



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0053466
(43) 공개일자 2025년04월22일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) G06N 3/042 (2023.01)
G06N 3/0464 (2023.01) G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/042 (2023.01)
(21) 출원번호 10-2023-0136760
(22) 출원일자 2023년10월13일
심사청구일자 2023년10월13일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
홍충선
경기도 용인시 수지구 상현로 30-10, 4813동 101호(상현동, 광고상떼빌파크뷰)
김녹원
경기도 용인시 기흥구 서천서로 27, 107동 1302호(서천동, 센트럴파크원)
(74) 대리인
두호특허법인

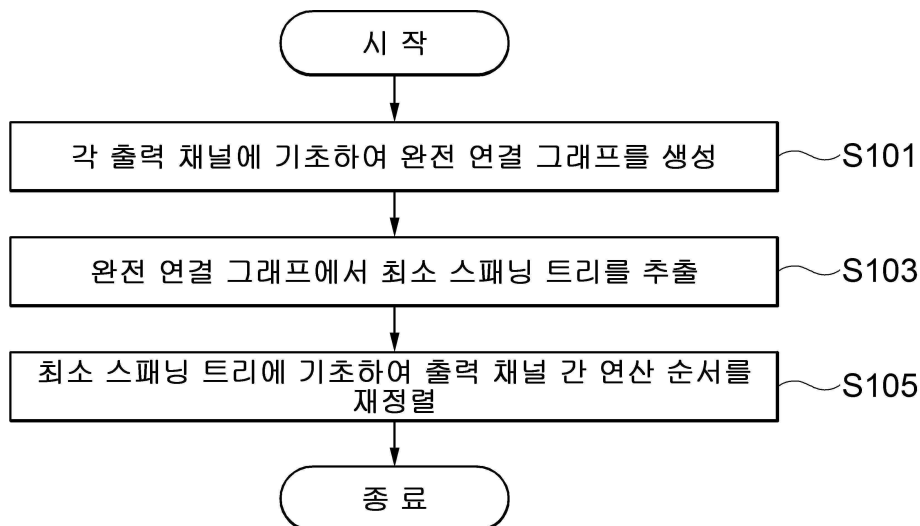
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 이진 신경망의 연산 방법 및 이를 수행하기 위한 컴퓨팅 장치

(57) 요약

이진 신경망의 연산 방법 및 이를 수행하기 위한 컴퓨팅 장치가 개시된다. 개시되는 일 실시예에 따른 컴퓨팅 장치는, 하나 이상의 프로세서들, 및 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 이진 신경망(Binary Neural Network)의 컨볼루션 계층의 각 출력 채널에 기초하여 완전 연결 그래프(fully connected graph)를 생성하는 단계, 완전 연결 그래프에서 최소 스패닝 트리(Minimum Spanning Tree)를 추출하는 단계, 및 최소 스패닝 트리에 기초하여 각 출력 채널 간 연산 순서를 재정렬하는 단계를 포함한다.

대표도 - 도2



(52) CPC특허분류

G06N 3/0464 (2023.01)

G06N 3/088 (2023.01)

(72) 발명자

보 우옹 허유

경기도 수원시 영통구 청명로59번길 59-10, 202호
(영통동)

김현수

경기도 성남시 분당구 정자로 143, 205동 1304호
(정자동, 한솔마을)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711179316
과제번호 00155911
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 인공지능융합혁신인재양성
연구과제명 인공지능융합혁신인재양성(경희대학교)
기 여 율 1/4
과제수행기관명 경희대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711193491
과제번호 2019-0-01287-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 SW컴퓨팅산업원천기술개발
연구과제명 (SW 스타랩) 분산 옛지를 위한 진화형 딥러닝 모델생성 플랫폼
기 여 율 1/4
과제수행기관명 경희대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711194179
과제번호 00207816
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 집단연구지원
연구과제명 Meta Federated Learning 기반 Satellite-Air-Ground 통합 네트워크 시스템 핵심 구
조 개발
기 여 율 1/4
과제수행기관명 경희대학교산학협력단
연구기간 2023.03.01 ~ 2024.02.29

이 발명을 지원한 국가연구개발사업

과제고유번호 1711193622
과제번호 2021-0-02068-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 인공지능 혁신 허브 연구 개발
기 여 율 1/4
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서,

이진 신경망(Binary Neural Network)의 컨볼루션 계층의 각 출력 채널에 기초하여 완전 연결 그래프(fully connected graph)를 생성하는 단계;

상기 완전 연결 그래프에서 최소 스패닝 트리(Minimum Spanning Tree)를 추출하는 단계; 및

상기 최소 스패닝 트리에 기초하여 각 출력 채널 간 연산 순서를 재정렬하는 단계를 포함하는, 이진 신경망의 연산 방법.

