상기 가중치 코드북에서의 대응되는 가중치 인덱스를 결정한 다음 가중치 인덱스를 통해 상기 가중치와 대응되는 중심 가중치를 결정하는 단계;

[0015] 상기 입력 뉴런에 근거하여 상기 뉴런 코드북에서의 대응되는 뉴런 인덱스를 결정한 다음 뉴런 인덱스를 통해 상기 입력 뉴런과 대응되는 중심 뉴런을 결정하는 단계; 및

[0016] 상기 중심 가중치와 중심 뉴런에 대해 연산작업을 진행하여 연산 결과를 얻고 상기 연산 결과를 매트릭스로 구성하여 상기 연산 코드북을 결정하는 단계를 포함한다.

[0017] 본원 발명의 하나의 가능한 실시예에서 상기 연산작업은 덧셈, 곱셈과 풀링에서의 적어도 하나를 포함하되, 여기서 상기 풀링은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함한다.

[0018] 본원 발명의 하나의 가능한 실시예에서는, 상기 가중치와 입력 뉴런에 대해 리트레이닝을 진행하되, 리트레이닝할 경우 상기 가중치 코드북과 뉴런 코드북만을 트레이닝하고 상기 가중치 사전과 뉴런 사전에서의 콘텐츠는 변하지 않으며 상기 리트레이닝은 오차역전파법을 사용하는 단계를 더 포함한다.

[0019] 본원 발명의 하나의 가능한 실시예에서 상기 가중치를 그루핑하는 상기 단계는,

[0020] 신경망에서의 모든 가중치를 한 그룹으로 하는 한 그룹으로의 그루핑;

[0021] 상기 신경망에서의 모든 합성곱 층의 가중치, 모든 풀 연결층의 가중치와 모든 장단기 메모리 네트워크층의 가중치를 각각 한 그룹으로 구획하는 레이어 타입 그루핑;

[0022] 상기 신경망에서의 하나 또는 다수의 합성곱 층의 가중치, 하나 또는 다수의 풀 연결층의 가중치와 하나 또는 다수의 장단기 메모리 네트워크층의 가중치를 각각 한 그룹으로 구획하는 층간 그루핑; 및

[0023] 상기 신경망의 한 층 내의 가중치를 분할하고 분할한 후의 매 한 부분을 한 그룹으로 그루핑하는 층내 그루핑을 포함한다.

[0024] 본원 발명의 하나의 가능한 실시예에서 상기 클러스터링 알고리즘은 K-means, K-medoids, Clara 및/또는 Clarans를 포함한다.

[0025] 본원 발명의 하나의 가능한 실시예에서 매 타입의 대응하는 중심 가중치의 선택방법은, 비용 함수 $J(w, w_0)$의 값이 제일 작도록 할 경우의 W0의 값을 결정하되, 이때 W0의 값은 상기 중심 가중치인 단계를 포함하는데,

[0026] 여기서 $J(w, w_0) = \sum_{i=1}^{n}(w_i - w_0)^2$ $J(w, w_0)$는 비용 함수이고 W는 이러한 타입에서의 모든 가중치이며 W0는 중심 가중치이고 n은 이러한 타입에서의 모든 가중치의 수량이며 Wi는 이러한 타입에서의 i번째 가중치이고 $1 \leq i \leq n$이며 i는 자연수이다.

[0027] 본원 발명의 다른 한 양태는,

[0028] 작동 명령을 저장하기 위한 메모리;

[0029] 메모리에서의 작동 명령을 실행하고 상기 작동 명령을 실행할 경우 청구항 1 내지 청구항 9에서의 임의의 한 항의 처리방법에 따라 작동하는 프로세서를 포함하는 처리장치를 제공한다.

[0030] 본원 발명의 하나의 가능한 실시예에서 상기 작동 명령은 이진수로서 옵 코드와 주소 코드를 포함하되, 옵 코드는 프로세서가 진행하게 될 작동을 지시하고 주소 코드는 프로세서로 하여금 메모리의 주소에서 작동에 참여한 데이터를 판독하도록 지시한다.

[0031] 본원 발명의 또 다른 양태는,

[0032] 수신한 명령을 디코딩하여 검색제어정보를 생성하기 위한 명령제어유닛; 및

[0033] 상기 검색제어정보 및 수신한 가중치 사전, 뉴런 사전, 연산 코드북, 가중치와 입력 뉴런에 근거하여 연산 코드북으로부터 출력 뉴런을 검색하기 위한 검색 테이블 유닛을 포함하는 연산장치를 제공한다.

[0034] 본원 발명의 하나의 가능한 실시예에서 상기 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 상기 뉴런 사전은 력 뉴런과 뉴런 인덱스를 포함하며 상기 연산 코드북은 가중치 인덱스, 뉴런 인덱스 및 입력 뉴런과 가중치의 연산 결과를 포함한다.

[0035] 본원 발명의 하나의 가능한 실시예에서 상기 연산장치는,

[0036] 외부에서 입력한 입력정보를 전처리하여 상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 얻기 위한 전처리 유닛;

[0037] 입력 뉴런, 가중치, 가중치 사전, 뉴런 사전, 연산 코드북과 명령을 저장하고 출력 뉴런을 수신하기 위한 저장 유닛;

[0038] 상기 명령, 입력 뉴런, 가중치, 가중치 인덱스, 뉴런 인덱스와 출력 뉴런을 캐시하기 위한 캐시 유닛;

[0039] 상기 저장 유닛과 캐시 유닛 사이에서 데이터 또는 명령의 판독 기록을 진행하기 위한 직접 메모리 액세스 유닛을 더 포함한다.

[0040] 본원 발명의 하나의 가능한 실시예에서 상기 캐시 유닛은,

[0041] 상기 명령을 캐시하고 캐시된 명령을 명령제어유닛에 출력하기 위한 명령 캐시;

[0042] 상기 가중치를 캐시하기 위한 가중치 캐시;

[0043] 상기 입력 뉴런을 캐시하기 위한 입력 뉴런 캐시; 및

[0044] 검색 테이블 유닛이 출력한 출력 뉴런을 캐시하기 위한 출력 뉴런 캐시를 포함한다.

[0045] 본원 발명의 하나의 가능한 실시예에서 상기 캐시 유닛은,

[0046] 가중치 인덱스를 캐시하기 위한 가중치 인덱스 캐시; 및

[0047] 뉴런 인덱스를 캐시하기 위한 뉴런 인덱스 캐시를 더 포함한다.

[0048] 본원 발명의 하나의 가능한 실시예에서 외부에서 입력한 입력정보를 전처리할 경우 상기 전처리 유닛은 구체적으로 분할, 가우스 필터링, 이진화, 규칙화 및/또는 정규화에 사용된다.

[0049] 본원 발명의 하나의 가능한 실시예에서 검색 테이블 유닛은,

[0050] 가중치 인덱스(in1)와 뉴런 인덱스(in2)를 입력하고 곱셈 검색 테이블을 통해 테이블 조사 동작(mult_lookup)으로 가중치 인덱스와 대응되는 중심 가중치(data1)와 뉴런 인덱스와 대응되는 중심 뉴런(data2)의 곱셈동작을 완성, 즉 테이블 조사 동작 out=mult_lookup(in1,in2)으로 곱셈기능 out=data1*data2을 완성하기 위한 곱셈 검색 테이블; 및/또는

[0051] 인덱스(in)에 근거하여 단계적 덧셈 검색 테이블을 통해 테이블 조사 동작(add_lookup)으로 인덱스와 대응되는 중심 데이터(data)의 덧셈 동작을 완성하도록 입력하되, 여기서 in과 data는 길이가 N인 벡터이고 N은 자연수, 즉 테이블 조사 동작 out=add_lookup(in)으로 덧셈기능 out＝data[1]+data[2]+...+data[N]을 완성하거나 및/또는 가중치 인덱스(in1)와 뉴런 인덱스(in2)가 덧셈 검색 테이블을 통해 테이블 조사 동작으로 가중치 인덱스와 대응되는 중심 가중치(data1)와 뉴런 인덱스와 대응되는 중심 뉴런(data2)의 덧셈 동작을 완성하도록 입력, 즉 테이블 조사 동작 out=add_lookup(in1,in2)으로 덧셈기능 out=data1+data2을 완성하는 덧셈 검색 테이블; 및/또는

[0052] 인덱스와 대응되는 중심 데이터(data)의 풀링 동작을 입력, 즉 테이블 조사 out=pool_lookup(in)로 풀링 동작 out=pool(data)을 완성하되, 풀링 동작은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함하는 풀링 검색 테이블을 포함한다.

[0053] 본원 발명의 하나의 가능한 실시예에서 상기 명령은 신경망 전용명령이고 상기 신경망 전용명령은,

[0054] 신경망 수행과정을 제어하기 위한 제어 명령;

[0055] 상이한 저장매체 사이의 데이터 전송을 완성하기 위한 것으로 데이터 양식은 매트릭스, 벡터와 스칼라를 포함하는 데이터 전송 명령;

[0056] 매트릭스 연산 명령, 벡터 연산 명령, 스칼라 연산 명령, 콘볼루션 신경망 연산 명령, 완전 연결 신경망 연산 명령, 풀링신경망 연산 명령, RBM 신경망 연산 명령, LRN 신경망 연산 명령, LCN 신경망 연산 명령, LSTM 신경망 연산 명령, RNN 신경망 연산 명령, RELU 신경망 연산 명령, PRELU 신경망 연산 명령, SIGMOID 신경망 연산 명령, TANH 신경망 연산 명령, MAXOUT 신경망 연산 명령을 포함하는 신경망의 산술 연산을 완성하기 위한 연산 명령; 및

[0057] 벡터논리 연산 명령과 스칼라 논리 연산 명령을 포함하는 신경망의 논리 연산을 완성하기 위한 논리 명령을 포함한다.

[0058] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 전용명령은 적어도 하나의 Cambricon 명령을 포함하고 상기 Cambricon 명령은 옵 코드와 피연산자를 포함하며 상기 Cambricon 명령은,

[0059] 수행과정을 제어하고 상기 Cambricon 제어 명령은 점프 명령과 조건부 분기 명령을 포함하는 Cambricon 제어 명령;

[0060] 로드 명령, 저장 명령, 운송 명령을 포함하는 상이한 저장매체 사이의 데이터 전송을 완성하되, 여기서 상기 로드 명령은 데이터를 메인 메모리로부터 캐시에 로딩하기 위한 것이고, 상기 저장 명령은 데이터를 캐시로부터 메인 메모리에 저장하기 위한 것이며 상기 운송 명령은 캐시와 캐시 또는 캐시와 레지스터 또는 레지스터와 레지스터 사이에서 데이터를 운송하기 위한 것인 Cambricon 데이터 전송 명령;

[0061] Cambricon 매트릭스 연산 명령, Cambricon 벡터 연산 명령과 Cambricon 스칼라 연산 명령을 포함하는 신경망 산술 연산을 완성하되, 여기서 상기 Cambricon 매트릭스 연산 명령은 매트릭스 곱셈 벡터 연산, 벡터 곱셈 매트릭스 연산, 매트릭스 곱셈 스칼라 연산, 외적 연산, 매트릭스 덧셈 매트릭스 연산과 매트릭스 뺄셈 매트릭스 연산을 포함하는 신경망에서의 매트릭스 연산을 완성하기 위한 것이고 상기 Cambricon 벡터 연산 명령은 벡터 기본 연산, 벡터 초월함수 연산, 내적 연산, 벡터 랜덤 생성 연산과 벡터에서의 최대/최소치 연산을 포함하는 신경망에서의 벡터 연산을 완성하기 위한 것이며 Cambricon 스칼라 연산 명령은 스칼라 기본 연산과 스칼라 초월함수 연산을 포함하는 신경망에서의 스칼라 연산을 완성하기 위한 것인 Cambricon 연산 명령; 및

[0062] 신경망의 논리 연산을 위한 것으로 Cambricon 벡터논리 연산 명령과 Cambricon스칼라 논리 연산 명령을 포함하되, 여기서 상기 Cambricon 벡터논리 연산 명령은 벡터 비교 연산, 벡터 논리 연산과 벡터 크기 합병 연산을 위한 것이고 여기서 벡터논리 연산은 그리고, 또는, 아님을 포함하며 상기 Cambricon 스칼라 논리 연산 명령은 스칼라 비교 연산과 스칼라 논리 연산 연산을 위한 것인 Cambricon 논리 명령을 포함한다.

[0063] 본원 발명의 하나의 가능한 실시예에서 상기 Cambricon 데이터 전송 명령은 매트릭스, 백터와 스칼라에서의 한 가지 또는 여러 가지 데이터 조직방식을 지지하고 상기 벡터 기본 연산은 벡터 더하기, 빼기, 곱하기, 나누기를 포함하며 벡터 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 스칼라 기본 연산은 스칼라 더하기, 빼기, 곱하기, 나누기를 포함하며 스칼라 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 벡터 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 벡터논리 연산은 그리고, 또는, 아님을 포함하고 상기 스칼라 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 스칼라 논리 연산은 그리고, 또는, 아님을 포함한다.

[0064] 본원 발명의 또 다른 양태는,

[0065] 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 수신하는 단계;

[0066] 상기 명령을 디코딩하여 검색제어정보를 결정하는 단계; 및

[0067] 상기 검색제어정보, 가중치, 가중치 사전, 뉴런 사전과 입력 뉴런에 근거하여 연산 코드북에서 출력 뉴런을 검색하는 단계를 포함하는 다른 연산방법을 제공한다.

[0068] 본원 발명의 하나의 가능한 실시예에서 상기 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 상기 뉴런 사전은 입력 뉴런과 뉴런 인덱스를 포함하며 상기 연산 코드북은 가중치 인덱스, 뉴런 인덱스 및 가중치와 입력 뉴런의 연산 결과를 포함한다.

[0069] 본원 발명의 하나의 가능한 실시예에서 상기 검색제어정보, 가중치와 입력 뉴런에 근거하여 연산 코드북에서 출력 뉴런을 검색하는 단계는,

[0070] 상기 가중치, 입력 뉴런, 가중치 사전과 뉴런 사전에 근거하여 뉴런 사전에서 뉴런 범위를 결정함으로써 뉴런 인덱스를 결정하고 가중치 사전에서 가중치 위치를 결정함으로써 가중치 인덱스를 결정하는 단계;

[0071] 상기 가중치 인덱스와 뉴런 인덱스에 근거하여 연산 코드북에서 상기 연산 결과를 검색함으로써 출력 뉴런을 결정하는 단계를 포함한다.

[0072] 본원 발명의 하나의 가능한 실시예에서 상기 연산 결과는 덧셈, 곱셈과 풀링에서의 적어도 한 연산작업의 결과를 포함하되, 여기서 풀링은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함한다.

[0073] 본원 발명의 하나의 가능한 실시예에서 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 수신하기 전에 외부에서 입력한 입력정보를 전처리하여 상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 얻는 단계를 더 포함; 및

[0074] 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 수신한 후 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 저장하고 출력 뉴런을 수신하는 단계; 및 상기 명령, 입력 뉴런, 가중치와 출력 뉴런을 캐시하는 단계를 더 포함한다.

[0075] 본원 발명의 하나의 가능한 실시예에서 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 수신한 후 상기 가중치 인덱스와 뉴런 인덱스를 캐시하는 단계를 더 포함한다.

[0076] 본원 발명의 하나의 가능한 실시예에서 상기 전처리는 분할, 가우스 필터링, 이진화, 규칙화 및/또는 정규화를 포함한다.

[0077] 본원 발명의 하나의 가능한 실시예에서 상기 명령은 신경망 전용명령이고 상기 신경망 전용명령은,

[0078] 신경망 수행과정을 제어하기 위한 제어 명령;

[0079] 상이한 저장매체 사이의 데이터 전송을 완성하되, 상기 데이터의 데이터 양식은 매트릭스, 벡터와 스칼라를 포함하는 데이터 전송 명령;

[0080] 매트릭스 연산 명령, 벡터 연산 명령, 스칼라 연산 명령, 콘볼루션 신경망 연산 명령, 완전 연결 신경망 연산

명령, 폴링신경망 연산 명령, RBM 신경망 연산 명령, LRN 신경망 연산 명령, LCN 신경망 연산 명령, LSTM 신경망 연산 명령, RNN 신경망 연산 명령, RELU 신경망 연산 명령, PRELU 신경망 연산 명령, SIGMOID 신경망 연산 명령, TANH 신경망 연산 명령, MAXOUT 신경망 연산 명령을 포함하는 신경망의 산술 연산을 완성하기 위한 연산 명령; 및

[0081] 벡터논리 연산 명령과 스칼라 논리 연산 명령을 포함하는 신경망의 논리 연산을 완성하기 위한 논리 명령을 포함한다.

[0082] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 전용명령은 적어도 하나의 Cambricon 명령을 포함하고 상기 Cambricon 명령은 옵 코드와 피연산자를 포함하며 상기 Cambricon 명령은,

[0083] 수행과정을 제어하고 상기 Cambricon 제어 명령은 점프 명령과 조건부 분기 명령을 포함하는 Cambricon 제어 명령;

[0084] 로드 명령, 저장 명령, 운송 명령을 포함하는 상이한 저장매체 사이의 데이터 전송을 완성하되, 여기서 상기 로드 명령은 데이터를 메인 메모리로부터 캐시에 로딩하기 위한 것이고 상기 저장 명령은 데이터를 캐시로부터 메인 메모리에 저장하기 위한 것이며 운송 실시 명령은 캐시와 캐시 또는 캐시와 레지스터 또는 레지스터와 레지스터 사이에서 데이터를 운송하기 위한 것인 Cambricon데이터 전송 명령;

[0085] Cambricon 매트릭스 연산 명령, Cambricon 벡터 연산 명령과 Cambricon 스칼라 연산 명령을 포함하는 신경망 산술 연산을 완성하되, 여기서 상기 Cambricon 매트릭스 연산 명은 매트릭스 곱셈 벡터 연산, 벡터 곱셈 매트릭스 연산, 매트릭스 곱셈 스칼라 연산, 외적 연산, 매트릭스 덧셈 매트릭스 연산과 매트릭스 뺄셈 매트릭스 연산을 포함하는 신경망에서의 매트릭스 연산을 완성하기 위한 것이고 상기 Cambricon 벡터 연산 명령은 벡터 기본 연산, 벡터 초월함수 연산, 내적 연산, 벡터 랜덤 생성 연산과 벡터에서의 최대/최소치 연산을 포함하는 신경망에서의 벡터 연산을 완성하기 위한 것이며 Cambricon 스칼라 연산 명령은 스칼라 기본 연산과 스칼라 초월함수 연산을 포함하는 신경망에서의 스칼라 연산을 완성하기 위한 것인 Cambricon 연산 명령; 및

[0086] 신경망의 논리 연산을 위한 것으로 Cambricon 벡터논리 연산 명령과 Cambricon 스칼라 논리 연산 명령을 포함하되, 여기서 상기 Cambricon 벡터논리 연산 명령은 벡터 비교 연산, 벡터 논리 연산과 벡터 크기 합병 연산을 위한 것이고 여기서 벡터논리 연산은 그리고, 또는, 아님을 포함하며 상기 Cambricon 스칼라 논리 연산 명령은 스칼라 비교 연산과 스칼라 논리 연산 연산을 위한 것인 Cambricon 논리 명령을 포함한다.

[0087] 본원 발명의 하나의 가능한 실시예에서 상기 Cambricon 데이터 전송 명령은 매트릭스, 벡터와 스칼라에서의 한 가지 또는 여러 가지 데이터 조합방식을 지지하고 상기 벡터 기본 연산은 벡터 더하기, 빼기, 곱하기, 나누기를 포함하며 벡터 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 스칼라 기본 연산은 스칼라 더하기, 빼기, 곱하기, 나누기를 포함하며 스칼라 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 벡터 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 벡터논리 연산은 그리고, 또는, 아님을 포함하고 상기 스칼라 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 스칼라 논리 연산은 그리고, 또는, 아님을 포함한다.

[0088] 본원 발명의 또 다른 양태는 연산장치를 제공하는데 상기 연산장치는,

[0089] 수신한 명령을 디코딩하여 검색제어정보를 생성하기 위한 명령제어유닛; 및

[0090] 상기 검색제어정보 및 수신한 가중치 사전, 뉴런 사전, 연산 코드북, 가중치와 입력 뉴런에 근거하여 연산 코드북으로부터 출력 뉴런을 검색하기 위한 검색 테이블 유닛을 포함한다.

[0091] 본원 발명의 하나의 가능한 실시예에서 상기 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 상기 뉴런 사전은 입력 뉴런과 뉴런 인덱스를 포함하며 상기 연산 코드북은 가중치 인덱스, 뉴런 인덱스 및 입력 뉴런과 가중치의 연산 결과를 포함한다.

[0092] 본원 발명의 하나의 가능한 실시예에서 상기 연산장치는,

[0093] 외부에서 입력한 입력정보를 전처리하여 상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 얻기 위한 전처리 유닛;

[0094] 입력 뉴런, 가중치, 가중치 사전, 뉴런 사전, 연산 코드북과 명령을 저장하고 출력 뉴런을 수신하기 위한 저장 유닛;

[0095] 상기 명령, 입력 뉴런, 가중치, 가중치 인덱스, 뉴런 인덱스와 출력 뉴런을 캐시하기 위한 캐시 유닛; 및

[0096] 상기 저장 유닛과 캐시 유닛 사이에서 데이터 또는 명령 판독 기록을 진행하기 위한 직접 메모리 액세스 유닛을 더 포함한다.

[0097] 본원 발명의 하나의 가능한 실시예에서 상기 캐시 유닛은,

[0098] 상기 명령을 캐시하고 캐시된 명령을 명령제어유닛에 출력하기 위한 명령 캐시;

[0099] 상기 가중치를 캐시하기 위한 가중치 캐시;

[0100] 상기 입력 뉴런을 캐시하기 위한 입력 뉴런 캐시;

[0101] 검색 테이블 유닛이 출력한 출력 뉴런을 캐시하기 위한 출력 뉴런 캐시를 포함한다.

[0102] 본원 발명의 하나의 가능한 실시예에서 상기 캐시 유닛은,

[0103] 가중치 인덱스를 캐시하기 위한 가중치 인덱스 캐시;

[0104] 뉴런 인덱스를 캐시하기 위한 뉴런 인덱스 캐시를 더 포함한다.

[0105] 본원 발명의 하나의 가능한 실시예에서 상기 전처리 유닛이 외부에서 입력한 입력정보를 전처리하는 단계는 분할, 가우스 필터링, 이진화, 규칙화 및/또는 정규화를 포함한다.

[0106] 본원 발명의 하나의 가능한 실시예에서 상기 검색 테이블 유닛은,

[0107] 가중치 인덱스(in1)와 뉴런 인덱스(in2)를 입력하고 곱셈 검색 테이블을 통해 테이블 조사 동작(mult_lookup)으로 가중치 인덱스와 대응되는 중심 가중치(data1)와 뉴런 인덱스와 대응되는 중심 뉴런(data2)의 곱셈동작을 완성, 즉 테이블 조사 동작 out=mult_lookup(in1,in2)으로 곱셈기능 out=data1*data2을 완성하는 곱셈 검색 테이블; 및/또는

[0108] 인덱스(in)에 근거하여 단계적 덧셈 검색 테이블을 통해 테이블 조사 동작(add_lookup)으로 인덱스와 대응되는 중심 데이터(data)의 덧셈 동작을 완성하도록 입력하되, 여기서 in과 data는 길이가 N인 벡터이고 N은 자연수, 즉 테이블 조사 동작 out=add_lookup(in)으로 덧셈기능 out＝data[1]+data[2]+...+data[N]을 완성하거나 및/또는 가중치 인덱스(in1)와 뉴런 인덱스(in2)가 덧셈 검색 테이블을 통해 테이블 조사 동작으로 가중치 인덱스와 대응되는 중심 가중치(data1)와 뉴런 인덱스와 대응되는 중심 뉴런(data2)의 덧셈 동작을 완성하도록 입력, 즉 테이블 조사 동작 out=add_lookup(in1,in2)으로 덧셈기능 out=data1+data2을 완성하는 덧셈 검색 테이블; 및/또는

[0109] 인덱스와 대응되는 중심 데이터(data)의 풀링 동작을 입력, 즉 테이블 조사 out=pool_lookup(in)로 풀링 동작 out=pool(data)을 완성하되, 풀링 동작은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함하는 풀링 검색 테이블을 포함한다.

[0110] 본원 발명의 하나의 가능한 실시예에서 상기 명령은 신경망 전용명령이고 상기 신경망 전용명령은,

[0111] 신경망 수행과정을 제어하기 위한 제어 명령;

[0112] 상이한 저장매체 사이의 데이터 전송을 완성하기 위한 것으로 데이터 양식은 매트릭스, 벡터와 스칼라를 포함하는 데이터 전송 명령;

[0113] 매트릭스 연산 명령, 벡터 연산 명령, 스칼라 연산 명령, 콘볼루션 신경망 연산 명령, 완전 연결 신경망 연산 명령, 풀링신경망 연산 명령, RBM 신경망 연산 명령, LRN 신경망 연산 명령, LCN 신경망 연산 명령, LSTM 신경망 연산 명령, RNN 신경망 연산 명령, RELU 신경망 연산 명령, PRELU 신경망 연산 명령, SIGMOID 신경망 연산 명령, TANH 신경망 연산 명령, MAXOUT 신경망 연산 명령을 포함하는 신경망의 산술 연산을 완성하기 위한 연산 명령; 및

[0114] 벡터논리 연산 명령과 스칼라 논리 연산 명령을 포함하는 신경망의 논리 연산을 완성하기 위한 논리 명령을 포함한다.

[0115] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 전용명령은 적어도 하나의 Cambricon 명령을 포함하고 상기 Cambricon 명령은 옵 코드와 피연산자를 포함하며 상기 Cambricon 명령은,

[0116] 수행과정을 제어하고 상기 Cambricon 제어 명령은 점프 명령과 조건부 분기 명령을 포함하는 Cambricon 제어 명

령;

[0117] 로드 명령, 저장 명령, 운송 명령을포함하는 상이한 저장매체 사이의 데이터 전송을 완성하되, 여기서 상기 로드 명령은 데이터를 메인 메모리로부터 캐시에 로딩하기 위한 것이고, 상기 저장 명령은 데이터를 캐시로부터 메인 메모리에 저장하기 위한 것이며 운송 실시 명령은 캐시와 캐시 또는 캐시와 레지스터 또는 레지스터와 레지스터 사이에서 데이터를 운송하기 위한 것인 Cambricon 데이터 전송 명령;

[0118] Cambricon 매트릭스 연산 명령, Cambricon 벡터 연산 명령과 Cambricon 스칼라 연산 명령을 포함하는 신경망 산술 연산을 완성하되, 여기서 상기 Cambricon 매트릭스 연산 명령은 매트릭스 곱셈 벡터 연산, 벡터 곱셈 매트릭스 연산, 매트릭스 곱셈 스칼라 연산, 외적 연산, 매트릭스 덧셈 매트릭스 연산과 매트릭스 뺄셈 매트릭스 연산을 포함하는 신경망에서의 매트릭스 연산을 완성하기 위한 것이고 상기 Cambricon 벡터 연산 명령은 벡터 기본 연산, 벡터 초월함수 연산, 내적 연산, 벡터 랜덤 생성 연산과 벡터에서의 최대/최소치 연산을 포함하는 신경망에서의 벡터 연산을 완성하기 위한 것이며 Cambricon 스칼라 연산 명령은 스칼라 기본 연산과 스칼라 초월함수 연산을 포함하는 신경망에서의 스칼라 연산을 완성하기 위한 것인 Cambricon 연산 명령; 및

[0119] 신경망의 논리 연산을 위한 것으로 Cambricon 벡터논리 연산 명령과 Cambricon 스칼라 논리 연산 명령을 포함하되, 여기서 상기 Cambricon 벡터논리 연산 명령은 벡터 비교 연산, 벡터 논리 연산과 벡터 크기 합병 연산을 위한 것이고 여기서 벡터논리 연산은 그리고, 또는, 아니를 포함하며 상기 Cambricon 스칼라 논리 연산 명령은 스칼라 비교 연산과 스칼라 논리 연산 연산을 위한 것인Cambricon 논리 명령을 포함한다.

[0120] 본원 발명의 하나의 가능한 실시예에서 상기 Cambricon 데이터 전송 명령은 매트릭스, 벡터와 스칼라에서의 한 가지 또는 여러 가지 데이터 조합방식을 지지하고 상기 벡터 기본 연산은 벡터 더하기, 빼기, 곱하기, 나누기를 포함하며 벡터 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 스칼라 기본 연산은 스칼라 더하기, 빼기, 곱하기, 나누기를 포함하며 스칼라 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 벡터 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 벡터논리 연산은 그리고, 또는, 아님을 포함하고 상기 스칼라 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 스칼라 논리 연산은 그리고, 또는, 아님을 포함한다.

[0121] 본원 발명의 또 다른 양태는,

[0122] 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 수신하는 단계;

[0123] 상기 명령을 디코딩하여 검색제어정보를 결정하는 단계;

[0124] 상기 검색제어정보, 가중치, 가중치 사전, 뉴런 사전과 입력 뉴런에 근거하여 연산 코드북에서 출력 뉴런을 검색하는 단계를 포함하는 처리방법을 제공한다.

[0125] 본원 발명의 하나의 가능한 실시예에서 상기 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 상기 뉴런 사전은 입력 뉴런과 뉴런 인덱스를 포함하며 상기 연산 코드북은 가중치 인덱스, 뉴런 인덱스 및 가중치와 입력 뉴런의 연산 결과를 포함한다.

[0126] 본원 발명의 하나의 가능한 실시예에서 상기 검색제어정보, 가중치와 입력 뉴런에 근거하여 연산 코드북에서 출력 뉴런을 검색하는 단계는,

[0127] 상기 가중치, 입력 뉴런, 가중치 사전과 뉴런 사전에 근거하여 뉴런 사전에서 뉴런 범위를 결정함으로써 뉴런 인덱스를 결정하고 가중치 사전에서 가중치 위치를 결정함으로써 가중치 인덱스를 결정하는 단계;

[0128] 상기 가중치 인덱스와 뉴런 인덱스에 근거하여 연산 코드북에서 상기 연산 결과를 검색함으로써 출력 뉴런을 결정하는 단계를 포함한다.

[0129] 본원 발명의 하나의 가능한 실시예에서 상기 연산 결과는 덧셈, 곱셈과 풀링에서의 적어도 한 연산작업의 결과를 포함하되, 여기서 풀링은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함한다.

[0130] 본원 발명의 하나의 가능한 실시예에서 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 수신하기 전에 외부에서 입력한 입력정보를 전처리하여 상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 얻는 단계를 더 포함; 및

[0131] 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 수신한 후 가중치, 입력 뉴런, 명령, 가중치

사전, 뉴런 사전, 연산 코드북을 저장하고 출력 뉴런을 수신하는 단계; 및 상기 명령, 입력 뉴런, 가중치와 출력 뉴런을 캐시하는 단계를 더 포함한다.

[0132] 본원 발명의 하나의 가능한 실시예에서 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 수신한 후 상기 가중치 인덱스와 뉴런 인덱스를 캐시하는 단계를 더 포함한다.

[0133] 본원 발명의 하나의 가능한 실시예에서 상기 전처리는 분할, 가우스 필터링, 이진화, 규칙화 및/또는 정규화를 포함한다.

[0134] 본원 발명의 하나의 가능한 실시예에서 상기 명령은 신경망 전용명령이고 상기 신경망 전용명령은,

[0135] 신경망 수행과정을 제어하기 위한 제어 명령;

[0136] 상이한 저장매체 사이의 데이터 전송을 완성하기 위한 것으로 데이터 양식은 매트릭스, 벡터와 스칼라를 포함하는 데이터 전송 명령;

[0137] 매트릭스 연산 명령, 벡터 연산 명령, 스칼라 연산 명령, 콘볼루션 신경망 연산 명령, 완전 연결 신경망 연산 명령, 풀링신경망 연산 명령, RBM 신경망 연산 명령, LRN 신경망 연산 명령, LCN 신경망 연산 명령, LSTM 신경망 연산 명령, RNN 신경망 연산 명령, RELU 신경망 연산 명령, PRELU 신경망 연산 명령, SIGMOID 신경망 연산 명령, TANH 신경망 연산 명령, MAXOUT 신경망 연산 명령을 포함하는 신경망의 산술 연산을 완성하기 위한 연산 명령; 및

[0138] 벡터논리 연산 명령과 스칼라 논리 연산 명령을 포함하는 신경망의 논리 연산을 완성하기 위한 논리 명령을 포함한다.

[0139] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 전용명령은 적어도 하나의 Cambricon 명령을 포함하고 상기 Cambricon 명령은 옵 코드와 피연산자를 포함하며 상기 Cambricon 명령은,

[0140] 수행과정을 제어하고 상기 Cambricon 제어 명령은 점프 명령과 조건부 분기 명령을 포함하는 Cambricon 제어 명령;

[0141] 로드 명령, 저장 명령, 운송 명령을포함하는 상이한 저장매체 사이의 데이터 전송을 완성하되, 여기서 상기 로드 명령은 데이터를 메인 메모리로부터 캐시에 로딩하기 위한 것이고, 상기 저장 명령은 데이터를 캐시로부터 메인 메모리에 저장하기 위한 것이며 운송 실시 명령은 캐시와 캐시 또는 캐시와 레지스터 또는 레지스터와 레지스터 사이에서 데이터를 운송하기 위한 것인 Cambricon 데이터 전송 명령;

[0142] Cambricon 매트릭스 연산 명령, Cambricon 벡터 연산 명령과 Cambricon 스칼라 연산 명령을 포함하는 신경망 산술 연산을 완성하되, 여기서 상기 Cambricon 매트릭스 연산 명령은 매트릭스 곱셈 벡터 연산, 벡터 곱셈 매트릭스 연산, 매트릭스 곱셈 스칼라 연산, 외적 연산, 매트릭스 덧셈 매트릭스 연산과 매트릭스 뺄셈 매트릭스 연산을 포함하는 신경망에서의 매트릭스 연산을 완성하기 위한 것이고 상기 Cambricon 벡터 연산 명령은 벡터 기본 연산, 벡터 초월함수 연산, 내적 연산, 벡터 랜덤 생성 연산과 벡터에서의 최대/최소치 연산을 포함하는 신경망에서의 벡터 연산을 완성하기 위한 것이며 Cambricon 스칼라 연산 명령은 스칼라 기본 연산과 스칼라 초월함수 연산을 포함하는 신경망에서의 스칼라 연산을 완성하기 위한 것인 Cambricon 연산 명령; 및

[0143] 신경망의 논리 연산을 위한 것으로 Cambricon 벡터논리 연산 명령과 Cambricon 스칼라 논리 연산 명령을 포함하되, 여기서 상기 Cambricon 벡터논리 연산 명령은 벡터 비교 연산, 벡터 논리 연산과 벡터 크기 합병 연산을 위한 것이고 여기서 벡터논리 연산은 그리고, 또는, 아니를 포함하며 상기 Cambricon 스칼라 논리 연산 명령은 스칼라 비교 연산과 스칼라 논리 연산 연산을 위한 것인Cambricon 논리 명령을 포함한다.

[0144] 본원 발명의 하나의 가능한 실시예에서 상기 Cambricon 데이터 전송 명령은 매트릭스, 벡터와 스칼라에서의 한 가지 또는 여러 가지 데이터 조합방식을 지지하고 상기 벡터 기본 연산은 벡터 더하기, 빼기, 곱하기, 나누기를 포함하며 벡터 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 스칼라 기본 연산은 스칼라 더하기, 빼기, 곱하기, 나누기를 포함하며 스칼라 초월함수는 다항식을 계수로 하는 다항방정식을 만족하지 않는 함수로서 지수함수, 로그함수, 삼각함수, 역삼각함수를 포함하고 상기 벡터 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 벡터논리 연산은 그리고, 또는, 아님을 포함하고 상기 스칼라 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하며 상기 스칼라 논리 연산은 그리고, 또는, 아님을 포함한다.

[0145] 신경망(neural network)은 이미 성공적인 응용을 획득하였으나 대규모적인 신경망의 파라미터는 저장에 매우 높

은 요구를 제출하였다. 한편으로 대량의 신경망 파라미터는 거대한 저장용량이 필요했다. 다른 한편으로 대량의 신경망 데이터를 액세스하는 것은 거대한 메모리 액세스 에너지 소모를 초래하였다.

[0146] 현재 신경망 파라미터를 저장하는 메모리는 오류 검출과 정정ECC(Error Correcting Code: 약칭 ECC)메모리인데 ECC메모리는 비록 데이터를 판독할 때 발생하는 오류를 정정할 수 있으나 ECC메모리가 별도의 저장용량 소비와 액세스 소비를 초래하게 된다. 신경망 알고리즘은 일정한 오차 허용 능력이 있으나 신경망의 모든 파라미터가 ECC 메모리를 사용하여 저장하게 되면 신경망의 오차 허용을 무시하게 되어 별도의 저장소비인 계산 소비와 액세스 저장 소비를 초래하게 되므로 신경망 오차 허용 능력과 결부하여 신경망 처리에 적합한 메모리를 선택하는가 하는 것은 하나의 해결이 필요한 과제가 되었다.

[0147] 본원 발명의 또 다른 양태는,

[0148] 데이터에서의 중요한 비트값을 저장하기 위한 정밀 저장 유닛;

[0149] 데이터에서의 중요하지 않은 비트값을 저장하기 위한 비정밀 저장 유닛을 포함하는 저장장치를 제공한다.

[0150] 본원 발명의 하나의 가능한 실시예에서 상기 정밀 저장 유닛은ECC메모리를 사용하고 상기 비정밀 저장 유닛은 비ECC메모리를 사용한다.

[0151] 본원 발명의 하나의 가능한 실시예에서 상기 데이터는 신경망 파라미터로서 입력 뉴런, 가중치와 출력 뉴런을 포함하고 상기 정밀 저장 유닛은 입력 뉴런의 중요한 비트값, 출력 뉴런의 중요한 비트값과 가중치의 중요한 비트값을 저장하기 위한 것이며 상기 비정밀 저장 유닛은 입력 뉴런의 중요하지 않은 비트값, 출력 뉴런의 중요하지 않은 비트값과 가중치의 중요하지 않은 비트값을 저장하기 위한 것이다.

[0152] 본원 발명의 하나의 가능한 실시예에서 상기 데이터는 부동 소수점 데이터와 고정 소수점 데이터를 포함하되, 상기 부동 소수점 데이터에서의 부호 비트와 지수 부분는 중요한 비트값이고 근 부분은 중요하지 않은 비트값이며 상기 고정 소수점 데이터에서의 부호 비트와 수치 부분의 앞의 x비트는 중요한 비트값이고 수치 부분의 나머지 비트는 중요하지 않은 비트값이며 여기서 x는 0보다 크거나 같고 m보다 작은 자연수이고 m는 데이터의 총 비트이다.

[0153] 본원 발명의 하나의 가능한 실시예에서 상기 ECC메모리는 ECC 체크가 있는 DRAM과 ECC 체크가 있는 SRAM을 포함하고 상기 ECC 체크가 있는 SRAM은 6T SRAM을 사용하거나 또는 4T SRAM 또는 3T SRAM을 사용한다.

[0154] 본원 발명의 하나의 가능한 실시예에서 상기 비 ECC메모리는ECC 체크가 아닌 DRAM과 ECC 체크가 아닌 SRAM을 포함하고 상기 ECC 체크가 아닌 SRAM은 6T SRAM을 사용하거나 또는 4T SRAM 또는 3T SRAM을 사용한다.

[0155] 본원 발명의 하나의 가능한 실시예에서 상기 6T SRAM에 매 하나의 비트를 저장하는 저장 유닛은 6개의 MOS관을 포함하고 상기 4T SRAM에 매 하나의 비트를 저장하는 저장 유닛은 4개의 MOS관을 포함하며 상기 3T SRAM에 매 하나의 비트를 저장하는 저장 유닛은 3개의 MOS관을 포함한다.

[0156] 본원 발명의 하나의 가능한 실시예에서 상기 4개의 MOS관은 제1 MOS관, 제2 MOS관, 제3 MOS관과 제4 MOS관을 포함하되, 제1 MOS관과 제2 MOS관은 게이팅을 위한 것이고 제3 MOS관과 제4 MOS관은 저장을 위한 것이며 여기서 제1 MOS관 그리드와 워드 라인(WL)은 전기적으로 연결되고 소스 전극과 비트 라인(BL)은 전기적으로 연결되며 제2 MOS관 그리드와 워드 라인(WL)은 전기적으로 연결되고 소스 전극과 비트 라인(BLB)은 전기적으로 연결되며 제3 MOS관 그리드는 제4 MOS관소스 전극과 제2 MOS관과 드레인 연결됨과 동시에 저항(R2)을 통해 작동 전압과 연결되고 제3 MOS관은 드레인 접지되며 제4 MOS관 그리드는 제3 MOS관소스 전극과 제1 MOS관과 드레인 연결됨과 동시에 저항(R1)을 통해 작동 전압과 연결되고 제4 MOS관은 드레인 접지되며 WL은 저장 유닛의 게이팅 액세스를 제어하고 BL은 저장 유닛의 판독 기록을 진행한다.

[0157] 본원 발명의 하나의 가능한 실시예에서 상기 3개의 MOS관은 제1 MOS관, 제2 MOS관과 제3 MOS관을 포함하되, 제1 MOS관은 게이팅을 위한 것이고 제2 MOS관과 제3 MOS관은 저장을 위한 것이며 여기서 제1 MOS관 그리드와 워드 라인(WL)은 전기적으로 연결되고 소스 전극과 비트 라인(BL)은 전기적으로 연결되며 제2 MOS관 그리드와 제3 MOS관소스 전극이 연결됨과 동시에 저항(R2)을 통해 작동 전압과 연결되고 제2 MOS관은 드레인 접지되며 제3 MOS관 그리드는 제2 MOS관소스 전극과 제1 MOS관과 드레인 연결됨과 동시에 저항(R1)을 통해 작동 전압과 연결되고 제3 MOS관은 드레인 접지되며 WL은 저장 유닛의 게이팅 액세스를 제어하고 BL은 저장 유닛의 판독 기록을 진행한다.

[0158] 본원 발명의 또 다른 양태는 데이터 처리장치를 제공하는데 이는,

[0159] 연산 유닛, 명령제어유닛과 상술한 저장장치를 포함하고 상기 저장장치는 입력한 명령과 연산 파라미터를 수신하고 연산 파라미터에서의 중요한 비트값과 명령을 정밀 저장 유닛에 저장하며 연산 파라미터에서의 중요하지 않은 비트값을 비정밀 저장 유닛에 저장하고 상기 명령제어유닛은 저장장치의 명령을 수신하고 제어정보를 디코딩하여 생성하며 상기 연산 유닛은 저장장치에서의 연산 파라미터를 수신하고 제어정보에 근거하여 연산을 진행하며 연산 결과를 저장장치에 전송한다.

[0160] 본원 발명의 하나의 가능한 실시예에서 상기 연산 유닛은 신경망 프로세서이다.

[0161] 본원 발명의 하나의 가능한 실시예에서 상기 연산 파라미터는 신경망 파라미터이고 상기 연산 유닛은 저장장치에서의 입력 뉴런과 가중치를 수신하며 제어정보에 따라 신경망 연산을 완성하여 출력 뉴런을 얻고 출력 뉴런을 저장장치에 전송한다.

[0162] 본원 발명의 하나의 가능한 실시예에서 상기 연산 유닛은 저장장치에서의 입력 뉴런의 중요한 비트값과 가중치의 중요한 비트값을 수신하여 계산하거나 또는 상기 연산 유닛은 중요한 비트값과 중요하지 않은 비트값을 이어 완성한 입력 뉴런과 가중치를 계산한다.

[0163] 본원 발명의 하나의 가능한 실시예에서는, 저장장치와 명령제어유닛 사이에 설치되어 전용 명령을 저장하기 위한 명령 캐시; 저장장치와 연산 유닛 사이에 설치되어 입력 뉴런을 캐시하되, 입력 뉴런 정밀 캐시와 입력 뉴런 비정밀 캐시를 포함하는 입력 뉴런 분층 캐시; 저장장치와 연산 유닛 사이에 설치되어 가중치 데이터를 캐시하되, 가중치 정밀 캐시와 가중치 비정밀 캐시를 포함하는 가중치 분층 캐시; 저장장치와 연산 유닛 사이에 설치되어 출력 뉴런을 캐시하되, 출력 뉴런 정밀 캐시와 출력 뉴런 비정밀 캐시를 포함하는 출력 뉴런 분층 캐시를 더 포함한다.

[0164] 본원 발명의 하나의 가능한 실시예에서는 상기 저장장치, 명령 캐시, 가중치 분층 캐시, 입력 뉴런 분층 캐시와 출력 뉴런 분층 캐시에서 데이터 또는 명령 판독 기록을 진행하기 위한 직접 데이터 액세스 유닛(DMA)를 더 포함한다.

[0165] 본원 발명의 하나의 가능한 실시예에서 상기 명령 캐시, 입력 뉴런 분층 캐시, 가중치 분층 캐시와 출력 뉴런 분층 캐시는 4T SRAM 또는 3T SRAM을 사용한다.

[0166] 본원 발명의 하나의 가능한 실시예에서는 입력된 데이터를 전처리하여 저장장치에 전송하기 위한 전처리모듈을 더 포함하되, 상기 전처리는 분할, 가우스 필터링, 이진화, 규칙화와 정규화를 포함한다.

[0167] 본원 발명의 하나의 가능한 실시예에서 상기 연산 유닛은 범용 연산 프로세서이다.

[0168] 본원 발명의 또 다른 양태는 상술한 데이터 처리장치를 포함하는 전자장치를 제공한다.

[0169] 본원 발명의 또 다른 양태는 데이터에서의 중요한 비트값을 정밀 저장하고 데이터에서의 중요하지 않은 비트값에 대해 비정밀 저장하는 단계를 포함하는 저장방법을 제공한다.

[0170] 본원 발명의 하나의 가능한 실시예에서 데이터에서의 중요한 비트값을 정밀 저장하는 상기 단계는 구체적으로, 데이터의 중요한 비트값을 추출하고 상기 데이터에서의 중요한 비트값을 ECC 메모리에 저장하여 정밀 저장하는 단계를 포함한다.

[0171] 본원 발명의 하나의 가능한 실시예에서 데이터에서의 중요하지 않은 비트값에 대해 비정밀 저장하는 상기 단계는 구체적으로, 데이터의 중요하지 않은 비트값을 추출하고 상기 데이터에서의 중요하지 않은 비트값을 비 ECC 메모리에 저장하여 비정밀 저장하는 단계를 포함한다.

[0172] 본원 발명의 하나의 가능한 실시예에서 상기 데이터는 신경망 파라미터로서 입력 뉴런, 가중치와 출력 뉴런을 포함하고 입력 뉴런의 중요한 비트값, 출력 뉴런의 중요한 비트값과 가중치의 중요한 비트값을 정밀 저장하며 입력 뉴런의 중요하지 않은 비트값, 출력 뉴런 의 중요하지 않은 비트값과 가중치의 중요하지 않은 비트값에 대해 비정밀 저장한다.

[0173] 본원 발명의 하나의 가능한 실시예에서 상기 데이터는 부동 소수점 데이터와 고정 소수점 데이터를 포함하되, 상기 부동 소수점 데이터에서의 부호 비트와 지수 부분은 중요한 비트값이고 근 부분은 중요하지 않은 비트값이며 상기 고정 소수점 데이터에서의 부호 비트와 수치 부분의 앞 x비트는 중요한 비트값이고 수치 부분의 나머지 비트는 중요하지 않은 비트값이며 여기서 x는 0보다 크거나 같고 m보다 작은 자연수이고 m는 파라미터의 총 비트이다.

[0174] 본원 발명의 하나의 가능한 실시예에서 상기 ECC메모리는 ECC 체크가 있는 DRAM와 ECC 체크가 있는 SRAM을 포함하고 상기 ECC 체크가 있는 SRAM은 6T SRAM, 4T SRAM 또는 3T SRAM을 사용한다.

[0175] 본원 발명의 하나의 가능한 실시예에서 상기 비 ECC메모리는ECC 체크가 아닌 DRAM와 ECC 체크가 아닌 SRAM을 포함하고 상기 ECC 체크가 아닌 SRAM은 6T SRAM, 4T SRAM 또는 3T SRAM을 사용한다.

[0176] 본원 발명의 또 다른 양태는,

[0177] 명령과 파라미터를 수신하고 상기 파라미터에서의 중요한 비트값과 명령을 정밀 저장하며 파라미터에서의 중요하지 않은 비트값에 대해 비정밀 저장하는 단계; 명령을 수신하고 명령을 제어정보로 디코딩하여 생성하는 단계; 파라미터를 수신하고 제어정보에 따라 연산하며 연산 결과를 저장하는 단계를 포함하는 데이터 처리방법을 제공한다.

[0178] 본원 발명의 하나의 가능한 실시예에서 상기 연산은 신경망 연산이고 상기 파라미터는 신경망 파라미터이다.

[0179] 본원 발명의 하나의 가능한 실시예에서 파라미터를 수신하고 제어정보에 따라 연산하며 연산 결과를 저장하는 상기 단계는, 입력 뉴런과 가중치를 수신하고 제어정보에 따라 신경망 연산을 완성하여 출력 뉴런을 얻으며 출력 뉴런을 저장 또는 출력하는 단계를 포함한다.

[0180] 본원 발명의 하나의 가능한 실시예에서 입력 뉴런과 가중치를 수신하고 제어정보에 따라 신경망 연산을 완성하여 출력 뉴런을 얻는 상기 단계는, 입력 뉴런의 중요한 비트값과 가중치의 중요한 비트값을 수신하여 계산하는 단계; 또는, 중요한 비트값과 중요하지 않은 비트값을 이어 완성한 입력 뉴런과 가중치를 수신하여 계산하는 단계를 포함한다.

