



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0152640
(43) 공개일자 2024년10월22일

(51) 국제특허분류(Int. Cl.)
G06N 3/098 (2023.01) G06N 3/0464 (2023.01)
H04L 67/2885 (2022.01)
(52) CPC특허분류
G06N 3/098 (2023.01)
G06N 3/0464 (2023.01)
(21) 출원번호 10-2023-0048905
(22) 출원일자 2023년04월13일
심사청구일자 2023년04월13일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
고한얼
세종특별자치시 달빛로 80, 1216동 1903호 (종촌동, 가재마을 12단지)
백상현
서울특별시 서초구 고무래로 35, 109동 1004호 (반포동, 반포리체아파트)
김보경
경기도 수원시 영통구 청명남로 25, 1004호 (영통동)
(74) 대리인
두호특허법인

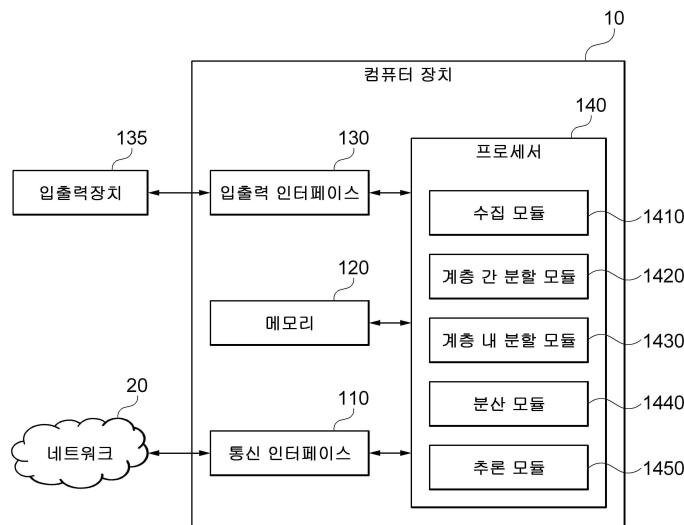
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 계층적 모바일 네트워크에서 기계학습 방법, 컴퓨터 장치 및 프로그램

(57) 요약

본 발명은 계층적 모바일 네트워크에서 계층 간 분할을 수행하는 분산 컴퓨팅 기술에 관한 것으로, 하나의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨터 장치에서 수행되는 방법으로서, 컴퓨팅 노드 정보를 수집하는 단계, DNN 모델의 계층(layer)간 분할 지점을 고려하여 분할하는 계층 내 분할 단계, DNN 모델의 계층 내 분할 지점을 고려하여 분할하는 계층 간 분할 단계 및 상기 수집된 컴퓨팅 노드 정보에 기반하여 상기 계층 간 분할 및 계층 내 분할된 DNN 모델의 모든 계층에 대한 실행 단위를 컴퓨팅 노드에 분산시키는 단계를 포함하는, 계층적 모바일 네트워크에서 기계학습 방법에 의해 추론 지연 시간을 최소화할 수 있다.

대표도



(52) CPC특허분류

H04L 67/2885 (2022.05)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193180
과제번호	2022-0-01015-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	6G핵심기술개발
연구과제명	지능형 6G 모바일 코어 네트워크 후보 요소기술 개발
기 여 율	1/1
과제수행기관명	승실대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

하나의 프로세서들, 및

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

컴퓨팅 노드 정보를 수집하는 단계;

DNN 모델의 계층(layer)간 분할 지점을 고려하여 분할하는 계층 내 분할 단계;

DNN 모델의 계층 내 분할 지점을 고려하여 분할하는 계층 간 분할 단계; 및

상기 수집된 컴퓨팅 노드 정보에 기반하여 상기 계층 간 분할 및 계층 내 분할된 DNN 모델의 모든 계층에 대한 실행 단위를 컴퓨팅 노드에 분산시키는 단계;를 포함하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 2

제 1 항에 있어서,

상기 수집하는 단계는,

참여 가능한 컴퓨팅 노드 정보 및 컴퓨팅 노드에 대한 가용 자원 정보를 수집하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 3

제 2 항에 있어서,

상기 계층 내 분할 단계는,

상기 수집하는 단계에서 수집된 컴퓨팅 노드 정보에 기반하여 컴퓨팅 노드의 가용 컴퓨팅 파워를 고려하여 분할하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 4

제 2 항에 있어서,

상기 계층 간 분할 단계는,

그래프 기반의 계층 간 분할 알고리즘을 이용하여 계층간 분할 지점을 결정하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 5

제 4 항에 있어서,

상기 계층 간 분할 알고리즘은,

상기 수집하는 단계에서 수집된 컴퓨팅 노드 정보에 기반하여 엣지 비용(edge cost)를 계산하고, 최단 경로 알고리즘을 기반으로 소스에서 목적지까지 최단 경로를 도출하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 6

제 5 항에 있어서,

상기 계층 간 분할 단계는,

상기 계층 간 분할 알고리즘에 의해 도출된 최단 경로를 기반으로 계층 간 분할 지점을 결정하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 7

제 4 항에 있어서,

그래프 기반의 그래프는 정점의 집합(V)과 엣지의 집합(E)으로 구성되며, 정점의 집합(V)은 하기 수학적 식 1과 같이 정의되고,

[수학적식 1]

$$V = \{v_l^t \mid \forall t, \forall l\}$$

여기서 v_l^t 는 계층 l 이전의 DNN 실행이 Tier l 이전 또는 Tier l 에서 완료되었음을 나타내는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 8

제 7 항에 있어서,

엣지의 집합(E)은 하기 수학적 식 2와 같이 정의되는, 계층적 모바일 네트워크에서 기계학습 방법.

[수학적식 2]

$$E = \left\{ (v_l^t, v_{l'}^{t'}) \mid t \leq t' \leq t+1 \leq T, l \leq l' \leq l+1 \leq L \right\} - \left\{ (v_l^t, v_{l'}^{t'}) \mid t = t', l = l' \right\} \cup \left\{ (v_l^t, v_D) \mid l = L \right\}$$

청구항 9

제 7 항에 있어서,

엣지 비용(edge cost)은 하기 수학적 식 3과 같이 정의되고,

[수학적식 3]

$$c(v_l^t, v_{l'}^{t'}) = \begin{cases} \tau_l^{tt'}, & \text{if } l = l', t \neq t' \\ \varepsilon_l^t, & \text{if } l \neq l', t = t' \\ \varepsilon_l^t + \tau_R, & \text{if } l = L \\ \varepsilon_l^t + \tau_l^{tt'}, & \text{otherwise} \end{cases}$$

여기서, $\tau_l^{tt'}$ 은 Tier l 에서 Tier l' 으로 계층 l 의 중간 데이터를 전송하기 위한 통신 지연, ε_l^t 는 계층 l 을 Tier l 에서 처리하는데 걸리는 지연시간, τ_R 은 DNN의 결과를 모바일 기기(MD)에게 전달하는데 걸리는 지연시간인, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 10

제 9 항에 있어서,

엣지 비용(edge cost)은 기본적으로 계층¹의 실행 단위가 각 컴퓨팅 노드의 가용 컴퓨팅 성능에 비례하여 할당된다는 가정하에 설정되는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 11

제 4 항에 있어서,

계층 간 분할 알고리즘은 그래프에서 소스에서 목적지까지의 최단 경로를 찾아 계층간 분할 지점을 결정하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 12

제 2 항에 있어서,

상기 분산시키는 단계는,

상기 수집된 컴퓨팅 노드 정보에 기반하여 컴퓨팅 노드의 계산 능력 또는 계층 간 링크 전송 속도를 반영하여 컴퓨팅 노드에 분산시키는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 13

제 1 항에 있어서,

상기 분산시키는 단계에서 실행 단위를 분산 받은 컴퓨팅 노드가 할당받은 실행 단위를 수행하여 추론하는 단계;를 더 포함하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 14

제 13 항에 있어서,

상기 추론하는 단계는,

컴퓨팅 노드가 실행 단위를 추론 완료 후 결과를 공유하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 15

제 13 항에 있어서,

상기 추론하는 단계는,

합성곱 레이어(convolutional layer)의 경우 분할된 이웃 특성맵(feature map)이 각 특성 맵(feature map)의 가장자리에 채워지도록 패딩을 수행하는, 계층적 모바일 네트워크에서 기계학습 방법.

