



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0142637
(43) 공개일자 2024년10월02일

(51) 국제특허분류(Int. Cl.)
G06N 3/098 (2023.01) G06N 3/063 (2023.01)
(52) CPC특허분류
G06N 3/098 (2023.01)
G06N 3/063 (2013.01)
(21) 출원번호 10-2023-0036486
(22) 출원일자 2023년03월21일
심사청구일자 2023년03월21일

(71) 출원인
한국항공대학교산학협력단
경기도 고양시 덕양구 항공대학로 76 (화전동, 한국항공대학교)
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
이재환
경기도 고양시 일산동구 일산로 205, 208동 101호
정하민
서울특별시 서대문구 연희로2안길 10-3, 605호 (창천동)
(74) 대리인
안병규

전체 청구항 수 : 총 15 항

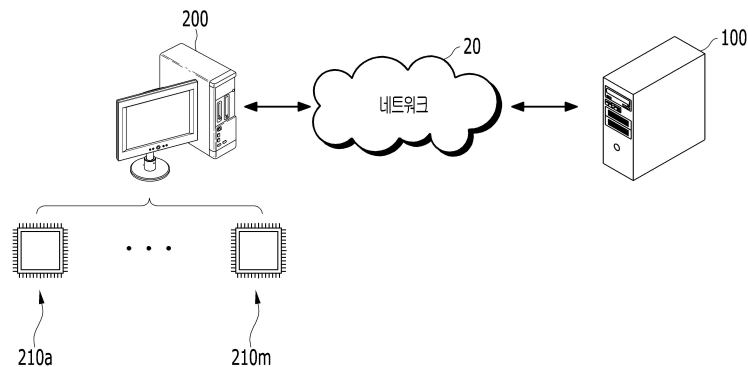
(54) 발명의 명칭 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법

(57) 요약

딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법이 개시되며, 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법은, (a) 딥러닝 모델을 이루는 레이어 정보 및 상기 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기에 대한 가속기 정보를 획득하는 단계 및 (b) 상기 레이어 정보 및 상기 가속기 정보에 기초하여 상기 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 상기 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정하는 단계를 포함할 수 있다.

대표도 - 도1

10



이 발명을 지원한 국가연구개발사업

과제고유번호	1711159975
과제번호	IITP-2023-2018-0-01423
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	지능형 비행로봇 융합기술 연구
기 여 율	1/1
과제수행기관명	세종대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법에 있어서,

(a) 딥러닝 모델을 이루는 레이어 정보 및 상기 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기에 대한 가속기 정보를 획득하는 단계; 및

(b) 상기 레이어 정보 및 상기 가속기 정보에 기초하여 상기 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 상기 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정하는 단계,

를 포함하고,

상기 (b) 단계는,

상기 복수의 가속기 각각에 대하여 할당되는 레이어 개수를 조정하기 위한 탐색 보폭에 대응하는 깊이 값을 조정하면서 상기 비용 함수를 연산하는 것을 특징으로 하는, 탐색 방법.

청구항 2

제1항에 있어서,

상기 비용 함수는 상기 모델 분할 시나리오 각각에 대응하는 네트워크 트래픽 및 메모리 사용량을 기초로 하여 연산되는 것인, 탐색 방법.

청구항 3

제1항에 있어서,

상기 (b) 단계는,

(b1) 상기 복수의 가속기 각각에 대응하는 플래그 정보에 기초하여 상기 복수의 가속기 중에서 상기 레이어 개수의 할당 값을 조정할 대상 가속기를 특정하는 단계;

(b2) 상기 대상 가속기에 대하여 기 적용된 상기 할당 값에 대한 상기 비용 함수의 제1연산 결과와 상기 깊이 값에 기초하여 갱신된 상기 할당 값에 대한 상기 비용 함수의 제2연산 결과를 비교하는 단계; 및

(b3) 상기 비교 결과에 기초하여 상기 할당 값 또는 상기 플래그 정보를 조정하는 단계,

를 포함하는 것인, 탐색 방법.

청구항 4

제3항에 있어서,

상기 (b) 단계는,

상기 플래그 정보에 따라 특정 탐색 회차에서 적용된 상기 깊이 값을 기초로 한 상기 복수의 가속기 각각에 대한 상기 할당 값 조정 및 상기 비용 함수 연산이 완결된 것으로 판단되면,

(b4) 상기 깊이 값을 미리 설정된 기준 크기만큼 감소시키는 단계,

를 더 포함하는 것인, 탐색 방법.

청구항 5

제4항에 있어서,

상기 (b) 단계는,

상기 감소된 깊이 값을 기초로 하여 상기 (b1) 단계 내지 상기 (b3) 단계를 재차 수행하는 것을 특징으로 하는, 탐색 방법.

청구항 6

제4항에 있어서,

상기 (b) 단계는,

상기 깊이 값이 미리 설정된 하한 값에 도달하면 종료되는 것이고,

상기 (b) 단계에 의해, 상기 복수의 가속기 각각에 대한 상기 할당 값에 대한 정보를 포함하는 상기 최적 분할 시나리오가 결정되는 것인, 탐색 방법.

청구항 7

제1항에 있어서,

상기 최적 분할 시나리오는 하기 식 1에 따른 개수만큼의 상기 복수의 모델 분할 시나리오 중에서 선정되는 것이고,

[식 1]

$${}_nC_2 \sum_{depth=1}^a \left\lfloor \frac{r}{depth} \right\rfloor$$

여기서, n은 상기 복수의 가속기의 개수이고, r은 상기 레이어 정보에 포함된 상기 딥러닝 모델의 전체 레이어의 개수이고, a는 상기 깊이 값의 최대 값인 것을 특징으로 하는, 탐색 방법.

청구항 8

제1항에 있어서,

상기 복수의 가속기 각각은 GPU(Graphics Processing Unit)인 것인, 탐색 방법.

청구항 9

제1항에 있어서,

(c) 상기 최적 분할 시나리오에 기초하여 상기 딥러닝 모델의 학습을 수행하는 단계,

를 더 포함하는 것인, 탐색 방법.

청구항 10

딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치에 있어서,

딥러닝 모델을 이루는 레이어 정보 및 상기 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기에 대한 가속기 정보를 획득하는 수집부; 및

상기 레이어 정보 및 상기 가속기 정보에 기초하여 상기 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 상기 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정하는 탐색 수행부,

를 포함하고,

상기 탐색 수행부는,

상기 복수의 가속기 각각에 대하여 할당되는 레이어 개수를 조정하기 위한 탐색 보폭에 대응하는 깊이 값을 조정하면서 상기 비용 함수를 연산하는 것을 특징으로 하는, 탐색 장치.