[0181] 본원 발명의 하나의 가능한 실시예에서 상기 데이터 처리방은, 전용 명령을 캐시하는 단계; 입력 뉴런에 대해 정밀 캐시와 비정밀 캐시를 진행하는 단계; 가중치 데이터에 대해 정밀 캐시와 비정밀 캐시를 진행하는 단계; 출력 뉴런에 대해 정밀 캐시와 비정밀 캐시를 진행하는 단계를 더 포함한다.

[0182] 본원 발명의 하나의 가능한 실시예에서 상기 연산은 범용 연산이다.

[0183] 본원 발명의 하나의 가능한 실시예에서 명령과 파라미터를 수신하고 파라미터에서의 중요한 비트값과 명령을 저장하여 정밀 저장을 진행하며 파라미터에서의 중요하지 않은 비트값에 대해 비정밀 저장하는 상기 단계 전에 입력 데이터를 전처리하여 저장하는 단계를 더 포함하되, 상기 전처리는 분할, 가우스 필터링, 이진화, 규칙화와 정규화를 포함한다.

[0184] 본원 발명의 또 다른 양태는 4T SRAM 또는 3T SRAM으로서, 신경망 파라미터를 저장하기 위한 저장 유닛을 제공한다.

[0185] 본원 발명의 하나의 가능한 실시예에서 상기 4T SRAM에 매 하나의 비트를 저장하는 저장 유닛은 4개의 MOS관을 포함하고 상기 3T SRAM에 매 하나의 비트를 저장하는 저장 유닛은 3개의 MOS관을 포함한다.

[0186] 본원 발명의 하나의 가능한 실시예에서 상기 4개의 MOS관은 제1 MOS관, 제2 MOS관, 제3 MOS관과 제4 MOS관을 포함하되, 제1 MOS관과 제2 MOS관은 게이팅을 위한 것이고 제3 MOS관과 제4 MOS관은 저장을 위한 것이며 여기서 제1 MOS관 그리드와 워드 라인(WL)은 전기적으로 연결되고 소스 전극과 비트 라인(BL)은 전기적으로 연결되며 제2 MOS관 그리드와 워드 라인(WL)은 전기적으로 연결되고 소스 전극과 비트 라인(BLB)은 전기적으로 연결되며 제3 MOS관 그리드는 제4 MOS관소스 전극과 제2 MOS관과 드레인 연결됨과 동시에 저항(R2)을 통해 작동 전압과 연결되고 제3 MOS관은 드레인 접지되며 제4 MOS관 그리드는 제3 MOS관소스 전극과 제1 MOS관과 드레인 연결됨과 동시에 저항(R1)을 통해 작동 전압과 연결되고 제4 MOS관은 드레인 접지되며 WL은 저장 유닛의 게이팅 액세스를 제어하고 BL은 저장 유닛의 판독 기록을 진행한다.

[0187] 본원 발명의 하나의 가능한 실시예에서 상기 3개의 MOS관은 제1 MOS관, 제2 MOS관과 제3 MOS관을 포함하되, 제1 MOS관은 게이팅을 위한 것이고 제2 MOS관과 제3 MOS관은 저장을 위한 것이며 여기서 제1 MOS관 그리드와 워드 라인(WL)은 전기적으로 연결되고 소스 전극과 비트 라인(BL)은 전기적으로 연결되며 제2 MOS관 그리드와 제3 MOS관소스 전극이 연결됨과 동시에 저항(R2)을 통해 작동 전압과 연결되고 제2 MOS관은 드레인 접지되며 제3 MOS관 그리드는 제2 MOS관소스 전극과 제1 MOS관과 드레인 연결됨과 동시에 저항(R1)을 통해 작동 전압과 연결되고 제3 MOS관은 드레인 접지되며 WL은 저장 유닛의 게이팅 액세스를 제어하고 BL은 저장 유닛의 판독 기록을 진행한다.

[0188] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 파라미터는 입력 뉴런, 가중치와 출력 뉴런을 포함한다.

[0189] 작동 주파수의 향상과 반도체 공법의 끊임없는 발전과 더불어 칩의 전력 손실 문제는 이미 딥 서브 나노미터 집적회로에서의 하나의 중요한 고려 요소로 되었고 동적 전압 및 주파수 스케일링(Dynamic Voltage Frequency scaling, 약칭 DVFS)은 현재 반도체 분야에서 광범위하게 사용하는 동적 전압 및 주파수 스케일링 기수로서 DVFS기술은 구체적으로 동적 조절 칩의 동작 주파수와 전압(동일한 칩에 대하여 주파수가 높을 수록 필요한 전압이 더 높다)에 있어 에너지 절약의 목적에 도달한다. 그러나 선행기술에는 스마트 칩에 응용되는 동적 전압 조절 및 주파수 변조 방법과 상응하는 장치의 디자인이 결여하고 시나리오 정보를 응용하여 칩에 대한 전압 주파수의 미리 조절을 진행할 수 없다.

[0190] 본 발명의 또 다른 양태는 동적 전압 조절 및 주파수 변조 장치를 제공하는데 이는,

[0191] 연결된 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보를 실시간으로 수집하되, 상기 애플리케이션 시나리오 정보는 상기 칩이 신경망 연산을 통해 얻어지거나 또는 상기 칩과 연결된 센서가 수집한 정보인 정보 수집 유닛;

[0192] 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 칩이 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하는 전압 조절 및 주파수 변조 유닛을 포함한다.

[0193] 본원 발명의 하나의 가능한 실시예에서 상기 칩의 작동 상태 정보는 상기 칩의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은,

[0194] 상기 칩의 운행속도가 타겟 속도보다 클 경우 상기 칩에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하고 상기 타겟 속도는 사용자 수요를 만족시킬 경우의 상기 칩의 운행속도이다.

[0195] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 적어도 제1 유닛과 제2 유닛을 포함하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이며 상기 칩의 작동 상태 정보는 상기 제1 유닛의 운행속도와 제2 유닛의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 주파수 변조 및 전압 조절 유닛은,

[0196] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제1 유닛의 운행시간이 상기 제2 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제2 유닛에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 제2 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0197] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하되, 상기 주파수 변조 및 전압 조절 유닛은,

[0198] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제2 유닛의 운행시간이 상기 제1 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제1 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 제1 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0199] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 적어도 N개의 유닛을 포함하고 상기 칩의 작동 상태 정보는 상기 적어도 N개의 유닛에서의 적어도 S개의 유닛의 작동 상태 정보를 포함하며 상기 N은 1보다 큰 정수이고 상기 S는 N보다 작거나 같은 정수이며 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은,

[0200] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이며,

[0201] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛에서의 임의의 하나이다.

[0202] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은,

[0203] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓일 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0204] 본원 발명의 하나의 가능한 실시예에서 상기 칩의 애플리케이션 시나리오는 이미지 인식이고 상기 애플리케이션 시나리오 정보는 인식 대기 이미지에서의 오브젝트의 개수이며 상기 전압 주파수 규제 정보는 제6 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0205] 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 작을 경우 상기 칩에 상기 제6 전압 주파수 규제 정보를 발송하되, 상기 제6 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0206] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 오브젝트 레이블 정보이고 상기 전압 주파수 규제 정보는 제7 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0207] 상기 오브젝트 레이블 정보가 기설정 오브젝트 태그 집합에 속한다고 결정할 경우 상기 칩에 상기 제7 전압 주파수 규제 정보를 발송하되, 상기 제7 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 하기 위한 것이다.

[0208] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 음성 인식에 응용되고 상기 애플리케이션 시나리오 정보는 음성 입력 속도이며 상기 전압 주파수 규제 정보는 제8 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0209] 상기 음성 입력 속도가 제2 임계값보다 작을 경우 상기 칩에 제8 전압 주파수 규제 정보를 발송하되, 상기 제8 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0210] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 상기 칩이 음성 인식을 진행하여 얻은 키워드이고 상기 전압 주파수 규제 정보는 제9 전압 주파수 규제 정보를 포함하며 상기 주파수 변조 및 전압 조절 유닛은 또,

[0211] 상기 키워드가 기설정 키워드 집합일 경우 상기 칩에 상기 제9 전압 주파수 규제 정보를 발송하되, 상기 제9 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0212] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 기계 번역에 응용되고 상기 애플리케이션 시나리오 정보는 문자 입력 속도 또는 번역 대기 이미지에서의 문자의 수량이며 상기 전압 주파수 규제 정보는 제10 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0213] 상기 문자 입력 속도가 제3 임계값 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 작을 경우 상기 칩에 상기 제10 전압 주파수 규제 정보를 발송하되, 상기 제10 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0214] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 외부의 광도이고 상기 전압 주파수 규제 정보는 제11 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0215] 상기 외부의 광도가 제5 임계값보다 작을 경우 상기 칩에 상기 제11 전압 주파수 규제 정보를 발송하되, 상기 제11 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0216] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 이미지 뷰티에 응용되고 상기 전압 주파수 규제 정보는 제12 전압 주파수 규제 정보와 제13 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0217] 상기 애플리케이션 시나리오 정보가 안면 이미지일 경우 상기 칩에 상기 제12 전압 주파수 규제 정보를 발송하되, 상기 제12 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압을 저하시키기 위한 것이고;

[0218] 상기 애플리케이션 시나리오 정보가 안면 이미지가 아닐 경우 상기 칩에 상기 제13 전압 주파수 규제 정보를 발송하되, 상기 제13 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0219] 본원 발명의 또 다른 양태는 동적 전압 조절 및 주파수 변조 방법을 제공하는데 이는,

[0220] 상기 동적 전압 조절 및 주파수 변조 장치와 연결된 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보를 실시간으로 수집하되, 상기 애플리케이션 시나리오 정보는 상기 칩이 신경망 연산을 통해 얻어지거나 또는 상기 칩과 연결된 센서가 수집한 정보인 단계;

[0221] 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 칩이 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하기 위한 것인 단계를 포함한다.

[0222] 본원 발명의 하나의 가능한 실시예에서 상기 칩의 작동 상태 정보는 상기 칩의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0223] 상기 칩의 운행속도가 타겟 속도보다 클 경우 상기 칩에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하고 상기 타겟 속도는 사용자 수요를 만족시킬 경우의 상기 칩의 운행속도인 단계를 포함한다.

[0224] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 적어도 제1 유닛과 제2 유닛을 포함하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이며 상기 칩의 작동 상태 정보는 상기 제1 유닛의 운행속도와 제2 유닛의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송는 상기 단계는,

[0225] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제1 유닛의 운행시간이 상기 제2 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제2 유닛에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 제2 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0226] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0227] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제2 유닛의 운행시간이 상기 제1 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제1 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 제1 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0228] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 적어도 N개의 유닛을 포함하고 상기 칩의 작동 상태 정보는 상기 적어도 N개의 유닛에서의 적어도 S개의 유닛의 작동 상태 정보를 포함하며 상기 N은 1보다 큰 정수이고 상기 S는 N보다 작거나 같은 정수이며 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0229] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함하고,

[0230] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛에서의 임의의 하나이다.

[0231] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0232] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓일 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0233] 본원 발명의 하나의 가능한 실시예에서 상기 칩의 애플리케이션 시나리오는 이미지 인식이고 상기 애플리케이션

시나리오 정보는 인식 대기 이미지에서의 오브젝트의 개수이며 상기 전압 주파수 규제 정보는 제6 전압 주파수 규제 정보를 포함하고 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0234] 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 작을 경우 상기 칩에 상기 제6 전압 주파수 규제 정보를 발송하되, 상기 제6 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0235] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 오브젝트 레이블 정보이고 상기 전압 주파수 규제 정보는 제7 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0236] 상기 오브젝트 레이블 정보가 기설정 오브젝트 태그 집합에 속한다고 결정할 경우 상기 칩에 상기 제7 전압 주파수 규제 정보를 발송하되, 상기 제7 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 하기 위한 것인 단계를 더 포함한다.

[0237] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 음성 인식에 응용되고 상기 애플리케이션 시나리오 정보는 음성 입력 속도이며 상기 전압 주파수 규제 정보는 제8 전압 주파수 규제 정보를 포함하고 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0238] 상기 음성 입력 속도가 제2 임계값보다 작을 경우 상기 칩에 상기 제8 전압 주파수 규제 정보를 발송하되, 상기 제8 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0239] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 상기 칩이 음성 인식을 진행하여 얻은 키워드이고 상기 전압 주파수 규제 정보는 제9 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0240] 상기 키워드가 기설정 키워드 집합일 경우 상기 칩에 상기 제9 전압 주파수 규제 정보를 발송하되, 상기 제9 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0241] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 기계 번역에 응용되고 상기 애플리케이션 시나리오 정보는 문자 입력 속도 또는 번역 대기 이미지에서의 문자의 수량이며 상기 전압 주파수 규제 정보는 제10 전압 주파수 규제 정보를 포함하고 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0242] 상기 문자 입력 속도가 제3 임계값 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 작을 경우 상기 칩에 상기 제10 전압 주파수 규제 정보를 발송하되, 상기 제10 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0243] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 외부의 광도이고 상기 전압 주파수 규제 정보는 제11 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0244] 상기 외부의 광도가 제5 임계값보다 작을 경우 상기 칩에 상기 제11 전압 주파수 규제 정보를 발송하되, 상기 제11 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0245] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 이미지 뷰티에 응용되고 상기 전압 주파수 규제 정보는 제12 전압 주파수 규제 정보와 제13 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0246] 상기 애플리케이션 시나리오 정보가 안면 이미지일 경우 상기 칩에 상기 제12 전압 주파수 규제 정보를 발송하되, 상기 제12 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압을 저하시키기 위한 것인 단계;

[0247] 상기 애플리케이션 시나리오 정보가 안면 이미지가 아닐 경우 상기 칩에 상기 제13 전압 주파수 규제 정보를 발송하되, 상기 제13 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0248] 작동 주파수의 향상과 반도체 공법의 끊임없는 발전과 더불어 칩의 전력 손실 문제는 이미 딥 서브 나노미터 집적회로에서의 하나의 중요한 고려 요소로 되었고 동적 전압 및 주파수 스케일링(Dynamic Voltage Frequency scaling, 약칭 DVFS)은 현재 반도체 분야에서 광범위하게 사용하는 동적 전압 및 주파수 스케일링 기수로서 DVFS기술은 구체적으로 동적 조절 칩의 동작 주파수와 전압(동일한 칩에 대하여 주파수가 높을 수록 필요한 전압이 더 높다)에 있어 에너지 절약의 목적에 도달한다. 그러나 선행기술에는 콘볼루션 연산장치와 같은 스마트 칩에 응용되는 동적 전압 조절 및 주파수 변조 방법과 상응하는 장치의 디자인이 결여한다.

[0249] 본원 발명의 또 다른 양태는 동적 전압 조절 및 주파수 변조 장치, 명령 저장 유닛, 컨트롤러 유닛, 데이터 액세스 유닛, 인터커넥트 모듈, 메인 연산 모듈 및 N개의 서브 연산 모듈을 포함하되, 상기 N은 1보다 큰 정수인 콘볼루션 연산장치를 제공하는데 여기서,

[0250] 상기 명령 저장 유닛은 상기 데이터 액세스 유닛에 의해 판독된 명령을 저장하고,

[0251] 상기 컨트롤러 유닛은 상기 명령 저장 유닛으로부터 명령을 판독하여 상기 명령을 기타 모듈의 행동을 제어하는 컨트롤 신호로 디코딩하며 상기 기타 모듈은 상기 데이터 액세스 유닛, 상기 메인 연산 모듈과 상기 N개의 서브 연산 모듈을 포함하고,

[0252] 상기 데이터 액세스 유닛은 외부 주소 공간과 상기 콘볼루션 연산장치 사이의 데이터 또는 명령 판독 기록 작업을 실행하며,

[0253] 상기 N개의 서브 연산 모듈은 콘볼루션 신경망 알고리즘의 입력 데이터와 콘볼루션 커널의 콘볼루션 연산을 실현하고,

[0254] 상기 인터커넥트 모듈은 상기 메인 연산 모듈과 상기 서브 연산 모듈 사이에서 데이터를 전송하며,

[0255] 상기 메인 연산 모듈은 모든 입력 데이터의 중간 벡터를 중간 결과로 연결하고 상기 중간 결과에 대하여 후속 연산을 실행하며,

[0256] 상기 동적 전압 조절 및 주파수 변조 장치는 상기 콘볼루션 연산장치의 작동 상태 정보를 수집하고 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하며 상기 전압 주파수 규제 정보는 상기 콘볼루션 연산장치가 작동 전압 또는 작동 주파수를 조절하도록 지시한다.

[0257] 본원 발명의 하나의 가능한 실시예에서 상기 메인 연산 모듈은 또 중간 결과에 바이어스 데이터를 가한 후 활성화 작업을 수행한다.

[0258] 본원 발명의 하나의 가능한 실시예에서 상기 N개의 서브 연산 모듈은 구체적으로 동일한 입력 데이터와 각각의 콘볼루션 커널을 이용하여 각각의 출력 스칼라를 병행으로 산출한다.

[0259] 본원 발명의 하나의 가능한 실시예에서 상기 메인 연산 모듈이 사용하는 활성화 함수 active는 비선형 함수 sigmoid, tanh, relu, softmax에서의 임의의 하나 또는 비선형 함수이다.

[0260] 본원 발명의 하나의 가능한 실시예에서 상기 인터커넥트 모듈은 상기 메인 연산 모듈과 상기 N개의 서브 연산 모듈 사이의 연속되거나 이산화된 데이터의 데이터 통로를 구성하고 상기 인터커넥트 모듈은 트리 구조, 환형 구조, 메쉬 구조, 분급 인터커넥트 구조 및 버스 구조 에서의 임의의 한가지 구조이다.

[0261] 본원 발명의 하나의 가능한 실시예에서 상기 메인 연산 모듈은,

[0262] 상기 메인 연산 모듈의 계산 과정에서 사용되는 입력 데이터와 출력 데이터를 캐시하는 제1 저장 유닛;

[0263] 상기 메인 연산 모듈의 여러 가지 연산 기능을 완성하는 제1 연산 유닛;

[0264] 제1 연산 유닛이 제1 저장 유닛을 판독 기록하는 포트이로서 상기 제1 저장 유닛의 데이터 판독 기록의 일치성을 담보하고 상기 제1 저장 유닛으로부터 입력된 뉴런 벡터를 판독하며 상기 인터커넥트 모듈을 통해 상기 N개의 서브 연산 모듈에 전송하고 상기 인터커넥트 모듈로부터의 중간 결과 벡터를 상기 제1 연산 유닛에 전송하는 제1 데이터 의존관계 판정 유닛을 포함한다.

[0265] 본원 발명의 하나의 가능한 실시예에서 상기 N개의 서브 연산 모듈에서의 매 하나의 서브 연산 모듈은,

[0266] 상기 컨트롤러 유닛이 발송한 컨트롤 신호를 수신하고 산술 논리 연산을 진행하는 제2 연산 유닛;

[0267] 계산 과정에서 제2 저장 유닛과 제3 저장 유닛의 판독 기록작업을 수행하여 상기 제2 저장 유닛과 상기 제3 저

장 유닛의 판독 기록의 일치성을 담보하는 제2 데이터 의존관계 판정 유닛;

[0268] 입력 데이터 및 상기 연산 모듈에 의해 산출된 출력 스칼라를 캐시하는 상기 제2 저장 유닛;

[0269] 상기 서브 연산 모듈이 계산 과정에서 요구되는 콘볼루션 커널을 캐시하는 상기 제3 저장 유닛을 포함한다.

[0270] 본원 발명의 하나의 가능한 실시예에서 상기 제1 데이터 의존관계 판정 유닛과 상기 제2 데이터 의존관계 판정 유닛은,

[0271] 실행되지 않은 컨트롤 신호와 실행 중인 컨트롤 신호의 데이터 사이에 의존관계가 존재하는 지의 여부를 판정하고 만약 존재하지 않으면 상기 컨트롤 신호를 즉시 발송하도록 허용하며 그렇지 않으면 상기 컨트롤 신호가 의존하는 모든 컨트롤 신호가 모두 실행을 완성할 때까지 대기한 후 상기 컨트롤 신호를 발송하도록 허용하는 방식으로 판독 기록의 일치성을 담보한다.

[0272] 본원 발명의 하나의 가능한 실시예에서 상기 데이터 액세스 유닛은 외부 주소 공간으로부터 입력 데이터, 바이어스 데이터 및 콘볼루션 커널에서의 적어도 하나를 판독한다.

[0273] 본원 발명의 하나의 가능한 실시예에서 상기 동적 전압 조절 및 주파수 변조 장치는,

[0274] 상기 콘볼루션 연산장치의 작동 상태 정보를 실시간으로 수집하는 정보 수집 유닛;

[0275] 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션운행 속도운행 속도운행 속도운행 속도운행 속도보는 상기 콘볼루션 연산장치가 작동 전압 또는 작동 주파수를 조절하도록 지시하기 위한 것인 전압 조절 및 주파수 변조 유닛을 포함한다.

[0276] 의 작동 상태 정보는 상기 콘볼루션 연산장치의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은,

[0277] 상기 콘볼루션 연산장치의 운행속도가 타겟 속도보다 클 경우 상기 콘볼루션 연산장치에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 콘볼루션 연산장치로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하며 상기 타겟 속도는 사용자의 요구를 만족시킬 경우의 상기 콘볼루션 연산장치의 운행속도이다.

[0278] 본원 발명의 하나의 가능한 실시예에서 상기 콘볼루션 연산장치의 작동 상태 정보는 상기 데이터 액세스 유닛의 운행속도와 메인 연산 모듈의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 주파수 변조 및 전압 조절 유닛은 또,

[0279] 상기 데이터 액세스 유닛의 운행속도와 상기 메인 연산 모듈의 운행속도에 근거하여 상기 데이터 액세스 유닛의 운행시간이 상기 메인 연산 모듈의 운행시간을 초과한다고 결정할 경우 상기 메인 연산 모듈에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 메인 연산 모듈로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시한다.

[0280] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하고 상기 주파수 변조 및 전압 조절 유닛은 또,

[0281] 상기 데이터 액세스 유닛의 운행속도와 상기 메인 연산 모듈의 운행속도에 근거하여 상기 메인 연산 모듈의 운행시간이 상기 데이터 액세스 유닛의 운행시간을 초과한다고 결정할 경우 상기 데이터 액세스 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 데이터 액세스 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시한다.

[0282] 본원 발명의 하나의 가능한 실시예에서 상기 콘볼루션 연산장치작동 상태 정보는 명령 저장 유닛, 컨트롤러 유닛, 데이터 액세스 유닛, 인터커넥트 모듈, 메인 연산 모듈 및 N개의 서브 연산 모듈에서의 적어도 S개의 유닛/모듈의 작동 상태 정보를 포함하되, 상기 S는 1보다 크며 N+5보다 작거나 같은 정수이고 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은,

[0283] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이고,

[0284] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛/모듈에서의 임의의 하나이다.

[0285] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0286] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓인다고 결정할 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0287] 본원 발명의 또 다른 양태는 신경망 프로세서를 제공하는데 상기 신경망 프로세서는 상술한 콘볼루션 연산장치와 같은 것을 포함한다.

[0288] 본원 발명의 또 다른 양태는 전자장치를 제공하는데 상기 전자장치는 상술한 신경망 프로세서와 같은 것을 포함한다.

[0289] 본원 발명의 또 다른 양태는 단일층 콘볼루션 신경망의 순방향 연산을 실행하기 위한 방법을 제공하는데 이는 상술한 콘볼루션 연산장치에 응용되고,

[0290] 명령 저장 유닛의 첫번째 주소에 하나의 입출력 IO명령을 미리 저장하는 단계;

[0291] 연산이 시작되면 컨트롤러 유닛은 상기 명령 저장 유닛의 첫번째 주소로부터 상기 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 데이터 액세스 유닛은 외부 주소 공간으로부터 대응되는 모든 콘볼루션 신경망 연산 명령을 판독하여 이를 상기 명령 저장 유닛에 캐시하는 단계;

[0292] 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 외부 주소 공간으로부터 메인 연산 모듈에 필요한 모든 데이터를 상기 메인 연산 모듈의 제1 저장 유닛에 발송하는 단계;

[0293] 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 외부 주소 공간으로부터 서브 연산 모듈에 필요한 콘볼루션 커널 데이터를 판독하는 단계;

[0294] 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 CONFIG 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 콘볼루션 연산장치는 상기 층의 신경망 계산에 필요한 여러 가지 상수를 배치하는 단계;

[0295] 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 COMPUTE 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 메인 연산 모듈은 먼저 인터커넥트 모듈을 통해 콘볼루션 윈도우 내의 입력 데이터를 N개의 서브 연산 모듈에 발송하여 상기 N개의 서브 연산 모듈의 제2 저장 유닛에 저장한 후 명령에 따라 콘볼루션 윈도우를 이동하는 단계;

[0296] COMPUTE 명령에 의해 디코딩된 컨트롤 신호에 근거하여 상기 N개의 서브 연산 모듈의 연산 유닛은 제3 저장 유닛으로부터 콘볼루션 커널을 판독하고 상기 제2 저장 유닛으로부터 입력 데이터를 판독하며 입력 데이터와 콘볼루션 커널의 콘볼루션 연산을 완성하고 획득한 출력 스칼라를 상기 인터커넥트 모듈을 통해 리턴시키는 단계;

[0297] 상기 인터커넥트 모듈에서 상기 N개의 서브 연산 모듈에 의해 리턴된 출력 스칼라는 완전한 중간 벡터로 단계적으로 연결되는 단계;

[0298] 상기 메인 연산 모듈은 인터커넥트 모듈에 의해 리턴된 중간 벡터를 획득하고 콘볼루션 윈도우는 모든 입력 데이터를 가로 지르며 상기 메인 연산 모듈은 모든 리턴된 백터를 중간 결과로 연결하고 COMPUTE 명령에 의해 디코딩된 컨트롤 신호에 근거하여 제1 저장 유닛으로부터 바이어스 데이터를 판독하며 벡터 가산 유닛에 의해 중간 결과와 가산되어 바이어스 결과를 획득한 후 활성화 유닛에 의해 바이어스 결과를 활성화시키고 최종 출력 데이터를 상기 제1 저장 유닛에 다시 기입하는 단계;

[0299] 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 상기 제1 저장 유닛의 출력 데이터를 외부 주소 공간의 지정된 주소에 저장하여 연산을 종료하는 단계를 포함한다.

[0300] 본원 발명의 하나의 가능한 실시예에서 상기 방법은,

[0301] 상기 콘볼루션 연산장치의 작동 상태 정보를 실시간으로 수집하는 단계;

[0302] 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 콘볼루션 연산장치로 하여금 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하는 단계를 더 포함한다.

[0303] 본원 발명의 한가지 가능한 실시예에서 상기 콘볼루션 연산장치의 작동 상태 정보는 상기 콘볼루션 연산장치의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0304] 상기 콘볼루션 연산장치의 운행속도가 타겟 속도보다 클 경우 상기 콘볼루션 연산장치에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 콘볼루션 연산장치로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하고 상기 타겟 속도는 사용자의 요구를 만족시킬 경우의 상기 칩의 운행속도인 단계를 포함한다.

[0305] 본원 발명의 하나의 가능한 실시예에서 상기 콘볼루션 연산장치의 작동 상태 정보는 상기 데이터 액세스 유닛의 운행속도와 메인 연산 모듈의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0306] 상기 데이터 액세스 유닛의 운행속도와 상기 메인 연산 모듈의 운행속도에 근거하여 상기 데이터 액세스 유닛의 운행시간이 상기 메인 연산 모듈의 운행시간을 초과한다고 결정할 경우 상기 메인 연산 모듈에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 메인 연산 모듈로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하는 단계를 더 포함한다.

[0307] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하고 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0308] 상기 데이터 액세스 유닛의 운행속도와 상기 메인 연산 모듈의 운행속도에 근거하여 상기 메인 연산 모듈의 운행시간이 상기 데이터 액세스 유닛의 운행시간을 초과한다고 결정할 경우 상기 데이터 액세스 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 데이터 액세스 유닛으로 하여금 작동 주파수 또는 작동 전압을 저하시키도록 지시하는 단계를 더 포함한다.

[0309] 본원 발명의 하나의 가능한 실시예에서 상기 콘볼루션 연산장치의 작동 상태 정보는 명령 저장 유닛, 컨트롤러 유닛, 데이터 액세스 유닛, 인터커넥트 모듈, 메인 연산 모듈 및 N개의 서브 연산 모듈 중 적어도 S개의 유닛/모듈의 작동 상태 정보를 포함하고 상기 S는 1보다 크며 N+5보다 작거나 같은 정수이고 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하며 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0310] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함하고,

[0311] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛/모듈에서의 임의의 하나이다.

[0312] 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0313] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓일 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0314] 본원 발명의 다른 양태는 다층 콘볼루션 신경망의 순방향 연산을 실행하기 위한 방법을 제공하는데 이는,

[0315] 매 한 층에 대하여 상술한 다층 콘볼루션 신경망의 순방향 연산을 실행하기 위한 방법을 수행하고 윗 층의 콘볼루션 신경망 실행이 완료된 후 본 층의 연산 명령은 메인 연산 모듈에 저장된 윗 층의 출력 데이터 주소를 본 층의 입력 데이터 주소로 하며 명령에서의 콘볼루션 커널과 바이어스 데이터 주소는 본 층에 대응되는 주소로 변경되는 단계를 포함한다.

[0316] 빅 데이터 시대가 다가옴에 따라 데이터는 폭발적인 속도로 자라고 있고 거대한 량의 데이터는 정보를 휴대하고

사람들 사이에서 전송되며 이미지는 인류가 세계를 감지하는 시각적 기초로서 인류가 정보를 얻고 정보를 표달하며 정보를 전달하는 중요한 수단으로 되었다.

[0317] 선행기술에서는 이미지 압축을 통해 데이터량을 효과적으로 절감하고 이미지의 전송속도를 향상시킨다. 그러나 이미지를 압축한 후 원본 이미지의 모든 정보를 보존하기 어려워 어떻게 이미지 압축을 진행할 것인지는 여전히 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자들이 해결해야 할 기술적 과제로 되었다.

[0318] 본원 발명의 또 다른 양태는 이미지 압축방법을 제공하는데 이는,

[0319] 제1 해상도의 원본 이미지를 획득하되, 상기 원본 이미지는 압축 신경망의 압축 트레이닝 이미지 집합에서의 임의의 트레이닝 이미지이고 상기 원본 이미지의 태그 정보를 타겟 태그 정보로 사용하는 단계;

[0320] 타겟 모델에 의해 상기 원본 이미지를 압축하여 제2 해상도의 압축 이미지를 얻되, 상기 제2 해상도는 상기 제1 해상도보다 작고 상기 타겟 모델은 상기 압축 신경망의 현재의 신경망 모델인 단계;

[0321] 인식 신경망 모델에 의해 상기 압축 이미지를 인식하여 참고 태그 정보를 얻되, 상기 인식 신경망 모델은 인식 신경망 트레이닝이 완성될 때 대응되는 신경망 모델인 단계;

[0322] 상기 타겟 태그 정보와 상기 참고 태그 정보에 근거하여 손실함수를 획득하는 단계;

[0323] 상기 손실함수가 제1 임계값에 수렴되거나 또는 상기 압축 신경망의 현재의 트레이닝 횟수가 제2 임계값보다 크거나 같을 경우 상기 제1 해상도의 타겟 원본 이미지를 획득하고 상기 타겟 모델을 상기 압축 신경망 트레이닝이 완성될 때 대응되는 압축 신경망 모델로 사용하는 단계;

[0324] 상기 압축 신경망 모델에 의해 상기 타겟 원본 이미지를 압축하여 상기 제2 해상도의 타겟 압축 이미지를 얻는 단계를 포함한다.

[0325] 본원 발명의 하나의 가능한 실시예에서 상기 이미지 압축방법은,

[0326] 상기 손실함수가 상기 제1 임계값에 수렴되지 않거나 또는 상기 압축 신경망의 현재의 트레이닝 횟수가 상기 제2 임계값보다 작을 경우 상기 손실함수에 근거하여 상기 타겟 모델을 업데이트하여 업데이트 모델을 얻고 상기 업데이트 모델을 상기 타겟 모델로 사용하며 그 다음의 한 트레이닝 이미지를 상기 원본 이미지로 사용하여 제1 해상도의 원본 이미지를 획득하는 상기 단계를 수행하는 단계를 더 포함한다.

[0327] 본원 발명의 하나의 가능한 실시예에서 인식 신경망 모델에 의해 상기 압축 이미지를 인식하여 참고 태그 정보를 획득하는 단계는 구체적으로,

[0328] 상기 압축 이미지를 전처리하여 인식 대기 이미지를 얻는 단계;

[0329] 상기 인식 신경망 모델에 의해 상기 인식 대기 이미지를 인식하여 상기 참고 태그 정보를 얻는 단계를 포함한다.

[0330] 본원 발명의 하나의 가능한 실시예에서 상기 전처리는 사이즈 처리를 포함하고 상기 압축 이미지를 전처리하여 인식 대기 이미지를 얻는 상기 단계는 구체적으로,

[0331] 상기 압축 이미지의 이미지 크기가 상기 인식 신경망의 기본 이미지 크기보다 작을 경우 상기 기본 이미지 크기에 따라 상기 압축 이미지에 대해 픽셀 포인트 충진을 진행하여 상기 인식 대기 이미지를 얻는 단계를 포함한다.

[0332] 본원 발명의 하나의 가능한 실시예에서 상기 압축 트레이닝 이미지 집합은 적어도 인식 트레이닝 이미지 집합을 포함하고 상기 방법은,

[0333] 상기 인식 트레이닝 이미지 집합을 사용하여 상기 인식 신경망을 트레이닝하여 상기 인식 신경망 모델을 얻되, 상기 인식 트레이닝 이미지 집합에서의 매 하나의 트레이닝 이미지는 적어도 상기 타겟 태그 정보의 유형과 일치한 태그 정보를 포함하는 단계를 더 포함한다.

[0334] 본원 발명의 하나의 가능한 실시예에서 상기 압축 신경망 모델에 의해 상기 타겟 원본 이미지를 압축하여 상기 제2 해상도의 타겟 압축 이미지를 얻는 단계 다음에 상기 방법은,

[0335] 상기 인식 신경망 모델에 의해 상기 타겟 압축 이미지를 압축하여 상기 타겟 원본 이미지의 태그 정보를 얻고 상기 타겟 원본 이미지의 태그 정보를 저장하는 단계를 더 포함한다.

[0336] 본원 발명의 하나의 가능한 실시예에서 상기 압축 트레이닝 이미지 집합은 다수의 차원을 포함하고 상기 타겟 모델에 의해 상기 원본 이미지를 압축하여 제2 해상도의 압축 이미지를 얻는 단계는,

[0337] 상기 타겟 모델에 의해 상기 원본 이미지를 인식하여 다수의 이미지 정보를 얻고 매 하나의 차원은 하나의 이미지 정보와 대응되는 단계;

[0338] 상기 타겟 모델과 상기 다수의 이미지 정보에 의해 상기 원본 이미지를 압축하여 상기 압축 이미지를 얻는 단계를 포함한다.

[0339] 본원 발명의 또 다른 양태는 프로세서 및 상기 프로세서와 연결된 메모리를 포함하는 이미지 압축 장치를 제공하는데 여기서,

[0340] 상기 메모리는 제1 임계값, 제2 임계값, 압축 신경망의 현재의 신경망 모델과 트레이닝 횟수, 상기 압축 신경망의 압축 트레이닝 이미지 집합과 상기 압축 트레이닝 이미지 집합에서의 매 하나의 트레이닝 이미지의 태그 정보, 인식 신경망 모델, 압축 신경망 모델을 저장하고 상기 압축 신경망의 현재의 신경망 모델을 타겟 모델로 사용하며 상기 압축 신경망 모델은 상기 압축 신경망 트레이닝이 완성될 때 대응되는 타겟 모델이고 상기 인식 신경망 모델은 인식 신경망 트레이닝이 완성될 때 대응되는 신경망 모델이며;

[0341] 상기 프로세서는 제1 해상도의 원본 이미지를 획득하되, 상기 원본 이미지는 상기 압축 트레이닝 이미지 집합에서의 임의의 한 트레이닝 이미지이고 상기 원본 이미지의 태그 정보를 타겟 태그 정보로 사용하며 상기 타겟 모델에 의해 상기 원본 이미지를 압축하여 제2 해상도의 압축 이미지를 얻되, 상기 제2 해상도는 상기 제1 해상도보다 작고 상기 인식 신경망 모델에 의해 상기 압축 이미지를 인식하여 참고 태그 정보를 획득하며 상기 타겟 태그 정보와 상기 참고 태그 정보에 근거하여 손실함수를 획득하고 상기 손실함수가 상기 제1 임계값에 수렴되거나 또는 상기 트레이닝 횟수가 상기 제2 임계값보다 크거나 같을 경우 상기 제1 해상도의 타겟 원본 이미지를 획득하여 상기 타겟 모델을 상기 압축 신경망 모델로 확인하며 상기 압축 신경망 모델에 의해 상기 타겟 원본 이미지를 압축하여 상기 제2 해상도의 타겟 압축 이미지를 얻는다.

[0342] 본원 발명의 하나의 가능한 실시예에서 상기 프로세서는 또 상기 손실함수가 상기 제1 임계값에 수렴되지 않거나 또는 상기 트레이닝 횟수가 상기 제2 임계값보다 작을 경우 상기 손실함수에 근거하여 상기 타겟 모델을 업데이트하여 업데이트 모델을 얻고 상기 업데이트 모델을 상기 타겟 모델로 사용하며 그 다음의 한 트레이닝 이미지를 상기 원본 이미지로 사용하여 제1 해상도의 원본 이미지를 획득하는 상기 단계를 수행한다.

[0343] 본원 발명의 하나의 가능한 실시예에서 상기 프로세서는 구체적으로 상기 압축 이미지를 전처리하여 인식 대기 이미지를 얻고 상기 인식 신경망 모델에 의해 상기 인식 대기 이미지를 인식하여 상기 참고 태그 정보를 얻는다.

[0344] 본원 발명의 하나의 가능한 실시예에서 상기 전처리는 사이즈 처리를 포함하고 상기 메모리는 상기 인식 신경망의 기본 이미지 크기를 저장하는데 더 사용되며 상기 프로세서는 구체적으로 상기 압축 이미지의 이미지 크기가 상기 기본 이미지 크기보다 작을 경우 상기 기본 이미지 크기에 따라 상기 압축 이미지에 대해 픽셀 포인트 충진을 진행하여 상기 인식 대기 이미지를 얻는다.

[0345] 본원 발명의 하나의 가능한 실시예에서 상기 압축 트레이닝 이미지 집합은 적어도 인식 트레이닝 이미지 집합을 포함하고 상기 프로세서는 또 상기 인식 트레이닝 이미지 집합을 사용하여 상기 인식 신경망을 트레이닝하여 상기 인식 신경망 모델을 얻되, 상기 인식 트레이닝 이미지 집합에서의 매 하나의 트레이닝 이미지는 적어도 상기 타겟 태그 정보의 유형과 일치한 태그 정보를 포함한다.

[0346] 본원 발명의 하나의 가능한 실시예에서 상기 프로세서는 또 상기 인식 신경망 모델에 의해 상기 타겟 압축 이미지를 인식하여 상기 타겟 원본 이미지의 태그 정보를 얻고, 상기 메모리는 또 상기 타겟 원본 이미지의 태그 정보를 저장한다.

[0347] 본원 발명의 하나의 가능한 실시예에서 상기 압축 트레이닝 이미지 집합은 다수의 차원을 포함하고 상기 프로세서는 구체적으로 상기 타겟 모델에 의해 상기 원본 이미지를 인식하여 다수의 이미지 정보를 얻고 매 하나의 차원은 하나의 이미지 정보와 대응되며 상기 타겟 모델과 상기 다수의 이미지 정보에 의해 상기 원본 이미지를 압축하여 상기 압축 이미지를 얻는다.

[0348] 본원 발명의 다른 양태는 프로세서,메모리, 통신 인터페이스 및 하나 또는 다수의 프로그램을 포함하는 다른 한 가지 전자기기를 제공하는데 여기서 상기 하나 또는 다수의 프로그램은 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 배치되며 상기 프로그램은 상술한 바와 같은 이미지 압축방법에서 설명한 일부 또는 모든

단계의 명령을 포함한다.

[0349] 본원 발명의 다른 양태는 한가지 컴퓨터 판독 가능 저장매체를 제공하는데 상기 컴퓨터 저장매체에는 컴퓨터 프로그램이 저장되고 상기 컴퓨터 프로그램은 프로그램 명령을 포함하며 상기 프로그램 명령은 프로세서에 의해 실행될 경우 상기 프로세서로 하여금 상술한 이미지 압축방법을 수행하도록 한다.

[0350] 본원 발명이 제공하는 처리방법 및 장치, 연산방법 및 장치는 선행기술과 비교하여 적어도 아래와 같은 장점을 구비한다.

[0351] 1. 양자화의 방법을 사용하여 신경망의 뉴런과 가중치를 양자화하고 가중치 사전과 가중치 코드북을 사용하여 양자화를 진행한 후의 가중치를 나타내며 뉴런 사전과 뉴런 코드북을 사용하여 양자화를 진행한 후의 뉴런을 나타낸 다음 신경망에서의 연산을 테이블 조사 동작으로 전환함으로써 신경망 파라미터의 저장량을 절감하고 메모리 액세스 에너지 소모와 계산 에너지 소모를 절감한다. 신경망 프로세서에는 검색 테이블에 의한 계산방법이 집적되어 테이블 조사 동작을 최적화하고 구조를 간략화하여 신경망 메모리 액세스 에너지 소모와 계산 에너지 소모를 절감함과 동시에 연산의 다원화를 더 실현할 수 있다.

[0352] 2. 신경망에 대해 리트레이닝을 진행할 수 있고 리트레이닝 할 경우 가중치 사전을 트레이닝 할 필요가 없이 코드북만 트레이닝하면 되므로 리트레이닝 작업을 간략화한다.

[0353] 3. 국부 양자화된 다층 인공 신경망에 대해 연산한 신경망 전용 명령과 원활한 연산 유닛을 사용하여 중앙 처리장치(CPU)와 그래픽 처리장치(GPU)의 연산기능의 부족함과 프런트 엔드 디코딩 지출이 큰 문제를 해결함으로써 다층 인공 신경망 연산 알고리즘에 대한 지지를 향상시켰다.

[0354] 4. 다층 인공 신경망 연산 알고리즘에 대해 사용한 전용 온-칩 캐시를 통해 입력 뉴런과 가중치 데이터의 중요성을 충분히 발굴하여 메모리에서 이러한 데이터를 반복적으로 판독하는 것을 방지하고 메모리 액세스 대역폭을 저하시켜 메모리 대역폭이 다층 인공 신경망 연산 및 그의 트레이닝 알고리즘에 가져온 기능적 난관의 문제를 방지하였다.

도면의 간단한 설명

[0355] 도 1a는 본원 발명의 실시예에서 제공하는 처리방법의 흐름모식도이다.

도 1b는 본원 발명의 실시예에서 제공하는 가중치를 양자화하는 과정모식도이다.

도 1c은 본원 발명의 실시예에서 제공하는 입력 뉴런을양자화하는 과정모식도이다.

도 1d는 본원 발명의 실시예에서 제공하는 연산 코드북을 결정하는 과정모식도이다.

도 1e는 본원 발명의 실시예에서 제공하는 처리장치의 구조모식도이다.

도 1f은 본원 발명의 실시예에서 제공하는 연산장치의 구조모식도이다.

도 1g은 본원 발명의 구체적인 실시예에서 제공하는 연산장치의 구조모식도이다.

도 1h은 본원 발명의 실시예에서 제공하는 연산방법의 흐름모식도이다.

도 1i는 본원 발명의 실시예에서 제공하는 구체적인 실시예的다른 연산방법의 흐름모식도이다.

도 2a은 본원 발명의 실시예에서 제공하는 분층 저장장치구조모식도이다.

도 2b는 본원 발명의 실시예에서 제공하는 4T SRAM저장 유닛의 구조모식도이다.

도 2c은 본원 발명의 실시예에서 제공하는 3T SRAM저장 유닛의 구조모식도이다.

도 2d는 본원 발명의 실시예에서 제공하는 데이터 처리장치의 구조모식도이다.

도 2e는 본원 발명의 실시예에서 제공하는 다른 데이터 처리장치의 구조모식도이다.

도 2f은 본원 발명의 실시예에서 제공하는 데이터 저장방법의 흐름도이다.

도 2g은 본원 발명의 실시예에서 제공하는 데이터 처리방법의 흐름도이다.

도 3a은 본원 발명의 실시예에서 제공하는 동적 전압 조절 및 주파수 변조 장치의 구조모식도이다.

도 3b는 본원 발명의 실시예에서 제공하는 동적 전압 조절 및 주파수 변조애플리케이션 시나리오 모식도이다.

도 3c은 본원 발명의 실시예에서 제공하는 다른 동적 전압 조절 및 주파수 변조애플리케이션 시나리오 모식도이다.

도 3d는 본원 발명의 실시예에서 제공하는 다른 동적 전압 조절 및 주파수 변조애플리케이션 시나리오 모식도이다.

도 3e는 본원 발명의 실시예에서 제공하는 인터커넥트 모듈(4)의 실시형태의 모식도이다.

도 3f은 본원 발명의 실시예에서 제공하는 콘볼루션 신경망의 순방향 연산을 수행하기 위한 장치에서의 메인 연산 모듈(5)의 구조의 예시적 블럭도이다.

도 3g은 본원 발명의 실시예에서 제공하는 콘볼루션 신경망의 순방향 연산을 수행하기 위한 장치에서의 서브 연산 모듈(6)의 구조의 예시적 블럭도이다.

도 3h은 본원 발명의 실시예에서 제공하는 동적 전압 조절 및 주파수 변조 방법의 흐름모식도이다.

도 4a은 본원 발명의 실시예에서 제공하는 콘볼루션 연산장치의 구조모식도이다.

도 4b는 본원 발명의 실시예에서 제공하는 콘볼루션 연산장치에서의 메인 연산 모듈의 구조의 예시적 블럭도이다.

도 4c은 본원 발명의 실시예에서 제공하는 콘볼루션 연산장치에서의 서브 연산 모듈의 구조의 예시적 블럭도이다.

도 4d는 본원 발명의 실시예에서 제공하는 콘볼루션 연산장치에서의 동적 전압 조절 및 주파수 변조 장치의 구조의 예시적 블럭도이다.

도 4e는 본원 발명의 실시예에서 제공하는 인터커넥트 모듈(4)의 실시형태의 모식도이다.

도 4f는 본원 발명의 실시예에서 제공하는 다른 콘볼루션 연산장치의 구조모식도이다.

도 4g는 본원 발명의 실시예에서 제공하는 단일층 콘볼루션 신경망의 순방향 연산방법을 수행하기 위한 흐름모식도이다.

도 5a는 본원 발명의 실시예에서 제공하는 신경망의 연산 모식도이다.

도 5b는 본원 발명의 실시예에서 제공하는 이미지 압축방법의 흐름모식도이다.

도 5c는 본원 발명의 실시예에서 제공하는 사이즈 처리방법의 시나리오 모식도이다.

도 5d는 본원 발명의 실시예에서 제공하는 단일층 신경망 연산방법의 흐름모식도이다.

도 5e는 본원 발명의 실시예에서 제공하는 압축 신경망의 역방향 트레이닝을 수행하기 위한 장치의 구조모식도이다.

도 5f는 본원 발명의 실시예에서 제공하는 H트리 모듈의 구조모식도이다.

도 5g는 본원 발명의 실시예에서 제공하는 메인 연산 모듈의 구조모식도이다.

도 5h는 본원 발명의 실시예에서 제공하는 연산 모듈의 구조모식도이다.

도 5i는 본원 발명의 실시예에서 제공하는 압축 신경망의 역방향 트레이닝의 예시적 블럭도이다.

도 5j는 본원 발명의 실시예에서 제공하는 이미지 압축방법의 흐름모식도이다.

도 5k는 본원 발명의 실시예에서 제공하는 전자장치의 구조모식도이다.

발명을 실시하기 위한 구체적인 내용

[0356] 선행기술에서 신경망의 데이터를 처리할 경우 아주 큰 계산량이 신경망의 애플리케이션으로 하여금 지장을 받도록 하는 기술적 단점에 기반하여 본원 발명은 처리방법 및 장치, 연산방법 및 장치를 제공한다. 여기서 처리방법 및 장치는 입력 뉴런과 가중치 이 두 가지 데이터를 양자화하여 층 간, 세그먼트 간의 데이터 사이의 유사성 및 층 내, 세그먼트 내의 대이터 국부 유사성을 각각 발굴하여 이 두 가지 데이터의 분포 특성을 발굴하도록 함으로써 낮은 비트의 양자화를 실현하고 매 하나의 데이터를 나타내기 위한 비트수를 절감함으로써 데이터 저장 지출과 액세스 지출을 저하시킨다. 처리방법 및 장치는 양자화 후의 뉴런과 가중치에 대해 테이블 조사 동작을

통해 양자의 연산작업을 실현하여 신경망 메모리 액세스 에너지 소모와 계산 에너지 소모를 절감한다.

