



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0107000
(43) 공개일자 2025년07월11일

(51) 국제특허분류(Int. Cl.)
H03M 7/30 (2006.01) G06N 3/04 (2023.01)
H03M 7/40 (2006.01)
(52) CPC특허분류
H03M 7/6017 (2013.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2024-0001656
(22) 출원일자 2024년01월04일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
심현욱
경기도 용인시 기흥구 흥덕1로62번길 7, 201호 (영덕동)
조현철
경기도 수원시
(뒷면에 계속)
(74) 대리인
특허법인 무한

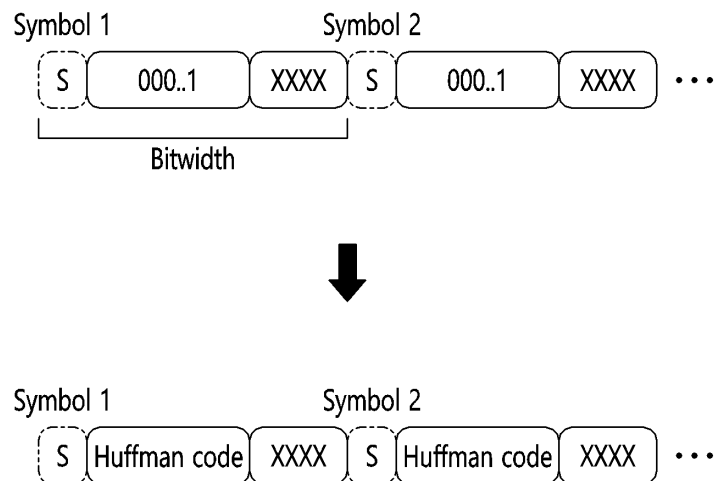
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 가중치의 인코딩 및 디코딩을 위한 인코더와 디코더 및 그 동작 방법

(57) 요약

실시예는, 가중치의 인코딩 및 디코딩을 위한 인코더와 디코더 및 그 동작 방법에 관한 것이다. 실시예에 디코딩 장치의 동작 방법은, 선행 코드 및 상기 선행 코드와 연결되는 후속 비트들로 이루어진 압축된 가중치를 수신하는 단계; 및 리딩 트리를 참조하여 상기 압축된 가중치의 선행 코드를 해시함으로써 상기 압축된 가중치를 디코딩하는 단계를 포함하고, 상기 리딩 트리는, 상기 선행 코드의 디코딩 정보를 포함할 수 있다.

대표도 - 도2



(52) CPC특허분류

H03M 7/405 (2013.01)

(72) 발명자

이수길

울산광역시 울주군 언양읍 유니스트길 50

이종은

울산광역시 울주군 언양읍 유니스트길 50

홍민욱

울산광역시 울주군 언양읍 유니스트길 50

명세서

청구범위

청구항 1

디코딩 장치의 동작 방법에 있어서,
선행 코드 및 상기 선행 코드와 연결되는 후속 비트들로 이루어진 압축된 가중치를 수신하는 단계; 및
리딩 트리를 참조하여 상기 압축된 가중치의 선행 코드를 해시함으로써 상기 압축된 가중치를 디코딩하는 단계
를 포함하고,
상기 리딩 트리는,
상기 선행 코드의 디코딩 정보를 포함하는,
디코딩 장치의 동작 방법.

청구항 2

제1항에 있어서,
상기 압축된 가중치를 디코딩하는 단계는,
상기 후속 비트의 값을 변환하지 않고 상기 해시된 코드에 연결시켜 표시하는 단계
를 포함하는,
디코딩 장치의 동작 방법.

청구항 3

제1항에 있어서,
상기 리딩 트리는,
상기 선행 코드의 모든 경우의 수에 매칭되는 디코딩 값을 기록하는,
디코딩 장치의 동작 방법.

청구항 4

제1항에 있어서,
상기 선행 코드는,
압축 이전의 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 개수에 따라 결정되는,
디코딩 장치의 동작 방법.

청구항 5

제1항에 있어서,
상기 압축된 가중치를 디코딩하는 단계는,

상기 압축된 가중치의 길이를 고정 길이 비트에 맞추도록 상기 압축된 가중치의 부족한 비트를 1로 채우는 단계를 포함하는,
디코딩 장치의 동작 방법.

청구항 6

제1항에 있어서,
상기 압축된 가중치를 수신하는 단계는,
SRAM에 저장된 상기 압축된 가중치를 수신하는 단계를 포함하는,
디코딩 장치의 동작 방법.

청구항 7

제1항에 있어서,
MAC 연산을 위해, 상기 디코딩된 가중치를 연산기로 입력하는 단계를 더 포함하는,
디코딩 장치의 동작 방법.

청구항 8

인코딩 장치의 동작 방법에 있어서,
가중치를 수신하는 단계;
상기 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 비트 수를 확인하는 단계;
상기 선행하는 0의 비트들을, 상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계; 및
상기 선행하는 0의 비트들 이후의 비트 값을 상기 압축된 코드에 연결하여 표시하는 단계를 포함하는,
인코딩 장치의 동작 방법.

청구항 9

제8항에 있어서,
상기 압축된 방식을 기록한 리딩 트리를 생성하는 단계를 더 포함하는,
인코딩 장치의 동작 방법.

청구항 10

제8항에 있어서,
상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계는,

최초의 비트가 1인 경우, 상기 최초의 비트 1을 압축하고, 상기 1 이후의 후속 비트의 비트 값을 상기 압축된 코드에 연결하여 표시하는 단계

를 포함하는,
인코딩 장치의 동작 방법.

청구항 11

제8항에 있어서,
상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계는,
상기 0의 비트들 중 첫 번째 비트를 제외한 비트들을 상기 0의 비트 수에 따라 미리 정해진 코드로 변환하는 단계
를 포함하는,
인코딩 장치의 동작 방법.

청구항 12

제8항에 있어서,
상기 선행하는 0의 비트 수를 확인하는 단계는,
상기 가중치의 비트 수에 대응하는 카운터를 생성하는 단계; 및
상기 카운터를 증가시키면서 상기 카운터에 해당하는 자리의 비트 값을 확인하는 단계
를 포함하는,
인코딩 장치의 동작 방법.

청구항 13

하드웨어와 결합되어 제9항 내지 제12항 중 어느 하나의 항의 방법을 실행시키기 위하여 컴퓨터 판독 가능 기록 매체에 저장된 컴퓨터 프로그램.

