(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0064044
(43) 공개일자 2021년06월02일

(51) 국제특허분류(Int. Cl.)
G06N 5/04 (2006.01) *G06N 3/04* (2006.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 5/04 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2020-0132458
(22) 출원일자 2020년10월14일
심사청구일자 2020년10월14일
(30) 우선권주장
1020190152092 2019년11월25일 대한민국(KR)

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
이경한
울산광역시 울주군 언양읍 유니스트길 50 울산과학기술원
빈경민
울산광역시 울주군 언양읍 유니스트길 50 울산과학기술원
(74) 대리인
제일특허법인(유)

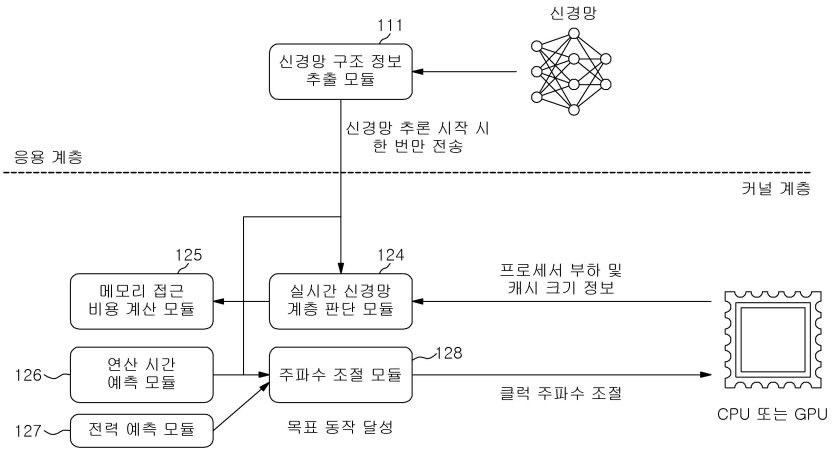
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치 및 방법**

(57) 요 약

본 발명에 따르면, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행 시 주파수 별로 필요한 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정하여 모바일 단말기에서 인공 신경망 추론을 수행할 시에 연산량에 따라 프로세싱 유닛의 클럭 주파수를 조절하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치가 개시된다.

대 표 도



(52) CPC특허분류
G06N 3/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711081170
과제번호	2017-0-00667-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	디지털콘텐츠원천기술개발(R&D)
연구과제명	이동 객체에 대한 증강현실 서비스 실현을 위한 정보-영상 정합기술 연구
기 여 율	1/1
과제수행기관명	울산과학기술원
연구기간	2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

모바일 단말기에 마련되며, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행하기 위해 상기 인공 신경망의 구조 정보를 추출하는 구조 정보 추출부; 및

프로세싱 유닛에 대한 부하 정보와 캐시 크기 정보를 입력 받고, 추출한 상기 인공 신경망의 구조 정보와 상기 부하 정보 및 캐시 크기 정보를 기반으로 기 지정된 목표 동작들 중 하나에 대하여 상기 프로세싱 유닛의 클럭 주파수를 결정하는 주파수 조절부;를 포함하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 2

제1항에 있어서,

상기 인공 신경망의 구조 정보는, 상기 인공 신경망 추론을 수행하는 연산에 필요한 행렬 크기를 포함하는 정보인 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 3

제1항에 있어서,

상기 구조 정보 추출부는, 상기 모바일 단말기의 어플리케이션 계층에서 구현되는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 4

제1항에 있어서,

상기 주파수 조절부는, 상기 모바일 단말기의 커널 계층에서 구현되는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 5

제1항에 있어서,

상기 주파수 조절부는, 상기 인공 신경망의 구조 정보와 상기 프로세싱 유닛에 대한 상기 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단하는 신경망 계층 판단부;를 포함하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 6

제5항에 있어서,

상기 주파수 조절부는, 상기 현재 연산중인 신경망 계층의 구조와 상기 캐시 크기 정보를 고려하여 상기 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 메모리 접근 비용 계산부;를 포함하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 7

제6항에 있어서,

상기 주파수 조절부는, 상기 메모리 접근 비용을 이용하여 상기 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 상기 주파수 별로 필요한 상기 프로세싱 유닛의 전력량을 산출하여, 상기 지정된 목표 동작들 중 하나에 대하여 상기 연산 시간과 상기 전력량에 따라 상기 프로세싱 유닛의 클럭 주파수를 결정하는 클럭 주파수 결정부;를 포함하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 8

제7항에 있어서,

상기 클럭 주파수 결정부는, 상기 지정된 목표 동작들 중 하나에 대하여 상기 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 상기 목표 주파수를 상기 클럭 주파수로 결정하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 9

모바일 단말기의 어플리케이션 계층에서 구현되며, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행 시 상기 인공 신경망에서 현재 연산중인 신경망 계층의 구조와 프로세싱 유닛의 캐시 크기 정보를 고려하여 상기 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 메모리 접근 비용 계산부; 및

상기 메모리 접근 비용을 이용하여 상기 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 상기 주파수 별로 필요한 상기 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 상기 연산 시간과 상기 전력량에 따라 상기 프로세싱 유닛의 클럭 주파수를 결정하는 클럭 주파수 결정부;를 포함하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 10

제9항에 있어서,

상기 클럭 주파수 결정부는, 상기 지정된 목표 동작들 중 하나에 대하여 상기 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 상기 목표 주파수를 상기 클럭 주파수로 결정하는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치.

청구항 11

프로세싱 유닛과 메모리를 포함하는 인공 신경망 추론을 수행하기 위한 장치의 인공 신경망 추론을 수행하기 위한 방법에 있어서,

상기 인공 신경망 추론을 수행하고자 하는 인공 신경망의 구조 정보를 추출하는 단계;

상기 인공 신경망의 구조 정보와 상기 프로세싱 유닛에 대한 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단하는 단계;

상기 현재 연산중인 신경망 계층의 구조와 캐시 크기 정보를 고려하여 상기 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 상기 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 단계; 및

상기 메모리 접근 비용을 이용하여 상기 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 상기 주파수 별로 필요한 상기 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 상기 연산 시간과 상기 전력량에 따라 상기 프로세싱 유닛의 클럭 주파수를 결정하는 단계;를 포함하여 수행되는 인공 신경망 추론을 수행하기 위한 방법.