청구항 2

청구항 1에 있어서,

상기 완전 연결 그래프는,

상기 출력 채널 간 거리에 기초하여 생성하는, 이진 신경망의 연산 방법.

청구항 3

청구항 2에 있어서,

상기 출력 채널은, 0 또는 1로 구성되는 가중치들의 목록을 포함하고,

상기 출력 채널 간 거리는, 해밍 거리(hamming distance)를 통해 산출하는, 이진 신경망의 연산 방법.

청구항 4

청구항 3에 있어서,

상기 완전 연결 그래프는,

각 출력 채널을 노드로 하고, 각 출력 채널 간 해밍 거리를 엣지로 하여 생성되는, 이진 신경망의 연산 방법.

청구항 5

청구항 4에 있어서,

상기 연산 순서를 재정렬하는 단계는,

상기 최소 스패닝 트리에서 최상위 노드에 해당하는 출력 채널에 대해 가장 먼저 연산을 수행하는 단계를 포함하는, 이진 신경망의 연산 방법.

청구항 6

청구항 5에 있어서,

상기 연산 순서를 재정렬하는 단계는,

상기 최소 스패닝 트리의 상기 최상위 노드에서 해밍 거리가 짧은 하위 노드를 따라 연산을 수행하는 단계를 더 포함하는, 이진 신경망의 연산 방법.

청구항 7

청구항 6에 있어서,

상기 컴퓨팅 장치는,

i번째 출력 채널로부터 j번째 출력 채널을 계산하기 위해 다음 수학적식을 사용하는, 이진 신경망의 연산 방법.

(수학적식)

$$Y_j = 2(P_i - d_{ij} + 2P_{ij}) - C_{in} \times M \times M$$

P_i : i번째 출력 채널의 팝 카운트(Popcount) 연산

d_{ij} : i번째 출력 채널과 j번째 출력 채널 사이의 해밍 거리

P_{ij} : i번째 출력 채널과 j번째 출력 채널 간 XNOR 연산의 합

C_{in} : 이진 신경망의 입력 채널

M : 커널 사이즈

청구항 8

청구항 1에 있어서,

상기 이진 신경망의 연산 방법은,

상기 이진 신경망의 컨볼루션 계층의 각 출력 채널의 가중치들을 클러스터링 하는 단계;

상기 클러스터링 된 각 클러스터에서 중심을 랜덤하게 선택하는 단계; 및

상기 각 클러스터의 중심과 해당 클러스터에 속하는 각 가중치 간의 거리가 최소화 되도록 상기 이진 신경망을 학습하는 단계를 더 포함하는, 이진 신경망의 연산 방법.

청구항 9

하나 이상의 프로세서들;

메모리; 및

하나 이상의 프로그램들을 포함하고,

상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며,

상기 하나 이상의 프로그램들은,

이진 신경망의 컨볼루션 계층의 각 출력 채널에 기초하여 완전 연결 그래프를 생성하기 위한 명령;

상기 완전 연결 그래프에서 최소 스패닝 트리를 추출하기 위한 명령; 및

상기 최소 스패닝 트리에 기초하여 각 출력 채널 간 연산 순서를 재정렬하기 위한 명령을 포함하는, 컴퓨팅 장

치.

청구항 10

청구항 9에 있어서,
상기 완전 연결 그래프는,
상기 출력 채널 간 거리에 기초하여 생성하는, 컴퓨팅 장치.

청구항 11

청구항 10에 있어서,
상기 출력 채널은, 0 또는 1로 구성되는 가중치들의 목록을 포함하고,
상기 출력 채널 간 거리는, 해밍 거리(hamming distance)를 통해 산출하는, 컴퓨팅 장치.

청구항 12

청구항 11에 있어서,
상기 완전 연결 그래프는,
각 출력 채널을 노드로 하고, 각 출력 채널 간 해밍 거리를 엣지로 하여 생성되는, 컴퓨팅 장치.

