



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0097530
(43) 공개일자 2025년06월30일

(51) 국제특허분류(Int. Cl.)
G06F 9/50 (2018.01) G06N 20/00 (2019.01)
(52) CPC특허분류
G06F 9/5077 (2013.01)
G06F 9/5016 (2013.01)
(21) 출원번호 10-2023-0188923
(22) 출원일자 2023년12월21일
심사청구일자 2023년12월21일

(71) 출원인
국립부경대학교 산학협력단
부산광역시 남구 용소로 45, 부경대학교내 (대연동)
(72) 발명자
한영선
경기도 성남시 분당구 서판교로 29, 905동 101호 (판교동, 판교원마을)
포브 김사이
부산광역시 남구 용소로 45, 국제교류부(대연동)
이상현
부산광역시 사하구 하신변영로 400 (하단동, 하단동에스케이뷰아파트)
(74) 대리인
오위환, 박진기, 정기택

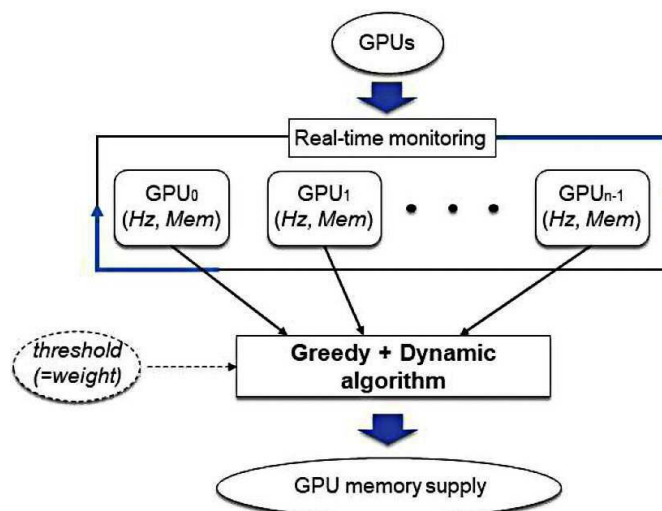
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법

(57) 요약

본 발명은 복수의 GPU가 존재하는 컴퓨팅 환경에서 DL 모델의 학습(Training) 및 추론(Inference) 작업에 요구되는 GPU 자원을 효율적으로 작업에 할당할 수 있도록 한 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법에 관한 것으로, 메모리 리소스(Memory resource) 추가 할당에 관한 기준치를 설정하는 추가 메모리 기준치 설정부;사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하는 실시간 자원 모니터링부;실시간 자원 모니터링을 통하여 각 GPU의 현재 상태 데이터를 획득하는 GPU 데이터 획득부;획득한 GPU 데이터를 바탕으로 그리디-다이나믹 알고리즘(Greedy + Dynamic Algorithm)을 실행하는 그리디-다이나믹 알고리즘 실행부;그리디-다이나믹 알고리즘 실행이 종료되면 GPU 메모리를 기존의 작업에 추가로 공급하여 DL 모델 학습 및 추론 작업을 재개하도록 하는 메모리 리소스 지원부;를 포함하는 것이다.

대표도 - 도1



(52) CPC특허분류

G06F 9/5027 (2013.01)

G06N 20/00 (2021.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193257
과제번호	2021-0-0031061782064340003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	(신통합 이지바로)
연구과제명	인공지능 학습/추론 효율성 향상을 위한 서버용 SW 프레임워크 개발
기 여 율	1/1
과제수행기관명	국립부경대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

메모리 리소스(Memory resource) 추가 할당에 관한 기준치를 설정하는 추가 메모리 기준치 설정부;

사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하는 실시간 자원 모니터링부;

실시간 자원 모니터링을 통하여 각 GPU의 현재 상태 데이터를 획득하는 GPU 데이터 획득부;

획득한 GPU 데이터를 바탕으로 그리디-다이나믹 알고리즘(Greedy + Dynamic Algorithm)을 실행하는 그리디-다이나믹 알고리즘 실행부;

그리디-다이나믹 알고리즘 실행이 종료되면 GPU 메모리를 기존의 작업에 추가로 공급하여 DL 모델 학습 및 추론 작업을 재개하도록 하는 메모리 리소스 지원부;를 포함하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치.

청구항 2

제 1 항에 있어서, 복수의 GPU 중 단 하나의 GPU 만을 학습 혹은 추론 작업에 할당하여 DL 모델 학습 및 추론을 수행하는 DL 모델 학습 및 추론부를 더 포함하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치.

청구항 3

제 1 항에 있어서, 데이터 획득부에서 획득되는 GPU의 현재 상태 데이터는,

Clock frequency, Memory usage 데이터를 포함하고,

메모리 할당을 위한 GPU의 우선순위를 결정하기 위하여 각 GPU의 value ($=\text{Clock frequency}/\text{memory usage}$)으로 산출하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치.