[0357] 본원 발명에서 제출한 입력 뉴런과 출력 뉴런은 전반 신경망의 입력층에서의 뉴런과 출력층에서의 뉴런을 가리키는 것이 아니라 네트워크에서의 임의의 인접한 두 층을 말하는 바, 네트워크 피드 포워드 연산 하층에서의 뉴런이 입력 뉴런이고 네트워크 피드 포워드 연산 상층에서의 뉴런이 출력 뉴런이다. 콘볼루션 신경망을 예로 하면 하나의 콘볼루션 신경망이 L층을 구비한다고 할 경우 $K=1,2,...,L-1$인데 제K층과 제K+1층에 있어서, 제K층을 입력층이라 부르고 여기서 뉴런은 상기 입력 뉴런이며 제K+1층을 출력층이라고 부르고 여기서 뉴런은 상기 출력 뉴런이다. 즉 맨 위층을 제외하고 매 한 층마다 모두 입력층이 될 수 있고 그 아래층이 대응하는 출력층이 된다.

[0358] 본원 발명의 목적, 기술적 해결수단과 장점이 더 뚜렷하고 명확해지도록 하기 위하여 아래에는 구체적인 실시예와 결부함과 동시에 도면을 참조하여 본원 발명에 대해 진일보로 상세히 설명한다.

[0359] 도 1a을 참조하면 도 1a은 본원 발명의 실시예에서 제공하는 처리방법의 흐름모식도 인 바, 도 1a에 도시된 바와 같이 처리방법은 다음과 같은 단계를 포함한다.

[0360] 단계 S1, 가중치와 입력 뉴런을 각각 양자화하여 가중치 사전, 가중치 코드북, 뉴런 사전과 뉴런 코드북을 결정;

[0361] 여기서 가중치를 양자화하는 과정은 구체적으로,

[0362] 가중치를 그루핑하되, 매 그룹의 가중치에 클러스터링 알고리즘을 이용하여 클러스터링 작업을 진행하고 한 그룹의 가중치를 m타입으로 분류하는데 m는 자연수이고 매 타입의 가중치는 하나의 가중치 인덱스와 대응하여 가중치 사전을 결정하며, 여기서 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 가중치 위치는 가중치가 신경망 구조에서의 위치를 가리키는 단계;

[0363] 매 타입의 모든 가중치를 하나의 중심 가중치로 대체하여 가중치 코드북을 결정하되, 상기 가중치 코드북은 가중치 인덱스와 중심 가중치를 포함하는 단계를 포함한다.

[0364] 도 1b를 참조하면 도 1b는 본원 발명의 실시예에서 제공하는 가중치를 양자화하는 과정모식도인 바, 도 1b에 도시된 바와 같이 기설정된 그루핑 전략에 따라 가중치를 그루핑하여 순서대로 배열된 가중치 매트릭스를 얻는다. 다음 그루핑 후의 가중치 매트릭스에 대해 그룹내의 샘플링 및 클러스터링 작업을 진행하고 수치가 근접한 가중치를 동일 타입으로 구분하며 손실함수에 근거하여 4개 타입의 중심 가중치 1.50, -0.13, -1.3과 0.23을 산출하여 4개 타입의 가중치와 각각 대응한다. 이미 알고 있는 가중치 코드북에서 중심 가중치가 -1.3인 타입의 가중치 인덱스는 00, 중심 가중치가 -0.13인 타입의 가중치 인덱스는 01, 중심 가중치가 0.23인 타입의 가중치 인덱스는 10, 중심 가중치가 1.50인 타입의 가중치 인덱스는 11이다. 이 외에 4개의 가중치가 대응하는 가중치 인덱스(00, 01, 10과 11)를 각각 더 사용하여 대응 타입에서의 가중치를 각각 나타냄으로써 가중치 사전을 얻는다. 유의해야 할 것은 가중치 사전은 가중치 위치, 즉 가중치가 신경망 구조에서의 위치를 더 포함하는데 가중치 사전에서 가중치 위치는 그 중의 제p행 제q열의 좌표, 즉(p, q)를 가리키고 본 실시예에서 $1\leq p\leq 4$, $1\leq q\leq 4$이다.

[0365] 보다시피, 상기 양자화 과정은 신경망 층간 가중치의 유사성 및 층내 가중치의 국부 유사성을 충분히 발굴하여 신경망의 가중치 분포 특성을 얻음으로써 낮은 비트 양자화를 진행하여 매 하나의 가중치를 나타내기 위한 비트수를 절감하여 가중치 저장 지출과 액세스 지출을 저하시킨다.

[0366] 선택적으로, 상기 기설정된 그루핑 전략은 신경망의 모든 가중치를 한 그룹으로 귀납하는 한 그룹으로의 그루핑; 신경망에서의 모든 합성곱 층의 가중치, 모든 풀 연결층의 가중치와 모든 장단기 모모리 네트워크층의 가중치를 각각 한 그룹으로 구획하는 레이어 타입 그루핑; 신경망에서의 하나 또는 다수의 합성곱 층의 가중치, 하나 또는 다수의 풀 연결층의 가중치와 하나 또는 다수의 장단기 메모리 네트워크층의 가중치를 각각 한 그룹으로 구획하는 층간 그루핑; 및 신경망의 한 층 내의 가중치를 분할하고 분할한 후의 매 하나의 부분을 한 그룹으로 그루핑하는 층내 그루핑을 포함하나 이에 한정되지 않는다.

[0367] 클러스터링 알고리즘은 K-means, K-medoids, Clara 및/또는 Clarans를 포함한다. 매 타입의 중심 가중치의 선택방법은 비용 함수 $J(w,w_0)$가 제일 작을 경우 W0의 값이 상기 중심 가중치이고 비용 함수는 평방거리 $J(w_i,w_0)=\sum_{i=1}^{n}(w_i-w_0)^2$일 수 있는데, 여기서 J는 비용 함수이고 W는 상기 타입의 모든 가중치이며 W0은 중심 가중치이고 n은 매 타입에서의 가중치수량이며 wi는 타입에서의 i번째 가중치이고 $1\leq i\leq n$이며 n은 자연수이다.

[0368] 진일보로, 입력 뉴런을 양자화하는 것에 대해 설명하는데 이는,

[0369] 입력 뉴런을 p단으로 나누되, 매 세그먼트의 입력 뉴런은 하나의 뉴런 범위 및 하나의 뉴런 인덱스와 대응하여 뉴런 사전을 결정하며, 여기서 p는 자연수인 단계; 및

[0370] 상기 입력 뉴런을 코딩하고 매 세그먼트의 모든 입력 뉴런을 하나의 중심 뉴런으로 대체하여 뉴런 코드북을 결정하는 단계를 포함한다.

[0371] 도 1c을 참조하면, 도 1c은 본원 발명의 실시예에서 제공하는 입력 뉴런을 양자화하는 과정모식도로서, 도 1c에 도시된 바와 같이 본 실시예는 ReLU활성화층 뉴런을 양자화하는 것을 예로 하여 구체적으로 설명하였다. 우선 ReLU함수를 분할하여 모두 4개의 세그먼트로 분할하고 각각 0.0, 0.2, 0.5와 0.7로 4개의 세그먼트의 중심 뉴런을 나타내며 00, 01, 10과 11로 뉴런 인덱스를 나타낸다. 마지막으로 뉴런 인덱스와 중심 뉴런을 포함하는 뉴런 코드북 및 뉴런 범위와 뉴런 인덱스를 포함하는 뉴런 사전을 생성하는데 여기서 뉴런 범위와 뉴런 인덱스는 대응되게 저장되고 x는 뉴런을 양자화하지 않을 경우의 뉴런의 값을 나타낸다. 상기 입력 뉴런의 양자화 과정은 실제 수요에 따라 입력 뉴런을 여러 세그먼트로 분할할 수 있고 매 세그먼트의 인덱스를 얻어 뉴런 사전을 조성할 수 있다. 다음 뉴런 인덱스에 근거하여 매 세그먼트에서의 입력 뉴런을 뉴런 코드북에서의 중심 뉴런으로 대체함으로써 입력 뉴런 사이의 유사성을 충분히 발굴할 수 있고 입력 뉴런의 분포 특성을 얻음으로써 낮은 비트 양자화를 진행하여 매 하나의 입력 뉴런을 나타내는 비트수를 절감함으로써 입력 뉴런의 저장지출과 액세스 지출을 저하시킨다.

[0372] 단계 S2, 상기 가중치 코드북과 뉴런 코드북에 근거하여 연산 코드북을 결정하되, 구체적으로는 아래와 같은 단계를 포함한다.

[0373] 단계 S21, 상기 가중치에 근거하여 가중치 코드북에서의 대응되는 가중치 인덱스를 결정하고 다시 가중치 인덱스를 통해 상기 가중치와 대응되는 중심 가중치를 결정;

[0374] 단계 S22, 상기 입력 뉴런에 근거하여 뉴런 코드북에서의 대응되는 뉴런 인덱스를 결정하고 다시 뉴런 인덱스를 통해 상기 입력 뉴런과 대응되는 중심 뉴런을 결정; 및

[0375] 단계 S23, 상기 중심 가중치와 중심 뉴런에 대해 연산작업을 진행하여 연산 결과를 얻고 상기 연산 결과를 매트릭스로 구성하여 상기 연산 코드북을 결정.

[0376] 도 1d를 참조하면 도 1d는 본원 발명의 실시예에서 제공하는 연산 코드북을 결정하는 과정모식도로서, 도 1d에 도시된 바와 같이 본 실시예는 곱셈 코드북을 예로 하고 기타 실시예에서 상기 연산 코드북은 덧셈 코드북, 풀링 코드북 등 일 수도 있으며 본원 발명은 이에 한정하지 않는다. 우선 가중치 사전에서 가중치와 대응되는 가중치 인덱스, 및 상기 가중치 인덱스와 대응되는 중심 가중치를 결정한 다음 뉴런 코드북에서 입력 뉴런에 근거하여 대응되는 뉴런 인덱스 및 상기 뉴런 인덱스와 대응되는 중심 뉴런을 결정한다. 마지막으로 상기 뉴런 인덱스와 가중치 인덱스를 연산 코드북의 행 인덱스와 열 인덱스로 하고 중심 뉴런과 중심 가중치에 대해 곱셈 연산하여 매트릭스를 조성, 즉 곱셈 코드북을 얻을 수 있다.

[0377] 단계 S2 다음에 가중치와 입력 뉴런에 대해 리트레이닝을 진행하되, 리트레이닝 할 경우 가중치 코드북과 뉴런 코드북만을 트레이닝하고 가중치 사전과 뉴런 사전에서의 콘텐츠는 변하지 않아 리트레이닝 작업을 간략화하여 작업량을 절감하는 단계 S3을 더 포함할 수 있다.

[0378] 도 1e를 참조하면, 도 1e는 본원 발명의 실시예에서 제공하는 처리장치의 구조모식도로서, 도 1e에 도시된 바와 같이 상기 처리장치는,

[0379] 작동 명령을 저장하기 위한 메모리(51);

[0380] 메모리(51)에서의 작동 명령을 실행하고 상기 작동 명령을 실행할 경우 상술한 처리방법에 따라 작업을 수행하는 프로세서(52)를 포함한다. 여기서 작동 명령은 옵 코드와 주소 코드를 포함하는 이진수 일 수 있는데 옵 코드는 프로세서(52)가 곧 실행할 조작을 지시하며 주소 코드는 프로세서(52)로 하여금 메모리(51)에서의 주소에서 상기 작업에 참여하는 데이터를 판독하도록 지시한다.

[0381] 본원 발명의 데이터의 처리장치에 있어서, 프로세서(52)는 메모리(51)에서의 작동 명령을 실행하는 것을 통해 상술한 데이터의 처리방법에 따라 작업을 수행함으로써 질서가 없는 가중치와 입력 뉴런을 양자화하여 낮은 비트와 규범화한 중심 가중치와 중심 뉴런을 얻을 수 있어 가중치와 입력 뉴런 사이의 국부 유사성을 발굴하여 양자의 분포 특성을 얻을 수 있으며 양자의 분포 특성에 근거하여 낮은 비트의 양자화를 진행함으로써 매 하나의

가중치와 입력 뉴런을 나타내는 비트수를 절감하여 양자의 저장 지출과 액세스 지출을 저하시킨다.

[0382] 도 1f를 참조하면, 도 1f는 본원 발명의 실시예에서 제공하는 연산장치의 구조모식도로서, 도 1f에 도시된 바와 같이 상기 연산장치는,

[0383] 수신한 명령을 디코딩하여 검색제어정보를 생성하는 명령제어유닛(1);

[0384] 명령제어유닛(1)이 생성한 검색제어정보 및 수신한 가중치 사전, 뉴런 사전, 연산 코드북, 가중치와 입력 뉴런에 근거하여 연산 코드북으로부터 출력 뉴런을 검색하는 검색 테이블 유닛(2)을 포함한다. 여기서 상기 가중치 사전은 가중치 위치(즉 가중치가 신경망 구조에서의 위치,(p, q)로 표시, 구체적으로는 가중치 사전에서 제p행 제q열의 위치를 표시)와 가중치 인덱스를 포함하고 상기 뉴런 사전은 입력 뉴런과 뉴런 인덱스를 포함하며 상기 연산 코드북은 가중치 인덱스, 뉴런 인덱스 및 입력 뉴런과 가중치의 연산 결과를 포함한다.

[0385] 여기서 상기 검색 테이블 유닛의 구체적인 작동 과정은 가중치에 근거하여 가중치가 가중치 사전에서 대응되는 가중치 위치를 결정하고 가중치 인덱스를 결정하며 입력 뉴런이 뉴런 사전에서 대응되는 뉴런 범위에 근거하여 뉴런 인덱스를 결정하고 가중치 인덱스와 뉴런 인덱스를 연산 코드북의 열 인덱스와 행 인덱스로 하여 연산 코드북에서 상기 열과 상기 행의 수치(연산 결과)를 검색하는데 상기 수치는 출력 뉴런이다.

[0386] 도 1b 내지 도 1d를 참조하면, 검색을 진행할 때 어느 뉴런의 뉴런 인덱스를 01이라고 가정하고 어느 가중치의 가중치 인덱스를 10이라고 가정하면 상기 뉴런과 가중치를 연산할 경우 곱셈 코드북에서의 제2행 제3열이 대응되는 수치 0.046, 즉 출력 뉴런을 검색한다. 유사하게 덧셈과 풀링 동작은 곱셈동작과 유사하므로 여기서 더이상 설명하지 않는다. 이해할 수 있는 것은 풀링은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함하나 이에 한정되지 않는다.

[0387] 더 구체적으로 상이한 연산작업에 근거하여 상기 검색 테이블은,

[0388] 가중치 인덱스(in1)와 뉴런 인덱스(in2)를 입력하고 곱셈 검색 테이블을 통해 테이블 조사 동작(mult_lookup)으로 가중치 인덱스와 대응되는 중심 가중치(data1)와 뉴런 인덱스와 대응되는 중심 뉴런(data2)의 곱셈동작을 완성, 즉 테이블 조사 동작 out=mult_lookup(in1,in2)으로 곱셈기능 out=data1*data2을 완성하는 곱셈 검색 테이블; 및/또는

[0389] 인덱스(in)에 근거하여 단계적 덧셈 검색 테이블을 통해 테이블 조사 동작(add_lookup)으로 인덱스와 대응되는 중심 데이터(data)의 덧셈 동작을 완성하도록 입력하되, 여기서 in과 data는 길이가 N인 벡터이고 N은 자연수, 즉 테이블 조사 동작 out=add_lookup(in)으로 덧셈 기능 out＝data[1]+data[2]+...+data[N]을 완성하거나 및/또는, 가중치 인덱스(in1)와 뉴런 인덱스(in2)가 덧셈 검색 테이블을 통해 테이블 조사 동작으로 가중치 인덱스와 대응되는 중심 가중치(data1)와 뉴런 인덱스와 대응되는 중심 뉴런(data2)의 덧셈 동작을 완성하도록 입력, 즉 테이블 조사 동작 out=add_lookup(in1,in2)으로 덧셈기능 out=data1+data2을 완성하는 덧셈 검색 테이블; 및/또는

[0390] 인덱스와 대응되는 중심 데이터(data)의 풀링 동작을 입력, 즉 테이블 조사 out=pool_lookup(in)로 풀링 동작 out=pool(data)을 완성하되, 풀링 동작은 평균값 풀링, 최대치 풀링과 중앙값 풀링을 포함하는 풀링 검색 테이블에서의 적어도 하나를 포함할 수 있다.

[0391] 도 1g을 참조하면 도 1g은 본원 발명의 실시예에서 제공하는 다른 연산장치의 구조모식도로서, 도 1g에 도시된 바와 같이 상기 구체적인 실시예의 연산장치는 도 1f에서의 연산장치와 비교하여 전처리 유닛(4), 저장 유닛(3), 캐시 유닛(6)과 직접 메모리 액세스 유닛(5)을 더 포함함으로써 본원 발명의 처리과정을 최적화하여 데이터의 처리에 더 순서가 있도록 할 수 있다.

[0392] 전처리 유닛(4), 외부에서 입력한 입력정보를 전처리하여 상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 얻고 전처리는 분할, 가우스 필터링, 이진화, 규칙화 및/또는 정규화를 포함하나 이에 한정되지 않는다.

[0393] 저장 유닛(3)은 입력 뉴런, 가중치, 가중치 사전, 뉴런 사전, 연산 코드북과 명령을 저장하고 출력 뉴런을 수신하고;

[0394] 캐시 유닛(6)은 상기 명령, 가중치 인덱스, 뉴런 인덱스와 출력 뉴런은 캐시하는데 이는,

[0395] 상기 명령을 캐시하고 캐시된 명령을 명령제어유닛(1)에 출력하는 명령 캐시(61);

[0396] 상기 가중치를 캐시하고 캐시된 가중치를 검색 테이블 유닛(2)에 출력하는 가중치 캐시(62);

[0397] 상기 입력 뉴런을 캐시하고 캐시된 입력 뉴런을 검색 테이블 유닛(2)에 출력하는 입력 뉴런 캐시(63);

[0398] 검색 테이블 유닛(2)이 출력한 출력 뉴런을 캐시하고 캐시된 출력 뉴런을 검색 테이블 유닛(2)에 출력하는 출력 뉴런 캐시(64);

[0399] 입력 뉴런에 근거하여 대응되는 뉴런 인덱스를 결정하고 상기 뉴런 인덱스를 캐시하며 캐시된 뉴런 인덱스를 검색 테이블 유닛(2)에 출력하는 뉴런 인덱스 캐시(65);

[0400] 가중치에 근거하여 대응되는 가중치 인덱스를 결정하고 상기 가중치 인덱스를 캐시하며 캐시된 가중치 인덱스를 검색 테이블 유닛(2)에 출력하는 가중치 인덱스 캐시(66)를 포함할 수 있으며,

[0401] 직접 메모리 액세스 유닛(5)은 상기 저장 유닛(3)과 캐시 유닛(6) 사이에서 데이터 또는 명령을 판독 기록할 수 있다.

[0402] 선택적으로, 명령에 있어서 상기 명령은 신경망 전용명령으로서 인공 신경망 연산을 완성하는데 전문적으로 사용될 수 있는 명령을 포함할 수 있다. 신경망 전용명령은 제어 명령, 데이터 전송 명령, 연산 명령과 논리 명령을 포함하나 이에 한정되지 않는다. 여기서 제어 명령은 신경망 실행과정을 제어한다. 데이터 전송 명령은 상이한 저장매체 사이의 데이터 전송을 완성하기 위한 것으로 데이터 양식은 매트릭스, 벡터와 스칼라를 포함하나 이에 한정되지 않는다. 연산 명령은 신경망의 산술 연산을 완성하기 위한 것으로 매트릭스 연산 명령, 벡터 연산 명령, 스칼라 연산 명령, 콘볼루션 신경망 연산 명령, 완전 연결 신경망 연산 명령, 풀링신경망 연산 명령, RBM 신경망 연산 명령, LRN 신경망 연산 명령, LCN 신경망 연산 명령, LSTM 신경망 연산 명령, RNN 신경망 연산 명령, RELU 신경망 연산 명령, PRELU 신경망 연산 명령, SIGMOID 신경망 연산 명령, TANH 신경망 연산 명령과 MAXOUT 신경망 연산 명령을 포함하나 이에 한정되지 않는다. 논리 명령은 신경망의 논리 연산을 완성하기 위한 것으 벡터논리 연산 명령과 스칼라 논리 연산 명령을 포함하나 이에 한정되지 않는다.

[0403] 여기서 RBM 신경망 연산 명령은 Restricted Boltzmann Machine(RBM) 신경망 연산을 실현한다.

[0404] LRN 신경망 연산 명령은 Local Response Normalization(LRN) 신경망 연산을 실현한다.

[0405] LSTM 신경망 연산 명령은 Long Short-Term Memory(LSTM) 신경망 연산을 실현한다.

[0406] RNN 신경망 연산 명령은 Recurrent Neural Networks(RNN) 신경망 연산을 실현한다.

[0407] RELU 신경망 연산 명령은 Rectified linear unit(RELU) 신경망 연산을 실현한다.

[0408] PRELU 신경망 연산 명령은 Parametric Rectified Linear Unit(PRELU) 신경망 연산을 실현한다.

[0409] SIGMOID 신경망 연산 명령은 S형 성장 곡선(SIGMOID) 신경망 연산을 실현한다.

[0410] TANH 신경망 연산 명령은 하이퍼 볼릭 탄젠트 함수(TANH) 신경망 연산을 실현한다.

[0411] MAXOUT 신경망 연산 명령은 MAXOUT 신경망 연산을 실현한다.

[0412] 더 진일보로, 상기 신경망 전용명령은 Cambricon 명령 집합을 포함하는데 여기서 상기 Cambricon 명령 집합은 적어도 하나의 Cambricon 명령을 포함하고 Cambricon 명령의 길이는 64bit이며 상기 Cambricon 명령은 옵 코드와 피연산자를 포함한다. Cambricon 명령은 4가지 유형의 명령을 포함하는데 각각 Cambricon 제어 명령(control instructions), Cambricon 데이터 전송 명령(data transfer instructions), Cambricon 연산 명령(computational instructions)과 Cambricon 논리 명령(logical instructions)이다.

[0413] 선택적으로, Cambricon 제어 명령은 수행과정을 제어한다. Cambricon 제어 명령은 점프(JUMP) 명령과 조건부 분기(conditional branch) 명령을 포함한다.

[0414] 선택적으로, Cambricon 데이터 전송 명령은 상이한 저장매체 사이의 데이터 전송을 완성한다. Cambricon 데이터 전송 명령은 로드(load)명령, 저장(store) 명령과 무브(move) 명령을 포함한다. load 명령은 데이터를 메인 메모리로부터 캐시에 로딩하기 위한 것이고 store 명령은 데이터를 캐시로부터 메인 메모리에 저장하기 위한 것이며 move 명령은 캐시와 캐시 또는 캐시와 레지스터 또는 레지스터와 레지스터 사이에서 데이터를 운송하기 위한 것이다. 데이터 전송 명령은 매트릭스, 벡터와 스칼라를 포함하는 세 가지 상이한 데이터 조직방식을 지지한다.

[0415] 선택적으로, Cambricon 연산 명령은 신경망 산술 연산을 완성하기 위한 것이다. Cambricon 연산 명령은

Cambricon 매트릭스 연산 명령, Cambricon 벡터 연산 명령과 Cambricon 스칼라 연산 명령을 포함한다.

[0416] 선택적으로, Cambricon 매트릭스 연산 명령은 신경망에서의 매트릭스 연산을 완성하기 위한 것으로 매트릭스 곱셈 벡터(matrix multiply vector), 벡터 곱셈 매트릭스(vector multiply matrix), 매트릭스 곱셈 스칼라(matrix multiply scalar), 외적(outer product), 매트릭스 덧셈 매트릭스(matrix add matrix)와 매트릭스 뺄셈 매트릭스(matrix subtract matrix)를 포함한다.

[0417] 선택적으로, Cambricon 벡터 연산 명령은 신경망에서의 벡터 연산을 완성하기 위한 것이로 벡터 기본 연산(vector elementary arithmetics), 벡터 초월함수 연산(vector transcendental functions), 내적(dot product), 벡터 랜덤 생성(random vector generator)과 벡터에서의 최대/최소치(maximum/minimum of a vector)를 포함한다. 여기서 벡터 기본 연산은 벡터 더하기, 빼기, 곱하기, 나누기(add、 subtract、multiply、divide)를 포함하고 벡터 초월함수는 다항식을 계수로 하는 그 어떤 다항식 방정식을 만족시키지 않는 함수를 가리키는데 지수함수, 로그함수, 삼각함수와 역삼각함수를 포함하나 이에 한정하지 않는다.

[0418] 선택적으로, Cambricon 스칼라 연산 명령은 신경망에서의 스칼라 연산을 완성하기 위한 것으로 스칼라 기본 연산(scalar elementary arithmetics)과 스칼라 초월함수 연산(scalar transcendental functions)을 포함한다. 여기서 스칼라 기본 연산은 스칼라 더하기, 빼기, 곱하기, 나누기(add、subtract、multiply、divide)를 포함하고 스칼라 초월함수는 다항식을 계수로 하는 그 어떤 다항식 방정식을 만족시키지 않는 함수를 가리키는데 지수함수, 로그함수, 삼각함수와 역삼각함수를 포함하나 이에 한정하지 않는다.

[0419] 선택적으로, Cambricon 논리 명령은 신경망의 논리 연산을 완성한다. Cambricon 논리 연산은 Cambricon 벡터논리 연산 명령과 Cambricon 스칼라 논리 연산 명령을 포함한다. 여기서 Cambricon 벡터논리 연산 명령은 벡터 비교(vector compare), 벡터논리 연산(vector logical operations)과 벡터 크기 병합(vector greater than merge)을 포함한다. 여기서 벡터 비교는 작기 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하나 이에 한정되지 않는다. 벡터논리 연산은 그리고, 또는, 아님을 포함한다.

[0420] 선택적으로, Cambricon 스칼라 논리 연산은 스칼라 비교(scalar compare), 스칼라 논리 연산(scalar logical operations)를 완성하기 위한 것이다. 여기서 스칼라 비교는 크기, 작기, 같기, 크거나 같기, 작거나 같기와 같지 않기를 포함하나 이에 한정되지 않는다. 스칼라 논리 연산은 그리고, 또는, 아님을 포함한다.

[0421] 도 1h을 참조하면 도 1h은 본원 발명의 실시예에서 제공하는 다른 연산방법의 흐름모식도로서 도 1h에 도시된 바와 같이 상기 연산방법은 아래와 같은 단계를 포함한다.

[0422] 단계 S81, 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 수신하되, 여기서 상기 가중치 사전은 가중치 위치와 가중치 인덱스를 포함하고 상기 뉴런 사전은 입력 뉴런과 뉴런 인덱스를 포함하며 상기 연산 코드북은 가중치 인덱스, 뉴런 인덱스 및 입력 뉴런과 가중치의 연산 결과를 포함한다.

[0423] 단계 S82, 상기 명령을 디코딩하여 검색제어정보를 결정;

[0424] 단계 S83, 상기 검색제어정보, 가중치, 가중치 사전, 뉴런 사전과 입력 뉴런에 근거하여 연산 코드북에서 출력 뉴런을 검색한다.

[0425] 여기서 단계 S83과 검색 테이블 유닛의 구체적인 작동 과정은 유사한 바, 구체적으로 아래와 같은 서브 단계를 포함한다.

[0426] 단계 S831, 상기 가중치, 입력 뉴런, 가중치 사전과 뉴런 사전에 근거하여 뉴런 사전에서 뉴런 범위를 결정함으로써 뉴런 인덱스를 결정하고 가중치 사전에서 가중치 위치를 결정함으로써 가중치 인덱스를 결정; 및

[0427] 단계 S832, 상기 가중치 인덱스와 뉴런 인덱스에 근거하여 연산 코드북에서 상기 연산 결과를 검색함으로써 출력 뉴런을 결정.

[0428] 본원 발명의 연산방법을 최적화하고 처리가 더 편리하고 질서가 있도록 하기 위하여 본원 발명의 실시예는 또 다른 연산방법을 제공하는데 도A9는 본원 발명의 실시예에서 제공하는 구체적인 실시예의 연산방법의 흐름모식도로서 상기 연산방법은 아래와 같은 단계를 포함한다.

[0429] 단계 S90, 외부에서 입력한 입력정보를 전처리한다.

[0430] 선택적으로, 외부에서 입력한 입력정보를 전처리하는 상기 단계는 구체적으로, 상기 입력정보와 대응되는 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 얻되, 상기 전처리는 분할, 가우스 필터링, 이

진화, 구칙화 및/또는 정규화를 포함하는 단계를 포함한다.

[0431] 단계 S91, 상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 수신한다.

[0432] 단계 S92, 상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전, 연산 코드북을 저장한다.

[0433] 단계 S93, 상기 가중치, 입력 뉴런, 명령, 가중치 인덱스, 뉴런 인덱스를 캐시한다.

[0434] 단계 S94, 상기 명령을 디코딩하여 검색제어정보를 결정한다.

[0435] 단계 S95, 상기 가중치, 입력 뉴런, 가중치 사전과 뉴런 사전에 근거하여 뉴런 사전에서 뉴런 범위를 결정함으로써 뉴런 인덱스를 결정하고 가중치 사전에서 가중치 위치를 결정함으로써 가중치 인덱스를 결정한다.

[0436] 단계 S96, 상기 가중치 인덱스와 뉴런 인덱스에 근거하여 연산 코드북에서 상기 연산 결과를 검색함으로써 출력 뉴런을 결정한다.

[0437] 도 2a을 참조하면 도 2a은 본원 발명의 실시예에서 제공하는 분층 저장장치의 구조모식도로서 도 2a에 도시된 바와 같이 상기 장치은 데이터에서의 중요한 비트값을 저장하기 위한 정밀 저장 유닛과 데이터에서의 중요하지 않은 비트값을 저장하기 위한 비정밀 저장 유닛을 포함한다.

[0438] 정밀 저장 유닛은 오류 검출과 정정ECC메모리를 사용하고 비정밀 저장 유닛은 비 ECC메모리를 사용한다.

[0439] 진일보로, 분층 저장장치저장의 데이터는 신경망 파라미터로서 입력 뉴런, 가중치와 출력 뉴런을 포함하고 정밀 저장 유닛은 입력 뉴런, 출력 뉴런 및 가중치의 중요한 비트값을 저장하며 비정밀 저장 유닛은 입력 뉴런, 출력 뉴런 및 가중치의 중요하지 않은 비트값을 저장한다.

[0440] 진일보로, 분층 저장장치가 저장한 데이터는 부동 소수점 데이터와 고정 소수점 데이터를 포함하는데 부동 소수점 데이터에서의 부호 비트와 지수 부분을 중요한 비트값으로 지정하고 근 부분을 중요하지 않은 비트값으로 지정하며 고정 소수점 데이터에서의 부호 비트와 수치 부분의 앞의 x비트를 중요한 비트값으로 지정하고 수치 부분의 나머지 비트를 중요하지 않은 비트값으로 지정하되, 여기서 x는 0보다 크거나 같고 m보다 작은 자연수이며 m는 고정 소수점 데이터의 총 비트이다. 중요한 비트값을 ECC메모리에 저장하여 정밀 저장을 진행하고 중요하지 않은 비트값을 비 ECC메모리에 저장하여 모두 비정밀 저장을 진행한다,

[0441] 진일보로, ECC메모리는 ECC 체크가 있는 DRAM(Dynamic Random Access Memory, 약칭 DRAM)동적 랜덤 액세스 메모리와 ECC 체크가 있는 SRAM(Static Random-Access Memory, 약칭SRAM) 정적 랜덤 액세스 메모리를 포함하되, 여기서 ECC 체크가 있는 SRAM은 6T SRAM을 사용하고 본원 발명의 기타 실시예에서는 4T SRAM 또는 3T SRAM을 사용할 수도 있다.

[0442] 진일보로, 비 ECC메모리는 ECC 체크가 아닌 DRAM와 ECC 체크가 아닌 SRAM을 포함하되, ECC 체크가 아닌 SRAM은 6T SRAM을 사용하고 본원 발의 기타 실시예에서는 4T SRAM 또는 3TSRAM을 사용할 수도 있다.

[0443] 여기서 6T SRAM에서 매 하나의 비트를 저장하는 유닛은 6개의 전계 효과 트랜지스터MOS(metal oxide semiconductor, 약칭 MOS)관으로 조성되고 4T SRAM에서 매 하나의 비트를 저장하는 유닛은 4개의 MOS관으로 조성되며 3T SRAM에서 매 하나의 비트를 저장하는 유닛은 3개의 MOS관으로 조성된다.

[0444] 신경망 가중치를 저장하는 SRAM은 일반적으로 6T SRAM을 사용하는데 비록 6T SRAM은 안정성이 높으나 점용한 면적이 커 판독 기록 전력 손실이 높다. 신경망 알고리즘은 일정한 오차 허용 능력을 가지나 6T SRAM이 신경망의 오차 허용 특성을 이용할 수 없으므로 본 실시예에서는 신경망의 오차 허용 능력을 충분히 발굴하기 위하여 4T SRAM 또는 3T SRAM 저장기술을 사용하여 6T SRAM을 대체함으로써 SRAM저장밀도를 향상시키고 SRAM 액세스 저장의 전력 손실을 절감시키며 신경망 알고리즘의 오차 허용성을 이용하여 4T SRAM의 소음 방지 능력이 약한 결점을 덮어버린다.

[0445] 도 2b를 참조하면 도 2b는 본원 발명의 실시예에서 제공하는 4T SRAM저장 유닛의 구조모식도로서 도 2b에 도시된 바와 같이 4T SRAM저장 유닛은 4개의 NMOS로 조성되는데 각각 M1(제1 MOS관), M2(제2 MOS관), M3(제3 MOS관), M4(제4 MOS관)이다. M1과 M2는 게이팅에 사용되고 M3과 M4는 저장에 사용된다.

[0446] M1그리드와 워드 라인(WL)(Word Line)은 전기적으로 연결되고소스 전극과 비트 라인(BL)(Bit Line)은 전기적으로 연결되며 M2그리드와 워드 라인(WL)은 전기적으로 연결되고 소스 전극과 비트 라인(BLB)은 전기적으로 연결되며 M3그리드와 M4소스 전극, M2는 드레인 연결됨과 동시에 저항(R2)을 통해 작동 전압(Vdd)과 연결되고 M3은 드레인 접지되며 M4그리드와 M3소스 전극, M1은 드레인 연결됨과 동시에 저항(R1)을 통해 작동 전압(Vdd)과 연

결되고 M4는 드레인 접지된다. WL은 저장 유닛의 게이팅 액세스를 제어하기 위한 것이고 BL은 저장 유닛의 판독 기록을 위한 것이다. 판독 작업을 진행할 경우 WL을 높여 BL에서 비트를 판독하면 된다. 기록 작업을 진행할 경우 WL을 높이고 BL을 높이거나 낮추는데 BL의 구동능력이 저장 유닛보다 강하므로 강제적으로 원래의 상태를 커버하게 된다.

[0447] 도 2c를 참조하면 도 2c는 본원 발명의 실시예에서 제공하는 3T SRAM저장 유닛의 구조모식도로서 도 2c에 도시된 바와 같이 3T SRAM저장 유닛은 3개의 MOS로 조성되는데 각각 M1(제1 MOS관), M2(제2 MOS관)와 M3(제3 MOS관)이다. M1은 게이팅에 사용되고 M2와 M3은 저장에 사용된다.

[0448] M1그리드와 워드 라인(WL)(Word Line)은 전기적으로 연결되고소스 전극과 비트 라인(BL)(Bit Line)은 전기적으로 연결되며 M2그리드와 M3소스 전극은 연결됨과 동시에 저항(R2)을 통해 작동 전압(Vdd)과 연결되고 M2는 드레인 접지되며 M3그리드와 M2소스 전극, M1은 드레인 연결됨과 동시에 저항(R1)을 통해 작동 전압(Vdd)과 연결되고 M3은 드레인 접지된다. WL은 저장 유닛의 게이팅 액세스를 제어하기 위한 것이고 BL은 저장 유닛의 판독 기록을 위한 것이다. 판독 작업을 진행할 경우 WL을 높여 BL에서 비트를 판독하면 된다. 기록 작업을 진행할 경우 WL을 높이고 BL을 높이거나 낮추는데 BL의 구동능력이 저장 유닛보다 강하므로 강제적으로 원래의 상태를 커버하게 된다.

[0449] 본원 발명의 저장장치는 근사 저장 기술을 사용하여 신경망의 오차 허용 능력을 충분히 발굴할 수 있고 신경 파라미터를 근사 저장할 수 있는데 파라미터에서의 중요한 비트는 정밀 저장을 사용하고 중요하지 않은 비트는 비정밀 저장을 사용함으로써 저장지출과 메모리 액세스 에너지 소모 지출을 절감한다.

[0450] 본원 발명의 실시예는 데이터 처리장치를 제공하는데 상기 장치는 근사 저장 기술과 대응되는 가속장치이고 도 ba4를 참조하면 도 2d는 본원 발명의 실시예에서 제공하는 데이터 처리장치의 구조모식도이며 상기 데이터 처리장치는 비정밀 연산 유닛, 명령제어유닛과 상술한 분층 저장장치를 포함한다.

[0451] 분층 저장장치는 명령과 연산 파라미터를 수신하고 연산 파라미터에서의 중요한 비트값과 명령을 정밀 저장 유닛에 저장하며 연산 파라미터에서의 중요하지 않은 비트값을 비정밀 저장 유닛에 저장한다.

[0452] 명령제어유닛은 분층 저장장치에서의 명령을 수신하고 명령을 디코딩하여 제어정보를 생성하며 비정밀 연산 유닛을 제어하여 계산 작업을 진행하도록 한다.

[0453] 비정밀 연산 유닛은 분층 저장장치에서의 연산 파라미터를 수신하고 제어정보에 따라 연산하며 연산 결과를 분층 저장장치에 전송하여 저장하거나 출력하도록 한다.

[0454] 진일보로, 비정밀 연산 유닛은 신경망 프로세서이다. 진일보로, 상기 연산 파라미터는 신경망 파라미터이고 분층 저장장치는 신경망의 뉴런, 가중치와 명령을 저장하며 뉴런의 중요한 비트값, 가중치의 중요한 비트값과 명령을 정밀 저장 유닛에 저장하고 뉴런의 중요하지 않은 비트값과 가중치의 중요하지 않은 비트값을 비정밀 저장 유닛에 저장한다. 비정밀 연산 유닛은 분층 저장장치에서의 입력 뉴런과 가중치를 수신하고 제어정보에 따라 신경망 연산을 완성하여 출력 뉴런을 얻으며 출력 뉴런을 분층 저장장치에 다시 전송하여 저장 또는 출력하도록 한다.

[0455] 진일보로, 비정밀 연산 유닛은(1)비정밀 연산 유닛은 분층 저장장치로부터의 정밀 저장 유닛에서의 입력 뉴런의 중요한 비트값과 가중치의 중요한 비트값을 직접 수신하여 계산;(2)비정밀 연산 유닛은 중요한 비트값과 중요하지 않은 비트값이 연결되어 완성한 입력 뉴런과 가중치를 수신하여 계산하는 두 가지 계산모드를 가질수 있는데 여기서 입력 뉴런과 가중치의 중요한 비트값과 중요하지 않은 비트값은 저장 유닛에서 판독될 때 연결된다.

[0456] 진일보로, 도 2e를 참조하면 도 2e에 도시된 바와 같이 데이터 처리장치는, 입력된 원시 데이터를 전처리하여 저장장치에 전송하기 위한 전처리모듈을 더 포함하는데 전처리는 분할, 가우스 필터링, 이진화, 규칙화, 정규화 등을 포함한다.

[0457] 진일보로, 데이터 처리장치는 명령 캐시, 입력 뉴런 분층 캐시, 가중치 분층 캐시와 출력 뉴런 분층 캐시를 더 포함하는데 여기서 명령 캐시는 분층 저장장치와 명령제어유닛 사이에 설치되어 전용 명령을 저장하고 입력 뉴런 분층 캐시는 저장장치와 비정밀 연산 유닛 사이에 설치되어 입력 뉴런을 캐시하며 입력 뉴런 분층 캐시는 입력 뉴런 정밀 캐시와 입력 뉴런 비정밀 캐시를 포함하여 입력 뉴런의 중요한 비트값과 중요하지 않은 비트값을 각각 캐시하고 가중치 분층 캐시는 저장장치와 비정밀 연산 유닛 사이에 설치되어 가중치 데이터를 캐시하며 가중치 분층 캐시는 가중치 정밀 캐시와 가중치 비정밀 캐시를 포함하여 가중치의 중요한 비트값과 중요하지 않은 비트값을 각각 캐시하고 출력 뉴런 분층 캐시는 저장장치와 비정밀 연산 유닛 사이에 설치되어 출력 뉴런을 캐

시하며 상기 출력 뉴런 분층 캐시는 출력 뉴런 정밀 캐시와 출력 뉴런 비정밀 캐시를 포함하여 출력 뉴런의 중요한 비트값과 중요하지 않은 비트값을 각각 캐시한다.

[0458] 진일보로, 데이터 처리장치는 저장장치, 명령 캐시, 가중치 분층 캐시, 입력 뉴런 분층 캐시와 출력 뉴런 분층 캐시에서 데이터 또는 명령의 판독 기록을 진행하기 위한 직접 데이터 액세스 유닛(DMA)(direct memory access)을 더 포함한다.

[0459] 진일보로, 상기 명령 캐시, 입력 뉴런 분층 캐시, 가중치 분층 캐시와 출력 뉴런 분층 캐시는 모두 4T SRAM 또는 3T SRAM을 사용한다.

[0460] 진일보로, 비정밀 연산 유닛은 제1 부분인 곱셈기, 제2 부분인 덧셈 트리, 제3 부분인 활성화 함수 유닛과 같은 3개의 부분을 포함하나 이에 한정되지 않는다. 제1 부분은 입력 데이터1(in1)와 입력 데이터2(in2)를 곱하여 곱한 후의 출력(out)을 얻고 그 과정은 out=in1*in2이며; 제2 부분은 입력 데이터(in)1가 덧셈 트리를 통해 단계적으로 더하여 출력 데이터(out)를 얻는데 여기서 in1은 하나의 길이가 N인 벡터이고 N은 1보다 크며 과정은 out＝in1[1]+in1[2]+...+in1[N]이거나; 또는 입력 데이터(in1)가 덧셈 트리를 통해 누가된 후 입력 데이터(in2)와 더하여 출력 데이터(out)를 얻고 그 과정은 out=in1[1]+in1[2]+...+in1[N]+in2이거나; 또는 입력 데이터(in1)와 입력 데이터(in2)를 더하여 출력 데이터(out)를 얻고 그 과정은 out=in1+in2이며; 제3 부분은 입력 데이터(in)가 활성화 함수(active) 연산을 통해 활성화 출력 데이터(out)를 얻고 그 과정은 out=active(in)인데 활성화 함수active는 sigmoid, tanh, relu, softmax 등 일 수 있고 활성화 작업을 제외하고 제3 부분은 기타 비선형 함수를 통해 입력 데이터(in)로 하여금 연산(f)에 의해 출력 데이터(out)를 얻으며 과정은 out=f(in)이다.

[0461] 비정밀 연산 유닛은 풀링 유닛을 더 포함할 수 있는데 풀링 유닛은 입력 데이터(in)가 풀링 연산을 통해 출력 데이터(out)를 얻고 과정은 out＝pool(in)이며 여기서 pool은 풀링 연산이고 풀링 연산은 평균값 풀링, 최대치 풀링, 중앙값 풀링을 포함하나 이에 한정되지 않으며 입력 데이터(in)는 출력(out)과 관련되는 하나의 풀링 핵에서의 데이터이다.

[0462] 비정밀 연산 유닛이 연산을 수행하는 것은 다음과 같은 몇개 부분을 포함하는데 제1 부분은 입력 데이터1과 입력 데이터2를 곱하여 곱한 후의 데이터를 얻는 것이고; 제2 부분은 덧셈 트리 연산을 수행하는 것인데 입력 데이터1을 덧셈 트리를 통해 단계적으로 더하거나 또는 상기 입력 데이터1을 덧셈 트리를 통해 단계적으로 더한 후 입력 데이터2와 더하여 출력 데이터를 얻는 것이며; 제3 부분은 활성화 함수 연산을 수행하는 것인데 입력 데이터에 활성화 함수(active) 연산을 통해 출력 데이터를 얻은 것이다. 이상의 몇개 부분의 연산은 자유롭게 조합되어 여러 가지 상이한 기능의 연산을 실현할 수 있다.

[0463] 본원 발명의 데이터 처리장치는 근사 저장 기술을 충분히 이용할 수 있고 신경망의 오차 허용 능력을 충분히 발굴하여 신경망의 계산량과 신경망 액세스 저장량을 절감함으로써 계산 에너지 소모와 메모리 액세스 에너지 소모를 절감할 수 있다. 다층 인공 신경망 연산에 대한 전용 SIMD명령과 제정된 연산 유닛을 통해 CPU와 GPU연산 성능이 부족한 문제를 해결하고 프런트 엔드 디코딩 지출이 큰 문제를 해결하여 다층 인공 신경망 연산 알고리즘에 대한 지지를 효과적으로 향상시키며, 다층 인공 신경망 연산 알고리즘에 대한 전용 비정밀 저장의 온-칩 캐시를 사용함으로써 입력 뉴런과 가중치 데이터의 중요성을 충분히 발굴하여 반복적으로 메모리에서 이러한 데이터를 판독하는 것을 방지하며 메모리 액세스 대역폭을 절감시키고 메모리 대역폭이 다층 인공 신경망 연산 및 그의 트레이닝 알고리즘의 기능적 난관이 되는 문제를 방지한다.

[0464] 이상은 단지 예시적인 설명일 뿐 본원 발명이 이에 한정되지 않는 바, 데이터 처리장치는 범용 연산 프로세서와 같은 비 신경망 프로세서를 포함할 수 있고 범용 연산은 스칼라산술 연산, 스칼라 논리 연산 등과 같은 상응하는 범용 연산 명령과 데이터를 구비하며 범용 연산 프로세서는 예하면 하나 또는 다수의 곱셈기, 하나 또는 다수의 덧셈기를 포함하여 덧셈, 곱셈 등 기본연산을 수행하나 이에 한정되지 않는다.

[0465] 본원 발명의 또 다른 실시예는 데이터 저장방법을 제공하는데 이는 근사 저장의 방식을 사용하여 데이터를 분층 저장하는 바, 도 2f을 참조하면 도 2f은 본원 발명의 실시예에서 제공하는 데이터 저장방법의 흐름도로서 다음과 같은 단계를 포함한다.

[0466] 단계 S601, 데이터에서의 중요한 비트값을 정밀 저장한다.

[0467] 단계 S602, 데이터에서의 중요하지 않은 비트값을 비정밀 저장한다.

[0468] 구체적으로 말하면 상기 데이터 저장방법은 다음과 같은 단계를 포함한다.

[0469] 데이터의 중요한 비트값과 중요하지 않은 비트값을 추출;

[0470] 데이터에서의 중요한 비트값을 ECC메모리에 저장하여 정밀 저장을 진행;

[0471] 데이터에서의 중요하지 않은 비트값을 비 ECC메모리에 저장하여 비정밀 저장을 진행.

[0472] 본 실시예에서 저장된 데이터는 신경망 파라미터이고 신경망 파라미터를 나타내는 비트수 비트를 중요한 비트값과 중요하지 않은 비트값으로 분류한다. 예를 들어 설명하면 신경망의 하나의 파라미터는 m개의 비트를 가지는데 여기서 N개의 비트는 중요한 비트값,(m-n)개의 비트는 중요하지 않은 비트값이며 여기서 m은 0보다 큰 정수이고 n은 0보다 크고 m보다 작거나 같은 정수이다.

[0473] 신경망 파라미터는 입력 뉴런, 가중치와 출력 뉴런을 포함하는데입력 뉴런의 중요한 비트값, 출력 뉴런의 중요한 비트값과 가중치의 중요한 비트값을 정밀 저장하고; 입력 뉴런의 중요하지 않은 비트값, 출력 뉴런 의 중요하지 않은 비트값과 가중치의 중요하지 않은 비트값을 비정밀 저장한다.

[0474] 데이터는 부동 소수점 데이터와 고정 소수점 데이터를 포함하는데 여기서 부동 소수점 데이터에서의 부호 비트와 지수 부분을 중요한 비트값으로 정의하고 근 부분을 중요하지 않은 비트값으로 정의하며; 고정 소수점 데이터에서의 부호 비트와 수치 부분의 앞 x비트를 중요한 비트값으로 정의하고 수치 부분의 나머지 비트를 중요하지 않은 비트값으로 정의하되, 여기서 x는 0보다 크거나 같고 m보다 작은 자연수이며 m는 파라미터의 총 비트이다.

[0475] ECC메모리는 ECC 체크가 있는 SRAM와 ECC 체크가 있는 DRAM을 포함하되, 상기 비 ECC메모리는 ECC 체크가 아닌 SRAM와 ECC 체크가 아닌 DRAM을 포함하고 상기 ECC 체크가 있는 SRAM와 ECC 체크가 아닌 SRAM은 6T SRAM을 사용하며 본원 발의 기타 실시예에서는 4T SRAM 또는 3T SRAM을 사용할 수도 있다.

[0476] 본원 발명의 또 다른 실시예는 데이터 처리방법을 제공하는데 도 2g은 본원 발명의 실시예에서 제공하는 데이터 처리방법의 흐름도로서 도 2g에 도시된 바와 같이 다음과 같은 단계를 포함한다.

[0477] 단계 S1, 명령과 파라미터를 수신하고 파라미터에서의 중요한 비트값과 명령을 정밀 저장하며 파라미터에서의 중요하지 않은 비트값을 비정밀 저장;

[0478] 단계 S2, 명령을 수신하고 명령을 제어정보로 디코딩하여 생성;

[0479] 단계 S3, 파라미터를 수신하고 제어정보에 따라 연산하며 연산 결과를 저장.

[0480] 여기서 상기 연산은 신경망 연산이고 파라미터는 신경망 파라미터로서 입력 뉴런, 가중치와 출력 뉴런을 포함한다.