청구항 13

청구항 12에 있어서,
상기 연산 순서를 재정렬하기 위한 명령은,
상기 최소 스패닝 트리에서 최상위 노드에 해당하는 출력 채널에 대해 가장 먼저 연산을 수행하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 14

청구항 13에 있어서,
상기 연산 순서를 재정렬하기 위한 명령은,
상기 최소 스패닝 트리의 상기 최상위 노드에서 해밍 거리가 짧은 하위 노드를 따라 연산을 수행하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 15

청구항 14에 있어서,
상기 하나 이상의 프로그램들은,
i번째 출력 채널로부터 j번째 출력 채널을 계산하기 위해 다음 수학적식을 사용하는, 컴퓨팅 장치.
(수학적식)

$$Y_j = 2(P_i - d_{ij} + 2P_{ij}) - C_{in} \times M \times M$$

P_i : i 번째 출력 채널의 팝 카운트(Popcount) 연산

d_{ij} : i 번째 출력 채널과 j 번째 출력 채널 사이의 해밍 거리

P_{ij} : i 번째 출력 채널과 j 번째 출력 채널 간 XNOR 연산의 합

C_{in} : 이진 신경망의 입력 채널

M : 커널 사이즈

청구항 16

청구항 9에 있어서,

상기 하나 이상의 프로그램들은,

상기 이진 신경망의 컨볼루션 계층의 각 출력 채널의 가중치들을 클러스터링 하기 위한 명령;

상기 클러스터링 된 각 클러스터에서 중심을 랜덤하게 선택하기 위한 명령; 및

상기 각 클러스터의 중심과 해당 클러스터에 속하는 각 가중치 간의 거리가 최소화 되도록 상기 이진 신경망을 학습하기 위한 명령을 더 포함하는, 컴퓨팅 장치.

청구항 17

비일시적 컴퓨터 판독 가능한 저장 매체(non-transitory computer readable storage medium)에 저장된 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은 하나 이상의 명령어들을 포함하고, 상기 명령어들은 하나 이상의 프로세서들을 갖는 컴퓨팅 장치에 의해 실행될 때, 상기 컴퓨팅 장치로 하여금,

이진 신경망의 컨볼루션 계층의 각 출력 채널에 기초하여 완전 연결 그래프를 생성하는 단계;

상기 완전 연결 그래프에서 최소 스패닝 트리를 추출하는 단계; 및

상기 최소 스패닝 트리에 기초하여 각 출력 채널 간 연산 순서를 재정렬하는 단계를 수행하도록 하는, 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명의 실시예는 이진 신경망의 연산 기술과 관련된다.

배경 기술

[0003] 심층 신경망(Deep Neural Network: DNN)은 컴퓨터 비전 및 음성 인식과 같은 많은 인공지능 응용분야에서 널리 적용되어 왔다. 그러나, 계산 복잡성과 막대한 용량 부담의 비용으로 인해 심층 신경망(DNN)을 모바일 장치 및 기타 자원 제약 플랫폼과 같은 임베디드 시스템에 적용하기에는 제약사항이 많다는 문제점이 있다.

[0004] 한편, 이진 신경망(Binary Neural Network: BNN)은 신경망의 가중치와 활성화를 1 비트 값으로 변환하는 특수한 형태의 양자화 방법으로서, 곱셈과 누적 연산을 XNOR(exclusive NOR) 연산과 팝 카운트(Popcount)와 같은 간단한 논리 연산으로 대체하기 때문에, 전력과 자원을 절감할 수 있게 된다. 그러나, 정확도 저하를 줄이면서 실용적인 요구사항을 충족하기 위해 인공 신경망이 더욱 깊어지고 넓어질수록 계산에 대한 부담은 이진 신경망에서도 여전히 해결해야 할 과제로 남게 된다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국공개특허공보 제10-2022-0090078호(2022.06.29)

발명의 내용

해결하려는 과제

[0007] 본 발명의 실시예는 이진 신경망의 연산 속도를 높이면서 연산 비용을 줄일 수 있는 이진 신경망의 연산 방법 및 이를 수행하기 위한 컴퓨팅 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0009] 개시되는 일 실시예에 따른 이진 신경망의 연산 방법은, 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로, 이진 신경망(Binary Neural Network)의 컨볼루션 계층의 각 출력 채널에 기초하여 완전 연결 그래프(fully connected graph)를 생성하는 단계; 상기 완전 연결 그래프에서 최소 스패닝 트리(Minimum Spanning Tree)를 추출하는 단계; 및 상기 최소 스패닝 트리에 기초하여 각 출력 채널 간 연산 순서를 재정렬하는 단계를 포함한다.

[0010] 상기 완전 연결 그래프는, 상기 출력 채널 간 거리에 기초하여 생성할 수 있다.

[0011] 상기 출력 채널은, 0 또는 1로 구성되는 가중치들의 목록을 포함하고, 상기 출력 채널 간 거리는, 해밍 거리(hamming distance)를 통해 산출할 수 있다.