청구항 16

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨터 장

치로서,

컴퓨팅 노드 정보를 수집하는 수집 모듈;

DNN 모델의 계층(layer)간 분할 지점을 고려하여 분할하는 계층 간 분할 모듈;

DNN 모델의 계층 내 분할 지점을 고려하여 분할하는 계층 내 분할 모듈; 및

상기 수집된 컴퓨팅 노드 정보에 기반하여 상기 계층 간 분할 및 계층 내 분할된 DNN 모델의 모든 계층에 대한 실행 단위를 컴퓨팅 노드에 분산시키는 분산 모듈;을 포함하는, 컴퓨터 장치.

청구항 17

제 16 항에 있어서,

상기 분산 모듈에서 실행 단위를 분산 받은 컴퓨팅 노드가 할당받은 실행 단위를 수행하여 추론하는 추론 모듈;을 더 포함하는, 컴퓨터 장치.

청구항 18

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 컴퓨터 장치에 의해 실행될 때, 상기 컴퓨터 장치로 하여금,

컴퓨팅 노드 정보를 수집하는 단계;

DNN 모델의 계층 내 분할 지점을 고려하여 분할하는 계층 내 분할 단계;

DNN 모델의 계층(layer)간 분할 지점을 고려하여 분할하는 계층 간 분할 단계; 및

상기 수집된 컴퓨팅 노드 정보에 기반하여 상기 계층 간 분할 및 계층 내 분할된 DNN 모델의 모든 계층에 대한 실행 단위를 컴퓨팅 노드에 분산시키는 단계;를 수행하도록 하는, 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명은 계층적 모바일 네트워크에서 계층 간 분할을 수행하는 분산 컴퓨팅 기술과 관련된다.

배경 기술

[0003] 심층 신경망(DNN, Deep Neural Network)은 지능형 모바일 어플리케이션에 사용되는 기계 학습 접근법으로 여러 개의 은닉층을 포함하는 신경망 모델이다.

[0004] 계층적인 모바일 네트워크에서 계층 간 분할(Hierarchical Splitting) 기계 학습 프레임 워크는, 분산형 머신러닝을 사용하여 모바일 기기에서 수행되는 AI 어플리케이션의 성능을 향상시키기 위한 것이다.

[0005] 계층 분할은 일반적으로 네트워크에서 여러 번 반복되고 이를 통해 DNN의 성능이 향상될 수 있다. 그러나 계층 분할이 많은 파라미터와 계산을 필요로 하기 때문에 정확한 분할 방법과 모델 크기에 대한 균형이 필요하다.

[0006] 기존의 분할 컴퓨팅 접근 방식은 계층 간 분할을 수행한다. 그러나 계층 간 분할만 사용하면 계층적 모바일 네트워크에서 분산 컴퓨팅 노드의 리소스를 효율적으로 활용할 수 없다는 한계가 있다.

[0007] 최근 모바일 기기(MD)의 컴퓨팅 능력이 향상되고 있지만 여전히 클라우드에는 미치지 못한다는 한계가 있다. 따라서 전체 심층 신경망(DNN, Deep Neural Network) 모델에 대한 추론을 수행할 때 MD는 충분한 낮은 지연 시간을 달성할 수 없다.

선행기술문헌

특허문헌

- [0009] (특허문헌 0001) 대한민국 등록특허공보 제10-2194280호(2020.12.16)
(특허문헌 0002) 대한민국 등록특허공보 제10-2376527호(2022.03.15)

발명의 내용

해결하려는 과제

- [0010] 본 발명은 이 같은 기술적 배경에서 도출된 것으로, 심층 신경망(DNN, Deep Neural Network)의 한 계층이 여러 개의 독립적인 실행 단위로 분할되고 이러한 단위가 다른 컴퓨팅 노드에 분산되도록 하여 분산 컴퓨팅 노드의 자원을 충분히 활용할 수 있도록 함으로써 추론 지연 시간을 최소화하고자 함에 그 목적이 있다.

과제의 해결 수단

- [0012] 본 발명의 예시적인 실시예에 따르면, 하나의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨터 장치에서 수행되는 방법으로서, 컴퓨팅 노드 정보를 수집하는 단계, DNN 모델의 계층(layer)간 분할 지점을 고려하여 분할하는 계층 내 분할 단계, DNN 모델의 계층 내 분할 지점을 고려하여 분할하는 계층 간 분할 단계 및 상기 수집된 컴퓨팅 노드 정보에 기반하여 상기 계층 간 분할 및 계층 내 분할된 DNN 모델의 모든 계층에 대한 실행 단위를 컴퓨팅 노드에 분산시키는 단계를 포함하는 계층적 모바일 네트워크에서 기계학습 방법이 제공된다.
- [0013] 상기 수집하는 단계는 참여 가능한 컴퓨팅 노드 정보 및 컴퓨팅 노드에 대한 가용 자원 정보를 수집할 수 있다.
- [0014] 상기 계층 내 분할 단계는 상기 수집하는 단계에서 수집된 컴퓨팅 노드 정보에 기반하여 컴퓨팅 노드의 가용 컴퓨팅 파워를 고려하여 분할할 수 있다.
- [0015] 상기 계층 간 분할 단계는 그래프 기반의 계층 간 분할 알고리즘을 이용하여 계층간 분할 지점을 결정할 수 있다.
- [0016] 상기 계층 간 분할 알고리즘은 상기 수집하는 단계에서 수집된 컴퓨팅 노드 정보에 기반하여 엣지 비용(edge cost)를 계산하고, 최단 경로 알고리즘을 기반으로 소스에서 목적지까지 최단 경로를 도출할 수 있다.
- [0017] 상기 계층 간 분할 단계는 상기 계층 간 분할 알고리즘에 의해 도출된 최단 경로를 기반으로 계층 간 분할 지점을 결정할 수 있다.
- [0018] 그래프 기반의 그래프는 정점의 집합(V)과 엣지의 집합(E)으로 구성되며, 정점의 집합(V)은 하기 수학적 식 1과 같이 정의되고,

[0019] [수학적식 1]

$$V = \{v_l^t \mid \forall t, \forall l\}$$

[0021] 여기서 v_l^t 는 계층 l 이전의 DNN 실행이 Tier l 이전 또는 Tier l 에서 완료되었음을 나타낼 수 있다.

[0022] 엣지의 집합(E)은 하기 수학적 식 2와 같이 정의될 수 있다.

[0023] [수학적식 2]

$$E = \left\{ (v_l^t, v_{l'}^{t'}) \mid t \leq t' \leq t+1 \leq T, l \leq l' \leq l+1 \leq L \right\} - \left\{ (v_l^t, v_{l'}^{t'}) \mid t = t', l = l' \right\} \cup \{ (v_l^t, v_D) \mid l = L \}$$

[0024]

[0025] 엣지 비용(edge cost)은 하기 수학적 식 3과 같이 정의되고,
 [0026] [수학적 식 3]

$$c(v_l^t, v_{l'}^{t'}) = \begin{cases} \tau_l^{tt'}, & \text{if } l = l', t \neq t' \\ \varepsilon_l^t, & \text{if } l \neq l', t = t' \\ \varepsilon_l^t + \tau_R, & \text{if } l = L \\ \varepsilon_l^t + \tau_l^{tt'}, & \text{otherwise} \end{cases}$$

[0027]
 [0028] 여기서, $\tau_l^{tt'}$ 은 Tier l 에서 Tier t' 으로 계층 l 의 중간 데이터를 전송하기 위한 통신 지연, ε_l^t 는 계층 l 을 Tier t 에서 처리하는데 걸리는 지연시간, τ_R 은 DNN의 결과를 모바일 기기(MD)에게 전달하는데 걸리는 지연시간일 수 있다.