청구항 14

디코딩 장치에 있어서,
하나 이상의 프로세서;
메모리; 및
상기 메모리에 저장되어 있으며 상기 하나 이상의 프로세서에 의하여 실행되도록 구성되는 하나 이상의 프로그램 램을 포함하고,
상기 프로그램은,
선행 코드 및 상기 선행 코드와 연결되는 후속 비트들로 이루어진 압축된 가중치를 수신하는 단계; 및
리딩 트리를 참조하여 상기 압축된 가중치의 선행 코드를 해시함으로써 상기 압축된 가중치를 디코딩하는 단계
를 포함하고,
상기 리딩 트리는,

상기 선행 코드의 디코딩 정보를 포함하는,
디코딩 장치.

청구항 15

제14항에 있어서,
상기 압축된 가중치를 디코딩하는 단계는,
상기 후속 비트의 값을 변환하지 않고 상기 해시된 코드에 연결시켜 표시하는 단계
를 포함하는,
디코딩 장치.

청구항 16

제14항에 있어서,
상기 선행 코드는,
압축 이전의 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 개수에 따라 결정되는,
디코딩 장치.

청구항 17

제14항에 있어서,
상기 압축된 가중치를 디코딩하는 단계는,
상기 압축된 가중치의 길이를 고정 길이 비트에 맞추도록 상기 압축된 가중치의 부족한 비트를 1로 채우는 단계
를 포함하는,
디코딩 장치.

청구항 18

인코딩 장치에 있어서,
하나 이상의 프로세서;
메모리; 및
상기 메모리에 저장되어 있으며 상기 하나 이상의 프로세서에 의하여 실행되도록 구성되는 하나 이상의 프로그램
램을 포함하고,
상기 프로그램은,
가중치를 수신하는 단계;
상기 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 비트 수를 확인하는 단계;
상기 선행하는 0의 비트들을, 상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계; 및
상기 선행하는 0의 비트들 이후의 비트 값을 상기 압축된 코드에 연결하여 표시하는 단계
를 포함하는,

인코딩 장치.

청구항 19

제18항에 있어서,

상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계는,

최초의 비트가 1인 경우, 상기 최초의 비트 1을 압축하고, 상기 1 이후의 비트 값을 상기 압축된 코드에 연결하여 표시하는 단계

를 포함하는,

인코딩 장치.

청구항 20

제18항에 있어서,

상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계는,

상기 0의 비트들 중 첫 번째 비트를 제외한 비트들을 상기 0의 비트 수에 따라 미리 정해진 코드로 변환하는 단계

를 포함하는,

인코딩 장치.

발명의 설명

기술 분야

[0001] 실시예는, 가중치의 인코딩 및 디코딩을 위한 인코더와 디코더 및 이의 동작 방법에 관한 것이다.

배경 기술

[0003] 데이터 압축은 데이터 저장 공간, 통신 대역폭 등의 소비를 줄이기 위해 다양한 분야에서 이용된다.

[0004] 인공 신경망의 학습에서 압축 기법은 학습 효율 및 신경망의 성능과 직결된다. 인공 신경망의 학습 시, 메모리 통신량과 모델 가중치의 크기를 줄이게 되면, 에너지 효율이 크게 향상될 수 있다.

[0005] 압축 기법은 방식에 따라 손실 압축과 비손실 압축 방법으로 구분할 수 있고, 고정 길이 및 가변 길이의 방법으로 구분할 수 있다. 고정 길이로 압축하는 방식에 대비하여 가변 길이로 압축하는 경우, 어디까지 하나의 값을 표현하기 위한 것인지 알아내는 방법이 동반되어야 한다.

[0006] 이와 관련하여 런-LENGTH 인코딩(Run-length encoding)은 연속된 값의 개수를 추가로 인코딩하여 연속되는 값을 하나의 값과 개수로 표현하는 비손실 압축 방법이다.

발명의 내용

과제의 해결 수단

[0008] 디코딩 장치의 동작 방법에 있어서, 선행 코드 및 상기 선행 코드와 연결되는 후속 비트들로 이루어진 압축된 가중치를 수신하는 단계; 및 리딩 트리를 참조하여 상기 압축된 가중치의 선행 코드를 해시함으로써 상기 압축된 가중치를 디코딩하는 단계를 포함하고, 상기 리딩 트리는, 상기 선행 코드의 디코딩 정보를 포함하는, 디코딩 장치의 동작 방법이 제공될 수 있다.

- [0009] 상기 압축된 가중치를 디코딩하는 단계는, 상기 후속 비트의 값을 변환하지 않고 상기 해시된 코드에 연결시켜 표시하는 단계를 포함할 수 있다.
- [0010] 상기 리딩 트리는, 상기 선행 코드의 모든 경우의 수에 매칭되는 디코딩 값을 기록할 수 있다.
- [0011] 상기 선행 코드는, 압축 이전의 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 개수에 따라 결정될 수 있다.
- [0012] 상기 압축된 가중치를 디코딩하는 단계는, 상기 압축된 가중치의 길이를 고정 길이 비트에 맞추도록 상기 압축된 가중치의 부족한 비트를 1로 채우는 단계를 포함할 수 있다.
- [0013] 상기 압축된 가중치를 수신하는 단계는, SRAM에 저장된 상기 압축된 가중치를 수신하는 단계를 포함할 수 있다.
- [0014] MAC 연산을 위해, 상기 디코딩된 가중치를 연산기로 입력하는 단계를 더 포함할 수 있다.
- [0015] 인코딩 장치의 동작 방법에 있어서, 가중치를 수신하는 단계; 상기 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 비트 수를 확인하는 단계; 상기 선행하는 0의 비트들을, 상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계; 및 상기 선행하는 0의 비트들 이후의 비트 값을 상기 압축된 코드에 연결하여 표시하는 단계를 포함하는, 인코딩 장치의 동작 방법이 제공될 수 있다.
- [0016] 상기 압축된 방식을 기록한 리딩 트리를 생성하는 단계를 더 포함할 수 있다.
- [0017] 상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계는, 최초의 비트가 1인 경우, 상기 최초의 비트 1을 압축하고, 상기 1 이후의 후속 비트의 비트 값을 상기 압축된 코드에 연결하여 표시하는 단계를 포함할 수 있다.
- [0018] 상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계는, 상기 0의 비트들 중 첫 번째 비트를 제외한 비트들을 상기 0의 비트 수에 따라 미리 정해진 코드로 변환하는 단계를 포함할 수 있다.
- [0019] 상기 선행하는 0의 비트 수를 확인하는 단계는, 상기 가중치의 비트 수에 대응하는 카운터를 생성하는 단계; 및 상기 카운터를 증가시키면서 상기 카운터에 해당하는 자리의 비트 값을 확인하는 단계를 포함할 수 있다.
- [0020] 디코딩 장치에 있어서, 하나 이상의 프로세서; 메모리; 및 상기 메모리에 저장되어 있으며 상기 하나 이상의 프로세서에 의하여 실행되도록 구성되는 하나 이상의 프로그램을 포함하고, 상기 프로그램은, 선행 코드 및 상기 선행 코드와 연결되는 후속 비트들로 이루어진 압축된 가중치를 수신하는 단계; 및 리딩 트리를 참조하여 상기 압축된 가중치의 선행 코드를 해시함으로써 상기 압축된 가중치를 디코딩하는 단계를 포함하고, 상기 리딩 트리는, 상기 선행 코드의 디코딩 정보를 포함하는, 디코딩 장치가 제공될 수 있다.
- [0021] 인코딩 장치에 있어서, 하나 이상의 프로세서; 메모리; 및 상기 메모리에 저장되어 있으며 상기 하나 이상의 프로세서에 의하여 실행되도록 구성되는 하나 이상의 프로그램을 포함하고, 상기 프로그램은, 가중치를 수신하는 단계; 상기 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 비트 수를 확인하는 단계; 상기 선행하는 0의 비트들을, 상기 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 압축하는 단계; 및 상기 선행하는 0의 비트들 이후의 비트 값을 상기 압축된 코드에 연결하여 표시하는 단계를 포함하는, 인코딩 장치가 제공될 수 있다.