[0012] 상기 완전 연결 그래프는, 각 출력 채널을 노드로 하고, 각 출력 채널 간 해밍 거리를 에지로 하여 생성될 수 있다.

[0013] 상기 연산 순서를 재정렬하는 단계는, 상기 최소 스패닝 트리에서 최상위 노드에 해당하는 출력 채널에 대해 가장 먼저 연산을 수행하는 단계를 포함할 수 있다.

[0014] 상기 연산 순서를 재정렬하는 단계는, 상기 최소 스패닝 트리의 상기 최상위 노드에서 해밍 거리가 짧은 하위 노드를 따라 연산을 수행하는 단계를 더 포함할 수 있다.

[0015] 상기 컴퓨팅 장치는, i 번째 출력 채널로부터 j 번째 출력 채널을 계산하기 위해 다음 수학적식을 사용할 수 있다.

[0016] (수학적식)

$$Y_j = 2(P_i - d_{ij} + 2P_{ij}) - C_{in} \times M \times M$$

[0017]

[0018] P_i : i 번째 출력 채널의 팝 카운트(Popcount) 연산

[0019] d_{ij} : i 번째 출력 채널과 j 번째 출력 채널 사이의 해밍 거리

[0020] P_{ij} : i 번째 출력 채널과 j 번째 출력 채널 간 XNOR 연산의 합

[0021] C_{in} : 이진 신경망의 입력 채널

[0022] M : 커널 사이즈

[0023] 상기 이진 신경망의 연산 방법은, 상기 이진 신경망의 컨볼루션 계층의 각 출력 채널의 가중치들을 클러스터링하는 단계; 상기 클러스터링 된 각 클러스터에서 중심을 랜덤하게 선택하는 단계; 및 상기 각 클러스터의 중심과 해당 클러스터에 속하는 각 가중치 간의 거리가 최소화 되도록 상기 이진 신경망을 학습하는 단계를 더 포함할 수 있다.

[0024] 개시되는 일 실시예에 따른 컴퓨팅 장치는, 하나 이상의 프로세서들; 메모리; 및 하나 이상의 프로그램들을 포함하고, 상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며, 상기 하나 이상의 프로그램들은, 이진 신경망의 컨볼루션 계층의 각 출력 채널에 기초하여 완전 연결 그래프를 생성하기 위한 명령; 상기 완전 연결 그래프에서 최소 스패닝 트리를 추출하기 위한 명령; 및 상기 최소 스패닝 트리에 기초하여 각 출력 채널 간 연산 순서를 재정렬하기 위한 명령을 포함한다.

발명의 효과

[0026] 개시되는 실시예에 의하면, 이진 신경망의 컨볼루션 계층의 각 출력 채널에 대해 완전 연결 그래프를 구성하고, 완전 연결 그래프에서 최소 스패닝 트리를 추출하며, 최소 스패닝 트리에 기초하여 출력 채널들의 연산 순서를 재정렬함으로써, 이진 신경망에서 연산 속도를 높이면서 연산 비용을 줄일 수 있게 된다.

도면의 간단한 설명

[0028] 도 1은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도

도 2는 본 발명의 일 실시예에 따른 이진 신경망의 연산 방법을 나타낸 흐름도

도 3은 본 발명의 일 실시예에 따른 이진 신경망의 컨볼루션 계층에서 계산 순서를 재정렬하는 상태를 나타낸 도면

도 4는 기존의 이진 신경망 연산과 본 발명의 일 실시예에 따른 이진 신경망 연산을 비교한 도면

발명을 실시하기 위한 구체적인 내용

[0029] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0030] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.

[0031] 또한, 제1, 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로 사용될 수 있다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[0033] 도 1은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경(10)을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0034] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 이진 신경망(Binary Neural Network: BNN)의 연산을 수행하기 위한 장치일 수 있다. 컴퓨팅 장치(12)는 이진 신경망에서 연산 속도를 높이기 위한 장치일 수 있다.

