(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년02월08일
(11) 등록번호 10-2360116
(24) 등록일자 2022년02월03일

(51) 국제특허분류(Int. Cl.)
*G06N 3/063* (2006.01) *G06N 5/04* (2006.01)
*H04N 19/13* (2014.01) *H04N 19/93* (2014.01)
(52) CPC특허분류
*G06N 3/063* (2013.01)
*G06N 5/04* (2013.01)
(21) 출원번호 10-2021-0039858
(22) 출원일자 2021년03월26일
심사청구일자 2021년03월26일
(56) 선행기술조사문헌
JP2020149156 A*
*(뒷면에 계속)*
전체 청구항 수 : 총 9 항

(73) 특허권자
**세종대학교산학협력단**
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
**이성주**
서울특별시 광진구 뚝섬로35길 32, 308-1110
**이준표**
서울특별시 광진구 능동로28길 23 202호
(74) 대리인
**이강민, 안준형, 남승희**

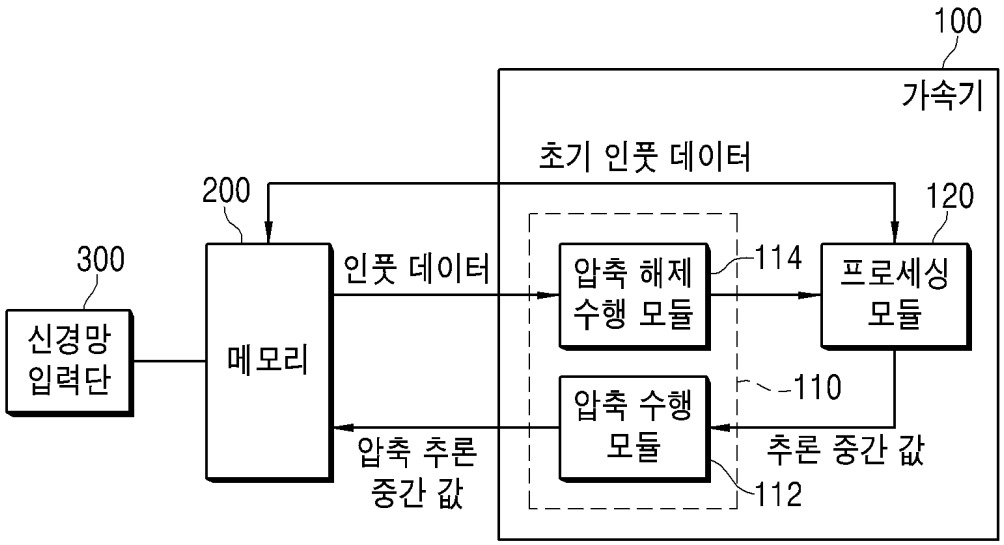
심사관 : 노지명

(54) 발명의 명칭 **압축 모듈을 포함하는 인공지능 가속기 및 이를 이용한 데이터 전달 방법**

## (57) 요 약

본 발명은 인공지능 가속기에 관한 것으로서, 가속기의 추론 중간 값을 런 렝스 코딩(Run Length Coding) 기법과 동적 허프만 코딩(Adaptive Huffman Coding) 기법을 사용하여 1, 2차 압축하여 메모리에 전달하도록 구성된 압축 모듈을 포함하는 인공지능 가속기 및 이를 이용한 데이터 전달 방법에 관한 것이다.

*대 표 도* - 도1



(52) CPC특허분류
*H04N 19/13* (2015.01)
*H04N 19/93* (2015.01)

(56) 선행기술조사문헌
KR1020090102598 A*
KR1020200093404 A*
JP4741317 B2
KR1020210053791 A
*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업
과제고유번호 1711116145
과제번호 2018-0-01423-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 대학ICT연구센터지원사업
연구과제명 지능형 비행로봇 융합기술 연구
기 여 율 5/10
과제수행기관명 세종대학교 산학협력단
연구기간 2021.01.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345321135
과제번호 2020R1A6A1A0303854011
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 대학중점연구소지원사업
연구과제명 자율지능무인비행체연구소
기 여 율 1/10
과제수행기관명 세종대학교 산학협력단
연구기간 2021.03.01 ~ 2022.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711108024
과제번호 2020R1A2C1007546
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 실내보안용 초고해상도 지능형 레이다센서 신호처리 연구
기 여 율 4/10
과제수행기관명 세종대학교 산합협력단
연구기간 2021.03.01 ~ 2022.02.28

# 명 세 서

## 청구범위

### 청구항 1

인공지능 가속기(artificial intelligent accelerator)에 있어서,

신경망(neural network) 입력단의 초기 인풋 데이터를 기반으로 연산하는 프로세싱 모듈로부터 출력되는 추론 중간 값을 압축하여 메모리에 저장하고, 상기 메모리에 저장된 압축된 추론 중간 값을 인풋 데이터로 획득하여 압축 해제하여 프로세싱 모델로 입력하는 압축모듈; 및

메모리에 저장된 초기 인풋 데이터를 획득하여 추론 연산을 수행하고, 그 추론 연산의 중간 값인 추론 중간 값을 출력하여 압축 모듈을 통해 메모리에 저장하고, 상기 저장된 추론 중간 값을 압축 모듈을 통해 인풋 데이터로 획득하여 추론 연산을 수행하는 프로세싱 모듈;을 포함하여 구성되며,

상기 압축 모듈은,

상기 프로세싱 모듈로부터 출력되는 추론 중간 값을 소정의 제1, 2 압축 방식으로 순차적으로 1차 및 2차 압축하여 메모리로 전달하여 저장하는 압축 수행 모듈; 및

상기 압축 수행 모듈에 의해 1, 2차 압축되어 저장된 압축 추론 중간 값을 리딩 하여 인풋 데이터로 획득하여, 상기 제1, 2차 압축 방식 각각에 대응하는 제1, 2 압축 해제 방식으로 1, 2차 압축 해제하여 프로세싱 모듈로 입력하는 압축 해제 수행 모듈; 을 포함하여 구성되고,

상기 1차 압축은,

반복되는 데이터의 값이 많은 추론 중간 값의 압축 효과를 높이기 위해, 중복되는 문자를 한 문자로 치환하는 런 렝스 압축 기법을 제1 압축방식으로 적용하여 수행하고,

상기 2차 압축은,

입력되는 데이터를 이용하여 실시간으로 트리 구조를 업데이트하여 압축을 진행하는 동적 허프만 코딩(Adaptive Huffman Coding) 기법을 제2 압축 방식으로 적용하여 수행하는 인공지능 가속기.

### 청구항 2

삭제

### 청구항 3

제1항에 있어서,

상기 압축 수행 모듈은,

상기 프로세싱 모듈로부터 출력되는 추론 중간 값을 소정의 제1 압축 방식으로 1차 압축하는 제1 압축 수행 모듈;

상기 제1 압축 수행 모듈에 의해 소정의 제1 압축 방식으로 1차 압축된 추론 중간 값을 소정의 제2 압축 방식으로 2차 압축하여 압축 추론 값을 생성하는 제2 압축 수행 모듈;

을 포함하여 구성되는 것을 특징으로 하는 인공지능 가속기.

### 청구항 4

제3항에 있어서,

상기 압축 해제 수행 모듈은,

메모리에 저장된 상기 압축 추론 중간 값을 인풋 데이터로 획득하여, 상기 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제하는 제1 압축 해제 수행 모듈;

상기 제1 압축 해제 수행 모듈에 의해 1차 압축 해제된 인풋 데이터를 상기 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제하여 프로세싱 모듈로 입력하는 제2 압축 해제 수행 모듈;

을 포함하여 구성되는 것을 특징으로 하는 인공지능 가속기.