도면의 간단한 설명

- [0023] 도 1은 실시예에서, 허프만 코딩 방법을 설명하기 위한 도면이다.
- 도 2는 실시예에서, 압축 방식을 도식화하여 설명하기 위한 것이다.
- 도 3은 실시예에서, 인코딩 장치에서 가중치를 압축하는 방법을 설명하기 위한 흐름도이다.
- 도 4는 일 실시예에 따른 인코딩 장치의 블록도이다.
- 도 5는 실시예에서, 연산기에서 연산 가능한 형태로 가중치를 디코딩하는 실시예를 설명하기 위한 도면이다.
- 도 6은 실시예에서, 디코딩 장치의 동작 방법을 설명하기 위한 흐름도이다.
- 도 7은 일 실시예에 따른 디코딩 장치의 블록도이다.

도 8은 실시예에서, 스마트폰 AP로 구현되는 방법을 설명하기 위한 도면이다.

도 9는 실시예에서, NPU를 구현하는 방법을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0024] 이하에서, 첨부된 도면을 참조하여 실시예들을 상세하게 설명한다. 그러나, 실시예들에는 다양한 변경이 가해질 수 있어서 특허출원의 권리 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 실시예들에 대한 모든 변경, 균등물 내지 대체물이 권리 범위에 포함되는 것으로 이해되어야 한다.
- [0025] 실시예에서 사용한 용어는 단지 설명을 목적으로 사용된 것으로, 한정하려는 의도로 해석되어서는 안된다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0026] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0028] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.
- [0029] 또한, 실시 예의 구성 요소를 설명하는 데 있어서, 제1, 제2, A, B, (a), (b) 등의 용어를 사용할 수 있다. 이러한 용어는 그 구성 요소를 다른 구성 요소와 구별하기 위한 것일 뿐, 그 용어에 의해 해당 구성 요소의 본질이나 차례 또는 순서 등이 한정되지 않는다. 어떤 구성 요소가 다른 구성요소에 "연결", "결합" 또는 "접속"된다고 기재된 경우, 그 구성 요소는 그 다른 구성요소에 직접적으로 연결되거나 접속될 수 있지만, 각 구성 요소 사이에 또 다른 구성 요소가 "연결", "결합" 또는 "접속"될 수도 있다고 이해되어야 할 것이다.
- [0030] 어느 하나의 실시 예에 포함된 구성요소와, 공통적인 기능을 포함하는 구성요소는, 다른 실시 예에서 동일한 명칭을 사용하여 설명하기로 한다. 반대되는 기재가 없는 이상, 어느 하나의 실시 예에 기재한 설명은 다른 실시 예에도 적용될 수 있으며, 중복되는 범위에서 구체적인 설명은 생략하기로 한다.
- [0032] 도 1은 실시예에서, 허프만 코딩 방법을 설명하기 위한 도면이다.
- [0033] 허프만 코딩(Huffman Coding) 방법은 각각의 값의 배열에서 개별 값의 분포를 이용하여 손실 없이 압축하는 방법이다. 데이터에서 자주 출현하는 문자는 적은 비트의 코드로 변환하고, 출현 빈도가 낮은 문자는 많은 비트의 코드로 변환함으로써 전체 데이터를 표현하는 데 필요한 비트 수를 절약할 수 있다.
- [0034] 허프만 코딩 방식을 통해, 데이터는 도 1과 같은 허프만 트리를 구성하면서 압축될 수 있다.
- [0035] 인코더는, 데이터를 입력 받은 후, 데이터를 구성하는 비트열을 문자 단위로 분리하여 각 문자의 출현 빈도수를 내림차순으로 정렬하고 빈도가 낮은 두 문자를 묶어 이진 트리를 구성하고 합산된 문자 수를 기재한다. 나머지 문자들 중 빈도수가 낮은 문자와 다시 동일한 방식으로 이진 트리를 구성하고 문자수를 기재하는 동작을 반복하여 허프만 트리가 구성될 수 있다.
- [0036] 상위 노드에서부터 순회하면서 1과 0을 부여하여 허프만 코딩이 이루어질 수 있다. 결과적으로는 입력 받은 데이터가 해시맵에 대조하여 압축된 2진 비트열이 출력될 수 있다.
- [0037] 허프만 코딩은 가변 길이를 가지므로, 압축된 데이터에 순차적으로 해시해야 하고, 그 길이가 문자의 종류에 따라 복잡도가 기하급수적으로 증가하게 된다. 예를 들어 8비트의 비트열을 허프만 코딩으로 압축하는 경우의 수는 256개이고, 비트 당 최대 256비트 길이로 인코딩될 수 있고, 허프만 트리의 크기도 깊이도 깊어질 수

있다.