[0481] 단계 S3은 진일보로, 입력 뉴런과 가중치를 수신하고 제어정보에 따라 신경망 연산을 완성하여 출력 뉴런을 얻으며 출력 뉴런을 저장 또는 출력하는 단계를 더 포함한다.

[0482] 진일보로, 입력 뉴런과 가중치를 수신하고 제어정보에 따라 신경망 연산을 완성하여 출력 뉴런을 얻는 상기 단계는, 입력 뉴런의 중요한 비트값과 가중치의 중요한 비트값을 수신하여 계산하는 단계; 또는 중요한 비트값과 중요하지 않은 비트값을 이어 완성한 입력 뉴런과 가중치를 수신하여 계산하는 단계를 포함한다.

[0483] 진일보로, 전용 명령을 캐시하는 단계; 입력 뉴런에 대해 정밀 캐시와 비정밀 캐시를 진행하는 단계; 가중치 데이터에 대해 정밀 캐시와 비정밀 캐시를 진행하는 단계; 출력 뉴런에 대해 정밀 캐시와 비정밀 캐시를 진행하는 단계를 더 포함한다.

[0484] 진일보로, 단계 S1 이전에 파라미터를 전처리하는 단계를 더 포함한다.

[0485] 본원 발명의 또 다른 실시예는 저장 유닛을 개시하는데 상기 저장 유닛은 4T SRAM 또는 3T SRAM이고 신경망 파라미터를 저장하며 여기서 상기 4T SRAM의 구체적인 구조는 도 2B에 도시된 구조를 참조하고 상기 3T SRAM의 구체적인 구조는 도 3B에 도시된 구조를 참조하면 되므로 여기서 더이상 설명하지 않는다.

[0486] 도 3a을 참조하면 도 3a은 본원 발명의 실시예에서 제공하는 동적 전압 조절 및 주파수 변조 장치(100)의 구조 모식도이다. 도 3a에 도시된 바와 같이 동적 전압 조절 및 주파수 변조 장치(100)는,

[0487] 상기 동적 전압 조절 및 주파수 변조와 연결된 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보를 실시간으로 수집하되, 상기 애플리케이션 시나리오 정보는 상기 칩이 신경망 연산을 통해 얻어지거나 또는 상기 칩과 연결된 센서가 수집한 정보인 정보 수집 유닛(101);

[0488] 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 칩이 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하기 위한 것인 전압 조절 및 주파수 변조 유닛(102)을 포함한다.

[0489] 본원 발명의 하나의 가능한 실시예에서 상기 칩의 작동 상태 정보는 상기 칩의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛(102)은,

[0490] 상기 칩의 운행속도가 타겟 속도보다 클 경우 상기 칩에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하고 상기 타겟 속도는 사용자 수요를 만족시킬 경우의 상기 칩의 운행속도이다.

[0491] 구체적으로 상기 정보 수집 유닛(101)은 그와 연결된 칩의 운행속도를 실시간으로 수집한다. 상기 칩의 운행속도는 상기 칩이 수행하는 임무가 상이함에 따라 상이한 유형의 속도일 수 있다. 상기 칩이 진행하는 작업이 동영상 이미지 처리일 경우 상기 칩의 운행속도는 상기 칩이 동영상 이미지 처리를 진행하는 프레임 레이트이고, 상기 칩이 진행하는 작업이 음성 인식일 경우 상기 칩의 운행속도는 상기 정보가 음성 인식을 진행하는 속도이다. 상기 전압 조절 및 주파수 변조 유닛(102)가 상기 칩의 운행속도가 상기 타겟 속도보다 크다고 결정할 경우, 즉 상기 칩의 운행속도가 사용자 수요를 만족시키는 상기 칩의 운행속도에 도달하였다고 결정할 경우 상기 칩에 제1 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 칩의 전력 손실을 저하시키도록 지시한다.

[0492] 예를 들어 설명하면 상기 칩이 진행하는 작업이 동영상 이미지 처리이고 상기 타겟 속도가 24프레임/초라고 가정한다. 상기 정보 수집 유닛은 상기 칩이 동영상 이미지 처리를 진행하는 프레임 레이트를 실시간으로 수집하고 현재 상기 칩이 동영상 이미지 처리를 진행하는 프레임 레이트는 54프레임/초이다. 상기 전압 조절 및 주파수 변조 유닛은 현재 상기 칩이 동영상 이미지 처리를 진행하는 프레임 레이트가 상기 타겟 속도보다 크다고 결정할 경우 칩에 제1 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 칩의 전력 손실을 저하시키도록 지시한다.

[0493] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 적어도 제1 유닛과 제2 유닛을 포함하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이며 상기 칩의 작동 상태 정보는 상기 제1 유닛의 운행속도와 제2 유닛의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 주파수 변조 및 전압 조절 유닛(102)은 또,

[0494] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제1 유닛의 운행시간이 상기 제2 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제2 유닛에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 제2 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0495] 구체적으로 상기 칩이 임무를 수행함에 있어서 상기 제1 유닛과상기 제2 유닛의 배합이 필요하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이다. 상기 정보 수집 유닛(101)은 상기 제1 유닛과 상기 제2 유닛의 운행속도를 실시간으로 수집한다. 상기 제1 유닛의 운행속도가 상기 제2 유닛의 운행속도보다 작다고 결정할 경우, 즉 상기 제1 유닛의 운행시간이 상기 제2 유닛의 운행시간을 초과한다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 제2 유닛에 상기 제2 전압 주파수 규제 정보를 발송하여 상기 제2 유닛으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 칩의 전반적인 운행속도에 영향을 주지 않는 전제하에 칩의 전반적인 전력 손실을 저하시키는 효과에 도달한다.

[0496] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하고 상기 주파수 변조 및 전압 조절 유닛(102)은 또,

[0497] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제2 유닛의 운행시간이 상기 제1 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제1 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 제1 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0498] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 적어도 N개의 유닛을 포함하고 상기 칩의 작동 상태 정보는 상기 적어도 N개의 유닛에서의 적어도 S개의 유닛의 작동 상태 정보를 포함하며 상기 N은 1보다 큰 정수이고 상기 S는 N보다 작거나 같은 정수이며 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하고 상기

전압 조절 및 주파수 변조 유닛(102)은,

[0499] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이고,

[0500] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛에서의 임의의 하나이다.

[0501] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛(102)은 또,

[0502] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓일 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0503] 구체적으로 상기 칩의 작동 과정에서 상기 정보 수집 유닛(101)은 상기 칩 내부의 적어도 S개의 유닛의 작동 상태 정보를 실시간으로 채집한다. 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 유닛(A)에 제4 전압 주파수 규제 정보를 발송하여 상기 유닛(A)으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시켜 상기 유닛(A)의 전력 손실을 저하시키도록 지시하고 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동 상태에 있을 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 유닛(A)에 제5 전압 주파수 규제 정보를 발송하여 상기 유닛(A)으로 하여금 그의 작동 주파수 또는 작동 전압을 향상시켜 상기 유닛(A)의 운행속도가 작동의 수요를 만족시키도록 한다.

[0504] 본원 발명의 하나의 가능한 실시예에서 상기 칩의 애플리케이션 시나리오가 이미지 인식일 경우 상기 애플리케이션 시나리오 정보는 인식 대기 이미지에서의 오브젝트의 개수이고 상기 전압 주파수 규제 정보는 제6 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛(102)은 또,

[0505] 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 작을 경우 상기 칩에 상기 제6 전압 주파수 규제 정보를 발송하되, 상기 제6 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0506] 구체적으로 상기 칩은 이미지 인식에 응용되고 상기 인식 대기 이미지에서의 오브젝트의 개수는 상기 칩이 신경망 알고리즘을 통해 얻은 것이며 상기 정보 수집 유닛(101)이 상기 칩에서 상기 인식 대기 이미지에서의 오브젝트의 개수(즉 상기 애플리케이션 시나리오 정보)를 획득한 후 상기 전압 조절 및 주파수 변조 유닛(102)이 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 작다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 제6 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하고 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 크다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키기 위한 전압 주파수 규제 정보를 발송한다.

[0507] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 오브젝트 레이블 정보이고 상기 전압 주파수 규제 정보는 제7 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛(102)은 또,

[0508] 상기 오브젝트 레이블 정보가 기설정 오브젝트 태그 집합에 속한다고 결정할 경우 상기 칩에 상기 제7 전압 주파수 규제 정보를 발송하되, 상기 제7 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 하기 위한 것이다.

[0509] 예를 들어 설명하면, 상기 기설정 오브젝트 태그 집합은 다수의 오브젝트 태그를 포함하고 상기 오브젝트 태그는 "사람", "개", "나무"와 "꽃" 일 수 있다. 상기 칩이 신경망 알고리즘을 통해 현재 애플리케이션 시나리오가 개를 포함한다고 할 경우 상기 칩은 상기 "개"를 포함하는 이 오브젝트 레이블 정보를 상기 정보 수집 유닛(101)에 전송한 다음 상기 주파수 변조 및 전압 조절 유닛(102)이 상기 오브젝트 레이블 정보가 "개"를 포함한다고 결정할 경우 상기 칩에 제7 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하고 상기 오브젝트 레이블 정보가 상기 기설정 오브젝트 태그 집합에 속하지 않는다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 전압 주파수 규제 정보를 발송한다.

[0510] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 음성 인식에 응용되고 상기 애플리케이션 시나리오 정보는 음성 입력 속도이며 상기 전압 주파수 규제 정보는 제8 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주

파수 변조 유닛은 또,

[0511] 상기 음성 입력 속도가 제2 임계값보다 작을 경우 상기 칩에 제8 전압 주파수 규제 정보를 발송하되, 상기 제8 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0512] 구체적으로 상기 칩의 애플리케이션 시나리오는 음성 인식이고 상기 칩의 입력 유닛은 일정한 속도로 칩에 음성을 입력한다. 상기 정보 수집 유닛(101)은 음성 입력 속도를 실시간으로 수집하고 상기 음성 입력 속도정보를 상기 전압 조절 및 주파수 변조 유닛(102)에 발송한다. 상기 전압 조절 및 주파수 변조 유닛(102)이 상기 음성 입력 속도가 제2 임계값보다 작다고 결정할 경우 상기 칩에 제8 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시한다. 상기 전압 조절 및 주파수 변조 유닛(102)이 상기 음성 입력 속도가 제2 임계값보다 크다고 결정할 경우 상기 칩에 상기 칩으로 하여금 그의 작동 전압을 향상시키기 위한 전압 주파수 규제 정보를 발송한다.

[0513] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 상기 칩이 음성 인식을 진행하여 얻은 키워드이고 상기 전압 주파수 규제 정보는 제9 전압 주파수 규제 정보를 포함하며 상기 주파수 변조 및 전압 조절 유닛은 또,

[0514] 상기 키워드가 기설정 키워드 집합일 경우 상기 칩에 상기 제9 전압 주파수 규제 정보를 발송하되, 상기 제9 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0515] 진일보로 상기 키워드가 상기 키워드 집합에 속하지 않을 경우상기 주파수 변조 및 전압 조절 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키기 위한 전압 조절 및 주파수 변조 정보를 발송한다.

[0516] 예를 들어 설명하면 상기 칩의 애플리케이션 시나리오가 음성 인식이 경우 상기 기설정 키워드 집합은 "이미지 뷰티", "신경망 알고리즘", "이미지 처리"와 "알리페이" 등 키워드를 포함한다. 만약 상기 애플리케이션 시나리오 정보가 "이미지 뷰티"라고 가정하면 상기 주파수 변조 및 전압 조절 유닛(102)은 상기 칩에 상기 제9 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하고 만약 상기 애플리케이션 시나리오 정보가 "촬영"일 경우 상기 주파수 변조 및 전압 조절 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키기 위한 전압 조절 및 주파수 변조 정보를 발송한다.

[0517] 본원 발명의 하나의 가능한 실시예에서 상기 칩이 기계 번역에 응용될 경우 상기 애플리케이션 시나리오 정보는 문자 입력 속도 또는 번역 대기 이미지에서의 문자의 수량이며 상기 전압 주파수 규제 정보는 제10 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0518] 상기 문자 입력 속도가 제3 임계값 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 작을 경우 상기 칩에 상기 제10 전압 주파수 규제 정보를 발송하되, 상기 제10 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0519] 구체적으로 상기 칩은 기계 번역에 응용되고 상기 정보 수집 유닛(101)가 수집한 애플리케이션 시나리오 정보는 문자 입력 속도 또는 번역 대기 이미지에서의 문자의 수량이며 상기 애플리케이션 시나리오 정보를 상기 전압 조절 및 주파수 변조 유닛(102)에 전송한다. 상기 문자 입력 속도가 제3 임계값보다 작거나 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 작다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 제10 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압을 저하시키도록 지시하고 상기 문자 입력 속도가 제3 임계값보다 크거나 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 크다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압을 향상시키기 위한 전압 주파수 규제 정보를 발송한다.

[0520] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 외부의 광도이고 상기 전압 주파수 규제 정보는 제11 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0521] 상기 외부의 광도가 제5 임계값보다 작을 경우 상기 칩에 상기 제11 전압 주파수 규제 정보를 발송하되, 상기 제11 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0522] 구체적으로 상기 외부의 광도는 상기 칩과 연결된 조도 센서가 수집하여 획득한 것이다. 상기 정보 수집 유닛(101)은 상기 광도를 획득한 후 상기 광도를 상기 전압 조절 및 주파수 변조 유닛(102)에 전송한다. 상기 광도가 제5 임계값보다 작다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 제11 전압 주파수 규제 정보를 발송하여 상기 칩으로 하여금 그의 작동 전압을 저하시키도록 지시하고 상기 광도가 제5 임계값보다 크다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 전압 주파수 규제 정보를 발송한다.

[0523] 본원 발명의 하나의 가능한 실시예에서 상기 칩은 이미지 뷰티에 응용되고 상기 전압 주파수 규제 정보는 제12 전압 주파수 규제 정보와 제13 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0524] 상기 애플리케이션 시나리오 정보가 안면 이미지일 경우 상기 칩에 상기 제12 전압 주파수 규제 정보를 발송하되, 상기 제12 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이고;

[0525] 상기 애플리케이션 시나리오 정보가 안면 이미지가 아닐 경우 상기 칩에 제13 전압 주파수 규제 정보를 발송하되, 상기 제13 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0526] 본원 발명의 하나의 가능한 실시예에서 상기 칩이 음성 인식에 응용 될 경우 상기 애플리케이션 시나리오 정보는 음성 강도이며 상기 음성 강도가 제6 임계값보다 클 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 전압 주파수 규제 정보를 발송하고 상기 음성 강도가 제6 임계값보다 작을 경우 상기 전압 조절 및 주파수 변조 유닛(102)은 상기 칩에 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 전압 주파수 규제 정보를 발송한다.

[0527] 설명해야 할 것은 상기 시나리오 정보는 광도, 음성 강도 등과 같은 센서가 수집한 외부 시나리오의 정보 일 수 있다. 상기 애플리케이션 시나리오 정보는 인공지능 알고리즘에 근거하여 산출한 정보 일 수도 있는데 예하면 오브젝트 인식 임무에서 칩의 실시간 계산결과정보를 정보 수집 유닛에 피드백하되, 상기 정보는 시나리오의 오브젝트 개수, 안면 이미지, 오브젝트 태그 키워드 등 정보를 포함한다.

[0528] 선택적으로, 상기 인공지능 알고리즘은 신경망 알고리즘을 포함하나 이에 한정되지 않는다.

[0529] 알 수 있다 시피 본 발명의 실시예의 방안에서 동적 전압 조절 및 주파수 변조 장치는 실시간으로 그와 연결된 칩 및 그 내부의 각 유닛의 작동 상태 정보 또는 칩의 애플리케이션 시나리오 정보를 수집하고 칩 및 그 내부의 각 유닛의 작동 상태 정보 또는 칩의 애플리케이션 시나리오 정보에 근거하여 칩 또는 그 내부의 각 유닛의 작동 주파수 또는 작동 전압을 조절함으로써 칩의 전반적인 운행의 전력 손실을 저하시키는 목적에 도달한다.

[0530] 도 2를 참조하면 도 3b는 본원 발명의 실시예에서 제공하는 동적 전압 조절 및 주파수 변조 애플리케이션 시나리오의 모식도이다. 도 3b에 도시된 바와 같이 상기 콘볼루션 연산장치는 동적 전압 조절 및 주파수 변조 장치(210)및 상기 동적 전압 조절 및 주파수 변조 장치와 연결된 칩(220)을 포함한다.

[0531] 여기서 상기 칩(220)은 제어 유닛(221), 저장 유닛(222)과 연산 유닛(223)을 포함한다. 상기 칩(220)은 이미지 처리, 음성 처리 등 임무를 진행할 수 있다.

[0532] 여기서 상기 동적 전압 조절 및 주파수 변조 장치(210)는 상기 칩(220)의 작동 상태 정보를 실시간으로 수집한다. 상기 칩(220)의 작동 상태 정보는 상기 칩(220)의 운행속도, 제어 유닛(221)의 운행속도, 저장 유닛(222)의 운행속도와 연산 유닛(223)의 운행속도를 포함한다.

[0533] 본원 발명의 하나의 가능한 실시예에서 칩(220)이 한차례의 임무를 수행할 경우 동적 전압 조절 및 주파수 변조 장치(210)는 저장 유닛(222)의 운행속도와 연산 유닛(223)의 운행속도에 근거하여 저장 유닛(222)의 운행시간이 연산 유닛(223)의 운행시간을 초과한다고 결정하고 동적 전압 조절 및 주파수 변조 장치(210)는 이번 임무를 수행하는 과정에서 저장 유닛(222)이 난관으로 되어 연산 유닛(223)이 현재의 연산작업을 완성한 후 저장 유닛(222)이 임무를 판독함과 동시에 그가 판독한 데이터를 연산 유닛(223)에 전송하여야만 연산 유닛(223)은 이번 저장 유닛(222)이 전송한 데이터에 근거하여 연산작업을 진행할 수 있다. 동적 전압 조절 및 주파수 변조 장치(210)는 연산 유닛(223)에 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 연산 유닛(223)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 연산 유닛(223)의 운행속도를 저하시키도록 지시하여 임무의 완성 시간에 영향을 미치지 않는 상황에서 칩(220)의 전반적인 운행의 전력 손실을

저하시킨다.

[0534] 본원 발명의 하나의 가능한 실시예에서 칩(220)이 한차례의 임무를 수행할 경우 동적 전압 조절 및 주파수 변조 장치(210)가 저장 유닛(222)의 운행속도와 연산 유닛(223)의 운행속도에 근거하여 저장 유닛(222)의 운행시간이 연산 유닛(223)의 운행시간보다 늦다고 결정하면 동적 전압 조절 및 주파수 변조 장치(210)는 이번 임무를 수행하는 과정에서 연산 유닛(223)이 난관으로 된다. 저장 유닛(222)이 데이터 판독을 완성한 후 연산 유닛(223)은 아직 현재의 연산작업을 완성하지 못하였으므로 저장 유닛(222)은 연산 유닛(223)이 현재의 연산작업을 완성할 때까지 대기한 후에야만 판독한 데이터를 연산 유닛(223)에 전송할 수 있다. 동적 전압 조절 및 주파수 변조 장치(210)는 저장 유닛(222)에 제2 전압 주파수 규제 정보를 발송하는데 상기 제2 전압 주파수 규제 정보는 저장 유닛(222)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 저장 유닛(222)의 운행속도를 저하시키도록 지시하여 임무의 완성 시간에 영향을 미치지 않는 상황에서 칩(220)의 전반적인 운행의 전력 손실을 저하시킨다.

[0535] 본원 발명의 하나의 가능한 실시예에서 동적 전압 조절 및 주파수 변조 장치(210)는 칩(220)의 운행속도를 실시간으로 수집한다. 동적 전압 조절 및 주파수 변조 장치(210)가 칩(220)의 운행속도가 타겟 운행속도보다 크다고 결정할 경우 상기 타겟 운행속도는 사용자 수요를 만족시킬 수 있는 운행속도이고 동적 전압 조절 및 주파수 변조 장치(210)는 칩(220)에 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 칩(220)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 칩(220)의 운행의 전력 손실을 저하시킨다.

[0536] 예를 들어 설명하면, 칩(220)은 영상 처리를 진행하기 위한 것으로, 예하면 정상적인 상황에서 사용자가 영상 처리를 진행하고자 하는 프레임 레이트는 30프레임보다 낮지 않고 만약 이때 칩(220)이 실제로 영상 처리를 진행한 프레임 레이트가 100이라고 가정하면 동적 전압 조절 및 주파수 변조 장치는 칩(220)에 전압 주파수 규제 정보를 발송하며 상기 전압 주파수 규제 정보는 칩(220)으로 하여금 작동 전압 또는 작동 주파수를 저하시켜 영상 처리된 프레임 레이트를 30프레임좌우로 저하시키도록 지시한다.

[0537] 본원 발명의 하나의 가능한 실시예에서 동적 전압 조절 및 주파수 변조 장치(210)는 칩(220)에서의 각 유닛(제어 유닛(221), 저장 유닛(222)과 연산 유닛(223)을 포함)의 작동 상태를 실시간으로 모니터링한다. 동적 전압 조절 및 주파수 변조 장치(220)가 각 유닛에서의 임의의 한 유닛이 유휴상태에 있따고 결정할 경우 상기 유닛에 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛의 작동 전압 또는 작동 주파수를 저하시켜 칩(220)의 전력 손실을 저하시키도록 지시한다. 상기 유닛이 다시 작동 상태에 있을 경우 동적 전압 조절 및 주파수 변조 장치(210)는 상기 유닛에 제5 전압 주파수 규제 정보를 발송하여 상기 유닛의 작동 전압 또는 작동 주파수를 향상시켜 상기 유닛의 운행속도가 작동 수요를 만족시키도록 한다. 알 수 있다 시피 본원 발명의 실시예의 방안에서 동적 전압 조절 및 주파수 변조 장치(210)는 칩 및 그 내부의 각 유닛의 운행속도정보를 실시간으로 수집하고 칩 및 그 내부의 각 유닛의 운행속도정보에 근거하여 칩 또는 그 내부의 각 유닛의 작동 주파수 또는 작동 전압을 저하시킴으로써 칩의 전반적인 운행의 전력 손실을 저하시키는 목적에 도달한다.

[0538] 도 3c을 참조하면 도 3c은 본원 발명의 실시예에서 제공하는 다른 동적 전압 조절 및 주파수 변조 애플리케이션 시나리오의 모식도이다. 3C에 도시된 바와 같이 상기 콘볼루션 연산장치는 동적 전압 조절 및 주파수 변조 장치(317), 레지스터 유닛(312), 인터커넥트 모듈(313), 연산 유닛(314), 제어 유닛(315)과 데이터 액세스 유닛(316)을 포함한다.

[0539] 여기서 연산 유닛(314)은 덧셈 계산기, 곱셈 계산기, 컴퍼레이터와 활성화 연산기에서의 적어도 두 가지를 포함한다.

[0540] 인터커넥트 모듈(313)은 연산 유닛(314)에서의 계산기의 연결관계를 제어하여 상기 적어도 두 가지 계산기로 하여금 상이한 계산 토폴로지 구조를 조성하도록 한다.

[0541] 레지스터 유닛(312)(레지스터 유닛, 명령 캐시, 스크래치패드 메모리 일 수 있음)은 상기 연산 명령, 데이터 블록이 저장매체에서의 주소, 연산 명령과 대응되는 계산 토폴로지 구조를 저장한다.

[0542] 선택적으로, 상기 콘볼루션 연산장치는 저장매체(311)를 더 포함한다.

[0543] 저장매체(311)는 오프 칩 메모리 일 수 있는데 물론 실제 응용에서는 데이터 블록을 저장하기 위한 온 칩 메모리 일 수도 있으며 상기 데이터 블록은 구체적으로 n차원 데이터 일 수 있고 n은 1보다 크거나 같은 정수 일 수 있는 바, 예하면 n=1일 경우 1차원 데이터, 즉 벡터이고 n=2일 경우 2차원 데이터, 즉 매트릭스이며 n=3 또는 3

이상 일 경우 다차원 데이터 일 수 있다.

[0544] 제어 유닛(315)은 레지스터 유닛(312) 내에서 연산 명령, 상기 연산 명령과 대응되는 작동 도메인 및 상기 연산 명령과 대응되는 제1 계산 토폴로지 구조를 추출하고 상기 연산 명령을 수행 명령으로 디코딩하며 상기 수행 명령은 연산 유닛(314)을 제어하여 연산작업을 수행하도록 하고 상기 작동 도메인을 데이터 액세스 유닛(316)에 전송하며 상기 계산 토폴로지 구조를 인터커넥트 모듈(313)에 전송한다.

[0545] 데이터 액세스 유닛(316)은 저장매체(311)에서 상기 작동 도메인과 대응되는 데이터 블록을 추출하고 상기 데이터 블록을 인터커넥트 모듈(313)에 전송한다.

[0546] 인터커넥트 모듈(313)은 제1 계산 토폴로지 구조의 데이터 블록을 수신한다.

[0547] 본원 발명의 하나의 가능한 실시예에서 인터커넥트 모듈(313)은 또 제1 계산 토폴로지 구조에 근거하여 데이터 블록을 다시 배치한다.

[0548] 연산 유닛(314)은 상기 명령을 수행하고 연산 유닛(314)의 계산기를 호출하여 상기 데이터 블록에 대해 연산작업을 수행함으로써 연산 결과를 얻으며 상기 연산 결과를 데이터 액세스 유닛(316)에 전송하여 저장매체(312) 내에 저장한다.

[0549] 본원 발명의 하나의 가능한 실시예에서 연산 유닛(314)은 또 제1 계산 토폴로지 구조 및 상기 수행 명령에 따라 계산기를 호출하여 다시 배치된 데이터 블록에 대해 연산작업을 수행하여 연산 결과를 얻고 상기 연산 결과를 데이터 액세스 유닛(316)에 전송하여 저장매체(312) 내에 저장한다.

[0550] 하나의 가능한 실시예에서 인터커넥트 모듈(313)은 또 연산 유닛(314)에서의 계산기의 연결관계를 제어함으로써 제1 계산 토폴로지 구조를 형성한다.

[0551] 동적 전압 조절 및 주파수 변조 장치(317)는 전반적인 콘볼루션 연산장치의 작동상태를 모니터링하고 그의 전압과 주파수에 대해 동적으로 규제한다.

[0552] 이하 상이한 연산 명령을 통해 상기 콘볼루션 연산장치의 구체적인 계산방법을 설명하는데 여기서 연산 명령은 콘볼루션 계산명령을 예로 들수 있고 상기 콘볼루션 계산명령은 신경망에 응용될 수 있으므로 상기 콘볼루션 계산명령은 콘볼루션 신경망으로 불릴 수도 있다. 콘볼루션 계산명령에 있어서 그가 실제로 수행해야 할 공식은:

[0553] $$s = s(\sum wx_i + b)$$

[0554] 여기서 콘볼루션 커널w(다수의 데이터를 포함할 수 있음)에 입력 데이터Xi를 곱하여 합계를 구한 다음 선택적으로 바이어스b를 가하고 그 다음 선택적으로 활성화 연산s(h)도 진행하여 최종적인 출력결과S를 얻는다. 상기 공식에 따라 얻은 상기 계산 토폴로지 구조는 곱셈 연산기-덧셈 연산기-(선택적으로)활성화 연산기이다. 상기 콘볼루션 계산명령은 명령 집합을 포함할 수 있는데 상기 명령 집합은 상이한 기능의 콘볼루션 신경망 COMPUTE 명령 및 CONFIG 명령, IO 명령, NOP 명령, JUMP 명령과 MOVE 명령을 포함한다.

[0555] 한가지 실시예에서 COMPUTE 명령은 다음과 같은 명령을 포함한다.

[0556] 콘볼루션 연산 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리 또는 스칼라 레지스터 파일)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한다.

[0557] 콘볼루션 신경망 sigmoid 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리 또는 스칼라 레지스터 파일)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 sigmoid 활성화를 진행;

[0558] 콘볼루션 신경망 TanH 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 TanH 활성화를 진행;

[0559] 콘볼루션 신경망 ReLU 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 ReLU활성화를 진행; 및

[0560] 콘볼루션 신경망 group 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래

치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 group을 나눈 다음 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 활성화를 진행한다.

[0561] CONFIG 명령, 매 층마다의 인공 신경망 계산이 시작되기 전에 현재 층의 계산에 필요한 여러 가지 상수를 배치한다.

[0562] IO 명령, 외부 저장공간으로부터 계산에 필요한 입력 데이터를 판독하고 계산이 완료된 후 데이터를 외부공간에 저장한다.

[0563] NOP 명령, 현재 상기 콘볼루션 연산장치 내부의 모든 컨트롤 신호 캐시 행렬에서의 컨트롤 신호를 정리하여 NOP 명령 이전의 모든 명령이 모두 수행 완료되도록 담보한다. NOP 명령 자체는 그 어떤 작업도 포함하지 않음;

[0564] JUMP 명령, 명령 저장 유닛으로부터 판독될 다음 명령 주소의 점프를 제어하는 것을 담당함으로써 제어 흐름의 점프를 실현;

[0565] MOVE 명령, 상기 콘볼루션 연산장치 내부 주소공간의 어느 한 주소의 데이터를 상기 콘볼루션 연산장치 내부 주소공간의 다른 한 주소로 이동시키는 것을 담당하되, 상기 과정은 연산 유닛과 별도로 수행과정에서 연산 유닛의 리소스를 점용하지 않는다.

[0566] 상기 콘볼루션 연산장치가 콘볼루션 계산명령을 수행하는 방법은 구체적으로 다음과 같을 수 있다.

[0567] 제어 유닛(315)은 레지스터 유닛(312) 내에서 콘볼루션 계산명령, 상기 콘볼루션 계산명령과 대응되는 작동 도메인 및 콘볼루션 계산명령과 대응되는 제1 계산 토폴로지 구조(곱셈 연산기-덧셈 연산기-덧셈 연산기-활성화 연산기)를 추출하고 제어 유닛은 상기 작동 도메인을 데이터 액세스 유닛에 전송하며 상기 제1 계산 토폴로지 구조를 인터커넥트 모듈에 전송한다.

[0568] 데이터 액세스 유닛(316)은 저장매체(311) 내에서 상기 작동 도메인과 대응되는 콘볼루션 커널 w과 바이어스 b(b가 0일 경우 바이어스 b를 추출할 필요가 없음)를 추출하고 콘볼루션 커널 w과 바이어스 b를 연산 유닛(314)에 전송한다.

[0569] 연산 유닛(314)의 곱셈 연산기는 콘볼루션 커널 w과 입력 데이터 Xi에 곱셈 연산을 진행하여 제1 결과를 얻고 제1 결과를 덧셈 연산기에 입력하여 덧셈 연산을 수행하여 제2 결과를 얻으며 제2 결과와 바이어스 b에 덧셈 연산을 진행하여 제3 결과를 얻고 제3 결과를 활성화 연산기에 입력하여 활성화 연산을 수행함으로써 출력 결과 S를 얻으며 출력 결과 S를 데이터 액세스 유닛에 전송하여 저장매체 내에 저장한다. 여기서 매 단계 뒤에는 모두 직접적으로 출력 결과를 데이터 액세스 유닛에 전송하여 저장매체 내에 저장할 수 있어 아래의 단계를 진행할 필요가 없다. 이 외에 제2 결과와 바이어스 b에 덧셈 연산을 진행하여 제3 결과를 얻는 이 단계는 선택적인 것으로서 즉 b가 0일 경우 이 단계는 필요하지 않게 된다. 이 외에 덧셈 연산과 곱셈 연산의 순서는 바뀔 수 있다.

[0570] 선택적으로, 상기 제1 결과는 다수의 곱셈 연산의 결과를 포함할 수 있다.

[0571] 본원 발명의 하나의 가능한 실시예에서 본원 발명의 실시예는 신경망 프로세서를 제공하는데 이는 상기 콘볼루션 연산장치를 포함한다.

[0572] 상기 신경망 프로세서는 인공 신경망 연산을 수행하고 음성 인식, 이미지 인식, 번역 등 인공지능의 응용을 실현한다.

[0573] 이 콘볼루션 계산 임무에서 상기 동적 전압 조절 및 주파수 변조 장치(317)의 작동 과정은 다음과 같다.

[0574] 상황1, 상기 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 동적 전압 조절 및 주파수 변조 장치(317)는 실시간으로 상기 신경망 프로세서의 데이터 액세스 유닛(316)과 연산 유닛(314)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(317)가 데이터 액세스 유닛(316)과 연산 유닛(314)의 운행속도에 근거하여 데이터 액세스 유닛(316)의 운행시간이 연산 유닛(314)의 운행시간을 초과한다고 결정할 경우 동적 전압 조절 및 주파수 변조 장치(317)는 콘볼루션 연산을 진행하는 과정에서 데이터 액세스 유닛(316)이 난관이 됨을 결정하고 연산 유닛(314)은 현재의 콘볼루션 연산작업을 완성한 후 데이터 액세스 유닛(316)이 판독 임무를 수행 완료함과 동시에 그가 판독한 데이터를 연산 유닛(314)에 전송하기를 대기하여야만 상기 연산 유닛(314)은 이번 데이터 액세스 유닛(316)이 전송한 데이터에 근거하여 콘볼루션 연산작업을 진행할 수 있다. 동적 전압 조절 및 주파수 변조 장치(317)는 연산 유닛(314)에 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 연산 유닛(314)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 연산 유닛(314)의 운행속도를

저하시키도록 지시함으로써 연산 유닛(314)의 운행속도와 데이터 액세스 유닛(316)의 운행속도가 매칭되도록 하여 연산 유닛(314)의 전력 손실을 저하시킴으로써 연산 유닛(314)이 유휴한 상황이 발생하는 것을 방지하고 최종적으로 임무의 완성시간에 영향을 미치지 않는 상황에서 상기 신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0575] 상황2, 상기 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 동적 전압 조절 및 주파수 변조 장치(317)는 실시간으로 상기 신경망 프로세서의 데이터 액세스 유닛(316)과 연산 유닛(314)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(317)가 데이터 액세스 유닛(316)과 연산 유닛(314)의 운행속도에 근거하여 연산 유닛(314)의 운행시간이 상기 데이터 액세스 유닛(316)의 운행시간을 초과한다고 결정할 경우 동적 전압 조절 및 주파수 변조 장치(317)는 콘볼루션 연산을 진행하는 과정에서 연산 유닛(314)이 난관이 됨을 결정할 수 있고 데이터 액세스 유닛(316)은 현재의 데이터 판독 작업을 완료한 후 연산 유닛(314)이 현재의 콘볼루션 연산 작업을 수행하기를 대기해야만이 데이터 액세스 유닛(316)은 판독한 데이터를 상기 연산 유닛(314)에 전송할 수 있다. 동적 전압 조절 및 주파수 변조 장치(317)는 데이터 액세스 유닛(316)에 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 데이터 액세스 유닛(316)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 데이터 액세스 유닛(316)의 운행속도를 저하시키도록 지시함으로써 데이터 액세스 유닛(316)의 운행속도와 연산 유닛(314)의 운행속도가 매칭되도록 하여 데이터 액세스 유닛(316)의 전력 손실을 저하시키고 데이터 액세스 유닛(316)이 유휴한 상황이 발생하는 것을 방지하여 최종적으로 임무의 완성 시간에 영향을 미치지 않는 상황에서 상기 신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0576] 상기 신경망 프로세서는 인공 신경망 연산을 수행하는데 인공지능 애플리케이션을 진행할 경우 동적 전압 조절 및 주파수 변조 장치(317)는 실시간으로 상기 신경망 프로세서가 인공지능 애플리케이션을 진행하는 동작 파라미터를 수집하고 상기 작동 파라미터에 근거하여 상기 신경망 프로세서의 작동 전압 또는 작동 주파수를 조절한다.

[0577] 구체적으로 상기 인공지능 애플리케이션은 동영상 이미지 처리, 오브젝트 인식, 기계 번역, 음성 인식과 이미지 뷰티 등 일 수 있다.

[0578] 상황3, 상기 신경망 프로세서가 동영상 이미지 처리를 진행할 경우 동적 전압 조절 및 주파수 변조 장치(317)는 실시간으로 상기 신경망 프로세서가 동영상 이미지 처리를 진행하는 프레임 레이트를 수집한다. 상기 동영상 이미지 처리의 프레임 레이트가 타겟 프레임 레이트를 초과할 경우 상기 타겟 프레임 레이트는 사용자가 정상적으로 수요하는 동영상 이미지 처리프레임 레이트이고 동적 전압 조절 및 주파수 변조 장치(317)는 상기 신경망 프로세서에 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하여 사용자의 정상적인 동영상 이미지 처리 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력손실을 저하시킨다.

[0579] 상황4, 상기 신경망 프로세서가 음성 인식을 진행할 경우 동적 전압 조절 및 주파수 변조 장치(317)는 실시간으로 상기 신경망 프로세서의 음성 인식 속도를 수집한다. 상기 신경망 프로세서의 음성 인식 속도가 사용자의 실제 음성 인식 속도를 초과할 경우 동적 전압 조절 및 주파수 변조 장치(317)는 상기 신경망 프로세서에 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하고 사용자의 정상적인 음성 인식 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력 손실을 저하시킨다.

[0580] 상황5, 동적 전압 조절 및 주파수 변조 장치(317)는 상기 신경망 프로세서에서의 각 유닛 또는 모듈(저장매체(311), 레지스터 유닛(312), 인터커넥트 모듈(313), 연산 유닛(314), 제어 유닛(315), 데이터 액세스 유닛(316))의 작동 상태를 실시간으로 모니터링한다. 상기 신경망 프로세서의 각 유닛 또는 모듈에서의 임의의 한 유닛 또는 모듈이 유휴상태에 있을 경우 동적 전압 조절 및 주파수 변조 장치(317)는 상기 유닛 또는 모듈에 제5 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 저하시킴으로써 상기 유닛 또는 모듈의 전력 손실을 저하시킨다. 상기 유닛 또는 모듈이 다시 작동 상태에 놓일 경우 동적 전압 조절 및 주파수 변조 장치(317)는 상기 유닛 또는 모듈에 제6 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 향상시킴으로써 상기 유닛 또는 모듈의 운행속도로 하여금 작업의 수요를 만족시키도록 한다.

[0581] 도 3d를 참조하면 도 3d는 본원 발명의 실시예에서 제공하는 또 다른 동적 전압 조절 및 주파수 변조 애플리케이션 시나리오의 모식도이다. 도 3d에 도시된 바와 같이 상기 콘볼루션 연산장치는 동적 전압 조절 및 주파수 변조 장치(7), 명령 저장 유닛(1), 컨트롤러 유닛(2), 데이터 액세스 유닛(3), 인터커넥트 모듈(4), 메인 연산

모듈(5)과 다수의 서브 연산 모듈(6)을 포함한다. 명령 저장 유닛(1), 컨트롤러 유닛(2), 데이터 액세스 유닛(3), 인터커넥트 모듈(4), 메인 연산 모듈(5)과 서브 연산 모듈(6)은 모두 하드웨어 회로(예하면 FPGA, CGRA, 특정 용도 지향 집적 회로ASIC, 아날로그 회로와 멤리스터 등을 포함하나 이에 한정되지 않음)를 통해 실현할 수 있다.

[0582] 명령 저장 유닛(1)은 데이터 액세스 유닛(3)을 통해 명령을 판독하고 판독된 명령을 저장한다.

[0583] 컨트롤러 유닛(2)은 명령 저장 유닛(1)으로부터 명령을 판독하고명령을 기타 모듈의 행동을 제어하는 컨트롤 신호로 디코딩하여 데이터 액세스 유닛(3), 메인 연산 모듈(5)과 서브 연산 모듈(6) 등과 같은 기타 모듈에 발송한다.

[0584] 데이터 액세스 유닛(3)은 외부 주소 공간을 액세스하여 상기 콘볼루션 연산장치 내부의 각 저장 유닛에 데이터를 판독 기록함으로써 데이터의 로딩과 저장을 완성한다.

[0585] 인터커넥트 모듈(4)은 메인 연산 모듈과 서브 연산 모듈을 연결하기 위한 것으로 상이한 인터커넥트 토폴로지(예하면 트리 구조, 환형 구조, 메쉬 구조, 분급 인터커넥트, 버스 구조 등)를 실현할 수 있다.

[0586] 동적 전압 조절 및 주파수 변조 장치(7)는 상기 데이터 액세스 유닛(3)과 상기 메인 연산 유닛(5)의 작동 상태 정보를 실시간으로 획득하고 상기 데이터 액세스 유닛(3)과 상기 메인 연산 유닛(5)의 작동 상태 정보에 근거하여 상기 데이터 액세스 유닛(3)과 상기 메인 연산 모듈(5)의 작동 전압 또는 작동 주파수를 조절한다.

[0587] 본원 발명의 하나의 가능한 실시예에서 본 발명의 실시예는 신경망 프로세서를 제공하는데 이는 상기 콘볼루션 연산장치를 포함한다.

[0588] 상기 신경망 프로세서는 인공 신경망 연산을 수행하고 음성 인식, 이미지 인식, 번역 등 인공지능의 응용을 실현한다.

[0589] 이 콘볼루션 계산임무에서 동적 전압 조절 및 주파수 변조 장치(7)의 작동 과정은 다음과 같다.

[0590] 상황1, 상기 콘볼루션 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서의 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(7)가 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도에 근거하여 데이터 액세스 유닛(3)의 운행시간이 메인 연산 모듈(5)의 운행시간을 초과한다고 결정할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 콘볼루션 연산을 진행하는 과정에서 데이터 액세스 유닛(3)이 난관이 됨을 결정하고 메인 연산 모듈(5)은 현재의 콘볼루션 연산작업을 완성한 후 상기 데이터 액세스 유닛(3)이 판독 임무를 수행 완료함과 동시에 그가 판독한 데이터를 메인 연산 모듈(5)에 전송하기를 대기하여야만 메인 연산 모듈(5)은 이번 데이터 액세스 유닛(3)이 전송한 데이터에 근거하여 콘볼루션 연산작업을 진행할 수 있다. 동적 전압 조절 및 주파수 변조 장치(7)는 메인 연산 모듈(5)에 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 메인 연산 모듈(5)로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 메인 연산 모듈(5)의 운행속도를 저하시키도록 지시함으로써메인 연산 모듈(5)의 운행속도와 데이터 액세스 유닛(3)의 운행속도가 매칭되도록 하여 메인 연산 모듈(5)의 전력 손실을 저하시킴으로써 메인 연산 모듈(5)이 유휴한 상황이 발생하는 것을 방지하고 최종적으로 임무의 완성시간에 영향을 미치지 않는 상황에서 상기 신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0591] 상황2, 상기 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서의 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(3)는 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도에 근거하여 메인 연산 모듈(5)의 운행시간이 데이터 액세스 유닛(3)의 운행시간을 초과한다고 결정할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 콘볼루션 연산을 진행하는 과정에서 메인 연산 모듈(5)이 난관이 됨을 결정할 수 있고 데이터 액세스 유닛(3)은 현재의 데이터 판독 작업을 완료한 후 메인 연산 모듈(5)이 현재의 콘볼루션 연산작업을 완성한 후에야만 데이터 액세스 유닛(3)이 그가 판독한 데이터를 메인 연산 모듈(5)에 전송할 수 있다. 동적 전압 조절 및 주파수 변조 장치(7)는 데이터 액세스 유닛(3)에 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 데이터 액세스 유닛(3)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 데이터 액세스 유닛(3)의 운행속도를 저하시키도록 지시함으로써 데이터 액세스 유닛(3)의 운행속도와 메인 연산 모듈(5)의 운행속도가 매칭되도록 하여 데이터 액세스 유닛(3)의 전력 손실을 저하시키고 데이터 액세스 유닛(3)이 유휴한 상황이 발생하는 것을 방지하여 최종적으로 임무의 완성 시간에 영향을 미치지 않는 상황에서 상기

신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0592] 상기 신경망 프로세서는 인공 신경망 연산을 수행하는데 인공지능 애플리케이션을 진행할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서가 인공지능 애플리케이션을 진행하는 동작 파라미터를 수집하고 상기 작동 파라미터에 근거하여 상기 신경망 프로세서의 작동 전압 또는 작동 주파수를 조절한다.

[0593] 구체적으로 상기 인공지능 애플리케이션은 동영상 이미지 처리, 오브젝트 인식, 기계 번역, 음성 인식과 이미지 뷰티 등 일 수 있다.

[0594] 상황3, 상기 신경망 프로세서가 동영상 이미지 처리를 진행할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서가 동영상 이미지 처리를 진행하는 프레임 레이트를 수집한다. 상기 동영상 이미지 처리의 프레임 레이트가 타겟 프레임 레이트를 초과할 경우 상기 타겟 프레임 레이트는 사용자가 정상적으로 수요하는 동영상 이미지 처리프레임 레이트이고 동적 전압 조절 및 주파수 변조 장치(7)는 상기 신경망 프로세서에 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하여 사용자의 정상적인 동영상 이미지 처리 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력손실을 저하시킨다.

[0595] 상황4, 상기 신경망 프로세서가 음성 인식을 진행할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서의 음성 인식 속도를 수집한다. 상기 신경망 프로세서의 음성 인식 속도가 사용자의 실제 음성 인식 속도를 초과할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 상기 신경망 프로세서에 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하고 사용자의 정상적인 음성 인식 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력 손실을 저하시킨다.

[0596] 상황5, 동적 전압 조절 및 주파수 변조 장치(7)는 상기 신경망 프로세서에서의 각 유닛 또는 모듈(명령(1), 컨트롤러 유닛(2), 데이터 액세스 유닛(3), 인터커넥트 모듈(4), 메인 연산 모듈(5)과 서브 연산 모듈(6)을 포함)의 작동 상태를 실시간으로 모니터링한다. 상기 신경망 프로세서의 각 유닛 또는 모듈에서의 임의의 한 유닛 또는 모듈이 유휴상태에 있을 경우 동적 전압 조절 및 주파수 변조 장치(7)는 상기 유닛 또는 모듈에 제5 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 저하시킴으로써 상기 유닛 또는 모듈의 전력 손실을 저하시킨다. 상기 유닛 또는 모듈이 다시 작동 상태에 놓일 경우 동적 전압 조절 및 주파수 변조 장치(7)는 상기 유닛 또는 모듈에 제6 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 향상시킴으로써 상기 유닛 또는 모듈의 운행속도로 하여금 작업의 수요를 만족시키도록 한다.

[0597] 도 3e를 참조하면 도 3e는 인터커넥트 모듈(4)의 한가지 실시형태인 H트리 모듈을 예시적으로 도시한다. 인터커넥트 모듈(4)은 메인 연산 모듈(5)과 다수의 서브 연산 모듈(6) 사이의 다수의 노드로 구성된 이진 트리 통로인 데이터 통로를 구성하고 매 노드마다 상류의 데이터를 마찬가지로 하류의 두 개의 노드에 발송하며 하류의 두 개의 노드가 리턴한 데이터를 합병하여 상류의 노드에 리턴한다. 예하면 콘볼루션 신경망이 계산을 시작하는 단계에서 메인 연산 모듈(5) 내의 뉴런 데이터는 인터커넥트 모듈(4)을 통해 각 서브 연산 모듈(6)에 발송되고; 서브 연산 모듈(6)의 계산 과정이 완성된 후 서브 연산 모듈의 계산 과정이 완성된 후 서브 연산 모듈마다 출력한 뉴런의 값은 인터커넥트 모듈(4)에서 단계적으로 하나의 완전한 뉴런으로 조성된 벡터로 결합된다. 예를 들어 설명하면 만약 장치에 모두 N개의 서브 연산 모듈이 있을 경우 입력 데이터Xi는 N개의 서브 연산 모듈에 각각 발송되고 매 하나의 서브 연산 모듈은 입력 데이터Xi와 상기 서브 연산 모듈과 대응되는 콘볼루션 커널을 콘볼루션 연산하여 스칼라 데이터를 얻으며 각 서브 연산 모듈의 스칼라 데이터는 인터커넥트 모듈(4)에 의해 N개의 요소를 포함하는 하나의 중간 벡터로 합병된다. 만약 콘볼루션 윈도우가 모두 A*B개(X방향은 A개, Y방향은 B개, X, Y는 3차원 직교 좌표계의 좌표축)의 입력 데이터Xi를 얻는다고 가정하면 A*B개의 xi에 상기 콘볼루션 작업을 수행하고 획득한 모든 벡터는 메인 연산 모듈에서 합병되어 A*B*N개의 3차원 중간 결과를 얻는다.

[0598] 도 3f을 참조하면 도 3f은 본원 발명의 실시예에 따른 콘볼루션 신경망의 순방향 연산을 수행하기 위한 장치에서의 메인 연산 모듈(5)의 구조의 예시적 블럭도를 도시한다. 도 3f에 도시된 바와 같이 메인 연산 모듈(5)은 제1 연산 유닛(51), 제1 데이터 의존관계 판정 유닛(52)과 제1 저장 유닛(53)을 포함한다.

[0599] 여기서 제1 연산 유닛(51)은 벡터 덧셈 유닛(511) 및 활성화 유닛(512)을 포함한다. 제1 연산 유닛(51)은 컨트롤러 유닛(2)에서 발송한 컨트롤 신호를 수신하여 메인 연산 모듈(5)의 여러 가지 연산 기능을 완성하고 벡터

덧셈 유닛(511)은 콘볼루션 신경망의 순방향 계산에서의 바이어스 추가 작업을 실현하며 상기 부품은 바이어스 데이터와 상기 중간 결과를 대립하여 더해 바이어스 결과를 얻고 활성화 연산 유닛(512)은 바이어스 결과에 대해 활성화 함수 작업을 수행한다. 상기 바이어스 데이터는 외부 주소 공간으로부터 판독한 것이거나 로컬에 저장된 것일 수 있다.

