(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년01월03일
(11) 등록번호 10-2346636
(24) 등록일자 2021년12월29일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06F 9/48* (2018.01)
G06F 9/50 (2018.01) *G06N 3/04* (2006.01)
G06N 3/063 (2006.01) *G06N 3/10* (2006.01)
G06Q 10/06 (2012.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06F 9/4881 (2013.01)
(21) 출원번호 10-2019-7027653
(22) 출원일자(국제) 2018년01월17일
심사청구일자 2019년09월20일
(85) 번역문제출일자 2019년09월20일
(65) 공개번호 10-2019-0118635
(43) 공개일자 2019년10월18일
(86) 국제출원번호 PCT/US2018/013939
(87) 국제공개번호 WO 2018/212799
국제공개일자 2018년11월22일
(30) 우선권주장
15/599,559 2017년05월19일 미국(US)
(56) 선행기술조사문헌
Jeffrey Dean et al., Large Scale Distributed Deep Networks, NIPS 1-9pages(2012)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
구글 엘엘씨
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
(72) 발명자
우 동혁
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
(74) 대리인
박장원

전체 청구항 수 : 총 12 항 심사관 : 양대경

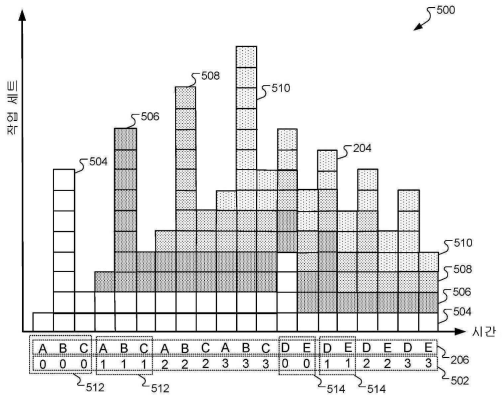
(54) 발명의 명칭 **신경 네트워크 프로세싱을 스케줄링하기**

(57) 요 약

컴퓨터로 구현되는 방법은 하드웨어 회로에서 신경 네트워크를 사용하여 프로세싱될 신경 네트워크 입력들의 배치를 수신하는 단계를 포함한다. 신경 네트워크는 방향 그래프로 배열된 다수의 레이어들을 가지며, 각 레이어는 각각의 파라미터들의 세트를 갖는다. 상기 방법은 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파

(뒷면에 계속)

대 표 도 - 도5



티셔닝을 결정하는 단계를 포함한다. 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 방향 그래프의 파티션이다. 상기 방법은 하드웨어 회로를 사용하여 입력들의 배치를 프로세싱하는 단계를 포함하며, 이는 상기 시퀀스에서 각 슈퍼레이어에 대해 i) 상기 슈퍼레이어의 상기 레이어들에 대한 상기 각각의 파라미터들의 세트를 상기 하드웨어 회로의 메모리에 로딩하는 단계; 및 ii) 상기 배치에서 각 입력에 대해, 입력에 대한 슈퍼레이어 출력을 생성하기 위해, 상기 하드웨어 회로의 상기 메모리에 있는 상기 파라미터들을 사용하여, 상기 슈퍼레이어 내의 상기 레이어들 각각을 통해 상기 입력을 프로세싱하는 단계를 포함한다.

(52) CPC특허분류
G06F 9/5016 (2013.01)
G06N 3/0427 (2013.01)
G06N 3/063 (2013.01)
G06N 3/10 (2013.01)
G06Q 10/06 (2020.05)

명 세 서

청구범위

청구항 1

하나 이상의 프로세싱 디바이스들 및 온칩 메모리를 포함하는 하드웨어 회로를 사용하여 수행되는 방법으로서,

상기 하드웨어 회로에서 신경 네트워크를 사용하여 프로세싱될 신경 네트워크 입력들의 배치(batch)를 수신하는 단계, 상기 신경 네트워크는 방향 그래프(directed graph)로 배열된 복수의 레이어들을 가지며, 각 레이어는 각각의 파라미터들의 세트를 가지며;

상기 신경 네트워크 레이어들의 슈퍼레이어(super layer)들의 시퀀스로의 파티셔닝을 결정하는 단계, 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 상기 방향 그래프의 파티션이며;

상기 신경 네트워크 입력들의 배치를 프로세싱하는 단계를 포함하며, 상기 프로세싱하는 단계는 상기 시퀀스에서 슈퍼레이어 각각에 대해:

상기 슈퍼레이어의 상기 레이어들에 대한 상기 각각의 파라미터들의 세트를 상기 하드웨어 회로의 상기 온칩 메모리에 로딩하는 단계; 및

상기 배치에서 신경 네트워크 입력 각각에 대해:

상기 신경 네트워크 입력에 대한 슈퍼레이어 출력을 생성하기 위해, 상기 하드웨어 회로의 상기 온칩 메모리에 있는 상기 파라미터들을 사용하여, 상기 슈퍼레이어 내의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력에 대응하는 슈퍼레이어 입력을 프로세싱하는 단계를 포함하며,

상기 하드웨어 회로의 상기 온칩 메모리는 임계 저장 용량을 가지며, 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 단계는:

상기 신경 네트워크의 레이어들을 그룹화하는 것,

레이어들의 각 그룹에 대한 총 온칩 메모리 사용량을 결정하는 것, 및

상기 하드웨어 회로가 신경 네트워크 입력들의 상기 배치를 프로세싱할 때 상기 레이어들의 각 그룹의 총 온칩 메모리 사용량이 상기 하드웨어 회로의 상기 온칩 메모리의 상기 임계 저장 용량을 초과하지 않도록 신경 네트워크 레이어들을 레이어들의 그룹들의 시퀀스로 파티셔닝하는 것을 포함하며, 상기 레이어들의 그룹들의 시퀀스는 슈퍼레이어들의 시퀀스를 형성하는 것을 특징으로 하는 방법.

청구항 2

청구항 1에 있어서, 상기 시퀀스의 제1 슈퍼레이어에 대해, 상기 신경 네트워크 입력에 대응하는 상기 슈퍼레이어 입력은 상기 신경 네트워크 입력인 것을 특징으로 하는 방법.

청구항 3

청구항 2에 있어서, 상기 제1 슈퍼레이어 출력 이후에 각 슈퍼레이어에 대한 상기 슈퍼레이어 입력은 상기 시퀀스에서 선행하는 슈퍼레이어에 의해 생성된 슈퍼레이어 출력인 것을 특징으로 하는 방법.

청구항 4

청구항 1에 있어서, 상기 하드웨어 회로를 사용하여, 상기 신경 네트워크 입력들의 배치를 프로세싱하는 단계는 슈퍼레이어 각각에 대해:

상기 배치에서 제2 신경 네트워크 입력에 대응하는 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 순차적으로 프로세싱되기 전에, 상기 배치에서 제1 신경 네트워크 입력에 대한 상기 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 프로세싱되도록, 상기 슈퍼레이어의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력들의 배치에 대응하는 상기 슈퍼레이어 입력들을 순차적으로 프로세싱하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 5

청구항 1에 있어서, 슈퍼레이어의 각 레이어는 작업 세트와 연관되며, 상기 작업 세트는 적어도:

i) 하드웨어 회로 상의 상기 신경 네트워크를 사용하여 프로세싱될 상기 신경 네트워크 입력들의 배치의 하나 이상의 입력들, 또는 상기 슈퍼레이어의 선행 레이어의 하나 이상의 출력들; 및

ii) 상기 레이어를 통해 하나 이상의 입력들을 프로세싱하는데 필요한 메모리 양을 표시하는 크기 파라미터에 의해 정의되며, 그리고

상기 레이어들의 그룹에 대한 총 온칩 메모리 사용량을 결정하는 것은:

i) 상기 그룹의 각 레이어와 연관된 상기 작업 세트에 대한 크기 파라미터를 결정하는 것;

ii) 상기 그룹에 대한 상기 온칩 메모리의 집합 파라미터 용량을 결정하는 것; 및

iii) 상기 결정된 크기 파라미터 및 상기 메모리의 결정된 집합 파라미터 용량에 기초하여, 상기 그룹에 대한 상기 총 온칩 메모리 사용량을 결정하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 6

청구항 1에 있어서, 상기 신경 네트워크 입력들의 배치 및 상기 각각의 파라미터들의 세트는 상기 하드웨어 회로 외부의 소스로부터 수신되며, 그리고 상기 슈퍼레이어의 각 레이어를 통해 상기 신경 네트워크 입력들에 대응하는 슈퍼레이어 입력들을 프로세싱하는 단계는 상기 외부 소스로부터 임의의 추가 파라미터들을 수신하지 않고 상기 슈퍼레이어 입력들을 프로세싱하는 것을 포함하는 것을 특징으로 하는 방법.

청구항 7

컴퓨팅 시스템으로서,

상기 컴퓨팅 시스템에 배치된 하드웨어 회로, 상기 하드웨어 회로는 하나 이상의 프로세싱 디바이스들 및 온칩 메모리를 포함하며,

상기 온칩 메모리는 동작들을 수행하기 위해 상기 하나 이상의 프로세싱 디바이스들에 의해 실행가능한 명령어들을 저장하도록 구성되며, 상기 동작들은:

상기 하드웨어 회로에서 신경 네트워크를 사용하여 프로세싱될 신경 네트워크 입력들의 배치(batch)를 수신하는 동작, 상기 신경 네트워크는 방향 그래프(directed graph)로 배열된 복수의 레이어들을 가지며, 각 레이어는 각각의 파라미터들의 세트를 가지며;

상기 신경 네트워크 레이어들의 슈퍼레이어(super layer)들의 시퀀스로의 파티셔닝을 결정하는 동작, 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 상기 방향 그래프의 파티션이며;

상기 신경 네트워크 입력들의 배치를 프로세싱하는 동작을 포함하며, 상기 프로세싱하는 동작은 상기 시퀀스에서 슈퍼레이어 각각에 대해:

상기 슈퍼레이어의 상기 레이어들에 대한 상기 각각의 파라미터들의 세트를 상기 하드웨어 회로의 상기 온칩 메모리에 로딩하는 동작; 및

상기 배치에서 신경 네트워크 입력 각각에 대해:

상기 신경 네트워크 입력에 대한 슈퍼레이어 출력을 생성하기 위해, 상기 하드웨어 회로의 상기 온칩 메모리에 있는 상기 파라미터들을 사용하여, 상기 슈퍼레이어 내의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력에 대응하는 슈퍼레이어 입력을 프로세싱하는 동작을 포함하며,

상기 하드웨어 회로의 상기 온칩 메모리는 임계 저장 용량을 가지며, 상기 하나 이상의 프로세싱 디바이스들은:

상기 신경 네트워크의 레이어들을 그룹화하는 것,

레이어들의 각 그룹에 대한 총 온칩 메모리 사용량을 결정하는 것, 및

상기 하드웨어 회로가 신경 네트워크 입력들의 상기 배치를 프로세싱할 때 상기 레이어들의 각 그룹의 총 온칩 메모리 사용량이 상기 하드웨어 회로의 상기 온칩 메모리의 상기 임계 저장 용량을 초과하지 않도록 신

경 네트워크 레이어들을 레이어들의 그룹들의 시퀀스로 파티셔닝하는 것에 의해, 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하기 위해 구성되며, 상기 레이어들의 그룹들의 시퀀스는 슈퍼레이어들의 시퀀스를 형성하는 것을 특징으로 하는 컴퓨팅 시스템.

청구항 8

청구항 7에 있어서, 상기 시퀀스의 제1 슈퍼레이어에 대해, 상기 신경 네트워크 입력에 대응하는 상기 슈퍼레이어 입력은 상기 신경 네트워크 입력인 것을 특징으로 하는 컴퓨팅 시스템.

청구항 9

청구항 8에 있어서, 상기 제1 슈퍼레이어 출력 이후에 각 슈퍼레이어에 대한 상기 슈퍼레이어 입력은 상기 시퀀스에서 선행하는 슈퍼레이어에 의해 생성된 슈퍼레이어 출력인 것을 특징으로 하는 컴퓨팅 시스템.

청구항 10

청구항 7에 있어서, 상기 하드웨어 회로를 사용하여, 상기 신경 네트워크 입력들의 배치를 프로세싱하는 단계는 슈퍼레이어 각각에 대해:

상기 배치에서 제2 신경 네트워크 입력에 대응하는 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 순차적으로 프로세싱되기 전에, 상기 배치에서 제1 신경 네트워크 입력에 대한 상기 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 프로세싱되도록, 상기 슈퍼레이어의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력들의 배치에 대응하는 상기 슈퍼레이어 입력들을 순차적으로 프로세싱하는 것을 포함하는 것을 특징으로 하는 컴퓨팅 시스템.

청구항 11

청구항 7에 있어서, 슈퍼레이어의 각 레이어는 작업 세트와 연관되며, 상기 작업 세트는 적어도:

i) 하드웨어 회로 상의 상기 신경 네트워크를 사용하여 프로세싱될 상기 신경 네트워크 입력들의 배치의 하나 이상의 입력들, 또는 상기 슈퍼레이어의 선행 레이어의 하나 이상의 출력들; 및

ii) 상기 레이어를 통해 하나 이상의 입력들을 프로세싱하는데 필요한 메모리 양을 표시하는 크기 파라미터에 의해 정의되며, 그리고

상기 하나 이상의 프로세싱 디바이스들은:

i) 상기 그룹의 각 레이어와 연관된 상기 작업 세트에 대한 크기 파라미터를 결정하는 것;

ii) 상기 그룹에 대한 상기 온칩 메모리의 집합 파라미터 용량을 결정하는 것; 및

iii) 상기 결정된 크기 파라미터 및 상기 메모리의 결정된 집합 파라미터 용량에 기초하여, 상기 그룹에 대한 상기 총 온칩 메모리 사용량을 결정하는 것에 의해 상기 레이어들의 그룹에 대한 총 온칩 메모리 사용량을 결정하도록 구성되는 것을 특징으로 하는 컴퓨팅 시스템.

청구항 12

청구항 7에 있어서, 상기 신경 네트워크 입력들의 배치 및 상기 각각의 파라미터들의 세트는 상기 하드웨어 회로 외부의 소스로부터 수신되며, 그리고 상기 슈퍼레이어의 각 레이어를 통해 상기 신경 네트워크 입력들에 대응하는 슈퍼레이어 입력들을 프로세싱하는 동작은 상기 외부 소스로부터 임의의 추가 파라미터들을 수신하지 않고 상기 슈퍼레이어 입력들을 프로세싱하는 것을 포함하는 것을 특징으로 하는 컴퓨팅 시스템.