- [0035] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.
- [0036] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 장치(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.
- [0037] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.
- [0038] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.
- [0040] 도 2는 본 발명의 일 실시예에 따른 이진 신경망의 연산 방법을 나타낸 흐름도이다. 도시된 흐름도에서는 상기 방법을 복수 개의 단계로 나누어 기재하였으나, 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 또는 도시되지 않은 하나 이상의 단계가 부가되어 수행될 수 있다. 또한, 도 3은 본 발명의 일 실시예에 따른 이진 신경망의 컨볼루션 계층(convolution layer)에서 계산 순서를 재정렬하는 상태를 나타낸 도면이다. 도 3에서는, 이진 컨볼루션 계층의 출력 채널이 4개인 경우를 도시하였으며, 커널 사이즈는 3×3 인 경우를 나타내었다.
- [0041] 도 2 및 도 3을 참조하면, 컴퓨팅 장치(12)는 이진 신경망의 컨볼루션 계층의 출력 채널에 기초하여 완전 연결 그래프(fully connected graph)를 생성할 수 있다(S 101). 컴퓨팅 장치(12)는 이진 신경망에서 컨볼루션 계층의 출력 채널(채널 1, 채널 2, 채널 3, 채널 4) 간 거리에 기초하여 완전 연결 그래프를 생성할 수 있다. 각 출력 채널은 커널 사이즈에 대응하는 개수(즉, $3 \times 3 = 9$)의 가중치 목록들로 구성될 수 있다. 이때, 가중치(weight)는 0 또는 1로 구성될 수 있다.
- [0042] 구체적으로, 컴퓨팅 장치(12)는 컨볼루션 계층의 출력 채널 간 거리에 기초하여 인접 매트릭스(adjacent matrix)를 생성할 수 있다. 일 실시예에서, 출력 채널 간 거리는 해밍 거리(hamming distance)를 통해 산출할 수 있다. 해밍 거리는 같은 비트 수를 갖는 2진 부호 사이에 대응되는 비트 값이 일치하지 않는 개수를 나타내는 것이다.
- [0043] 예를 들어, 도 3에서 출력 채널 1과 출력 채널 2 사이의 해밍 거리는 $5(d_{12}=5)$ 가 된다. 즉, 출력 채널 1과 출력 채널 2 사이에 비트 값이 일치하지 않는 개수는 5이다. 출력 채널 1은 {100111011}의 비트 값을 가지고, 출력 채널 2는 {111100111}의 비트 값을 가지는데, 2번째, 3번째, 5번째, 6번째, 및 7번째 비트에서 그 값이 일치하지 않으므로 출력 채널 1과 출력 채널 2 사이의 해밍 거리는 5가 된다. 그리고, 출력 채널 1과 출력 채널 3 사이의 해밍 거리는 $4(d_{13}=4)$ 가 된다. 또한, 출력 채널 1과 출력 채널 4 사이의 해밍 거리는 $2(d_{14}=2)$ 가 된다. 이와 같이, 컴퓨팅 장치(12)는 각 출력 채널 간 해밍 거리를 산출하고, 각 출력 채널 간 해밍 거리 값을 행렬 요

소로 하여 4×4의 인접 매트릭스를 생성할 수 있다.