[0600] 제1 데이터 의존관계 판정 유닛(52)은 제1 연산 유닛(51)이 제1 저장 유닛(53)을 판독 기록하는 포트로서 제1 저장 유닛(53)에서의 데이터의 판독 기록의 일치성을 담보한다. 이와 동시에 제1 데이터 의존관계 판정 유닛(52)은 또 제1 저장 유닛(53)으로부터 판독한 데이터가 인터커넥트 모듈(4)을 통해 서브 연산 모듈에 발송하고 서브 연산 모듈(6)의 출력 데이터는 인터커넥트 모듈(4)을 통해 제1 연산 유닛(51)에 직접적으로 발송하도록 담보한다. 컨트롤러 유닛(2)이 출력한 명령은 계산유닛(51)과 제1 데이터 의존관계 판정 유닛(52)에 발송하여 그의 동작을 제어한다.

[0601] 저장 유닛(53)은 메인 연산 모듈(5)이 계산 과정에서 사용하는 입력 데이터와 출력 데이터를 캐시한다.

[0602] 도 3g을 참조하면 도 3g은 본원 발명의 실시예에 따른 콘볼루션 신경망의 순방향 연산을 수행하기 위한 장치에서의 서브 연산 모듈(6)의 구조의 예시적인 블럭도를 도시한다. 도 3g에 도시된 바와 같이 매 하나의 서브 연산 모듈(6)은 제2 연산 유닛(61), 데이터 의존관계 판정 유닛(62), 제2 저장 유닛(63)과 제3 저장 유닛(64)을 포함한다.

[0603] 제2 연산 유닛(61)은 컨트롤러 유닛(2)이 발송한 컨트롤 신호를수신하여 콘볼루션 연산을 진행한다. 제2 연산 유닛은 벡터 곱셈 유닛(611)과 누적 유닛(612)을 포함하여 각각 콘볼루션 연산에서의 벡터 곱셈 연산과 누적 연산을 담당한다.

[0604] 제2 데이터 의존관계 판정 유닛(62)은 계산 과정에서 제2 저장 유닛(63)의 판독 기록 작업을 담당한다. 제2 데이터 의존관계 판정 유닛(62)은 판독 기록 작업을 수행하기 전에 우선 명령 사이에 사용되는 데이터가 판독 기록 일치성 충돌이 존재하지 않음을 담보한다. 예하면 데이터 의존관계 유닛(62)에 발송하는 모든 컨트롤 신호는 모두 데이터 의존관계 유닛(62) 내부의 명령 행렬에 저장되는데 상기 행렬에서 명령을 판독하는 판독 데이터의 범위가 행렬의 앞지리의 명령을 기입하고 데이터를 기입하는 범위와 충돌되면 상기 명령은 반드시 의존하는 명령 기입이 수행된 후에야만 수행될 수 있다.

[0605] 제2 저장 유닛(63)은 상기 서브 연산 모듈(6)의 입력 데이터와 출력 스칼라 데이터를 캐시한다.

[0606] 제3 저장 유닛(64)은 상기 서브 연산 모듈(6)이 계산 과정에서 필요한 콘볼루션 커널 데이터를 캐시한다.

[0607] 알 수 있다 시피 본 발명의 실시예의 방안에서 상기 동적 전압 조절 및 주파수 변조 장치는 상기 신경망 프로세서 및 그 내부의 각 유닛과 모듈의 운행속도를 실시간으로 수집하는데 신경망 프로세서 및 그 내부의 각 유닛과 모듈의 운행속도에 근거하여 신경망 프로세서 또는 그 내부의 각 유닛의 작동 주파수 또는 작동 전압이 저하됨을 결정하면 실제 작업에서의 사용자의 수요를 만족시키는 동시에 칩의 전반적인 운행의 전력 손실을 저하시키는 목적에 도달할 수 있다.

[0608] 도 3h을 참조하면 도 3h은 본원 발명의 실시예에서 제공하는 동적 전압 조절 및 주파수 변조 방법의 흐름모식도이다. 도 8에 도시된 바와 같이 상기 방법은 다음과 같은 단계를 포함한다.

[0609] 단계 S801, 동적 전압 조절 및 주파수 변조 장치상기 동적 전압 조절 및 주파수 변조와 연결된 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보를 실시간으로 수집하되, 상기 애플리케이션 시나리오 정보는 상기 칩이 신경망 연산을 통해 얻어지거나 또는 상기 칩과 연결된 센서가 수집한 정보이다.

[0610] 단계 S802, 동적 전압 조절 및 주파수 변조 장치는 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 칩이 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하기 위한 것이다.

[0611] 여기서 상기 칩의 작동 상태 정보는 상기 칩의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0612] 상기 칩의 운행속도가 타겟 속도보다 클 경우 상기 칩에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하고 상기 타겟 속도는 사용자 수요를 만족시킬 경우의 상기 칩의 운행속도인 단계를 포함한다.

[0613] 진일보로, 상기 칩은 적어도 제1 유닛과 제2 유닛을 포함하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이며 상기 칩의 작동 상태 정보는 상기 제1 유닛의 운행속도와 제2 유닛의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0614] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제1 유닛의 운행시간이 상기 제2 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제2 유닛에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 제2 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0615] 진일보로, 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하고 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0616] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제2 유닛의 운행시간이 상기 제1 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제1 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 제1 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0617] 선택적으로, 상기 칩은 적어도 N개의 유닛을 포함하고 상기 칩의 작동 상태 정보는 상기 N개의 유닛에서의 적어도 S개의 유닛의 작동 상태 정보를 포함하며 상기 N은 1보다 큰 정수이고 상기 S는 N보다 작거나 같은 정수이며 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하고 상기 칩의 작동 상태 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0618] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함하고,

[0619] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛에서의 임의의 하나이다.

[0620] 선택적으로, 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 칩의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 칩에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0621] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓일 경우 상기 유닛(A)에 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0622] 선택적으로, 상기 칩의 애플리케이션 시나리오는 이미지 인식이고 상기 애플리케이션 시나리오 정보는 인식 대기 이미지에서의 오브젝트의 개수이며 상기 전압 주파수 규제 정보는 제6 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0623] 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 작을 경우 상기 칩에 상기 제6 전압 주파수 규제 정보를 발송하되, 상기 제6 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0624] 선택적으로, 상기 애플리케이션 시나리오 정보는 오브젝트 레이블 정보이고 상기 전압 주파수 규제 정보는 제7 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0625] 상기 오브젝트 레이블 정보가 기설정 오브젝트 태그 집합에 속한다고 결정할 경우 상기 칩에 상기 제7 전압 주파수 규제 정보를 발송하되, 상기 제7 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 하기 위한 것이다.

[0626] 선택적으로, 상기 칩은 음성 인식에 응용되고 상기 애플리케이션 시나리오 정보는 음성 입력 속도이며 상기 전압 주파수 규제 정보는 제8 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0627] 상기 음성 입력 속도가 제2 임계값보다 작을 경우 상기 칩에 제8 전압 주파수 규제 정보를 발송하되, 상기 제8 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0628] 선택적으로, 상기 애플리케이션 시나리오 정보는 상기 칩이 음성 인식을 진행하여 얻은 키워드이고 상기 전압

주파수 규제 정보는 제9 전압 주파수 규제 정보를 포함하며 상기 주파수 변조 및 전압 조절 유닛은 또,

[0629] 상기 키워드가 기설정 키워드 집합일 경우 상기 칩에 상기 제9 전압 주파수 규제 정보를 발송하되, 상기 제9 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0630] 선택적으로, 상기 칩은 기계 번역에 응용되고 상기 애플리케이션 시나리오 정보는 문자 입력 속도 또는 번역 대기 이미지에서의 문자의 수량이며 상기 전압 주파수 규제 정보는 제10 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛은 또,

[0631] 상기 문자 입력 속도가 제3 임계값 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 작을 경우 상기 칩에 상기 제10 전압 주파수 규제 정보를 발송하되, 상기 제10 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0632] 선택적으로, 상기 애플리케이션 시나리오 정보는 외부의 광도이고 상기 전압 주파수 규제 정보는 제11 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0633] 상기 외부의 광도가 제5 임계값보다 작을 경우 상기 칩에 상기 제11 전압 주파수 규제 정보를 발송하되, 상기 제11 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0634] 선택적으로, 상기 칩은 이미지 뷰티에 응용되고 상기 전압 주파수 규제 정보가 제12 전압 주파수 규제 정보와 제13 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛은 또,

[0635] 상기 애플리케이션 시나리오 정보가 안면 이미지일 경우 상기 칩에 상기 제12 전압 주파수 규제 정보를 발송하되, 상기 제12 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압을 저하시키기 위한 것이고;

[0636] 상기 애플리케이션 시나리오 정보가 안면 이미지가 아닐 경우 상기 칩에 제13 전압 주파수 규제 정보를 발송하되, 상기 제13 전압 주파수 규제 정보는 상기 칩으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0637] 설명해야 할 것은 상기 방법 실시예의 구체적인 실현과정은 도 3a에 도시된 실시예의 관련된 설명을 참조할 수 있으므로 여기서 더이상 설명하지 않는다.

[0638] 도 4a을 참조하면 도 4a은 본원 발명의 실시예에서 제공하는 콘볼루션 연산장치의 구조모식도이다. 도 4a에 도시된 바와 같이 상기 콘볼루션 연산장치는 동적 전압 조절 및 주파수 변조 장치(7), 명령 저장 유닛(1), 컨트롤러 유닛(2), 데이터 액세스 유닛(3), 인터커넥트 모듈(4), 메인 연산 모듈(5)과 N개의 서브 연산 모듈(6)을 포함한다.

[0639] 여기서 명령 저장 유닛(1), 컨트롤러 유닛(2), 데이터 액세스 유닛(3), 인터커넥트 모듈(4), 메인 연산 모듈(5)과 N개의 서브 연산 모듈(6) 은 모두 하드웨어 회로(예하면 FPGA, CGRA, 특정 용도 지향 집적 회로ASIC, 아날로그 회로와 멤리스터 등을 포함하나 이에 한정되지 않음)를 통해 실현할 수 있다.

[0640] 여기서 명령 저장 유닛(1)은 데이터 액세스 유닛(3)이 판독한 명령을 저장한다.

[0641] 컨트롤러 유닛(2)은 명령 저장 유닛(1)으로부터 명령을 판독하고명령을 기타 모듈의 행동을 제어하는 컨트롤 신호로 디코딩하여 데이터 액세스 유닛(3), 메인 연산 모듈(5)과 서브 연산 모듈(6) 등과 같은 기타 모듈에 발송한다.

[0642] 데이터 액세스 유닛(3)은 외부 주소 공간과 콘볼루션 연산장치 사이의 데이터 또는 명령 판독 기록 작업을 수행한다.

[0643] 구체적으로 데이터 액세스 유닛(3)은 외부 주소 공간을 액세스하여 장치 내부의 각 저장 유닛에 데이터를 직접적으로 판독 기록하여 데이터의 로딩과 저장을 완성한다.

[0644] N개의 서브 연산 모듈(6)은 콘볼루션 신경망 알고리즘에서의 입력 데이터와 콘볼루션 커널의 콘볼루션 연산을 실행한다.

[0645] 여기서 N개의 서브 연산 모듈(6)은 구체적으로, 동일한 입력 데이터와 각각의 콘볼루션 커널을 이용하여 각각의 출력 스칼라를 병행 산출한다.

[0646] 인터커넥트 모듈(4)은 메인 연산 모듈(5)과 N개의 서브 연산 모듈(6)을 연결하기 위한 것으로 상이한 인터커넥트 토폴로지(예하면 트리 구조, 환형 구조, 메쉬 구조, 분급 인터커넥트, 버스 구조 등)를 실현할 수 있다. 인터커넥트 모듈(4)은 메인 연산 모듈(5)과 N개의 서브 연산 모듈(6) 사이의 데이터 전송을 실현할 수 있다.

[0647] 다시 말하면 인터커넥트 모듈(4)은 메인 연산 모듈(5)과 N개의 서브 연산 모듈(6) 사이의 연속되거나 또는 이산화된 데이터의 데이터 통로를 구성하고 인터커넥트 모듈(4)은 트리 구조, 환형 구조, 메쉬 구조, 분급 인터커넥트와 버스 구조에서의 임의의 한가지 구조이다.

[0648] 메인 연산 모듈(5)은 모든 입력 데이터의 중간 벡터를 중간 결과로 결합하고 상기 중간 결과에 대해 후속 연산을 수행한다.

[0649] 여기서 메인 연산 모듈(5)은 중간 결과와 바이어스 데이터를 가한 다음 활성화 작업을 수행하는데 더 사용된다. 메인 연산 모듈이 사용하는 활성화 함수active는 비선형 함수sigmoid, tanh, relu, softmax에서의 임의의 한 비선형 함수이다.

[0650] 여기서 메인 연산 모듈(5)은,

[0651] 메인 연산 모듈(5)이 계산 과정에서 얻은 입력 데이터와 출력 데이터를 캐시하기 위한 제1 저장 유닛(53);

[0652] 메인 연산 모듈(%)의 여러 가지 연산 기능을 완성하기 위한 제1 연산 유닛(51);

[0653] 제1 연산 유닛(51)이 제1 저장 유닛(53)을 판독 기록하는 포트로서 제1 저장 유닛(53)의 데이터의 판독 기록의 일치성을 담보하고 제1 저장 유닛(53)으로부터 입력된 뉴런 벡터를 판독하며 인터커넥트 모듈(4)을 통해 N개의 서브 연산 모듈(6)에 발송하고 인터커넥트 모듈(4)에서 발송한 중간 결과 벡터를 상기 제1 연산 유닛(51)에 발송하는 제1 데이터 의존관계 판정 유닛(52)을 포함한다.

[0654] 여기서 N개의 서브 연산 모듈(6)에서의 매 하나의 서브 연산 모듈은,

[0655] 상기 컨트롤러 유닛(2)이 발송한 컨트롤 신호를 수신하여 산술 논리 연산을 진행하는 제2 연산 유닛(61);

[0656] 계산과정에서 제2 저장 유닛(63)과 제3 저장 유닛(64)의 판독 기록 작업을 진행하여 제2 저장 유닛(63)과 제4 저장 유닛(64)의 판독 기록의 일치성을 담보하는 제2 데이터 의존관계 판정 유닛(62);

[0657] 입력 데이터 및 상기 서브 연산 모듈이 산출한 출력 스칼라를 캐시하는 제2 저장 유닛(63);

[0658] 상기 서브 연산 모듈이 계산 과정에서 필요한 콘볼루션 커널을 캐시하는 제3 저장 유닛(64)을 포함한다.

[0659] 진일보로, 제1 데이터 의존관계 판정 유닛(52)과 제2 데이터 의존관계 판정 유닛(62)은 아래의 방식을 통해 판독 기록의 일치성을 담보한다.

[0660] 실행되지 않은 컨트롤 신호와 실행 중인 컨트롤 신호의 데이터 사이에 의존관계가 존재하는 지의 여부를 판정하고 만약 존재하지 않으면 상기 컨트롤 신호를 즉시 발송하도록 허용하며 그렇지 않으면 상기 컨트롤 신호가 의존하는 모든 컨트롤 신호가 모두 실행을 완성할 때까지 대기한 후 상기 컨트롤 신호를 발송하도록 허용한다.

[0661] 선택적으로, 데이터 액세스 유닛(3)은 외부 주소 공간으로부터 입력 데이터, 바이어스 데이터와 콘볼루션 커널에서의 적어도 하나를 판독한다.

[0662] 신경망 풀 연결층 순방향 연산이 시작되기 전에 메인 연산 모듈(5)은 인터커넥트 모듈(4)을 통해 입력 데이터를 N개의 서브 연산 모듈(6)의 매 하나의 서브 연산 모듈에 전송하고 N개의 서브 연산 모듈(6)의 계산 과정이 완료된 후 인터커넥트 모듈(4)은 단계적으로 N개의 서브 연산 모듈(6)의 출력 스칼라를 중간 벡터로 결합하여 메인 연산 모듈(5)에 전송한다.

[0663] 이하 상이한 연산 명령을 통해 상기 콘볼루션 연산장치의 구체적인 계산방법을 설명하는데 여기서 연산 명령은 콘볼루션 계산명령을 예로 들수 있고 상기 콘볼루션 계산명령은 신경망에 응용될 수 있으므로 상기 콘볼루션 계산명령은 콘볼루션 신경망으로 불릴 수도 있다. 콘볼루션 계산명령에 있어서 그가 실제로 수행해야 할 공식은:

[0664] $$s = s(\sum wx_i + b)$$

[0665] 여기서 콘볼루션 커널w(다수의 데이터를 포함할 수 있음)에 입력 데이터Xi를 곱하여 합계를 구한 다음 선택적으로 바이어스b를 가하고 그 다음 선택적으로 활성화 연산s(h)도 진행하여 최종적인 출력결과S를 얻는다. 상기 공식에 따라 얻은 상기 계산 토폴로지 구조는 곱셈 연산기-덧셈 연산기-(선택적으로)활성화 연산기이다. 상기 콘

볼루션 계산명령은 명령 집합을 포함할 수 있는데 상기 명령 집합은 상이한 기능의 콘볼루션 신경망COMPUTE명령 및 CONFIG명령, IO명령, NOP 명령, JUMP 명령과 MOVE명령을 포함한다.

[0666] 한가지 실시예에서 COMPUTE명령은 다음과 같은 명령을 포함한다.

[0667] 콘볼루션 연산 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리 또는 스칼라 레지스터 파일)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한다.

[0668] 콘볼루션 신경망 sigmoid 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리 또는 스칼라 레지스터 파일)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 sigmoid 활성화를 진행;

[0669] 콘볼루션 신경망 TanH 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 TanH 활성화를 진행;

[0670] 콘볼루션 신경망 ReLU 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 ReLU 활성화를 진행;

[0671] 콘볼루션 신경망 group 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 group을 나눈 다음 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 활성화를 진행한다.

[0672] CONFIG 명령, 매 층마다의 인공 신경망 계산이 시작되기 전에 현재 층의 계산에 필요한 여러 가지 상수를 배치한다.

[0673] IO 명령, 외부 저장공간으로부터 계산에 필요한 입력 데이터를 판독하고 계산이 완료된 후 데이터를 외부공간에 저장한다.

[0674] NOP 명령, 현재 상기 콘볼루션 연산장치 내부의 모든 컨트롤 신호 캐시 행렬에서의 컨트롤 신호를 정리하여 NOP 명령 이전의 모든 명령이 모두 수행 완료되도록 담보한다. NOP 명령 자체는 그 어떤 작업도 포함하지 않음;

[0675] JUMP 명령, 명령 저장 유닛으로부터 판독될 다음 명령 주소의 점프를 제어하는 것을 담당함으로써 제어 흐름의 점프를 실현;

[0676] MOVE 명령, 상기 콘볼루션 연산장치 내부 주소공간의 어느 한 주소의 데이터를 상기 콘볼루션 연산장치 내부 주소공간의 다른 한 주소로 이동시키는 것을 담당하되, 상기 과정은 연산 유닛과 별도로 수행과정에서 연산 유닛의 리소스를 점용하지 않는다.

[0677] 상기 콘볼루션 연산장치가 콘볼루션 계산명령을 수행하는 방법은 구체적으로 다음과 같을 수 있다.

[0678] 컨트롤러 유닛(2)은 명령 저장 유닛(1) 내에서 콘볼루션 계산명령, 콘볼루션 계산명령과 대응되는 작동 도메인 및 콘볼루션 계산명령과 대응되는 제1 계산 토폴로지 구조(곱셈 연산기-덧셈 연산기-덧셈 연산기-활성화 연산기)를 추출하고 제어 유닛은 상기 작동 도메인을 데이터 액세스 유닛에 전송하며 상기 제1 계산 토폴로지 구조를 인터커넥트 모듈(4)에 전송한다.

[0679] 데이터 액세스 유닛(3)은 외부 저장매체로부터 상기 작동 도메인과 대응되는 콘볼루션 커널 w과 바이어스 b(b가 0일 경우 바이어스b를 추출할 필요가 없음)를 추출하고 콘볼루션 커널 w과 바이어스 b를 메인 연산 모듈(5)에 전송한다.

[0680] 선택적으로, 상기 제1 결과는 다수의 곱셈 연산의 결과를 포함할 수 있다.

[0681] 동적 전압 조절 및 주파수 변조 장치(7)는 상기 콘볼루션 연산장치의 작동 상태 정보를 수집하고 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 콘볼루션 연산장치로 하여금 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하기 위한 것이다.

[0682] 구체적으로 동적 전압 조절 및 주파수 변조 장치(7)는,

[0683] 상기 콘볼루션 연산장치의 작동 상태 정보를 수집하기 위한 정보 수집 유닛(71);

[0684] 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 콘볼루션 연산장치(71)에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 콘볼루션 연산장치(71)로 하여금 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하기 위한 것인 전압 조절 및 주파수 변조 유닛(72)을 포함한다.

[0685] 본원 발명의 하나의 가능한 실시예에서 상기 콘볼루션 연산장치의 작동 상태 정보는 상기 콘볼루션 연산장치의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 전압 조절 및 주파수 변조 유닛(72)은,

[0686] 상기 콘볼루션 연산장치의 운행속도가 타겟 속도보다 클 경우 상기 콘볼루션 연산장치에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 콘볼루션 연산장치로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하고 상기 타겟 속도는 사용자 수요를 만족시킬 경우의 상기 콘볼루션 연산장치의 운행속도이다.

[0687] 본원 발명의 하나의 가능한 실시예에서 상기 콘볼루션 연산장치의 작동 상태 정보는 데이터 액세스 유닛(3)의 운행속도와 메인 연산 모듈(5)의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 주파수 변조 및 전압 조절 유닛(72)은 또,

[0688] 뎠데이터 액세스 유닛(3)의 운행속도와 메인 연산 모듈(5)의 운행속도에 근거하여 상기 데이터 액세스 유닛(3)의 운행시간이 상기 메인 연산 모듈(5)의 운행시간을 초과한다고 결정할 경우 상기 메인 연산 모듈(5)에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 메인 연산 모듈(5)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0689] 진일보로, 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하고 주파수 변조 및 전압 조절 유닛(72)은 또,

[0690] 상기 데이터 액세스 유닛(3)의 운행속도와 상기 메인 연산 모듈(5)의 운행속도에 근거하여 상기 메인 연산 모듈(5)의 운행시간이 상기 데이터 액세스 유닛(3)의 운행시간을 초과한다고 결정할 경우 상기 데이터 액세스 유닛(3)에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 데이터 액세스 유닛(3)으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0691] 본원 발명의 하나의 가능한 실시예에서 상기 콘볼루션 연산장치의 작동 상태 정보는 명령 저장 유닛(1), 컨트롤러 유닛(2), 데이터 액세스 유닛(3), 인터커넥트 모듈(4), 메인 연산 모듈(5) 및 N개의 서브 연산 모듈(6)에서의 적어도 S개의 유닛/모듈의 작동 상태 정보를 포함하되, 상기 S는 1보다 크며 N+5보다 작거나 같은 정수이고 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하며 상기 전압 조절 및 주파수 변조 유닛(72)은,

[0692] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이고,

[0693] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛/모듈에서의 임의의 하나이다.

[0694] 진일보로, 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 전압 조절 및 주파수 변조 유닛(72)은 또,

[0695] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓인다고 결정할 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0696] 본원 발명의 하나의 가능한 실시예에서 본 발명의 실시예는 신경망 프로세서를 제공하는데 이는 상기 콘볼루션 연산장치를 포함한다.

[0697] 상기 신경망 프로세서는 인공 신경망 연산을 수행하고 음성 인식, 이미지 인식, 번역 등 인공지능의 응용을 실현한다.

[0698] 이 콘볼루션 계산 임무에서 도 4a에서의 동적 전압 조절 및 주파수 변조 장치(7)의 작동 과정은 다음과 같다.

[0699] 상황1, 상기 콘볼루션 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 동적 전압 조절 및 주파수 변조

장치(7)는 실시간으로 도 4a에서의 신경망 프로세서의 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(7)가 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도에 근거하여 데이터 액세스 유닛(3)의 운행시간이 메인 연산 모듈(5)의 운행시간을 초과한다고 결정할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 콘볼루션 연산을 진행하는 과정에서 상기 데이터 액세스 유닛(3)이 난관이 됨을 결정하고 메인 연산 모듈(5)은 현재의 콘볼루션 연산작업을 완성한 후 상기 데이터 액세스 유닛(3)이 판독 임무를 수행 완료함과 동시에 그가 판독한 데이터를 메인 연산 모듈(5)에 전송하기를 대기하여야만 메인 연산 모듈(5)은 이번 데이터 액세스 유닛(3)이 전송한 데이터에 근거하여 콘볼루션 연산작업을 진행할 수 있다. 동적 전압 조절 및 주파수 변조 장치(7)는 메인 연산 모듈(5)에 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 메인 연산 모듈(5)로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 메인 연산 모듈(5)의 운행속도를 저하시키도록 지시함으로써메인 연산 모듈(5)의 운행속도와 데이터 액세스 유닛(3)의 운행속도가 매칭되도록 하여 메인 연산 모듈(5)의 전력 손실을 저하시킴으로써 메인 연산 모듈(5)이 유휴한 상황이 발생하는 것을 방지하고 최종적으로 임무의 완성시간에 영향을 미치지 않는 상황에서 상기 신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0700] 상황2, 상기 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 도 4a에서의 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서의 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(7)는 데이터 액세스 유닛(3)과 메인 연산 모듈(5)의 운행속도에 근거하여 메인 연산 모듈(5)의 운행시간이 데이터 액세스 유닛(3)의 운행시간을 초과한다고 결정할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 콘볼루션 연산을 진행하는 과정에서 메인 연산 모듈(5)이 난관이 됨을 결정할 수 있고 데이터 액세스 유닛(3)은 현재의 데이터 판독 작업을 완료한 후 메인 연산 모듈(5)이 현재의 콘볼루션 연산작업을 완성한 후에야만 데이터 액세스 유닛(3)이 그가 판독한 데이터를 메인 연산 모듈(5)에 전송할 수 있다. 동적 전압 조절 및 주파수 변조 장치(7)는 데이터 액세스 유닛(3)에 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 데이터 액세스 유닛(3)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 데이터 액세스 유닛(3)의 운행속도를 저하시키도록 지시함으로써 데이터 액세스 유닛(3)의 운행속도와 메인 연산 모듈(5)의 운행속도가 매칭되도록 하여 데이터 액세스 유닛(3)의 전력 손실을 저하시키고 데이터 액세스 유닛(3)이 유휴한 상황이 발생하는 것을 방지하여 최종적으로 임무의 완성 시간에 영향을 미치지 않는 상황에서 상기 신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0701] 상기 신경망 프로세서는 인공 신경망 연산을 수행하는데 인공지능 애플리케이션을 진행할 경우 도 4a에서의 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서가 인공지능 애플리케이션을 진행하는 동작 파라미터를 수집하고 상기 작동 파라미터에 근거하여 상기 신경망 프로세서의 작동 전압 또는 작동 주파수를 조절한다.

[0702] 구체적으로 상기 인공지능 애플리케이션은 동영상 이미지 처리, 오브젝트 인식, 기계 번역, 음성 인식과 이미지 뷰티 등 일 수 있다.

[0703] 상황3, 상기 신경망 프로세서가 동영상 이미지 처리를 진행할 경우 도 4a에서의 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서가 동영상 이미지 처리를 진행하는 프레임 레이트를 수집한다. 상기 동영상 이미지 처리의 프레임 레이트가 타겟 프레임 레이트를 초과할 경우 상기 타겟 프레임 레이트는 사용자가 정상적으로 수요하는 동영상 이미지 처리프레임 레이트이고 동적 전압 조절 및 주파수 변조 장치(7)는 상기 신경망 프로세서에 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하여 사용자의 정상적인 동영상 이미지 처리 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력손실을 저하시킨다.

[0704] 상황4, 상기 신경망 프로세서가 음성 인식을 진행할 경우 도 4a에서의 동적 전압 조절 및 주파수 변조 장치(7)는 실시간으로 상기 신경망 프로세서의 음성 인식 속도를 수집한다. 상기 신경망 프로세서의 음성 인식 속도가 사용자의 실제 음성 인식 속도를 초과할 경우 동적 전압 조절 및 주파수 변조 장치(7)는 상기 신경망 프로세서에 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하고 사용자의 정상적인 음성 인식 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력 손실을 저하시킨다.

[0705] 상황5, 도 4a에서의 동적 전압 조절 및 주파수 변조 장치(7)는 상기 신경망 프로세서에서의 각 유닛 또는 모듈(명령 저장 유닛(1), 컨트롤러 유닛(2), 데이터 액세스 유닛(3), 인터커넥트 모듈(4), 메인 연산 모듈(5)과 N개의 서브 연산 모듈(6)을 포함)의 작동 상태를 실시간으로 모니터링한다. 상기 신경망 프로세서의 각 유닛 또는

모듈에서의 임의의 한 유닛 또는 모듈이 유휴상태에 있을 경우 동적 전압 조절 및 주파수 변조 장치(7)는 상기 유닛 또는 모듈에 제5 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 저하시킴으로써 상기 유닛 또는 모듈의 전력 손실을 저하시킨다. 상기 유닛 또는 모듈이 다시 작동 상태에 놓일 경우 동적 전압 조절 및 주파수 변조 장치(7)는 상기 유닛 또는 모듈에 제6 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 향상시킴으로써 상기 유닛 또는 모듈의 운행속도로 하여금 작업의 수요를 만족시키도록 한다.

[0706] 도 4e를 참조하면 도 4e는 인터커넥트 모듈(4)의 한가지 실시형태인 H트리 모듈을 예시적으로 도시한다. 인터커넥트 모듈(4)은 메인 연산 모듈(5)과 다수의 서브 연산 모듈(6) 사이의 다수의 노드로 구성된 이진 트리 통로인 데이터 통로를 구성하고 매 노드마다 상류의 데이터를 마찬가지로 하류의 두 개의 노드에 발송하며 하류의 두 개의 노드가 리턴한 데이터를 합병하여 상류의 노드에 리턴한다. 예하면 콘볼루션 신경망이 계산을 시작하는 단계에서 메인 연산 모듈(5) 내의 뉴런 데이터는 인터커넥트 모듈(4)을 통해 각 서브 연산 모듈(6)에 발송되고; 서브 연산 모듈(6)의 계산 과정이 완성된 후 서브 연산 모듈의 계산 과정이 완성된 후 서브 연산 모듈마다 출력한 뉴런의 값은 인터커넥트 모듈에서 단계적으로 하나의 완전한 뉴런으로 조성된 벡터로 결합된다. 예를 들어 설명하면 만약 콘볼루션 연산장치에 모두 N개의 서브 연산 모듈이 있을 경우 입력 데이터Xi는 N개의 서브 연산 모듈에 각각 발송되고 매 하나의 서브 연산 모듈은 입력 데이터Xi와 상기 서브 연산 모듈과 대응되는 콘볼루션 커널을 콘볼루션 연산하여 스칼라 데이터를 얻으며 각 서브 연산 모듈의 스칼라 데이터는 인터커넥트 모듈(4)에 의해 N개의 요소를 포함하는 하나의 중간 벡터로 합병된다. 만약 콘볼루션 윈도우가 모두 A*B개(X방향은 A개, Y방향은 B개, X, Y는 3차원 직교 좌표계의 좌표축)의 입력 데이터Xi를 얻는다고 가정하면 A*B개의 xi에 상기 콘볼루션 작업을 수행하고 획득한 모든 벡터는 메인 연산 모듈에서 합병되어 A*B*N개의 3차원 중간 결과를 얻는다.

[0707] 도 4b를 참조하면 도 4b는 본원 발명의 실시예에 따른 콘볼루션 신경망의 순방향 연산을 수행하기 위한 장치에서의 메인 연산 모듈(5)의 구조의 예시적 블럭도를 도시한다. 도 4b에 도시된 바와 같이 메인 연산 모듈(5)은 제1 연산 유닛(51), 제1 데이터 의존관계 판정 유닛(52)과 제1 저장 유닛(53)을 포함한다.

[0708] 여기서 제1 연산 유닛(51)은 벡터 덧셈 유닛(511) 및 활성화 유닛(512)을 포함한다. 제1 연산 유닛(51)은 도 4a에서의 컨트롤러 유닛(2)에서 발송한 컨트롤 신호를 수신하여 메인 연산 모듈(5)의 여러 가지 연산 기능을 완성하고 벡터 덧셈 유닛(511)은 콘볼루션 신경망의 순방향 계산에서의 바이어스 추가 작업을 실현하며 상기 부품은 바이어스 데이터와 상기 중간 결과를 대립하여 더해 바이어스 결과를 얻고 활성화 연산 유닛(512)은 바이어스 결과에 대해 활성화 함수 작업을 수행한다. 상기 바이어스 데이터는 외부 주소 공간으로부터 판독한 것이거나 로컬에 저장된 것일 수 있다.

[0709] 제1 데이터 의존관계 판정 유닛(52)은 제1 연산 유닛(51)이 제1 저장 유닛(53)을 판독 기록하는 포트로서 제1 저장 유닛(53)에서의 데이터의 판독 기록의 일치성을 담보한다. 이와 동시에 제1 데이터 의존관계 판정 유닛(52)은 또 제1 저장 유닛(53)으로부터 판독한 데이터가 인터커넥트 모듈(4)을 통해 서브 연산 모듈에 발송하고 서브 연산 모듈(6)의 출력 데이터는 인터커넥트 모듈(4)을 통해 제1 연산 유닛(51)에 직접적으로 발송하도록 담보한다. 컨트롤러 유닛(2)이 출력한 명령은 계산유닛(51)과 제1 데이터 의존관계 판정 유닛(52)에 발송하여 그의 동작을 제어한다.

[0710] 저장 유닛(53)은 메인 연산 모듈(5)이 계산 과정에서 사용하는 입력 데이터와 출력 데이터를 캐시한다.

[0711] 도 4c을 참조하면 도 4c은 본원 발명의 실시예에 따른 콘볼루션 신경망의 순방향 연산을 수행하기 위한 장치에서의 서브 연산 모듈(6)의 구조의 예시적인 블럭도를 도시한다. 도 4c에 도시된 바와 같이 매 하나의 서브 연산 모듈(6)은 제2 연산 유닛(61), 데이터 의존관계 판정 유닛(62), 제2 저장 유닛(63)과 제3 저장 유닛(64)을 포함한다.

[0712] 제2 연산 유닛(61)은 도 4a에서의 컨트롤러 유닛(2)이 발송한 컨트롤 신호를 수신하여 콘볼루션 연산을 진행한다. 제2 연산 유닛은 벡터 곱셈 유닛(611)과 누적 유닛(612)을 포함하여 각각 콘볼루션 연산에서의 벡터 곱셈 연산과 누적 연산을 담당한다.

[0713] 제2 데이터 의존관계 판정 유닛(62)은 계산 과정에서 제2 저장 유닛(63)의 판독 기록 작업을 담당한다. 제2 데이터 의존관계 판정 유닛(62)은 판독 기록 작업을 수행하기 전에 우선 명령 사이에 사용되는 데이터가 판독 기록 일치성 충돌이 존재하지 않음을 담보한다. 예하면 데이터 의존관계 유닛(62)에 발송하는 모든 컨트롤 신호는 모두 데이터 의존관계 유닛(62) 내부의 명령 행렬에 저장되는데 상기 행렬에서 명령을 판독하는 판독 데이터의

범위가 행렬의 앞지리의 명령을 기입하고 데이터를 기입하는 범위와 충돌되면 상기 명령은 반드시 의존하는 명령 기입이 수행된 후에야만 수행될 수 있다.

[0714] 제2 저장 유닛(63)은 상기 서브 연산 모듈(6)의 입력 데이터와 출력 스칼라 데이터를 캐시한다.

[0715] 제3 저장 유닛(64)은 상기 서브 연산 모듈(6)이 계산 과정에서 필요한 콘볼루션 커널 데이터를 캐시한다.

[0716] 본원 발명의 하나의 가능한 실시예에서 본원 발명의 실시예는 신경망 프로세서를 제공하는데 이는 상기 콘볼루션 연산장치를 포함한다.

[0717] 상기 신경망 프로세서는 인공 신경망 연산을 수행하고 음성 인식, 이미지 인식, 번역 등 인공지능의 응용을 실현한다.

[0718] 한가지 구체적인 애플리케이션 시나리오에서 콘볼루션 계산 임무를 수행할 경우 도 4a에서의 동적 전압 조절 및 주파수 변조 장치(7)의 작동 과정은 다음과 같다.

[0719] 동적 전압 조절 및 주파수 변조 장치(7)의 정보 수집 유닛(71)은 동적 전압 조절 및 주파수 변조 장치(7)와 연결된 신경망 프로세서의 작동 상태 정보 또는 애플리케이션 시나리오 정보를 실시간으로 수집하는데 상기 애플리케이션 시나리오 정보는 상기 신경망 프로세서가 신경망 연산을 통해 얻은 정보이거나 또는 상기 신경망 프로세서와 연결된 센서가 수집한 정보이고 동적 전압 조절 및 주파수 변조 장치(7)의 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서의 작동 상태 정보 또는 애플리케이션 시나리오 정보에 근거하여 상기 신경망 프로세서에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 조절하도록 지시한다.

[0720] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서의 작동 상태 정보는 상기 신경망 프로세서의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 전압 조절 및 주파수 변조 유닛(72)은,

[0721] 상기 신경망 프로세서의 운행속도가 타겟 속도보다 클 경우 상기 신경망 프로세서에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이고 상기 타겟 속도는 사용자 수요를 만족시킬 때 상기 신경망 프로세서의 운행속도이다.

[0722] 구체적으로 정보 수집 유닛(71)은 이와 연결된 신경망 프로세서의 운행속도를 실시간으로 수집한다. 상기 신경망 프로세서의 운행속도는 상기 신경망 프로세서가 수행한 임무의 상이함에 근거하여 상이한 유형의 속도가 될 수 있다. 상기 신경망 프로세서가 진행하는 작업이 동영상 이미지 처리일 경우 상기 신경망 프로세서의 운행속도는 상기 신경망 프로세서가 동영상 이미지 처리를 진행할 경우의 프레임 레이트일 수 있고 상기 신경망 프로세서가 진행하는 작업이 음성 인식일 경우 상기 신경망 프로세서의 운행속도는 상기 정보가 음성 인식을 진행하는 속도이다. 전압 조절 및 주파수 변조 유닛(72)이 상기 신경망 프로세서의 운행속도가 상기 타겟 속도보다 크다고, 즉 상기 신경망 프로세서의 운행속도가 사용자 수요를 만족할 때의 상기 신경망 프로세서의 운행속도에 도달한다고 결정할 경우 상기 신경망 프로세서에 제1 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 신경망 프로세서의 전력 손실을 저하시키도록 지시한다.

[0723] 예를 들어 설명하면 상기 신경망 프로세서가 진행하는 작업이 동영상 이미지 처리이고 상기 타겟 속도가 24프레임/초라고 가정한다. 정보 수집 유닛(71)은 상기 신경망 프로세서가 동영상 이미지 처리를 진행하는 프레임 레이트를 실시간으로 수집하고 현재 상기 신경망 프로세서가 동영상 이미지 처리를 진행하는 프레임 레이트는 54프레임/초이다. 전압 조절 및 주파수 변조 유닛(72)은 현재 상기 신경망 프로세서가 동영상 이미지 처리를 진행하는 프레임 레이트가 상기 타겟 속도보다 크다는 것을 결정할 경우 상기 신경망 프로세서에 제1 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 신경망 프로세서의 전력 손실을 저하시키도록 지시한다.

[0724] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서는 적어도 제1 유닛과 제2 유닛을 포함하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이며 상기 신경망 프로세서의 작동 상태 정보는 상기 제1 유닛의 운행속도와 제2 유닛의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 주파수 변조 및 전압 조절 유닛(72)은 또,

[0725] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제1 유닛의 운행시간이 상기 제2 유닛

의 운행시간을 초과하였다고 결정할 경우 상기 제2 유닛에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 제2 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0726] 구체적으로 상기 신경망 프로세서가 임무를 수행함에 있어서 상기 제1 유닛과상기 제2 유닛의 배합이 필요하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이다. 정보 수집 유닛(71)은 상기 제1 유닛과 상기 제2 유닛의 운행속도를 실시간으로 수집한다. 상기 제1 유닛의 운행속도가 상기 제2 유닛의 운행속도보다 작다고 결정할 경우, 즉 상기 제1 유닛의 운행시간이 상기 제2 유닛의 운행시간을 초과한다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(72)은 상기 제2 유닛에 상기 제2 전압 주파수 규제 정보를 발송하여 상기 제2 유닛으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 신경망 프로세서의 전반적인 운행속도에 영향을 주지 않는 전제하에 상기 신경망 프로세서의 전반적인 전력 손실을 저하시키는 효과에 도달한다.

[0727] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하고 주파수 변조 및 전압 조절 유닛(72)은 또,

[0728] 상기 제1 유닛의 운행속도와 상기 제2 유닛의 운행속도에 근거하여 상기 제2 유닛의 운행시간이 상기 제1 유닛의 운행시간을 초과하였다고 결정할 경우 상기 제1 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 제1 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이다.

[0729] 구체적으로 상기 신경망 프로세서가 임무를 수행함에 있어서 상기 제1 유닛과상기 제2 유닛의 배합이 필요하고 상기 제1 유닛의 출력 데이터는 상기 제2 유닛의 입력 데이터이다. 정보 수집 유닛(71)은 상기 제1 유닛과 상기 제2 유닛의 운행속도를 실시간으로 수집한다. 상기 제1 유닛의 운행속도가 상기 제2 유닛의 운행속도보다 크다고 결정할 경우, 즉 상기 제2 유닛의 운행시간이 상기 제1 유닛의 운행시간을 초과한다고 결정할 경우 전압 조절 및 주파수 변조 유닛(72)은 상기 제1 유닛에 상기 제3 전압 주파수 규제 정보를 발송하여 상기 제1 유닛으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 신경망 프로세서의 전반적인 운행속도에 영향을 주지 않는 전제하에 상기 신경망 프로세서의 전반적인 전력 손실을 저하시키는 효과에 도달한다.

[0730] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서包括적어도 N개의 유닛, 상기 신경망 프로세서의 작동 상태 정보는 상기 적어도 N개의 유닛에서의 적어도 S개의 유닛의 작동 상태 정보를 포함하며 상기 N은 1보다 큰 정수이고 상기 S는 N보다 작거나 같은 정수이며 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하고 전압 조절 및 주파수 변조 유닛(72)은,

[0731] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것이고,

[0732] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛에서의 임의의 하나이다.

[0733] 본원 발명의 하나의 가능한 실시예에서 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 전압 조절 및 주파수 변조 유닛(72)은 또,

[0734] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓일 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이다.

[0735] 구체적으로 상기 신경망 프로세서의 작동 과정에서 정보 수집 유닛(71)은 상기 신경망 프로세서 내부의 적어도 S개의 유닛의 작동 상태 정보를 실시간으로 수집한다. 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 전압 조절 및 주파수 변조 유닛(72)은 상기 유닛(A)에 제4 전압 주파수 규제 정보를 발송하여 상기 유닛(A)으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시켜 상기 유닛(A)의 전력 손실을 저하시키도록 지시하고 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동 상태에 있을 경우 상기 전압 조절 및 주파수 변조 유닛(72)은 상기 유닛(A)에 제5 전압 주파수 규제 정보를 발송하여 상기 유닛(A)으로 하여금 그의 작동 주파수 또는 작동 전압을 향상시켜 상기 유닛(A)의 운행속도가 작동의 수요를 만족시키도록 한다.

[0736] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서의 애플리케이션 시나리오가 이미지 인식일 경우 상기 애플리케이션 시나리오 정보는 인식 대기 이미지에서의 오브젝트의 개수이고 상기 전압 주파수 규제 정보

는 제6 전압 주파수 규제 정보를 포함하며 전압 조절 및 주파수 변조 유닛(72)은 또,

[0737] 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 작을 경우 상기 신경망 프로세서에 상기 제6 전압 주파수 규제 정보를 발송하되, 상기 제6 전압 주파수 규제 정보는 상기 신경망 프로세서으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0738] 구체적으로 상기 신경망 프로세서가 이미지 인식에 응용되고 상기 인식 대기 이미지에서의 오브젝트의 개수는 상기 신경망 프로세서가 신경망 알고리즘을 통해 얻은 것이며 상기 정보 수집 유닛(71)이 상기 신경망 프로세서에서 상기 인식 대기 이미지에서의 오브젝트의 개수(즉 상기 애플리케이션 시나리오 정보)를 획득한 후 전압 조절 및 주파수 변조 유닛(72)이 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 작다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 상기 제6 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하고 상기 인식 대기 이미지에서의 오브젝트의 개수가 제1 임계값보다 크다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키기 위한 전압 주파수 규제 정보를 발송한다.

[0739] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 오브젝트 레이블 정보이고 상기 전압 주파수 규제 정보는 제7 전압 주파수 규제 정보를 포함하며 전압 조절 및 주파수 변조 유닛(72)은 또,

[0740] 상기 오브젝트 레이블 정보가 기설정 오브젝트 태그 집합에 속한다고 결정할 경우 상기 신경망 프로세서에 상기 제7 전압 주파수 규제 정보를 발송하되, 상기 제7 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 하기 위한 것이다.

[0741] 예를 들어 설명하면, 상기 기설정 오브젝트 태그 집합은 다수의 오브젝트 태그를 포함하고 상기 오브젝트 태그는 "사람", "개", "나무"와 "꽃" 일 수 있다. 상기 신경망 프로세서가 신경망 알고리즘을 통해 현재 애플리케이션 시나리오가 개를 포함한다고 할 경우 상기 신경망 프로세서는 상기 "개"를 포함하는 이 오브젝트 레이블 정보를 상기 정보 수집 유닛(71)에 전송한 다음 상기 주파수 변조 및 전압 조절 유닛(72)이 상기 오브젝트 레이블 정보가 "개"를 포함한다고 결정할 경우 상기 신경망 프로세서에 제7 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하고 상기 오브젝트 레이블 정보가 상기 기설정 오브젝트 태그 집합에 속하지 않는다고 결정할 경우 상기 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 전압 주파수 규제 정보를 발송한다.

[0742] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서는 음성 인식에 응용되고 상기 애플리케이션 시나리오 정보는 음성 입력 속도이며 상기 전압 주파수 규제 정보는 제8 전압 주파수 규제 정보를 포함하고 전압 조절 및 주파수 변조 유닛(72)은 또,

[0743] 상기 음성 입력 속도가 제2 임계값보다 작을 경우 상기 신경망 프로세서에 제8 전압 주파수 규제 정보를 발송하되, 상기 제8 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0744] 구체적으로 상기 신경망 프로세서의 애플리케이션 시나리오는 음성 인식이고 상기 신경망 프로세서의 입력 유닛은 일정한 속도로 신경망 프로세서에 음성을 입력한다. 정보 수집 유닛(71)은 음성 입력 속도를 실시간으로 수집하고 상기 음성 입력 속도정보를 상기 전압 조절 및 주파수 변조 유닛(72)에 발송한다. 상기 전압 조절 및 주파수 변조 유닛(72)이 상기 음성 입력 속도가 제2 임계값보다 작다고 결정할 경우 상기 신경망 프로세서에 제8 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시한다. 상기 전압 조절 및 주파수 변조 유닛(72)이 상기 음성 입력 속도가 제2 임계값보다 크다고 결정할 경우 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압을 향상시키기 위한 전압 주파수 규제 정보를 발송한다.

[0745] 본원 발명의 하나의 가능한 실시예에서 상기 애플리케이션 시나리오 정보는 상기 신경망 프로세서가 음성 인식을 진행하여 얻은 키워드이고 상기 전압 주파수 규제 정보는 제9 전압 주파수 규제 정보를 포함하며 상기 주파수 변조 및 전압 조절 유닛은 또,

[0746] 상기 키워드가 기설정 키워드 집합일 경우 상기 신경망 프로세서에 상기 제9 전압 주파수 규제 정보를 발송하되, 상기 제9 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수

를 향상시키도록 지시하기 위한 것이다.

[0747] 진일보로 상기 키워드가 상기 키워드 집합에 속하지 않을 경우 주파수 변조 및 전압 조절 유닛(72)은 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키기 위한 전압 조절 및 주파수 변조 정보를 발송한다.

[0748] 예를 들어 설명하면 상기 칩의 애플리케이션 시나리오가 음성 인식이 경우 상기 기설정 키워드 집합은 "이미지 뷰티", "신경망 알고리즘", "이미지 처리"와 "알리페이" 등 키워드를 포함한다. 만약 상기 애플리케이션 시나리오 정보가 "이미지 뷰티"라고 가정하면 주파수 변조 및 전압 조절 유닛(72)은 상기 신경망 프로세서에 상기 제9 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하고 마약 상기 애플리케이션 시나리오 정보가 "촬영"일 경우 주파수 변조 및 전압 조절 유닛(72)으 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키기 위한 전압 조절 및 주파수 변조 정보를 발송한다.

[0749] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서가 기계 번역에 응용될 경우 상기 애플리케이션 시나리오 정보는 문자 입력 속도 또는 번역 대기 이미지에서의 문자의 수량이며 상기 전압 주파수 규제 정보는 제10 전압 주파수 규제 정보를 포함하고 전압 조절 및 주파수 변조 유닛(72)은 또,

[0750] 상기 문자 입력 속도가 제3 임계값 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 작을 경우 상기 신경망 프로세서에 상기 제10 전압 주파수 규제 정보를 발송하되, 상기 제10 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0751] 구체적으로 상기 신경망 프로세서는 기계 번역에 응용되고 정보 수집 유닛(71)이 수집한 애플리케이션 시나리오 정보는 문자 입력 속도 또는 번역 대기 이미지에서의 문자의 수량이며 상기 애플리케이션 시나리오 정보를 전압 조절 및 주파수 변조 유닛(72)에 전송한다. 상기 문자 입력 속도가 제3 임계값보다 작거나 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 작다고 결정할 경우 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 제10 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압을 저하시키도록 지시하고 상기 문자 입력 속도가 제3 임계값보다 크거나 또는 번역 대기 이미지에서의 문자의 수량이 제4 임계값보다 크다고 결정할 경우 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압을 향상시키기 위한 전압 주파수 규제 정보를 발송한다.