청구항 11

제10항에 있어서,

상기 비용 함수는 상기 모델 분할 시나리오 각각에 대응하는 네트워크 트래픽 및 메모리 사용량을 기초로 하여 연산되는 것인, 탐색 장치.

청구항 12

제10항에 있어서,

상기 탐색 수행부는,

상기 복수의 가속기 각각에 대응하는 플래그 정보에 기초하여 상기 복수의 가속기 중에서 상기 레이어 개수의 할당 값을 조정할 대상 가속기를 특정하는 가속기 식별부;

상기 대상 가속기에 대하여 기 적용된 상기 할당 값에 대한 상기 비용 함수의 제1연산 결과와 상기 깊이 값에 기초하여 갱신된 상기 할당 값에 대한 상기 비용 함수의 제2연산 결과를 비교하는 연산 수행부; 및

상기 비교 결과에 기초하여 상기 할당 값 또는 상기 플래그 정보를 조정하는 탐색 설정부,

를 포함하는 것인, 탐색 장치.

청구항 13

제12항에 있어서,

상기 탐색 수행부는,

상기 플래그 정보에 따라 특정 탐색 회차에서 적용된 상기 깊이 값을 기초로 한 상기 복수의 가속기 각각에 대한 상기 할당 값 조정 및 상기 비용 함수 연산이 완결된 것으로 판단되면, 상기 깊이 값을 미리 설정된 기준 크기만큼 감소시키는 깊이 조정부,

를 더 포함하는 것인, 탐색 장치.

청구항 14

제10항에 있어서,

상기 최적 분할 시나리오는 하기 식 1에 따른 개수만큼의 상기 복수의 모델 분할 시나리오 중에서 선정되는 것이고,

[식 1]

$${}_nC_2 \sum_{depth=1}^a \left\lfloor \frac{r}{depth} \right\rfloor$$

여기서, n은 상기 복수의 가속기의 개수이고, r은 상기 레이어 정보에 포함된 상기 딥러닝 모델의 전체 레이어의 개수이고, a는 상기 깊이 값의 최대 값인 것을 특징으로 하는, 탐색 장치.

청구항 15

제10항에 있어서,

상기 최적 분할 시나리오에 기초하여 상기 딥러닝 모델의 학습을 수행하는 학습부,

를 더 포함하는 것인, 탐색 장치.

발명의 설명

기술 분야

본원은 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법에 관한 것이다.

[0001]

배경 기술

- [0002] 최근 딥러닝 모델들은 응용의 인식 성능을 높이기 위해 모델의 계층이 깊어지고, 고려해야 할 특징이 많아지는 대규모 네트워크로 구축되고 있다. 대규모 딥러닝 네트워크를 효율적이고 빠르게 트레이닝 하기 위해 다중 GPGPU(General-Purpose Computing on Graphics Processing Units)를 활용한 딥러닝 분산 처리 기술을 딥러닝 프레임워크 개발에 적용하고 있는 추세이다.
- [0003] 현재까지 연구된 딥러닝 분산 처리 기술은 크게 트레이닝 데이터를 다수의 컴퓨터에서 분산 처리하는 데이터 병렬 처리(Data Parallelism)와 딥러닝 네트워크를 다수의 컴퓨터에서 분산 처리하는 모델 병렬 처리(Model Parallelism)로 구분될 수 있다.
- [0004] 이중 딥러닝 모델 병렬 처리는 딥러닝 네트워크를 하나의 컴퓨팅 디바이스에서 처리하지 못할 경우, 네트워크를 분할하여 다수의 컴퓨터에서 처리하는 기법을 의미한다.
- [0005] 이와 관련하여, 모델 병렬화 시 모델의 GPU 메모리 사용량의 경우 각 GPU에 최대한 균등하게 분배한다면, 배치 사이즈(batch size)를 늘릴 수 있는데, 이렇게 배치 사이즈가 커지게 되면 학습 시의 노이즈를 줄일 수 있고, 파라미터를 갱신(업데이트)하는 빈도가 줄어들어 에포크(epoch) 당 수행시간을 줄일 수 있다는 장점이 있고, 네트워크 트래픽을 최소화 하게 되면, 전체 학습시간을 줄일 수 있게 된다.
- [0006] 또한, 모델 병렬화 시 분할한 레이어들을 가속기(Accelerator)에 적절히 분배하는 것이 중요한데, 이를 위하여 각각의 모델 분할 시나리오에 대한 성능 평가를 수행하며 최적화를 적용할 수 있으나, 각 모델 분할 시나리오에 대한 전수 조사에는 다소 큰 시간이 소요되므로, 이를 효율적으로 줄일 수 있는 알고리즘의 개발이 필요하다.
- [0007] 본원의 배경이 되는 기술은 한국공개특허공보 제10-2022-0056621호에 개시되어 있다.

발명의 내용

해결하려는 과제

- [0008] 본원은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 모델 병렬화 적용 시 파티션을 적절히 분배하기 위하여 비용 함수를 이용하여 최적의 모델 분할 시나리오를 탐색할 수 있는 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법을 제공하려는 것을 목적으로 한다.
- [0009] 본원은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 가속기 간의 상관성을 고려한 정밀한 탐색 및 깊이 기반의 탐색을 통해 탐색 보폭을 가변적으로 적용하면서 최적 모델 분할 시나리오를 탐색하기 위한 전체 탐색 횟수를 줄일 수 있는 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법을 제공하려는 것을 목적으로 한다.
- [0010] 다만, 본원의 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

- [0011] 상기한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법은, (a) 딥러닝 모델을 이루는 레이어 정보 및 상기 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기에 대한 가속기 정보를 획득하는 단계 및 (b) 상기 레이어 정보 및 상기 가속기 정보에 기초하여 상기 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 상기 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정하는 단계를 포함할 수 있다.
- [0012] 또한, 상기 (b) 단계는, 상기 복수의 가속기 각각에 대하여 할당되는 레이어 개수를 조정하기 위한 탐색 보폭에 대응하는 깊이 값을 조정하면서 상기 비용 함수를 연산할 수 있다.
- [0013] 또한, 상기 비용 함수는 상기 모델 분할 시나리오 각각에 대응하는 네트워크 트래픽 및 메모리 사용량을 기초로 하여 연산될 수 있다.
- [0014] 또한, 상기 (b) 단계는, (b1) 상기 복수의 가속기 각각에 대응하는 플래그 정보에 기초하여 상기 복수의 가속기 중에서 상기 레이어 개수의 할당 값을 조정할 대상 가속기를 특정하는 단계, (b2) 상기 대상 가속기에 대하여 적용된 상기 할당 값에 대한 상기 비용 함수의 제1연산 결과와 상기 깊이 값에 기초하여 갱신된 상기 할당

값에 대한 상기 비용 함수의 제2연산 결과를 비교하는 단계 및 (b3) 상기 비교 결과에 기초하여 상기 할당 값 또는 상기 플래그 정보를 조정하는 단계를 포함할 수 있다.