청구항 12

제11항에 있어서,

상기 프로세싱 유닛의 클럭 주파수를 결정하는 단계는, 상기 지정된 목표 동작들 중 하나에 대하여 상기 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 상기 목표 주파수를 상기 클럭 주파수로 결정하는 인공 신경망 추론을 수행하기 위한 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 일반적으로 딥러닝 신경망은 큰 구조를 가질수록 인공 신경망의 성능이 좋지만 추론(Inference)시 연산량이 많

고 이에 따른 에너지 소모량이 많다.

[0003] 이러한 인공 신경망의 특성 때문에 상대적으로 하드웨어의 성능이 낮고 배터리 지속시간이 중요한 모바일 단말기에서 인공 신경망 추론이 필요할 시 서버로 데이터를 전송하여 인공 신경망 추론을 진행 후 결과를 모바일 단말기로 받아오는 방식으로 보완하고 있다.

[0004] 하지만 위와 같은 방식으로 인공 신경망 추론을 수행할 경우, 얼굴 인식 등과 같은 사용자의 개인정보를 이용한 인공 신경망 추론이 필요한 경우에는 보안상의 문제가 발생할 수 있고 네트워크 상황에 따른 데이터 전송 시간의 증가 등의 문제가 발생하므로 서버와 데이터를 주고 받는 형식이 아닌 모바일 단말기에서 직접 인공 신경망 추론을 하는 방식이 필요하다.

발명의 내용

해결하려는 과제

[0005] 본 발명의 일 실시예에 따른 해결하고자 하는 과제는, 모바일 단말기에서 인공 신경망 추론을 수행할 시에 연산량에 따라 프로세싱 유닛의 클럭 주파수를 조절하는 것을 포함한다.

[0006] 본 명세서에 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0007] 상기 과제를 해결하기 위해, 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치는, 모바일 단말기에 마련되며, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행하기 위해 상기 인공 신경망의 구조 정보를 추출하는 구조 정보 추출부 및 프로세싱 유닛에 대한 부하 정보와 캐시 크기 정보를 입력 받고, 추출한 상기 인공 신경망의 구조 정보와 상기 부하 정보 및 캐시 크기 정보를 기반으로 기 지정된 목표 동작들 중 하나에 대하여 상기 프로세싱 유닛의 클럭 주파수를 결정하는 주파수 조절부를 포함한다.

[0008] 여기서, 상기 인공 신경망의 구조 정보는, 상기 인공 신경망 추론을 수행하는 연산에 필요한 행렬 크기를 포함하는 정보이다.

[0009] 여기서, 상기 구조 정보 추출부는, 상기 모바일 단말기의 어플리케이션 계층에서 구현된다.

[0010] 여기서, 상기 주파수 조절부는, 상기 모바일 단말기의 커널 계층에서 구현된다.

[0011] 여기서, 상기 주파수 조절부는, 상기 인공 신경망의 구조 정보와 상기 프로세싱 유닛에 대한 상기 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단하는 신경망 계층 판단부를 포함한다.

[0012] 여기서, 상기 주파수 조절부는, 상기 현재 연산중인 신경망 계층의 구조와 상기 캐시 크기 정보를 고려하여 상기 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 메모리 접근 비용 계산부를 포함한다.

[0013] 여기서, 상기 주파수 조절부는, 상기 메모리 접근 비용을 이용하여 상기 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 상기 주파수 별로 필요한 상기 프로세싱 유닛의 전력량을 산출하여, 상기 지정된 목표 동작들 중 하나에 대하여 상기 연산 시간과 상기 전력량에 따라 상기 프로세싱 유닛의 클럭 주파수를 결정하는 클럭 주파수 결정부를 포함한다.

[0014] 여기서, 상기 클럭 주파수 결정부는, 상기 지정된 목표 동작들 중 하나에 대하여 상기 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 상기 목표 주파수를 상기 클럭 주파수로 결정한다.

[0015] 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치는, 모바일 단말기의 어플리케이션 계층에서 구현되며, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행 시 상기 인공 신경망에서 현재 연산중인 신경망 계층의 구조와 프로세싱 유닛의 캐시 크기 정보를 고려하여 상기 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 메모리 접근 비용 계산부 및 상기 메모리 접근 비용을 이용하여 상기 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 상기 주파수 별로 필요한 상기 프로세싱 유닛의 전력량을 산출하여, 상기 지정된 목표 동작들 중 하나에 대하여 상기 연산 시간과 상기 전력량에 따라 상기 프로세싱 유닛의 클럭 주파수를 결정하는 클럭 주파수

결정부를 포함한다.

[0016] 여기서, 상기 클럭 주파수 결정부는, 상기 지정된 목표 동작들 중 하나에 대하여 상기 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 상기 목표 주파수를 상기 클럭 주파수로 결정한다.

[0017] 본 발명의 또 다른 실시예에 따른 프로세싱 유닛과 메모리를 포함하는 인공 신경망 추론을 수행하기 위한 장치의 인공 신경망 추론을 수행하기 위한 방법에 있어서, 상기 인공 신경망 추론을 수행하고자 하는 인공 신경망의 구조 정보를 추출하는 단계, 상기 인공 신경망의 구조 정보와 상기 프로세싱 유닛에 대한 상기 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단하는 단계, 상기 현재 연산중인 신경망 계층의 구조와 상기 캐시 크기 정보를 고려하여 상기 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 상기 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 단계 및 상기 메모리 접근 비용을 이용하여 상기 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 상기 주파수 별로 필요한 상기 프로세싱 유닛의 전력량을 산출하여, 상기 지정된 목표 동작들 중 하나에 대하여 상기 연산 시간과 상기 전력량에 따라 상기 프로세싱 유닛의 클럭 주파수를 결정하는 단계를 포함하여 수행된다.

[0018] 여기서, 상기 프로세싱 유닛의 클럭 주파수를 결정하는 단계는, 상기 지정된 목표 동작들 중 하나에 대하여 상기 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 상기 목표 주파수를 상기 클럭 주파수로 결정한다.