- [0044] 여기서, 컴퓨팅 장치(12)는 출력 채널 간 거리에 따른 인접 매트릭스에 기초하여 완전 연결 그래프를 생성할 수 있다. 컴퓨팅 장치(12)는 각 출력 채널을 노드(node)로 하고, 각 출력 채널 간 거리를 에지(edge)로 하여 각 출력 채널들이 완전히 연결된 완전 연결 그래프를 생성할 수 있다.
- [0045] 다음으로, 컴퓨팅 장치(12)는 완전 연결 그래프에서 최소 스패닝 트리(Minimum Spanning Tree)를 추출할 수 있다(S 103). 여기서, 최소 스패닝 트리는 각 에지에 가중치(여기서는 해밍 거리)가 있는 무방향 그래프에서 모든 노드들을 연결하는 가중치의 합이 최소가 되는 신장 트리를 의미한다.
- [0046] 컴퓨팅 장치(12)는 그래프에서 최소 스패닝 트리를 추출하기 위한 기 설정된 알고리즘(예를 들어, Kruskal 알고리즘, Prim 알고리즘 등)을 이용하여 완전 연결 그래프에서 최소 스패닝 트리를 추출할 수 있다. 도 3에서는 노드 4(즉, 채널 4)를 중심으로 모든 노드를 연결하는 가중치(즉, 해밍 거리)의 합(2+3+2 = 8)이 최소가 되는 최소 스패닝 트리를 추출한 상태를 도시하였다.
- [0047] 다음으로, 컴퓨팅 장치(12)는 완전 연결 그래프의 최소 스패닝 트리에 기초하여 각 출력 채널 간 연산 순서를 재정렬할 수 있다(S 105).
- [0048] 구체적으로, 컴퓨팅 장치(12)는 최소 스패닝 트리에서 최상위 노드(루트 노드)에 해당하는 출력 채널에 대해 가장 먼저 연산을 수행할 수 있다. 여기서, 최상위 노드는 최소 스패닝 트리에서 인접 노드와 연결되는 개수가 가장 많은 노드일 수 있다. 도 3에서는 노드 4가 최상위 노드가 될 수 있다.
- [0049] 컴퓨팅 장치(12)는 최소 스패닝 트리의 최상위 노드에서 해밍 거리가 짧은 하위 노드를 따라 연산을 수행할 수 있다. 이때, 하위 노드와의 해밍 거리가 동일한 경우 각 하위 노드는 동시에 연산이 수행될 수 있다. 도 3을 참조하면, 최상위 노드인 노드 4(출력 채널 4)가 연산이 된 후, 노드 4로부터 해밍 거리가 짧은 노드 1(출력 채널 1) 및 노드 3(출력 채널 3)이 동시에 연산이 되고, 노드 4로부터 해밍 거리가 가장 긴 노드 2(출력 채널 2)가 가장 나중에 연산이 될 수 있다.
- [0050] 컴퓨팅 장치(12)는 i번째 출력 채널로부터 j번째 출력 채널을 계산하기 위해 다음 수학적 식 1을 사용할 수 있다.
- [0051] (수학적 식 1)
- [0052]
$$Y_j = 2(P_i - d_{ij} + 2P_{ij}) - C_{in} \times M \times M$$
- [0053] P_i : i번째 출력 채널의 팝 카운트(Popcount) 연산
- [0054] d_{ij} : i번째 출력 채널과 j번째 출력 채널 사이의 해밍 거리
- [0055] P_{ij} : i번째 출력 채널과 j번째 출력 채널 간 XNOR 연산의 합
- [0056] C_{in} : 이진 신경망의 입력 채널
- [0057] M : 커널 사이즈
- [0058] 개시되는 실시예에 의하면, 이진 신경망의 컨볼루션 계층의 각 출력 채널에 대해 완전 연결 그래프를 구성하고, 완전 연결 그래프에서 최소 스패닝 트리를 추출하며, 최소 스패닝 트리에 기초하여 출력 채널들의 연산 순서를 재정렬함으로써, 이진 신경망에서 연산 속도를 높이면서 연산 비용을 줄일 수 있게 된다.
- [0060] 도 4는 기존의 이진 신경망 연산과 본 발명의 일 실시예에 따른 이진 신경망 연산을 비교한 도면이다. 도 4의 (a)는 기존의 이진 신경망 연산을 나타낸 도면이고, 도 4의 (b)는 본 발명의 일 실시예에 따른 이진 신경망 연산을 나타낸 도면이다.
- [0061] 도 4를 참조하면, 본 발명의 일 실시예에 따른 이진 신경망(Our BNN)의 연산은 각 레이어에 대한 모든 $C_{in} \times M \times M$ XNOR 연산을 모두 계산하는 대신 최소 스패닝 트리(MST)를 사용하여 XNOR 연산 수와 전체 파라미터의 수를 최소화 하는 최적의 계산 순서를 제공하게 된다.
- [0062] 즉, 도 4에서 기존의 이진 신경망 연산의 경우, 전체 파라미터의 수는 $C_{out} \times C_{in} \times M \times M = 4 \times 1 \times 3 \times 3 = 36$ 개가 된다. 반면, 본 실시예에 따른 이진 신경망 연산의 전체 파라미터 개수는

$\sum_2^{C_{in} \times M \times M} d_{ij} + C_{in} \times M \times M = 2 + 3 + 2 + 9 = 16$ 개로서 그 개수가 줄어들게 된다. 여기서, C_{in} 은 입력 채널을 나타내고, C_{out} 은 출력 채널을 나타내며, M 은 커널 사이즈를 나타내고, d_{ij} 는 출력 채널 i 와 출력 채널 j 사이의 해밍 거리를 나타낸다.

[0063] 또한, 기존의 이진 신경망 연산에서 XNOR 연산은 $H_{out} \times W_{out} \times C_{in} \times M \times M$ 이 되나, 본 실시예에 따른 이진 신경망 연산에서 XNOR 연산은 $H_{out} \times W_{out} \times (\sum_2^{C_{in} \times M \times M} d_{ij} + C_{in} \times M \times M)$ 이 된다. 여기서, $H_{out} \times W_{out}$ 는 출력 피쳐맵의 사이즈를 나타낸다.

[0064] 여기서, 기존의 이진 신경망 연산 대비 본 발명의 일 실시예에 따른 이진 신경망 연산의 압축 비율(R)은 다음 수학적 식 2로 나타낼 수 있다. 압축 비율은 전체 연산의 수가 기존의 이진 신경망 연산 대비 줄어든 비율을 나타낼 수 있다.