**청구항 5**

삭제

**청구항 6**

삭제

**청구항 7**

제4항에 있어서,

상기 제1 압축 해제 방식은,

상기 동적 허프만 코딩(Adaptive Huffman Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 하는 인공지능 가속기.

**청구항 8**

제4항에 있어서,

상기 제2 압축 해제 방식은,

상기 런 렝스 코딩(Run Length Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 하는 인공지능 가속기.

**청구항 9**

인공지능 가속기(artificial intelligent accelerator)에서 메모리 간의 데이터 전달 방법에 있어서,

가속기의 프로세싱 모듈에서, 메모리에 미리 저장된 신경망 입력단으로부터의 초기 인풋 데이터를 리딩하여 획득하는 초기 인풋 데이터 획득 단계;

가속기의 프로세싱 모듈에서, 상기 초기 인풋 데이터 획득 단계에서 획득한 초기 인풋 데이터를 이용하여 연산하여 추론 결과의 중간 값인 추론 중간 값을 출력하는 추론 중간 값 출력 단계;

가속기의 압축 모듈에서, 상기 추론 중간 값 출력 단계에서 출력된 프로세싱 모듈로부터의 추론 중간 값을 소정의 제1 압축 방식을 이용하여 1차 압축을 수행하고, 1차 압축된 추론 중간 값을 상기 제1 압축 방식과는 다른 소정의 제2 압축 방식을 이용하여 2차 압축하는 추론 중간 값 압축 단계;

가속기의 압축 모듈에서, 상기 압축 추론 중간 값 저장 단계에 의해 메모리 에 저장된 압축 추론 중간 값을 리딩하여 인풋 데이터로 획득하는 인풋 데이터 획득 단계;

가속기의 압축 모듈에서, 상기 인풋 데이터 획득 단계에서 인풋 데이터로 획 득한 압축 추론 중간 값을 상기 소정의 제1, 2 압축 방식에 각각 대응하는 제1, 2 압축 해제 방식으로 압축 추론 중간 값을 1, 2차 압축 해제하는 인풋 데이터 압축 해제 단계;

가속기의 압축 모듈에서, 상기 인풋 데이터 압축 해제 단계에서 2차 압축 해 제된 인풋 데이터를 프로세싱 모듈로 전달하는 인풋 데이터 전달 단계;

를 포함하여 구성되고,

상기 1차 압축의 수행은, 반복되는 데이터의 값이 많은 추론 중간 값의 압축 효과를 높이기 위해, 중복되는 문자를 한 문자로 치환하는 런 렝스 압축 기법을 제1 압축방식으로 적용하여 수행하고,

상기 2차 압축의 수행은, 입력되는 데이터를 이용하여 실시간으로 트리 구조를 업데이트하여 압축을 진행하는 동적 허프만 코딩(Adaptive Huffman Coding) 기법을 제2 압축 방식으로 적용하여 수행하는 것;

을 특징으로 하는 인공지능 가속기와 메모리 간 데이터 전달 방법.

**청구항 10**

삭제

**청구항 11**

삭제

**청구항 12**

제9항에 있어서,

상기 인풋 데이터 압축 해제 단계는,

상기 인풋 데이터 획득 단계에서 인풋 데이터로 획득한 압축 추론 중간 값을 상기 소정의 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제하는 제1 압축 해제 수행 단계;

상기 제1 압축 해제 단계에서 제2 압축 해제 방식으로 1차 압축 해제한 압축 추론 중간 값을 상기 소정의 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제하는 제2 압축 해제 수행 단계;

를 포함하여 구성되는 것을 특징으로 하는 인공지능 가속기와 메모리 간 데이터 전달 방법.

**청구항 13**

삭제

**청구항 14**

삭제

**청구항 15**

제12항에 있어서,

상기 제1 압축 해제 방식은,

동적 허프만 코딩(Adaptive Huffman Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 하는 인공지능 가속기와 메모리 간 데이터 전달 방법.

**청구항 16**

제12항에 있어서,

상기 제2 압축 해제 방식은,

런 렝스 코딩(Run Length Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 하는 인공지능 가속기와 메모리 간 데이터 전달 방법.

## *발명의 설명*

### *기 술 분 야*

[0001] 본 발명은 인공지능 가속기에 관한 것으로서, 보다 구체적으로는 2가지의 압축 기술을 사용하여 압축하도록 구성된 압축 모듈을 포함하는 인공지능 가속기 및 이를 이용한 데이터 전달 방법에 관한 것이다.

### *배 경 기 술*

[0002] 현대 사회에서 인공지능(artificial intelligent)은 4차 산업혁명을 견인하는 핵심 기술로 자리 잡았다. 인공지능의 산업적 활용을 위해서는 막대한 계산량으로 인해 발생하는 전력 소비 문제를 해결하는 것인데, 이러한 문제를 해결하기 위한 방안 중 하나로서 연산 속도를 높이기 위해 인공신경망(artificial neural network)의 연산에 최적화되도록 설계된 연산 장치인 인공지능 가속기(accelerator)를 사용하고 있다.

[0003] 통상의 인공지능 가속기를 이용한 추론 결과 값 획득 과정을 살펴보면, 신경망(neural network) 입력단(예, pc, 카메라 등)에서 추론 결과를 얻고자 하는 새로운 데이터를 메모리에 입력하면, 가속기의 프로세싱 유닛(processing unit)은 이를 초기 입력 데이터로 입력 받아 추론 결과 값을 도출하고, 이를 다시 메모리로 전달하면 신경망 입력단에서 해당 메모리 주소로 추론 결과 값을 읽어 들여 확인하는 것으로 구성된다. 여기서, 가속기의 프로세싱 유닛은 내부 연산 프로그램에 따라 연산을 진행하는데, 전체 데이터를 한 번에 연산할 경우 연산량이 너무 많아지기 때문에 일반적으로 연산 과정 중에 일부 데이터(즉, 추론 중간 값)를 출력하여 메모리에 저장하였다가 이를 다시 입력 데이터로 입력 받아 연산하여, 단계 별로 나누어서 연산하는 방식으로 구성된다.

[0004] 그런데, 메모리와 인공지능 가속기 사이에서 일어나는 데이터 전달은 많은 시간과 에너지를 소모하며, 이는 곧 전력 소비 문제로 이어지게 된다. 이를 해결하기 위해 메모리와 가속기 사이의 데이터 전달을 최소화하기 위해 많은 연구들이 진행되고 있다.

[0005] 대표적인 해결 방안으로서, 통상적으로 가속기 내에 프로세싱 유닛으로부터 출력되는 추론 중간 값을 소정의 압축 기법으로 압축하여 메모리에 전달하고, 이를 다시 입력 받으면 압축 해제하여 프로세싱 유닛으로 전달하는 압축 모듈을 구성하여 메모리와 인공지능 가속기 사이의 데이터 전달 속도를 최소화 하는 방식을 사용하고 있다.

[0006] 종래에는 압축 모듈의 압축 기법으로서, 주로 무손실 압축 기법 중 하나인 허프만 코딩 기법을 채택하여 사용하고 있다.