[0029] 엣지 비용(edge cost)은 기본적으로 계층 l 의 실행 단위가 각 컴퓨팅 노드의 가용 컴퓨팅 성능에 비례하여 할당된다는 가정하에 설정될 수 있다.

[0030] 계층 간 분할 알고리즘은 그래프에서 소스에서 목적지까지의 최단 경로를 찾아 계층간 분할 지점을 결정할 수 있다.

[0031] 상기 분산시키는 단계는 상기 수집된 컴퓨팅 노드 정보에 기반하여 컴퓨팅 노드의 계산 능력 또는 계층 간 링크 전송 속도를 반영하여 컴퓨팅 노드에 분산시킬 수 있다.

[0032] 상기 분산시키는 단계에서 실행 단위를 분산 받은 컴퓨팅 노드가 할당받은 실행 단위를 수행하여 추론하는 단계를 더 포함할 수 있다.

[0033] 상기 추론하는 단계는 컴퓨팅 노드가 실행 단위를 추론 완료 후 결과를 공유하되, 동일 계층에서 링크 용량이 최대인 컴퓨팅 노드가 해당 계층의 중간 데이터를 다음 계층의 컴퓨팅 노드로 전송할 수 있다.

[0034] 상기 추론하는 단계는 합성곱 레이어(convolutional layer)의 경우 분할된 이웃 특성맵(feature map)이 각 특성맵(feature map)의 가장자리에 채워지도록 패딩을 수행할 수 있다.

발명의 효과

[0036] 본 발명에 의하면 심층 신경망(DNN, Deep Neural Network)의 한 계층이 여러 개의 독립적인 실행 단위로 분할되고 이러한 단위가 다른 컴퓨팅 노드에 분산되도록 하여 분산 컴퓨팅 노드의 자원을 충분히 활용할 수 있도록 함으로써 추론 지연 시간을 최소화할 수 있는 효과가 도출된다.

도면의 간단한 설명

[0038] 도 1은 본 발명의 일 실시예에 따른 컴퓨터 장치의 구성을 나타낸 도면
 도 2는 본 발명의 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법을 수행하는 프레임워크의 예시도
 도 3는 본 발명의 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법을 도시한 흐름도
 도 4a 및 도 4b는 가정을 통한 간소화 전후 그래프의 예시도
 도 5a, 도 5b 및 도 5c는 계층 내 분할시 특성맵의 변형과정을 설명하기 위한 예시도
 도 6은 컴퓨팅 노드 수 증가에 따른 추론시간 변화를 도시한 그래프

발명을 실시하기 위한 구체적인 내용

[0039] 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.

- [0040] 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.
- [0041] 이하의 설명에 있어서, 신호 또는 정보의 "전송", "통신", "송신", "수신" 기타 이와 유사한 의미의 용어는 일 구성요소에서 다른 구성요소로 신호 또는 정보가 직접 전달되는 것뿐만이 아니라 다른 구성요소를 거쳐 전달되는 것도 포함한다. 특히 신호 또는 정보를 일 구성요소로 "전송" 또는 "송신"한다는 것은 그 신호 또는 정보의 최종 목적지를 지시하는 것이고 직접적인 목적지를 의미하는 것이 아니다. 이는 신호 또는 정보의 "수신"에 있어서도 동일하다. 또한 본 명세서에 있어서, 2 이상의 데이터 또는 정보가 "관련"된다는 것은 하나의 데이터(또는 정보)를 획득하면, 그에 기초하여 다른 데이터(또는 정보)의 적어도 일부를 획득할 수 있음을 의미한다.
- [0042] 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.
- [0044] 이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명한다.
- [0045] 도 1은 본 발명의 일 실시예에 따른 컴퓨터 장치의 구성을 나타낸 도면이다.
- [0046] 본 발명의 실시예들에 따른 계층적 모바일 네트워크에서 기계학습 방법은 적어도 하나의 컴퓨터 장치를 통해 수행될 수 있다. 이때, 컴퓨터 장치에는 본 발명의 일 실시예에 따른 컴퓨터 프로그램이 설치 및 구동될 수 있고, 컴퓨터 장치는 구동된 컴퓨터 프로그램의 제어에 따라 본 발명의 실시예들에 따른 계층적 모바일 네트워크에서 기계학습 방법을 수행할 수 있다. 상술한 컴퓨터 프로그램은 컴퓨터 장치와 결합되어 계층적 모바일 네트워크에서 기계학습 방법을 컴퓨터에 실행시키기 위해 컴퓨터 판독 가능한 기록매체에 저장될 수 있다.
- [0047] 통신 인터페이스(110)는 네트워크(20)를 통해 컴퓨터 장치(10)가 다른 장치와 서로 통신하기 위한 기능을 제공할 수 있다. 일례로, 컴퓨터 장치(10)의 프로세서(140)가 메모리(120)와 같은 기록 장치에 저장된 프로그램 코드에 따라 생성한 요청이나 명령, 데이터, 파일 등이 통신 인터페이스(110)의 제어에 따라 네트워크(20)를 통해 다른 장치들로 전달될 수 있다. 역으로, 다른 장치로부터의 신호나 명령, 데이터, 파일 등이 네트워크(20)를 거쳐 컴퓨터 장치(10)의 통신 인터페이스(110)를 통해 컴퓨터 장치(10)로 수신될 수 있다. 통신 인터페이스(110)를 통해 수신된 신호나 명령, 데이터 등은 프로세서(140)나 메모리(120)로 전달될 수 있고, 파일 등은 컴퓨터 장치(10)가 더 포함할 수 있는 저장 매체(상술한 영구 저장 장치)로 저장될 수 있다.
- [0048] 메모리(120)는 컴퓨터에서 판독 가능한 기록매체로서, RAM(random access memory), ROM(read only memory) 및 디스크 드라이브와 같은 비소멸성 대용량 기록장치(permanent mass storage device)를 포함할 수 있다. 여기서 ROM과 디스크 드라이브와 같은 비소멸성 대용량 기록장치는 메모리(120)와는 구분되는 별도의 영구 저장 장치로서 컴퓨터 장치(10)에 포함될 수도 있다.
- [0049] 또한, 메모리(120)에는 운영체제와 적어도 하나의 프로그램 코드가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 메모리(120)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 메모리(120)로 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 인터페이스(110)를 통해 메모리(120)에 로딩될 수도 있다. 예를 들어, 소프트웨어 구성요소들은 네트워크(20)를 통해 수신되는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여

컴퓨터 장치(10)의 메모리(120)에 로딩될 수 있다.