청구항 13

삭제

청구항 14

삭제

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

발명의 설명

기 술 분 야

배 경 기 술

[0001] 본 명세서는 신경 네트워크 계산을 수행하기 위한 메모리 관리 프로세스들에 관한 것이다.

[0002] 신경 네트워크들은 동작들의 하나 이상의 레이어들을 이용하여 수신된 입력에 대한 출력 예를 들면, 분류를 생성하는 기계 학습 모델들이다. 일부 신경 네트워크들은 출력 레이어에 더하여 하나 이상의 히든 레이어들을 포함한다. 각 히든 레이어의 출력은 네트워크에서 다음 레이어 즉, 다음 히든 레이어 또는 네트워크의 출력 레이어에 대한 입력으로서 사용된다. 네트워크의 레이어들 전부 또는 일부는 각각의 파라미터들의 세트의 현재 값들에 따라 수신된 입력으로부터 출력을 생성한다.

[0003] 일부 신경 네트워크는 하나 이상의 컨볼루션 신경 네트워크 레이어들을 포함한다. 각 컨벌루션 신경 네트워크 레이어는 연관된 커널들의 세트를 가진다. 각 커널은 사용자가 만든 신경 네트워크 모델에 의해 설정된 값을 포함한다. 일부 구현예에서, 커널들은 특정한 이미지 윤곽, 모양 또는 색상을 식별한다. 커널들은 가중 입력들의 행렬 구조로서 표현될 수 있다. 각 컨벌루션 레이어는 또한 액티베이션 입력들의 세트를 프로세싱할 수 있다. 액티베이션 입력들의 세트 또한 행렬 구조로서 표현될 수 있다.

발명의 내용

[0004] 본 명세서에 기술된 발명은 하드웨어 회로에서 신경 네트워크를 사용하여 프로세싱될 신경 네트워크 입력들의 배치를 수신하는 시스템 및 방법을 포함한다. 신경 네트워크는 방향 그래프로 배열된 다수의 레이어들을 가지며, 각 레이어는 각각의 파라미터들의 세트를 갖는다. 설명된 기술들에 따른 방법은 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 단계를 포함한다. 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 방향 그래프의 파티션일 수 있다.

[0005] 설명된 방법은 하드웨어 회로를 사용하여 입력들의 배치를 프로세싱하는 단계를 포함할 수 있다. 예를 들어, 입력들의 배치를 프로세싱하는 단계는 시퀀스의 각 슈퍼레이어에서 레이어들에 대한 각각의 파라미터들의 세트를 하드웨어 회로의 메모리에 로딩하는 단계를 포함할 수 있다. 추가적으로, 상기 배치에서 각 입력에 대해, 설명된 방법은 입력에 기초하여 슈퍼레이어 출력을 생성하기 위해, 상기 하드웨어 회로의 상기 메모리에 있는 상기 파라미터들을 사용하여, 상기 슈퍼레이어 내의 상기 레이어들 각각을 통해 상기 입력을 프로세싱하는 단계를 포함할 수 있다.

[0006] 본 명세서에 기술된 발명의 일 양태는 컴퓨터로 구현되는 방법으로 이용될 수 있다. 상기 방법은 하드웨어 회로

에서 신경 네트워크를 사용하여 프로세싱될 신경 네트워크 입력들의 배치(batch)를 수신하는 동작, 상기 신경 네트워크는 방향 그래프(directed graph)로 배열된 복수의 레이어들을 가지며, 각 레이어는 각각의 파라미터들의 세트를 가지며; 및 상기 신경 네트워크 레이어들의 슈퍼레이어(super layer)들의 시퀀스로의 파티셔닝을 결정하는 동작을 포함하며, 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 상기 방향 그래프의 파티션이다.

[0007] 상기 방법은 상기 하드웨어 회로를 사용하여 상기 신경 네트워크 입력들의 배치를 프로세싱하는 단계를 포함하며, 상기 시퀀스에서 각 슈퍼레이어에 대해: 상기 슈퍼레이어의 상기 레이어들에 대한 상기 각각의 파라미터들의 세트를 상기 하드웨어 회로의 메모리에 로딩하는 단계; 및 상기 배치에서 신경 네트워크 입력 각각에 대해: 상기 신경 네트워크 입력에 대한 슈퍼레이어 출력을 생성하기 위해, 상기 하드웨어 회로의 상기 메모리에 있는 상기 파라미터들을 사용하여, 상기 슈퍼레이어 내의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력에 대응하는 슈퍼레이어 입력을 프로세싱하는 단계를 포함한다.

[0008] 이들 또는 다른 실시예들은 다음 구성들 중 하나 이상을 각각 선택적으로 포함할 수 있다. 예를 들면, 일부 구현예에서, 상기 시퀀스의 제1 슈퍼레이어에 대해, 상기 신경 네트워크 입력에 대응하는 상기 슈퍼레이어 입력은 상기 신경 네트워크 입력이다. 일부 구현예에서, 상기 제1 슈퍼레이어 출력 이후에 각 슈퍼레이어에 대한 상기 슈퍼레이어 입력은 상기 시퀀스에서 선행하는 슈퍼레이어에 의해 생성된 슈퍼레이어 출력이다.

[0009] 일부 구현예에서, 상기 하드웨어 회로를 사용하여 상기 신경 네트워크 입력들의 배치를 프로세싱하는 단계는, 각 슈퍼레이어에 대해, 상기 배치에서 제2 신경 네트워크 입력에 대응하는 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 순차적으로 프로세싱되기 전에, 상기 배치에서 제1 신경 네트워크 입력에 대한 상기 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 프로세싱되도록, 상기 슈퍼레이어의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력들의 배치에 대응하는 상기 슈퍼레이어 입력들을 순차적으로 프로세싱하는 것을 포함한다.

[0010] 일부 구현예에서, 슈퍼레이어의 각각의 레이어들은 작업 세트와 연관되며, 각 작업 세트는 적어도: i) 하드웨어 회로 상의 상기 신경 네트워크를 사용하여 프로세싱될 상기 신경 네트워크 입력들의 배치의 하나 이상의 입력들, 또는 상기 슈퍼레이어의 선행 레이어의 하나 이상의 출력들; 및 ii) 상기 슈퍼레이어의 레이어들 각각을 통해 하나 이상의 입력들을 프로세싱하는데 필요한 메모리 양을 표시하는 크기 파라미터에 의해 정의된다.

[0011] 일부 구현예에서, 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 단계는: i) 적어도 하나의 작업 세트에 대한 특정한 크기 파라미터를 결정하는 것; ii) 상기 하드웨어 회로의 메모리의 특정한 집합 파라미터 용량을 결정하는 것; 및 iii) 상기 적어도 하나의 작업 세트에 대한 특정한 크기 파라미터 또는 상기 하드웨어 회로의 메모리의 특정한 집합 파라미터 용량 중 적어도 하나에 기초하여, 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 것을 포함한다.

[0012] 일부 구현예에서, 하드웨어 회로의 메모리는 임계 저장 용량을 가지며, 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 단계는: 상기 하드웨어 회로의 메모리의 임계 저장 용량에 기초하여, 상기 신경 네트워크 레이어들을 슈퍼레이어들의 시퀀스로 파티셔닝하는 것을 포함한다.

[0013] 일부 구현예에서, 상기 신경 네트워크 레이어들이 슈퍼레이어들의 시퀀스로 파티셔닝되어, 상기 하드웨어 회로가 상기 신경 네트워크 입력들의 배치를 프로세싱하는 경우 상기 메모리의 임계 저장 용량을 초과하지 않도록 한다.

[0014] 일부 구현예에서, 상기 신경 네트워크 입력들의 배치 및 상기 각각의 파라미터들의 세트는 상기 하드웨어 회로 외부의 소스로부터 수신되며, 그리고 상기 슈퍼레이어의 각 레이어를 통해 상기 신경 네트워크 입력들에 대응하는 슈퍼레이어 입력들을 프로세싱하는 단계는 상기 외부 소스로부터 임의의 추가 파라미터들을 수신하지 않고 상기 슈퍼레이어 입력들을 프로세싱하는 것을 포함한다.

[0015] 상기 그리고 다른 양태들의 다른 구현예들은 상기 방법들의 액션들을 수행하도록 구성된 대응 시스템들, 장치들 및 컴퓨터 저장 디바이스들에 인코딩된 컴퓨터 프로그램들을 포함한다. 하나 이상의 컴퓨터들 또는 하드웨어 회로들의 컴퓨팅 시스템은 동작 중에 시스템이 상기 동작들을 수행하게 하는 소프트웨어, 펌웨어, 하드웨어 또는 이들의 조합에 의해 구성될 수 있다. 하나 이상의 컴퓨터 프로그램들은 데이터 프로세싱 장치에 의해 실행될 때 상기 장치로 하여금 동작들을 수행하게 하는 명령어들을 갖도록 구성될 수 있다.

[0016] 본 명세서에 기술된 본 발명은 다음의 이점들을 실현하기 위해 특정한 실시예들에서 구현될 수 있다. 신경 네트워크 레이어들을 슈퍼레이어들의 시퀀스로 파티셔닝함으로써, 신경 네트워크가 파라미터들의 세트를 사용하여 입력을 프로세싱할 때 신경 네트워크 하드웨어 회로에 의한 외부 통신이 최소화될 수 있다. 계산 프로세스들에

서 하드웨어 회로에 의한 외부 통신을 최소화하면 하드웨어 회로에 의한 대역폭 소비 및 에너지 최적화가 개선될 수 있다.

[0017] 또한, 슈퍼레이어들의 시퀀스는 신경 네트워크 모델의 "배치(batch)" 및 "레이어(layer)" 차원들을 혼합하는 글로벌 스케줄링 프로세스를 제공하여, 신경 네트워크 레이어들을 통한 입력들의 프로세싱을 위한 하나 이상의 메모리 작업 세트들을 최적화할 수 있다. 예를 들어, 배치 및 레이어 차원에 대한 글로벌 스케줄링을 수행함으로써, 신경 네트워크 어플리케이션들의 라이브 메모리 작업 세트가 최소화되어, 주어진 하드웨어 회로에 대한 입력들의 배치없는 실행을 향상시킬 수 있다. 라이브 메모리 작업 세트들은 신경 네트워크의 레이어들을 통한 프로세싱을 위한 데이터에 대응할 수 있으며, 여기서 데이터는 현재 데이터 프로세싱 장치 또는 프로세서 하드웨어 회로의 물리적 메모리 공간에 상주한다.

[0018] 또한, 예시적 하드웨어 회로는 최소화된 작업 세트들의 입력들 및 파라미터들이 SRAM 용량을 사용하여 온칩에 저장될 수 있도록 온칩 메모리(예를 들어, SRAM)를 포함할 수 있다. 따라서, SRAM 용량이 슈퍼레이어들의 시퀀스들을 제공하는 글로벌 스케줄링 프로세스에 기초하여 효율적으로 활용될 때, 입력 및 파라미터를 저장하기 위해 추가 메모리 리소스들이 더 이상 필요하지 않는 경우 비용 절감이 실현될 수 있다. 일부 구현예에서, 온칩 SRAM 용량은 특정한 설계 요건들을 충족시키고 슈퍼레이어 시퀀스들을 형성하는 것을 포함하거나 포함하지 않을 수 있는 스케줄링 프로세스들을 제공하기 위해 필요에 따라 스케일 업 또는 다운될 수 있다.

[0019] 본 명세서에 기술된 본 발명의 하나 이상의 구현예들의 세부 사항은 첨부 도면과 아래의 설명에서 기술된다. 본 발명의 다른 잠재적 구성들, 양태들 및 이점들은 설명, 도면 및 청구항으로부터 명백해질 것이다.

도면의 간단한 설명

[0020] 도 1은 각각의 파라미터들의 세트를 각각 가지는 신경 네트워크의 레이어들을 통해 신경 네트워크 입력들을 프로세싱하기 위한 예시적 하드웨어 회로를 도시한다.

도 2a는 신경 네트워크의 각각의 레이어들을 사용하여 단일 배치 엘리먼트의 프로세싱과 관련된 예시적 그래프를 도시한다.

도 2b는 신경 네트워크의 주어진 레이어에 대한 다수의 배치 엘리먼트들의 프로세싱과 관련된 예시적 그래프를 도시한다.

도 3은 슈퍼레이어를 형성하는 신경 네트워크의 다수의 레이어들 중에서의 단일 배치 엘리먼트의 프로세싱과 관련된 예시적 그래프를 도시한다.

도 4는 신경 네트워크의 슈퍼레이어들을 통해 신경 네트워크 입력들을 프로세싱하는 방법에 대한 예시적 흐름도이다.

도 5는 슈퍼레이어의 다수의 레이어들을 사용하여 단일 배치 엘리먼트를 프로세싱하기 위해 슈퍼레이어들의 시퀀스로 파티셔닝된 신경 네트워크 레이어들을 표현하는 예시적 그래프를 도시한다.

도 6a는 신경 네트워크 레이어에 대한 작업 세트 크기를 표현하는 예시적 그래프를 도시한다.

도 6b는 신경 네트워크의 슈퍼레이어에 대한 작업 세트 크기를 표현하는 예시적 그래프를 도시한다.

다양한 도면들에서 동일한 참조 번호 및 기호는 동일한 구성요소를 표시한다.

발명을 실시하기 위한 구체적인 내용

[0021] 다수의 레이어들을 갖는 신경 네트워크는 추론들을 계산하는데 사용될 수 있다. 예를 들어, 입력이 주어지면, 신경 네트워크는 입력에 대한 추론을 계산할 수 있다. 신경 네트워크는 신경 네트워크의 레이어들 각각을 통해 입력을 프로세싱함으로써 이러한 추론을 계산한다. 특히, 신경 네트워크의 레이어들은 방향 그래프로 배열될 수 있으며, 레이어들의 일부 또는 전부는 각각의 파라미터들의 세트를 갖는다. 각 레이어는 입력을 수신하고, 레이어에 대한 파라미터들의 세트에 따라 입력을 프로세싱하여 출력을 생성한다. 출력은 다음 신경 네트워크 레이어에서 입력으로 사용될 수 있다.