[0007] 허프만 코딩(Huffman coding) 기법은, 정해진 데이터의 횟수들을 측정하여 가장 많이 등장하는 데이터가 가장 적은 비트를 사용하여 보낼 수 있도록 가변적인 데이터 비트를 사용하여 압축하는 알고리즘이다. 이러한 허프만 코딩은 필터 설정 값에 따라 특정 개수의 데이터들을 가지고 있어야 압축 알고리즘 수행하도록 구성된다. 따라서, 가속기의 프로세싱 유닛으로부터 추론 중간 값이 특정 개수가 채워질 때까지 대기하고 있다가 특정 개수가 충족되면 압축 알고리즘을 이용하여 압축시켜 메모리로 전달하게 된다. 이로 인해, 대기 시간 동안의 공백으로 데이터 전달 속도가 저하되는 문제가 발생된다.

[0008] (특허문헌 1) KR10-2020-0093404 A

## 발명의 내용

### 해결하려는 과제

[0009] 본 발명은 상술한 문제점을 해결하고자 하는 것으로서, 기존의 허프만 코딩이 갖는 속도 저하를 개선하기 위하여 압축 모듈을 2가지의 압축 기법을 결합하여 압축하도록 구성하여, 데이터 전달 속도 및 압축률이 모두 향상된 인공지능 가속기를 제공하고자 한다.

### 과제의 해결 수단

[0010] 본 발명에 따른 인공지능 가속기(artificial intelligent accelerator)는, 신경망(neural network) 입력단의 초기 인풋 데이터를 기반으로 연산하는 프로세싱 모듈로부터 출력되는 추론 중간 값을 소정의 서로 다른 압축 방식으로 1, 2차 압축하여 메모리에 저장하고, 상기 메모리에 저장된 압축된 추론 중간 값을 인풋 데이터로 획득하여 상기 압축 방식 각각에 대응하는 압축 해제 방식으로 1, 2차 압축 해제하여 상기 프로세싱 모듈로 입력하는 압축 모듈; 및 메모리에 저장된 초기 인풋 데이터를 획득하여 추론 연산을 수행하고, 그 추론 연산의 중간 값인 추론 중간 값을 출력하여 압축 모듈을 통해 메모리에 저장하고, 상기 저장된 추론 중간 값을 압축 모듈을 통해 인풋 데이터로 획득하여 추론 연산을 수행하는 프로세싱 모듈; 을 포함하여 구성된다.

[0011] 보다 구체적으로, 상기 압축 모듈은, 프로세싱 모듈로부터 출력되는 추론 중간 값을 소정의 제1, 2 압축 방식으로 1, 2차 압축하여 메모리로 전달하여 저장하는 압축 수행 모듈; 및 상기 압축 수행 모듈에 의해 2차 압축되어 저장된 압축 추론 중간 값을 리딩하여 인풋 데이터로 획득하여, 상기 제1, 2차 압축 방식 각각에 대응하는 제1, 2 압축 해제 방식으로 1, 2차 압축 해제하여 프로세싱 모듈로 입력하는 압축 해제 수행 모듈; 을 포함하여 구성되는 것을 특징으로 한다.

[0012] 상기 압축 수행 모듈은, 프로세싱 모듈로부터 출력되는 추론 중간 값을 소정의 제1 압축 방식으로 1차 압축하는 제1 압축 수행 모듈; 상기 제1 압축 수행 모듈에 의해 소정의 제1 압축 방식으로 1차 압축된 추론 중간 값을 소정의 제2 압축 방식으로 2차 압축하여 압축 추론 값을 생성하는 제2 압축 수행 모듈; 을 포함하여 구성되는 것

을 특징으로 한다.

[0013] 한편, 상기 압축 해제 수행 모듈은, 메모리에 저장된 상기 압축 추론 중간 값을 인풋 데이터로 획득하여, 상기 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제하는 제1 압축 해제 수행 모듈; 상기 제1 압축 해제 수행 모듈에 의해 1차 압축 해제된 인풋 데이터를 상기 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제하여 프로세싱 모듈로 입력하는 제2 압축 해제 수행 모듈; 을 포함하여 구성되는 것을 특징으로 한다.

[0014] 여기서, 상기 소정의 제1 압축 방식은, 런 렝스 코딩(Run Length Coding) 기법인 것을 특징으로 한다.

[0015] 한편, 상기 소정의 제2 압축 방식은, 동적 허프만 코딩(Adaptive Huffman Coding) 기법인 것을 특징으로 한다.

[0016] 한편, 상기 제1 압축 해제 방식은, 상기 동적 허프만 코딩(Adaptive Huffman Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 한다.

[0017] 한편, 상기 제2 압축 해제 방식은, 상기 런 렝스 코딩(Run Length Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 한다.

[0018] 본 발명에 따른 인공지능 가속기(artificial intelligent accelerator)에서 메모리 간의 데이터 전달 방법은, 가속기의 프로세싱 모듈에서, 메모리에 미리 저장된 신경망 입력단으로부터의 초기 인풋 데이터를 리딩하여 획득하는 초기 인풋 데이터 획득 단계; 가속기의 프로세싱 모듈에서, 상기 초기 인풋 데이터 획득 단계에서 획득한 초기 인풋 데이터를 이용하여 연산하여 추론 결과의 중간 값인 추론 중간 값을 출력하는 추론 중간 값 출력 단계; 가속기의 압축 모듈에서, 상기 추론 중간 값 출력 단계에서 출력된 프로세싱 모듈로부터의 추론 중간 값을 소정의 서로 다른 압축 방식으로 1, 2차 압축하여 압축 추론 중간 값을 생성하는 추론 중간 값 압축 단계; 가속기의 압축 모듈에서, 상기 추론 중간 값 압축 단계에서 생성된 압축 추론 중간 값을 메모리로 전달하여 저장하는 압축 추론 중간 값 저장 단계; 를 포함하여 구성된다.

[0019] 한편, 가속기의 압축 모듈에서, 상기 압축 추론 중간 값 저장 단계에 의해 메모리에 저장된 압축 추론 중간 값을 리딩하여 인풋 데이터로 획득하는 인풋 데이터 획득 단계; 가속기의 압축 모듈에서, 상기 인풋 데이터 획득 단계에서 인풋 데이터로 획득한 압축 추론 중간 값을 상기 소정의 제1, 2 압축 방식에 각각 대응하는 제1, 2 압축 해제 방식으로 압축 추론 중간 값을 1, 2차 압축 해제하는 인풋 데이터 압축 해제 단계; 가속기의 압축 모듈에서, 상기 인풋 데이터 압축 해제 단계에서 2차 압축 해제된 인풋 데이터를 프로세싱 모듈로 전달하는 인풋 데이터 전달 단계; 를 더 포함하여 구성된다.

[0020] 보다 구체적으로, 상기 추론 중간 값 압축 단계는, 상기 추론 중간 값 출력 단계에서 출력된 프로세싱 모듈로부터의 추론 중간 값을 소정의 제1 압축 방식을 사용하여 1차 압축하는 제1 압축 수행 단계; 상기 제1 압축 수행 단계에서 소정의 제1 압축 방식으로 1차 압축된 추론 중간 값을 소정의 제2 압축 방식을 사용하여 2차 압축하는 제2 압축 수행 단계; 를 포함하여 구성되는 것을 특징으로 한다.

[0021] 한편, 상기 인풋 데이터 압축 해제 단계는, 상기 인풋 데이터 획득 단계에서 인풋 데이터로 획득한 압축 추론 중간 값을 상기 소정의 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제하는 제1 압축 해제 수행 단계; 상기 제1 압축 해제 단계에서 제2 압축 해제 방식으로 1차 압축 해제한 압축 추론 중간 값을 상기 소정의 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제하는 제2 압축 해제 수행 단계; 를 포함하여 구성되는 것을 특징으로 한다.