- [0050] 입출력 인터페이스(130)는 입출력 장치(135)와의 인터페이스를 위한 수단일 수 있다. 예를 들어, 입력 장치는 마이크, 키보드 또는 마우스 등의 장치를, 그리고 출력 장치는 디스플레이, 스피커와 같은 장치를 포함할 수 있다. 다른 예로 입출력 인터페이스(130)는 터치스크린과 같이 입력과 출력을 위한 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수도 있다. 입출력 장치(135)는 컴퓨터 장치(10)와 하나의 장치로 구성될 수도 있다.
- [0052] 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법은 DNN 모델을 두 종류의 분할 지점 즉, 계층 간 분할 지점과 계층 내 분할 지점을 고려하여 나누고 여러 컴퓨팅 노드에 적절히 분산시켜 추론 지연 시간을 최소화할 수 있다.
- [0053] 도 2는 본 발명의 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법을 수행하는 프레임워크의 예시 도이다.
- [0054] 도 2는 컴퓨팅 노드가 여러 티어(Tier)로 구축된 시스템 모델을 도시한 것이다.
- [0055] 구체적으로, T 개의 티어가 있다고 가정하면, 티어 1(Tier 1)에서 모바일 기기(MD)는 컴퓨팅 노드의 역할을 한다. 한편, radio unit(RU)는 안테나일 뿐 계산을 수행할 수 없으므로, 컴퓨팅 노드로 간주되지 않는다. 티어 2와 티어 3에서는 각각 분산 유닛(distributed unit, DU) 및 중앙 유닛(central unit, CU)이 각각 컴퓨팅 노드의 역할을 수행할 수 있다. 티어(Tier) 4부터 티어(Tier) T까지는 다양한 유형의 클라우드(예: �지 및 중앙 클라우드)가 설치되어 컴퓨팅 노드로 활용될 수 있다.
- [0056] 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법을 수행하는 프레임워크는 DNN 모델의 추론(inference) 지연을 줄이기 분할 컴퓨팅(split computing)을 활용한다. 즉, L개의 레이어를 가진 DNN을 여러 실행 단위로 분할하고 이러한 실행 단위들을 컴퓨팅 노드(모바일 기기, 분산 유닛, 중앙 유닛, �지 클라우드, 중앙 클라우드)에 분산시킨다. 분산 컴퓨팅 노드의 자원을 완전히 활용하여 추론 지연을 효율적으로 줄이기 위해, 두 가지 종류의 분할(즉, 계층 간 분할과 계층 내 분할)이 수행된다.
- [0057] 그리고 최적의 분할을 위해, 중앙 클라우드는 계층 간 링크의 전송 속도와 컴퓨팅 노드의 가용 컴퓨팅 파워와 같은 몇 가지 정보를 수집한다. 그런 다음 수집된 정보를 기반으로 중앙 클라우드는 그래프 기반의 계층 간 분할 알고리즘(GRAPH-BASED INTER-LAYER SPLITTING ALGORITHM; GISA)을 사용하여 최적의 계층 간 분할 지점을 결정할 수 있다. 한편, 계층 내 분할의 경우, 특정 레이어의 실행 단위는 계층 내 각 컴퓨팅 노드의 가용 컴퓨팅 파워에 비례하여 분할된다.
- [0058] 그리고 계층 간 및 계층 내 분할 지점을 결정한 후, 중앙 클라우드는 실행 단위 할당 메시지를 전송하여 실행 단위를 컴퓨팅 노드에 할당한다. 실행 단위 할당 메시지에는 계층 간 및 계층 내 분할 정책이 포함된다.
- [0059] 즉, 분산 컴퓨팅 노드의 자원을 최대한 활용하여 추론 지연 시간을 효율적으로 줄이기 위해 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법을 수행하는 프레임워크는 계층 간 분할 및 계층 내 분할 즉 두 가지 분할을 모두 수행한다.
- [0060] 도 3 은 본 발명의 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법을 도시한 흐름도이다.
- [0061] 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법은 컴퓨팅 노드 정보를 수집한다(S200).
- [0062] 수집하는 단계는 참여 가능한 컴퓨팅 노드 정보 및 컴퓨팅 노드에 대한 가용 자원 정보를 수집한다
- [0063] 한편, 계층 간 분할을 위해 중앙 클라우드는 계층 간 링크의 전송 속도 및 컴퓨팅 노드의 사용 가능한 컴퓨팅 성능과 같은 일부 정보를 수집한다.
- [0064] 그리고 DNN 모델의 계층 내 분할 지점을 고려하여 분할한다(S210).
- [0065] 계층 내 분할 단계는 수집하는 단계에서 수집된 컴퓨팅 노드 정보에 기반하여 컴퓨팅 노드의 가용 컴퓨팅 파워를 고려하여 분할한다. 컴퓨팅 노드의 가용 컴퓨팅 파워를 고려하여 계층 내 분할점을 결정한다. 즉, 각 컴퓨팅 노드의 가용 컴퓨팅 성능에 비례하여 특정 계층의 실행 단위를 컴퓨팅 노드에 나누어 할당한다. 구체적으로, 계층 내 분할 대상 데이터의 크기를 측정하여 분할 대상 데이터가 각 노드의 메모리에 로드될 수 있는지 확인하는 과정을 수행한다. 그리고 계층 내 분할 대상 데이터의 처리량을 측정하여 각 노드에서 처리할 수 있는 최대 처

리량을 계산한다.

- [0066] 각 노드에서 실제로 사용 가능한 컴퓨팅 파워를 측정하는데, 이때 노드의 CPU 성능, 메모리 용량, 네트워크 대역폭 등을 고려할 수 있다.
- [0067] 그리고 분할 대상 데이터의 크기, 처리량 및 각 노드의 가용 컴퓨팅 파워를 종합하여 계층 내 분할점을 결정한다. 이 때, 분할 대상 데이터의 처리량이 가용 컴퓨팅 파워보다 높을 경우, 노드가 처리 능력을 제대로 발휘하지 못할 가능성이 있으므로 분할점을 더 작게 설정하고, 분할 대상 데이터의 처리량이 가용 컴퓨팅 파워보다 낮을 경우, 노드의 처리 능력이 남아있으므로 분할점을 더 크게 설정하도록 조절할 수 있다. 즉, 계층 내 분할 후 각 노드에서의 처리 속도를 측정하여 최적의 분할점을 찾을 수 있다.
- [0068] 이후에 DNN 모델의 계층(layer) 간 분할 지점을 고려하여 분할한다(S220).
- [0069] 계층 간 분할을 위해 중앙 클라우드는 계층 간 링크의 전송 속도 및 컴퓨팅 노드의 사용 가능한 컴퓨팅 성능과 같은 일부 정보를 수집한다.
- [0070] 한편, 계층 간 분할을 위해 중앙 클라우드는 계층 간 링크의 전송 속도 및 컴퓨팅 노드의 사용 가능한 컴퓨팅 성능과 같은 일부 정보를 수집한다.
- [0071] 그래프 기반의 계층 간 분할 알고리즘은 수집된 컴퓨팅 노드 정보에 기반하여 엣지 비용(edge cost)를 계산하고, 최단 경로 알고리즘을 기반으로 소스에서 목적지까지 최단 경로를 도출한다. 예를 들어, 최단 경로 알고리즘은 다익스트라(Dijkstra) 알고리즘일 수 있다. 그러나, 이에 한정되는 것은 아니다.
- [0072] 일 실시예에 있어서, 주어진 모바일 네트워크 및 DNN에 맞는 그래프를 구성하여 그래프 기반의 계층 간 분할 알고리즘에서 활용한다. 그래프 G는 정점의 집합 V와 엣지의 집합 E로 구성된다.
- [0073] 정점의 집합 V는 수학식 1과 같이 정의된다.

수학식 1

[0074]
$$V = \{v_i^t \mid \forall t, \forall l\}$$

- [0075] 여기서, 정점 v_i^t 는 계층 l 이전의 DNN 실행이 Tier t 이전 또는 Tier t에서 완료되었음을 나타낸다. 즉, v_1^t 는 Tier t 이전 또는 Tier t에서 어떠한 계층도 완료되지 않은 경우를 나타낸다.
- [0076] 그리고, 엣지의 집합 E는 다음 수학식 2와 같이 정의된다.