[0065] (수학적 식 2)

$$R = \frac{\sum_{j \neq root}^{C_{out}} d_{ij} + C_{in} \times M \times M}{C_{out} \times C_{in} \times M \times M}$$

[0068] 한편, 컴퓨팅 장치(12)는 이진 신경망의 컨볼루션 계층의 각 출력 채널에 대해 완전 연결 그래프를 구성하고, 완전 연결 그래프에서 최소 스패닝 트리(MST)를 추출하는 경우, 최소 스패닝 트리의 깊이와 총 거리를 더욱 줄일 수 있도록 하기 위해 이진 신경망을 학습할 수 있다.

[0069] 구체적으로, 컴퓨팅 장치(12)는 이진 신경망의 컨볼루션 계층의 각 출력 채널의 가중치들을 클러스터링 할 수 있다. 이때, 클러스터링은 기존의 다양한 클러스터링 기법들을 사용할 수 있다. 컴퓨팅 장치(12)는 가중치들의 각 클러스터에서 중심을 랜덤하게 선택할 수 있다. 컴퓨팅 장치(12)는 각 클러스터의 중심과 해당 클러스터에 속하는 각 가중치 간의 거리가 최소화 되도록 이진 신경망을 학습할 수 있다. 여기서, 이진 신경망에 대한 전체 손실 함수는 하기 수학적 식 3으로 나타낼 수 있다.

[0070] (수학적 식 3)

$$L = L_0(input, w_b) + \gamma \sum_i \|C(w_b^i) - w_b^i\|^2$$

[0072] L_0 : 이진 신경망의 원래 태스크에 따른 손실 함수

[0073] γ : 기 설정되는 하이퍼 파라미터

[0074] $C(w_b^i)$: i 번째 클러스터의 중심

[0075] w_b^i : i 번째 클러스터에 속하는 가중치

[0076] 수학적 식 3의 우측 항에서, 첫 번째 부분이 이진 신경망의 원래 태스크에 따른 손실 함수를 나타내고, 두 번째 부분이 최소 스패닝 트리(MST)의 깊이와 총 거리를 줄이기 위한 손실 함수를 나타낸다.

[0078] 이상에서 본 발명의 대표적인 실시예들을 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허 청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

부호의 설명

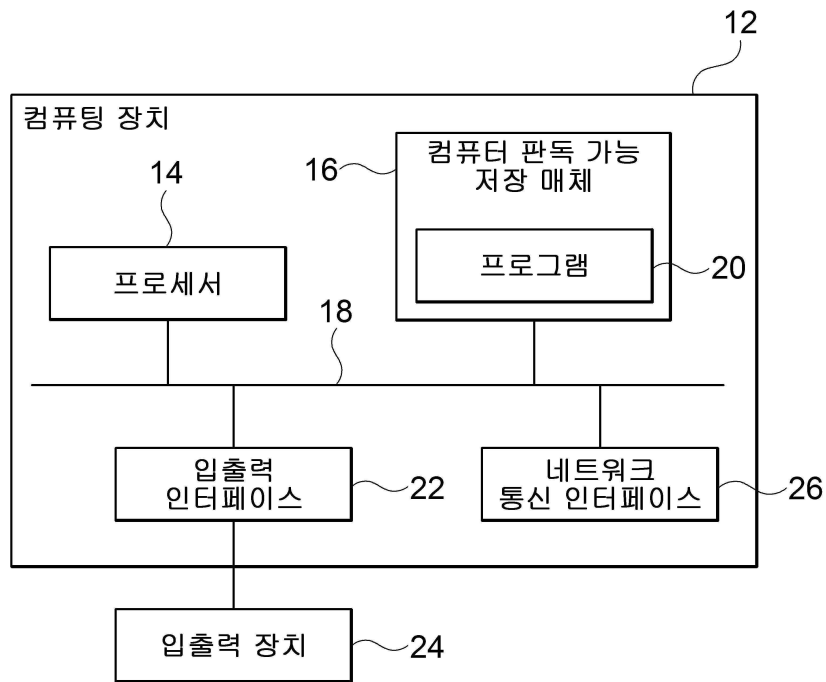
[0080]

- 10 : 컴퓨팅 환경
- 12 : 컴퓨팅 장치
- 14 : 프로세서
- 16 : 컴퓨터 판독 가능 저장 매체
- 18 : 통신 버스
- 20 : 프로그램
- 22 : 입출력 인터페이스
- 24 : 입출력 장치
- 26 : 네트워크 통신 인터페이스