[0022] 여기서, 상기 소정의 제1 압축 방식은, 런 렝스 코딩(Run Length Coding) 기법인 것을 특징으로 한다.

[0023] 한편, 상기 소정의 제2 압축 방식은, 동적 허프만 코딩(Adaptive Huffman Coding) 기법인 것을 특징으로 한다.

[0024] 한편, 상기 제1 압축 해제 방식은, 동적 허프만 코딩(Adaptive Huffman Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 한다.

[0025] 한편, 상기 제2 압축 해제 방식은, 런 렝스 코딩(Run Length Coding) 기법의 압축 과정을 역순으로 수행하는 것; 을 특징으로 한다.

## 발명의 효과

[0026] 본 발명은 가속기의 압축 모듈을 런 렝스 코딩(Run Lenght Coding) 기법으로 1차 압축하고, 동적 허프만 코딩(Adaptive Huffman Coding) 기법으로 2차 압축하여 메모리로 전달하도록 구성함으로써, 각 기법의 단점은 상호

보완하고 장점은 극대화 하는 효과를 발휘하여 압축률 및 데이터 전달 속도가 모두 향상된 인공지능 가속기를 제공할 수 있다.

## 도면의 간단한 설명

[0027] 도 1은 본 발명에 따른 인공지능 가속기를 포함하는 전체적인 데이터 전달 시스템을 도시한 도면이다.

도 2는 도 1의 압축 모듈의 세부 구성을 도시한 도면이다.

도 3은 런 렝스 코딩 기법의 예시를 보여주는 도면이다.

도 4는 동적 허프만 코딩 기법의 예시를 보여주는 도면이다.

도 5는 허프만 코딩 기법의 예시를 보여주는 도면이다.

도 6은 본 발명에 따른 인공지능 가속기와 메모리 간 데이터 전달 방법의 흐름을 보여주는 도면이다.

## 발명을 실시하기 위한 구체적인 내용

[0028] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시 예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면부호를 붙였다.

[0029] 제1, 제2 등과 같이 서수를 포함하는 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되지는 않는다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예컨대, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 본 출원에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

[0030] 명세서 전체에서, 어떤 부분이 다른 부분과 “연결”되어 있다고 할 때, 이는 “직접적으로 연결”되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 “전기적으로 연결”되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 “포함”한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 본원 명세서 전체에서 사용되는 정도의 용어 “~(하는) 단계” 또는 “~의 단계”는 “~를 위한 단계”를 의미하지 않는다.

[0031] 본 발명에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0032] 이하, 도면을 참조하여 본 발명에 대하여 상세하게 설명한다.

[0033] 1. 본 발명에서 사용하는 용어

[0034] 1.1. 추론 중간 값

[0035] 본 발명에서 사용하는 추론 중간 값은, 가속기의 프로세싱 모듈에서 인풋 데이터와 가중치를 이용하여 연산하는 과정 중에 중간마다 출력하는 부분합(partial sum)을 말한다.

[0036] 통상적으로 가속기의 프로세싱 모듈은 미리 설정된 내부 연산 프로그램에 기반하여 신경망 입력단(PC, 카메라 등)으로부터 입력하는 새로운 입력 데이터에 대한 추론 결과 값을 도출하기 위한 연산을 수행하는데, 한 번에 모든 연산을 처리하기에는 연산량이 너무 많아지기 때문에 연산 처리 과정에서 중간 중간마다 연산 결과를 저장하고, 이를 다시 입력 데이터로 하여 다음 연산 단계를 수행하도록 구성된다. 여기서, 연산 과정 중에 중간마다 출력되는 연산 결과 값을 본 명세서에서는 추론 중간 값으로 지칭하는 것이다.

[0037] 즉, 추론 중간 값이란 가속기의 추론 결과 값을 도출하기 위한 연산 과정 중 출력하는 중간 값을 의미한다.

[0038] 1.2. 압축 추론 중간 값

[0039] 압축 추론 중간 값이란, 본 발명에 따른 가속기의 압축 모듈에 의해 두 번 압축 완료된 추론 중간 값을 의미한다.

[0040] 따라서, 본 발명은 가속기에서 메모리에 추론 중간 값을 저장하고 이를 다시 읽어 들일 때 두 번 압축된 압축 추론 중간 값의 형태로 전달하여, 종래보다 빠른 속도로 데이터 전달이 이루어질 수 있다.

[0041] 2. 본 발명에 따른 인공지능 가속기

[0042] 본 발명에 따른 인공지능 가속기(artificial intelligent accelerator)는, 가속기로 들어오는 인풋 데이터를 이용하여 연산한 추론 결과를 중간 중간마다 저장하기 위해 메모리에 접근할 때 최대한 압축하여 빠르게 전달하여 메모리와 가속기 간 데이터 전달 속도를 높일 수 있도록 구성된 압축 모듈을 포함하여 구성된다. 도 1은 본 발명에 따른 가속기를 포함하는 전체적인 데이터 전달 시스템을 도시한 도면이고, 도 2는 압축 모듈의 세부 구성을 도시한 도면이다.

[0043] 도 1 및 2를 참조하면, 본 발명에 따른 인공지능 가속기(100)는 하기의 구성을 포함하여 구성된다.

[0044] 2.1. 압축 모듈(110)

[0045] 압축 모듈(110)은, 메모리(200)와 가속기(100) 내의 프로세싱 모듈(120) 사이에 구비되어, 프로세싱 모듈(120)로부터 출력되는 추론 중간 값을 소정의 서로 다른 압축 방식으로 두 번에 걸쳐 압축하여 메모리(200)로 전달하여 저장하고, 상기 메모리(200)에 저장되어 있는 압축된 추론 중간 값을 인풋 데이터로 다시 획득하여 상기 압축 방식 각각에 대응하는 압축 해제 방식으로 두 번에 걸쳐 복원하여 프로세싱 모듈(120)로 입력하도록 구성된다.

[0046] 이와 같은 압축 모듈은, 아래와 같은 세부 구성을 포함하여 구성될 수 있다.

[0047] 가. 압축 수행 모듈(112)

[0048] 압축 수행 모듈(112)은, 프로세싱 모듈(120)로부터 출력되는 추론 중간 갓을 소정의 서로 다른 압축 방식으로 1, 2차 압축하여 메모리(200)로 전달하여 저장하는 구성이다.

[0049] 1) 제1 압축 수행 모듈(1122)

[0050] 제1 압축 수행 모듈(1122)은, 프로세싱 모듈(120)로부터 출력되는 추론 중간 값을 소정의 제1 압축 방식으로 1차 압축하도록 구성될 수 있다.

[0051] 여기서, 소정의 제1 압축 방식은 런 렝스 코딩(Run Length Coding) 기법일 수 있다.

[0052] 런 렝스 코딩 기법은, 중복되는 문자를 한 문자로 치환하는 방식으로 데이터의 길이를 압축시키는 개념이다. 쉽게 말해, ‘AAAAABBCCCDEEFFGG’ 라는 텍스트가 있을 때 이것을 각각 ‘문자 X 반복 횟수’로 표현하는 방법이다.

[0053] 본 발명의 제1 압축 수행 모듈(1122)은, 이와 같은 런 렝스 코딩 기법을 사용하여 프로세싱 모듈(120)로부터의 추론 중간 값을 1차 압축한다.

[0054] 도 3은 런 렝스 코딩 기법으로 1차 압축하는 예시를 보여주는 도면이다. 도 3을 참조하여 설명하면, 예를 들어 프로세싱 모듈(120)로부터 출력되는 추론 중간 값이 도 3의 (a)와 같은 경우, 이를 런 렝스 코딩 기법으로 압축하면 도 3의 (b)와 같이 추론 중간 값의 첫 번째에 배열된 ‘3’ 부터 값, 횟수, 값, 횟수, 값, 횟수 순서로 데이터가 압축 표현되는 것이다.