[0015] 또한, 상기 (b) 단계는, 상기 플래그 정보에 따라 특정 탐색 회차에서 적용된 상기 깊이 값을 기초로 한 상기 복수의 가속기 각각에 대한 상기 할당 값 조정 및 상기 비용 함수 연산이 완결된 것으로 판단되면, (b4) 상기 깊이 값을 미리 설정된 기준 크기만큼 감소시키는 단계를 포함할 수 있다.

[0016] 또한, 상기 (b) 단계는, 상기 감소된 깊이 값을 기초로 하여 상기 (b1) 단계 내지 상기 (b3) 단계를 재차 수행할 수 있다.

[0017] 또한, 상기 (b) 단계는, 상기 깊이 값이 미리 설정된 하한 값에 도달하면 종료될 수 있다.

[0018] 또한, 상기 (b) 단계에 의해, 상기 복수의 가속기 각각에 대한 상기 할당 값에 대한 정보를 포함하는 상기 최적 분할 시나리오가 결정될 수 있다.

[0019] 또한, 상기 최적 분할 시나리오는 하기 식 1에 따른 개수만큼의 상기 복수의 모델 분할 시나리오 중에서 선정될 수 있다.

[0020] 또한, 상기 복수의 가속기 각각은 GPU(Graphics Processing Unit)일 수 있다.

[0021] 또한, 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법은, (c) 상기 최적 분할 시나리오에 기초하여 상기 딥러닝 모델의 학습을 수행하는 단계를 포함할 수 있다.

[0022] 한편, 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치는, 딥러닝 모델을 이루는 레이어 정보 및 상기 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기에 대한 가속기 정보를 획득하는 수집부 및 상기 레이어 정보 및 상기 가속기 정보에 기초하여 상기 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 상기 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정하는 탐색 수행부를 포함할 수 있다.

[0023] 또한, 상기 탐색 수행부는, 상기 복수의 가속기 각각에 대하여 할당되는 레이어 개수를 조정하기 위한 탐색 보폭에 대응하는 깊이 값을 조정하면서 상기 비용 함수를 연산할 수 있다.

[0024] 또한, 상기 탐색 수행부는, 상기 복수의 가속기 각각에 대응하는 플래그 정보에 기초하여 상기 복수의 가속기 중에서 상기 레이어 개수의 할당 값을 조정할 대상 가속기를 특정하는 가속기 식별부, 상기 대상 가속기에 대하여 기 적용된 상기 할당 값에 대한 상기 비용 함수의 제1연산 결과와 상기 깊이 값에 기초하여 갱신된 상기 할당 값에 대한 상기 비용 함수의 제2연산 결과를 비교하는 연산 수행부 및 상기 비교 결과에 기초하여 상기 할당 값 또는 상기 플래그 정보를 조정하는 탐색 설정부를 포함할 수 있다.

[0025] 또한, 상기 탐색 수행부는, 상기 플래그 정보에 따라 특정 탐색 회차에서 적용된 상기 깊이 값을 기초로 한 상기 복수의 가속기 각각에 대한 상기 할당 값 조정 및 상기 비용 함수 연산이 완결된 것으로 판단되면, 상기 깊이 값을 미리 설정된 기준 크기만큼 감소시키는 깊이 조정부를 포함할 수 있다.

[0026] 또한, 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치는 상기 최적 분할 시나리오에 기초하여 상기 딥러닝 모델의 학습을 수행하는 학습부를 포함할 수 있다.

[0027] 상술한 과제 해결 수단은 단지 예시적인 것으로서, 본원을 제한하려는 의도로 해석되지 않아야 한다. 상술한 예시적인 실시예 외에도, 도면 및 발명의 상세한 설명에 추가적인 실시예가 존재할 수 있다.

발명의 효과

[0028] 진술한 본원의 과제 해결 수단에 의하면, 모델 병렬화 적용 시 파티션을 적절히 분배하기 위하여 비용 함수를 이용하여 최적의 모델 분할 시나리오를 탐색할 수 있는 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법을 제공할 수 있다.

[0029] 진술한 본원의 과제 해결 수단에 의하면, 가속기 간의 상관성을 고려한 정밀한 탐색 및 깊이 기반의 탐색을 통해 탐색 보폭을 가변적으로 적용하면서 최적 모델 분할 시나리오를 탐색하기 위한 전체 탐색 횟수를 줄일 수 있는 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법을 제공할 수 있다.

[0030] 진술한 본원의 과제 해결 수단에 의하면, 다수의 가속기를 사용하는 환경에서 모델 병렬화를 수행할 때 파티션들을 적절히 분배하고, 모든 경우의 수에 대해 탐색하는 것에 비해 짧은 시간 내에 최적의 파티션 분할 시나리오

오를 탐색해 내어 효율적인 모델 병렬화 작업을 수행할 수 있다.

[0031] 전문한 본원의 과제 해결 수단에 의하면, 깊이 기반의 탐색을 통해 필요한 연산의 경우의 수를 줄여, 분할 시나리오 탐색을 위한 테스트 횟수를 줄일 수 있다.

[0032] 다만, 본원에서 얻을 수 있는 효과는 상기된 바와 같은 효과들로 한정되지 않으며, 또 다른 효과들이 존재할 수 있다.

도면의 간단한 설명

[0033] 도 1은 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치를 포함하는 딥러닝 모델 학습 시스템의 개략적인 구성도이다.

도 2는 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치의 최적 분할 시나리오 탐색 프로세스를 설명하기 위한 개념도이다.

도 3은 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치의 개략적인 구성도이다.

도 4는 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치의 탐색 수행부의 세부 구성도이다.

도 5는 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법에 대한 동작 흐름도이다.

도 6은 본원의 일 실시예에 따른 최적 분할 시나리오 선정 프로세스에 대한 세부 동작 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0034] 아래에서는 첨부한 도면을 참조하여 본원이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본원의 실시예를 상세히 설명한다. 그러나 본원은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본원을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0035] 본원 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐만 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결" 또는 "간접적으로 연결"되어 있는 경우도 포함한다.

[0036] 본원 명세서 전체에서, 어떤 부재가 다른 부재 "상에", "상부에", "상단에", "하에", "하부에", "하단에" 위치하고 있다고 할 때, 이는 어떤 부재가 다른 부재에 접해 있는 경우뿐만 아니라 두 부재 사이에 또 다른 부재가 존재하는 경우도 포함한다.

[0037] 본원 명세서 전체에서, 어떤 부분이 어떤 구성 요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성 요소를 제외하는 것이 아니라 다른 구성 요소를 더 포함할 수 있는 것을 의미한다.