- [0039] 도 2는 실시예에서, 압축 방식을 도식화하여 설명하기 위한 것이다.
- [0040] 앞서 설명한 바와 같이, 허프만 코딩은 연속된 비트들의 배열에 따라 개별적인 코드로 인코딩하는 방식이다.
- [0041] 한편, 인코딩 장치는 수신된 가중치에서 선행되는 0의 숫자를 카운트하여, 0의 길이에 따라 압축되는 코드를 결정하는 방식으로 가중치를 압축할 수 있다.
- [0042] 인코더는, 수신된 가중치에서 '1'값을 가지는 데이터가 발생하기 이전에 '0'의 개수가 몇 개인지 카운트할 수 있다. 이렇게 '0'이 카운트된 비트들을 대응하는 코드로 압축하고, '1'이후의 후속 비트들의 값을 연결하는 (Concatenate) 방식으로 가중치가 압축될 수 있다.
- [0043] 예를 들어, 8비트의 가중치를 해당하는 방식으로 압축하면, 모든 비트가 '0'인 경우 및 모든 비트가 '1'인 경우를 포함하여 9가지의 경우의 수로 리딩 트리가 생성될 수 있다. 생성된 리딩 트리는 디코더 장치에서 디코딩하는 데에 활용될 수 있다.
- [0044] 고정 소수점 포맷의 예시에서 음수는 부호 비트 값이 '1'으로 표현되므로, 첫 번째 비트 값이 '1'이 나타나게 되고, 선행되는 1의 개수를 카운트하여 압축하는 방식이 적용되는 것이 설명된 인코딩 방식과 대등한 효과를 가져올 수 있다.
- [0045] 이하에서는 선행하는 0의 개수를 카운트하여 압축하는 방식에 대해서 설명하도록 한다.
- [0047] 도 3은 실시예에서, 인코딩 장치에서 가중치를 압축하는 방법을 설명하기 위한 흐름도이다.
- [0048] 이하 실시예에서 각 단계들은 순차적으로 수행될 수도 있으나, 반드시 순차적으로 수행되는 것은 아니다. 예를 들어, 각 단계들의 순서가 변경될 수도 있으며, 적어도 두 단계들이 병렬적으로 수행될 수도 있다.
- [0049] 인코딩 장치는 단계(310) 내지 단계(340)을 통해 인공 신경망의 가중치를 압축할 수 있다. 인코딩 장치는 인공 신경망에서 가중치를 전달할 때에 통신 효율을 높이고, 인공 신경망의 동작 속도를 높이기 위한 것이다.
- [0050] 단계(310)에서 인코딩 장치는, 가중치를 수신한다.
- [0051] 인코딩 장치는 W 길이의 대역폭을 가지는 가중치를 수신할 수 있다.
- [0052] 단계(320)에서 인코딩 장치는, 가중치에서 처음으로 1이 등장하기 이전까지 선행하는 0의 비트 수를 확인한다.
- [0053] 인코딩 장치는, W+1 개의 카운터를 만들어 0으로 초기화한 후, 카운터를 1씩 증가시키면서 해당하는 위치에 대한 비트 값을 확인할 수 있다. 예를 들어, 8비트의 가중치가 '001xxxxx'인 경우, 선행하는 0의 비트가 2비트임을 확인할 수 있다.
- [0054] 단계(330)에서 인코딩 장치는, 선행하는 0의 비트들을, 0의 비트 수에 따라 미리 정해진 방식으로 정해진 코드로 변환한다.
- [0055] 인코딩 장치는, 카운트에 누적에 따라 비트 값이 '0'인 연속되는 비트들의 수를 카운트하고, 카운트된 비트들의 수에 따라 해당 비트들을 특정 이진 코드로 변환할 수 있다.
- [0056] 인코딩 장치는, 이를 위해 미리 저장된 알고리즘을 이용할 수 있다. 알고리즘에 의해, 선행하는 0의 비트들의 경우의 수에 따라 각각 다른 이진 코드로 변환될 수 있다. 이때, 첫 번째로 나타나는 0을 제외한 나머지 비트들 및 첫 번째 1이 등장하는 비트까지의 비트열을 이진 코드로 변환할 수 있다.
- [0057] 단계(340)에서 인코딩 장치는, 선행하는 0의 비트들 이후의 후속 비트의 비트 값을 압축된 코드에 연결하여 표시할 수 있다.
- [0058] 앞서, 변환된 이진 코드에 1의 비트 값이 출현하는 비트 이후의 포함하는 후속 비트들은 변환하지 않고 해당 이진 코드의 뒤에 연결하여 표시할 수 있다. 예컨대, '00001xxx'의 값을 가지는 가중치에 대해서, '00001'에 대응하는 이진 코드를 1의 위치에 표시하고, xxx의 값을 해당 이진 코드의 후속 비트로 그대로 연결하여 표시할 수 있다.
- [0059] 인코딩 장치는, 이러한 방식으로 가중치에 따라 가변 길이를 가지는 형태의 압축 방식을 제공할 수 있다. 인코

딩이 가능한 경우의 수를 리딩 트리로 생성하여 디코딩에 참조할 수 있도록 한다.