[0022] 따라서, 수신된 입력으로부터 추론을 계산하기 위해, 신경 네트워크는 입력을 수신하고 방향 그래프에서 각각의 신경 네트워크 레이어들을 통해 입력을 프로세싱하여 다음 신경 네트워크 레이어에 대한 입력으로서 제공되는 하나의 신경 네트워크 레이어로부터의 출력을 가지는 추론을 생성한다. 신경 네트워크 레이어에 대한 데이터 입

력, 예를 들어, 신경 네트워크에 대한 입력 또는 신경 네트워크 레이어에 대한 방향 그래프에서 레이어에 연결된 하나 이상의 레이어들의 출력은 레이어에 대한 액티베이션 입력들로서 지칭될 수 있다.

[0023] 방향 그래프에서 특정한 레이어는 다수의 입력들을 수신하거나, 다수의 출력들을 생성하거나 둘 모두를 수행할 수 있다. 신경 네트워크의 레이어들은 또한 레이어의 출력이 입력으로서 이전 레이어에 다시 전송될 수 있도록 배열될 수 있다. 설명된 기술들에 따른 방법은 상기 신경 네트워크 레이어들의 슈퍼레이어(super layer)들의 시퀀스로의 파티셔닝을 결정하는 단계를 포함하며, 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 상기 방향 그래프의 파티션이도록 한다.

[0024] 설명된 방법은 하드웨어 회로상의 신경 네트워크에 대한 시퀀스에서 각각의 슈퍼레이어들의 레이어들을 통해 입력들의 배치를 프로세싱하는 단계를 포함할 수 있다. 입력들의 배치를 프로세싱하는 단계는 레이어들에 대한 파라미터들을 하드웨어 회로의 메모리로 로딩하는 단계, 및 파라미터들을 사용하여 신경 네트워크 입력을 프로세싱하여 입력에 대한 각각의 슈퍼레이어 출력들을 생성하는 단계를 포함할 수 있다.

[0025] 일부 구현예에서, 본 명세서에서 기술된 하나 이상의 기능들은 시스템의 하드웨어 회로 또는 전자 컴포넌트를 사용하여 수행될 수 있다. 하드웨어 회로는 하드웨어 회로에 전기적으로 연결된 제어 디바이스로부터 제어 신호들을 수신할 수 있다. 하드웨어 회로는 신경 네트워크 레이어에 입력들 및 상기 입력들을 프로세싱하는데 사용되는 파라미터들을 저장하기 위한 하나 이상의 비일시적 기계 판독가능 저장 매체(예를 들어, 메모리)를 포함하는 패키징된 전자 디바이스일 수 있다.

[0026] 하드웨어 회로는 프로세서 마이크로 칩(예를 들어, CPU 또는 GPU)과 같은 패키징된 집적 회로 또는 프로세서 디바이스를 형성하는 다수의 컴포넌트들을 포함할 수 있다. 따라서, 이 경우에, 하드웨어 회로의 메모리는 마이크로 칩을 형성하는 다수의 다른 컴포넌트들과 관련하여 "온칩" 메모리일 수 있다. 본 명세서에서 사용되는 바와 같이, 패키징된 하드웨어 회로 또는 전자 디바이스는 실리콘 웨이퍼와 같은 반도체 물질을 포함할 수 있으며, 이는 지지 케이스 내에 캡슐화되거나 둘러싸여 있다. 지지 케이스는 디바이스를 인쇄된 회로 보드에 연결하기 위해 케이스의 주변부로부터 연장되는 하나의 도체선들을 포함할 수 있다.

[0027] 제어 디바이스는 하드웨어 회로와 이격되어 있고 하드웨어 회로의 컴포넌트 패키지(예를 들어, 지지 케이스)에 의해 둘러싸인 적어도 온칩 메모리 외부에 있는 외부 제어기일 수 있다. 외부 제어기는 하드웨어 회로에 제어 신호를 제공하여, 하드웨어 회로로 하여금 전술 한 입력 및 파라미터를 사용하여 신경 네트워크 추론 계산을 수행하게 하는 시스템 레벨 제어기일 수 있다. 외부 컨트롤러는 "오프 칩"메모리를 포함할 수 있으며, 여기서 적어도 메모리는 패키징된 하드웨어 회로의 온칩 메모리와 같은 위치에 있지 않기 때문에, 메모리는 오프 칩이다.

[0028] 일부 구현예에서, 오프-칩 메모리를 사용하지 않고 추론 계산을 수행하는 경우, 외부 제어기는 입력 및 파라미터를 저장하기 위해 하드웨어 회로의 온-칩 메모리를 사용할 수 있다. 시스템의 적어도 하나의 제어기로부터 제어 신호를 수신하는 것에 응답하여, 하드웨어 회로는 온칩 메모리에 액세스하고 저장된 입력 및 파라미터를 사용하여 신경 네트워크 계산을 수행한다.

[0029] 도 1은 신경 네트워크 계산을 수행하는데 사용될 수 있는 하드웨어 회로(100)의 예시를 도시한다. 신경 네트워크 계산을 수행하는 것은 각각의 파라미터들의 세트를 각각 가지는 신경 네트워크의 레이어들을 통해 신경 네트워크 입력들을 프로세싱하는 회로(100)를 포함할 수 있다. 일부 구현예에서, 회로(100)는 하나 이상의 프로세서들, 프로세서 마이크로 칩들 또는 신경 네트워크를 구현하는 다른 회로 컴포넌트들을 포함하는 하드웨어 회로에 대응한다. 다른 구현예에서, 회로(100)는 하나 이상의 신경 네트워크들을 형성하는 하나 이상의 하드웨어 회로들, 프로세서들 및 다른 관련 회로 컴포넌트들을 포함할 수 있다. 일반적으로, 설명된 기술에 따른 방법은 CPU, GPU, 디지털 신호 프로세서(DSP) 또는 다른 관련 프로세서 아키텍처와 같은 다양한 프로세서 아키텍처에 적용되거나 이를 사용하여 구현될 수 있다.

[0030] 회로(100)는 일반적으로 메모리(104)와 연관된 입력이 메모리(102)의 메모리 주소에 저장되거나 메모리 주소로부터 검색되게 하는 하나 이상의 제어 신호(110)들을 제공하는 제어기(108)를 포함한다. 유사하게, 제어기(108)는 파라미터 메모리(106)에 대한 파라미터들로 하여금 메모리(102)의 메모리 주소에 저장되거나 메모리 주소로부터 검색되게 하는 하나 이상의 제어 신호(110)들을 제공한다.

[0031] 회로(100)는 하나 이상의 MAC(multiply accumulate cell) 셀/유닛(들)(107), 입력 액티베이션 버스(112) 및 출력 액티베이션 버스(114)를 더 포함한다. 제어 신호들(110)은 예를 들어, 메모리(102)로 하여금 입력 액티베이션 버스(112)에 하나 이상의 입력들을 제공하게 하고, 메모리(102)로 하여금 파라미터 메모리(106)로부터 하나 이상의 파라미터들을 제공하게 하고 및/또는 MAC 셀/유닛(107)으로 하여금 입력들과 파라미터들을 사용하게 하

여 출력 액티베이션 버스(114)에 제공되는 출력 액티베이션들을 생성하는 계산을 수행하게 할 수 있다.

[0032] 제어기(108)는 하나 이상의 프로세싱 유닛들 및 메모리를 포함할 수 있다. 제어기(108)의 프로세싱 유닛은 하나 이상의 프로세서들(예를 들어, 마이크로 프로세서 또는 중앙 처리 장치(CPU)), 그래픽 처리 장치(GPU), 주문형 집적 회로(ASIC), 또는 상이한 프로세서들의 조합을 포함할 수 있다. 제어기(108)는 또한 본 명세서에 기술된 결정 및 계산 중 하나 이상을 수행하기 위한 추가 프로세싱 옵션을 제공하는 다른 스토리지 또는 컴퓨팅 리소스/디바이스들(예를 들어, 버퍼, 레지스터, 제어 회로 등)을 포함할 수 있다.

[0033] 일부 구현예에서, 제어기(108)의 프로세싱 유닛(들)은 제어기(108) 및 회로(100)로 하여금 본 명세서에서 설명된 하나 이상의 기능들을 수행하게 하기 위해 메모리에 저장된 명령어들을 실행한다. 제어기(108)의 메모리는 하나 이상의 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 본 명세서에 기술된 비일시적 기계 판독가능 저장 매체는 솔리드 스테이트 메모리, 자기 디스크, 광 디스크, 휴대용 컴퓨터 디스켓, 랜덤 액세스 메모리(RAM), 읽기 전용 메모리(ROM), 소거가능한 프로그래머블 읽기 전용 메모리(예: EPROM, EEPROM 또는 플래시 메모리) 또는 정보를 저장할 수 있는 임의의 다른 유형의 매체를 포함할 수 있다.

[0034] 회로(100)는 예시적 계산 유닛 또는 계산 타일일 수 있고, 텐서, 매트릭스 및/또는 데이터 어레이들과 같은 다차원 데이터 구조들과 연관된 계산을 수행하기 위한 추가 하드웨어 구조를 포함할 수 있다. 일부 구현예에서, 입력 값들은 액티베이션 메모리(104)에 사전-로딩될 수 있고, 파라미터/가중치 값들은 신경 네트워크 컴퓨팅 시스템과 연관된 외부 또는 상위 레벨 제어 디바이스로부터 회로(100)에 의해 수신된 데이터 값들을 사용하여 파라미터 메모리(106)에 사전-로딩될 수 있다.

[0035] 회로(100)는 시스템의 신경 네트워크를 사용함으로써 수행될 특정한 계산 동작을 정의하는 명령어들을 수신할 수 있다. 일반적으로, 메모리(102)에 저장된 데이터 값들은 일반적으로 각각의 메모리 주소 위치에 기록된다. 메모리(102)의 주소 위치는 특정한 계산 동작을 수행하기 위해 입력과 같은 데이터 값이 필요할 때 예시적 제어 디바이스(예를 들어, 제어기(108))에 의해 액세스될 수 있다.

[0036] 제어기(108)는 하나 이상의 제어 신호들(110)을 메모리(102)에 제공하여 메모리(102)로부터 입력 액티베이션 버스(112)에 입력들을 로딩하고, 그 값들을 MAC(107)을 포함하는 계산 유닛의 어레이에 제공할 수 있다. 액티베이션 메모리의 인덱스(104)는 입력들을 갖는 모든 메모리 주소 위치들을 포함할 수 있다. 데이터 버스(112)는 하나 이상의 계산 어레이 유닛들에 의해 액세스 가능하다. 계산 어레이의 유닛들은 수신된 액티베이션 값들에 기초하여 행렬 곱셈과 관련된 계산을 수행하기 위해 하나 이상의 액티베이션 값들을 데이터 버스(112)로부터 수신할 수 있다.

[0037] 주어진 계산 사이클에 대해, 회로(100)는 신경 네트워크 레이어에 대한 추론 계산과 연관된 곱셈 연산을 실행하기 위해 액티베이션 메모리(104) 및 파라미터 메모리(106)의 엘리먼트에 대한 액세스를 요구할 수 있다. 계산이 수행되는 사이클에 대해, 제어기(108)는 한 번에 하나의 입력 값을 제공할 수 있고, MAC 셀(107)을 포함하는 계산 유닛의 어레이는 주어진 입력에 대해 상이한 출력 액티베이션들을 생성하기 위해 액티베이션에 가중치/파라미터를 곱할 것이다.

[0038] 일부 구현예에서, 계산 유닛들의 어레이의 각 MAC 셀(107)은 신경 네트워크 레이어의 상이한 출력 깊이를 담당할 수 있다. 계산 유닛들의 어레이는 제어기(108)에 의해 완전히 제어될 수 있고, 제어기(108)는 특정한 계산을 수행할 필요가 있을 때 액티베이션 값의 검출에 기초하여 결정할 수 있다.

[0039] 또한, 입력 값들은 메모리(102)에 저장하기 위해 회로(100)에 도달하면 분석될 수 있다. 입력들을 분석하는 것에 응답하여, 제어기(108)는 특정한 입력 값들만 메모리(102)에 저장함으로써(예를 들어, 0이 아닌 액티베이션 값들만), 액티베이션 데이터를 효율적으로 압축하기 위해 프로그래밍된 명령어들을 실행할 수 있으며, 이에 의해 메모리 저장 공간 및 대응하는 대역폭을 절약할 수 있다.

[0040] 회로(100)가 입력 및 파라미터를 수신할 때, 제어기(108)는 예를 들어, 하나 이상의 직접 메모리 액세스 동작을 실행할 수 있다. 이들 메모리 액세스 동작의 실행은 메모리(102)의 주소 위치에, 액티베이션 메모리(104)의 차원적 엘리먼트들에 대응하는 입력을 저장하는 것을 포함한다. 유사하게, 제어기(108)는 메모리(102)의 주소 위치에, 파라미터 메모리(106)의 차원적 엘리먼트들에 대응하는 파라미터들을 저장할 수 있다. 제어기(108)는 특정한 입력이 페치될 메모리 주소들을 유지하는 하나 이상의 주소 레지스터들을 더 포함할 수 있다. 또한, 하나 이상의 레지스터들은 대응하는 파라미터가 특정한 입력과 곱해지기 위해 페치되는 메모리 주소들을 저장할 것이다.

[0041] 제어기(108)는 제1 입력에 대한 대응하는 파라미터(및 메모리 주소)를 결정하고, 제1 및 제2 입력이 순차적으로

프로세싱될 때 제2 입력에 대한 대응하는 파라미터(및 메모리 주소)를 결정하기 위해 상기 언급된 레지스터들을 참조할 수 있다. 일부 구현예에서, 제1 신경 네트워크 레이어에서 계산된 출력 액티베이션은 네트워크의 다음/후속 제2 레이어, 예를 들어, 네트워크의 다음 히든 레이어 또는 출력 레이어에 대한 입력들로서 사용된다. 일반적으로, 신경 네트워크의 각 레이어는 각각의 세트의 현재 값들에 따라 수신된 입력으로부터 출력을 생성한다.