[0752] 본원 발명의 하나의 가능한 실시예에서 在상기 애플리케이션 시나리오 정보는 외부의 광도 일 경우 상기 전압 주파수 규제 정보는 제11 전압 주파수 규제 정보를 포함하며 전압 조절 및 주파수 변조 유닛(72)은 또,

[0753] 상기 외부의 광도가 제5 임계값보다 작을 경우 상기 신경망 프로세서에 상기 제11 전압 주파수 규제 정보를 발송하되, 상기 제11 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0754] 구체적으로 상기 외부의 광도는 상기 신경망 프로세서와 연결된 조도 센서가 수집하여 획득한 것이다. 정보 수집 유닛(71)은 상기 광도를 획득한 후 상기 광도를 전압 조절 및 주파수 변조 유닛(72)에 전송한다. 상기 광도가 제5 임계값보다 작다고 결정할 경우 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 상기 제11 전압 주파수 규제 정보를 발송하여 상기 신경망 프로세서로 하여금 그의 작동 전압을 저하시키도록 지시하고 상기 광도가 제5 임계값보다 크다고 결정할 경우 전압 조절 및 주파수 변조 유닛(72)은 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 전압 주파수 규제 정보를 발송한다.

[0755] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서는 이미지 뷰티에 응용되고 상기 전압 주파수 규제 정보는 제12 전압 주파수 규제 정보와 제13 전압 주파수 규제 정보를 포함하며 전압 조절 및 주파수 변조 유닛은 또,

[0756] 상기 애플리케이션 시나리오 정보가 안면 이미지일 경우 상기 신경망 프로세서에 상기 제12 전압 주파수 규제 정보를 발송하되, 상기 제12 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것이고;

[0757] 상기 애플리케이션 시나리오 정보가 안면 이미지가 아닐 경우 상기 신경망 프로세서에 제13 전압 주파수 규제 정보를 발송하되, 상기 제13 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 것이다.

[0758] 본원 발명의 하나의 가능한 실시예에서 상기 신경망 프로세서가 음성 인식에 응용 될 경우 상기 애플리케이션 시나리오 정보는 음성 강도이며 상기 음성 강도가 제6 임계값보다 클 경우 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하기 위한 전압 주파수 규제 정보를 발송하고 상기 음성 강도가 제6 임계값보다 작을 경우 전압 조절 및 주파수 변조 유닛(72)은 상기 신경망 프로세서에 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 전압 주파수 규제 정보를 발송한다.

[0759] 설명해야 할 것은 상기 시나리오 정보는 광도, 음성 강도 등과 같은 센서가 수집한 외부 시나리오의 정보 일 수 있다. 상기 애플리케이션 시나리오 정보는 인공지능 알고리즘에 근거하여 산출한 정보 일 수도 있는데 예하면 오브젝트 인식 임무에서 신경망 프로세서의 실시간 계산결과정보를 정보 수집 유닛에 피드백하되, 상기 정보는 시나리오의 오브젝트 개수, 안면 이미지, 오브젝트 태그 키워드 등 정보를 포함한다.

[0760] 선택적으로, 상기 인공지능 알고리즘은 신경망 알고리즘을 포함하나 이에 한정되지 않는다.

[0761] 도 4f을 참조하면 도 4f은 본원 발명의 실시예에서 제공하는 다른 콘볼루션 연산장치의 구조모식도이다. 도 4f에 도시된 바와 같이 상기 콘볼루션 연산장치는 동적 전압 조절 및 주파수 변조 장치(617), 레지스터 유닛(612), 인터커넥트 모듈(613), 연산 유닛(614), 제어 유닛(615)과 데이터 액세스 유닛(616)을 포함한다.

[0762] 여기서 연산 유닛(614)은 덧셈 계산기, 곱셈 계산기, 컴퍼레이터와 활성화 연산기에서의 적어도 두 가지를 포함한다.

[0763] 인터커넥트 모듈(613)은 연산 유닛(614)에서의 계산기의 연결관계를 제어하여 적어도 두 가지 계산기로 하여금 상이한 계산 토폴로지 구조를 조성하도록 한다.

[0764] 레지스터 유닛(612)(레지스터 유닛, 명령 캐시, 스크래치패드 메모리 일 수 있음)은 상기 연산 명령, 데이터 블록이 저장매체에서의 주소, 연산 명령과 대응되는 계산 토폴로지 구조를 저장한다.

[0765] 선택적으로, 상기 콘볼루션 연산장치는 저장매체(611)를 더 포함한다.

[0766] 저장매체(611)는 오프 칩 메모리 일 수 있는데 물론 실제 응용에서는 데이터 블록을 저장하기 위한 온 칩 메모리 일 수도 있으며 상기 데이터 블록은 구체적으로 n차원 데이터 일 수 있고 n은 1보다 크거나 같은 정수 일 수 있는 바, 예하면 n=1일 경우 1차원 데이터, 즉 벡터이고 n=2일 경우 2차원 데이터, 즉 매트릭스이며 n=3 또는 3 이상 일 경우 다차원 데이터 일 수 있다.

[0767] 제어 유닛(615)은 레지스터 유닛(612) 내에서 연산 명령, 상기 연산 명령과 대응되는 작동 도메인 및 상기 연산 명령과 대응되는 제1 계산 토폴로지 구조를 추출하고 상기 연산 명령을 수행 명령으로 디코딩하며 상기 수행 명령은 연산 유닛(614)을 제어하여 연산작업을 수행하도록 하고 상기 작동 도메인을 데이터 액세스 유닛(616)에 전송하며 상기 계산 토폴로지 구조를 인터커넥트 모듈(613)에 전송한다.

[0768] 데이터 액세스 유닛(616)은 저장매체(611)에서 상기 작동 도메인과 대응되는 데이터 블록을 추출하고 상기 데이터 블록을 인터커넥트 모듈(613)에 전송한다.

[0769] 인터커넥트 모듈(613)은 제1 계산 토폴로지 구조의 데이터 블록을 수신한다.

[0770] 본원 발명의 하나의 가능한 실시예에서 인터커넥트 모듈(613)은 또 제1 계산 토폴로지 구조에 근거하여 데이터 블록을 다시 배치한다.

[0771] 연산 유닛(614)은 명령을 수행하고 연산 유닛(614)의 계산기를 호출하여 데이터 블록에 대해 연산작업을 수행함으로써 연산 결과를 얻으며 상기 연산 결과를 데이터 액세스 유닛(616)에 전송하여 저장매체(611) 내에 저장한다.

[0772] 본원 발명의 하나의 가능한 실시예에서 연산 유닛(614)은 또 제1 계산 토폴로지 구조 및 상기 수행 명령에 따라 계산기를 호출하여 다시 배치된 데이터 블록에 대해 연산작업을 수행하여 연산 결과를 얻고 상기 연산 결과를 데이터 액세스 유닛(616)에 전송하여 저장매체(611) 내에 저장한다.

[0773] 하나의 가능한 실시예에서 인터커넥트 모듈(613)은 또 연산 유닛(614)에서의 계산기의 연결관계를 제어함으로써 제1 계산 토폴로지 구조를 형성한다.

[0774] 동적 전압 조절 및 주파수 변조 장치(617)는 전반적인 콘볼루션 연산장치의 작동상태를 모니터링하고 그의 전압과 주파수에 대해 동적으로 규제한다.

[0775] 이하 상이한 연산 명령을 통해 상기 콘볼루션 연산장치의 구체적인 계산방법을 설명하는데 여기서 연산 명령은 콘볼루션 계산명령을 예로 들수 있고 상기 콘볼루션 계산명령은 신경망에 응용될 수 있으므로 상기 콘볼루션 계산명령은 콘볼루션 신경망으로 불릴 수도 있다. 콘볼루션 계산명령에 있어서 그가 실제로 수행해야 할 공식은:

[0776] $$s = s(\sum wx_i + b)$$

[0777] 여기서 콘볼루션 커널w(다수의 데이터를 포함할 수 있음)에 입력 데이터Xi를 곱하여 합계를 구한 다음 선택적으로 바이어스b를 가하고 그 다음 선택적으로 활성화 연산s(h)도 진행하여 최종적인 출력결과S를 얻는다. 상기 공식에 따라 얻은 상기 계산 토폴로지 구조는 곱셈 연산기-덧셈 연산기-(선택적으로)활성화 연산기이다. 상기 콘볼루션 계산명령은 명령 집합을 포함할 수 있는데 상기 명령 집합은 상이한 기능의 콘볼루션 신경망 COMPUTE 명령 및 CONFIG 명령, IO 명령, NOP 명령, JUMP 명령과 MOVE 명령을 포함한다.

[0778] 한가지 실시예에서 COMPUTE 명령은 다음과 같은 명령을 포함한다.

[0779] 콘볼루션 연산 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리 또는 스칼라 레지스터 파일)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한다.

[0780] 콘볼루션 신경망 sigmoid 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리 또는 스칼라 레지스터 파일)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 sigmoid 활성화를 진행;

[0781] 콘볼루션 신경망 TanH 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 TanH 활성화를 진행;

[0782] 콘볼루션 신경망 ReLU 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 ReLU 활성화를 진행; 및

[0783] 콘볼루션 신경망 group 명령, 상기 명령에 근거하여 상기 콘볼루션 연산장치가 각각 메모리(바람직하게 스크래치패드 메모리)의 지정된 주소로부터 지정된 크기의 입력 데이터와 콘볼루션 커널을 취하고 group을 나눈 다음 콘볼루션 연산부품에서 콘볼루션 작업을 진행한 다음 출력결과에 대해 활성화를 진행한다.

[0784] CONFIG 명령, 매 층마다의 인공 신경망 계산이 시작되기 전에 현재 층의 계산에 필요한 여러 가지 상수를 배치한다.

[0785] IO 명령, 외부 저장공간으로부터 계산에 필요한 입력 데이터를 판독하고 계산이 완료된 후 데이터를 외부공간에 저장한다.

[0786] NOP 명령, 현재 상기 콘볼루션 연산장치 내부의 모든 컨트롤 신호 캐시 행렬에서의 컨트롤 신호를 정리하여 NOP 명령 이전의 모든 명령이 모두 수행 완료되도록 담보한다. NOP 명령 자체는 그 어떤 작업도 포함하지 않음;

[0787] JUMP 명령, 명령 저장 유닛으로부터 판독될 다음 명령 주소의 점프를 제어하는 것을 담당함으로써 제어 흐름의 점프를 실현;

[0788] MOVE명령, 상기 콘볼루션 연산장치 내부 주소공간의 어느 한 주소의 데이터를 상기 콘볼루션 연산장치 내부 주소공간의 다른 한 주소로 이동시키는 것을 담당하되, 상기 과정은 연산 유닛과 별도로 수행과정에서 연산 유닛의 리소스를 점용하지 않는다.

[0789] 상기 콘볼루션 연산장치가 콘볼루션 계산명령을 수행하는 방법은 구체적으로 다음과 같을 수 있다.

[0790] 제어 유닛(615)은 레지스터 유닛(612) 내에서 콘볼루션 계산명령, 상기 콘볼루션 계산명령과 대응되는 작동 도메인 및 콘볼루션 계산명령과 대응되는 제1 계산 토폴로지 구조(곱셈 연산기-덧셈 연산기-덧셈 연산기-활성화 연산기)를 추출하고 제어 유닛은 상기 작동 도메인을 데이터 액세스 유닛(616)에 전송하며 상기 제1 계산 토폴로지 구조를 인터커넥트 모듈(613)에 전송한다.

[0791] 데이터 액세스 유닛(616)은 저장매체(611) 내에서 상기 작동 도메인과 대응되는 콘볼루션 커널w과 바이어스b(b가 0일 경우 바이어스b를 추출할 필요가 없음)를 추출하고 콘볼루션 커널w과 바이어스b를 연산 유닛(614)에 전

송한다.

[0792] 연산 유닛(614)의 곱셈 연산기는 콘볼루션 커널w과 입력 데이터Xi에 곱셈 연산을 진행하여 제1 결과를 얻고 제1 결과를 덧셈 연산기에 입력하여 덧셈 연산을 수행하여 제2 결과를 얻으며 제2 결과와 바이어스b에 덧셈 연산을 진행하여 제3 결과를 얻고 제3 결과를 활성화 연산기에 입력하여 활성화 연산을 수행함으로써 출력결과S를 얻으며 출력 결과S를 데이터 액세스 유닛(616)에 전송하여 저장매체 내(611) 내에 저장한다. 여기서 매 단계 뒤에는 모두 직접적으로 출력 결과를 데이터 액세스 유닛에 전송하여 저장매체(611) 내에 저장할 수 있어 아래의 단계를 진행할 필요가 없다. 이 외에 제2 결과와 바이어스b에 덧셈 연산을 진행하여 제3 결과를 얻는 이 단계는 선택적인 것으로서 즉 b가 0일 경우 이 단계는 필요하지 않게 된다. 이 외에 덧셈 연산과 곱셈 연산의 순서는 바뀔 수 있다.

[0793] 선택적으로, 상기 제1 결과는 다수의 곱셈 연산의 결과를 포함할 수 있다.

[0794] 본원 발명의 하나의 가능한 실시예에서 본원 발명의 실시예는 신경망 프로세서를 제공하는데 이는 상기 콘볼루션 연산장치를 포함한다.

[0795] 상기 신경망 프로세서는 인공 신경망 연산을 수행하고 음성 인식, 이미지 인식, 번역 등 인공지능의 응용을 실현한다.

[0796] 이 콘볼루션 계산 임무에서 동적 전압 조절 및 주파수 변조 장치(617)의 작동 과정은 다음과 같다.

[0797] 상황1, 상기 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 도 4f에서의 동적 전압 조절 및 주파수 변조 장치(617)는 실시간으로 신경망 프로세서의 데이터 액세스 유닛(616)과 연산 유닛(614)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(617)가 데이터 액세스 유닛(616)과 연산 유닛(614)의 운행속도에 근거하여 데이터 액세스 유닛(616)의 운행시간이 상기 연산 유닛(614)의 운행시간을 초과한다고 결정할 경우 상기 동적 전압 조절 및 주파수 변조 장치(617)는 콘볼루션 연산을 진행하는 과정에서 데이터 액세스 유닛(616)이 난관이 됨을 결정하고 연산 유닛(614)은 현재의 콘볼루션 연산작업을 완성한 후 상기 데이터 액세스 유닛(616)이 판독 임무를 수행 완료함과 동시에 그가 판독한 데이터를 연산 유닛(614)에 전송하기를 대기하여야만 상기 연산 유닛(614)은 이번 데이터 액세스 유닛(616)이 전송한 데이터에 근거하여 콘볼루션 연산작업을 진행할 수 있다. 동적 전압 조절 및 주파수 변조 장치(617)는 연산 유닛(614)에 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 연산 유닛(614)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 상기 연산 유닛(614)의 운행속도를 저하시키도록 지시함으로써 연산 유닛(614)의 운행속도와 데이터 액세스 유닛(616)의 운행속도가 매칭되도록 하여 연산 유닛(614)의 전력 손실을 저하시킴으로써 연산 유닛(614)이 유휴한 상황이 발생하는 것을 방지하고 최종적으로 임무의 완성시간에 영향을 미치지 않는 상황에서 상기 신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0798] 상황2, 상기 신경망 프로세서가 콘볼루션 연산을 진행하는 과정에서 동적 전압 조절 및 주파수 변조 장치(617)는 실시간으로 상기 신경망 프로세서의 데이터 액세스 유닛(616)과 연산 유닛(614)의 운행속도를 획득한다. 동적 전압 조절 및 주파수 변조 장치(617)가 데이터 액세스 유닛(616)과 연산 유닛(614)의 운행속도에 근거하여 연산 유닛(614)의 운행시간이 데이터 액세스 유닛(616)의 운행시간을 초과한다고 결정할 경우 동적 전압 조절 및 주파수 변조 장치(617)는 콘볼루션 연산을 진행하는 과정에서 연산 유닛(614)이 난관이 됨을 결정할 수 있고 데이터 액세스 유닛(616)은 현재의 데이터 판독 작업을 완료한 후 연산 유닛(614)이 현재의 콘볼루션 연산작업을 수행하기를 대기해야만이 데이터 액세스 유닛(616)은 판독한 데이터를 상기 연산 유닛(614)에 전송할 수 있다. 동적 전압 조절 및 주파수 변조 장치(617)는 데이터 액세스 유닛(316)에 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 데이터 액세스 유닛(616)으로 하여금 그의 작동 전압 또는 작동 주파수를 저하시켜 데이터 액세스 유닛(616)의 운행속도를 저하시키도록 지시함으로써 데이터 액세스 유닛(616)의 운행속도와 상기 연산 유닛(614)의 운행속도가 매칭되도록 하여 데이터 액세스 유닛(616)의 전력 손실을 저하시키고 데이터 액세스 유닛(616)이 유휴한 상황이 발생하는 것을 방지하여 최종적으로 임무의 완성 시간에 영향을 미치지 않는 상황에서 상기 신경망 프로세서의 전반적인 운행의 전력 손실을 저하시킨다.

[0799] 상기 신경망 프로세서는 인공 신경망 연산을 수행하는데 인공지능 애플리케이션을 진행할 경우 동적 전압 조절 및 주파수 변조 장치(617)는 실시간으로 상기 신경망 프로세서가 인공지능 애플리케이션을 진행하는 동작 파라미터를 수집하고 상기 작동 파라미터에 근거하여 상기 신경망 프로세서의 작동 전압 또는 작동 주파수를 조절한다.

[0800] 구체적으로 상기 인공지능 애플리케이션은 동영상 이미지 처리, 오브젝트 인식, 기계 번역, 음성 인식과 이미지

뷰티 등 일 수 있다.

[0801] 상황3, 상기 신경망 프로세서가 동영상 이미지 처리를 진행할 경우 동적 전압 조절 및 주파수 변조 장치(617)는 실시간으로 상기 신경망 프로세서가 동영상 이미지 처리를 진행하는 프레임 레이트를 수집한다. 상기 동영상 이미지 처리의 프레임 레이트가 타겟 프레임 레이트를 초과할 경우 상기 타겟 프레임 레이트는 사용자가 정상적으로 수요하는 동영상 이미지 처리프레임 레이트이고 동적 전압 조절 및 주파수 변조 장치(617)는 상기 신경망 프로세서에 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하여 사용자의 정상적인 동영상 이미지 처리 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력손실을 저하시킨다.

[0802] 상황4, 상기 신경망 프로세서가 음성 인식을 진행할 경우 동적 전압 조절 및 주파수 변조 장치(617)는 실시간으로 상기 신경망 프로세서의 음성 인식 속도를 수집한다. 상기 신경망 프로세서의 음성 인식 속도가 사용자의 실제 음성 인식 속도를 초과할 경우 동적 전압 조절 및 주파수 변조 장치(617)는 상기 신경망 프로세서에 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 신경망 프로세서로 하여금 그의 작동 전압 또는 작동 주파수를 저하시키도록 지시하고 사용자의 정상적인 음성 인식 수요를 만족시킴과 동시에 상기 신경망 프로세서의 전력 손실을 저하시킨다.

[0803] 상황5, 동적 전압 조절 및 주파수 변조 장치(617)는 상기 신경망 프로세서에서의 각 유닛 또는 모듈(저장매체(611), 레지스터 유닛(612), 인터커넥트 모듈(613), 연산 유닛(614), 제어 유닛(615), 데이터 액세스 유닛(616))의 작동 상태를 실시간으로 모니터링한다. 상기 신경망 프로세서의 각 유닛 또는 모듈에서의 임의의 한 유닛 또는 모듈이 유휴상태에 있을 경우 동적 전압 조절 및 주파수 변조 장치(617)는 상기 유닛 또는 모듈에 제5 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 저하시킴으로써 상기 유닛 또는 모듈의 전력 손실을 저하시킨다. 상기 유닛 또는 모듈이 다시 작동 상태에 놓일 경우 동적 전압 조절 및 주파수 변조 장치(317)는 상기 유닛 또는 모듈에 제6 전압 주파수 규제 정보를 발송하여 상기 유닛 또는 모듈의 작동 전압 또는 작동 주파수를 향상시킴으로써 상기 유닛 또는 모듈의 운행속도로 하여금 작업의 수요를 만족시키도록 한다.

[0804] 도 4g를 참조하면 도 4g는 본원 발명의 실시예에서 제공하는 단일층 콘볼루션 신경망의 순방향 연산 방법을 수행하기 위한 흐름 모식도로서 상기 방법은 상기 콘볼루션 연산장치에 응용된다. 도 4g에 도시된 바와 같이 상기 방법은 다음과 같은 단계를 포함한다.

[0805] 단계 S701, 명령 저장 유닛의 첫번째 주소에 하나의 입출력 IO명령을 미리 저장;

[0806] 단계 S702, 연산이 시작되면 컨트롤러 유닛은 상기 명령 저장 유닛의 첫번째 주소로부터 상기 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 데이터 액세스 유닛은 외부 주소 공간으로부터 대응되는 모든 콘볼루션 신경망 연산 명령을 판독하여 이를 상기 명령 저장 유닛에 캐시;

[0807] 단계 S703, 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 외부 주소 공간으로부터 메인 연산 모듈에 필요한 모든 데이터를 상기 메인 연산 모듈의 제1 저장 유닛에 발송;

[0808] 단계 S704, 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 외부 주소 공간으로부터 서브 연산 모듈에 필요한 콘볼루션 커널 데이터를 판독;

[0809] 단계 S705, 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 CONFIG 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 콘볼루션 연산장치는 상기 층의 신경망 계산에 필요한 여러 가지 상수를 배치;

[0810] 단계 S706, 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 COMPUTE 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 메인 연산 모듈은 먼저 인터커넥트 모듈을 통해 콘볼루션 윈도우 내의 입력 데이터를 N개의 서브 연산 모듈에 발송하여 상기 N개의 서브 연산 모듈의 제2 저장 유닛에 저장한 후 명령에 따라 콘볼루션 윈도우를 이동;

[0811] 단계 S707, COMPUTE 명령에 의해 디코딩된 컨트롤 신호에 근거하여 상기 N개의 서브 연산 모듈의 연산 유닛은 제3 저장 유닛으로부터 콘볼루션 커널을 판독하고 상기 제2 저장 유닛으로부터 입력 데이터를 판독하며 입력 데이터와 콘볼루션 커널의 콘볼루션 연산을 완성하고 획득한 출력 스칼라를 상기 인터커넥트 모듈을 통해 리턴;

[0812] 단계 S708, 상기 인터커넥트 모듈에서 상기 N개의 서브 연산 모듈에 의해 리턴된 출력 스칼라는 완전한 중간 벡터로 단계적으로 연결;

[0813] 단계 S709, 상기 메인 연산 모듈은 인터커넥트 모듈에 의해 리턴된 중간 벡터를 획득하고 콘볼루션 윈도우는 모든 입력 데이터를 가로 지르며 상기 메인 연산 모듈은 모든 리턴된 백터를 중간 결과로 연결하고 COMPUTE 명령에 의해 디코딩된 컨트롤 신호에 근거하여 제1 저장 유닛으로부터 바이어스 데이터를 판독하며 벡터 가산 유닛에 의해 중간 결과와 가산되어 바이어스 결과를 획득한 후 활성화 유닛에 의해 바이어스 결과를 활성화시키고 최종 출력 데이터를 상기 제1 저장 유닛에 다시 기입;

[0814] 단계 S710, 상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 상기 제1 저장 유닛의 출력 데이터를 외부 주소 공간의 지정된 주소에 저장하여 연산을 종료.

[0815] 선택적으로, 상기 방법은,

[0816] 상기 콘볼루션 연산장치의 작동 상태 정보를 실시간으로 수집하는 단계;

[0817] 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하되, 상기 전압 주파수 규제 정보는 상기 콘볼루션 연산장치로 하여금 그의 작동 전압 또는 작동 주파수를 조절하도록 지시하는 단계를 더 포함한다.

[0818] 선택적으로, 상기 콘볼루션 연산장치의 작동 상태 정보는 상기 콘볼루션 연산장치의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제1 전압 주파수 규제 정보를 포함하며 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0819] 상기 콘볼루션 연산장치의 운행속도가 타겟 속도보다 클 경우 상기 콘볼루션 연산장치에 상기 제1 전압 주파수 규제 정보를 발송하되, 상기 제1 전압 주파수 규제 정보는 상기 콘볼루션 연산장치로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하고 상기 타겟 속도는 사용자의 요구를 만족시킬 경우의 상기 칩의 운행속도인 단계를 포함한다.

[0820] 선택적으로, 상기 콘볼루션 연산장치의 작동 상태 정보는 상기 데이터 액세스 유닛의 운행속도와 메인 연산 유닛의 운행속도를 포함하고 상기 전압 주파수 규제 정보는 제2 전압 주파수 규제 정보를 포함하며 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0821] 상기 데이터 액세스 유닛의 운행속도와 상기 메인 연산 유닛의 운행속도에 근거하여 상기 데이터 액세스 유닛의 운행시간이 상기 메인 연산 유닛의 운행시간을 초과한다고 결정할 경우 상기 메인 연산 유닛에 상기 제2 전압 주파수 규제 정보를 발송하되, 상기 제2 전압 주파수 규제 정보는 상기 메인 연산 유닛으로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하는 단계를 더 포함한다.

[0822] 선택적으로, 상기 전압 주파수 규제 정보는 제3 전압 주파수 규제 정보를 포함하고 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0823] 상기 데이터 액세스 유닛의 운행속도와 상기 메인 연산 유닛의 운행속도에 근거하여 상기 메인 연산 유닛의 운행시간이 상기 데이터 액세스 유닛의 운행시간을 초과한다고 결정할 경우 상기 데이터 액세스 유닛에 상기 제3 전압 주파수 규제 정보를 발송하되, 상기 제3 전압 주파수 규제 정보는 상기 데이터 액세스 유닛으로 하여금 작동 주파수 또는 작동 전압을 저하시키도록 지시하는 단계를 더 포함한다.

[0824] 선택적으로, 상기 콘볼루션 연산장치의 작동 상태 정보는 명령 저장 유닛, 컨트롤러 유닛, 데이터 액세스 유닛, 인터커넥트 모듈, 메인 연산 모듈 및 N개의 서브 연산 모듈 중 적어도 S개의 유닛/모듈의 작동 상태 정보를 포함하고 상기 S는 1보다 크며 N+5보다 작거나 같은 정수이고 상기 전압 주파수 규제 정보는 제4 전압 주파수 규제 정보를 포함하며 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0825] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 유휴상태에 있다고 결정할 경우 상기 유닛(A)에 상기 제4 전압 주파수 규제 정보를 발송하되, 상기 제4 전압 주파수 규제 정보는 상기 유닛(A)로 하여금 그의 작동 주파수 또는 작동 전압을 저하시키도록 지시하기 위한 것인 단계를 더 포함하고,

[0826] 여기서 상기 유닛(A)는 상기 적어도 S개의 유닛/모듈에서의 임의의 하나이다.

[0827] 선택적으로, 상기 전압 주파수 규제 정보는 제5 전압 주파수 규제 정보를 포함하고 상기 콘볼루션 연산장치의 작동 상태 정보에 근거하여 상기 콘볼루션 연산장치에 전압 주파수 규제 정보를 발송하는 상기 단계는,

[0828] 상기 유닛(A)의 작동 상태 정보에 근거하여 상기 유닛(A)이 다시 작동상태에 놓일 경우 상기 유닛(A)에 기 제5 전압 주파수 규제 정보를 발송하되, 상기 제5 전압 주파수 규제 정보는 상기 유닛(A)으로 하여금 그의 작동 전압 또는 작동 주파수를 향상시키도록 지시하기 위한 것인 단계를 더 포함한다.

[0829] 설명해야 할 것은 상기 방법 실시예의 구체적인 실현과정은 도 4a-도 4f에 도시된 실시예의 관련 설명을 참조할 수 있으므로 여기서 더이상 설명하지 않는다.

[0830] 본원 발명의 하나의 가능한 실시예에서는 다층 콘볼루션 신경망의 순방향 연산을 수행하기 위한 방법을 제공하는데 이는 매 한층에 도 4g에 도시된 바와 같은 신경망의 순방향 연산의 방법인 윗 층의 콘볼루션 신경망 실행이 완료된 후 본 층의 연산 명령은 메인 연산 모듈에 저장된 윗 층의 출력 데이터 주소를 본 층의 입력 데이터 주소로 하며 명령에서의 콘볼루션 커널과 바이어스 데이터 주소는 본 층에 대응되는 주소로 변경하는 것을 포함한다.

[0831] 본원 발명의 또 다른 양태는 이미지 압축방법 및 관련 장치를 제공하는데 이미지 압축을 위한 압축 신경망을 트레이닝하여 이미지 압축의 유효성과 인식의 정확성을 향상시킨다.

[0832] 도 5a을 참조하면 도 5a은 본원 발명에서 제공하는 한가지 신경망 연산 과정인 바. 도 5a에 도시된 바와 같이 도면에서 점선으로 된 화살표는 역방향 연산을 나타내고 실선으로 된 화살표는 순방향 연산을 나타낸다. 순방향 연산에서 윗 층의 인공 신경망이 수행 완료된 후 윗 층이 얻은 출력 뉴런을 아래 층의 입력 뉴런으로 사용하여 연산(또는 상기 출력 뉴런에 어떠한 작업을 진행한 다음 아래 층의 입력 뉴런으로 사용)을 진행함과 동시에 가중치를 아래 층의 가중치로 대치한다. 역방향 연산에서 윗 층의 인공 신경망의 역방향 연산이 수행 완료된 후 윗 층이 얻은 입력 뉴런 그래디언트를 아래 층의 출력 뉴런 그래디언트로 사용하여 연산(또는 상기 입력 뉴런 그래디언트에 어떠한 작업을 진행한 다음 아래 층의 출력 뉴런 그래디언트로 사용)을 진행함과 동시에 가중치를 아래 층의 가중치로 대체한다.

[0833] 신경망의 순방향 전파 단계는 순방향 연산과 대응되고 출력 데이터가 출력될 때까지 입력 데이터를 입력하는 과정이며 역방향 전파 단계는 역방향 연산과 대응되고 최종 결과 데이터와 희망 출력 데이터 사이의 오차가 역방향으로 순방향 전파 단계를 통과하는 과정이며 계속 순환하는 순방향 전파와 역방향 전파를 통해 오차 그래디언트가 하강하는 방식에 따라 각 층의 가중치를 수정하고 각 층의 가중치를 조절하는 것도 신경망 학습 트레이닝의 과정인 바, 네트워크 출력의 오차를 절감할 수 있다.

[0834] 본원 발명에서 압축 신경망에 대한 압축 트레이닝 이미지 집합의 유형과 매 유형의 트레이닝 이미지 집합이 포함하는 트레이닝 이미지의 수량은 한정되지 않는 바, 유형이 많을 수록 수량도 더 많고 트레이닝 횟수가 많을수록 이미지 압축의 소모률도 더 낮아 이미지 인식의 정확도를 향상시키는데 편리하다.

[0835] 압축 트레이닝 이미지 집합은 여러 각도의 이미지, 여러 가지 광도에서의 이미지 또는 여러 가지 상이한 유형의 이미지 수집 기기가 수집한 이미지 등 다수의 차원을 포함할 수 있다. 상기 상이한 차원과 대응되는 압축 트레이닝 이미지 집합은 압축 신경망을 트레이닝하여 상이한 상황에서의 이미지 압축의 유효성을 향상시키고 이미지 압축방법의 적용범위를 확대한다.

[0836] 압축 트레이닝 이미지 집합에서 트레이닝 이미지가 포함하는 태그 정보에 있어서, 본원 발명은 태그 정보의 구체적인 내용을 한정하지 않는 바, 트레이닝한 이미지 부분을 표기하여 압축 신경망의 트레이닝이 완성되었는지 여부를 검출할 수 있다. 예하면 도로 영상 모니터링이 촬영한 주행 이미지에서 태그 정보는 타겟 차량번호 정보이고 주행 이미지를 압축 신경망에 입력하여 압축 이미지를 얻으며 인식 신경망 모델에 기반하여 압축 이미지를 인식함으로써 참조 차량번호 정보를 얻는데 만약 참조 차량번호 정보와 타겟 차량번호 정보가 매칭되면 압축 신경망의 트레이닝을 완성하였음을 결정하고 그렇지 않으면 압축 신경망의 현재의 트레이닝 횟수가 기설정 임계값보다 작을 경우 압축 신경망을 트레이닝 할 필요가 더 있다.

[0837] 본원 발명은 태그 정보의 유형에 대해 한정하지 않는 바, 차량 정보 일 수도 있고 안면 정보, 교통표지정보, 오브젝트 분류 정보 등 일 수도 있다.

[0838] 본원 발명에 관한 인식 신경망 모델은 이미지 인식을 위한 인식 신경망 트레이닝이 완성될 때 얻은 데이터로서 인식 신경망의 트레이닝 방법은 한정되지 않고 배치 경사 하강 알고리즘(Batch Gradient Descent, 약칭: BGD),

랜덤 경사 하강 알고리즘(Stochastic Gradient Descent, 약칭: SGD) 또는 미니 배치 경사 하강 알고리즘(mini-batch SGD)등을 이용하여 트레이닝 할 수 있으며 하나의 트레이닝 주기는 한차례의 순방향 연산과 역방향 역방향 그래디언트 전파에 의해 완성된다.

[0839] 여기서 인식 트레이닝 이미지 집합에서의 매 하나의 트레이닝 이미지는 적어도 상기 압축 트레이닝 이미지에서의 매 하나의 트레이닝 이미지의 타겟 태그 정보의 유형과 일치한 태그 정보를 포함한다. 바꾸어 말하면 인식 신경망 모델은 압축 신경망(트레이닝 대기 또는 트레이닝 완성)이 출력한 압축 이미지를 인식할 수 있다.

[0840] 예를 들어 설명하면 만약 압축 트레이닝 이미지의 태그 정보의 유형이 차량번호이면 트레이닝 이미지의 태그 정보의 유형에 적어도 차량번호가 포함됨을 인식함으로써 인식 신경망 모델이 압축 신경망에 대해 출력한 압축 이미지를 인식하도록 담보하여 차량번호 정보를 얻는다.

[0841] 선택적으로, 압축 트레이닝 이미지 집합은 적어도 인식 트레이닝 이미지 집합을 포함한다.

[0842] 트레이닝 이미지 집합에서의 이미지가 각도, 광선 또는 이미지 수집 기기 등 요소의 영향을 받으므로 인식 트레이닝 이미지 집합을 이용하여 트레이닝 할 경우 인식 신경망 모델의 정확도를 향상시킬 수 있어 압축 신경망의 트레이닝 효율을 향상, 즉 이미지 압축의 유효성을 향상시키는데 편리하다.

[0843] 도 5b를 참조하면 도 5b는 본원 발명의 실시예에서 제공하는 이미지 압축방법의 흐름모식도이다. 도 5b에 도시된 바와 같이 상기 이미지 압축방법은 다음과 같은 단계를 포함한다.

[0844] 단계 S201, 제1 해상도의 원본 이미지를 획득한다.

[0845] 여기서 제1 해상도는 압축 신경망의 입력 해상도이고 제2 해상도는 제1 해상도보다 작으며 압축 신경망의 출력 해상도, 즉 압축 신경망을 출력하는 이미지의 압축비(제2 해상도와 제1 해상도의 비)는 고정된 것인 바, 바꾸어 말하면 동일한 압축 신경망 모델에 기반하여 상이한 이미지에 대해 압축을 진행하여 동일한 압축비의 이미지를 얻을 수 있다.

[0846] 원본 이미지는 압축 신경망의 압축 트레이닝 이미지 집합에서의 임의의 트레이닝 이미지이고 원본 이미지의 태그 정보를 타겟 태그 정보로 사용한다. 본원 발명의 태그 정보는 한정되지 않는 바, 인위적으로 인식하여 표기함으로써 얻은 것일 수 있고 원본 이미지를 인식 신경망에 입력하여 인식 신경망 모델에 기빈하여 인식함으로써 얻은 것 등 일 수도 있다.

[0847] 단계 S202, 타겟 모델에 기반하여 상기 원본 이미지를 압축하여제2 해상도의 압축 이미지를 얻는다.

[0848] 여기서 타겟 모델은 상기 압축 신경망의 현재의 신경망 모델, 즉 타겟 모델는 압축 신경망의 현재 파라미터이다. 타겟 모델에 기반하여 해상도와 압축 신경망이 같은 입력 해상도의 원본 이미지를 압축함으로써 해상도와 압축 신경망이 같은 출력 해상도의 압축 이미지를 얻을 수 있다.

[0849] 선택적으로, 타겟 모델에 기반하여 상기 원본 이미지를 압축하여 제2 해상도의 압축 이미지를 얻는 상기 단계는, 상기 타겟 모델에 의해 상기 원본 이미지를 인식하여 다수의 이미지 정보를 얻는 단계; 상기 타겟 모델과 상기 다수의 이미지 정보에 의해 상기 원본 이미지를 압축하여 상기 압축 이미지를 얻는 단계를 포함한다.

[0850] 상술한 바와 같은 트레이닝 이미지는 다수의 차원을 포함하는데 우선 타겟 모델에 기반하여 원본 이미지를 인식하여 매 하나의 차원과 대응되는 이미지 정보를 결정한 다음 매 하나의 이미지 정보에 대해 원본 이미지를 압축하여 상이한 차원의 이미지 압축의 정확도를 향상시킬 수 있다.

[0851] 단계 S203, 인식 신경망 모델에 기반하여 상기 압축 이미지를 인식함으로써 참고 태그 정보를 획득한다.

[0852] 본원 발명은 인식 방법을 한정하지 않는 바, 특징 추출과 특징 인식 두 부분을 포함할 수 있고 특징 인식을 진행하여 얻은 결과를 참고 태그 정보로 사용하는데, 예하면 주행 이미지 압축 뒤에 얻은 주행 압축 이미지와 대응되는 참고 태그 정보는 차량번호이고 안면 이미지 압축 뒤에 얻은 안면 압축 이미지와 대응되는 참고 태그 정보는 안면 인식 결과이다.

[0853] 선택적으로, 인식 신경망 모델에 기반하여 상기 압축 이미지를 인식함으로써 참고 태그 정보를 획득하는 상기 단계는, 상기 압축 이미지를 전처리하여 인식 대기 이미지를 얻는 단계; 상기 인식 신경망 모델에 기반하여 상기 인식 대기 이미지를 인식함으로써 상기 참고 태그 정보를 얻는 단계를 포함한다.

[0854] 전처리는 데이터 양식 변환처리(예하면 정규화 처리, 정수형 데이터 변환 등), 데이터 중복 제거 처리, 데이터 이상 처리, 데이터 결여 보충 처리 등에서의 임의의 하나 또는 다수를 포함하나 이에 한정되지 않는다. 압축 이

미지를 전처리함으로써 이미지 인식의 인식효율과 정확도를 향상시킬 수 있다.

[0855] 마찬가지로 제1 해상도의 원본 이미지를 획득하는 상기 단계는, 입력 이미지를 수신하는 단계; 상기 입력 이미지를 전처리하여 상기 원본 이미지를 얻는 단계를 포함한다. 입력 이미지에 대한 전처리를 통해 이미지 압축의 압축효율을 향상시킬 수 있다.

[0856] 상술한 전처리는 사이즈 처리를 더 포함하는데 신경망이 고정된 사이즈 요구가 있으므로 상기 신경망의 기본 이미지 크기와 동일한 이미지를 처리할 수밖에 없다. 압축 신경망의 기본 이미지 크기를 제1 기본 이미지 크기로 하고 인식 신경망의 기본 이미지 크기를 제2 기본 이미지 크기로 하는 바, 즉 압축 신경망이 입력 이미지 사이즈에 대한 요구는 이미지 크기가 제1 기본 이미지 크기와 같은 것이고 인식 신경망이 입력 이미지 사이즈에 대한 요구는 이미지 크기가 제2 기본 이미지 크기와 같은 것이다. 압축 신경망은 제1 기본 이미지 크기를 만족하는 인식 대기 이미지를 인식함으로써 참고 태그 정보를 획득할 수 있다.

[0857] 본원 발명은 사이즈 처리에 대한 구체적인 방식을 한정하지 않는 바, 픽셀 포인트를 클리핑 또는 충진하는 방식을 포함할 수 있고 기본 이미지 크기에 따라 스케일링 하는 방법을 포함할 수도 있으며 또 입력 이미지를 다운 샘플링하는 방법 등을 포함할 수도 있다.

[0858] 여기서 외곽 픽셀 포인트 클리핑은 이미지 외곽의 중요하지 않은 정보영역이고 다운 샘플링 처리는 특정 신호의 샘플링 레이트를 저하시키는 과정으로서 예하면 4개의 이웃하는 픽셀 포인트에서 평균값을 취하여 처리 후의 이미지의 대응 위치에서의 하나의 픽셀 포인트의 값으로 함으로써 이미지 크기를 절감한다.

[0859] 선택적으로 상기 압축 이미지를 전처리하여 인식 대기 이미지를 얻는 상기 단계는, 상기 압축 이미지의 이미지 크기가 인식 신경망의 기본 이미지 크기보다 작을 경우 상기 기본 이미지 크기에 따라 상기 압축 이미지에 대해 픽셀 포인트를 충진하여 상기 인식 대기 이미지를 얻는 단계를 포함한다.

[0860] 본원 발명은 픽셀 포인트에 대해 한정하지 않는 바, 임의의 커러 모드가 대응하는 것일 수 있는데 예하면 rgb(0, 0, 0) 일 수 있다. 픽셀 포인트에 충진하는 구체적인 위치도 한정하지 않는데 압축 이미지를 제외한 임의의 위치일 수 있는 바, 즉 압축 이미지를 처리하지 않고 픽셀 포인트를 충진한느 방식으로 이미지 확장을 진행함으로써 압축 이미지를 변형시키지 않아 이미지 인식의 인식효율과 정확도를 향상시키는데 편리하다.

[0861] 예를 들어 설명하면 도 5c에 도시된 바와 같이 압축 이미지를 인식 대기 이미지의 좌측 상단에 위치시키는데 인식 대기 이미지는 압축 이미지를 제외한 위치에 픽셀 포인트를 충진시키는 것이다.

[0862] 마찬가지로 상기 입력 이미지를 전처리하여 상기 원본 이미지를 얻는 상기 단계는, 상기 입력 이미지의 이미지 크기가 상기 압축 신경망의 제1 기본 이미지 크기보다 작을 경우 상기 제1 기본 이미지 크기에 따라 상기 입력 이미지에 픽셀 포인트를 충진함으로써 상기 원본 이미지를 얻는 단계를 포함한다. 픽셀 포인트를 충진함으로써 압축된 원본 이미지가 인식 신경망에 의해 인식되어 참고 태그 정보를 얻도록 하고 픽셀 포인트는 변화되지 않은 입력 이미지의 압축률을 충진하여 압축 신경망을 트레이닝하는 효율과 정확도를 향상시키는데 편리하다.

[0863] 단계 S204, 상기 타겟 태그 정보와 상기 참고 태그 정보에 근거하여 손실함수를 획득한다.

[0864] 본원 발명에서 손실함수는 타겟 태그 저오와 참고 태그 정보 사이의 오차 크기를 설명하기 위한 것이고 태그 정보는 다수의 차원을 포함하며 일반적으로 평방차 공식을 사용하여 계산한다.

[0865] $$E=\frac{1}{2}\sum_{k=1}^{c}\left(t_k-y_k\right)^2$$

[0866] 여기서 c는 태그 정보의 차원이고 t_k는 참고 태그 정보의 k번째 차원이며 y_k는 타겟 태그 정보의 k번째 차원이다.

[0867] 단계 S205, 상기 손실함수가 제1 임계값에 수렴되었는지 여부 또는 상기 압축 신경망의 현재의 트레이닝 횟수가 제2 임계값보다 크거나 같은지 여부를 판정하여 만약 그렇다면 단계 S206을 수행하고 만약 아니면 단계 S207을 수행한다.

[0868] 본원 발명에 관련된 압축 신경망의 트레이닝 방법에서 매 하나의 트레이닝 이미지와 대응되는 트레이닝 주기는 한차례의 순반향 연산과 역방향 그래디언트 전파에 의해 완성되는데 손실함수의 임계값을 제1 임계값으로 설치하고 압축 신경망의 트레이닝 횟수의 임계값을 제2 임계값으로 설치한다. 바꾸어 말하면 만약 손실함수가 제1 임계값에 수렴되거나 또는 트레이닝 횟수가 제2 임계값보다 크거나 같으면 압축 신경망의 트레이닝을 완성하고

상기 타겟 모델을 상기 압축 신경망 트레이닝이 완성될 때 대응되는 압축 신경망 모델로 사용하며 그렇지 않을 경우 손실함수에 근거하여 압축 신경망의 역방향 전파 단계에 진입, 즉 손실함수에 근거하여 타겟 모델을 업데이트하고 그 다음의 트레이닝 이미지를 트레이닝, 즉 단계 S202-단계 S205를 수행하며 상기 조건을 만족시킬 경우 트레이닝을 완료하여 단계 S206을 수행하기를 대기한다.

[0869] 본원 발명은 압축 신경망의 역방향 트레이닝 방법을 한정하지 않는 바, 선택적으로 도 5d에서 제공하는 단일층 신경망 연산방법의 흐름모식도를 참조할 수 있으며 도 5d는 도 5e에 도시된 압축 신경망의 역방향 트레이님을 수행하기 위한 장치의 구조모식도에 응용될 수 있다.

[0870] 도 5e에 도시된 바와 같이 상기 장치는 명령 캐시 유닛(21), 컨트롤러 유닛(22), 직접 메모리 액세스 유닛(23), H트리 모듈(24), 메인 연산 모듈(25)과 다수의 서브 연산 모듈(26)을 포함하고 상기 장치는 하드웨어 회로(에하면 특정 용도 지향 직접 회로ASIC)를 통해 실현될 수 있다.

[0871] 여기서 명령 캐시 유닛(21)은 직접 메모리 액세스 유닛(23)을 통해 명령을 판독하고 판독된 명령을 캐시하며; 컨트롤러 유닛(22)은 명령 캐시 유닛(21)에서 명령을 판독하고 명령을 직접 메모리 액세스 유닛(23), 메인 연산 모듈(25)과 서브 연산 모듈(26)과 같은 기타 모듈의 행동을 제어하기 위한 마이크로 명령으로 디코딩하며; 직접 메모리 액세스 유닛(23)는 외부 주소 공간을 액세스하여 장치 내부의 각 캐시 유닛에 데이터를 판독 기록함으로써 데이터의 로딩과 저장을 완성할 수 있다.

[0872] 도 5f을 참조하면 도 5f은 H트리 모듈(24)의 구조를 도시하는바 도 5f에 도시된 바와 같이 H트리 모듈(24)은 메인 연산 모듈(25)과 다수의 서브 연산 모듈(26) 사이의 데이터 통로를 구성하고 H트리 모양의 구조를 가진다. H트리는 다수의 노드로 구성된 이진 트리 통로로서 매 하나의 노드는 상류의 데이터를 마찬가지로 하류의 두 개의 노드에 발송하고 하류의 두 개의 노드가 리턴한 데이터를 합병하여 상류의 노드에 리턴한다. 예하면 신경망 역방향 연산과정에서 하류의 두 개의 노드가 리턴한 벡터는 현재 노드를 가하여 하나의 벡터로 하고 상류 노드에 리턴한다. 매 층의 인공 신경망이 계산을 시작하는 단계에서 메인 연산 모듈(25) 내의 입력 그래디언트는 H트리 모듈(24)을 통해 각 서브 연산 모듈(26)에 발송되고 서브 연산 모듈(26)의 계산 과정이 완성된 후 매 하나의 서브 연산 모듈(26)이 출력한 출력 그래디언트 벡터 부분은 H트리 모듈(24)에서 단계적으로 둘씩 가하게 되는 바, 즉 모든 출력 그래디언트 벡터 부분에 대해 합을 구하여 최종적인 출력 그래디언트 벡터로 사용한다.

[0873] 도 5g을 참조하면 도 5g은 메인 연산 모듈(25)의 구조모식도로서 도 5g에 도시된 바와 같이 메인 연산 모듈(25)은 연산 유닛(251), 데이터 의존관계 판정 유닛(252)과 뉴런 캐시 유닛(253)을 포함한다.

[0874] 여기서 뉴런 캐시 유닛(253)은(캐시)메인 연산 모듈(25)이 계산 과정에서 사용하는 입력 데이터와 출력 데이터를 캐시한다. 연산 유닛(251)은 메인 연산 모듈의 여러 가지 연산 기능을 완성한다. 데이터 의존관계 판정 유닛(252)은 연산 유닛(251)이 뉴런 캐시 유닛(253)을 판독 기록하는 포트인 동시에 뉴런 캐시 유닛(253)에서의 데이터의 판독 기록이 일치성 충돌이 발생하지 않도록 담보한다. 구체적으로 데이터 의존관계 판정 유닛(252)은 실행되지 않은 마이크로 명령과 수행과정에 있는 마이크로 명령의 데이터 사이에 의존관계가 있는지 여부를 판정하고 만약 의존관계가 존재하지 않으면 상기 마이크로 명령을 즉시 발송하도록 허용하고 그렇지 않으면 상기 마이크로 명령이 의존하는 모든 마이크로 명령이 모두 수행된 후 상기 마이크로 명령이 발송되되록 허용한다. 예하면 데이터 의존관계 판정 유닛(252)에 발송되는 모든 마이크로 명령이 모두 데이터 의존관계 판정 유닛(252) 내부의 명령 행렬에 저장되는데 상기 행렬에서 명령을 판독하는 판독 데이터의 범위가 행렬의 앞지리의 명령을 기입하고 데이터를 기입하는 범위와 충돌되면 상기 명령은 반드시 의존하는 명령 기입이 수행된 후에야만 수행될 수 있다. 이와 동시에 데이터 의존관계 판정 유닛(252)은 뉴런 캐시 유닛(253)으로부터 입력 그래디언트 벡터를 판독하고 H트리 모듈(24)을 통해 서브 연산 모듈(26)에 발송하며 서브 연산 모듈(26)의 출력 데이터는 H트리 모듈(24)을 통해 연산 유닛(251)에 직접 발송한다. 컨트롤러 유닛(22)이 출력한 명령은 연산유닛(251)과 데이터 의존관계 판정 유닛(252)에 발송되어 이의 행동을 제어한다.