수학식 2

[0077]
$$E = \left\{ (v_i^t, v_{i'}^{t'}) \mid t \leq t' \leq t+1 \leq T, l \leq l' \leq l+1 \leq L \right\} - \left\{ (v_i^t, v_{i'}^{t'}) \mid t = t', l = l' \right\} \cup \left\{ (v_i^t, v_D) \mid l = L \right\}$$

- [0078] 여기서, T는 티어의 수, L은 계층의 수, v_D 는 목적지를 나타낸다. 두 개의 정점(vertices) 간 전이(엣지, edge)는 레이어의 추가 실행과 중간 데이터가 다음 티어로 전송되는 것을 나타낸다. 따라서 모든 정점에는 자신을 가리키는 엣지가 없다.
- [0079] 그리고 엣지 $(v_i^t, v_{i'}^{t'})$ 에서의 엣지 비용 $(c(v_i^t, v_{i'}^{t'}))$ 은 수학식 3과 같이 정의된다.

수학식 3

$$c(v_l^t, v_{l'}^{t'}) = \begin{cases} \tau_l^{tt'}, & \text{if } l = l', t \neq t' \\ \varepsilon_l^t, & \text{if } l \neq l', t = t' \\ \varepsilon_l^t + \tau_R, & \text{if } l = L \\ \varepsilon_l^t + \tau_l^{tt'}, & \text{otherwise} \end{cases}$$

여기서, $\tau_l^{tt'}$ 은 Tier l 에서 Tier l' 으로 계층 l 의 중간 데이터를 전송하기 위한 통신 지연을 나타낸다. ε_l^t 는 계층 l 을 Tier l 에서 처리하는데 걸리는 지연시간을 나타내며, τ_R 은 DNN의 결과를 모바일 기기(MD)에게 전달하는데 걸리는 지연시간을 나타낸다.

한편, 그래프의 최단 거리 알고리즘의 복잡도는 정점과 엣지의 수에 의해 결정된다. 알고리즘의 복잡성을 줄이기 위해(즉, 정점과 엣지의 수를 줄이기 위해) 각 티어(Tier)가 최소한 하나의 계층을 실행한다는 가정한다. 각 티어(Tier)가 하나 이상의 계층을 실행하더라도 유사한 중간 데이터 크기를 갖는 분할점 후보가 많기 때문에 이러한 가정이 성능(즉, DNN 모델의 추론 지연)에 거의 영향을 미치지 않는다.

도 4a 및 도 4b는 가정을 통한 간소화 전후 그래프의 예시도이다.

그래프 구성시에 엣지 비용(edge cost)은 기본적으로 계층 l 의 실행 단위가 각 컴퓨팅 노드의 가용 컴퓨팅 성능에 비례하여 할당된다는 가정하에 설정되므로 계층 내 분할 지점(즉, 실행 단위)은 최단 경로를 찾았을 때 결정될 수 있다.

알고리즘의 복잡성을 줄이기 위해, 각 티어가 적어도 하나의 계층을 실행한다는 가정을 도입하여 정점과 엣지의 수를 감소시키도록 도 4a 그래프를 가지치기하여 도 4b의 간소화된 그래프를 도출할 수 있다.

즉, 경우의 수 중 일반적으로 발생하지 않는 경우의 수를 간소화하여 티어별로 참여할 디바이스를 미리 정해서 참여하는 노드만 계산을 진행함으로써 지연 시간을 줄일 수 있다.

이 같은 과정을 거쳐서 정점의 수는 거의 절반으로, 엣지의 수는 약 3분의 2로 감소한다. 각 티어 별로 적어도 하나의 계층을 실행한다는 가정이 있지만, 충분한 전송 속도가 보장된다면(즉, 계산 노드 중 하나가 중간 데이터를 짧은 시간 내에 전송할 수 있는 경우) 이 가정이 성능(즉, DNN 모델의 추론 지연)에 거의 영향을 미치지 않는다.

각 티어(Tier)가 최소한 하나의 계층을 실행한다고 가정했을 경우 정점 집합 V' 와 엣지 집합 E' 은 다음 수학식 4 및 수학식 5와 같이 재정의된다.

수학식 4

$$V' = \{v_l^t \mid 0 \leq t \leq L \leq L\}$$

수학식 5

$$E' = \{(v_l^t, v_{l'}^{t'}) \mid t \leq t' \leq t+1 \leq T, l' = l+1 \leq L\} \cup \{(v_l^t, D) \mid l = L\}$$

그리고 수집된 컴퓨팅 노드 정보에 기반하여 계층 간 분할 및 계층 내 분할된 DNN 모델의 모든 계층에 대한 실행 단위를 컴퓨팅 노드에 분산시킨다(S230).

- [0092] 이후에, 모든 계층에 대한 모든 실행 단위를 컴퓨팅 노드에 할당한다. 분산시키는 단계는 수집된 컴퓨팅 노드 정보에 기반하여 컴퓨팅 노드의 계산 능력 또는 계층 간 링크 전송 속도를 반영하여 컴퓨팅 노드에 분산시킨다.
- [0093] DNN 모델은 하나의 디바이스에서 계산하기에는 크기 때문에 여러 개의 디바이스에서 참여를 시킨다.
- [0094] 본 발명의 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법은 기존의 레이어 별로 디바이스들이 나누어 계산하는 방식에서 추론 가속화를 위해 레이어를 동일한 계층 내에서 더 세분화하여 더 많은 디바이스들을 참여시킨다.
- [0095] 그리고 분산시키는 단계에서 실행 단위를 분산 받은 컴퓨팅 노드가 할당받은 실행 단위를 수행하여 추론한다(S240).
- [0096] 추론하는 단계는 컴퓨팅 노드가 실행 단위를 추론 완료 후 결과를 공유할 수 있다.
- [0097] 예를 들어, 모든 계층에 대한 모든 실행 단위를 컴퓨팅 노드에 할당한 후, 첫 번째 DNN 계층 l 에 대한 실행 단위를 수행하는 Tier t 의 컴퓨팅 노드가 추론을 시작한다. Tier t 의 모든 컴퓨팅 노드는 자신에게 할당된 실행 단위를 완료하면 결과를 공유한다.
- [0098] 그리고 다음 계층 $l+1$ 에 대한 실행 단위도 동일한 계층 티어(Tier) t 에 할당되면 컴퓨팅 노드는 다음 계층 $l+1$ 에 대한 실행 단위를 수행한다. 만약, 다음 계층 $l+1$ 에 대한 실행 단위가 다음 티어 $t+1$ 에 할당되었을 때는 다음 티어 $t+1$ 에 대한 링크 용량이 가장 큰 컴퓨팅 노드가 계층 l 의 중간 데이터를 다음 티어 $t+1$ 로 전송한다. 마지막 계층 L 에 대한 실행 단위가 특정 Tier t' 에서 완료될 때까지 이러한 절차가 반복된다. 마지막 계층 L 을 수행한 Tier t' 컴퓨팅 노드 중 하나가 결과를 MD로 전송한다.
- [0099] 또한, 추론하는 단계는 합성곱 레이어(convolutional layer)의 경우 분할된 이웃 특성맵(feature map)이 각 특성맵(feature map)의 가장자리에 채워지도록 패딩을 수행한다.
- [0100] 도 5는 계층 내 분할시 특성맵의 변형과정을 설명하기 위한 예시도이다.
- [0101] Fully-connected layer의 경우 계층 l 의 입력(즉, 특성맵)을 조작 없이 단순히 여러 실행 단위로 분할하더라도 계층 내 분할을 고려하지 않은 경우와 동일한 결과를 얻을 수 있다. 한편, convolutional layer의 경우, 계층 내 분할이 고려되지 않았을 때와 같은 결과를 얻기 위해 각 계산 노드는 특성맵(feature map)의 조작이 필요하다. 도 5에서(a)와 (b)는 특성맵(feature map)이 분할되지 않은 경우와 특성맵(feature map)이 두 개의 특성맵(feature map) 즉, 실행 단위로 나누어졌으나 추가 조작을 수행하지 않은 경우를 보여준다.
- [0102] 도 5에서와 같이 특성맵(feature map)의 분할로 인해 동일한 영역에 컨볼루션 필터가 중첩되지 않는 경우 결과 값이 서로 다른 상황이 발생한다. 이러한 상황을 방지하기 위해, 도 5의(c)와 같이 분할된 이웃 특성맵(feature map)의 각 특성맵(feature map)의 가장자리에 채워져야 한다. 여기서, 특성맵의 빗금친 셀은 제로 패딩을 나타낸다.
- [0103] 도 6은 컴퓨팅 노드 수 증가에 따른 추론시간 변화를 도시한 그래프이다.
- [0104] 본 발명을 통해 모바일 네트워크에 분산된 컴퓨팅 노드의 자원을 효과적으로 활용하여 추론 지연을 최소화할 수 있다.
- [0105] 도 6에 기반하여 본 발명의 성능 평가를 위해 시뮬레이션을 진행하고 기존의 방법들과 비교할 수 있다.
- [0106] 도 6에서 "본 발명"은 본 발명의 일 실시예에 따른 계층적 모바일 네트워크에서 기계학습 방법이 적용된 경우의 결과를 나타낸 것이고, "계층 간 분할"은 Tier에서 가장 높은 컴퓨팅 파워를 갖는 컴퓨팅 노드를 기반으로 계층 간 분할만을 수행하는 기존방법에 의한 결과를 나타낸 것이다.
- [0107] 도 6은 티어(Tier)의 컴퓨팅 노드 수가 증가함에 따른 추론시간 변화를 나타낸 것으로, 도 6에서 볼 수 있듯이 본 발명은 모바일 네트워크에 분산된 컴퓨팅 노드의 자원을 효과적으로 활용하기 때문에 추론시간이 기존 방법보다 짧아지는 것을 확인할 수 있다. 특히, 컴퓨팅 노드 수가 증가함에 따라 더 많은 성능 차이를 보여주는 것을 확인할 수 있는데, 이는 컴퓨팅 노드 수에 비례하여 전체 분산 컴퓨팅 자원이 증가하기 때문이다.
- [0109] 진술한 방법은 애플리케이션으로 구현되거나 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어