발명의 효과

[0019] 이상에서 설명한 바와 같이 본 발명의 실시예들에 의하면, 모바일 단말기에서 인공 신경망 추론을 수행할 시에 연산량에 따라 프로세싱 유닛의 클럭 주파수를 조절하여 에너지 소모량을 낮출 수 있다.

[0020] 이에 따라, 모바일 단말기에서 직접 인공 신경망 추론을 수행할 수 있다.

[0021] 여기에서 명시적으로 언급되지 않은 효과라 하더라도, 본 발명의 기술적 특징에 의해 기대되는 이하의 명세서에서 기재된 효과 및 그 잠정적인 효과는 본 발명의 명세서에 기재된 것과 같이 취급된다.

도면의 간단한 설명

[0022] 도 1 및 도 2는 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치를 나타낸 블록도이다.

도 3 및 도 4는 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치를 나타낸 블록도이다.

도 5는 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 방법을 나타낸 흐름도이다.

도 6은 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 방법을 나타낸 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0023] 이하, 본 발명에 관련된 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치 및 방법에 대하여 도면을 참조하여 보다 상세하게 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0024] 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다.

[0025] 본 발명의 일 실시예는 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치 및 방법에 관한 것이다.

[0026] 도 1 및 도 2는 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치를 나타낸 블록도이다.

[0027] 도 1을 참조하면, 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(10)

는 구조 정보 추출부(110) 및 주파수 조절부(120)를 포함한다.

[0028] 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(10)는 모바일 단말기에서 사용자의 입력에 따라 인공 신경망 추론을 수행하는 경우 인공 신경망 구조에 따른 프로세싱 유닛의 클럭 주파수의 조절을 통해 에너지 효율적으로 인공 신경망 추론을 수행하기 위한 장치이다.

[0029] 여기서, 프로세싱 유닛은 CPU(Central Processing Unit), GPU(Graphic Processing Unit) 및 범용 그래픽 프로세싱 유닛(General Purpose Graphic Processing Unit, GPGPU)등을 포함하는 프로세서를 지칭한다.

[0030] 구조 정보 추출부(110)는 모바일 단말기에 마련되며, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행하기 위해 인공 신경망의 구조 정보를 추출한다.

[0031] 여기서, 적어도 하나의 사용자 입력은 사용자가 모바일 단말기를 이용하여 특정 처리를 수행하기 위한 지시를 의미하며, 예를 들어 얼굴 인식, 사진 인식 및 지문 인식 등을 포함할 수 있다. 얼굴 인식 등과 같은 사용자의 개인정보를 이용한 인공 신경망 추론이 필요한 경우 보안상의 문제가 발생할 수 있고 네트워크 상황에 따른 데이터 전송 시간의 증가 등의 문제로 서버와 데이터를 주고 받는 형식이 아닌 모바일 단말기에서 직접 신경망 추론을 하는 방식이 필요하게 된다.

[0032] 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(10)는 서버와 데이터를 주고 받는 형식이 아닌 모바일 단말기에서 직접 신경망 추론을 하는 방식을 이용하기 위해 인공 신경망 구조에 따라 연산량을 계산하여, 프로세싱 유닛의 클럭 주파수의 조절을 통해 에너지 효율적으로 인공 신경망 추론을 수행할 수 있다.

[0033] 구체적으로, 인공 신경망의 구조 정보는, 인공 신경망 추론을 수행하는 연산에 필요한 행렬 크기를 포함하는 정보이다.

[0034] 신경망 중에서 가장 대표적인 신경망인 합성곱 신경망에서 이미지의 중요한 특징을 추출해내는 역할을 하는 계층은 합성곱 계층(Convolutional Layer)이며, 합성곱(Convolution) 연산은 행렬 (3차원 또는 4차원) 곱셈 누산(Multiply-addition) 을 반복적으로 수행한다.

[0035] 합성곱(Convolution) 연산 시 합성곱 필터(Convolutional Filter)를 반복적으로 곱셈 누산(Multiply-addition)에 사용하는데 이 필터의 크기가 CPU 또는 GPU 의 캐시 크기보다 클 경우 CPU 또는 GPU 가 메모리에 접근하는 횟수가 증가하게 되어 클럭 주파수를 높이더라도 메모리에 접근하는 시간을 기다려야 되기 때문에 클럭 주파수를 높이는 만큼 연산 시간이 줄어들지 않게 되고 클럭 주파수를 높임에 따라 발생하는 에너지 소모량은 증가하게 된다.

[0036] 이에 따라, 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(10)는 합성곱(Convolution) 연산에 필요한 행렬 크기를 포함하는 정보를 추출하여 메모리에 접근하는 횟수를 미리 산출하여 클럭 주파수를 조절하게 된다.

[0037] 본 발명의 일 실시예에 따르면, 구조 정보 추출부(110)는 모바일 단말기의 어플리케이션 계층에서 구현되며, 주파수 조절부(120)는 모바일 단말기의 커널 계층에서 구현된다.

[0038] 주파수 조절부(120)는 신경망 계층 판단부(121), 메모리 접근 비용 계산부(122) 및 클럭 주파수 결정부(123)를 포함한다.

[0039] 주파수 조절부(120)는 프로세싱 유닛에 대한 부하 정보와 캐시 크기 정보를 입력 받고, 추출한 인공 신경망의 구조 정보와 부하 정보 및 캐시 크기 정보를 기반으로 기 지정된 목표 동작들 중 하나에 대하여 프로세싱 유닛의 클럭 주파수를 결정한다.

[0040] CPU를 예로 들어 설명하면, 안드로이드에서는 모바일 기기의 에너지 소모량을 줄이기 위해 DVFS(Dynamic Voltage & Frequency Scaling) 이라는 현재 CPU 에 요청된 일의 양에 따라 CPU의 클럭 주파수를 조절하는 기법을 사용하며, CPU의 클럭 주파수를 높일 경우 시간당 처리할 수 있는 연산량이 증가하지만 전력이 기하 급수적으로 높아져 에너지 소모량 역시 증가하고 반대로 클럭 주파수를 낮출 경우 연산량은 감소하지만 전력이 감소해 에너지 소모량 역시 감소하게 된다.