청구항 4

제 1 항에 있어서, 그리디-다이나믹 알고리즘 실행부는,

그리디-다이나믹 알고리즘이 실행되면 현재 GPU의 GPU의 현재 상태 데이터를 입력으로 받는 데이터 수신부와,

작업에 추가로 필요한 메모리 사용량(Threshold)을 추가로 전달하기 위해 'Greedy algorithm'을 적용하는 그리디 알고리즘 실행부와,

각 GPU의 value($=\text{Clock frequency}/\text{memory usage}$)를 비교하는 GPU 데이터 비교부와,

GPU 데이터 비교 결과를 기준으로 메모리 할당을 위한 GPU의 우선순위를 결정하고, 순위에 따라 선택된 GPU로부터 메모리 리소스를 획득하는 GPU 우선순위 결정부와,

'Greedy algorithm'이 종료된 이후, 추가된 자원량이 기준치를 만족시키는지 확인하는 추가 메모리 확인부를 포함하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치.

청구항 5

제 4 항에 있어서, 그리디-다이나믹 알고리즘 실행부는,

추가 메모리 확인부의 확인 과정에서 추가된 자원량이 기준치를 만족시키지 못하면,

이전 알고리즘에서 선택되지 않았던, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 하는 Dynamic Programming을 적용하는 다이나믹 프로그래밍 실행부를 더 포함하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치.

청구항 6

복수의 GPU 중 단 하나의 GPU 만을 학습 혹은 추론 작업에 할당되면, 메모리 리소스(Memory resource) 추가 할당에 관한 기준치를 설정하는 단계;

DL 모델 학습 및 추론 작업을 수행하는 과정에서 GPU의 Memory가 부족한 상황이 발생하는지 모니터링을 하는 단계;

모니터링 과정에서 추가적인 GPU 자원이 필요한 경우, 작업에 사용되지 않은 GPU로부터 사용자가 요구하는 추가 메모리 기준치(Threshold)만큼의 메모리 리소스 (Memory resource)을 추가로 가져와 작업에 할당하는 단계;

사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하고 이를 통하여 각 GPU의 현재 상태 데이터를 획득하는 단계;

획득한 GPU 데이터를 바탕으로 그리디-다이나믹 알고리즘(Greedy + Dynamic Algorithm)을 실행하고, 알고리즘이 종료되면 GPU 메모리를 기존의 작업에 추가로 공급하는 단계;를 포함하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치.

청구항 7

제 6 항에 있어서, GPU의 현재 상태 데이터는,

Clock frequency, Memory usage 데이터를 포함하고,

메모리 할당을 위한 GPU의 우선순위를 결정하기 위하여 각 GPU의 value ($=\text{Clock frequency}/\text{memory usage}$)으로 산출하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법.

청구항 8

제 6 항에 있어서, 그리디-다이나믹 알고리즘(Greedy + Dynamic Algorithm)을 실행하는 단계는,

현재 GPU의 현재 상태 데이터를 입력으로 받는 단계와,

작업에 추가로 필요한 메모리 사용량(Threshold)을 추가로 전달하기 위해, 먼저 'Greedy algorithm'을 실행하는 단계와,

각 GPU의 value ($=\text{Clock frequency}/\text{memory usage}$)를 비교하여 메모리 할당을 위한 GPU의 우선순위를 결정하고, 순위에 따라 선택된 GPU로부터 메모리 리소스를 획득하는 단계와,

'Greedy algorithm'이 종료된 이후, 추가된 자원량이 기준치를 만족시키는지 판단하는 단계와,

추가된 자원량이 기준치를 만족시키지 못한다면, 'Dynamic Programming'을 실행하는 단계를 포함하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법.

청구항 9

제 8 항에 있어서, 'Dynamic Programming'은 이전 알고리즘에서 선택되지 않았던, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 하고,

Dynamic Programming을 통하여 기준치만큼의 추가 메모리(GPU memory resource)를 획득하면, 알고리즘을 종료하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법.

청구항 10

제 8 항에 있어서, Value가 동일한 경우 우선순위는 GPU 장치의 순서(order)를 통해 순위를 결정하고,

여러 개의 GPU가 존재하는 시스템에서 각 GPU에 순서가 매겨지는 순서(order)의 더 작은 번호를 갖는 GPU가 높은 우선순위를 갖도록 하는 것을 특징으로 하는 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법.

발명의 설명

기술분야

[0001] 본 발명은 DL 모델 학습 및 추론에 관한 것으로, 구체적으로 복수의 GPU가 존재하는 컴퓨팅 환경에서 DL 모델의 학습(Training) 및 추론(Inference) 작업에 요구되는 GPU 자원을 효율적으로 할당할 수 있도록 한 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법에 관한 것이다.