[0038] 본원은 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치 및 방법에 관한 것이다.

[0039] 도 1은 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치를 포함하는 딥러닝 모델 학습 시스템의 개략적인 구성도이다.

[0040] 도 1을 참조하면, 본원의 일 실시예에 따른 딥러닝 모델 학습 시스템(10)은 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치(100)(이하, '파티션 탐색 장치(100)'라 한다.) 및 디바이스(200)를 포함할 수 있다.

[0041] 파티션 탐색 장치(100) 및 디바이스(200) 상호간은 네트워크(20)를 통해 통신할 수 있다. 네트워크(20)는 단말들 및 서버들과 같은 각각의 노드 상호간에 정보 교환이 가능한 연결 구조를 의미하는 것으로, 이러한 네트워크(20)의 일 예에는, 3GPP(3rd Generation Partnership Project) 네트워크, LTE(Long Term Evolution) 네트워크, 5G 네트워크, WIMAX(World Interoperability for Microwave Access) 네트워크, 인터넷(Internet), LAN(Local Area Network), Wireless LAN(Wireless Local Area Network), WAN(Wide Area Network), PAN(Personal Area Network), wifi 네트워크, 블루투스(Bluetooth) 네트워크, 위성 방송 네트워크, 아날로그 방송 네트워크, DMB(Digital Multimedia Broadcasting) 네트워크 등이 포함되나 이에 한정되지는 않는다.

[0042] 디바이스(200)는 예를 들면, 스마트폰(Smartphone), 스마트패드(SmartPad), 태블릿 PC 등과 PCS(Personal

Communication System), GSM(Global System for Mobile communication), PDC(Personal Digital Cellular), PHS(Personal Handyphone System), PDA(Personal Digital Assistant), IMT(International Mobile Telecommunication)-2000, CDMA(Code Division Multiple Access)-2000, W-CDMA(W-Code Division Multiple Access), Wibro(Wireless Broadband Internet) 단말기 같은 모든 종류의 무선 통신 장치일 수 있다.

[0043] 또한, 도 1을 참조하면, 딥러닝 모델의 학습을 위한 디바이스(200)에는 적어도 하나 이상의 가속기(210)가 탑재될 수 있다. 예를 들어, m 개의 프로세싱 유닛을 구비하는 디바이스(200)의 경우, 첫 번째 가속기(210a) 내지 m 번째 가속기(210m)를 구비하는 것일 수 있다.

[0044] 이와 관련하여, 본원에서 개시하는 디바이스(200)에 탑재되는 가속기(210)는 GPU(Graphics Processing Unit)일 수 있다. 딥러닝(Deep Learning)은 일반적으로 행렬 곱과 같은 벡터 연산으로 이루어진 복수의 레이어(Layer)로 이루어져 있다. 여기서, 복수의 레이어 각각은 사용하는 연산, 파라미터 등에 따라 연산량 및 연산시 필요한 필요 메모리 크기가 제각각이다. 따라서, 단순히 전체 레이어의 수(즉, 복수의 레이어의 개수)를 GPU의 수(즉, 멀티 GPU의 개수)만큼 나누어서 딥러닝 모델에 대한 작업을 멀티 GPU에 할당하는 경우, 특정 GPU에 부하를 주거나 또는 연산량이 적을 경우 해당 GPU의 연산 자원을 최대로 활용하기 어려운 문제가 있다. 그러므로, 멀티 GPU 기반의 효율적인 딥러닝 모델의 병렬화를 위해서는 딥러닝 모델에 대한 작업을 GPU 별로 적절하게 분배해야 할 필요가 있다.

[0045] 한편, 본원의 실시예에 관한 설명에서 파티션 탐색 장치(100)는 디바이스(200)와 연동하여 디바이스(200)에서 수행되는 딥러닝 모델의 학습을 위한 모델 병렬 처리(Model Parallelism)를 최적화 하도록 동작하는 디바이스(200)에 대한 별도의 장치일 수 있으나, 이에만 한정되는 것은 아니고, 디바이스(200)에 탑재(내장)되어 해당 디바이스(200)에서 수행될 딥러닝 모델의 학습을 위한 모델 병렬 처리(Model Parallelism)를 최적화 하도록 동작하는 것일 수 있다. 또한, 본원의 구현예에 따라서는 디바이스(200)는 특정 딥러닝 모델의 학습에 관여하는 적어도 하나 이상의 하드웨어 장치를 통합적으로 지칭하는 것일 수 있다.

[0046] 이하에서는 파티션 탐색 장치(100)의 구체적인 기능 및 동작에 대하여 상세히 설명하도록 한다.

[0047] 먼저 파티션 탐색 장치(100)는 딥러닝 모델을 이루는 레이어 정보 및 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기(210)에 대한 가속기 정보를 획득할 수 있다.

[0048] 구체적으로 본원의 실시예에 관한 설명에서 딥러닝 모델을 이루는 레이어 정보는 해당 딥러닝 모델의 전체 레이어 개수에 대한 정보, 각각의 레이어의 구조(아키텍처) 정보 등을 포함할 수 있으며, 가속기 정보는 모델 병렬화 기반의 딥러닝 모델 학습을 수행하기 위한 복수의 가속기(210)의 개수, 각각의 가속기(210)의 제원 정보 등을 포함할 수 있다.

[0049] 또한, 파티션 탐색 장치(100)는 획득한 레이어 정보 및 가속기 정보에 기초하여 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정할 수 있다.

[0050] 달리 말해, 파티션 탐색 장치(100)는 모델 병렬화 시 파티션을 어떻게 분할하는지에 따라 전체 학습 성능에 차이가 존재하는 점을 고려하여, 학습에 소요되는 각종 자원에 대한 파라미터를 포함하는 비용 함수(Cost Function)를 설정한 후, 비용 함수의 연산 결과가 가장 작게 도출되는 파티션 분할 시나리오(최적 분할 시나리오)를 탐색하도록 동작할 수 있다.

[0051] 이와 관련하여, 본원에서 개시하는 비용 함수는 모델 분할 시나리오 각각에 대응하는 네트워크 트래픽(Network Traffic) 및 메모리 사용량(예를 들면, GPU의 메모리 사용량 등)을 기초로 하여 연산되는 것일 수 있으나, 이에만 한정되는 것은 아니고, 딥러닝 모델의 학습에 소요되는 다양한 자원(리소스)과 연계된 파라미터를 기초로 하여 정의될 수 있다. 또한, 본원의 실시예에 관한 설명에서 모델 분할 시나리오 각각은 레이어 정보 및 가속기 정보에 기초하여 복수의 가속기(210) 각각에 대하여 모델 병렬화를 통해 할당되는 레이어의 개수가 서로 상이한 패턴으로 적용되는 것으로 상호 구분되는 것일 수 있다.