[0041] 여기서 각 CPU 에서 클럭 주파수를 조절하는 역할을 하는 주체를 Governor 라고 하는데, 이 Governor가 현재 CPU에 요청된 일이 많을 경우 CPU 클럭 주파수를 높여서 일을 빠르게 처리하고 요청된 일이 거의 없을 경우 낮은 주파수로 조절해 에너지 소모량을 낮추게 된다. 기존의 안드로이드 Governor는 사용자의 반응성에 맞추어 설

계되어 있기 때문에 이러한 신경망 연산의 특성을 잘 반영하고 있지 못하므로 에너지 소모량 측면에서 비효율적이다.

[0042] 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(10)의 주파수 조절부(120)는 프로세싱 유닛에 대한 부하 정보와 캐시 크기 정보를 입력 받고, 추출한 인공 신경망의 구조 정보와 부하 정보 및 캐시 크기 정보를 기반으로 기 지정된 목표 동작들 중 하나에 대하여 프로세싱 유닛의 클럭 주파수를 결정하도록 설계되므로, 인공 신경망 연산의 특성을 반영할 수 있다.

[0043] 신경망 계층 판단부(121)는 인공 신경망의 구조 정보와 프로세싱 유닛에 대한 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단한다.

[0044] 메모리 접근 비용 계산부(122)는 현재 연산중인 신경망 계층의 구조와 캐시 크기 정보를 고려하여 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산한다.

[0045] 클럭 주파수 결정부(123)는 메모리 접근 비용을 이용하여 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 주파수 별로 필요한 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정한다.

[0046] 클럭 주파수 결정부(123)는 지정된 목표 동작들 중 하나에 대하여 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 목표 주파수를 클럭 주파수로 결정한다.

[0047] 도 2를 참조하면, 어플리케이션 계층은 응용 계층을 지칭하며, 구조 정보 추출부(110)는 신경망 구조 정보 추출 모듈(111)을 포함한다.

[0048] 신경망 구조 정보 추출 모듈(111)은 응용 계층으로부터 신경망 추론 시작 전에 신경망 전체 구조에 대한 정보를 커널 계층에서 주파수 조절부가 필요로 하는 정보 형태로 추출하여 전달한다. 여기서, 주파수 조절부는 Governor를 포함한다.

[0049] 신경망 계층 판단부(121)는 실시간 신경망 계층 판단 모듈(124)을 포함하며, 인공 신경망의 구조 정보와 프로세싱 유닛에 대한 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단한다.

[0050] 본 발명의 일 실시예에 따르면, 응용 계층에서 신경망 구조 정보를 추출하고, 커널 계층으로 전달함에 따라, 응용 계층과 커널 계층 간에 정보를 실시간으로 주고 받음으로써 발생하는 지연 시간을 최소화할 수 있다.

[0051] 실시간 신경망 계층 판단 모듈(124)는 현재 프로세싱 유닛에 대기 중인 연산의 종류(데이터 읽기/쓰기, 곱/합 연산 등)와 연산의 양과 아직 연산이 진행되지 않은 신경망 구조 정보(연산에 필요한 행렬 크기 등)를 비교하여 현재 연산중인 신경망 계층을 판단한다.

[0052] 메모리 접근 비용 계산부(122)는 메모리 접근 비용 계산 모듈(125)을 포함하며, 현재 연산중인 신경망 계층의 구조와 캐시 크기 정보를 고려하여 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산한다.

[0053] 여기서, 메모리 접근 비용은 Last-level cache misses/instructions으로 계산된다. 즉, 연산한 명령 개수 대비 메인 메모리에 접근한 횟수를 의미한다.

[0054] 구체적으로, 메모리 접근 비용은 신경망 계층 구조와 캐시 크기에 따라 달라지므로 현재 연산중인 신경망 계층 구조 정보와 프로세싱 유닛의 캐시 크기를 이용하여 예측한다. 만일, 캐시 크기가 작다면 동일한 신경망 계층 구조이더라도 메모리 접근 비용이 증가하게 되며, 신경망 계층 구조가 많은 양의 파라미터가 필요하거나 연산해야 할 행렬 크기가 큰 구조라면 동일한 캐시 크기에서도 다른 계층에 비해 메모리 접근 비용이 증가한다.

[0055] 메모리 접근 비용 계산 모듈(125)은 현재 연산중인 신경망 계층에 대한 정보를 이용해 메모리 접근 비용을 계산하여 클럭 주파수 결정부(123)의 연산 시간 예측 모듈(126)과 전력 예측 모듈(127)에 전달한다.

[0056] 클럭 주파수 결정부(123)는 연산 시간 예측 모듈(126), 전력 예측 모듈(127) 및 주파수 조절 모듈(128)을 포함하며, 메모리 접근 비용을 이용하여 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 주파수 별로 필요한 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정한다.

[0057] 구체적으로, 연산 시간 예측 모듈(126)은 메모리 접근 비용을 이용해 현재 신경망의 연산에 필요한 연산 시간을 각 주파수별로 예측 후 주파수 조절 모듈(128)에 테이블 형태로 전달한다.

[0058] 여기서, 메모리 접근에 필요한 cycle 수는 모바일 기기에 따라 정해져 있기 때문에 전달 받은 메모리 접근 비용을 이용하여 연산 시간을 예측할 수 있다.

[0059] 예를 들어, 한 신경망 계층의 메모리 접근 비용이 0.01, 메모리 접근에 필요한 cycle 수가 10, 연산에 필요한 cycle 수가 1일 때, 신경망 연산에 필요한 명령의 개수가 10,000개라면 10,000 * 0.01 = 100 개의 명령이 메모리 접근에 필요한 명령의 개수이고 9,900 개의 명령이 연산의 개수라고 예측할 수 있다. 이에 따라 100 * 10 = 1000 cycle 이 메모리 접근에 필요한 cycle 회수이고 연산에 필요한 cycle 은 9900 * 1 = 9900 cycle 이라고 예측 가능하다. 프로세싱 유닛의 주파수는 시간당 cycle 의 회수이기 때문에 예측된 cycle 수와 주파수의 역수를 곱하게 된다면 연산 시간 예측이 가능하다.