[0055] 이와 같이 런 렝스 코딩 기법은 연속적으로 같은 데이터가 많이 들어올수록 압축률이 높아지는 알고리즘으로서, 특히 활성화 함수가 ‘Relu’ 함수일 경우 발생하는 대부분의 중간 값이 0이므로 압축의 효과를 더욱 극대화 할 수 있다.

[0056] 2) 제2 압축 수행 모듈(1124)

[0057] 제2 압축 수행 모듈은, 상기 제1 압축 수행 모듈(1122)에 의해 소정의 제1 방식으로 1차 압축된 추론 중간 값을 소정의 제2 압축 방식으로 다시 압축하여 압축 추론 중간 값을 생성할 수 있다.

[0058] 여기서, 소정의 제2 압축 방식은 동적 허프만 코딩(Adaptive Huffman Coding) 기법일 수 있다.

[0059] 동적 허프만 코딩 기법은, 문자의 빈도수를 만들어나가면서 코딩을 하는 방법으로서, 트리가 실시간으로 만들어

지며 입력 값을 읽고 트리를 만들어지는 과정이 동시에 이루어지는 방식이다.

[0060] 도 4는 동적 허프만 코딩 기법을 사용하여 1차 압축된 추론 중간 값을 2차 압축하는 예시를 보여주는 도면이다. 이를 참조하면, 런 렝스 코딩 기법으로 1차 압축된 추론 중간 값이 도 3의 (b)과 같은 경우, 제2 압축 수행 모듈(1124)에 '3, 1, 0, 9, …' 순서로 들어오면서 트리 구조가 실시간으로 업데이트 되며, 압축된 값의 표현도 달라진다. 예를 들어, 도 3의 (b)에서 마지막 숫자인 '4'가 들어올 때 트리 구조는 도 4의 (a)와 같이 업데이트 되며 해당 트리를 이용한 테이블은 도 4의 (b)에 보이는 표와 같다.

[0061] 이에 따라, 제2 압축 수행 모듈(1124)로 '4' 뒤에 들어오는 값이 도 4 (b)의 테이블에 존재하는 숫자라면 테이블에 존재하는 압축된 값 표현으로 바뀌고, 테이블에 존재하지 않는 새로운 값이 입력될 경우 해당 값은 압축되지 않고 그대로 메모리(200)에 저장하고 도 4 (a), (b)의 트리 및 테이블에는 해당 새로운 값을 추가되어 업데이트 된다.

[0062] 앞서 언급하였던 것과 같이, 종래에는 정해진 데이터의 횟수들을 측정하여 가장 많이 등장하는 데이터가 가장 적은 비트를 사용하여 보낼 수 있도록 가변적인 데이터 비트를 사용하여 압축하는 알고리즘인 허프만 코딩(Huffman coding)을 사용하였다. 도 5는 허프만 코딩 기법의 예시를 보여주는 도면이다. 이를 참조하면, 그림 (a)처럼 'A', 'B', 'C', 'D', 'E' 가 각각 15번, 7번, 6번, 6번, 5번 등장한 데이터를 가지고 허프만 코딩을 하면, (b) 단계처럼 가장 적은 횟수로 등장하는 'D'와 'E'를 묶어 트리로 표현한 후, (c) 단계에서는 'C'와 'D', 'E' 횟수 합이 'B'의 횟수보다 많기 때문에 'C'와 'B'를 묶는다. 이후 (d) 단계와 (e) 단계에서는 남은 'A'가 가장 많은 횟수이기 때문에 가장 위쪽 최상단 트리로 넣어 가장 적은 비트가 포함되도록 만든다. 그러므로, (e) 단계의 완성된 그림으로 보았을 때 'A'는 '1', 'B'는 '001', 'C'는 '010', 'D'는 '001', 'E'는 '000' 표현된다.

[0063] 이와 같은 알고리즘으로 압축하는 허프만 코딩 기법은, (a) 단계에서 미리 설정된 특정 개수의 데이터, 즉 'A', 'B', 'C, 'D', 'E' 처럼 5개의 데이터가 쌓여야지만 알고리즘을 이용하여 압축하는 것이 가능하다. 이러한 특징은, 쌓인 데이터의 개수가 많을수록 압축 효과가 극대화되어 높은 압축률을 가지는 반면, 특정 개수의 데이터가 쌓일 때까지 대기해야 하므로 대기시간 동안의 공백이 발생하게 된다. 이로 인해 가속기에서 메모리로의 데이터 전달에 속도 저하 문제가 있지만, 종래에는 이러한 문제점을 감안하고 압축률에 주안점을 두는 방향으로 하여 허프만 코딩 기법을 채택하여 사용하였다.

[0064] 본 발명은 이를 개선하기 위하여, 상술한 바와 같이 프로세싱 모듈(120)로부터 출력되는 추론 중간 값을 반복되는 데이터의 값이 많을수록 압축률의 효과가 극대화되는 런 렝스 기법으로 1차 압축하고, 이를 허프만 코딩 기법과 달리 입력되는 데이터를 이용하여 실시간으로 트리 구조를 업데이트하며 압축을 진행하는 동적 허프만 기법으로 2차 압축하여 메모리(200)로 전달하도록 구성함으로써, 상호 간 단점 보완으로 각 기법의 장점을 극대화시켜 향상된 압축률 및 시간 단축 효과를 제공할 수 있다.

[0065] 나. 압축 해제 수행 모듈(114)

[0066] 압축 해제 수행 모듈(114)는, 상기 압축 수행 모듈(112)에 의해 2차 압축되어 메모리(200)에 저장된 압축 추론 중간 값을 리딩하여 다시 인풋 데이터로 획득하여, 상기 제1, 2 압축 방식 각각에 대응하는 제1, 2 압축 해제 방식으로 1, 2차 압축 해제하여 프로세싱 모듈(120)로 입력하는 구성이다.

[0067] 1) 제1 압축 해제 수행 모듈(1142)

[0068] 제1 압축 해제 수행 모듈(1142)은, 메모리(200)에 저장되어 있는 압축 추론 중간 값을 인풋 데이터로 획득하여, 이를 상기 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제하도록 구성될 수 있다.

[0069] 여기서, 제2 압축 해제 방식이라 함은, 압축 수행 모듈(112)에서 메모리(200)로 전달할 시 마지막 압축 방식인 동적 허프만 코딩(Adaptive Huffman Coging) 기법의 압축 과정을 역순으로 수행하는 것을 말한다. 즉, 앞서 설명한 도 4의 예시와 같은 알고리즘을 역순으로 수행하는 것이다.

[0070] 압축 추론 중간 값은 런 렝스 코딩 기법으로 1차 압축, 동적 허프만 코딩 기법으로 2차 압축된 것이므로, 이를 다시 복원하기 위해서는 역순으로 먼저 2차 압축 방식으로 해제한 후 1차 압축 방식으로 해제되어야 한다.

[0071] 따라서, 압축 추론 중간 값의 1차 압축 해제를 수행하는 제1 압축 해제 수행 모듈(1142)은, 제2 압축 방식에 대응하는 동적 허프만 코딩 기법의 압축 과정을 역순으로 수행하여 압축 추론 중간 값을 1차적으로 압축 해제할 수 있다.