[0052] 예를 들어, 각각의 모델 분할 시나리오는 각 가속기(210)에 할당되는 레이어 개수가 상이하게 결정되는 것은 물론, 동일한 타입의 레이어 개수들의 패턴으로 전체 레이어가 분할되는 경우에도 각 레이어 그룹을 할당받는 가속기의 유형이 상이하게 결정되는 것 역시 상호 구분되는 모델 분할 시나리오로 취급될 수 있다.

[0053] 구체적으로 파티션 탐색 장치(100)는 복수의 가속기(210) 각각에 대하여 할당되는 레이어 개수를 조정하기 위한 탐색 보폭에 대응하는 깊이 값을 조정하면서 비용 함수를 연산할 수 있다.

[0054] 한편, 본원의 일 실시예에 따르면, 파티션 탐색 장치(100)는 복수의 가속기(210) 각각에 대응하는 식별 정보를 나타내는 플래그 정보를 미리 정의하고, 모델 병렬화 시 각 가속기(210)에 대하여 적용될 최적 레이어 개수를 탐색하기 위한 레이어 개수의 할당 값을 탐색 과정에서 가변(조정)하는 정도(탐색 보폭)을 나타내는 깊이 값을 각각의 탐색 회차마다 조정하여 최적 분할 시나리오를 탐색하는 알고리즘을 적용할 수 있다.

[0055] 이와 관련하여, 파티션 탐색을 수행마다 초기화 작업(Initializing) 및 전달 작업(Forwarding)이 수행되어야 하므로, 모델 분할 시나리오의 전체 경우의 수를 테스트하는 것은 학습 시간(Training time) 측면에서 큰 부담으로 작용할 수 있으며, 이는, 모델 병렬화를 통해 할당할 레이어의 수 및 가속기의 개수가 커질수록 더욱 큰 폭으로 증가하게 되므로, 본원에서 개시하는 파티션 탐색 장치(100)는 깊이 기반의 탐색을 적용함으로써 최적 분할 시나리오를 도출하기 위한 탐색 횟수를 조절하고, 개별 가속기에 대한 테스트 결과를 종합한 최적의 파티션 분할 시나리오에 도달할 수 있도록 하는 탐색 알고리즘을 적용한 것이다.

[0056] 참고로, 이러한 깊이 기반의 탐색 방식을 적용한 최적 분할 시나리오 탐색을 통해 본원에서 개시하는 파티션 탐색 장치(100)는 하기 식 1에 따른 개수만큼의 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 탐색할 수 있다.

[0057] [식 1]

$${}_nC_2 \sum_{depth=1}^a \left\lfloor \frac{r}{depth} \right\rfloor$$

[0058] 상기 식 1에서, n은 복수의 가속기(210)의 개수이고, r은 레이어 정보에 포함된 딥러닝 모델의 전체 레이어의 개수이고, a는 미리 설정되는 깊이 값의 최대 값일 수 있다.

[0060] 이와 대비하여, 본원에서 개시하는 파티션 탐색 알고리즘을 미적용 하는 탐색 방식(달리 말해, 모든 분할 시나리오를 고려하는 방식)의 경우, 하기 식 2에 따른 개수만큼의 전체 모델 분할 시나리오를 탐색하여야 한다.

[0061] [식 2]

$${}_nH_r = {}_{n+r-1}C_r = {}_{r+n-1}C_{n-1}$$

[0063] 보다 구체적으로, 이해를 돕기 위해 예시하면, 복수의 가속기(210)의 개수가 4개이고(n = 4), 딥러닝 모델의 전체 레이어 개수가 24개이고(r = 24), 깊이 값의 최대 값이 4인 경우(a = 4), 식 2에 따른 종래의 탐색 알고리즘에 의할 때, 2,925 가지의 경우의 수가 최적 분할 시나리오 탐색을 위하여 고려되어야 하는 반면, 본원에서 개시하는 파티션 탐색 장치(100)는 300가지의 경우의 수만을 탐색하여 최적 분할 시나리오를 도출할 수 있게 된다.

[0064] 이하에서는 도 2를 참조하여 각 탐색 회차마다 파티션 탐색 장치(100)가 수행하는 테스트 프로세스를 설명하도록 한다.

[0065] 도 2는 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치의 최적 분할 시나리오 탐색 프로세스를 설명하기 위한 개념도이다.

[0066] 도 2를 참조하면, 파티션 탐색 장치(100)에 의해 도출되는 'n_gpu' 값은 탐색 완료 후 딥러닝 모델의 실제 훈련 시 각 가속기(GPU)에 할당할 레이어의 개수(달리 말해, 최적 분할 시나리오)를 나타내고, 'temp' 값은 해당 탐색 회차(테스트) 중 각 가속기(GPU)에 할당할 레이어의 개수를 나타내고, 'depth' 값은 탐색 보폭에 대응하는 깊이 값으로써 각 탐색 회차(테스트)마다 미리 설정된 기준 크기만큼 감소(도 2의 경우, 1씩 감소)하도록 하여 탐색 초반부에는 상대적으로 넓은(큰) 보폭으로 탐색이 이루어지고, 탐색 후반부에는 상대적으로 좁은(작은) 보폭으로 정밀하게 탐색이 이루어지도록 한 것일 수 있다.

[0067] 또한, 'flag' 값은 테스트 내에서 레이어 개수를 가변할 가속기(GPU)를 변경하기 위한 식별자를 나타내고, 이에 따라 플래그 정보(flag 값)을 통해 모든 경우의 수에 대한 탐색을 마치면, 파티션 탐색 장치(100)는 해당 탐색 회차(테스트)를 종료하고, 깊이 값을 기준 크기(예를 들면, '1' 등)만큼 감소시켜 탐색 보폭을 줄인 후 다시 탐색 회차(테스트)를 수행할 수 있다. 또한, 깊이 값이 미리 설정된 하한 값(도 2의 경우, 0에 해당)에 도달하면, 전체 탐색 회차가 종료된 것으로 판단하여, 비용 함수가 최소인 것으로 도출된 레이어 할당 개수가 각각의 가속기마다 결정되어 최적 분할 시나리오가 특정될 수 있다.