[0875] 도 5h을 참조하면 도 5h은 서브 연산 모듈(26)의 구조모식도로서 도 5h에 도시된 바와 같이 매 하나의 서브 연산 모듈(26)은 연산 유닛(261), 데이터 의존관계 판정 유닛(262), 뉴런 캐시 유닛(263), 가중치 캐시 유닛(264)과 가중치 그래디언트 캐시 유닛(265)을 포함한다.

[0876] 여기서 연산 유닛(261)은 컨트롤러 유닛(22)이 발송한 마이크로 명령을 수신하고 산술 논리 연산을 진행한다.

[0877] 데이터 의존관계 판정 유닛(262)은 계산 과정에서 캐시 유닛에 대한 판독 기록 작업을 담당한다. 데이터 의존관계 판정 유닛(262)은 캐시 유닛의 판독 기록에 일치성 충돌이 존재하지 않음을 담보한다. 구체적으로 데이터 의존관계 판정 유닛(262)는 실행되지 않은 마이크로 명령과 수행 과정에 있는 마이크로 명령의 데이터 사이에 의

존관계가 있는지 여부를 판정하여 만약 의존관계가 존재하지 않으면 상기 마이크로 명령이 즉시 발송되도록 허용하고 그렇지 않으면 상기 마이크로 명령이 의존하는 모든 마이크로 명령이 모두 실행을 완성할 때까지 대기한 후 상기 마이크로 명령을 발송하도록 허용한다. 예하면 데이터 의존관계 판정 유닛(262)에 발송되는 모든 마이크로 명령은 모두 데이터 의존관계 판정 유닛(262) 내부의 명령행렬에 저장되고 상기 행렬에서 명령을 판독하는 판독 데이터의 범위가 행렬의 앞지리의 명령을 기입하고 데이터를 기입하는 범위와 충돌되면 상기 명령은 반드시 의존하는 명령 기입이 수행된 후에야만 수행될 수 있다.

[0878] 뉴런 캐시 유닛(263)은 입력 그래디언트 벡터 데이터 및 상기 서브 연산 모듈(26)이 계산하여 얻은 출력 그래디언트 벡터의 부분합을 캐시한다.

[0879] 가중치 캐시 유닛(264)은 상기 서브 연산 모듈(26)이 계산 과정에서 필요로 하는 가중치 벡터를 캐시한다. 매 하나의 서브 연산 모듈에 대하여 모두 가중치 매트릭스에서 상기 서브 연산 모듈(26)과 대응되는 열만 저장하게 된다.

[0880] 가중치 그래디언트 캐시 유닛(265)은 상응한 서브 연산 모듈이 가중치를 업데이트하는 과정에서 필요로 하는 가중치 그래디언트 데이터를 캐시한다. 매 하나의 서브 연산 모듈(26)이 저장하는 가중치 그래디언트 데이터는 그가 저장한 가중치 벡터와 서로 대응된다.

[0881] 서브 연산 모듈(26)은 매 층의 인공 신경망의 역방향 트레이닝이 그래디언트 벡터를 출력하는 과정에서 병행할 수 있는 앞부분 및 가중치의 업데이트를 실현한다. 인공 신경망 풀 연결층(MLP)을 예로 하면 과정은 out_gradient = w*in_gradient이고 여기서 가중치 매트릭스w와 입력 그래디언트 벡터in_gradient의 곱셈은 관련되지 않는 병행 계산 서브 임무로 구획될 수 있으며 out_gradient와 in_gradient는 열 벡터이고 매 하나의 서브 연산 모듈은 in_gradient에서 상응한 일부 스칼라 요소와 가중치 매트릭스w가 대응되는 열의 곱만을 계산는데 얻은 매 하나의 출력 벡터는 모두 최종 결과의 하나의 누적 대기 부분합이고 이러한 부분합은 H트리에서 단계적으로 둘씩 가해져 최종 결과를 얻는다. 그러므로 계산 과정은 병행하는 계산 부분합의 과정과 그 뒤의 누적 과정으로 변한다. 매 하나의 서브 연산 모듈(26)은 출력 그래디언트 벡터의 부분합을 산출하고 모든 부분합은 H트리 모듈(24)에서 합계 연산을 완성하여 최후의 출력 그래디언트 벡터를 얻는다. 매 하나의 서브 연산 모듈(26)은 동시에 입력 그래디언트 벡터와 순방향 연산을 진행할 경우의 매 층의 출력 값을 곱하여 가중치 그래디언트를 산출함으로써 서브 연산 모듈(26)에 저장된 가중치를 업데이트한다. 순방향 연산과 역방향 트레이닝은 신경망 알고리즘의 두 개의 주요한 과정인 바, 신경망은이 네트워크에서의 가중치를 트레이닝(업데이트) 하려면 우선 입력 벡턱가 현재 가중치로 구성된 네트워크에서의 순방향 출력을 계산해야 하는데 이는 순방향 과정이고, 다음 출력 값과 입력 벡터 자체의 타겟 값 사이의 차이값에 근거하여 역방향으로 차례로 매 층의 가중치를 트레이닝(업데이트)한다. 순방향 계산 과정에서는 매 한 층의 출력 벡터 및 활성화 함수의 도함수 값을 저장하게 되는데 이러한 데이터는 역방향 트레이닝 과정에 필요한 것이므로 역방향 트레이닝이 시작될 경우 이러한 데이터는 이미 존재하도록 담보된다. 순방향 연산에서 매 층의 출력 값은 역방향 연산이 시작될 경우 이미 존재하는 데이터로서 직접 메모리 액세스 유닛을 통해 메인 연산 모듈에 캐시되고 H트리를 통해 서브 연산 모듈에 발송될 수 있다. 메인 연산 모듈(25)은 출력 그래디언트 벡터에 기반하여 후속적인 계산을 진행하는 바, 예하면 출력 그래디언트 벡터와 순방향 연산을 진행할 경우의 활성화 함수의 도함수를 곱하여 다음 한 층의 입력 그래디언트 값을 얻는다. 순방향 연산을 진행할 경우의 활성화 함수의 도함수는 역방향 연산이 시작될 경우에 이미 존재하는 데이터로서 직접 메모리 액세스 유닛을 통해 메인 연산 모듈에 캐시될 수 있다.

[0882] 본 발명의 실시예에 근거하면 상술한 장치에서 인공 신경망의 순방향 연산을 수행하는 명령 집합을 더 제공한다. 명령 집합에는 CONFIG명령, COMPUTE명령, IO명령, NOP 명령, JUMP 명령과 MOVE명령이 포함되는데 여기서,

[0883] CONFIG명령, 매 층마다의 인공 신경망 계산이 시작되기 전에 현재 층의 계산에 필요한 여러 가지 상수를 배치;

[0884] COMPUTE명령, 매 층의 인공 신경망의 산술 논리 계산을 완성;

[0885] IO명령, 외부 저장공간으로부터 계산에 필요한 입력 데이터를 판독하고 계산이 완료된 후 데이터를 외부공간에 저장한;

[0886] NOP 명령, 현재 장치 내지 내부의 모든 마이크로 명령 캐시 행렬에서의 마이크로 명령을 정리하여 NOP 명령 이전의 모든 명령이 모두 수행 완료되도록 담보한다. NOP 명령 자체는 그 어떤 작업도 포함하지 않음;

[0887] JUMP 명령, 컨트롤러가 명령 캐시 유닛으로부터 판독될 다음 명령 주소의 점프를 제어하는 것을 담당함으로써

제어 흐름의 점프를 실현;

[0888] MOVE명령, 상기 장치 내부 주소공간의 어느 한 주소의 데이터를 상기 장치 내부 주소공간의 다른 한 주소로 이동시키는 것을 담당하되, 상기 과정은 연산 유닛과 별도로 수행과정에서 연산 유닛의 리소스를 점용하지 않는다.

[0889] 도 5i를 참조하면 도 5i는 본원 발명의 실시예에서 제공하는 압축 신경망의 역방향 트레이닝의 예시적 블럭도이다. 출력 그래디언트 벡터를 계산하는 과정은 out_gradient=w*in_gradient이고 여기서 가중치 매트릭스w와 입력 그래디언트 벡터in_gradient의 매트릭스 벡터 곱셈은 관련되지 않는 병행 계산 서브 임무로 구획될 수 있으며 매 하나의 서브 연산 모듈(26)은 출력 그래디언트 벡터의 부분합을 산출하고 모든 부분합은 H트리 모듈(24)에서 합계 연산을 완성하여 최후의 출력 그래디언트 벡터를 얻는다. 도 5i에서 윗 층의 출력 그래디언트 벡터input gradient는 대응되는 활성화 함수 도함수를 곱하여 본 층의 입력 데이터를 얻고, 다음 가중치 매트릭스를 곱하여 출력 그래디언트 벡터를 얻는다. 가중치 업데이트 그래디언트를 계산하는 과정은 dw = x * in_gradient이고 여기서 매 하나의 서브 연산 모듈(26)은 본 모듈의 대응 부분의 가중치의 업데이트 그래디언트를 계산한다. 서브 연산 모듈(26)은 입력 그래디언트와 순방향 연산을 진행할 경우의 입력 뉴런을 곱하여 가중치 업데이트 그래디언트dw를 산출한 다음 w, dw와 윗 층의 가중치를 업데이트 할 경우 사용하는 가중치 업데이트 그래디언트dw'를 사용하여 명령이 설치한 학습률에 근거하여 가중치w를 업데이트한다.

[0890] 도 5i에 도시된 내용을 참조하면 input gradient(도 5i에서의[input gradient0, …, input gradient3])는 n+1번째 층의 출력 그래디언트 벡터이고 상기 벡터는 우선 순방향 연산 과정에서의 n번째 층의 도함수 값(도 5i中的[f'(out0), …, f'(out3)])과 곱하여 n번째 층의 입력 그래디언트 벡터를 얻으며 상기 과정은 메인 연산 모듈(25)에서 완성되고 H트리 모듈(24)에서 서브 연산 모듈(26)에 발송되어 서브 연산 모듈(26)의 뉴런 캐시 유닛(263)에 임시 저장된다. 다음 입력 그래디언트 벡터는 가중치 매트릭스와 곱하여 n번째 층의 출력 그래디언트 벡터를 얻는다. 이 과정에서 i번째 서브 연산 모듈은 입력 그래디언트 벡터에서의 i번째 스칼라와 가중치 매트릭스에서의 열벡터[w_i0, …, w_iN]의 곱을 계산하고 획득한 출력 벡터는 H트리 모듈(24)에서 단계적으로 둘씩 더하여 최후의 출력 그래디언트 벡터output gradient(도 5i에서의 [output gradient0, …, output gradient3])를 얻는다.

[0891] 이와 동시에 서브 연산 모듈(26)은 본 모듈에 저장된 가중치를 업데이트할 필요가 있는데 가중치 업데이트 그래디언트를 계산하는 과정은 dw_ij = x_j * in_gradient_i이고 여기서 x_j는 순방향 연산을 진행할 경우의 n번째 층의 입력(즉 n-1번째 층의 출력) 벡터의 j번째 요소이며 in_gradient_i는 역방향 연산의 n번째 층의 입력 그래디언트 벡터(즉 도 5i에서의 input gradient와 도함수 f'의 곱)의 i번째 요소이다. 순방향 연산을 진행할 경우 n번째 층의 입력은 역방향 트레이닝이 시작될 때 이미 존재하는 데이터로서 H트리 모듈(24)을 통해 서브 연산 모듈(26)에 전송됨과 동시에 뉴런 캐시 유닛(263)에 임시 저자○된다. 서브 연산 모듈(26)에서는 출력 그래디언트 벡터의 부분합의 계산을 완성한 후 입력 그래디언트 벡터의 i번째 스칼라와 순방향 연산의 n번째 층의 입력 벡터를 곱하여 업데이트 가중치의 그래디언트 벡터dw를 얻고 이로써 가중치를 업데이트한다.

[0892] 도 5d에 도시된 바와 같이 명령 캐시 유닛의 첫번째 주소가 미리 하나의 IO명령을 저장한 상태에 놓이면 컨트롤러 유닛은 명령 캐시 유닛의 첫번째 주소로부터 상기 IO명령을 판독하고 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 외부 주소 공간으로부터 상기 단일층 인공 신경망의 역방향 트레이닝과 관련된 모든 명령을 판독하며 이를 명령 캐시 유닛에 캐시하고, 컨트롤러 유닛은 이어서 명령 캐시 유닛으로부터 다음의 IO명령을 판독하며 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 외부 주소 공간으로부터 메인 연산 모듈에 필요한 모든 데이터를 판독하여 메인 연산 모듈의 뉴런 캐시 유닛에 발송하는데 상기 데이터는 그 전의 순방향 연산을 진행할 경우의 입력 뉴런과 활성화 함수의 도함수 값 및 입력 그래디언트 벡터를 포함하며 컨트롤러 유닛은 이어서 명령 캐시 유닛으로부터 다음의 IO명령을 판독하고 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 외부 주소 공간으로부터 서브 연산 모듈에 필요한 모든 가중치 데이터와 가중치 그래디언트 데이터를 판독하여 상응한 서브 연산 모듈의 가중치 캐시 윤닛과 가중치 그래디언트 캐시 유닛에 각각 저장하며 컨트롤러 유닛은 이이서 명령 캐시 유닛으로부터 다음의 CONFIG명령을 판독하고 연산 유닛은 디코딩된 마이크로 명령에서의 파라미터에 근거하여 연산 유닛 내부의 레지스터의 값을 배치하는데 이는 상기 층의 신경망 계산에 필요한 여러 가지 상수, 본 층 계산의 정확도 설치, 가중치를 업데이트 할 때의 학습률 등을 포함하고 컨트롤러 유닛은 이어서 명령 캐시 유닛으로부터 다음의 COMPUTE명령을 판독하며 디코딩된 마이크로 명령에 근거하여 메인 연산 모듈은 H트리 모듈을 통해 입력 그래디언트 벡터와 순방향 연산을 진행할 때의 입력 뉴런을 각 서브 연산 모듈에 발송하고 상기 입력 그래디언트 벡터와 순방향 연산을 진행할 때의 입력 뉴런은 서브 연산 모듈의 뉴런 캐시 유닛에 저장되며 COMPUTE명령에 근거하여 마이크로 명령을 디코딩하고 서브 연산 모듈의 연산

유닛은 가중치 캐시 유닛으로부터 가중치 벡터(즉 상기 서브 연산 모듈이 저장한 가중치 매트릭스의 부분 열)를 판독하여 가중치 벡터와 입력 그래디언트 벡터의 벡터 곱셈 스칼라 연산을 완성하며 출력 벡터 부분합을 H트리를 통해 리턴함과 동시에 서브 연산 모듈은 입력 그래디언트 벡터와 입력 뉴런을 곱하여 가중치 그래디언트를 얻어 가중치 그래디언트 캐시 유닛에 저장하며 H트리 모듈에 있어서 각 서브 연산 모듈이 리턴한 출력 그래디언트 부분합은 단계적으로 둘씩 더하여 완전한 출력 그래디언트 벡터를 얻고 메인 연산 모듈은 H트리 모듈의 리턴 값을 얻으며 COMPUTE명령에 근거하여 마이크로 명령을 디코딩하며 뉴런 캐시 유닛으로부터 순방향 연산을 진행할 때의 활성화 함수의 도함수 값을 얻고 도함수 값에 리턴된 출력 벡터를 곱하여 다음의 역방향 트레이닝된 입력 그래디언트 벡터를 얻으며 이를 뉴런 캐시 유닛에 기입하고 컨트롤러 유닛은 이어서 명령 캐시 유닛으로부터 다음의 COMPUTE명령을 판독하며 디코딩된 마이크로 명령에 근거하여 서브 연산 모듈은 가중치 캐시 유닛으로부터 가중치w를 판독하고 가중치 그래디언트 캐시 유닛으로부터 이번 가중치 그래디언트dw와 이 전의 업데이트 가중치가 사용한 가중치 그래디언트dw’를 판독하며 가중치w를 업데이트하고 컨트롤러 유닛은 이어서 명령 캐시 유닛으로부터 다음의 IO명령을 판독하며 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 뉴런 캐시 유닛에서의 출력 그래디언트 벡터를 외부 주소 공간이 지정한 주소에 저장함으로써 연산이 완료된다.

[0893] 다층 인공 신경망에 있어서, 이의 실현과정은 단일층 신경망과 유사한 바, 윗 층의 인공 신경망이 수행된 후 다음 층의 연산 명령은 메인 연산 모듈에서 산출된 출력 그래디언트 벡터를 다음 층에서 트레이닝된 입력 그래디언트 벡터로 하여 상기와 같은 계산 과정을 진행하고 명령에서의 가중치 주소와 가중치 그래디언트 주소도 본 층과 대응되는 주소로 변경된다.

[0894] 신경망의 역방향 트레이닝을 수행하기 위한 장치를 사용함으로써 다층 인공 신경망의 순방향 연산에 대한 지지를 효과적으로 향상시킨다. 또한 다층 인공 신경망 연산 알고리즘에 대해 사용한 전용 온-칩 캐시를 통해 입력 뉴런과 가중치 데이터의 중요성을 충분히 발굴하여 메모리에서 이러한 데이터를 반복적으로 판독하는 것을 방지하고 메모리 액세스 대역폭을 저하시켜 메모리 대역폭이 다층 인공 신경망 연산 및 그의 트레이닝 알고리즘에 가져온 기능적 난관의 문제를 방지하였다.

[0895] 단계 S206, 상기 제1 해상도의 타겟 원본 이미지를 획득하고 압축 신경망 모델에 기반하여 상기 타겟 원본 이미지를 압축하여 상기 제2 해상도의 타겟 압축 이미지를 얻는다.

[0896] 여기서 타겟 원본 이미지는 트레이닝 이미지의 태그 정보의 유형과 일치한 이미지(동일한 데이터 집합의 이미지에 속함)이다. 만약 손실함수가 제1 임계값에 수렴되거나 또는 트레이닝 횟수가 제2 임계값보다 크거나 같으면 압축 신경망은 트레이닝을 완성하고 압축 신경망에 직접 입력되어 이미지 압축을 진행함으로써 타겟 압축 이미지를 얻으며 상기 타겟 압축 이미지는 인식 신경망에 의해 인식될 수 있다.

[0897] 선택적으로 압축 신경망 모델이 상기 타겟 원본 이미지를 압축하여 상기 제2 해상도의 타겟 압축 이미지를 얻는 상기 단계 다음에 상기 방법은, 상기 인식 신경망 모델에 기반하여 상기 타겟 압축 이미지를 인식하여 상기 타겟 원본 이미지의 태그 정보를 얻고 상기 타겟 원본 이미지의 태그 정보를 저장하는 단계를 더 포함한다.

[0898] 바꾸어 말하면 압축 신경망 트레이닝이 완성된 후 인식 신경망 모델에 기반하여 압축 이미지를 인식함으로써 인공적으로 태그 정보를 인식하는 효율과 정확도를 향상시킬 수 있다.

[0899] 단계 S207, 상기 손실함수에 근거하여 상기 타겟 모델을 업데이트하여 업데이트 모델을 얻고 상기 업데이트 모델을 상기 타겟 모델로 사용하며 그 다음의 한 트레이닝 이미지를 상기 원본 이미지로 사용하여 단계 S202를 수행한다.

[0900] 이해할 수 있는 것은, 이미 트레이닝을 완성하여 얻은 인식 신경망 모델이 획득한 참고 태그값과 원본 이미지가 포함하는 타겟 태그값을 통해 손실함수를 획득하고 손실함수가 기설정 조건을 만족하거나 또는 압축 신경망의 현재의 트레이닝 횟수가 기설정 임계값을 초과할 경우 트레이닝을 완성하며 그렇지 않을 경우 압축 신경망을 트레이닝하는 것을 통해 그 가중치를 반복적으로 조절, 즉 동일한 이미지에서 매 하나의 픽셀에 대해 나타내는 이미지 콘텐츠를 조절하여 압축 신경망의 손실을 절감한다. 또한 트레이닝을 완성하여 얻은 압축 신경망 모델에 대해 이미지 압축을 진행하여 이미지 압축의 유효성을 향상시킴으로써 인식의 정확도를 향상시키는데 편리하다.

[0901] 도 5j을 참조하면 도 5j은 본원 발명의 실시예에서 제공하는 이미지 압축 장치(300)의 구조모식도로서 도 5j에 도시된 바와 같이 이미지 압축 장치(300)는 프로세서(301), 메모리(302)를 포함한다.

[0902] 본원 발명의 실시예에서 메모리(302)는 제1 임계값, 제2 임계값, 압축 신경망의 현재의 신경망 모델과 트레이닝 횟수, 상기 압축 신경망의 압축 트레이닝 이미지 집합과 상기 압축 트레이닝 이미지 집합에서의 매 하나의 트레이닝 이미지의 태그 정보, 인식 신경망 모델, 압축 신경망 모델을 저장하고 상기 압축 신경망의 현재의 신경망

모델을 타겟 모델로 사용하며 상기 압축 신경망 모델은 상기 압축 신경망 트레이닝이 완성될 때 대응되는 타겟 모델이고 상기 인식 신경망 모델은 인식 신경망 트레이닝이 완성될 때 대응되는 신경망 모델이다.

[0903] 프로세서(301)는 제1 해상도의 원본 이미지를 획득하되, 상기 원본 이미지는 상기 압축 트레이닝 이미지 집합에서의 임의의 한 트레이닝 이미지이고 상기 원본 이미지의 태그 정보를 타겟 태그 정보로 사용하며 상기 타겟 모델에 의해 상기 원본 이미지를 압축하여 제2 해상도의 압축 이미지를 얻되, 상기 제2 해상도는 상기 제1 해상도보다 작고 상기 인식 신경망 모델에 의해 상기 압축 이미지를 인식하여 참고 태그 정보를 획득하며 상기 타겟 태그 정보와 상기 참고 태그 정보에 근거하여 손실함수를 획득하고 상기 손실함수가 상기 제1 임계값에 수렴되거나 또는 상기 트레이닝 횟수가 상기 제2 임계값보다 크거나 같을 경우 상기 제1 해상도의 타겟 원본 이미지를 획득하여 상기 타겟 모델을 상기 압축 신경망 모델로 확인하며 상기 압축 신경망 모델에 의해 상기 타겟 원본 이미지를 압축하여 상기 제2 해상도의 타겟 압축 이미지를 얻는다.

[0904] 선택적으로, 프로세서(301)는 또 상기 손실함수가 상기 제1 임계값에 수렴되지 않거나 또는 상기 트레이닝 횟수가 상기 제2 임계값보다 작을 경우 상기 손실함수에 근거하여 상기 타겟 모델을 업데이트하여 업데이트 모델을 얻고 상기 업데이트 모델을 상기 타겟 모델로 사용하며 그 다음의 한 트레이닝 이미지를 상기 원본 이미지로 사용하여 제1 해상도의 원본 이미지를 획득하는 상기 단계를 수행한다.

[0905] 선택적으로, 프로세서(301)는 구체적으로, 상기 압축 이미지를 전처리하여 인식 대기 이미지를 얻고 상기 인식 신경망 모델에 의해 상기 인식 대기 이미지를 인식하여 상기 참고 태그 정보를 얻는다.

[0906] 선택적으로, 상기 전처리는 사이즈 처리를 포함하고 메모리(302)는 또 상기 인식 신경망의 기본 이미지 크기를 저장하며 프로세서(301)는 구체적으로 상기 압축 이미지의 이미지 크기가 상기 기본 이미지 크기보다 작을 경우 상기 기본 이미지 크기에 따라 상기 압축 이미지에 대해 픽셀 포인트 충진을 진행하여 상기 인식 대기 이미지를 얻는다.

[0907] 선택적으로, 상기 압축 트레이닝 이미지 집합은 적어도 인식 트레이닝 이미지 집합을 포함하고 프로세서(301)는 또 상기 인식 트레이닝 이미지 집합을 사용하여 상기 인식 신경망을 트레이닝하여 상기 인식 신경망 모델을 얻되, 상기 인식 트레이닝 이미지 집합에서의 매 하나의 트레이닝 이미지는 적어도 상기 타겟 태그 정보의 유형과 일치한 태그 정보를 포함한다.

[0908] 선택적으로, 프로세서(301)는 또 상기 인식 신경망 모델에 의해 상기 타겟 압축 이미지를 인식하여 상기 타겟 원본 이미지의 태그 정보를 얻고;

[0909] 메모리(302)는 또 상기 타겟 원본 이미지의 태그 정보를 저장한다.

[0910] 선택적으로, 상기 압축 트레이닝 이미지 집합은 다수의 차원을 포함하고 프로세서(301)는 구체적으로, 상기 타겟 모델에 의해 상기 원본 이미지를 인식하여 다수의 이미지 정보를 얻고 매 하나의 차원은 하나의 이미지 정보와 대응하고 상기 타겟 모델과 상기 다수의 이미지 정보에 의해 상기 원본 이미지를 압축하여 상기 압축 이미지를 얻는다.

[0911] 이해할 수 있는 것은, 타겟 모델에 기반하여 원본 이미지의 압축 이미지를 획득하고 인식 신경망 모델에 기반하여 압추 이미지의 참고 태그 정보를 획득하며 원본 이미지가 포함하는 타겟 태그 정보와 참고 태그 정보에 근거하여 손실함수를 획득하고 손실함수가 제1 임계값에 수렴되거나 또는 압축 신경망의 현재의 트레이닝 횟수가 제2 임계값보다 크거나 같을 경우 즉시 이미지 압축을 위한 압축 신경망의 트레이닝을 완성하며 타겟 모델을 압축 신경망 모델로 하고 압축 신경망 모델에 기반하여 타겟 원본 이미지의 타겟 압축 이미지를 획득할 수 있다. 바꾸어 말하면 이미 트레이닝을 완성하여 얻은 인식 신경망 모델이 획득한 참고 태그값과 원본 이미지가 포함하는 타겟 태그값을 통해 손실함수를 획득하고 손실함수가 기설정 조건을 만족하거나 또는 압축 신경망의 현재의 트레이닝 횟수가 기설정 임계값을 초과할 경우 트레이닝을 완성하며 그렇지 않을 경우 압축 신경망을 트레이닝하는 것을 통해 그 가중치를 반복적으로 조절, 즉 동일한 이미지에서 매 하나의 픽셀에 대해 나타내는 이미지 콘텐츠를 조절하여 압축 신경망의 손실을 절감하고 이미지 압축의 유효성을 향상함으로써 인식의 정확도를 향상시키는데 편리하다.

[0912] 본원 발명의 하나의 가능한 실시예에서는 전자장치(400)를 제공하는데 전자장치(400)는 이미지 압축 장치(300)를 포함하고 도 5k에 도시된 바와 같이 전자장치(400)는 프로세서(401), 메모리(402), 통신 인터페이스(403) 및 하나 또는 다수의 프로그램(404)을 포함하며 여기서 하나 또는 다수의 프로그램(404)은 메모리(402)에 저장되고 프로세서(401)에 의해 수행되도록 배치되며 프로그램(404)은 상기 이미지 압축 방법을 수행하기 위한 설명의 일

부 또는 모든 단계의 명령을 포함한다.

[0913] 설명해야 할 것은 상기 각 유닛 또는 모듈은 모두 회로 일 수 있고 디지털 회로, 아날로그 회로 등을 포함한다. 상기 각 유닛 또는 모듈 구조의 물리적 실현은 물리소자를 포함하나 이에 한정되지 않고 물리소자는 트랜지스터, 멤리스터 등을 포함하나 이에 한정되지 않는다. 상기 칩 또는 상기 신경망 프로세서는 임의의 적당한 하드웨어 프로세서, 예하면 CPU, GPU, FPGA, DSP와 ASIC 등 일 수 있다. 상기 저장 유닛은 임의의 정당한 자기 저장매체 또는 광자기 저장매체, 예하면 저항성 랜덤 액세스 메모리(Resistive Random Access Memory, RRAM), 동적 랜덤 액세스 메모리(Dynamic Random Access Memory, DRAM), 정적 랜덤 액세스 메모리(Static Random Access Memory, SRAM), 증강된 동적 랜덤 액세스 메모리(Enhanced Dynamic Random Access Memory, EDRAM), 고 대역폭 메모리(High Bandwidth Memory, HBM), 하이브리드 메모리 큐브(Hybrid Memory Cube, HMC) 등 일 수 있다.

[0914] 본원 발명은 수많은 범용 또는 전용 계산 시스템 환경 또는 배치에 사용될 수 있다. 예하면 개인 컴퓨터, 서버 컴퓨터, 핸드 헬드 기기 또는 휴대용 기기, 태블릿 기기, 다중 프로세서 시스템, 마이크로 프로세서에 기반한 시스템, 탑 핏, 프로그램 가능한 소비 전자기기, 네트워크 개인 컴퓨터(personal computer, PC), 소형 컴퓨터, 대형 컴퓨터, 이상의 임의의 시스템 또는 기기를 포함하는 분포식 계산환경 등이다.

[0915] 하나의 실시예에서 본원 발명은 칩을 제공하는데 이는 상술한 연산장치를 포함하고 상기 칩은 가중치와 입력 뉴런에 대해 동시에 여러 가지 연산을 진행할 수 있어 연산의 다원화를 실현할 수 있다. 이 외에 다층 인공 신경망 연산 알고리즘에 대해 사용한 전용 온-칩 캐시를 통해 입력 뉴런과 가중치 데이터의 중요성을 충분히 발굴하여 메모리에서 이러한 데이터를 반복적으로 판독하는 것을 방지하고 메모리 액세스 대역폭을 저하시켜 메모리 대역폭이 다층 인공 신경망 연산 및 그의 트레이닝 알고리즘에 가져온 기능적 난관의 문제를 방지하였다.

[0916] 본원 발명의 하나의 가능한 실시예에서 본 발명의 실시예는 칩 패키지 구조를 제공하는데 이는 상기 신경망 프로세서를 포함한다.

[0917] 본원 발명의 하나의 가능한 실시예에서 본 발명의 실시예는 보드 카드를 제공하는데 이는 상기 칩 패키지 구조를 포함한다.

[0918] 본원 발명의 하나의 가능한 실시예에서 본 발명의 실시예는 전자장치를 제공하는데 이는 상기 보드 카드를 포함한다.

[0919] 상기 전자장치는 로봇, 컴퓨터, 프린터, 스캐너, 태블릿 피씨, 스마트 단말기, 휴대폰, 블랙박스, 네비게이터, 센서, 캠, 클라우드 서버, 사진기, 카메라, 프로젝터, 손목시계, 이어폰, 모바일 메모리, 휴대용 기기, 교통수단, 가전제품, 의료기기를 포함하나 이에 한정되지 않는다.

[0920] 상기 교통수단은 비행기, 기선 및/또는 차량을 포함하고 상기 가전제품은 텔레비전, 에어컨, 전자레인지, 냉장고, 전기밥솥, 가습기, 세탁기, 전등, 가스레인지, 주방 환풍기를 포함하며 상기 의료기기는 MRI, 초음파기기 및/또는 일렉트로카르디오그래프를 포함한다.

[0921] 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자들은 본 명세서에서 출원하는 실시예가 설명하는 각 예시적인 유닛 및 알고리즘 단계와 결부하여 전자 하드웨어, 컴퓨터 소프트웨어 또는 양자의 결합으로 실현할 수 있고 하드웨어와 소프트웨어의 호환 가능성을 뚜렷이 설명하기 위하여 상기 설명에서는 이미 기능에 따라 각 예시적인 조성 및 단계를 통상적으로 설명하였음을 알 수 있다. 이러한 기능이 하드웨어 방식으로 실행될 것인지 아니면 소프트웨어 방식으로 실행될 것인지는 기술적 해결수단의 특정된 응용과 디자인의 제약조건에 의해 결정된다. 통상의 지식을 가진 자들은 매 하나의 특정된 응용에 상이한 방법을 사용하여 설명된 기능을 실현하나 이러한 실현은 본 발명의 범위를 벗어난 것으로 간주되지 말아야 한다.

[0922] 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자들은 설명의 편리와 간결함을 위하여 상기에서 설명한 단말기와 유닛의 구체적인 작동 과정은 상술한 방법 실시예에서의 대응되는 과정을 참조할 수 있으므로 여기서 더 이상 설명하지 않음을 뚜렷이 알고 있다.

[0923] 본원 발명이 제공하는 몇 개의 실시예에서 개시된 단말기와 방법은 기타 방식으로 실현될 수 있음을 이해할 수 있다. 예하면 이상에서 설명한 장치 실시예는 단지 예시적인 것으로, 예하면 상기 유닛의 구획은 단지 한가지 논리 기능의 구획으로서 실제 실현과정에서는 별도의 구획방식이 존재할 수 있는 바, 예하면 다수의 유닛 또는 어셈블리가 또 다른 시스템에 결합되거나 집적될 수 있거나 또는 일부 특징이 무시될 수 있거나 또는 수행되지 않을 수 있다. 이 외에 표시되거나 토론된 서로 간의 커플링 또는 직접 커플링 또는 통신 연결은 일부 인터페이

스, 장치 또는 유닛의 간접적 커플링 또는 통신 연결일 수도 있고 전기적, 기계적 또는 기타 형식의 연결일 수도 있다.

[0924] 상기 분리부재로 설명된 유닛은 물리적으로 분리된 것이거나 아닐 수 있고 유닛으로 표시된 부품은 물리적 유닛이거나 아닐 수 있는 바, 즉 한 곳에 위치될 수도 있거나 다수의 네트워크에 분포될 수도 있다. 실제 수요에 근거하여 그 중의 일부 또는 전부 유닛을 선택하여 본 발명의 실시예의 방안의 목적을 실현할 수 있다.

[0925] 이 외에 본 발명의 각 실시예의 각 기능유닛은 하나의 처리 유닛에 집적될 수도 있고 각 유닛이 단독으로 물리적으로 존재할 수도 있고 두 개 또는 두 개 이상의 유닛이 하나의 유닛에 집적될 수도 있다. 상기 집적된 유닛은 하드웨어 형식을 사용하여 실현할 수 있을 뿐만 아니라 소프트웨어 기능유닛의 형식을 사용하여 실현할 수도 있다.

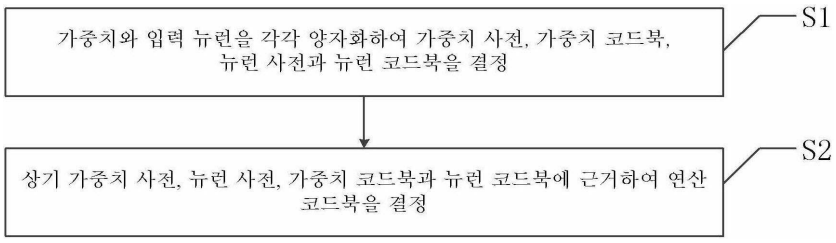
[0926] 상기 집적된 유닛이 만약 소프트웨어 기능유닛의 형식으로 실현됨과 동시에 독립된 제품으로 판매 또는 사용될 경우 하나의 컴퓨터 판독 가능 저장매체에 저장될 수 있다. 이러한 이유에 기반하여 본 발명의 기술적 해결수단은 실질적으로 또는 선행기술에 대하여 기여하는 부분 또는 상기 기술적 해결수단의 전부 또는 일부가 소프트웨어 제품의 형식으로 구현되고 상기 컴퓨터 소프트웨어 제품은 하나의 저장매체에 저장되며 약간의 명령을 포함하여 한 대의 컴퓨터 기기(개인 컴퓨터, 서버 또는 네트워크 기기 등 일 수 있음)가 본 발명의 각 실시예에서 설명하는 방법의 전부 또는 일부 단계를 수행하도록 한다. 상술한 저장매체는 USB메모리, 모바일 하드디스크, 롬(Read-Only Memory, ROM), 램(Random Access Memory, RAM), 자기 디스크 또는 시디롬 등 여러 가지 프로그램 코드를 저장할 수 있는 메체를 포함한다.

[0927] 설명해야 할 것은 도면 또는 전반 명세서에서 도시되지 않았거나 설명되지 않은 실현방식은 모두 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자들이 알고 있는 형식이므로 상세히 설명하지 않는다. 이 외에 상기 각 소자와 방법에 대한 정의는 실시예에서 제출한 여러 가지 구체적인 구조, 모양 또는 방식에 한정되지 않는 바, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자들은 이에 대해 간단히 변경하거나 대체할 수 있다.

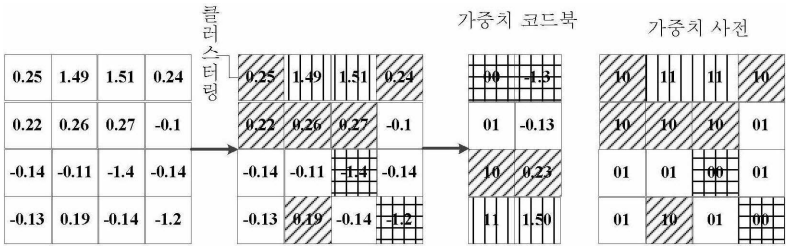
[0928] 이상에서 설명한 구체적인 실시예는 본원 발명의 목적, 기술적 해결수단과 유리한 효과를 진일보로 상세히 설명하였는데 이상에서의 설명은 단지 본원 발명의 구체적인 실시예일 뿐 본원 발명을 한정하기 위한 것이 아니고 본원 발명의 정신과 원칙 내에서 진행한 그 어떤 수정, 등가적인 대체, 개선은 모두 본원 발명의 보호범위 내에 포함되는 것으로 이해해야 한다.

도면

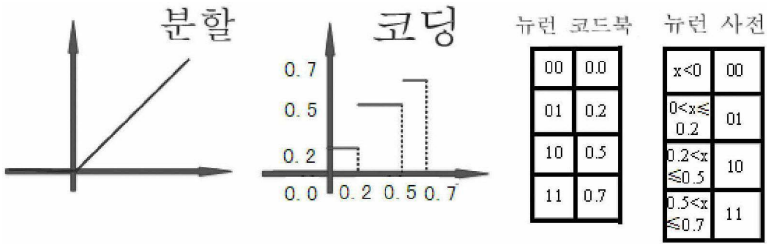
도면1a



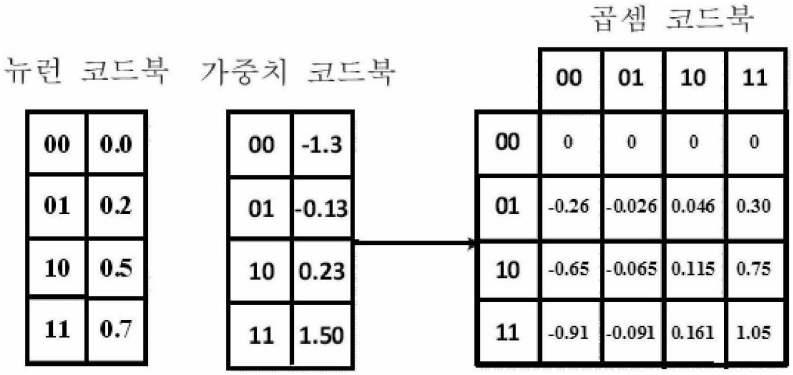
도면1b



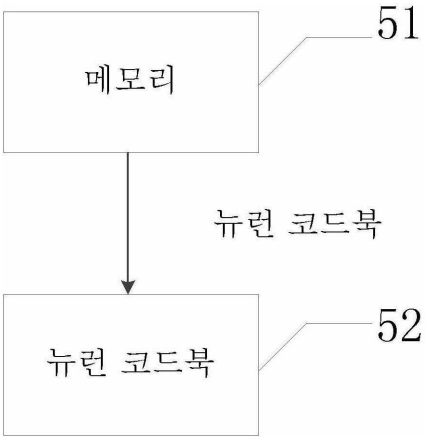
도면1c



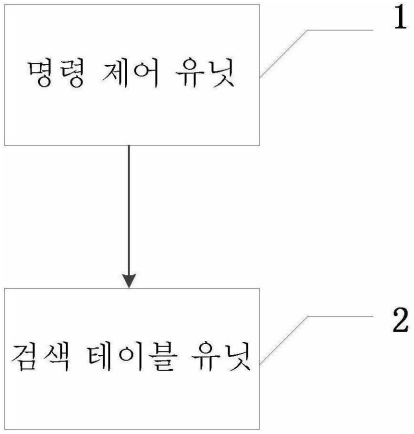
도면1d



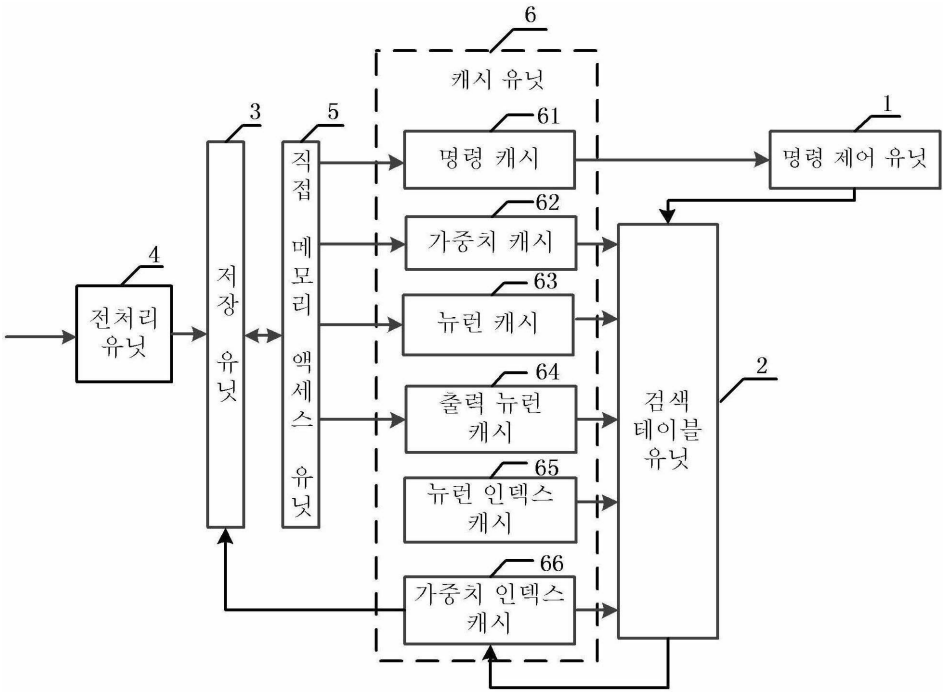
도면1e



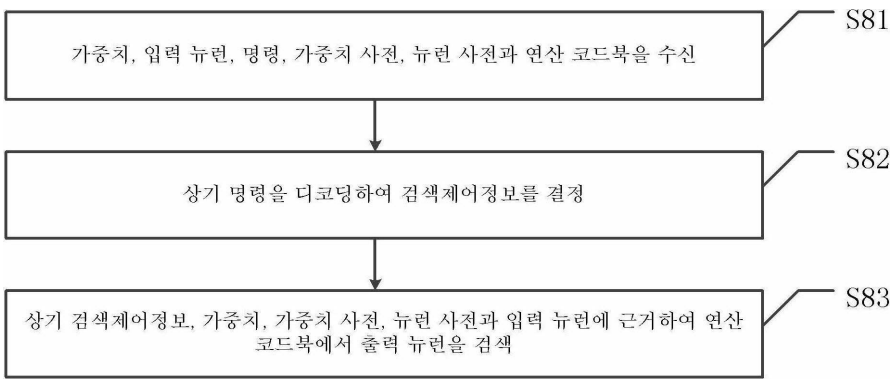
도면1f



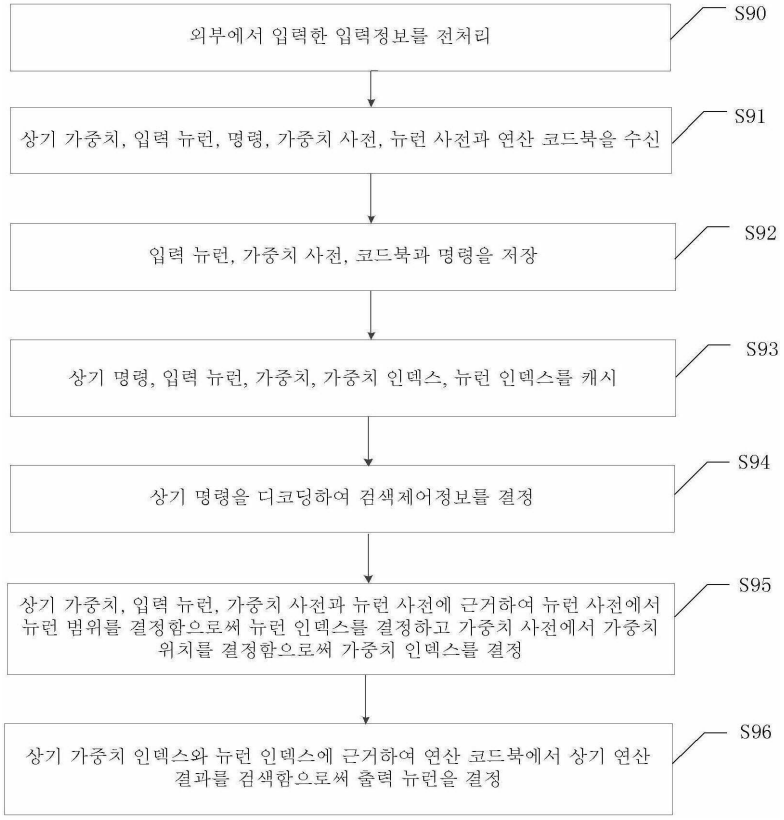
도면1g


6
캐시 유닛
61
명령 캐시
62
가중치 캐시
63
뉴런 캐시
64
출력 뉴런 캐시
65
뉴런 인덱스 캐시
66
가중치 인덱스 캐시
1
명령 제어 유닛
2
검색 테이블 유닛
3
저장 유닛
4
전처리 유닛
5
직접 메모리 액세스 유닛

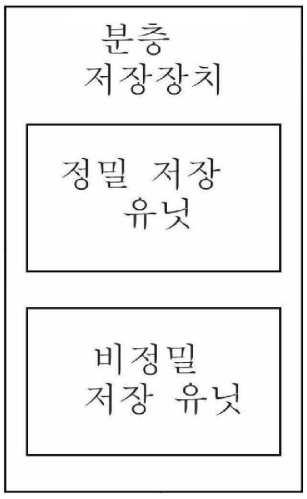
도면1h


S81
가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 수신
S82
상기 명령을 디코딩하여 검색제어정보를 결정
S83
상기 검색제어정보, 가중치, 가중치 사전, 뉴런 사전과 입력 뉴런에 근거하여 연산 코드북에서 출력 뉴런을 검색

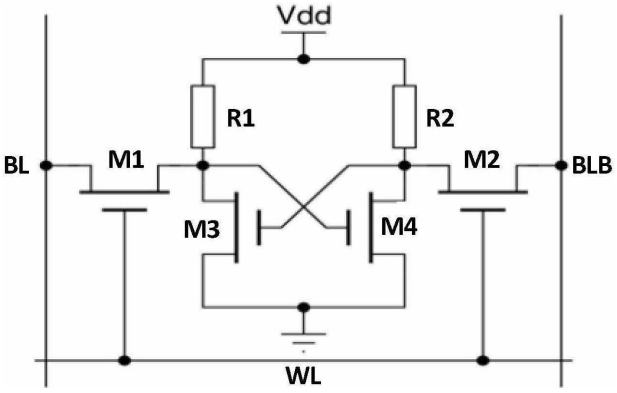
도면1i


외부에서 입력한 입력정보를 전처리
S90
상기 가중치, 입력 뉴런, 명령, 가중치 사전, 뉴런 사전과 연산 코드북을 수신
S91
입력 뉴런, 가중치 사전, 코드북과 명령을 저장
S92
상기 명령, 입력 뉴런, 가중치, 가중치 인덱스, 뉴런 인덱스를 캐시
S93
상기 명령을 디코딩하여 검색제어정보를 결정
S94
상기 가중치, 입력 뉴런, 가중치 사전과 뉴런 사전에 근거하여 뉴런 사전에서 뉴런 범위를 결정함으로써 뉴런 인덱스를 결정하고 가중치 사전에서 가중치 위치를 결정함으로써 가중치 인덱스를 결정
S95
상기 가중치 인덱스와 뉴런 인덱스에 근거하여 연산 코드북에서 상기 연산 결과를 검색함으로써 출력 뉴런을 결정
S96