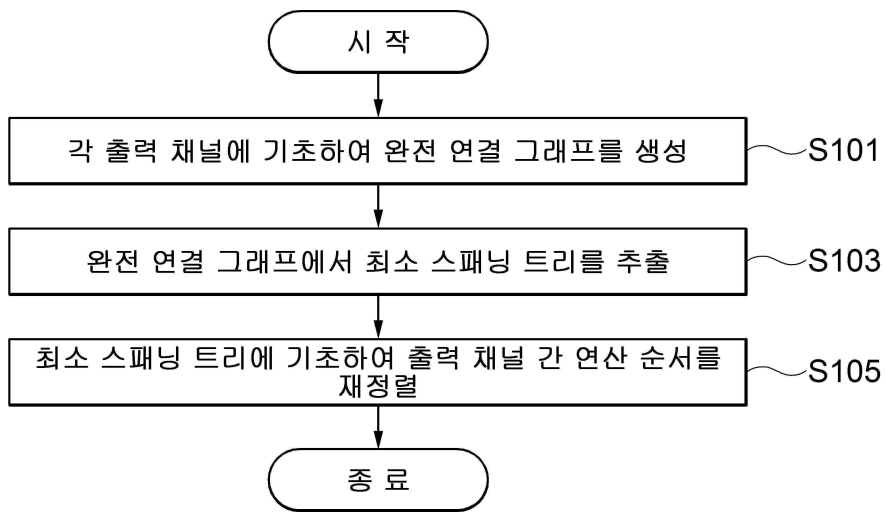
도면

도면1

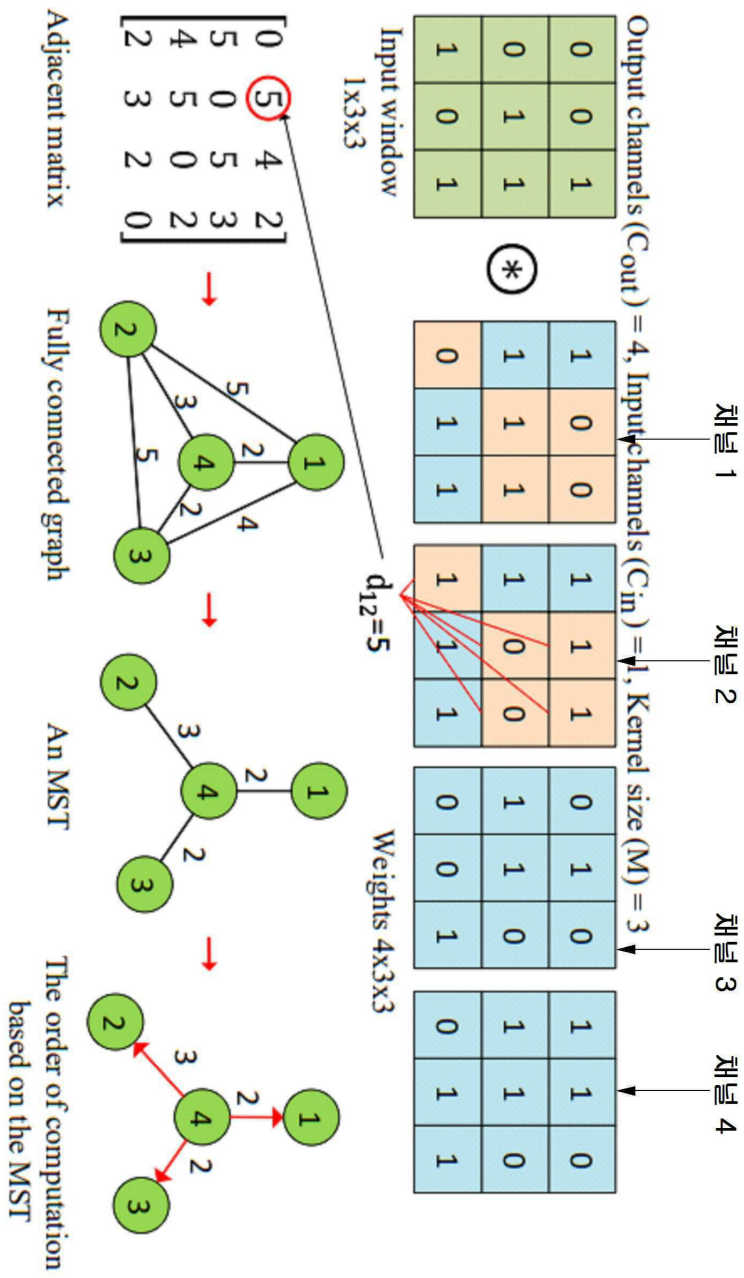
10



도면2



도면3



도면4