- [0068] 즉, 파티션 탐색 장치(100)는 할당할 레이어의 개수를 포함하는 레이어 정보, 복수의 가속기(210)의 가속기 개수를 포함하는 가속기 정보 및 미리 정의된 깊이 값(depth) 내지 깊이 값의 최대값에 대한 정보를 획득할 수 있다.
- [0069] 각 탐색 회차에서 수행되는 탐색 프로세스를 살펴보면, 먼저 파티션 탐색 장치(100)는 복수의 가속기(210) 각각에 대응하는 플래그 정보에 기초하여 복수의 가속기(210) 중에서 모델 병렬화를 위한 레이어 개수의 할당 값을 조정할 대상 가속기를 특정할 수 있다.
- [0070] 이와 관련하여, 도 2의 하측에 도시된 'depth -=1, flag = 0' 블록을 참조하면, 각각의 탐색 회차마다 플래그 정보가 디바이스(200)에 탑재되는 첫 번째 가속기(210a)에 대응하는 시작 번호(0번)으로 갱신되는 것을 확인할 수 있고, 도 2에 가운데에 도시된 'flag += 1' 블록을 참조하면, 특정 깊이 값이 적용되는 탐색 회차 내에서 플래그 정보가 변경(1씩 증가)되면서 각 가속기에 대한 레이어 개수 할당 값 조정 및 비용 함수 연산이 반복적으로 수행되는 것을 확인할 수 있다.
- [0071] 이어서, 파티션 탐색 장치(100)는 특정된 대상 가속기에 대하여 기 적용된 할당 값에 대한 비용 함수의 제1연산 결과와 해당 탐색 회차에서 적용되는 깊이 값에 기초하여 갱신된 할당 값에 대한 비용 함수의 제2연산 결과를 비교할 수 있다.
- [0072] 또한, 파티션 탐색 장치(100)는 제1연산 결과 및 제2연산 결과에 대한 비교 결과에 기초하여 대상 가속기에 대하여 적용되는 레이어 개수의 할당 값 또는 대상 가속기를 식별하기 위한 플래그 정보를 조정할 수 있다.
- [0073] 달리 말해, 파티션 탐색 장치(100)는 도 2의 4번째 블록 및 5번째 블록에 도시된 바와 같이, 해당 탐색 회차에 대하여 적용되는 깊이 값을 레이어 개수가 조정되는 폭(보폭)으로 하여 temp0 및 temp1 값을 변경하면서 비용 함수를 연산하여 변경된 레이어 개수에 따른 비용 함수의 연산 결과가 최소가 되도록 하는 레이어 개수 할당 패턴이 결정되도록 하고, 해당 가속기에 대한 레이어 개수 할당 패턴이 비용 함수가 최소가 되도록 하는 패턴으로 결정되면, 플래그 정보를 변경(도 2의 'flag += 1' 블록 참조)하여 다음 가속기에 대한 레이어 개수 할당을 위한 테스트가 이루어지도록 할 수 있다.
- [0074] 또한, 파티션 탐색 장치(100)는 플래그 정보에 따라 특정 탐색 회차에서 적용된 깊이 값을 기초로 한 복수의 가속기(210) 각각에 대한 레이어 개수의 할당 값 조정 및 비용 함수 연산을 완결하면, 깊이 값을 미리 설정된 기준 크기만큼 감소시킬 수 있다.
- [0075] 또한, 본원의 일 실시예에 따르면, 의해 깊이 값이 조정(감소)되고 나면, 파티션 탐색 장치(100)는 감소된 깊이 값을 기초로 하여 다음 탐색 회차를 수행할 수 있다.
- [0076] 또한, 본원의 일 실시예에 따르면, 파티션 탐색 장치(100)에 의해 탐색 회차가 반복 수행됨으로써 깊이 값이 미리 설정된 하한 값에 도달하면, 최적 분할 시나리오 탐색 프로세스가 종료되고, 이에 따라 복수의 가속기(210) 각각에 대한 레이어 개수의 할당 값(최적 할당 값)에 대한 정보를 포함하는 최적 분할 시나리오가 결정되는 것일 수 있다.
- [0077] 또한, 파티션 탐색 장치(100)는 앞서 상세히 설명한 최적 분할 시나리오 탐색 프로세스를 통해 선정된 최적 분할 시나리오에 기초하여 해당 딥러닝 모델의 학습을 수행할 수 있다.
- [0078] 도 3은 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치의 개략적인 구성도이다.
- [0079] 도 3을 참조하면, 파티션 탐색 장치(100)는 수집부(110), 탐색 수행부(120) 및 학습부(130)를 포함할 수 있다.
- [0080] 수집부(110)는 딥러닝 모델을 이루는 레이어 정보 및 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기(210)에 대한 가속기 정보를 획득할 수 있다.
- [0081] 탐색 수행부(120)는 획득한 레이어 정보 및 가속기 정보에 기초하여 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정할 수 있다.
- [0082] 구체적으로 탐색 수행부(120)는 복수의 가속기(210) 각각에 대하여 할당되는 레이어 개수를 조정하기 위한 탐색 보폭에 대응하는 깊이 값을 조정하면서 비용 함수를 연산할 수 있다.
- [0083] 한편, 본원의 일 실시예에 따르면, 탐색 수행부(120)는 복수의 가속기(210) 각각에 대응하는 식별 정보를 나타내는 플래그 정보를 미리 정의하고, 모델 병렬화 시 각 가속기(210)에 대하여 적용될 최적 레이어 개수를 탐색

하기 위한 레이어 개수의 할당 값을 탐색 과정에서 가변(조정)하는 정도(탐색 폭폭)을 나타내는 깊이 값을 각각의 탐색 회차마다 조정하여 최적 분할 시나리오를 탐색하는 알고리즘을 적용할 수 있다. 이하에서는 각 탐색 회차마다 탐색 수행부(120)가 수행하는 레이어 개수 할당, 비용 함수 연산, 탐색 파라미터 조정 등에 대한 세부 알고리즘 동작 방식을 설명하도록 한다.