[0072] 2) 제2 압축 해제 수행 모듈(1144)

[0073] 제2 압축 해제 수행 모듈은, 상기 제1 압축 해제 수행 모듈(1142)에 의해 1차 압축 해제된 압축 추론 중간 값을 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제하도록 구성될 수 있다.

[0074] 여기서, 제1 압축 해제 방식이라 함은, 압축 수행 모듈(112)에서 프로세싱 모듈(120)로부터의 추론 중간 값을 1차 압축한 압축 방식인 런 렝스 코딩(Run Length Coding) 기법의 압축 과정을 역순으로 수행하는 것을 말한다. 즉, 앞서 설명한 도 3의 예시와 같은 알고리즘을 역순으로 수행하는 것이다.

[0075] 제2 압축 해제 수행 모듈(1144)에서 상기 제1 압축 해제 수행 모듈(1142)에 의해 1차 압축된 압축 추론 중간 값을 런 렝스 코딩 기법의 압축 과정을 역순으로 수행하여 2차 압축 해제하면, 압축 수행 모듈(112)에서 추론 중간 값의 압축을 수행하기 전, 즉 프로세싱 모듈(120)로부터 출력되었을 시의 본래의 추론 중간 값 상태로 복원된다.

[0076] 이와 같이 복원된 추론 중간 값은 프로세싱 모듈(120)로 입력하여 다음 단계의 연산을 수행할 수 있도록 한다.

[0077] 2.2. 프로세싱 모듈(120)

[0078] 프로세싱 모듈(120)은, 미리 설정된 내부 연산 프로그램에 기반으로 메모리(200)에 저장된 인풋 데이터를 이용하여 추론 연산을 수행하여 추론 결과 값을 도출하는 구성이다.

[0079] 구체적으로, 초기에는 메모리(200)에 저장된 초기 인풋 데이터를 리딩하여 획득하여 추론 연산을 수행하되, 연산 수행 중 추론 결과의 중간 값인 추론 중간 값을 출력하여 압축 모듈(110)을 통해 메모리(200)에 저장하고, 저장된 추론 중간 값을 압축 모듈(110)을 통해 다시 인풋 데이터로 하여 추론 연산을 단계적으로 수행하도록 구성된다.

[0080] 아래에는 본 발명의 가속기(100)를 설명하는 데에 언급된 공지 구성으로서 메모리(200)와 신경망 입력단(300)에 대해 간략하게 설명한다.

[0081] 2.3. 메모리(200)

[0082] 메모리(200)는, 신경망 입력단(300)으로부터 입력 받은 초기 인풋 데이터를 저장하고, 가속기(100)의 압축 모듈(110)로부터 전달되는 압축 추론 중간 값을 저장한다.

[0083] 이러한 메모리(200)에 의해, 가속기(100)에서 추론 연산 단계에 따라 중간 값을 저장하였다가, 다시 인풋 데이터로서 읽어 들여 다음 단계의 연산을 수행할 수 있어 가속기(100)의 연산 과정이 효율적으로 운영되도록 한다.

[0084] 2.4. 신경망(neural network) 입력단(300)

[0085] 신경망 입력단(300)은, 가속기(100)를 이용하여 새로운 데이터에 대한 추론 결과 값을 획득하고자 하는 구성으로서, 예를 들어 PC, 카메라 등을 포함할 수 있다.

[0086] 신경망 입력단(300)은, 추론 결과 값을 얻고자 하는 새로운 데이터를 메모리(200)에 초기 인풋 데이터로서 저장한다. 이후 초기 인풋 데이터를 입력 받은 가속기(100)로부터 추론 연산을 통해 추론 중간 값이 출력되어 메모리(200)에 저장되면, 해당 메모리 주소로 추론 중간 값을 읽어 들여 인식/확인하는 형태로 구성된다.

[0087] 3. 본 발명에 따른 인공지능 가속기의 데이터 전달 방법

[0088] 도 6은 본 발명에 따른 인공지능 가속기를 이용한 메모리로의 데이터 전달 방법의 흐름을 보여주는 도면이다.

[0089] 도 6을 참조하면, 본 발명의 데이터 전달 방법은, 아래와 같은 단계를 포함하여 구성된다.

[0090] 3.1. 초기 인풋 데이터 획득 단계(S100)

[0091] 먼저, 가속기의 프로세싱 모듈(120)에서, 신경망 입력단(300)에 의해 메모리(200)에 저장된 초기 인풋 데이터를 리딩하여 획득하는 단계이다. 가속기의 프로세싱 모듈(120)은 획득한 초기 인풋 데이터를 시작으로 추론 연산을 수행할 수 있다.

[0092] 여기서, 초기 인풋 데이터란, 신경망 입력단(300)에서 추론 결과를 얻고자 입력하는 새로운 데이터를 의미한다.

[0093] 3.2. 추론 중간 값 출력 단계(S200)

[0094] 가속기의 프로세싱 모듈(120)은, 상기 초기 인풋 데이터 획득 단계(S100)에서 획득한 초기 인풋 데이터를 이용하여 연산하며 추론 결과의 중간 값인 추론 중간 값을 출력하는 추론 중간 값 출력 단계(S200)를 수행한다.

[0095] 3.3. 추론 중간 값 압축 단계(S300)

[0096] 가속기의 압축 모듈(110)은, 가속기의 프로세싱 모듈(120)에 의해 추론 중간 값 출력 단계(S200)에서 추론 중간 값이 출력되면, 상기 출력된 추론 중간 값을 소정의 서로 다른 압축 방식으로 1, 2차 압축하여 압축 추론 중간 값을 생성하는 추론 중간 값 압축 단계(S300)를 수행한다.

[0097] 이러한 추론 중간 값 압축 단계(S300)는 아래의 세부 단계를 포함하여 구성된다.

[0098] 가. 제1 압축 수행 단계(S310)

[0099] 먼저, 상기 추론 중간 값 출력 단계(S200)에서 출력된 프로세싱 모듈(120)로부터의 추론 중간 값을 소정의 제1 압축 방식을 사용하여 1차 압축하는 제1 압축 수행 단계(S310)를 수행한다.

[0100] 여기서, 소정의 제1 압축 방식은 런 렝스 코딩(Run Length Coding) 기법일 수 있다.

[0101] 런 렝스 기법을 사용하여 데이터를 1차 압축하는 방식은 앞서 시스템 구성에서 상세하게 설명하였으므로, 구체적인 설명은 생략한다.

[0102] 이와 같은 제1 압축 수행 단계(S310)는, 가속기(100)의 제1 압축 수행 모듈(1122)에 의해 실행된다.

[0103] 나. 제2 압축 수행 단계(S320)

[0104] 상기 제1 압축 수행 단계(S310)에 의해 소정의 제1 압축 방식을 사용하여 프로세싱 모듈(120)로부터의 추론 중간 값이 1차 압축되면, 상기 1차 압축된 추론 중간 값을 소정의 제2 압축 방식을 사용하여 2차 압축하는 제2 압축 수행 단계(S320)를 수행한다.

[0105] 여기서, 소정의 제2 압축 방식은 동적 허프만 코딩(Adaptive Huffman Coding) 기법일 수 있다.

[0106] 동적 허프만 코딩 기법을 사용하여 데이터를 압축하는 방식은, 앞서 시스템 구성에서 상세하게 설명하였으므로, 구체적인 설명은 생략한다.

[0107] 이와 같은 제2 압축 수행 단계(S320)는, 가속기의 제2 압축 수행 모듈(1124)에 의해 실행된다.