[0060] 전력 예측 모듈(127)은 각 주파수별로 필요한 프로세싱 유닛의 전력량을 주파수 조절 모듈(128)에 전달한다.

[0061] 클럭 주파수 결정부(123)는 지정된 목표 동작들 중 하나에 대하여 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 목표 주파수를 클럭 주파수로 결정한다.

[0062] 지정된 목표 동작은 클럭 주파수를 효율적으로 조절하기 위해 설정된 동작이다. 구체적으로, 커널 계층과 응용 계층 모두에서, 주파수 조절 모듈(128)은 아래와 같은 지정된 목표 동작 중 하나를 달성하는 방향으로 클럭 주파수를 조절하여 프로세싱 유닛에 할당한다.

[0063] 목표 동작은 다음과 같은 제1 목표 동작 내지 제4 목표 동작을 포함할 수 있다.

[0064] 제1 목표 동작은 최고 속도 연산으로, 연산 시간 예측 모듈로부터 받은 주파수별 연산 시간에서 달성할 수 있는 가장 빠른 속도의 클럭 주파수를 지정한다. 이때, 가장 짧은 연산 시간을 갖는 주파수를 할당하게 된다.

[0065] 제2 목표 동작은 최저 에너지 연산으로, 전력 예측 모듈로부터 받은 주파수별 전력과 연산 시간 예측 모듈로부터 받은 주파수별 연산 시간에서 달성할 수 있는 가장 적은 에너지로 지정한다. 이때, 전력과 연산 시간을 곱하면 에너지가 되는데 이 값이 가장 적은 주파수를 할당하게 된다.

[0066] 제3 목표 동작은 목표 속도에 맞춘 최저 에너지 연산으로, 현재 남은 계층 수, 종류 및 구조와 목표 속도를 고려한 현재 계층에서 허용 가능한 속도의 주파수 중 가장 낮은 에너지의 주파수를 지정한다.

[0067] 제4 목표 동작은 목표 에너지에 맞춘 최고 속도 연산으로, 현재 남은 계층 수, 종류 및 구조와 목표 에너지를 고려한 현재 계층에서 허용 가능한 에너지의 주파수 중 가장 낮은 속도의 주파수를 지정한다.

[0068] 도 3 및 도 4는 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치를 나타낸 블록도이다.

[0069] 도 3을 참조하면, 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(20)는 메모리 접근 비용 계산부(210) 및 클럭 주파수 결정부(220)를 포함한다.

[0070] 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(20)는 모바일 단말기에서 사용자의 입력에 따라 인공 신경망 추론을 수행하는 경우 인공 신경망 구조에 따른 프로세싱 유닛의 클럭 주파수의 조절을 통해 에너지 효율적으로 인공 신경망 추론을 수행하기 위한 장치이다.

[0071] 여기서, 프로세싱 유닛은 CPU(Central Processing Unit), GPU(Graphic Processing Unit) 및 범용 그래픽 프로세싱 유닛(General Purpose Graphic Processing Unit, GPGPU)등을 포함하는 프로세서를 지칭한다.

[0072] 본 발명의 또 다른 실시예에 따르면, 메모리 접근 비용 계산부(210) 및 클럭 주파수 결정부(220)는 모바일 단말기의 어플리케이션 계층에서 구현된다.

[0073] 앞서 설명한 본 발명의 일 실시예에 따른 커널 계층에서 커널 주파수를 조절하는 점과 달리 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(20)는 클럭 주파수 결정부(220)가 어플리케이션 계층에 있기 때문에 현재 연산중인 신경망 계층을 바로 알 수 있다.

[0074] 메모리 접근 비용 계산부(210)는 모바일 단말기의 어플리케이션 계층에서 구현되며, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행 시 인공 신경망에서 현재 연산중인 신경망 계층의 구조와 프로세싱 유닛의 캐시 크기 정보를 고려하여 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산한다.

[0075] 클럭 주파수 결정부(220)는 메모리 접근 비용을 이용하여 현재 연산중인 신경망 계층의 연산 시간을 주파수 별

로 산출하고, 주파수 별로 필요한 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정한다.

[0076] 클럭 주파수 결정부(220)는 지정된 목표 동작들 중 하나에 대하여 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 목표 주파수를 클럭 주파수로 결정한다.

[0077] 도 4를 참조하면, 어플리케이션 계층은 응용 계층을 지칭하며, 메모리 접근 비용 계산부(210)는 메모리 접근 비용 계산 모듈(211)을 포함하여, 적어도 하나의 사용자 입력에 따라 인공 신경망 추론을 수행 시 인공 신경망에서 현재 연산중인 신경망 계층의 구조와 프로세싱 유닛의 캐시 크기 정보를 고려하여 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산한다.

[0078] 메모리 접근 비용 계산 모듈(211)은 현재 연산중인 신경망 계층에 대한 정보를 이용해 메모리 접근 비용을 계산하여 연산 시간 예측 모듈(221)과 전력 예측 모듈(222)에 전달한다.

[0079] 여기서, 메모리 접근 비용은 Last-level cache misses/instructions으로 계산된다. 즉, 연산한 명령 개수 대비 메인 메모리에 접근한 횟수를 의미한다.

[0080] 구체적으로, 메모리 접근 비용은 신경망 계층 구조와 캐시 크기에 따라 달라지므로 현재 연산중인 신경망 계층 구조 정보와 프로세싱 유닛의 캐시 크기를 이용하여 예측한다. 만일, 캐시 크기가 작다면 동일한 신경망 계층 구조이더라도 메모리 접근 비용이 증가하게 되며, 신경망 계층 구조가 많은 양의 파라미터가 필요하거나 연산해야 할 행렬 크기가 큰 구조라면 동일한 캐시 크기에서도 다른 계층에 비해 메모리 접근 비용이 증가한다.

[0081] 클럭 주파수 결정부(220)는 연산 시간 예측 모듈(221), 전력 예측 모듈(222) 및 주파수 조절 모듈(223)을 포함하며, 메모리 접근 비용을 이용하여 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 주파수 별로 필요한 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정한다.