- [0060] 인코딩 장치로 수신되는 복수의 가중치들에 대해 분석하면, 가중치들의 분포를 알 수 있게 되고, 해당 분포를 이용하여, 동일한 분포에 해당하는 가중치에 대해서 동일한 방식으로 압축함으로써 압축 효율을 높일 수 있다.
- [0061] 인코딩된 가중치는 메모리에 저장될 수 있고, 이후 디코딩을 위해 디코딩 장치로 제공될 수 있다.
- [0063] 도 4는 일 실시예에 따른 인코딩 장치의 블록도이다.
- [0064] 도 4를 참조하면, 일 실시예에 따른 인코딩 장치(400)는 통신 인터페이스(410), 프로세서(430), 및 메모리(450)를 포함할 수 있다. 통신 인터페이스(410), 프로세서(430) 및 메모리(450)는 통신 버스(405)를 통해 서로 통신할 수 있다.
- [0065] 통신 인터페이스(410)는 가중치를 수신한다.
- [0066] 프로세서(430)는 통신 인터페이스(410)를 통해 수신된 가중치를 미리 정해진 방식으로 압축할 수 있다. 프로세서(430)는 가중치의 비트열에 선행하는 0의 비트수에 따라 압축 코드를 결정하고, 후속 비트를 연결하여 표시하는 방법으로 가중치를 압축할 수 있다.
- [0067] 메모리(450)는 상술한 프로세서(430)의 처리 과정에서 생성되는 다양한 정보들을 저장할 수 있다. 이 밖에도, 메모리(450)는 각종 데이터와 프로그램 등을 저장할 수 있다. 메모리(450)는 휘발성 메모리 또는 비휘발성 메모리를 포함할 수 있다. 메모리(450)는 하드 디스크 등과 같은 대용량 저장 매체를 구비하여 각종 데이터를 저장할 수 있다.
- [0068] 또한, 프로세서(430)는 도 2 내지 도 3을 통해 전술한 적어도 하나의 방법 또는 적어도 하나의 방법에 대응되는 알고리즘을 수행할 수 있다. 프로세서(530)는 목적하는 동작들(desired operations)을 실행시키기 위한 물리적인 구조를 갖는 회로를 가지는 하드웨어로 구현된 데이터 처리 장치일 수 있다. 예를 들어, 목적하는 동작들은 프로그램에 포함된 코드(code) 또는 인스트럭션들(instructions)을 포함할 수 있다. 프로세서(430)는 예를 들어, CPU(Central Processing Unit), GPU(Graphics Processing Unit), 또는 NPU(Neural network Processing Unit)으로 구성될 수 있다. 예를 들어, 하드웨어로 구현된 예측 장치(500)는 마이크로프로세서(microprocessor), 중앙 처리 장치(central processing unit), 프로세서 코어(processor core), 멀티-코어 프로세서(multi-core processor), 멀티프로세서(multiprocessor), ASIC(Application-Specific Integrated Circuit), FPGA(Field Programmable Gate Array)를 포함할 수 있다.
- [0069] 프로세서(430)는 프로그램을 실행하고, 인코딩 장치(400)를 제어할 수 있다. 프로세서(430)에 의하여 실행되는 프로그램 코드는 메모리(450)에 저장될 수 있다.
- [0071] 도 5는 실시예에서, 연산기에서 연산 가능한 형태로 가중치를 디코딩하는 실시예를 설명하기 위한 도면이다.
- [0072] 디코딩 장치(500)는 도 2 내지 도 4의 방법으로 압축된 가중치를 연산 장치(510)에서 연산이 가능한 형태로 디코딩할 수 있다.
- [0073] 디코딩 장치(500)는 리딩 트리(501)를 참조하여 압축되어 가변 길이를 가지는 가중치를 디코딩할 수 있다.
- [0074] 리딩 트리(501)은 메모리에 저장되거나, 다른 장치로부터 불러올 수 있다. 리딩 트리(501)는 압축된 가중치에 대한 해시 정보를 포함할 수 있다. 예컨대, 압축 시 가중치의 선행되는 0의 개수에 따라 압축된 이진 코드에 대한 정보를 지시할 수 있다.
- [0075] 디코딩 장치(500)는 수신된 압축된 가중치의 길이를 고정 길이의 비트수로 먼저 변환할 수 있다. 압축된 가중치는 가변 길이를 가지기 때문에, 압축 이전의 고정 길이에 맞추어 비트 값을 추가할 수 있다.
- [0076] 리딩 트리(501)를 참조하여 따르면, 압축된 가중치의 선행 코드가 0인 경우, 심볼(symbol) A로 변환할 수 있고, 압축된 가중치의 선행 코드가 1110인 경우, 심볼 D로 복호할 수 있다.
- [0077] 압축된 가중치는 이를 저장하는 메모리, 예컨대 SRAM 등의 메모리를 통해 수신할 수 있다. 압축된 가중치는 메모리를 읽어서 가져올 수 있고, 디코딩 장치(500)는 선행 코드에 해당하는 부분을 리딩 트리(501)를 해시하여 표시할 수 있고, 선행 코드의 후속 비트는 해시하지 않고 해시된 코드 이후에 연결하여 표시할 수 있다.

- [0078] 이후 연산 가능한 형태가 되어 연산기(510)로 입력될 수 있다. 연산기(510)는 디코딩 장치(500)에 의해 획득한 원래의 가중치를 이용하여 입력에 대한 MAC 연산을 실행할 수 있다.
- [0079] 여기서, 압축된 가중치를 저장한 메모리는 연산을 위해 재사용될 수 있다. 해당 메모리 크기를 줄임으로써 메모리의 크기 및 소모 전력을 줄일 수 있다.
- [0080] 연산기(510)에서 압축 이전의 가중치(Weight) 및 입력(Activation)에 대해 MAC 연산이 수행될 수 있으며, 연산기를 통해 획득한 출력은, 다음 레이어의 입력이 되거나, 최종 출력 값이 될 수 있다.
- [0082] 도 6은 실시예에서, 디코딩 장치의 동작 방법을 설명하기 위한 흐름도이다.
- [0083] 단계(610)에서 디코딩 장치는, 압축된 가중치를 수신한다.
- [0084] 압축된 가중치는 선행 코드와 선행 코드와 연결되는 후속 비트들로 구성될 수 있다. 선행 코드가 압축된 비트에 해당한다.
- [0085] 디코딩 장치는, 압축된 가중치를 수신하고, 압축된 가중치의 길이를 고정 길이에 대응하도록 변환할 수 있다. 예를 들어, 8비트의 고정 길이를 가지는 가중치에 대해 압축된 가중치의 길이가 5비트인 경우, 8비트가 되도록 부족한 3비트의 비트 값(111)을 추가할 수 있다. 해당 비트 값은 선행 코드에 선행되어 추가될 수 있다.
- [0086] 단계(620)에서 디코딩 장치는, 리딩 트리를 참조하여 압축된 가중치의 선행 코드를 해시함으로써 압축된 가중치를 디코딩한다.
- [0087] 리딩 트리는 메모리에 저장되거나, 메모리로부터 불러올 수 있다. 리딩 트리는 선행 코드에 대한 해시 정보를 포함할 수 있다. 리딩 트리는, 압축 이전의 가중치에 선행되는 0의 개수에 따라 압축된 이진 코드에 대한 정보를 지시할 수 있다. 예컨대, 리딩 트리를 고정 길이 비트에 맞추어 추가된 1의 비트열을 포함하여 참조함으로써, 각각의 압축 이전의 비트열에 대한 압축 정보가 구분될 수 있도록 한다. 리딩 트리에 대한 예시는 도 5를 참조할 수 있다.
- [0088] 디코딩 장치는 선행 코드에 해당하는 부분을 리딩 트리로부터 해시하여 표시할 수 있고, 선행 코드의 후속 비트는 해시하지 않고 해시된 코드 이후에 연결하여 표시할 수 있다. 즉, 디코딩된 가중치는 선행 코드의 해시 정보 + 후속 비트의 형태로 표현될 수 있고, 디코딩된 가중치는 고정 길이에 해당할 수 있다.
- [0089] 디코딩된 가중치는 MAC 연산에 활용될 수 있다.
- [0091] 도 7은 일 실시예에 따른 디코딩 장치의 블록도이다.
- [0092] 도 7을 참조하면, 일 실시예에 따른 디코딩 장치(700)는 통신 인터페이스(710), 프로세서(730), 및 메모리(750)를 포함할 수 있다. 통신 인터페이스(710), 프로세서(730) 및 메모리(750)는 통신 버스(705)를 통해 서로 통신할 수 있다.
- [0093] 통신 인터페이스(710)는 압축된 가중치를 수신한다.
- [0094] 프로세서(730)는 통신 인터페이스(710)를 통해 압축된 가중치를 미리 정해진 방식으로 디코딩할 수 있다. 프로세서(730)는 가중치의 선행 코드에 대응하는 비트열을 이용하여 선행 코드를 변환하고, 이후의 후속 비트를 연결하여 표시하는 방법으로 압축된 가중치를 디코딩할 수 있다.
- [0095] 메모리(750)는 상술한 프로세서(730)의 처리 과정에서 생성되는 다양한 정보들을 저장할 수 있다. 이 밖에도, 메모리(750)는 각종 데이터와 프로그램 등을 저장할 수 있다. 메모리(750)는 휘발성 메모리 또는 비휘발성 메모리를 포함할 수 있다. 메모리(750)는 하드 디스크 등과 같은 대용량 저장 매체를 구비하여 각종 데이터를 저장할 수 있다.
- [0096] 또한, 프로세서(730)는 전술한 적어도 하나의 방법 또는 적어도 하나의 방법에 대응되는 알고리즘을 수행할 수 있다. 프로세서(530)는 목적하는 동작들(desired operations)을 실행시키기 위한 물리적인 구조를 갖는 회로를 가지는 하드웨어로 구현된 데이터 처리 장치일 수 있다. 예를 들어, 목적하는 동작들은 프로그램에 포함된 코드(code) 또는 인스트럭션들(instructions)을 포함할 수 있다. 프로세서(430)는 예를 들어, CPU(Central Processing Unit), GPU(Graphics Processing Unit), 또는 NPU(Neural network Processing Unit)으로 구성될