[0108] 3.4. 압축 추론 중간 값 저장 단계(S400)

[0109] 가속기의 압축 모듈(110)은, 상기 추론 중간 값 압축 단계(S300)에서 두 번에 걸쳐 압축된 추론 중간 값을 메모리(200)로 전달하여 저장하는 압축 추론 중간 값 저장 단계(S400)를 수행한다.

[0110] 3.5. 인풋 데이터 획득 단계(S500)

[0111] 가속기의 압축 모듈(110)에서, 상기 압축 추론 중간 값 저장 단계(S400)를 통해 메모리(200)에 저장되어 있는 압축 추론 중간 값을 리딩하여 다시 인풋 데이터로 획득하는 인풋 데이터 획득 단계(S500)를 수행한다.

[0112] 이 때, 획득하는 인풋 데이터는 앞서 설명한 제1, 2 압축 수행 단계(S310, S320)에 의해 2차 압축되어 있는 상태이다.

[0113] 3.6. 인풋 데이터 압축 해제 단계(S600)

[0114] 가속기의 압축 모듈(110)은, 상기 인풋 데이터 획득 단계(S500)에서 인풋 데이터로서 획득한 압축 추론 중간 값을 상기 추론 중간 값 압축 단계(S300)에서의 소정의 제1, 제2 압축 방식에 각각 대응하는 제1, 제2 압축 해제 방식으로 압축 해제하는 단계를 수행한다.

[0115] 가. 제1 압축 해제 수행 단계(S610)

[0116] 먼저, 상기 인풋 데이터 획득 단계(S500)에서 메모리(200)로부터 인풋 데이터로 획득한 압축 추론 중간 값을 상기 소정의 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제한다.

[0117] 상술한 것처럼 소정의 제2 압축 방식은 동적 허프만 코딩 기법으로, 이에 대응하는 제2 압축 해제 방식이라 함은 동적 허프만 코딩 기법의 압축 과정을 역순으로 수행하는 것을 말한다.

[0118] 이와 같은 단계는, 압축 모듈(110)의 제1 압축 해제 수행 모듈(1142)에 의해 이루어진다.

[0119] 나. 제2 압축 해제 수행 단계(S620)

[0120] 다음, 상기 제1 압축 해제 수행 단계(S610)에서 제2 압축 해제 방식으로 메모리(200)로부터의 압축 추론 중간 값을 1차 압축 해제하면, 상기 소정의 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제한다.

[0121] 상술한 것처럼 소정의 제1 압축 방식은 런 렝스 코딩 기법으로, 이에 대응하는 제1 압축 해제 방식이라 함은 런 렝스 코딩 기법의 압축 과정을 역순으로 수행하는 것을 말한다.

[0122] 즉, 프로세싱 모듈(120)로부터의 추론 중간 값을 런 렝스 코딩 기법으로 1차 압축하고, 동적 허프만 코딩 기법으로 2차 압축하여 메모리(200)에 저장 후, 이를 다시 인풋 데이터로 획득하여 역순으로 동적 허프만 코딩 기법으로 1차 압축 해제하고, 런 렝스 코딩 기법으로 2차 압축 해제하여 본래의 추론 중간 값 상태로 복원하는 것이다.

[0123] 이와 같은 단계는, 압축 모듈(110)의 제2 압축 해제 수행 모듈(1144)에 의해 이루어진다.

[0124] 3.7. 인풋 데이터 전달 단계(S700)

[0125] 가속기의 압축 모듈(110)은, 상기 제1, 2 압축 해제 수행 단계(S510, S520)를 거쳐 메모리(200)로부터의 압축 추론 중간 값이 압축 전 상태의 추론 중간 값으로 복원되면, 이를 프로세싱 모듈(120)로 입력하는 인풋 데이터 전달 단계(S700)를 수행한다.

[0126] 그러면, 프로세싱 모듈(120)은 인풋 데이터로 입력 받은 압축 해제된 추론 중간 값을 이용하여 다음 연산 과정을 수행할 수 있다.

[0127] 이후, 앞서 설명한 추론 중간 값 출력 단계(S200) 내지 인풋 데이터 전달 단계(S700)가 반복적으로 수행되어 가속기(100)에서 추론 결과 값을 도출할 수 있다.

[0128] 한편, 본 발명의 기술적 사상은 상기 실시 예에 따라 구체적으로 기술되었으나, 상기 실시 예는 그 설명을 위한 것이며, 그 제한을 위한 것이 아님을 주지해야 한다. 또한, 본 발명의 기술분야에서 당업자는 본 발명의 기술 사상의 범위 내에서 다양한 실시 예가 가능함을 이해할 수 있을 것이다.