[0082] 연산 시간 예측 모듈(221)은 메모리 접근 비용을 이용해 현재 신경망의 연산에 필요한 연산 시간을 각 주파수별로 예측 후 주파수 조절 모듈(223)에 테이블 형태로 전달한다. 연산 시간 예측 방법은 도 2에서 설명한 방법을 이용하여 예측한다.

[0083] 전력 예측 모듈(222)은 각 주파수별로 필요한 프로세싱 유닛의 전력량을 주파수 조절 모듈(223)에 전달한다.

[0084] 구체적으로, 프로세싱 유닛의 주파수별로 필요한 전력량은 미리 측정한 값을 토대로 테이블 형태로 저장한 뒤에 주파수별로 요청이 왔을 때 테이블에 저장된 값을 전달한다.

[0085] 주파수 조절 모듈(223)은 커널 계층으로부터 현재 프로세싱 유닛의 부하 정보를 받아 주파수 조절을 다음 페이지에서 제안하는 목표 동작을 달성하기 위해 동작한다.

[0086] 목표 동작은 다음과 같은 제1 목표 동작 내지 제4 목표 동작을 포함할 수 있다.

[0087] 제1 목표 동작은 최고 속도 연산으로, 연산 시간 예측 모듈로부터 받은 주파수별 연산 시간에서 달성할 수 있는 가장 빠른 속도의 클럭 주파수를 지정한다. 이때, 가장 짧은 연산 시간을 갖는 주파수를 할당하게 된다.

[0088] 제2 목표 동작은 최저 에너지 연산으로, 전력 예측 모듈로부터 받은 주파수별 전력과 연산 시간 예측 모듈로부터 받은 주파수별 연산 시간에서 달성할 수 있는 가장 적은 에너지로 지정한다. 이때, 전력과 연산 시간을 곱하면 에너지가 되는데 이 값이 가장 적은 주파수를 할당하게 된다.

[0089] 제3 목표 동작은 목표 속도에 맞춘 최저 에너지 연산으로, 현재 남은 계층 수, 종류 및 구조와 목표 속도를 고려한 현재 계층에서 허용 가능한 속도의 주파수 중 가장 낮은 에너지의 주파수를 지정한다.

[0090] 제4 목표 동작은 목표 에너지에 맞춘 최고 속도 연산으로, 현재 남은 계층 수, 종류 및 구조와 목표 에너지를 고려한 현재 계층에서 허용 가능한 에너지의 주파수 중 가장 낮은 속도의 주파수를 지정한다.

[0091] 도 5는 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 방법을 나타낸 흐름도이다.

[0092] 도 5는 본 발명의 일 실시예에서 구조 정보 추출부(110)가 모바일 단말기의 어플리케이션 계층에서 구현되며, 주파수 조절부(120)가 모바일 단말기의 커널 계층에서 구현되는 경우에 인공 신경망 추론을 수행하기 위한 방법을 나타낸 것이다.

[0093] 도 5를 참조하면, 본 발명의 일 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 방법은 프로세싱 유닛과 메모리를 포함하는 인공 신경망 추론을 수행하기 위한 장치가 단계 S110에서 인공 신경망 추론을 수행하고자 하는 인공 신경망의 구조 정보를 추출한다.

[0094] 단계 S120에서 인공 신경망의 구조 정보와 프로세싱 유닛에 대한 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단한다.

[0095] 단계 S130에서 현재 연산중인 신경망 계층의 구조와 캐시 크기 정보를 고려하여 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산한다.

[0096] 단계 S140에서 메모리 접근 비용을 이용하여 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 주파수 별로 필요한 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정한다.

[0097] 프로세싱 유닛의 클럭 주파수를 결정하는 단계(S140)는, 지정된 목표 동작들 중 하나에 대하여 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 목표 주파수를 클럭 주파수로 결정한다.

[0098] 도 6은 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 방법을 나타낸 흐름도이다.

[0099] 도 6은 본 발명의 또 다른 실시예에서, 메모리 접근 비용 계산부(210) 및 클럭 주파수 결정부(220)는 모바일 단말기의 어플리케이션 계층에서 구현되는 경우에 인공 신경망 추론을 수행하기 위한 방법을 나타낸 것이다.

[0100] 앞서 설명한 본 발명의 일 실시예에 따른 커널 계층에서 커널 주파수를 조절하는 점과 달리 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치(20)는 클럭 주파수 결정부(220)가 어플리케이션 계층에 있기 때문에 현재 연산중인 신경망 계층을 바로 알 수 있다.

[0101] 도 6을 참조하면, 본 발명의 또 다른 실시예에 따른 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 방법은 프로세싱 유닛과 메모리를 포함하는 인공 신경망 추론을 수행하기 위한 장치가 단계 S210에서 현재 연산중인 신경망 계층의 구조와 캐시 크기 정보를 고려하여 현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산한다.

[0102] 단계 S220에서 메모리 접근 비용을 이용하여 현재 연산중인 신경망 계층의 연산 시간을 주파수 별로 산출하고, 주파수 별로 필요한 프로세싱 유닛의 전력량을 산출하여, 지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정한다.

[0103] 프로세싱 유닛의 클럭 주파수를 결정하는 단계(S220)는, 지정된 목표 동작들 중 하나에 대하여 주파수 별로 산출된 연산 시간들 중 가장 짧은 연산 시간을 갖는 목표 주파수를 선택하여 목표 주파수를 클럭 주파수로 결정한다.

[0104] 본 발명의 또 다른 실시예에 따르면, 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 방법을 실행하기 위한 프로그램을 기록한 컴퓨터 판독 가능한 기록매체를 제공할 수 있다.

[0105] 이상의 설명은 본 발명의 일 실시예에 불과할 뿐, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명의 본질적 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현할 수 있을 것이다. 따라서 본 발명의 범위는 전술한 실시예에 한정되지 않고 특허 청구 범위에 기재된 내용과 동등한 범위 내에 있는 다양한 실시 형태가 포함되도록 해석되어야 할 것이다.