의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다.

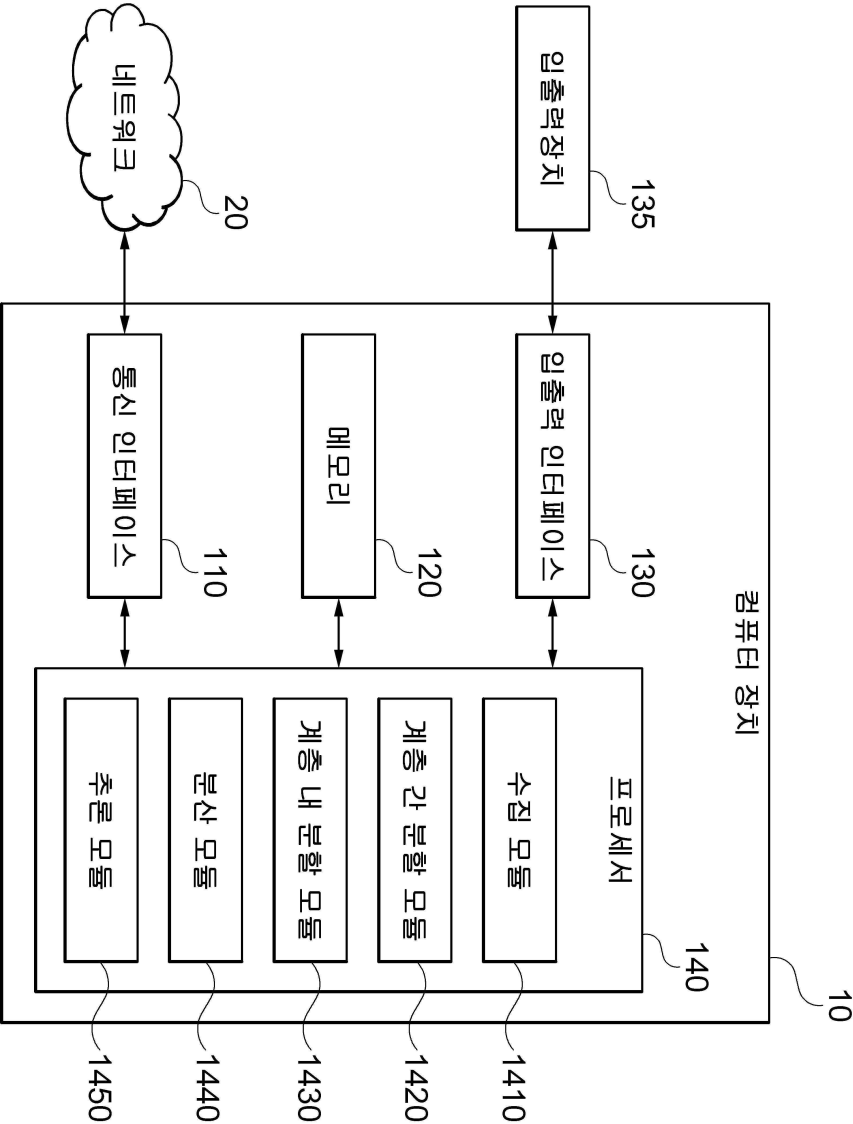
- [0110] 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.
- [0111] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD 와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.
- [0112] 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0113] 이상에서는 실시예들을 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