## *부호의 설명*

[0129] 100: 인공지능 가속기

110: 압축 모듈

112: 압축 수행 모듈

1122: 제1 압축 수행 모듈

1124: 제2 압축 수행 모듈

114: 압축 해제 수행 모듈

1142: 제1 압축 해제 수행 모듈

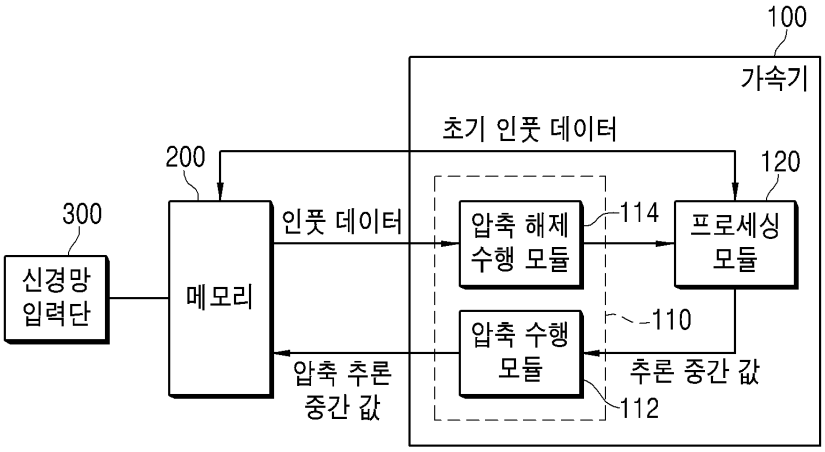
1144: 제2 압축 해제 수행 모듈

200: 메모리

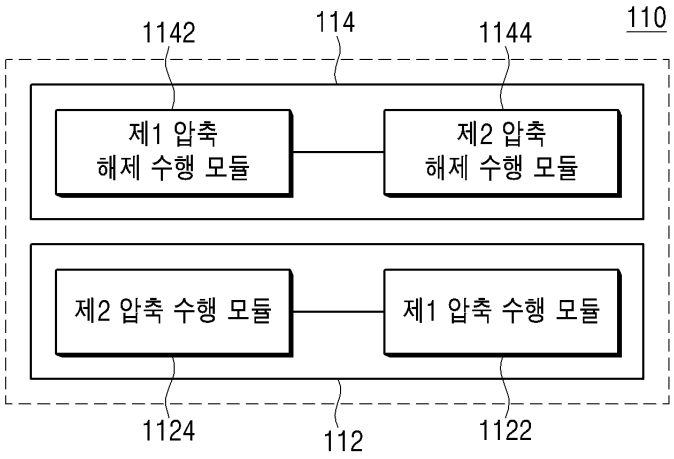
300: 신경망 입력단

## 도면

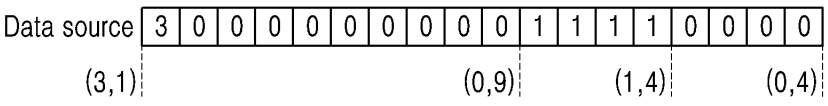
### 도면1



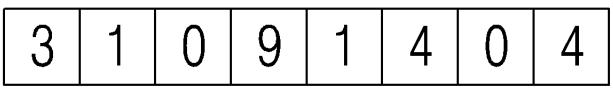

### 도면2




### 도면3

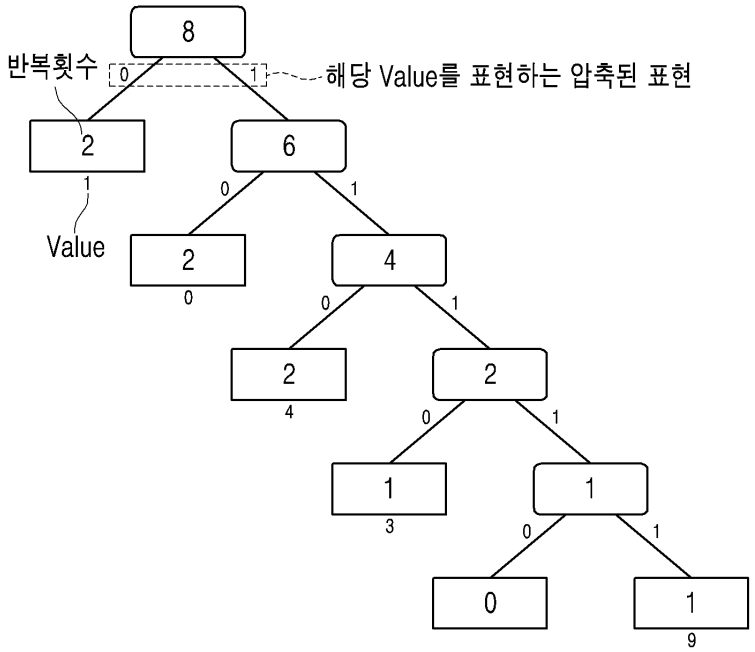



(a)




(b)

도면4

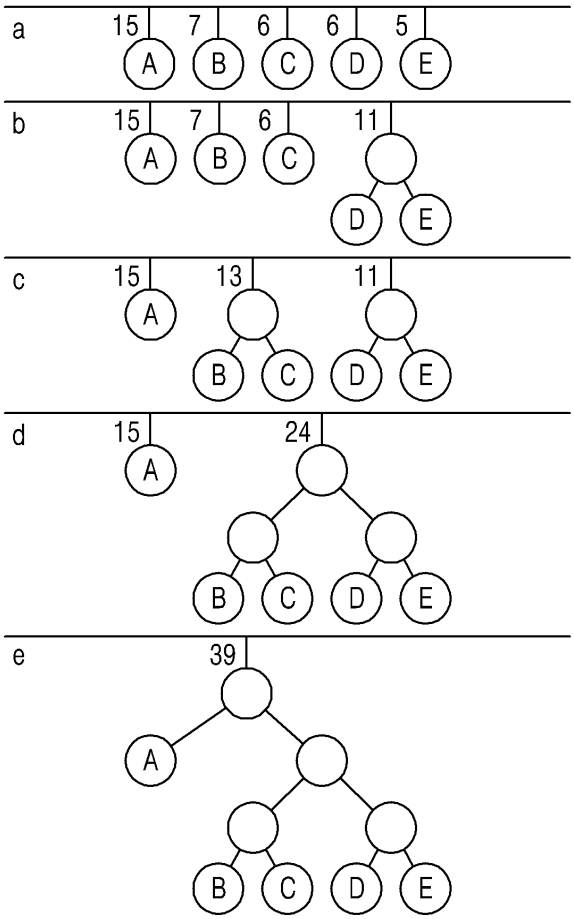



(a)

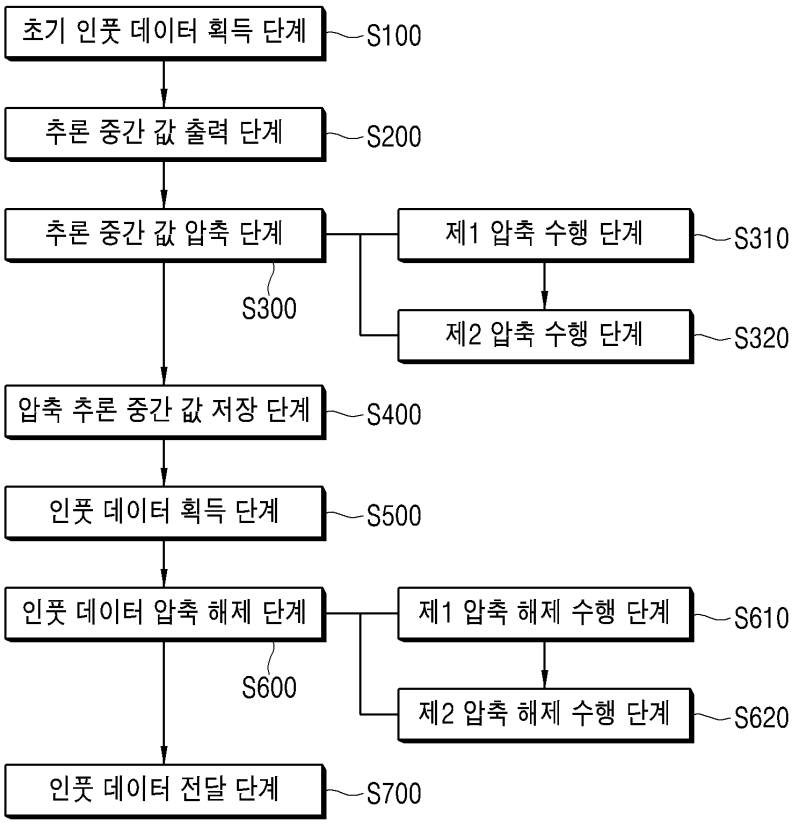
| 값 | 횟수 | 압축된 값 (2진수) |
|---|---|---|
| 3 | 1 | 1110 |
| 1 | 2 | 0 |
| 0 | 2 | 10 |
| 9 | 1 | 11110 |
| 4 | 2 | 110 |

(b)

도면5


a
15
7
6
6
5
A
B
C
D
E
b
15
7
6
11
A
B
C
D
E
c
15
13
11
A
B
C
D
E
d
15
24
A
B
C
D
E
e
39
A
B
C
D
E

도면6




【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 4

【변경전】

제3항에 있어서,

상기 압축 해제 수행 모듈은,

메모리에 저장된 상기 압축 추론 중간 값을 인풋 데이터로 획득하여, 상기 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제하는 제1 압축 해제 수행 모듈;

상기 제1 압축 해제 수행 모듈에 의해 1차 압축 해제된 인풋 데이터를 상기 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제하여 프로세싱 모듈로 입력하는 제2 압축 해제 수행 모듈;

을 포함하여 구성되는 것을 특징으로 하는 인공지능 가속기.

【변경후】

제3항에 있어서,

상기 압축 해제 수행 모듈은,

메모리에 저장된 상기 압축 추론 중간 값을 인풋 데이터로 획득하여, 상기 제2 압축 방식에 대응하는 제2 압축 해제 방식으로 1차 압축 해제하는 제1 압축 해제 수행 모듈;

상기 제1 압축 해제 수행 모듈에 의해 1차 압축 해제된 인풋 데이터를 상기 제1 압축 방식에 대응하는 제1 압축 해제 방식으로 2차 압축 해제하여 프로세싱 모듈로 입력하는 제2 압축 해제 수행 모듈;

을 포함하여 구성되는 것을 특징으로 하는 인공지능 가속기.