수 있다. 예를 들어, 하드웨어로 구현된 예측 장치(500)는 마이크로프로세서(microprocessor), 중앙 처리 장치(central processing unit), 프로세서 코어(processor core), 멀티-코어 프로세서(multi-core processor), 멀티프로세서(multiprocessor), ASIC(Application-Specific Integrated Circuit), FPGA(Field Programmable Gate Array)를 포함할 수 있다.

[0097] 프로세서(730)는 프로그램을 실행하고, 디코딩 장치(700)를 제어할 수 있다. 프로세서(730)에 의하여 실행되는 프로그램 코드는 메모리(750)에 저장될 수 있다.

[0099] 도 8은 실시예에서, 스마트폰 AP로 구현되는 방법을 설명하기 위한 도면이다.

[0100] 도 8은 스마트폰 AP의 내부 구성을 개략적으로 나타낸 것이다. AP와 같은 곳에서 인코딩이 이루어지고, 디코딩 장치(800)가 AP에 내장되는 실시예를 도시한 것이다.

[0101] 스마트폰 AP는 NPU(Neural Processing Unit)를 포함할 수 있다. NPU는 MAC등의 연산을 위해 구비될 수 있다. NPU는 디코딩 장치(800)에서 디코딩되어 획득한 가중치를 수신하여 입력에 대한 MAC 연산을 수행할 수 있다.

[0102] 스마트폰의 NPU가 신경망을 연산할 때에, 많은 전력이 외부 메모리에서 가중치를 읽어오는 데에 소모되는데, 앞서 설명된 인코딩 방식으로 압축된 가중치를 수신하게 되면, 소모되는 통신량 및 전력을 줄일 수 있게 된다.

[0103] 만약, 가중치를 읽어오는 데에 신경망 처리 속도가 저하되는 경우, 이를 위한 통신량을 줄임으로써 신경망 처리 속도를 높일 수 있다.

[0105] 도 9는 실시예에서, NPU를 구현하는 방법을 설명하기 위한 도면이다.

[0106] 도 9의 NPU는 소비 전력을 절감하기 위한 것이다. 압축된 가중치를 내부 메모리(On-chip memory, SRAM)에서 수신하고, 연산기를 사용할 때만 압축된 가중치를 디코딩하도록 디코딩 장치(900)가 메모리에 연결되어 구현될 수 있다.

[0107] 내부 메모리의 비용과 데이터를 읽는 전력을 추가적으로 더 절약할 수 있다. 한편, 내부 메모리에서 외부 메모리보다 읽는 횟수가 많기 때문에 디코딩하는 전력이 더 소모될 수 있다.

[0109] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

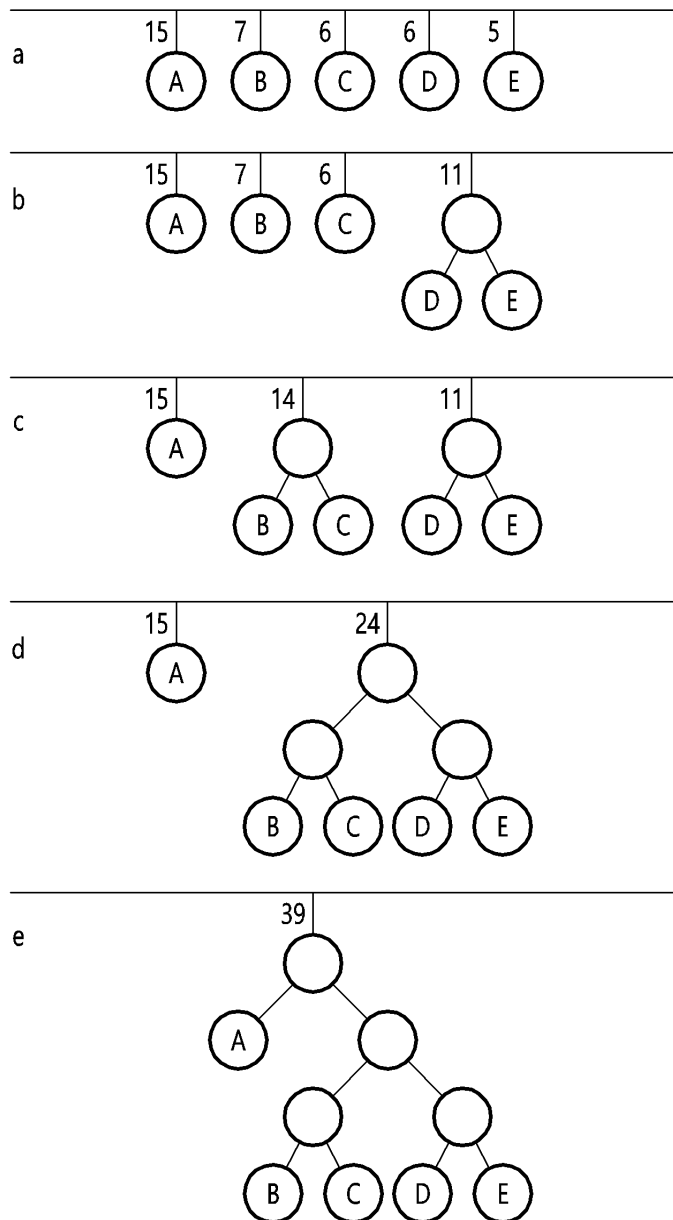
[0111] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0113] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