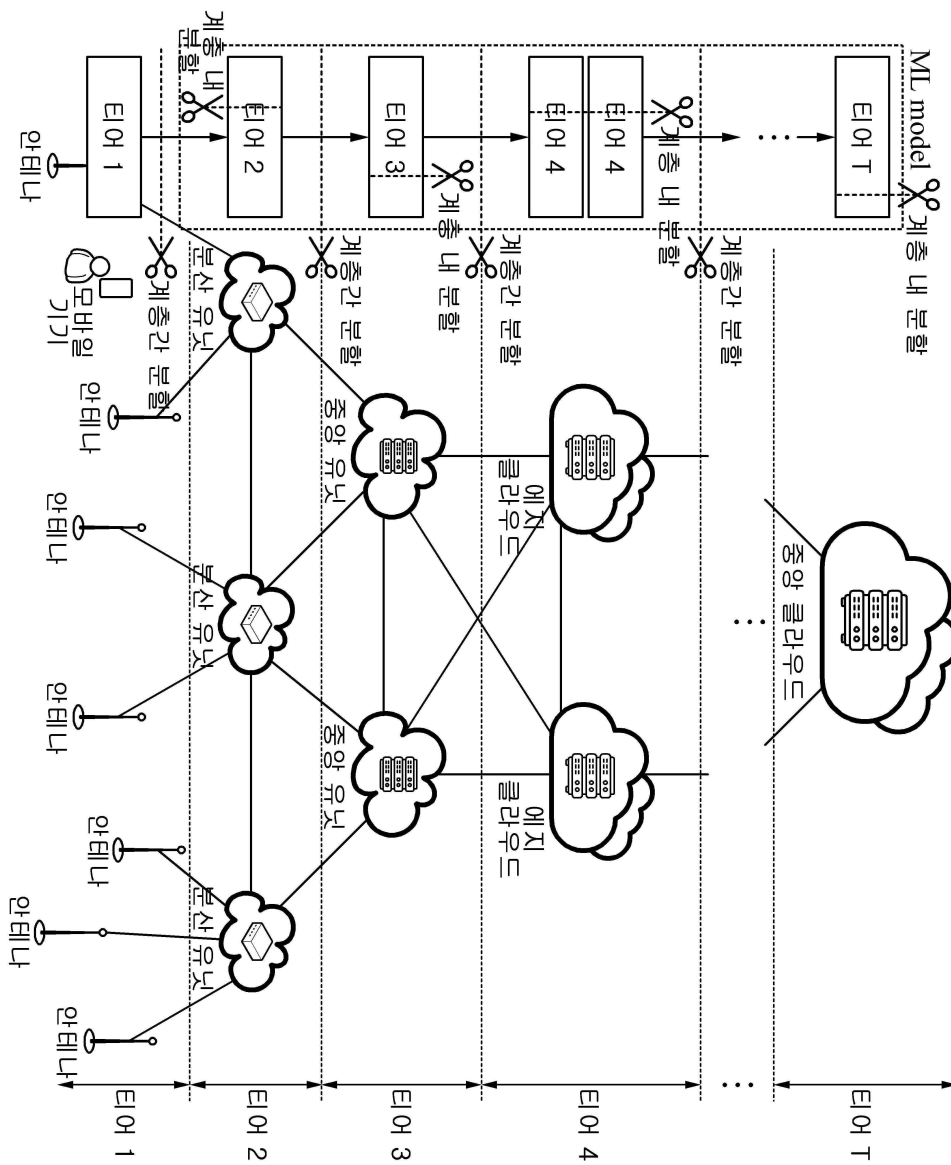
- [0115] 10 : 컴퓨터 장치
- 20 : 네트워크
- 110 : 통신 인터페이스
- 120 : 메모리
- 130 : 입출력 인터페이스
- 140 : 프로세서
- 1410 : 수집 모듈
- 1420 : 계층 간 분할 모듈
- 1430 : 계층 내 분할 모듈
- 1440 : 분산 모듈
- 1450 : 추론 모듈