[0042] 대안적 구현예에서, 단일 입력이 파라미터 메모리(106)의 주어진 차원 엘리먼트에 대한 다양한 가중치들을 커버하는 몇몇 곱셈 연산들에 대한 피연산자로 사용되는 일부 계산 연산들이 있을 수 있다(예를 들어, "X" 또는 "Y" 차원을 반복하기 위해). 설명된 기술들에 따르면, 회로(100)는 컴퓨팅 시스템 또는 기계 학습 시스템의 외부 제어기로부터 제어 신호들을 수신하도록 구성될 수 있다. 외부 제어기는 회로(100)의 온칩 메모리에 저장된 신경 네트워크 입력들 및 파라미터들의 배치들을 제공할 수 있다. 아래에서 더 상세히 설명되는 바와 같이, 외부 제어기는 회로(100) 상의 신경 네트워크에 의한 배치 엘리먼트 프로세싱을 위한 스케줄링 정책을 구현하도록 구성될 수 있다.

[0043] 예를 들어, 시스템의 외부 제어기는 회로(100)에 제어 신호들을 제공하여, 회로(100)로 하여금 회로(100)의 온칩 메모리에 저장된 입력들 및 파라미터들을 사용하여 신경 네트워크의 레이어들을 통해 신경 네트워크 입력들을 프로세싱하게 할 수 있다. 설명된 기술들에 따르면, 특정한 스케줄링 정책은 신경 네트워크의 레이어들을 슈퍼레이어들의 하나 이상의 시퀀스들을 형성하는 레이어들의 그룹핑으로 파티셔닝하기 위해 사용될 수 있다(아래에 설명됨). 그 다음, 시스템 제어기는 회로(100)를 사용하여 온칩 메모리에 저장된 입력들 및 파라미터들에 액세스한 다음 슈퍼레이어들의 시퀀스에서 각 레이어를 통해 신경 네트워크 입력들의 배치를 프로세싱할 수 있다.

[0044] 도 2a는 신경 네트워크의 각각의 레이어들을 사용하여 단일 배치 엘리먼트의 프로세싱과 관련된 예시적 그래프(200A)를 도시한다. 일부 구현예에서, 아래에 설명되는 그래프(200A/B) 및 그래프(300, 500, 600A/B)는 신경 네트워크의 토폴로지를 표현할 수 있는 예시적 방향 그래프와는 상이하다.

[0045] 그래프(200A)는 신경 네트워크의 레이어들을 통한 배치 엘리먼트의 프로세싱 동안 작업 세트들의 크기가 어떻게 변하는지를 도시한다. 작업 세트의 크기는 저장 유닛들(204)로 표현된다. 일반적으로, 주어진 신경 네트워크 레이어에 대한 작업 세트는 신경 네트워크 레이어로의 입력들, 신경 네트워크 레이어으로부터의 출력들 및 신경 네트워크 레이어에 의해 입력들을 프로세싱하는데 사용되는 파라미터들을 포함한다. 작업 세트들은 일반적으로 주어진 신경 네트워크 계산에 필요하고 아래에 더 상세히 설명되는 하나 이상의 데이터 구조들의 그룹을 포함한다.

[0046] 하나 이상의 저장 유닛들(204)은 신경 네트워크 레이어에 대한 작업 세트의 입력들과 및 연관된 파라미터들을 저장하는데 사용된다. 저장 유닛들(204)은 상술한 메모리(102)의 메모리 리소스와 연관될 수 있다. 배치 엘리먼트는 하드웨어 회로에서 예시적 신경 네트워크를 사용하여 프로세싱되는 단일 신경 네트워크 입력이다.

[0047] 위에서 언급한 바와 같이, 신경 네트워크는 추론을 계산하는데 사용되는 다수의 레이어들을 포함할 수 있고, 신경 네트워크의 레이어들을 통해 신경 네트워크 입력을 프로세싱함으로써 추론이 계산된다. 따라서, 그래프(200A)는 레이어 A, 레이어 B, 레이어 C, 레이어 D 및 레이어 E를 포함하는 신경 네트워크 레이어들(206)을 추가로 도시한다. 그래프(200A)는 배치 엘리먼트가 먼저 레이어 A를 통해, 이어서 레이어 B를 통해, 이어서 레이어 C를 통해, 이어서 레이어 D를 통해, 이어서 레이어 E를 통해 프로세싱됨을 나타낸다. 일부 구현예에서, 레이어들(206)의 각각의 레이어들은 컨볼루션 레이어, 감소 레이어, 완전히 연결된(FC) 레이어, 분류기 레이어, 엘리먼트별 곱셈 레이어, 또는 풀링 레이어, 예를 들어 평균 풀링 레이어 또는 최대 풀링 레이어 중 하나의 유형의 신경 네트워크 레이어일 수 있다.

[0048] 신경 네트워크 레이어에 대한 작업 세트는 신경 네트워크의 각각의 레이어를 통해 배치 엘리먼트를 프로세싱하는데 사용되는 하나 이상의 배치 엘리먼트들 및 파라미터들을 포함할 수 있다. 작업 세트는: i) 하드웨어 회로 상의 신경 네트워크를 사용하여 프로세싱될 입력들의 배치의 하나 이상의 입력/배치 엘리먼트; 및 ii) 입력들 및 파라미터들을 저장하는데 필요한 메모리의 양을 표시하는 크기 파라미터 또는 저장 유닛(204)들의 수에 의해 정의될 수 있다. 입력들에 더하여, 작업 세트는 출력 액티베이션들을 포함할 수 있다. 일부 구현예에서, 신경 네트워크는 전술된 배치 엘리먼트들과 연관된 "배치" 차원 및 레이어들(206)에 대응하는 "레이어" 차원을 갖는 것으로 기술될 수 있다.

[0049] 일반적으로, 도 2a의 다음 설명은 예를 들어, 도 3 내지 6을 참조하여 아래에 설명되는 개선된 신경 네트워크 스케줄링 프로세스들에 대한 컨텍스트를 제공한다. 예를 들어, 레이어들(206)은 적어도 5개의 레이어(예를

들어, 레이어 A, B, C, D 및 E)를 포함하는 예시적 기계 학습 모델의 신경 네트워크 레이어일 수 있다. 기계 학습 모델에 의해 수행된 추론 계산은 피처 깊이 또는 출력 스트라이드가 갑자기 또는 예기치 않게 증가할 수 있다. 이러한 상황이 발생하면, 신경 네트워크 계산 프로세스의 주어진 포인트에서 활성 작업 세트는 시간에 따라 입력 및 출력 액티베이션 수량을 증가시키거나 입력 및 출력 액티베이션 수량을 감소시킬 수 있다.

[0050] 예를 들어, 도 2a에 도시된 바와 같이, 기계 학습 모델에 의해 프로세싱된 단일 배치 엘리먼트의 작업 세트는 레이어 A에서 발생하는 배치 프로세싱을 위해 단일 저장 유닛(204)을 필요로 할 수 있다. 주어진 작업 세트에 대해 프로세싱된 입력 액티베이션들의 증가는 레이어 B에서의 배치 프로세싱 동안 발생할 수 있다. 따라서, 기계 학습 모델은 레이어 A에서의 단일 저장 유닛(204)보다는 레이어 B에서의 배치 프로세싱 동안 8개의 저장 유닛들(204)의 사용을 요구할 수 있다. 추가로, 도 2a에 도시된 바와 같이, 레이어 C, D 및 E에서 프로세싱된 작업 세트들은 각각 2, 4 및 1개의 저장 유닛을 요구할 수 있다.

[0051] 일부 구현예에서, 입력/출력 액티베이션 수량 및 대응하는 필요한 저장 유닛들의 증가 또는 감소은 각각 상이한 수의 파라미터들 또는 가중치들을 갖는 신경 네트워크의 레이어들에 기초하여 발생할 수 있다. 따라서, 레이어 A에 대한 작업 세트는 레이어 B에 비해 더 적은 액티베이션들 및 파라미터들을 포함할 수 있고, 따라서 레이어 A에 대한 작업 세트는 더 많은 저장 리소스를 요구할 수 있는 레이어 B에 대한 더 큰 작업 세트에 비해 더 적은 저장 리소스를 요구할 수 있다.

[0052] 일부 구현예에서, 저장 유닛들(204)은 입력 메모리(104) 및 파라미터 메모리(106)의 메모리 리소스들에 대응할 수 있다. 예를 들어, 저장 유닛들(204)은 회로(100)의 하드웨어 회로의 상술된 전자 컴포넌트의 온칩 메모리와 연관된 정적 랜덤 액세스 메모리(SRAM)의 메모리 리소스들에 대응할 수 있다. 메모리(104, 106)를 포함하는 온칩 메모리 리소스는 고정 또는 임계 저장 용량을 가질 수 있다. 이 임계 저장 용량은 회로(100)의 오프 칩 메모리와 연관된 DRAM(Dynamic Random Access Memory) 리소스의 저장 용량보다 작거나 실질적으로 작을 수 있다. 상술한 바와 같이, 오프 칩 메모리는 상위 레벨 외부 제어 디바이스의 메모리일 수 있다.

[0053] 도 2b는 신경 네트워크의 주어진 레이어에 대한 다수의 배치 엘리먼트들의 프로세싱과 관련된 예시적 그래프(200B)를 도시한다. 그래프(200B)는 배치(212)의 각각의 배치 엘리먼트들과 연관된 작업 세트들의 입력을 저장하기 위한 저장 유닛들(208)의 제1 집합을 포함한다. 그래프(200B)는 배치(214)의 각각의 배치 엘리먼트들과 연관된 작업 세트들의 입력을 저장하기 위한 저장 유닛들(210)의 제2 집합을 더 포함한다.

[0054] 도 2b의 구현예에서, 2개 이상의 배치들은 각각 다수의 배치 엘리먼트들을 포함할 수 있으며, 즉 배치(212)는 적어도 하나의 개별 배치 엘리먼트 "0"을 가질 수 있고, 배치(214)는 적어도 하나의 개별 배치 엘리먼트 "1"을 가질 수 있다. 적어도 2개의 배치들(212, 214)의 프로세싱은 주어진 작업 세트의 상대 크기가 배치 크기의 팩터에 의해 증폭되게 할 수 있다. 예를 들어, 도 2b에 도시된 바와 같이, 각각의 레이어들(206)(레이어 A-레이어 E)에서의 작업 세트 크기는 대응하는 배치 크기들을 갖는 적어도 2개의 배치들(배치 212 및 배치 214)의 프로세싱 입력들에 기초하여 예를 들어 2배로 증폭될 수 있다.

[0055] 상기 논의된 바와 같이, 시스템 제어기는 입력들의 배치가 신경의 하나 이상의 레이어들을 통해 프로세싱되는 방식을 정의하는 신경 네트워크 스케줄링 프로세스 또는 정책을 구현하기 위한, 컴파일 타임 스케줄링 또는 다른 컴퓨팅 로직을 포함하도록 구성될 수 있다. 예를 들어, 회로(100)는 신경 네트워크 입력들의 배치를 수신하고, 시스템 제어기는 입력들이 배치의 각 입력에 대한 추론을 수행하기 위해 어떻게 프로세싱되어야 하는지에 대한 스케줄링 프로세스를 결정한다. 입력들의 프로세싱은 신경 네트워크로 하여금 신경 네트워크의 후속 레이어에 제공될 수 있는 입력 액티베이션들과 같은 중간 입력들을 생성하게 한다. 중간 입력들은 후속 신경 네트워크 레이어에 입력 액티베이션들로서 제공되는 제1 신경 네트워크 레이어의 출력 액티베이션들에 대응할 수 있다.

[0056] 종래의 스케줄링 정책에서, 신경 네트워크는 제1 신경 네트워크 레이어를 통해 배치에서 각 입력 또는 배치 엘리먼트를 프로세싱하여, 각 배치 엘리먼트에 대한 레이어 출력(출력 액티베이션)을 생성한다. 각 레이어 출력은 배치 내의 배치 엘리먼트들의 프로세싱이 완료될 때까지 제2 신경 네트워크 레이어를 통해 프로세싱된다. 즉, 주어진 레이어의 프로세싱은 신경 네트워크에서 다음 레이어에 대한 프로세싱이 발생하기 전에 배치의 모든 배치 엘리먼트에 대해 수행된다. 이러한 종래의 신경 네트워크 스케줄링 정책은 메모리 용량과 같은 제약에 의해 제한될 수 있고, 따라서 기계 학습 시스템의 가용 메모리 및 컴퓨팅 리소스의 사용을 최대화하는데 비효율적일 수 있다.

[0057] 일부 구현예에서, 예시적 하드웨어 회로의 온칩 메모리, 예를 들어, 메모리(104, 106)의 저장 유닛(204)의 사용

과 관련하여, 온칩 메모리 리소스에 의해 지원될 수 있는 최대 배치 크기는 작업 세트의 크기에 기초하여 결정될 수 있다. 특히, 저장 유닛(204)에 의해 지원되는 최대 배치 크기는 주어진 신경 네트워크 레이어에 의해 프로세싱되는 입력들 및 파라미터들의 최대 작업 세트에 부분적으로 기초하여 결정될 수 있다.

[0058] 예를 들어, 도 2b를 참조하면, 메모리(102 및 104)와 연관된 총 온칩 저장 용량은 20개의 저장 유닛(204)으로 제한될 수 있다. 도 2b에서, 레이어 B에 의해 프로세싱된 2개의 배치 엘리먼트들의 작업 세트는 16개의 저장 유닛(204)을 필요로 하기 때문에, 제3 배치 엘리먼트의 프로세싱은 24개의 저장 유닛(204)을 필요로 하므로, 20개의 저장 유닛 용량을 초과할 것이다. 따라서, 이 예에서, 신경 네트워크는 각 배치 엘리먼트를 프로세싱하는 것이 적어도 8개의 저장 유닛이 필요한 경우 2개의 배치 엘리먼트들을 포함하는 특정한 최대 작업 세트 크기만을 지원할 수 있다.

[0059] 구체적으로, 도 2b의 구현예에서, 작업 세트에서 배치 엘리먼트 "0"의 프로세싱은 참조 피처(208)에 의해 표시된 바와 같이 8개의 저장 유닛를 필요로 하고, 배치 엘리먼트 "1"의 프로세싱은 참조 피처(210)에 의해 표시된 바와 같이 8개의 저장 유닛를 필요로 한다. 따라서, 배치 엘리엄트들 0 및 1을 프로세싱하는 것은 집합적으로 16 개의 저장 유닛(204)을 필요로 하기 때문에, 4개 보다 많은 저장 유닛(204)을 필요로 하는 적어도 하나의 추가 배치 엘리먼트의 프로세싱은 신경 네트워크의 하드웨어 회로의 사용 가능한 메모리 리소스들의 온칩 저장 용량(여기서는 20개로 제한됨)을 초과할 것이다.