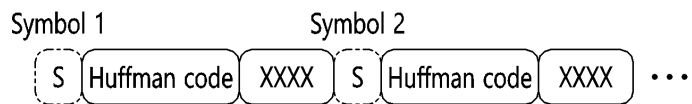
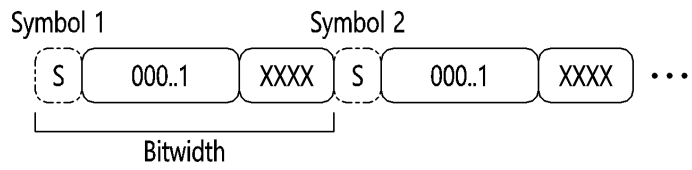
[0114] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속한다.

도면

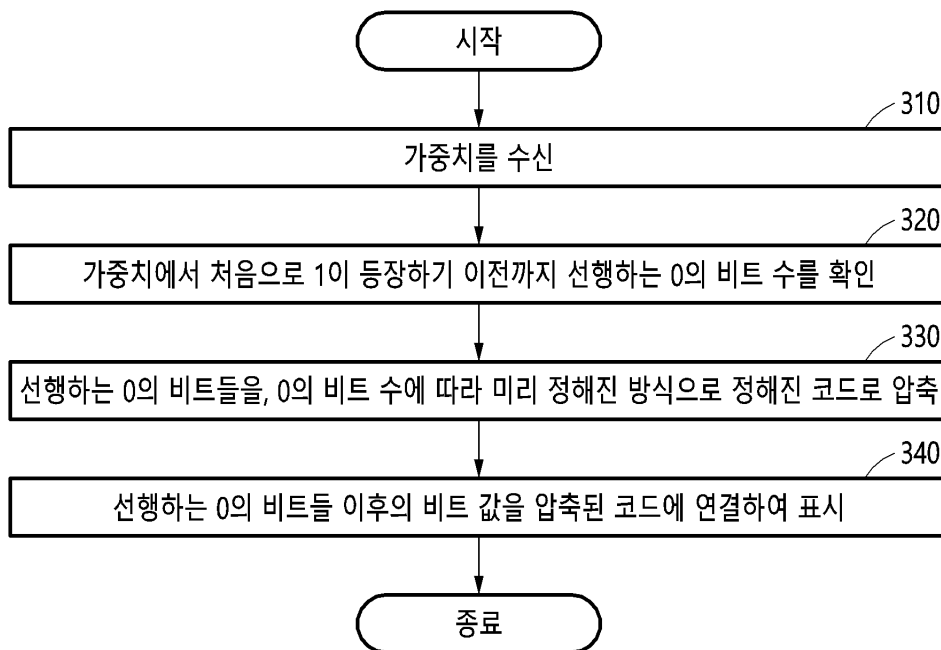
도면1



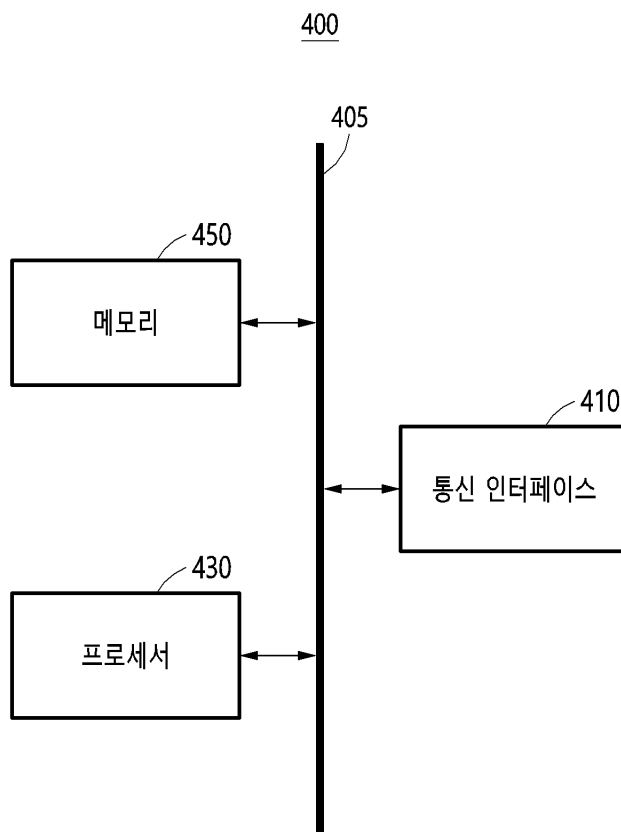