- [0084] 이와 관련하여 도 4는 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치의 탐색 수행부의 세부 구성도이다.
- [0085] 도 4를 참조하면, 탐색 수행부(120)는 가속기 식별부(121), 연산 수행부(122), 탐색 설정부(123) 및 깊이 조정부(124)를 포함할 수 있다.
- [0086] 가속기 식별부(121)는 복수의 가속기(210) 각각에 대응하는 플래그 정보에 기초하여 복수의 가속기(210) 중에서 모델 병렬화를 위한 레이어 개수의 할당 값을 조정할 대상 가속기를 특정할 수 있다.
- [0087] 연산 수행부(122)는 특정된 대상 가속기에 대하여 기 적용된 할당 값에 대한 비용 함수의 제1연산 결과와 해당 탐색 회차에서 적용되는 깊이 값에 기초하여 갱신된 할당 값에 대한 비용 함수의 제2연산 결과를 비교할 수 있다.
- [0088] 탐색 설정부(123)는 제1연산 결과 및 제2연산 결과에 대한 비교 결과에 기초하여 대상 가속기에 대하여 적용되는 레이어 개수의 할당 값 또는 대상 가속기를 식별하기 위한 플래그 정보를 조정할 수 있다.
- [0089] 또한, 플래그 정보에 따라 특정 탐색 회차에서 적용된 깊이 값을 기초로 한 복수의 가속기(210) 각각에 대한 레이어 개수의 할당 값 조정 및 비용 함수 연산이 완료된 것으로 판단되면, 깊이 조정부(124)는 깊이 값을 미리 설정된 기준 크기만큼 감소시킬 수 있다.
- [0090] 또한, 본원의 일 실시예에 따르면, 깊이 조정부(124)에 의해 깊이 값이 조정(감소)되고 나면, 탐색 수행부(120)는 감소된 깊이 값을 기초로 하여 다음 탐색 회차를 수행할 수 있다.
- [0091] 또한, 본원의 일 실시예에 따르면, 탐색 수행부(120)가 수행하는 탐색 회차가 반복 수행됨으로써 깊이 값이 미리 설정된 하한 값에 도달하면, 최적 분할 시나리오 탐색 프로세스가 종료되고, 이에 따라 복수의 가속기(210) 각각에 대한 레이어 개수의 할당 값(최적 할당 값)에 대한 정보를 포함하는 최적 분할 시나리오가 결정되는 것일 수 있다.
- [0092] 학습부(130)는 탐색 수행부(120)에 의해 선정된 최적 분할 시나리오에 기초하여 딥러닝 모델의 학습을 수행할 수 있다. 이와 관련하여, 본원의 일 실시예에 따르면, 학습부(150)는 선정된 최적 분할 시나리오에 따른 모델 병렬화 적용 시의 학습 파라미터(즉, 가속기 별 레이어 할당 개수 등)를 도출하고, 도출된 학습 파라미터를 디바이스(200)에 전송하도록 동작할 수 있다.
- [0093] 이하에서는 상기에 자세히 설명된 내용을 기반으로, 본원의 동작 흐름을 간단히 살펴보기로 한다.
- [0094] 도 5는 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법에 대한 동작 흐름도이다.
- [0095] 도 5에 도시된 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법은 앞서 설명된 파티션 탐색 장치(100)에 의하여 수행될 수 있다. 따라서, 이하 생략된 내용이라고 하더라도 파티션 탐색 장치(100)에 대하여 설명된 내용은 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법에 대한 설명에도 동일하게 적용될 수 있다.
- [0096] 도 5를 참조하면, 단계 S11에서 수집부(110)는 (a) 딥러닝 모델을 이루는 레이어 정보 및 딥러닝 모델에 대한 모델 병렬화를 수행하기 위한 복수의 가속기(210)에 대한 가속기 정보를 획득할 수 있다.
- [0097] 다음으로, 단계 S12에서 탐색 수행부(120)는 단계 S11에서 획득한 (b) 레이어 정보 및 가속기 정보에 기초하여 딥러닝 모델의 학습을 위한 복수의 모델 분할 시나리오 각각에 대응하는 비용 함수를 연산하여 복수의 모델 분할 시나리오 중에서 최적 분할 시나리오를 선정할 수 있다.
- [0098] 구체적으로 단계 S12에서 탐색 수행부(120)는 복수의 가속기(210) 각각에 대하여 할당되는 레이어 개수를 조정하기 위한 탐색 폭폭에 대응하는 깊이 값을 조정하면서 비용 함수를 연산할 수 있다.
- [0099] 또한, 본원의 일 실시예에 따르면, 단계 S12에서 탐색 수행부(120)는 모델 분할 시나리오 각각에 대응하는 네트워크 트래픽 및 메모리 사용량을 기초로 하여 비용 함수를 연산할 수 있다.
- [0100] 다음으로, 단계 S13에서 학습부(130)는 (c) 선정된 최적 분할 시나리오에 기초하여 딥러닝 모델의 학습을 수행

할 수 있다.

- [0101] 상술한 설명에서, 단계 S11 내지 S13은 본원의 구현예에 따라서, 추가적인 단계들로 더 분할되거나, 더 적은 단계들로 조합될 수 있다. 또한, 일부 단계는 필요에 따라 생략될 수도 있고, 단계 간의 순서가 변경될 수도 있다.
- [0102] 도 6은 본원의 일 실시예에 따른 최적 분할 시나리오 선정 프로세스에 대한 세부 동작 흐름도이다.
- [0103] 도 6에 도시된 최적 분할 시나리오 선정 프로세스는 앞서 설명된 파티션 탐색 장치(100)에 의하여 수행될 수 있다. 따라서, 이하 생략된 내용이라고 하더라도 파티션 탐색 장치(100)에 대하여 설명된 내용은 도 6에 대한 설명에도 동일하게 적용될 수 있다.
- [0104] 도 6을 참조하면, 단계 S121에서 가속기 식별부(121)는 (b1) 복수의 가속기(210) 각각에 대응하는 플래그 정보에 기초하여 복수의 가속기(210) 중에서 모델 병렬화를 위한 레이어 개수의 할당 값을 조정할 대상 가속기를 특정할 수 있다.
- [0105] 다음으로, 단계 S122에서 연산 수행부(122)는 (b2) 단계 S121에서 특정된 대상 가속기에 대하여 기 적용된 할당 값에 대한 비용 함수의 제1연산 결과와 해당 탐색 회차에서 적용되는 깊이 값에 기초하여 갱신된 할당 값에 대한 비용 함수의 제2연산 결과를 비교할 수 있다.
- [0106] 다음으로, 단계 S123에서 탐색 설정부(123)는 (b3) 단계 S122에서의 비교 결과에 기초하여 대상 가속기에 대하여 적용되는 레이어 개수의 할당 값 또는 대상 가속기를 식별하기 위한 플래그 정보를 조정할 수 있다.
- [0107] 만일, 플래그 정보에 따라 특정 탐색 회차에서 적용된 깊이 값을 기초로 한 복수의 가속기(210) 각각에 대한 레이어 개수의 할당 값 조정 및 비용 함수 연산이 완결된 것으로 판단되면, 다음으로, 단계 S124에서 깊이 조정부(124)는 (b4) 깊이 값을 미리 설정된 기준 크기만큼 감소시킬 수 있다.
- [0108] 또한, 본원의 일 실시예에 따르면, 단계 S124를 통해 깊이 값이 조정(감소)되고 나면, 탐색 수행부(120)는 감소된 깊이 값을 기초로 하여 전술한 단계 S121 내지 단계 S123을 수행(다시 말해, 다음 탐색 회차를 수행)할 수 있다.
- [0109] 또한, 본원의 일 실시예에 따르면, 전술한 단계 S121 내지 단계 S124를 수행하는 탐색 회차가 반복 수행됨으로써 깊이 값이 미리 설정된 하한 값에 도달하면, 단계 S12는 종료되고, 이에 따라 복수의 가속기(210) 각각에 대한 레이어 개수의 할당 값에 대한 최적 값에 대한 정보를 포함하는 최적 분할 시나리오가 결정되는 것일 수 있다.
- [0110] 상술한 설명에서, 단계 S121 내지 S124는 본원의 구현예에 따라서, 추가적인 단계들로 더 분할되거나, 더 적은 단계들로 조합될 수 있다. 또한, 일부 단계는 필요에 따라 생략될 수도 있고, 단계 간의 순서가 변경될 수도 있다.
- [0111] 본원의 일 실시예에 따른 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 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 본 발명을 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 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0112] 또한, 전술한 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 방법은 기록 매체에 저장되는 컴퓨터에 의해 실행되는 컴퓨터 프로그램 또는 애플리케이션의 형태로도 구현될 수 있다.
- [0113] 전술한 본원의 설명은 예시를 위한 것이며, 본원이 속하는 기술분야의 통상의 지식을 가진 자는 본원의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만