[0060] 도 3은 하나 이상의 슈퍼레이어들(308 및 310)을 형성하는 신경 네트워크의 다수의 레이어들(206) 중 배치 엘리먼트들의 프로세싱에 관한 예시적 그래프(300)를 도시하며, 여기서 슈퍼레이어(308)는 예를 들어 레이어 A, B 및 C를 포함한다. 그래프(300)는 각각의 배치 엘리먼트들(302)의 배치 엘리먼트들과 연관된 작업 세트들의 입력들 및 파라미터들을 저장하기 위한 저장 유닛들(304)의 제1 집합을 포함한다. 마찬가지로, 그래프(300)는 각각의 배치 엘리먼트들(302)의 배치 엘리먼트(1)과 연관된 작업 세트들의 입력들 및 파라미터들을 저장하기 위한 저장 유닛들(306)의 제2 집합을 더 포함하며, 이는 도 3에 회색으로 도시된다.

[0061] 표시된 바와 같이, 회로(100)는 회로(100)의 다른 컴포넌트들 또는 회로들에 비해 더 적은 온칩 또는 SRAM 저장 리소스들을 가질 수 있는 예시적 전자 컴포넌트 또는 하드웨어 회로를 포함할 수 있다. 그러나, 본 명세서에 설명된 바와 같이, 회로(100)는 사용가능한 온칩 메모리를 사용하여 계산 집약적 기계 학습 알고리즘을 실행하도록 구성될 수 있다. 이러한 경우에, 기계 학습 시스템의 신경 네트워크는 하드웨어 회로의 온칩 메모리의 저장 유닛(204)에 의해 지원될 수 있는 최소 또는 최대 배치 크기에 대한 불필요한 제약을 부과하지 않는 가속기 아키텍처를 포함할 수 있다.

[0062] 설명된 기술에 따르면, 개선된 신경 네트워크 스케줄링 프로세스는 회로(100)의 하드웨어 회로의 로컬 온칩 저장 리소스들의 사용을 통해 제공되는 배치 로컬성을 효율적으로 이용하기 위해 사용될 수 있다. 또한, 이 온칩 저장과 다른 로컬 컴퓨팅 리소스들을 사용하면, 사용가능한 대역폭을 최적화하고 대역폭 및 에너지에 민감한 컴퓨팅 환경에서 컴포넌트 에너지 소비를 절약할 수 있다. 또한, 이 온칩 스토리지 및 다른 로컬 리소스들의 사용은 신경 네트워크의 레이어들을 통한 입력들의 프로세싱 동안 하드웨어 회로에 의한 외부 통신을 최소화하는 역할을 할 수 있다.

[0063] 예를 들어, 위에서 간략히 언급된 바와 같이, 신경 네트워크를 구현하는 하드웨어 회로는 추론을 계산하기 위해 신경 네트워크에 의해 사용되는 신경 네트워크 입력들 및 파라미터들을 수신하기 위해 호스트 디바이스/외부 제어기와 외부적으로 통신할 수 있다. 이러한 외부 통신은 하드웨어 회로의 온칩 컴퓨팅 리소스를 사용을 요구할 수 있다. 따라서, 외부 통신은 하드웨어 회로의 가용 컴퓨팅 대역폭을 감소시키고, 시스템 지연을 증가시키고, 하드웨어 회로의 전자 컴포넌트들에 의한 에너지 소비를 증가시킬 수 있다.

[0064] 대역폭 및 에너지 소비와 관련된 이러한 제약을 고려하여, 본 명세서는 예시적 신경 네트워크 모델의 "배치" 및 "레이어" 차원들을 혼합하여 특정한 메모리 작업 세트들의 사용을 최적화하는 글로벌 스케줄링 정책 또는 프로세스를 설명한다. 특히, 설명된 기술의 구현예는 신경 네트워크에 의해 프로세싱되는 배치 엘리먼트들에 대한 활성 작업 세트의 크기를 최소화하기 위해 기계 학습 모델의 배치 및 레이어 차원을 활용하는 유연한 신경 네트워크 스케줄링 정책을 포함할 수 있다.

[0065] 예를 들어, 설명된 교시에 따른 개선된 신경 네트워크 스케줄링 프로세스는 온-칩 메모리(104, 106) 내의 파라미터들을 포함하는 작업 세트들의 저장이 온칩 메모리 리소스의 임계 저장 용량을 초과하지 않도록 활성 작업 세트가 사이징되도록 할 수 있다. 따라서, 여기에 설명된 방법들은 신경 네트워크에 의한 배치 엘리먼트 프로세싱의 효율적인 스케줄링을 가능하게 한다. 예를 들어, 입력들을 프로세싱하는데 사용되는 입력들 및 파라미터들

의 배치 크기 에 불필요한 제약을 부과하지 않는 방식으로 작업 세트로 하여금 하드웨어 회로의 온칩 스토리지에 저장되게 하는 스케줄링 정책에 기초하여 효율성을 실현할 수 있다.

[0066] 또한, 기술된 교시에 따른 개선된 스케쥴링 정책은 입력들 및 파라미터들 저장하기 위해 사용가능한 온칩 리소스들의 효율적 사용을 최대화하여, 오프-칩 리소스에 액세스하기 위한 외부 통신이 최소화될 수 있다. 온칩 리소스들의 효율적 사용 및 감소된 외부 통신은 사용가능한 시스템 대역폭이 증가하게 하고, 시스템 컴포넌트들의 에너지 소비를 전반적으로 감소시킬 수 있다.

[0067] 일부 구현예에서, 개선된 스케줄링 프로세스 또는 정책의 양태들은 소프트웨어 명령어들 또는 프로그램 코드를 사용하여 인코딩될 수 있다. 명령들은 회로(100)의 적어도 하나의 프로세서, 제어기(108)의 적어도 하나의 프로세서 또는 회로(100) 또는 제어기(108)의 예시적 하드웨어 회로의 적어도 하나의 프로세서 또는 둘 모두에 의해 실행 가능할 수 있다.

[0068] 도 4는 회로(100)를 사용하는 신경 네트워크의 슈퍼레이어들을 통해 신경 네트워크 입력들을 프로세싱하는 방법(400)에 대한 예시적 흐름도이다. 방법 또는 프로세스(400)는 신경 네트워크에 의한 배치 엘리먼트 프로세싱을 위한 개선된 스케줄링 정책에 대응한다. 블록(402)에서, 회로(100)는 시스템의 하드웨어 회로에서 신경 네트워크를 사용하여 프로세싱될 신경 네트워크 입력들의 배치를 수신한다. 신경 네트워크는 방향 그래프로 배열된 다수의 레이어들을 가지며, 각 레이어는 각각의 파라미터들의 세트를 갖는다. 상기 논의된 바와 같이, 일부 구현예에서, 회로(100)의 하드웨어 회로는 예시적 신경 네트워크 하드웨어 시스템의 호스트 인터페이스 디바이스 또는 상위 레벨 제어기로부터 입력들을 수신할 수 있다.

[0069] 블록(404)에서, 회로(100)는 상기 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정한다. 예를 들어, 회로(100)는 신경 네트워크 레이어의 슈퍼레이어들의 시퀀스로의 하나 이상의 파티션들을 결정하도록 구성된 컴파일러 로직을 포함하거나 그에 액세스할 수 있다. 대안적으로 또는 컴파일러 로직에 추가하여, 회로(100)는 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스들로의 하나 이상의 파티션들을 결정하도록 구성된 적어도 하나의 하드웨어 블록을 포함하거나 그에 액세스할 수 있다. 일부 구현예에서, 슈퍼레이어들의 시퀀스에서 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 방향 그래프의 파티션이다.

[0070] 블록(406)에서, 회로(100)는 시스템의 하드웨어 회로를 사용하여 신경 네트워크 입력들의 배치를 프로세싱한다. 일부 구현예에서, 하드웨어 회로를 사용하여 신경 네트워크 입력들의 배치를 프로세싱하는 단계는 슈퍼레이어의 레이어들에 대한 각각의 파라미터들의 세트를 메모리(106)에 로딩하는 단계를 포함할 수 있다. 일부 경우에, 슈퍼레이어의 레이어들에 대한 파라미터들은 슈퍼레이어들의 시퀀스에서 각 슈퍼레이어 각각에 대해 로딩된다. 추가로, 하드웨어 회로를 사용하여 신경 네트워크 입력들의 배치를 프로세싱하는 단계는 상기 배치에서 각 신경 네트워크 입력에 대해, 상기 신경 네트워크 입력에 대한 슈퍼레이어 출력을 생성하기 위해, 상기 하드웨어 회로의 상기 메모리에 있는 상기 파라미터들을 사용하여, 상기 슈퍼레이어 내의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력을 프로세싱하는 단계를 포함한다.

[0071] 시퀀스의 제1 슈퍼레이어에 대해, 슈퍼레이어에 대한 신경 네트워크 입력의 출력(예를 들어, 슈퍼레이어 입력)은 제1 슈퍼레이어 출력이다. 추가적으로, 상기 제1 슈퍼레이어 이후에 각 슈퍼레이어에 대한 상기 슈퍼레이어 입력은 상기 시퀀스에서 선행하는 슈퍼레이어에 의해 생성된 슈퍼레이어 출력이다. 일부 구현예에서, 신경 네트워크 입력들의 배치를 프로세싱하는 단계는 시퀀스에서 제1 슈퍼레이어의 모든 레이어들을 통한 입력들을 프로세싱한 다음, 배치 내의 모든 입력들이 신경 네트워크에서 슈퍼레이어들 모두를 통해 프로세싱될 때까지, 시퀀스의 각 후속 슈퍼레이어의 모든 레이어들을 통한 입력들을 프로세싱하는 단계를 포함한다.

[0072] 다시 도 3을 참조하면, 개선된 신경 네트워크 스케줄링 프로세스를 사용하는 경우, 하나의 배치 엘리먼트는 다수의 레이어들(308 및 310)에 대해 배치없는 방식으로 실행될 수 있다. 설명된 기술에 따르면, 다수의 레이어들(308)은 제1 슈퍼레이어를 형성할 수 있고, 다수의 레이어들(310)은 제1 슈퍼레이어와 상이한 제2 슈퍼레이어를 형성할 수 있다. 슈퍼레이어들을 형성하기 위해 파티셔닝된 다수의 레이어들의 그룹화는 도 4를 참조하여 아래에 보다 상세하게 설명된다.

[0073] 도 3에 도시된 바와 같이, 일부 구현예에서, 예시적 기계 학습 모델의 레이어 B는 더 작은 작업 세트가 프로세싱되는 레이어 C에서 요구되는 저장 유닛의 양에 비해 큰 작업 세트를 프로세싱하기 위해 많은 양의 저장 유닛(204)을 요구할 수 있다. 배치 엘리먼트에 대한 작업 세트가 충분히 작은 경우, 개선된 스케줄링 프로세스는 슈퍼레이어/레이어(308)와 같은 다수의 레이어들(예를 들어, 슈퍼레이어)의 특정한 그룹화에 의해 프로세싱되는 다음 배치 엘리먼트로 스위칭하는 기계 학습 모델을 포함할 수 있다.

[0074] 예를 들어, 회로(100)의 하드웨어 회로 상에 구현된 신경 네트워크는 신경 네트워크의 "배치" 및 "레이어" 차원에 대해 글로벌 스케줄링을 수행하도록 구성될 수 있다. 특히, 신경 네트워크 레이어에 대한 입력들의 배치 프로세싱은 제1 프로세스 반복에서 엘리먼트 0의 제1 배치에 대한 레이어들의 그룹(308)(A, B, C))을 실행하고, 그 다음 제2 프로세스 반복에서 엘리먼트 1의 제2 배치에 대한 레이어들의 동일한 그룹(308)(A, B, C))을 실행함으로써 수행될 수 있다.

[0075] 도 3에 도시된 바와 같이, 개선된 스케줄링 정책에 따라 상이한 배치 엘리먼트들 사이에서 교번(alternating)하는 것은 전술한 종래의 스케줄링 정책의 최대 작업 세트 크기에 비해 작업 세트의 최대 크기를 감소시킨다. 예를 들어, 적어도 배치 엘리먼트 1에 대한 레이버 B에서의 배치 프로세싱과 관련하여, 상이한 배치 엘리먼트들 사이에서 교번하는 것은, 상기 기술된 종래의 스케줄링 정책을 사용하는 경우 요구되는 최대 작업 세트 크기가 16 유닛이 아니라 레이어 B의 최대 작업 세트 크기를 10 유닛으로 감소시킬일 수 있다. 예를 들어, 배치 엘리먼트 1에 대한 레이어 B에서 배치 프로세싱에 8개의 유닛이 사용될 수 있고, 2 유닛이 배치 엘리먼트 0의 레이어 A, B, C에서 이전의 배치 프로세싱의 출력 및/또는 레이어 D 및 E에서 프로세싱하기 위한 배치 엘리먼트 0과 연관된 작업 세트들의 입력들과 파라미터들을 저장하는데 사용될 수 있다.

[0076] 도 5는 슈퍼레이어들을 형성하기 위해 파티셔닝된 다수의 레이어들을 사용하여 적어도 단일 배치 엘리먼트를 프로세싱하기 위해 슈퍼레이어들의 시퀀스로 파티셔닝된 신경 네트워크 레이어들을 표현하는 예시적 그래프를 도시한다. 그래프(500)는 각각의 배치 엘리먼트들(502)의 배치 엘리먼트 0에 대한 작업 세트들의 입력들을 저장하기 위한 저장 유닛들(504)의 제1 집합을 포함한다.

[0077] 마찬가지로, 그래프(500)는 a) 각각의 배치 엘리먼트들(502)의 배치 엘리먼트 1에 대한 작업 세트의 입력들을 저장하기 위한 저장 유닛들의 제2 집합(506); b) 각각의 배치 엘리먼트(502)의 배치 엘리먼트 2에 대한 작업 세트들의 입력들을 저장하기 위한 저장 유닛들(508)의 제3 집합(508); 및 c) 각각의 배치 요소(502)의 배치 엘리먼트 3에 대한 작업 세트들의 입력들을 저장하기 위한 저장 유닛들의 제4 집합(510)을 더 포함한다.