도면2



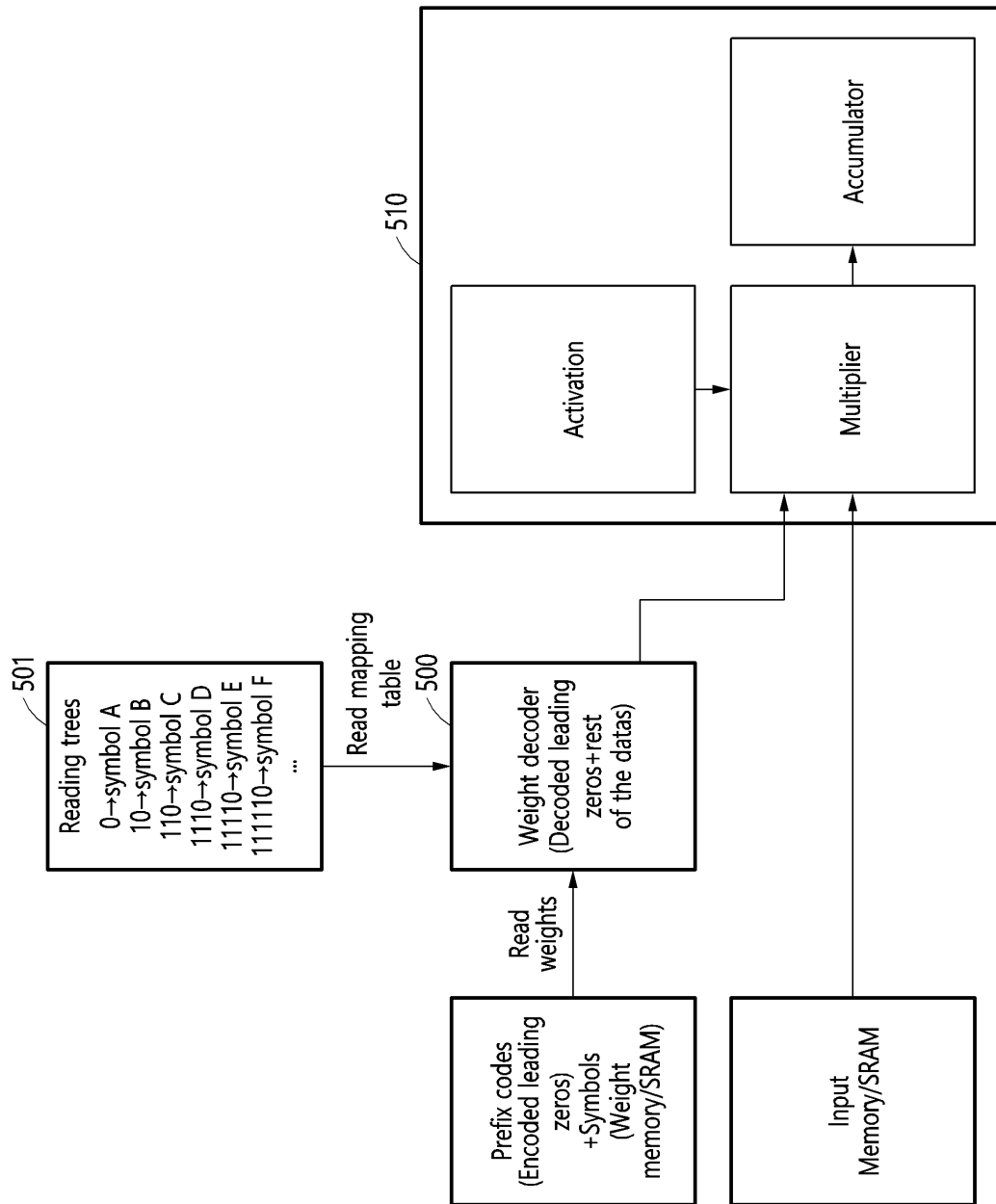
도면3



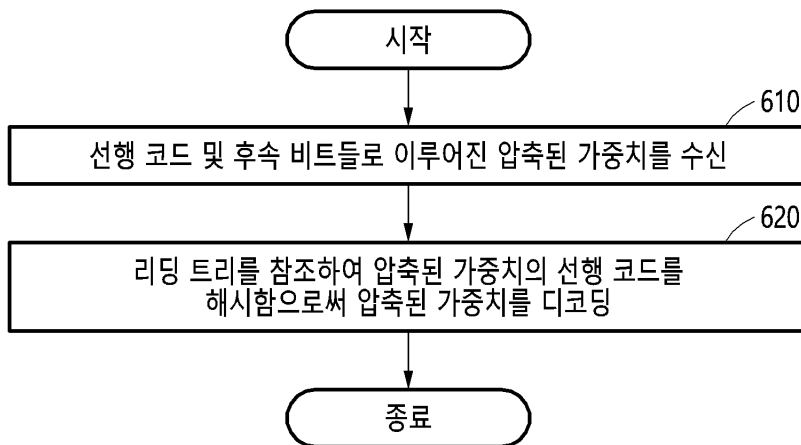
도면4



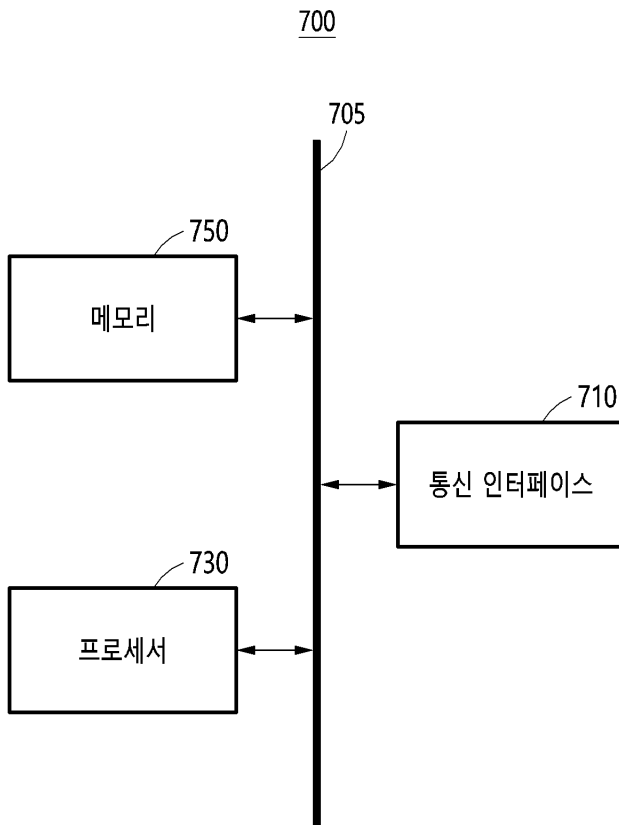
도면5



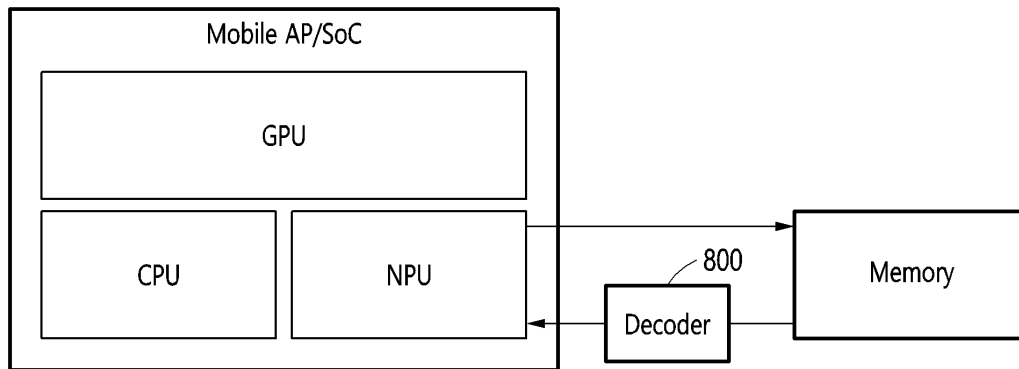
도면6



도면7



도면8



도면9