부호의 설명

[0106] 10: 모바일 단말기에서 인공 신경망 추론을 수행하기 위한 장치

110: 구조 정보 추출부

120: 주파수 조절부

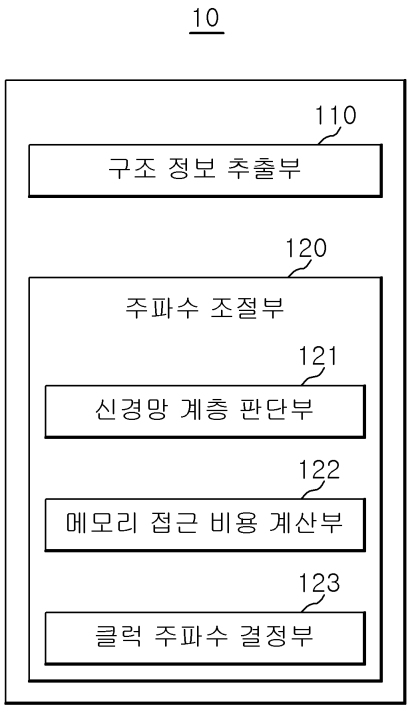
121: 신경망 계층 판단부

122: 메모리 접근 비용 계산부

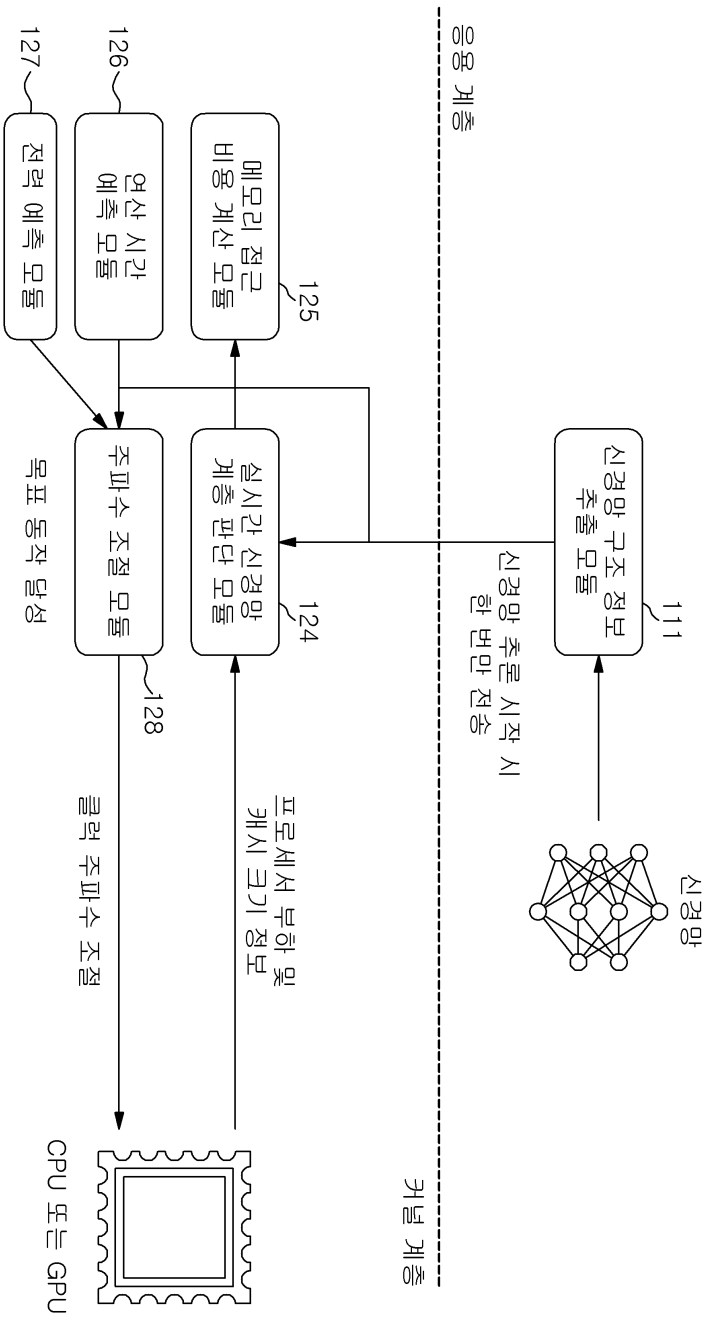
123: 클럭 주파수 결정부

도면

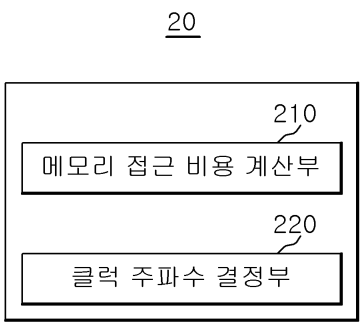
도면1



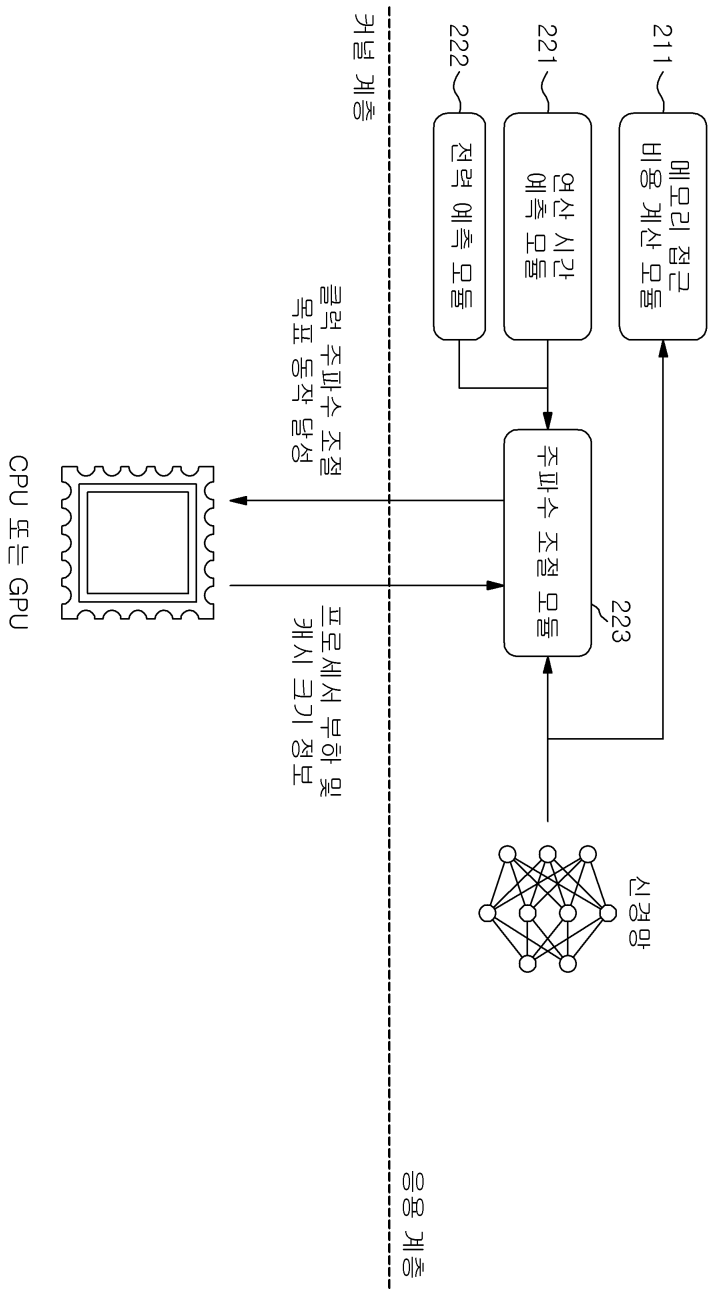
도면2


응용 계층
커널 계층
신경망 추론 시작 시 한 번만 전송
신경망 구조 정보 추출 모듈
111
신경망
실시간 신경망 계층 판단 모듈
124
프로세서 부하 및 캐시 크기 정보
메모리 접근 비용 계산 모듈
125
연산 시간 예측 모듈
126
전력 예측 모듈
127
주파수 조절 모듈
128
클럭 주파수 조절
목표 동작 달성
CPU 또는 GPU

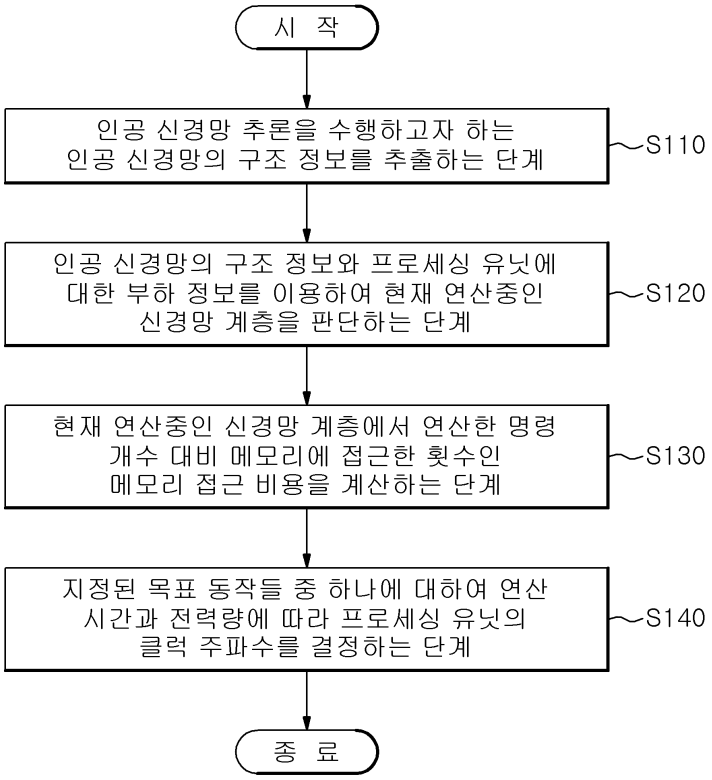
도면3