한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0114] 본원의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본원의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0115]

10: 딥러닝 모델 학습 시스템

100: 딥러닝 모델의 병렬 처리를 위한 파티션 탐색 장치

110: 수집부

120: 탐색 수행부

121: 가속기 식별부

122: 연산 수행부

123: 탐색 설정부

124: 깊이 조정부

130: 학습부

200: 디바이스

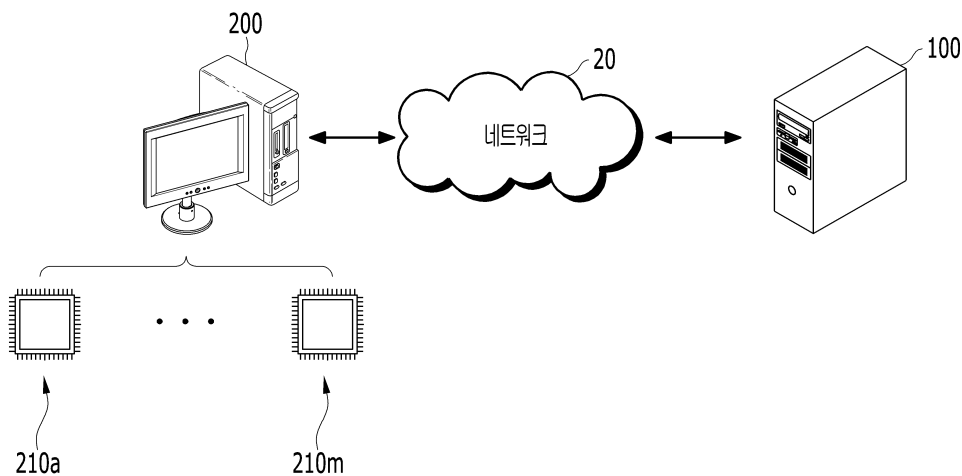
210: 가속기

20: 네트워크

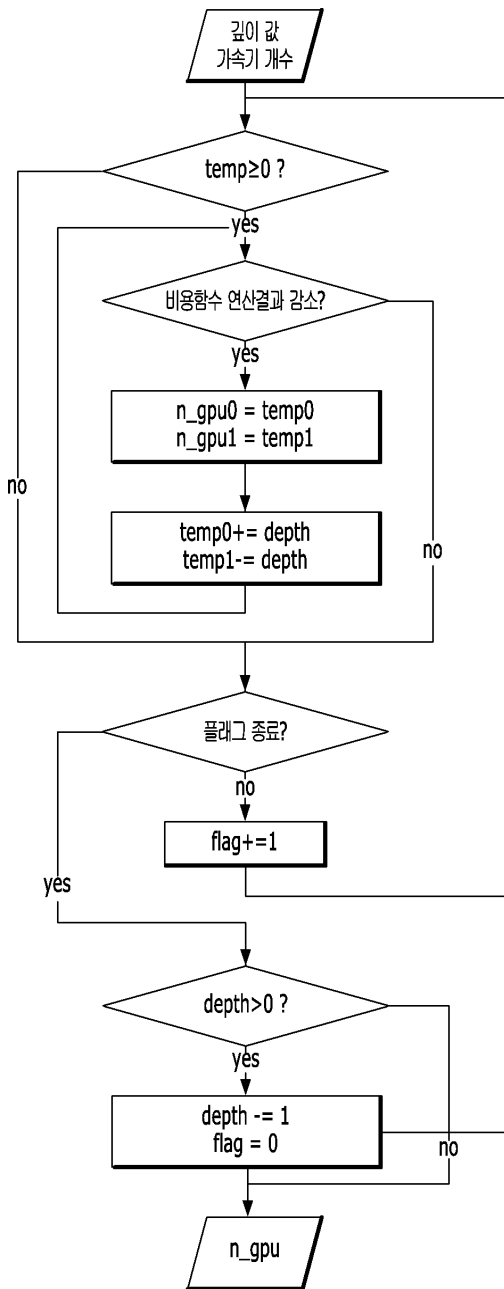
도면

도면1

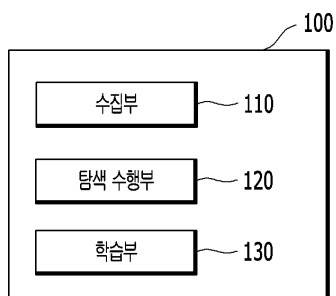
10



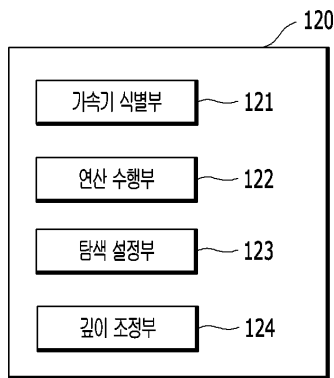
도면2



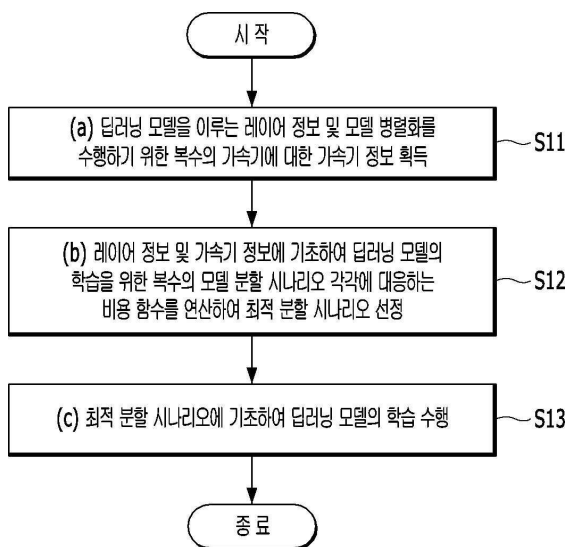
도면3



도면4



도면5



도면6