[0078] 그래프(500)는 그래프의 X-축을 따라 슈퍼레이어들의 시퀀스를 더 포함한다. 예를 들어, 그래프(500)는 i) 레이어들 A, B, C 각각을 통해 배치 엘리먼트 0, 1, 2 및 3을 프로세싱하기 위한 제1 슈퍼레이어(512); 및 ii) 레이어들 D, E 각각을 통해 배치 엘리먼트 0, 1, 2 및 3을 프로세싱하기 위한 제2 슈퍼레이어(514)를 포함한다. 설명된 교시에 따르면, 개선된 신경 네트워크 스케줄링 정책에 기초하여 정의된 슈퍼레이어들의 시퀀스는 신경 네트워크를 실행하는 하드웨어 회로의 온칩 메모리 용량 또는 임계 용량을 초과하지 않고 상대적으로 높은 작업 세트 배치 크기를 지원할 수 있다.

[0079] 예를 들어, 도 5에 도시된 바와 같이, 예시적 "B3" 레이어 및 배치 페이즈 동안 입력들이 프로세싱되는 경우, 작업 세트의 최대 크기는 각각의 저장 유닛(204)의 쉐이트 패턴들을 구별함으로써 표시되는 바와 같이 4개의 배치 엘리먼트, 예를 들어 배치 엘리먼트 0, 1, 2 및 3에 대해 단지 14개의 저장 유닛을 필요로 할 수 있다. (예를 들어, 16개의 저장 유닛을 필요로 하는) 종래의 스케쥴링 프로세스와 비교할 때, 필요한 저장 유닛의 이러한 감소는 하드웨어 회로의 온-칩 메모리를 통해 수신 및 저장된 입력 및 파라미터의 로컬성을 개선하는 것을 가능하게 한다. 온칩 리소스의 이러한 개선된 활용은 오프 칩 또는 DRAM 메모리 리소스의 사용 감소에 부분적으로 기초하여 실현되는 증가된 대역폭 및 에너지 절약을 결과로 할 수 있다.

[0080] 또한, 위에서 간략히 언급된 바와 같이, 개선된 스케줄링 정책은 회로(100)의 하드웨어 회로의 온칩 메모리 용량을 초과하지 않고 하나 이상의 배치의 입력 또는 입력들을 프로세싱하는데 사용될 수 있다. 일부 구현예에서, 시퀀스에서 슈퍼레이어의 레이어들을 통해 신경 네트워크 입력들의 하나 이상의 배치들을 프로세싱하는 단계는 시퀀스에서 제1 슈퍼레이어(512)에 의해, 적어도 신경 네트워크의 후속 레이어에 의한 수신을 위해 제1 슈퍼레이어 출력을 후속 레이어에 대한 입력으로서 생성하는 것을 포함할 수 있다.

[0081] 일부 경우에, 슈퍼레이어들의 시퀀스에서 제2 슈퍼레이어에 대한 신경 네트워크 입력은 시퀀스에서 제1 슈퍼레이어에 의해 생성된 제1 슈퍼레이어 출력에 해당할 수 있다. 또한, 시퀀스에서 슈퍼레이어의 레이어들을 통해 입력들의 배치를 프로세싱하는 단계는 제1 슈퍼레이어 출력에 대등하는 신경 네트워크 입력에 대한 제2 슈퍼레이어 출력을 생성하기 위해, 하드웨어 회로의 메모리에서 파라미터들을 사용하여 제2 슈퍼레이어에서 레이어들 각각을 통해 신경 네트워크 입력을 프로세싱하는 것을 포함할 수 있다.

[0082] 일부 구현예에서, 슈퍼레이어들의 시퀀스에서 슈퍼레이어의 레이어들을 통해 신경 네트워크 입력들의 배치를 프로세싱하는 단계는 슈퍼레이어의 각 레이어를 통해 배치 엘리먼트 하나씩에 대한 입력들을 프로세싱하는 것을 포함할 수 있다. 예를 들어, 입력들의 배치를 프로세싱하는 것은 슈퍼레이어에서 레이어들 각각을 통해 2개 이

상의 신경 네트워크 입력들을 순차적으로 프로세싱하는 것을 포함할 수 있다. 이러한 순차적 프로세싱은 슈퍼레이어의 각 레이어를 통해 제1 신경 네트워크 입력을 프로세싱하고, 그 다음 슈퍼레이어의 각 레이어를 통해 제2 신경 네트워크 입력을 프로세싱하는 것을 포함할 수 있다.

[0083] 일부 구현예에서, 시퀀스에서 슈퍼레이어 각각에 대해, 슈퍼레이어의 레이어들을 통해 입력들을 프로세싱하는 단계는, 상기 배치에서 제2 신경 네트워크 입력에 대응하는 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 순차적으로 프로세싱되기 전에, 상기 배치에서 제1 신경 네트워크 입력에 대한 상기 슈퍼레이어 입력이 상기 슈퍼레이어의 상기 레이어들 각각을 통해 프로세싱되도록, 상기 슈퍼레이어의 상기 레이어들 각각을 통해 상기 신경 네트워크 입력들의 배치에 대응하는 상기 슈퍼레이어 입력들을 순차적으로 프로세싱하는 것을 포함한다.

[0084] 일부 구현예에서, 슈퍼레이어들의 시퀀스에서 제1 슈퍼레이어는 단일 신경 네트워크 레이어를 포함할 수 있다. 이 구현예에서, 슈퍼레이어들의 시퀀스를 통해 입력들을 프로세싱하는 것은 단일 신경 네트워크 레이어를 포함하는 제1 슈퍼레이어를 통해 제1 입력을 프로세싱하는 것을 포함할 수 있다. 이 제1 입력이 제1 슈퍼레이어의 단일 레이어를 통해 프로세싱된 후, 제1 입력이 시퀀스에서 제1 슈퍼레이어에 뒤따르는 후속 슈퍼레이어의 모든 레이어들을 통해 프로세싱되기 전에, 제1 슈퍼레이어에 의해 제2 입력이 즉시 프로세싱될 수 있다. 시퀀스에서 후속 슈퍼레이어에 의해 프로세싱되는 제1 입력은 단일 신경 네트워크 레이어를 포함하는 제1 슈퍼레이어의 슈퍼레이어 출력일 수 있다.

[0085] 개선된 신경 네트워크 스케줄링 정책에 따라, 레이어들의 파티셔닝 그룹에 기초하여 슈퍼레이어 및 하나 이상의 슈퍼레이어들의 시퀀스가 형성될 수 있다. 일부 구현예에서, 회로(100)는 개선된 스케줄링 정책을 위한 프로그래밍된 명령어들을 포함하고, 이들 명령어들은 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 것을 포함할 수 있다. 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 방향 그래프의 파티션일 수 있다.

[0086] 개선된 스케줄링 프로세스의 양태는 신경 네트워크 레이어들이 다수의 슈퍼레이어들로 형성되도록 하여, 주어진 슈퍼레이어에 대한 모든 입력들 및 파라미터들이 회로(100)의 하드웨어 회로의 온칩 스토리지로부터 액세스될 수 있도록 한다. 상기 지적한 바와 같이, 입력 및 파라미터에 대한 온칩 액세스는 하드웨어 회로에 의한 외부 통신을 최소화할 수 있다. 예를 들어, 하드웨어 회로는 오프 칩 인터페이스로부터 추가적인 양의 입력 및 파라미터를 얻기 위해 반복적인 페치 연산과 연관된 컴퓨팅 프로세스들을 피할 수 있기 때문에 외부 통신이 최소화될 수 있다.

[0087] 일부 구현예에서, 오프-칩 인터페이스는 하드웨어 회로를 회로(100)에 입력 및 파라미터를 제공하는 외부 제어 디바이스에 결합할 수 있다. 특히, 슈퍼레이어의 시퀀스에서 각 슈퍼레이어는 슈퍼레이어에 대한 하나 이상의 신경 네트워크 입력들을 프로세싱하기 위한 특정한 양의 파라미터들을 수신할 수 있다. 일부 경우에, 슈퍼레이어의 레이어들을 통해 하나 이상의 신경 네트워크 입력들을 프로세싱하는 것은 슈퍼레이어에 대한 특정한 양의 입력들을 프로세싱하기 위해 후속적인 양의 파라미터들을 수신하지 않고 입력을 프로세싱하는 것을 포함할 수 있다.

[0088] 일부 구현예에서, 회로(100)는 슈퍼레이어들의 시퀀스의 하나 이상의 슈퍼레이어 파티션들 또는 경계들을 결정하기 위해 프로그램 코드를 실행한다. 예를 들어, 회로(100)는 주어진 레이어에 대한 액티베이션 작업 세트 및 집계 파라미터 용량의 합을 결정하거나 계산할 수 있다. 회로(100)는 결정된 합을 사용하여 하드웨어 회로의 메모리 리소스들의 미리 정의된 또는 임계 온칩 저장 용량(예를 들어, 메모리(104 및 106))에 부분적으로 기초하여 신경 네트워크 레이어의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정할 수 있다. 따라서, 신경 네트워크 레이어들이 슈퍼레이어들의 시퀀스로 파티셔닝되어, 회로(100)의 하드웨어 회로가 상기 신경 네트워크 입력들의 배치를 프로세싱하는 경우 온칩 메모리의 임계 저장 용량을 초과하지 않도록 한다.

[0089] 일부 구현예에서, 신경 네트워크 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 단계는: i) 신경 네트워크에 의한 프로세싱을 위한 입력들을 포함하는 적어도 하나의 작업 세트에 대한 특정한 크기 파라미터를 결정하는 회로(100); ii) 하드웨어 회로의 메모리의 특정한 집계 입력 액티베이션 및 파라미터 용량을 결정하는 회로(100); 및 iii) 적어도 하나의 작업 세트에 대한 특정한 크기 파라미터 또는 하드웨어 회로의 메모리의 특정한 집계 입력 액티베이션 및 파라미터 용량에 적어도 기초하여 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하는 회로(100)를 포함한다.

[0090] 예를 들어, 온칩 메모리의 저장 용량 또는 임계 용량은 500MB일 수 있다. 회로(100)는 식 1[총 사용량=(작업 세

트 * N) + 파라미터들]에 기초하여, 총 온칩 메모리 사용량을 결정할 수 있고, 여기서 식 1의 변수 N은 배치 크기이다. 회로(100)는 신경 네트워크의 각 레이어에 대한 각각의 파라미터들의 세트를 저장하는데 필요한 메모리의 양을 결정할 수 있다. 일부 구현예에서, 도 5를 참조하여, 회로(100)는: i) 레이어 A에 대한 파라미터들의 세트가 25MB의 메모리를 필요로 하고; ii) 레이어 B에 대한 파라미터들의 세트가 125MB의 메모리를 필요로 하고, iii) 레이어 C에 대한 파라미터들의 세트가 50MB의 메모리를 필요로 한다고 결정할 수 있다.

[0091] 따라서, 이 예에서, 회로(100)는 레이어 A, B 및 C에 대한 각각의 파라미터들의 세트에 대한 집계 메모리 사용량이 200MB임을 결정하고, 입력들을 저장하는 것에 사용하기 위해 300MB의 이용 가능한 온칩 메모리를 남긴다(예를 들어, 500MB 온칩 메모리에서 200MB의 집계 메모리 사용량을 뺌). 각 레이어 A, B, C에 대해, 회로(100)는 각각의 레이어들에 의해 프로세싱될 작업 세트의 입력들을 위한 특정한 크기 파라미터 및 작업 세트에 대한 대응하는 배치 크기를 결정할 수 있다. 작업 세트에 대한 입력들의 크기 파라미터 및 대응하는 배치 크기를 사용하여, 회로(100)는 메모리의 집계 액티베이션 및 파라미터 용량을 결정할 수 있다. 회로(100)는 메모리의 집계 액티베이션 및 파라미터 용량을 사용하여 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정할 수 있다.

[0092] 일부 구현예에서, 회로(100)는 식 1, 입력들의 크기 파라미터(예를 들어, 메모리 유닛들에서), 배치 크기 및 파라미터들에 대해 사용된 집계 메모리를 사용하여 레이어들의 하나 이상의 그룹들에 대한 총 온칩 메모리 사용량을 결정한다. 회로(100)는 레이어들의 각 그룹에 대한 총 메모리 사용량을 500MB 온칩 저장 용량과 비교할 수 있다. 그 다음, 회로(100)는 비교 결과에 기초하여, 슈퍼레이어들의 시퀀스를 형성하는 레이어들의 파티셔닝 또는 그룹화를 결정할 수 있다. 회로(100)는 레이어들의 슈퍼레이어들의 시퀀스로의 파티셔닝을 결정하여, 하드웨어 회로가 작업 세트들에 대한 신경 네트워크 입력들의 배치를 프로세싱하는 경우 온칩 메모리의 임계 저장 용량(500MB)을 초과하지 않도록 한다.

[0093] 도 6a는 신경 네트워크 레이어에 대한 액티베이션 작업 세트 크기를 표현하는 예시적 그래프(600A)를 도시하며, 도 6b는 신경 네트워크의 슈퍼레이어에 대한 액티베이션 작업 세트 크기를 표현하는 예시적 그래프(600B)를 도시한다. 전술한 바와 같이, 그리고 그래프(600A 및 600B)에 의해 표시된 바와 같이, 슈퍼레이어들로 배열되지 않은 신경 네트워크 레이어들에 대한 작업 세트들은 슈퍼레이어들로서 배열된 신경 네트워크 레이어들에 대한 작업 세트의 크기와 비교할 때 실질적으로 보다 큰 작업 세트 크기를 포함할 수 있다.