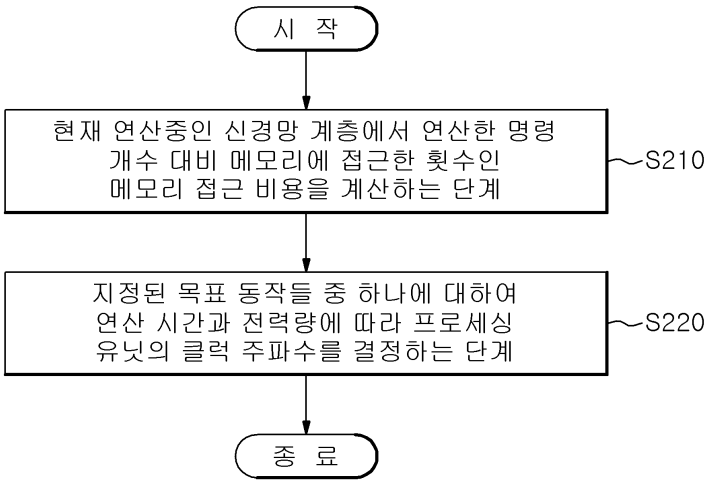
도면4


응용 계층
커널 계층
211
메모리 접근
비용 계산 모듈
221
연산 시간
예측 모듈
222
전력 예측 모듈
223
주파수 조절 모듈
신경망
프로세서 부하 및
캐시 크기 정보
클럭 주파수 조절
목표 동작 달성
CPU 또는 GPU

도면5


시 작
인공 신경망 추론을 수행하고자 하는 인공 신경망의 구조 정보를 추출하는 단계
S110
인공 신경망의 구조 정보와 프로세싱 유닛에 대한 부하 정보를 이용하여 현재 연산중인 신경망 계층을 판단하는 단계
S120
현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 단계
S130
지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정하는 단계
S140
종 료

도면6


시 작
현재 연산중인 신경망 계층에서 연산한 명령 개수 대비 메모리에 접근한 횟수인 메모리 접근 비용을 계산하는 단계
S210
지정된 목표 동작들 중 하나에 대하여 연산 시간과 전력량에 따라 프로세싱 유닛의 클럭 주파수를 결정하는 단계
S220
종 료