도면

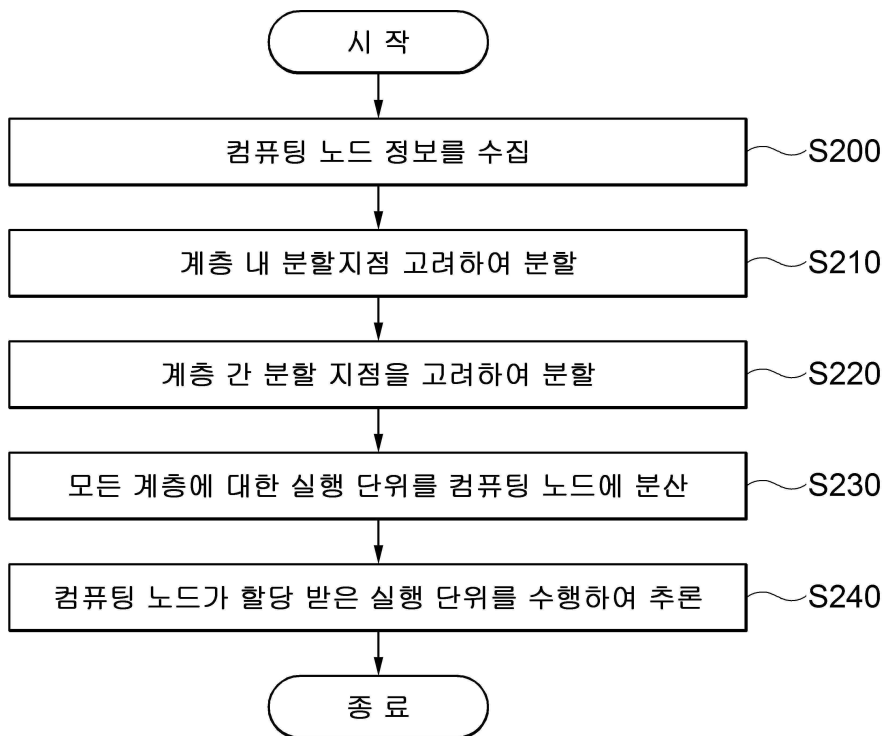
도면1



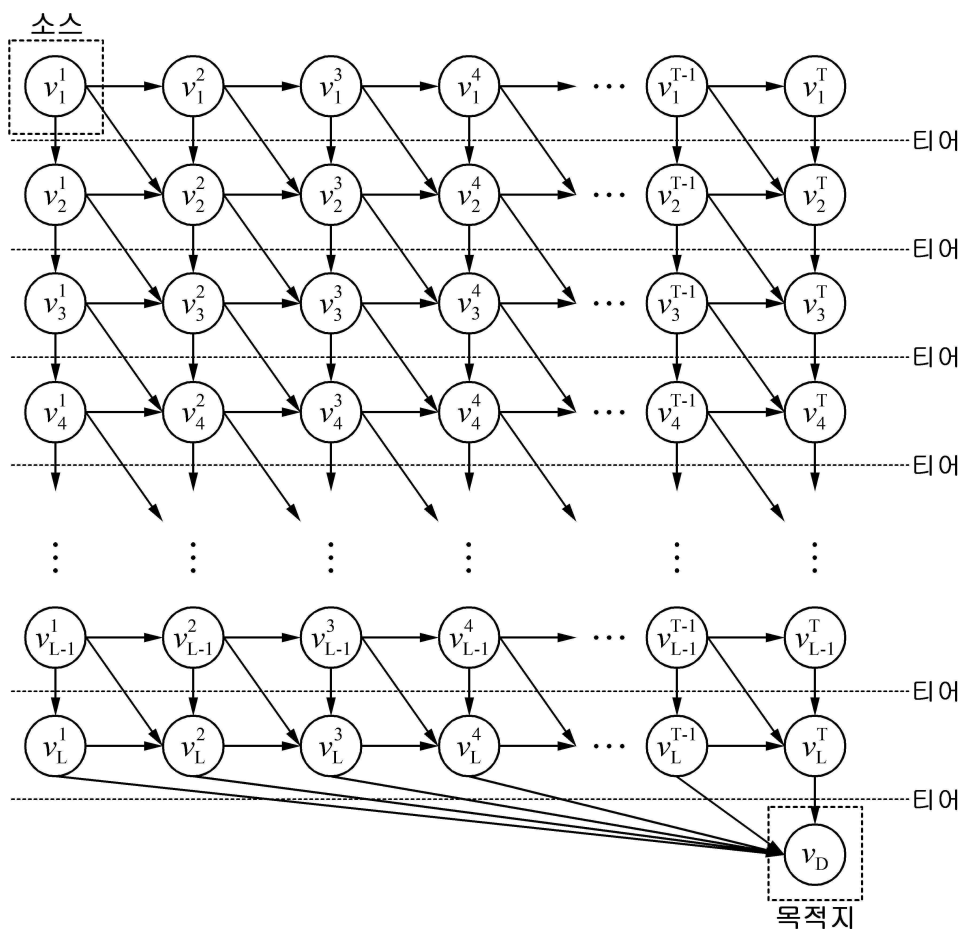
도면2



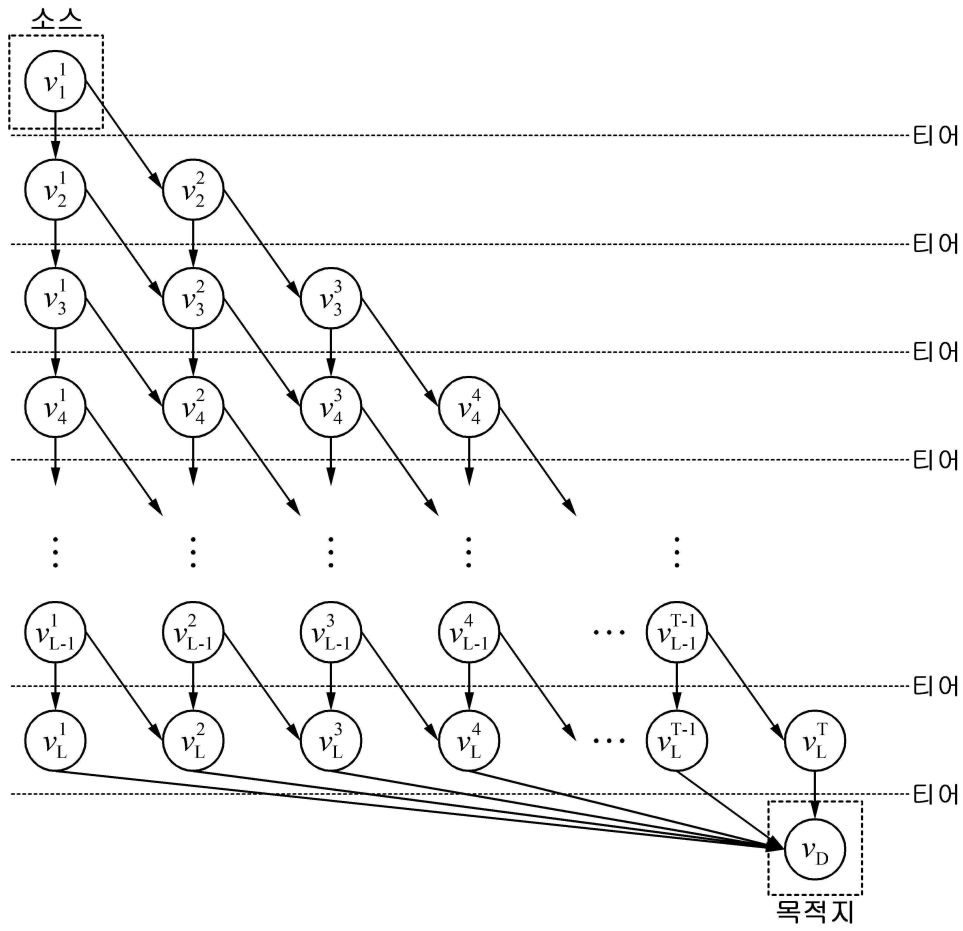
도면3



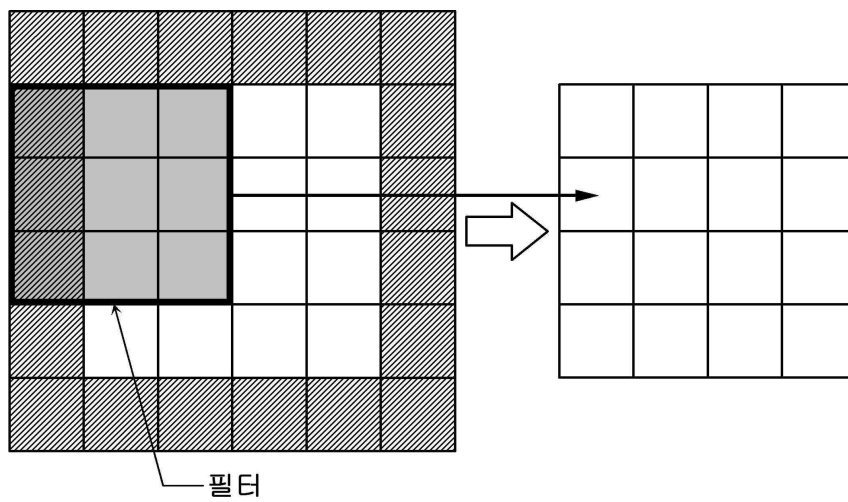
도면4a



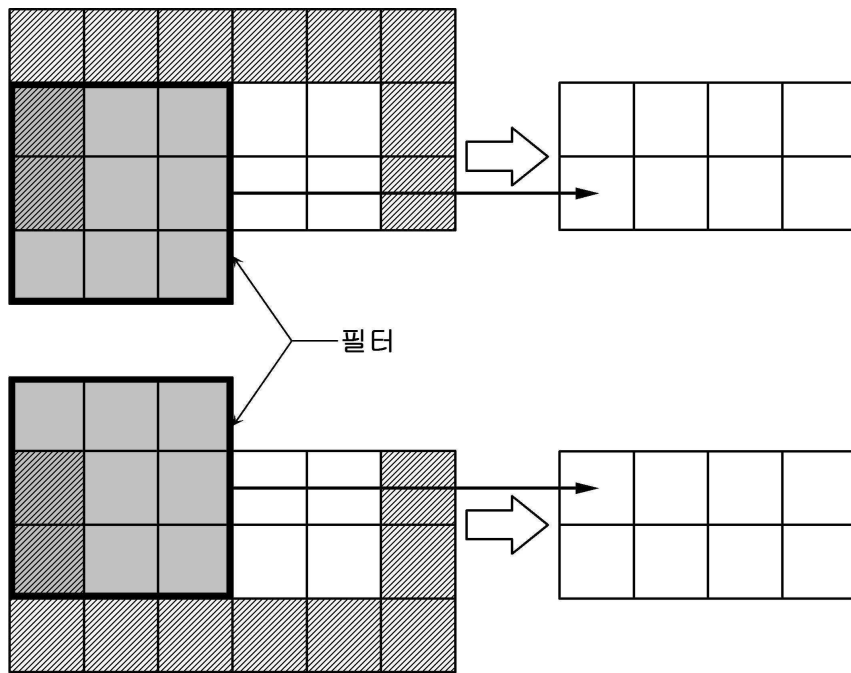
도면4b



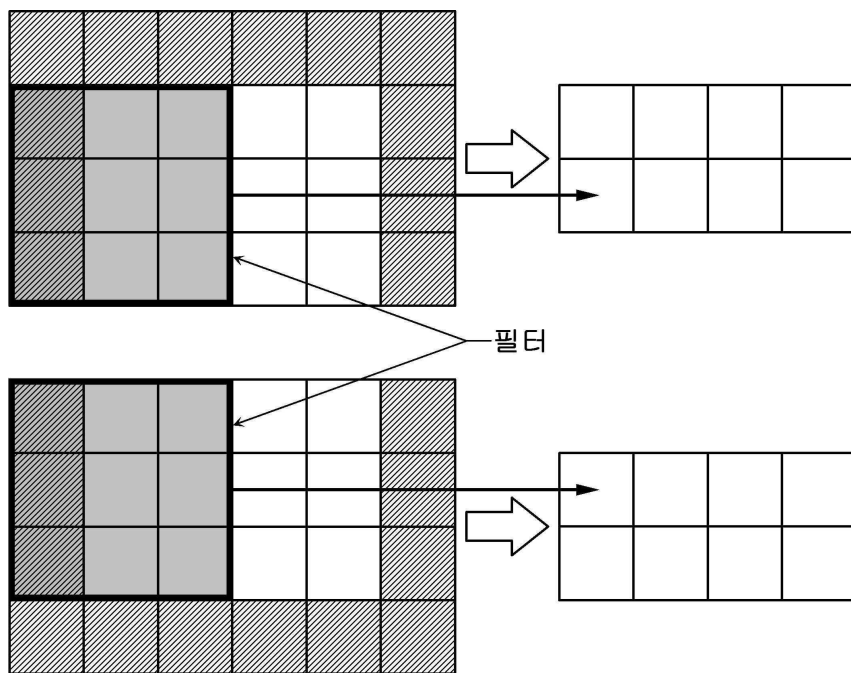
도면5a



도면5b



도면5c



도면6