[0094] 예를 들어, 전술된 종래의 스케줄링 정책을 사용하는 배치 프로세싱을 위한 작업 세트는 수백만 개의 입력들을 포함하는 작업 세트 크기들을 결과로 할 수 있다. 이러한 대량의 입력들은 온칩 저장 유닛(204)이 입력들을 프로세싱하는데 사용되는 입력들 및 파라미터들을 저장하기 위해 사용될 때 하드웨어 회로의 온칩 메모리 리소스의 저장 또는 임계 용량을 초과할 수 있다. 반대로, 본 명세서에 기술된 개선된 스케줄링 정책에 기초하여, 슈퍼레이어 파티션들을 사용하는 배치 프로세싱 위한 작업 세트는 실질적으로 더 적은 입력들을 포함하는 작업 세트 크기를 결과로 할 수 있다. 온칩 메모리 용량을 초과하지 않도록 실질적으로 더 적은 양의 입력들이 온칩 저장 유닛(204)을 사용하여 효율적으로 저장될 수 있다.

[0095] 본 발명의 실시예들과 본 명세서에 기술된 기능적 동작들은 본 발명에 개시된 구조들 및 그들의 구조적 균등물들 또는 그들 중 하나 이상의 조합들을 포함하는, 디지털 전자회로에서, 유형적으로 수록된 컴퓨터 소프트웨어 또는 펌웨어에서, 컴퓨터 하드웨어에서 구현될 수 있다. 본 명세서에 기술된 본 발명의 실시예들은 하나 이상의 컴퓨터 프로그램들로서 구현될 수 있다. 즉, 데이터 프로세싱 장치에 의해 실행 또는 데이터 프로세싱 장치의 동작을 제어하기 위한 유형적 비일시적인 프로그램 캐리어에 인코딩된 컴퓨터 프로그램 명령어들의 하나 이상의 모듈들.

[0096] 대안적으로 또는 추가로, 프로그램 명령어들은 데이터 프로세싱 장치에 의해 실행하기 위한 적절한 수신기 장치에 전송하기 위한 정보를 인코딩하기 위해 생성된 인공적으로 생성된 전파된 신호 즉, 기계-생성 전기, 광학 또는 전자기적 신호에 인코딩될 수 있다. 컴퓨터 저장 매체는 기계 판독가능 저장 디바이스, 기계 판독가능 저장 기판, 랜덤 또는 직렬 엑세스 메모리 디바이스 또는 그들 중 하나 이상의 조합일 수 있다.

[0097] 본 명세서에 기술된 프로세스들 및 논리 흐름들은 입력 데이터를 동작하고 출력(들)을 생성함으로써 기능들을 수행하기 위해 하나 이상의 컴퓨터 프로그램들을 실행하는 하나 이상의 프로그래머블 컴퓨터들에 의해 수행될 수 있다. 프로세스들 및 논리 흐름들은 또한 FPGA(field programmable gate array), ASIC(application specific integrated circuit), GPGPU(General purpose graphics processing unit)와 같은 특수 목적 논리 회로 또는 일부 다른 프로세싱 유닛에 의해 수행될 수 있고, 장치는 또한 특수 목적 논리 회로 또는 일부 다른 프로세싱 유닛으로서 구현될 수 있다.

[0098] 컴퓨터 프로그램의 실행에 적절한 컴퓨터들은 예시로서, 일반적 또는 특수 목적 마이크로프로세서들 또는 둘 모두, 또는 임의의 기타 종류의 중앙 프로세싱 유닛을 포함하거나 이에 기초할 수 있다. 일반적으로, 중앙 프로세싱 유닛은 읽기-전용 메모리 또는 랜덤 액세스 메모리 또는 둘 모두로부터 명령어들 및 데이터를 수신할 것이다. 컴퓨터의 필수 엘리먼트들은 명령어들을 수행하거나 실행하기 위한 중앙 프로세싱 유닛 및 명령어들 및 데이터를 저장하기 위한 하나 이상의 메모리 디바이스들이다. 일반적으로, 컴퓨터는 데이터를 저장하기 위한 하나 이상의 대형 저장 디바이스들 예를 들면, 자기적, 자기-광학 디스크들 또는 광학적 디스크들 또한 포함하거나 또는 그로부터 데이터를 수신하거나 그에 데이터를 전송하기 위해 동작적으로 결합될 수 있다. 그러나, 컴퓨터는 상기 디바이스들을 반드시 가져야하는 것은 아니다.

[0099] 컴퓨터 프로그램 명령어들 및 데이터를 저장하기에 적합한 컴퓨터 판독가능 매체는 예를 들어, EPROM, EEPROM 및 플래시 메모리 디바이스들과 같은 반도체 메모리 디바이스들; 예를 들어, 내부 하드 디스크들 또는 이동식 디스크들과 같은 자기 디스크들을 포함하는 모든 형태의 비휘발성 메모리, 매체 및 메모리 디바이스들을 포함한다. 프로세서 및 메모리는 특수 목적 논리 회로에 의해 보충되거나 그 안에 통합될 수 있다.

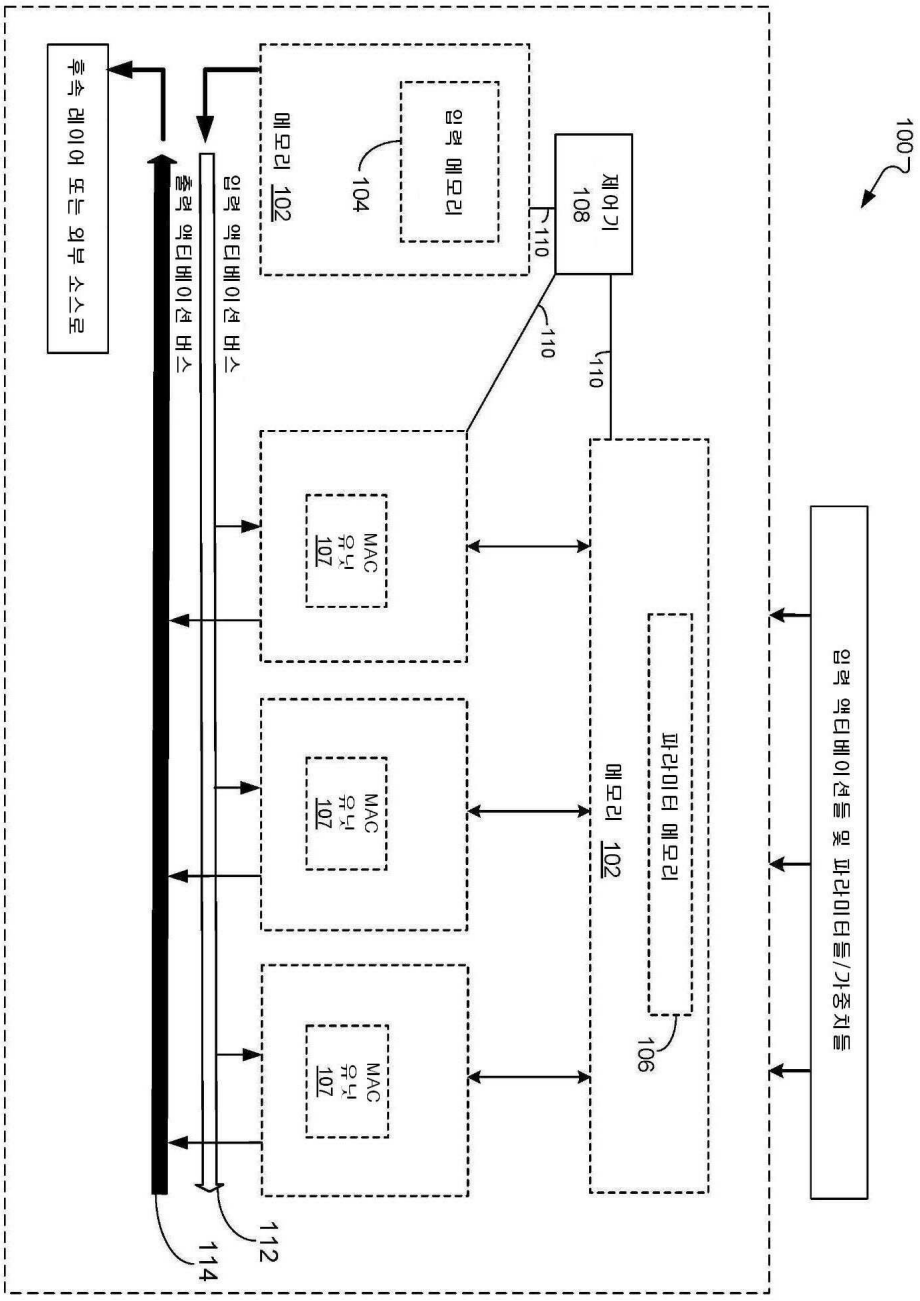
[0100] 본 명세서는 많은 특정 구현 세부내용을 포함하지만, 이들은 임의의 발명의 범위 또는 청구될 수 있는 범위에 대한 제한으로서 해석되어서는 안되며, 오히려 특정한 발명의 특정한 실시예에 특정적일 수 있는 구성들에 대한 설명으로 해석되어야 한다. 별개의 실시예의 맥락에서 본 명세서에서 기술되는 일정 구성들은 또한 단일 실시예에서 조합하여 구현될 수 있다. 반대로, 단일 실시예의 맥락에서 기술된 다양한 구성들은 또한 다수의 실시예에서 개별적으로 또는 임의의 적합한 서브 조합으로 구현될 수 있다. 게다가, 구성들은 일정 조합으로 동작하고 심지어 초기적으로 그렇게 청구되는 것으로서 상기에서 기술될 수 있지만, 청구된 조합으로부터의 하나 이상의 구성들은 일부 경우, 조합으로부터 제거될 수 있고, 청구된 조합은 서브 조합 또는 서브 조합의 변형으로 안내될 수 있다.

[0101] 유사하게, 동작들이 특정한 순서로 도면에서 도시되었지만, 이는 상기 동작들이 도시된 특정한 순서로 또는 시계열적 순서로 수행되어야 함을 요구하는 것으로서 또는 모든 도시된 동작들이 수행되어야 하는 것으로 이해되어서는 안된다. 특정 환경에서, 멀티태스킹과 병렬 프로세싱은 이점이 있다. 게다가, 상기 기술된 실시예에서 다양한 시스템 모듈들 및 컴포넌트들의 분리는 모든 실시예에서 그러한 분리가 필요한 것으로서 이해되어서는 안되며, 일반적으로 기술된 프로그램 컴포넌트들 및 시스템들은 단일의 소프트웨어 제품에 함께 통합되거나 다수의 소프트웨어 제품들에 패키징될 수 있다고 이해되어야 한다.

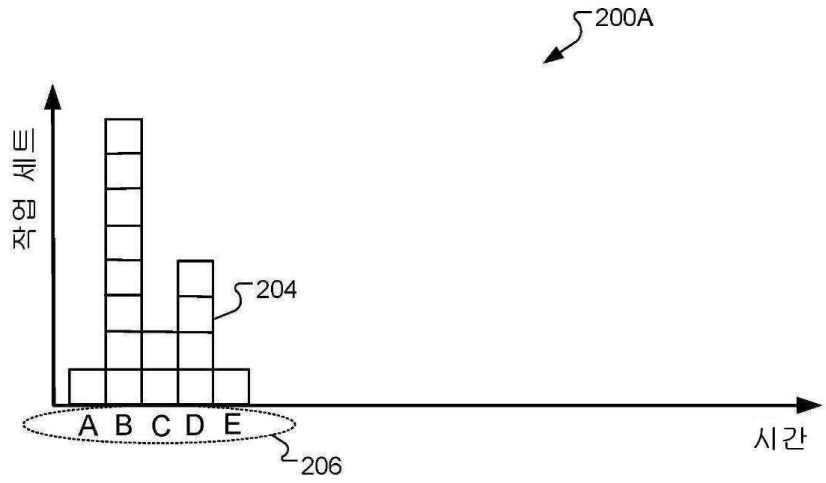
[0102] 본 발명의 특정한 실시예들이 기술되었다. 다른 실시예들도 다음의 청구항들의 범위 내에 있다. 예를 들면, 청구항들에서 기재된 액션들은 상이한 순서로 수행되고 여전히 원하는 결과들을 달성할 수 있다. 일 예시로서, 첨부 도면들에 도시된 프로세스들은 원하는 결과들을 달성하기 위해 특정한 도시된 순서, 또는 시계열적 순서를 반드시 필요로 하지 않는다. 특정 구현예에서, 멀티태스킹과 병렬 프로세싱은 이점이 있다.