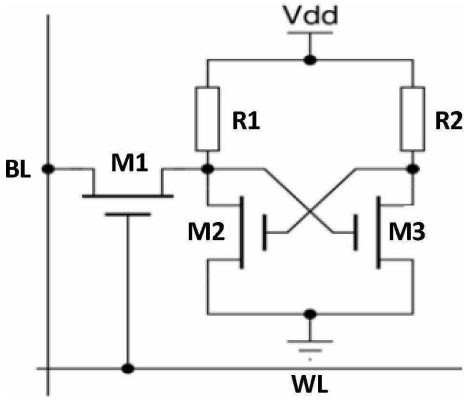
도면2a


분층
저장장치
정밀 저장
유닛
비정밀
저장 유닛

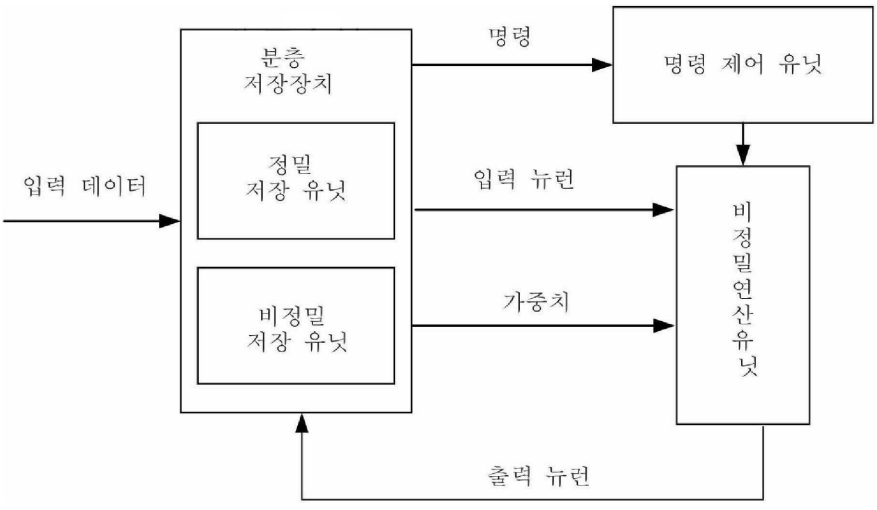
도면2b



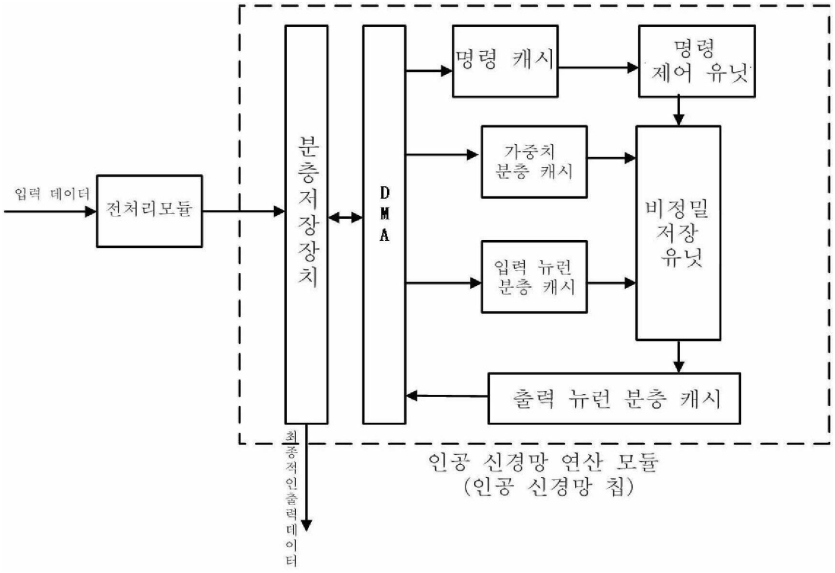
도면2c



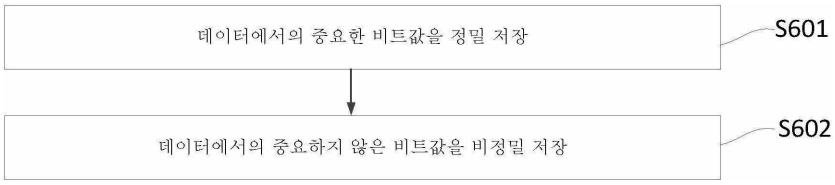
도면2d



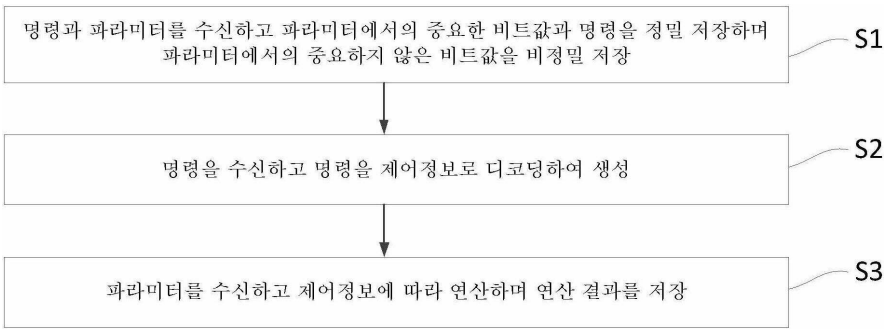
도면2e



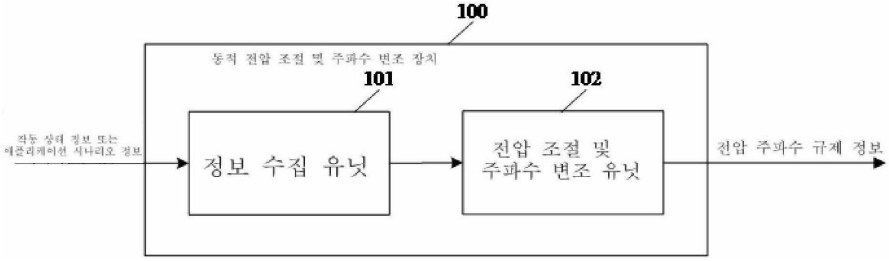
도면2f



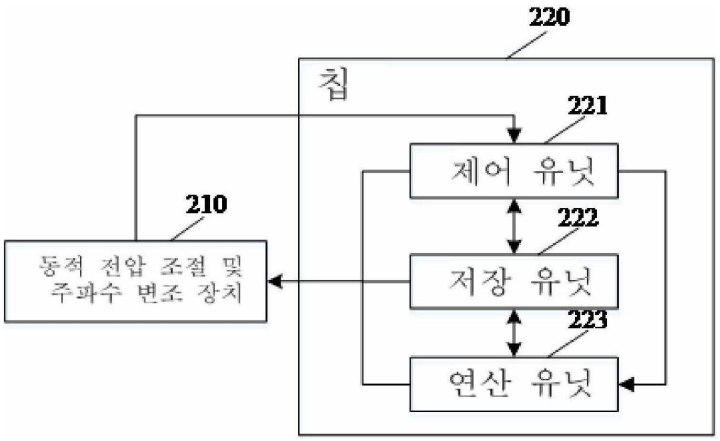
도면2g



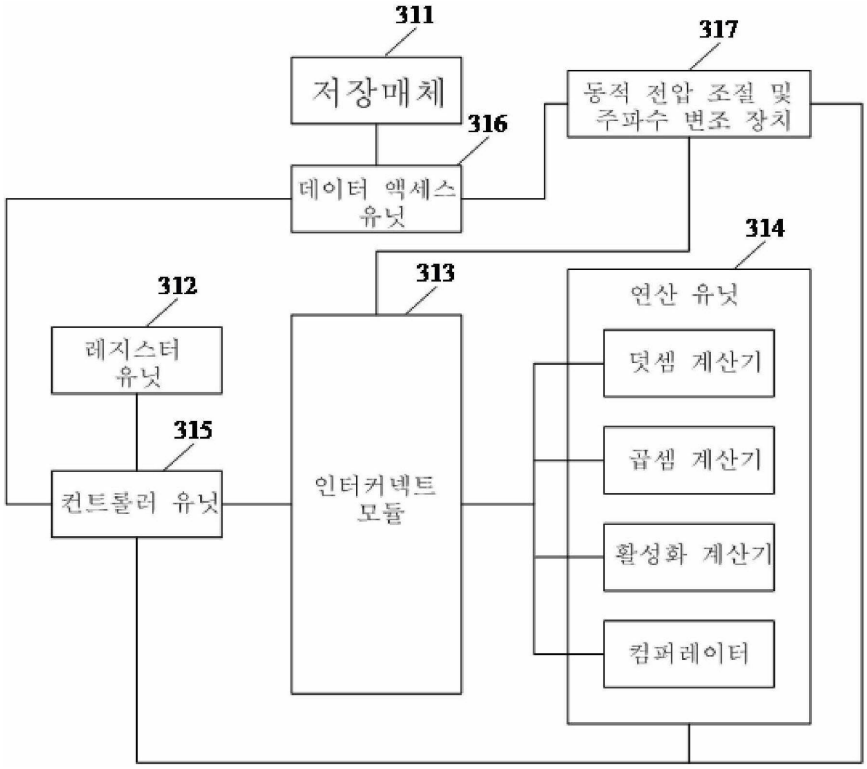
도면3a



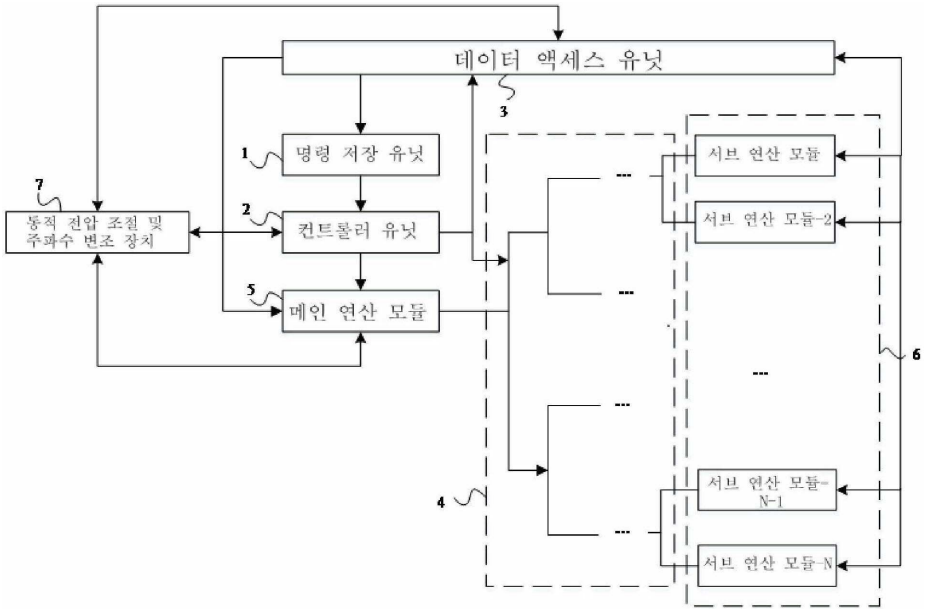
도면3b


220
칩
221
제어 유닛
222
저장 유닛
223
연산 유닛
210
동적 전압 조절 및
주파수 변조 장치

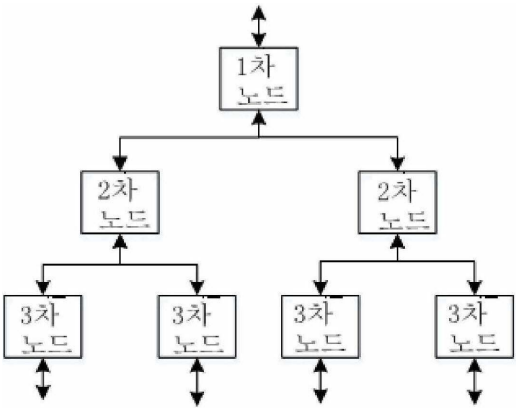
도면3c


311
저장매체
317
동적 전압 조절 및
주파수 변조 장치
316
데이터 액세스
유닛
314
연산 유닛
313
312
레지스터
유닛
덧셈 계산기
315
곱셈 계산기
인터커넥트
모듈
컨트롤러 유닛
활성화 계산기
컴퍼레이터

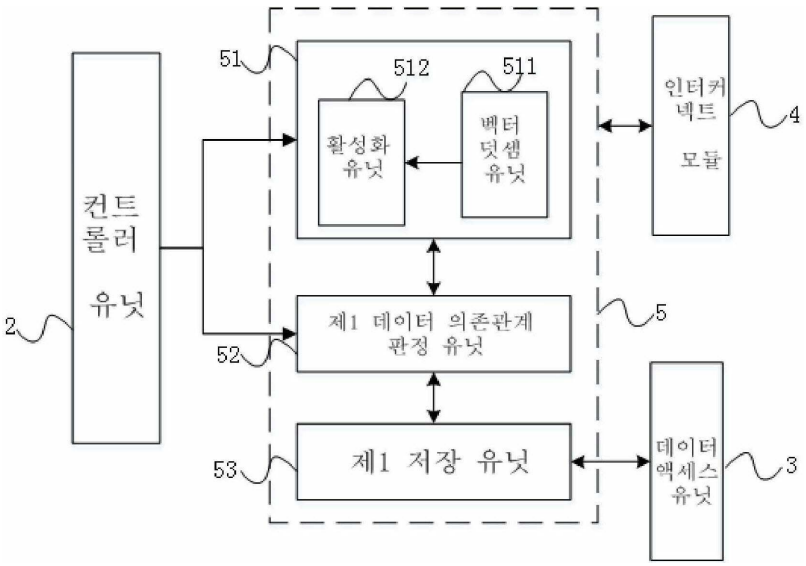
도면3d



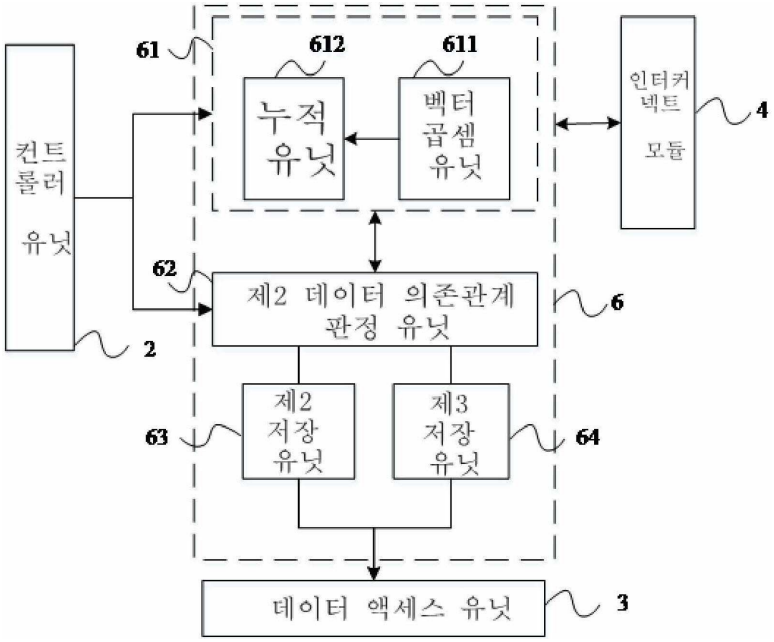
도면3e



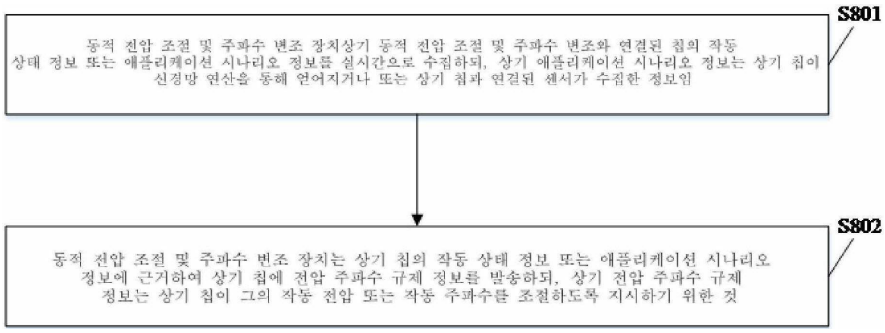
도면3f



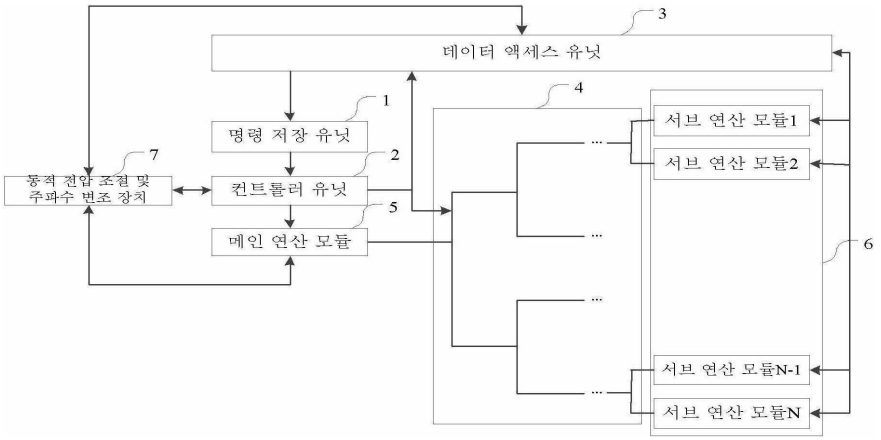
도면3g



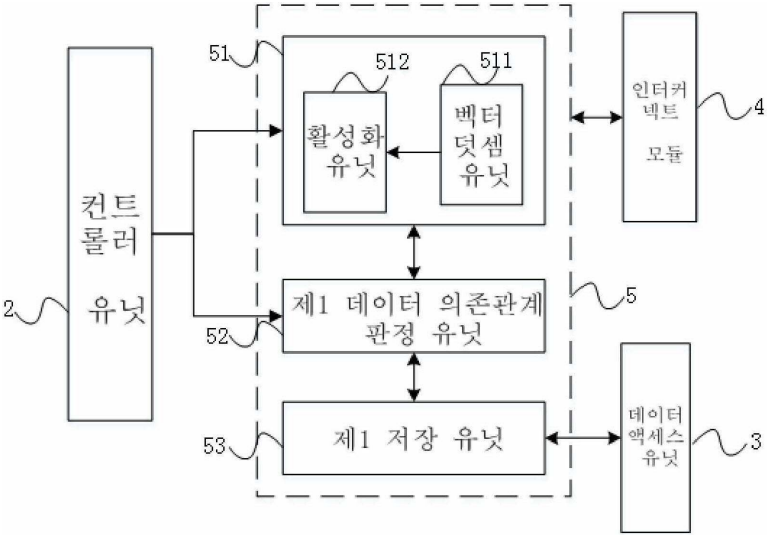
도면3h



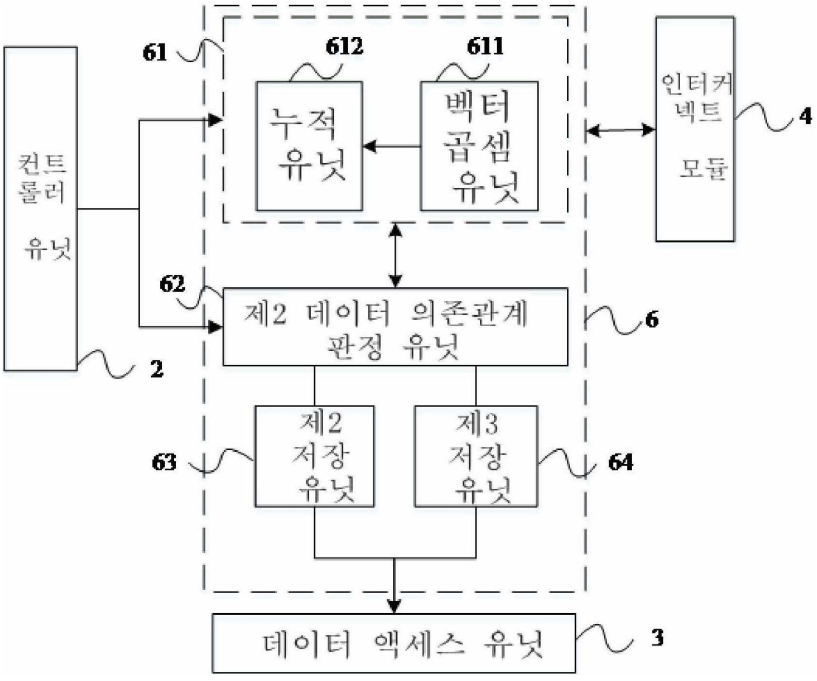
도면4a



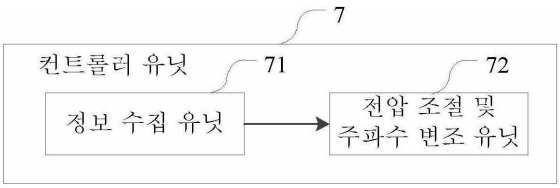
도면4b



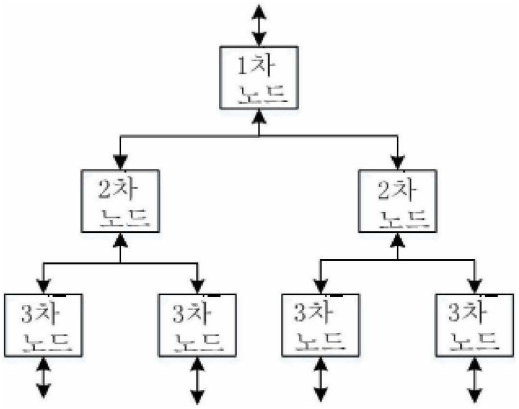
도면4c



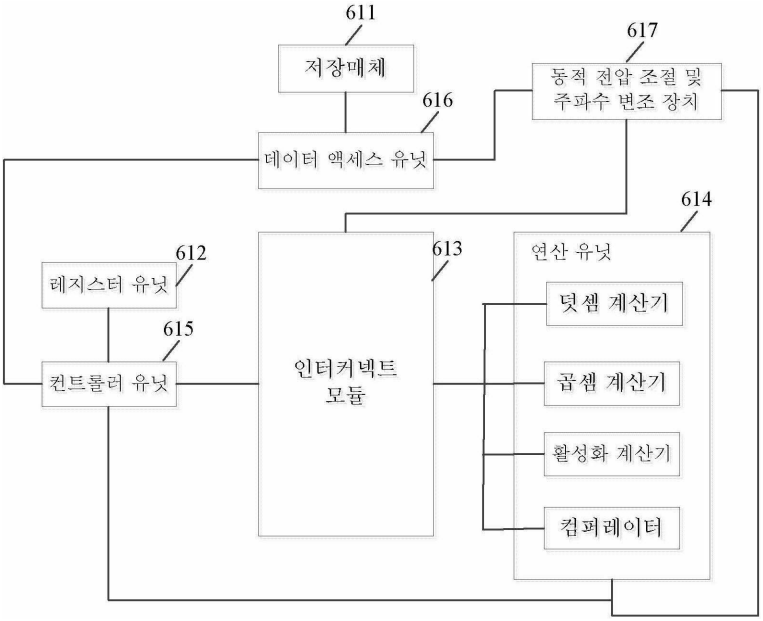
도면4d



도면4e


1차
노드
2차
노드
2차
노드
3차
노드
3차
노드
3차
노드
3차
노드

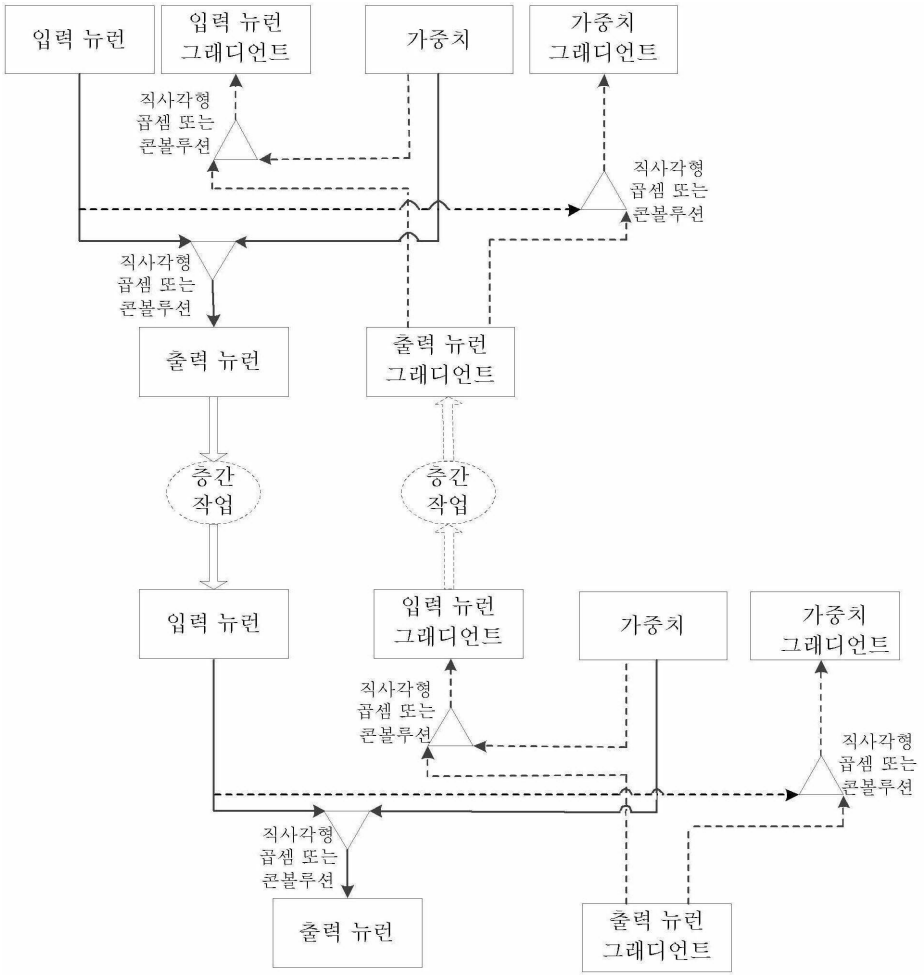
도면4f


611
저장매체
617
동적 전압 조절 및
주파수 변조 장치
616
데이터 액세스 유닛
614
612
레지스터 유닛
613
연산 유닛
615
컨트롤러 유닛
인터커넥트
모듈
덧셈 계산기
곱셈 계산기
활성화 계산기
컴퍼레이터

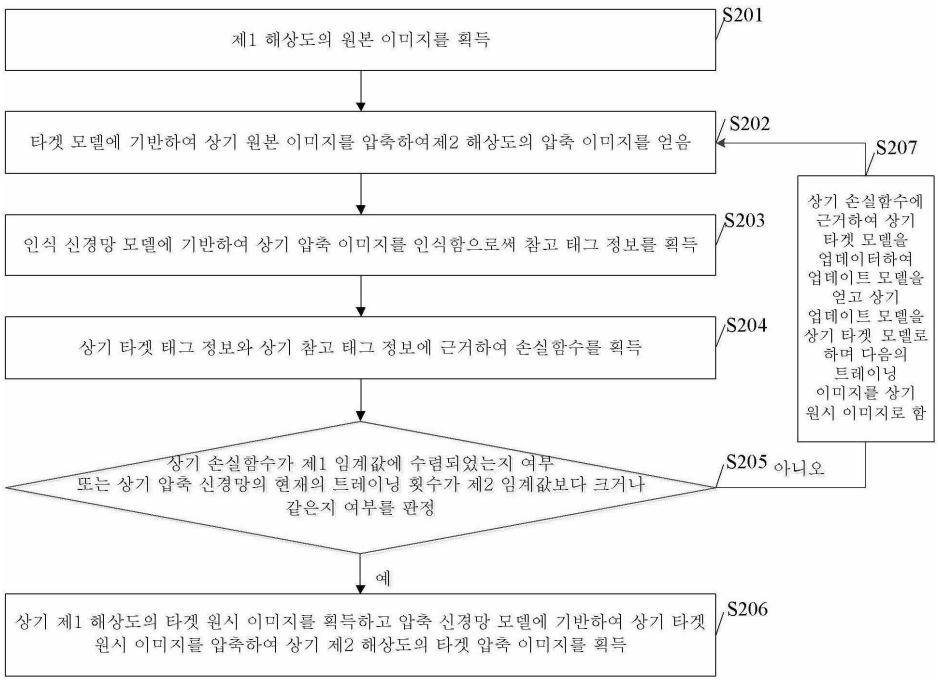
도면4g


S701
명령 저장 유닛의 첫번째 주소에 하나의 입출력 IO명령을 미리 저장
S702
연산이 시작되면 컨트롤러 유닛은 상기 명령 저장 유닛의 첫번째 주소로부터 상기 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 데이터 액세스 유닛은 외부 주소 공간으로부터 대응되는 모든 콘볼루션 신경망 연산 명령을 판독하여 이를 상기 명령 저장 유닛에 캐시
S703
상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 외부 주소 공간으로부터 메인 연산 모듈에 필요한 모든 데이터를 상기 메인 연산 모듈의 제1 저장 유닛에 발송
S704
상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 외부 주소 공간으로부터 서브 연산 모듈에 필요한 콘볼루션 커널 데이터를 판독
S705
상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 CONFIG 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 콘볼루션 연산장치는 상기 층의 신경망 계산에 필요한 여러 가지 상수를 배치
S706
상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 COMPUTE 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 메인 연산 모듈은 먼저 인터커넥트 모듈을 통해 콘볼루션 윈도우 내의 입력 데이터를 N개의 서브 연산 모듈에 발송하여 상기 N개의 서브 연산 모듈의 제2 저장 유닛에 저장한 후 명령에 따라 콘볼루션 윈도우를 이동
S707
COMPUTE 명령에 의해 디코딩된 컨트롤 신호에 근거하여 상기 N개의 서브 연산 모듈의 연산 유닛은 제3 저장 유닛으로부터 콘볼루션 커널을 판독하고 상기 제2 저장 유닛으로부터 입력 데이터를 판독하며 입력 데이터와 콘볼루션 커널의 콘볼루션 연산을 완성하고 획득한 출력 스칼라를 상기 인터커넥트 모듈을 통해 리턴
S708
상기 인터커넥트 모듈에서 상기 N개의 서브 연산 모듈에 의해 리턴된 출력 스칼라는 완전한 중간 벡터로 단계적으로 연결
S709
상기 메인 연산 모듈은 인터커넥트 모듈에 의해 리턴된 중간 벡터를 획득하고 콘볼루션 윈도우는 모든 입력 데이터를 가로 지르며 상기 메인 연산 모듈은 모든 리턴된 벡터를 중간 결과로 연결하고 COMPUTE 명령에 의해 디코딩된 컨트롤 신호에 근거하여 제1 저장 유닛으로부터 바이어스 데이터를 판독하며 벡터 가산 유닛에 의해 중간 결과와 가산되어 바이어스 결과를 획득한 후 활성화 유닛에 의해 바이어스 결과를 활성화시키고 최종 출력 데이터를 상기 제1 저장 유닛에 다시 기입ㄹ
S710
상기 컨트롤러 유닛은 계속하여 상기 명령 저장 유닛으로부터 다음의 한 IO 명령을 판독하고 디코딩된 컨트롤 신호에 근거하여 상기 데이터 액세스 유닛은 상기 제1 저장 유닛의 출력 데이터를 외부 주소 공간의 지정된 주소에 저장하여 연산을 종료

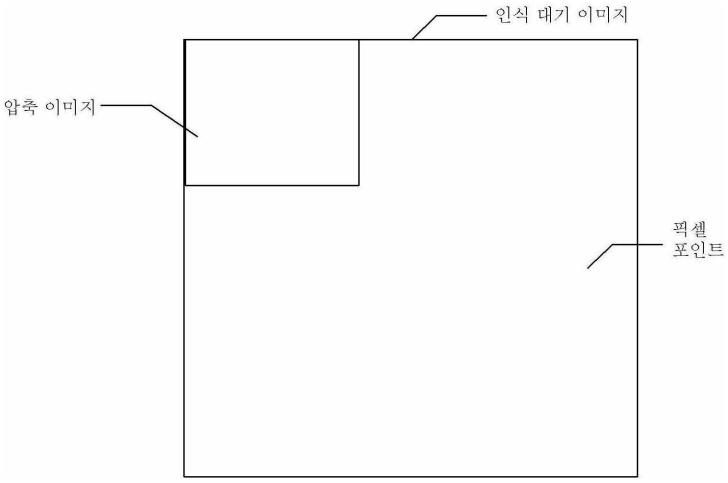
도면5a


입력 뉴런
입력 뉴런 그래디언트
가중치
가중치 그래디언트
직사각형 곱셈 또는 콘볼루션
직사각형 곱셈 또는 콘볼루션
직사각형 곱셈 또는 콘볼루션
출력 뉴런
출력 뉴런 그래디언트
층간 작업
층간 작업
입력 뉴런
입력 뉴런 그래디언트
가중치
가중치 그래디언트
직사각형 곱셈 또는 콘볼루션
직사각형 곱셈 또는 콘볼루션
직사각형 곱셈 또는 콘볼루션
출력 뉴런
출력 뉴런 그래디언트

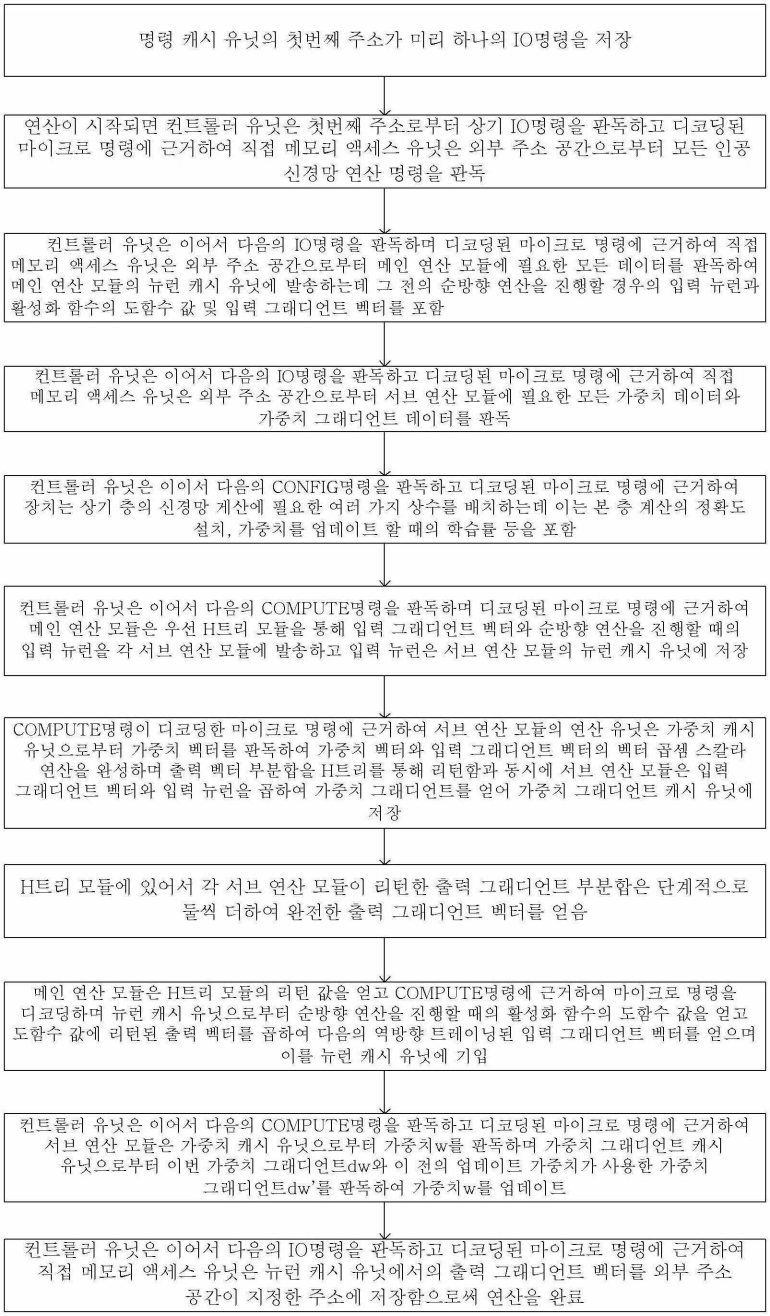
도면5b


제1 해상도의 원본 이미지를 획득
S201
타겟 모델에 기반하여 상기 원본 이미지를 압축하여제2 해상도의 압축 이미지를 얻음
S202
인식 신경망 모델에 기반하여 상기 압축 이미지를 인식함으로써 참고 태그 정보를 획득
S203
상기 타겟 태그 정보와 상기 참고 태그 정보에 근거하여 손실함수를 획득
S204
상기 손실함수가 제1 임계값에 수렴되었는지 여부 또는 상기 압축 신경망의 현재의 트레이닝 횟수가 제2 임계값보다 크거나 같은지 여부를 판정
S205
아니오
예
상기 제1 해상도의 타겟 원시 이미지를 획득하고 압축 신경망 모델에 기반하여 상기 타겟 원시 이미지를 압축하여 상기 제2 해상도의 타겟 압축 이미지를 획득
S206
S207
상기 손실함수에 근거하여 상기 타겟 모델을 업데이터하여 업데이트 모델을 얻고 상기 업데이트 모델을 상기 타겟 모델로 하며 다음의 트레이닝 이미지를 상기 원시 이미지로 함

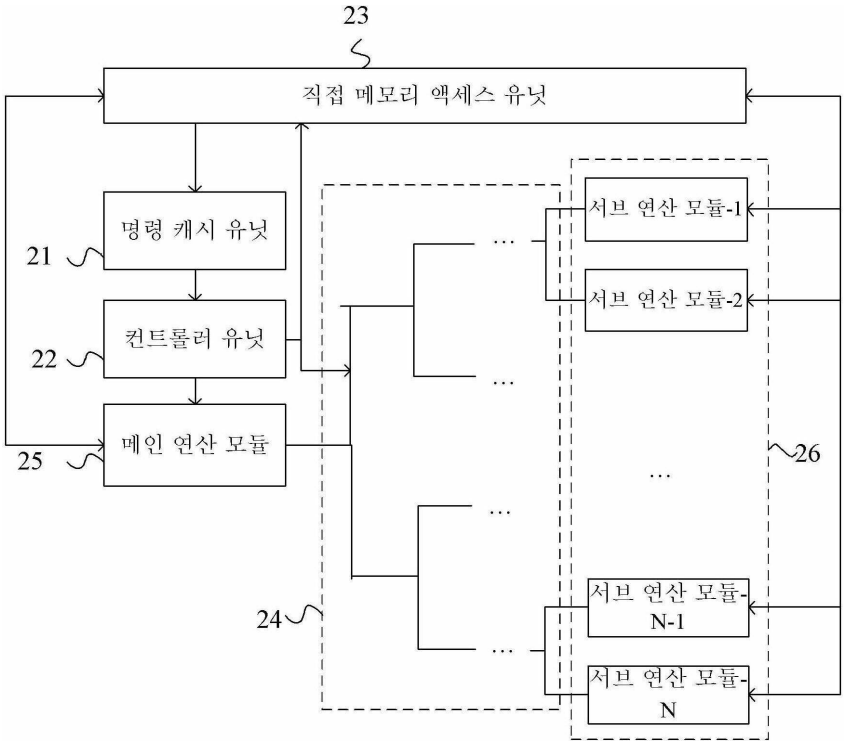
도면5c



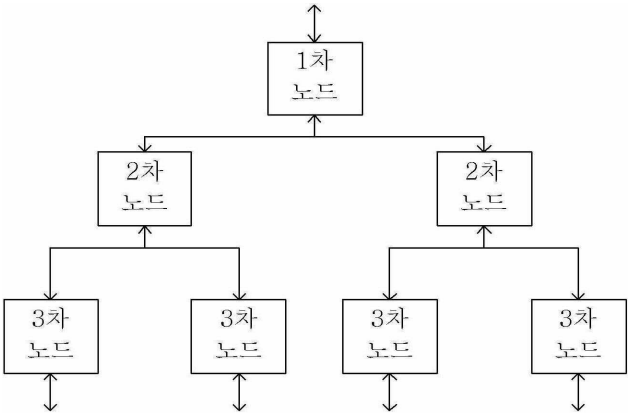
도면5d


명령 캐시 유닛의 첫번째 주소가 미리 하나의 IO명령을 저장
연산이 시작되면 컨트롤러 유닛은 첫번째 주소로부터 상기 IO명령을 판독하고 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 외부 주소 공간으로부터 모든 인공 신경망 연산 명령을 판독
컨트롤러 유닛은 이어서 다음의 IO명령을 판독하며 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 외부 주소 공간으로부터 메인 연산 모듈에 필요한 모든 데이터를 판독하여 메인 연산 모듈의 뉴런 캐시 유닛에 발송하는데 그 전의 순방향 연산을 진행할 경우의 입력 뉴런과 활성화 함수의 도함수 값 및 입력 그래디언트 벡터를 포함
컨트롤러 유닛은 이어서 다음의 IO명령을 판독하고 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 외부 주소 공간으로부터 서브 연산 모듈에 필요한 모든 가중치 데이터와 가중치 그래디언트 데이터를 판독
컨트롤러 유닛은 이어서 다음의 CONFIG명령을 판독하고 디코딩된 마이크로 명령에 근거하여 장치는 상기 층의 신경망 계산에 필요한 여러 가지 상수를 배치하는데 이는 본 층 계산의 정확도 설치, 가중치를 업데이트 할 때의 학습률 등을 포함
컨트롤러 유닛은 이어서 다음의 COMPUTE명령을 판독하며 디코딩된 마이크로 명령에 근거하여 메인 연산 모듈은 우선 H트리 모듈을 통해 입력 그래디언트 벡터와 순방향 연산을 진행할 때의 입력 뉴런을 각 서브 연산 모듈에 발송하고 입력 뉴런은 서브 연산 모듈의 뉴런 캐시 유닛에 저장
COMPUTE명령이 디코딩한 마이크로 명령에 근거하여 서브 연산 모듈의 연산 유닛은 가중치 캐시 유닛으로부터 가중치 벡터를 판독하여 가중치 벡터와 입력 그래디언트 벡터의 벡터 곱셈 스칼라 연산을 완성하며 출력 벡터 부분합을 H트리를 통해 리턴함과 동시에 서브 연산 모듈은 입력 그래디언트 벡터와 입력 뉴런을 곱하여 가중치 그래디언트를 얻어 가중치 그래디언트 캐시 유닛에 저장
H트리 모듈에 있어서 각 서브 연산 모듈이 리턴한 출력 그래디언트 부분합은 단계적으로 둘씩 더하여 완전한 출력 그래디언트 벡터를 얻음
메인 연산 모듈은 H트리 모듈의 리턴 값을 얻고 COMPUTE명령에 근거하여 마이크로 명령을 디코딩하며 뉴런 캐시 유닛으로부터 순방향 연산을 진행할 때의 활성화 함수의 도함수 값을 얻고 도함수 값에 리턴된 출력 벡터를 곱하여 다음의 역방향 트레이닝된 입력 그래디언트 벡터를 얻으며 이를 뉴런 캐시 유닛에 기입
컨트롤러 유닛은 이어서 다음의 COMPUTE명령을 판독하고 디코딩된 마이크로 명령에 근거하여 서브 연산 모듈은 가중치 캐시 유닛으로부터 가중치w를 판독하며 가중치 그래디언트 캐시 유닛으로부터 이번 가중치 그래디언트dw와 이 전의 업데이트 가중치가 사용한 가중치 그래디언트dw'를 판독하여 가중치w를 업데이트
컨트롤러 유닛은 이어서 다음의 IO명령을 판독하고 디코딩된 마이크로 명령에 근거하여 직접 메모리 액세스 유닛은 뉴런 캐시 유닛에서의 출력 그래디언트 벡터를 외부 주소 공간이 지정한 주소에 저장함으로써 연산을 완료

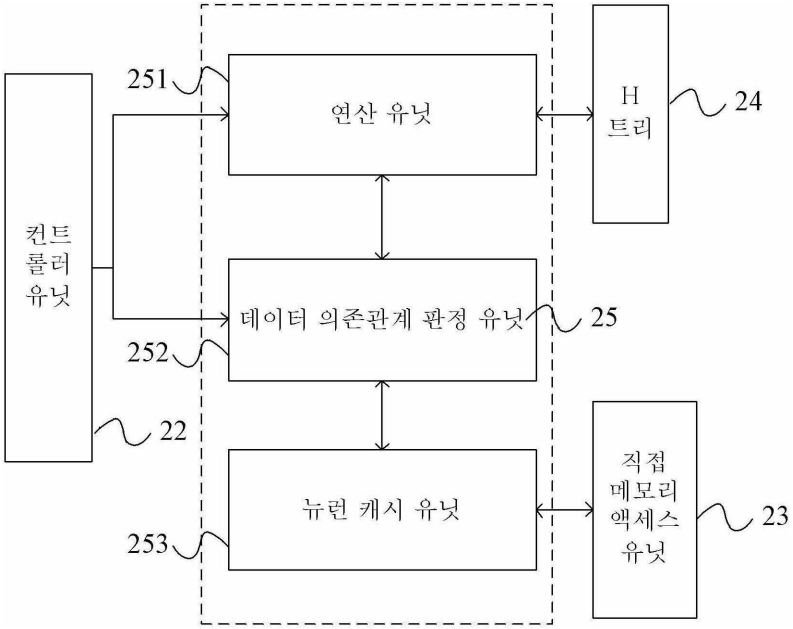
도면5e


23
직접 메모리 액세스 유닛
명령 캐시 유닛
21
컨트롤러 유닛
22
메인 연산 모듈
25
서브 연산 모듈-1
서브 연산 모듈-2
서브 연산 모듈-N-1
서브 연산 모듈-N
24
26

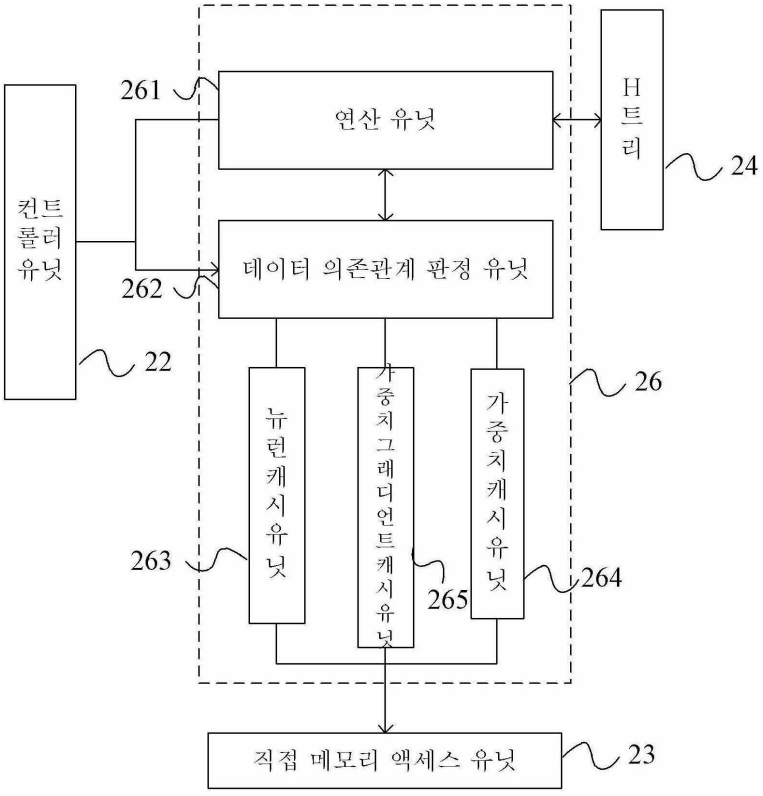
도면5f


1차 노드
2차 노드
2차 노드
3차 노드
3차 노드
3차 노드
3차 노드

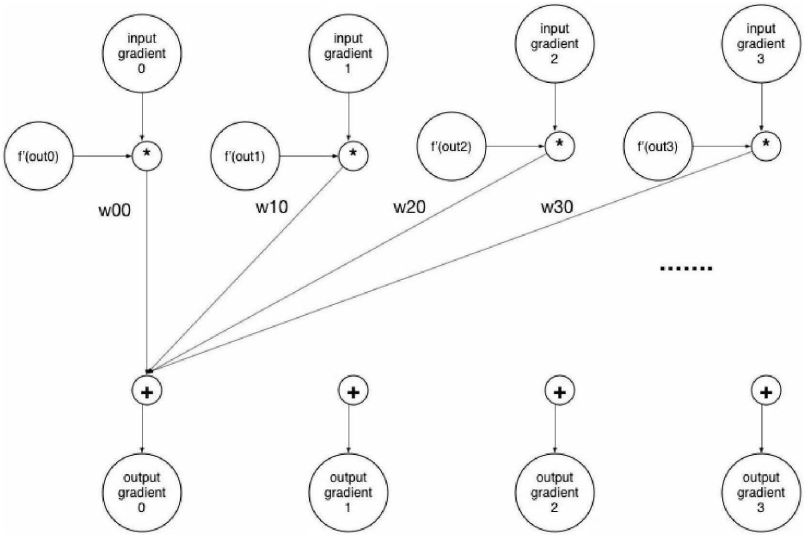
도면5g


251
연산 유닛
H 트리
24
컨트롤러 유닛
데이터 의존관계 판정 유닛
25
252
22
뉴런 캐시 유닛
직접 메모리 액세스 유닛
253
23

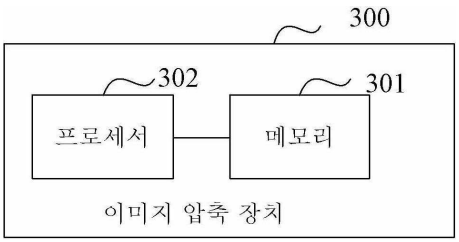
도면5h


261
연산 유닛
H 트리
24
컨트롤러 유닛
데이터 의존관계 판정 유닛
262
22
26
뉴런캐시유닛
가중치그래디언트캐시유닛
가중치캐시유닛
263
265
264
직접 메모리 액세스 유닛
23

도면5i



도면5j



도면5k