도면

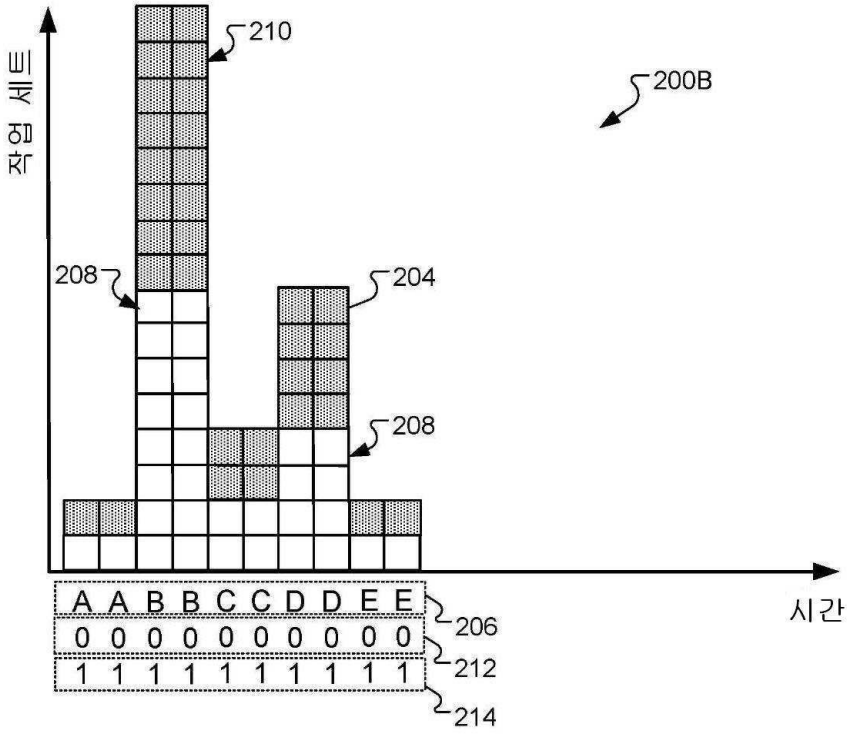
도면1


100
메모리 102
입력 메모리
104
제어기
108
110
110
입력 액티베이션 버스
출력 액티베이션 버스
후속 레이어 또는 외부 소스로
MAC 유닛 107
MAC 유닛 107
MAC 유닛 107
메모리 102
파라미터 메모리
106
입력 액티베이션 및 파라미터들/가중치들
112
114

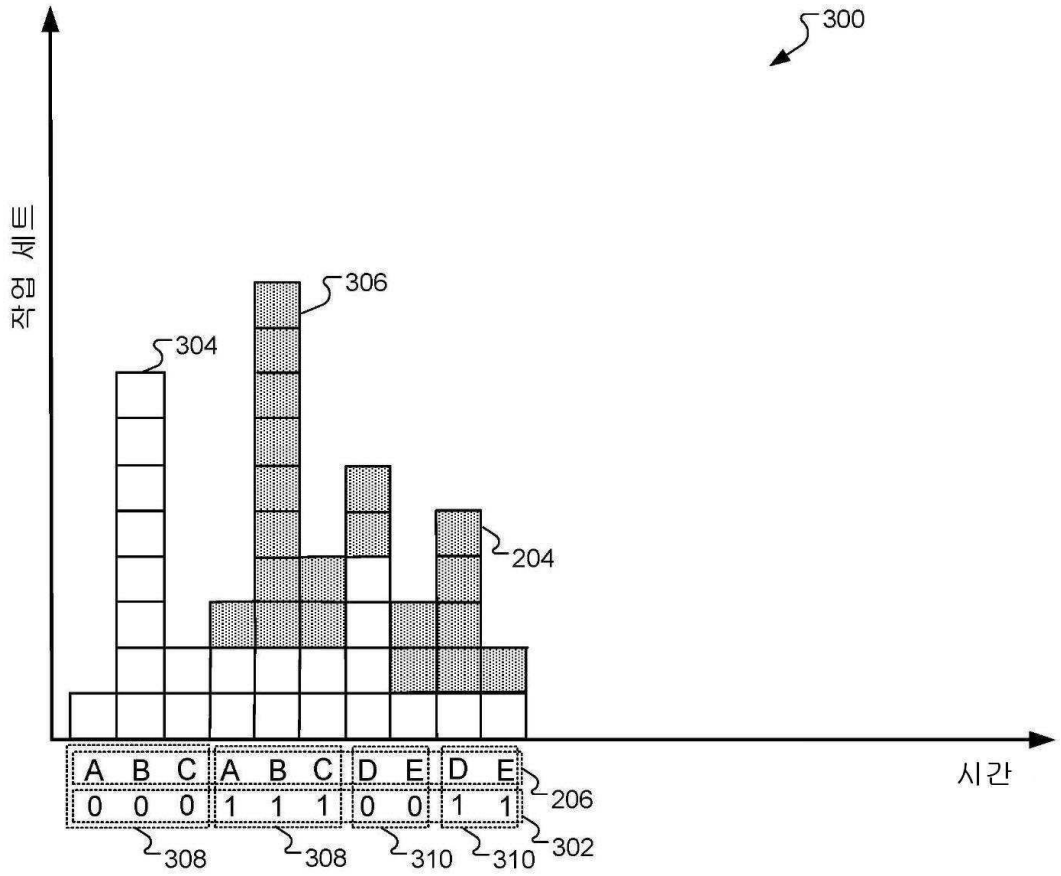
도면2a


200A
작업 세트
204
A B C D E
206
시간

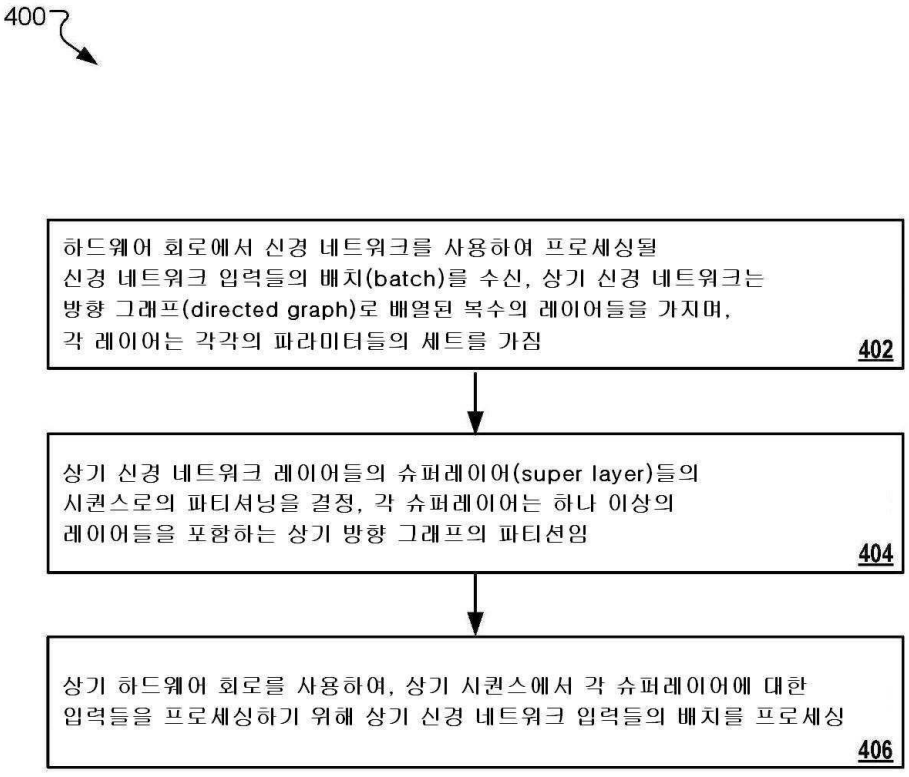
도면2b


210
200B
208
204
208
작업 세트
A A B B C C D D E E
206
0 0 0 0 0 0 0 0 0 0
212
1 1 1 1 1 1 1 1 1 1
214
시간

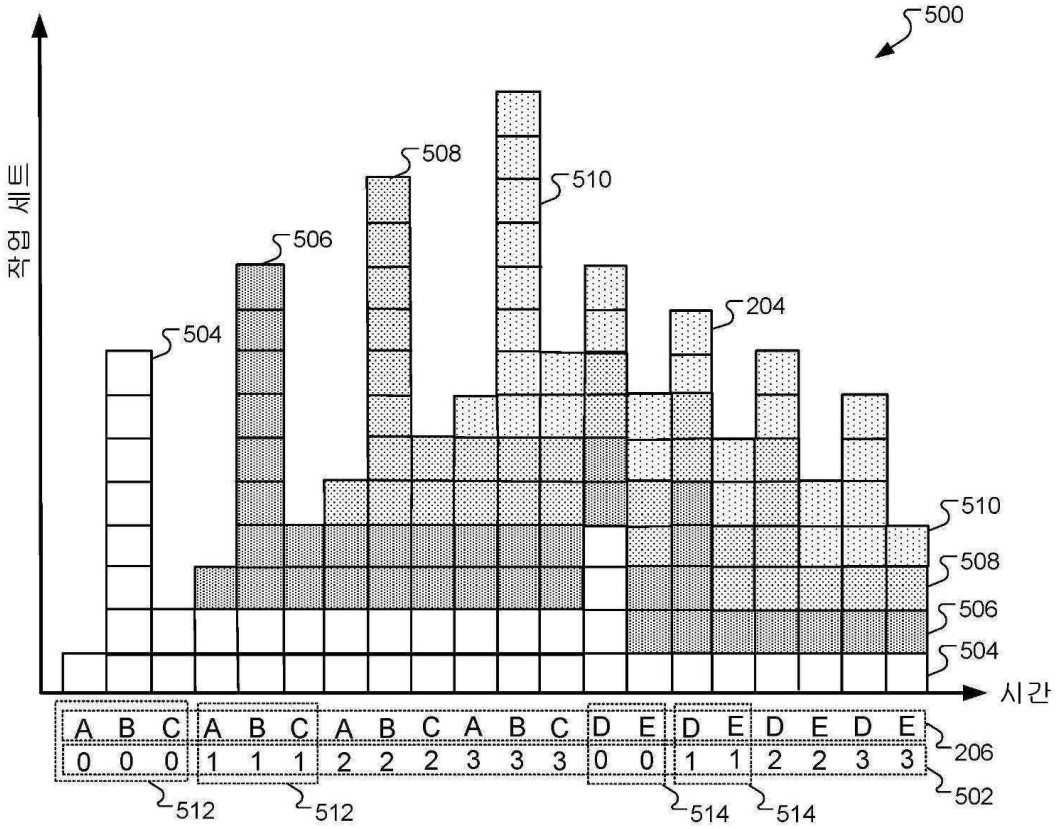
도면3


300
작업 세트
306
304
204
시간
A B C A B C D E D E
0 0 0 1 1 1 0 0 1 1
206
308
308
310
310
302

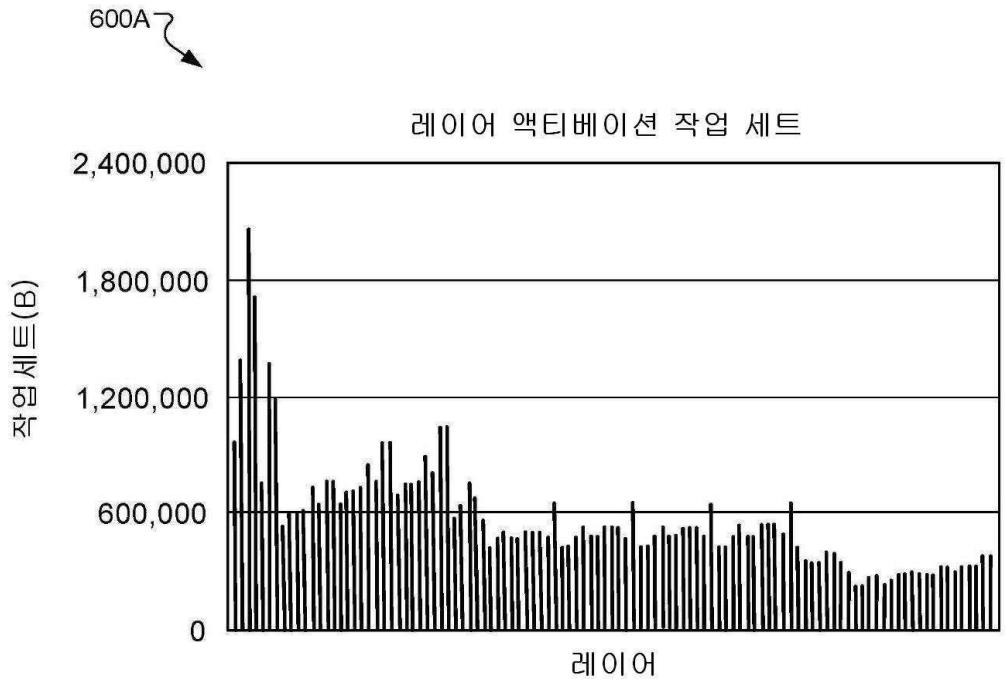
도면4


400
하드웨어 회로에서 신경 네트워크를 사용하여 프로세싱될 신경 네트워크 입력들의 배치(batch)를 수신, 상기 신경 네트워크는 방향 그래프(directed graph)로 배열된 복수의 레이어들을 가지며, 각 레이어는 각각의 파라미터들의 세트를 가짐
402
상기 신경 네트워크 레이어들의 슈퍼레이어(super layer)들의 시퀀스로의 파티셔닝을 결정, 각 슈퍼레이어는 하나 이상의 레이어들을 포함하는 상기 방향 그래프의 파티션임
404
상기 하드웨어 회로를 사용하여, 상기 시퀀스에서 각 슈퍼레이어에 대한 입력들을 프로세싱하기 위해 상기 신경 네트워크 입력들의 배치를 프로세싱
406

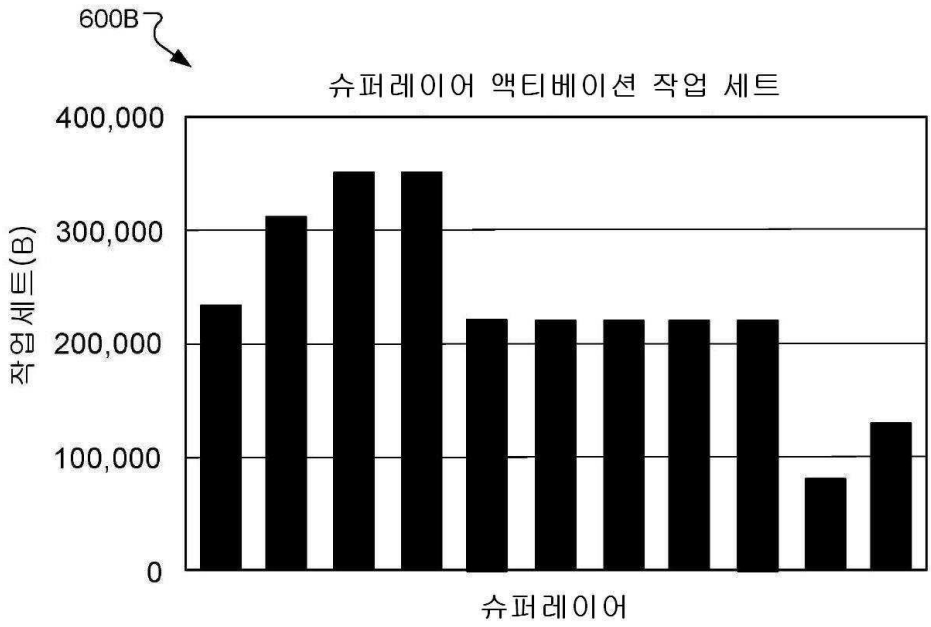
도면5


500
작업 세트
508
510
506
504
204
510
508
506
504
시간
A B C A B C A B C A B C D E D E D E D E
0 0 0 1 1 1 2 2 2 3 3 3 0 0 1 1 2 2 3 3
206
502
512
512
514
514

도면6a


600A
레이어 액티베이션 작업 세트
2,400,000
1,800,000
1,200,000
600,000
0
작업 세트(B)
레이어

도면6b


600B
슈퍼레이어 액티베이션 작업 세트
400,000
300,000
200,000
100,000
0
작업세트(B)
슈퍼레이어