배경기술

- [0002] 현재 딥러닝, 블록체인, 유전자 분석 등 다양한 분야에서 GPU가 병렬 연산장치로 활용되고 있다.
- [0003] 그러나 GPU 관리는 워크로드의 다양성을 반영하기보다 간결한 제어 구조를 통한 개별 워크로드의 병렬성 극대화에 초점을 맞추어왔다.
- [0004] 최근 다양한 분야의 워크로드가 GPU에서 동시에 수행되는 것이 지원되면서 개별 워크로드의 병렬 수행보다 멀티태스킹으로 인한 효율성을 높이는 것이 GPU 관리의 중요한 이슈로 부각되고 있다.
- [0005] 한편, 딥 러닝은 인공신경망(artificial neural network)에 많은 수의 레이어(layer)를 만들고 학습시키는 여러 가지 방법을 통틀어 일컫는다.
- [0006] 인공신경망에 대한 연구는 오래전부터 진행되어 왔으나 방대한 계산량 때문에 2000년대 중반에 이르기까지 실용화 되지 못하였다. 특히 GPU를 활용하여 딥 뉴럴 네트워크(DNN; Deep Neural Network)를 이용한 딥러닝을 수행함에 있어서, GPU 메모리 용량의 한계라는 어려움이 있어 왔다.
- [0007] 이와 같이 최근 GPU가 그래픽 처리뿐 아니라 다양한 분야의 병렬 처리로 그 영역을 넓혀가고 있다.
- [0008] 그러나 현재 GPU는 워크로드의 다양성을 반영하기보다 간결한 제어 구조를 통한 개별 워크로드의 병렬성 극대화에 초점을 맞추고 있다.
- [0009] 3D 처리를 하던 GPU는 스마트디바이스의 저전력 고성능 처리를 위해 OpenCL, CUDA, OpenMP 등 GPGPU(General Purpose computing on GPU)를 도입하면서 기존 CPU가 맡았던 응용프로그램들의 계산을 담당하게 되었다.
- [0010] 특히, DL 모델의 학습 및 추론 작업을 보다 빠르게 처리하기 위해서는 다수의 Graphic Processing Unit(GPU)가 존재하는 컴퓨팅 환경이 필수적이며 일반적으로 주어진 컴퓨팅 환경에 존재하는 모든 GPU를 실험에 사용한다.
- [0011] 그러나 이는 불필요한 GPU까지 실험에 사용되는 문제점이 있다.
- [0012] 따라서, 복수의 GPU가 존재하는 컴퓨팅 환경에서 DL 모델의 학습(Training) 및 추론(Inference) 작업에 요구되는 GPU 자원을 효율적으로 할당할 수 있도록 하는 새로운 기술의 개발이 요구되고 있다.

선행기술문헌

특허문헌

- [0013] (특허문헌 0001) 대한민국 공개특허 제10-2023-0068918호
(특허문헌 0002) 대한민국 공개특허 제10-2022-0141592호
(특허문헌 0003) 대한민국 등록특허 제10-2064882호

발명의 내용

해결하려는 과제

- [0014] 본 발명은 종래 기술의 DL 모델 학습 및 추론 기술의 문제점을 해결하기 위한 것으로, 복수의 GPU가 존재하는 컴퓨팅 환경에서 DL 모델의 학습(Training) 및 추론(Inference) 작업에 요구되는 GPU 자원을 효율적으로 작업에 할당할 수 있도록 한 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법을 제공하는데 그 목적이 있다.
- [0015] 본 발명은 작업을 수행하는 과정에서 GPU의 Memory가 부족한 상황이 발생하여, 추가적인 GPU 자원이 필요한 경우, 작업에 사용되지 않은 GPU로부터 사용자가 요구하는 추가 메모리 기준치(Threshold)만큼의 메모리 리소스

(Memory resource)을 추가로 가져와 할당하여 전체 학습에서 사용되는 자원(Hardware resource)을 효과적으로 줄일 수 있도록 한 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법을 제공하는데 그 목적이 있다.

[0016] 본 발명은 사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하여 각 GPU의 현재 Clock frequency, Memory usage 데이터를 획득하고, 이를 바탕으로 'Greedy + Dynamic Algorithm'을 실행하고 알고리즘이 종료되면 GPU memory를 기존의 작업에 추가로 공급하여 효율적인 자원 관리가 가능하도록 한 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법을 제공하는데 그 목적이 있다.

[0017] 본 발명의 다른 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0018] 상기와 같은 목적을 달성하기 위한 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치는 메모리 리소스(Memory resource) 추가 할당에 관한 기준치를 설정하는 추가 메모리 기준치 설정부; 사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하는 실시간 자원 모니터링부; 실시간 자원 모니터링을 통하여 각 GPU의 현재 상태 데이터를 획득하는 GPU 데이터 획득부; 획득한 GPU 데이터를 바탕으로 그리디-다이나믹 알고리즘(Greedy + Dynamic Algorithm)을 실행하는 그리디-다이나믹 알고리즘 실행부; 그리디-다이나믹 알고리즘 실행이 종료되면 GPU 메모리를 기존의 작업에 추가로 공급하여 DL 모델 학습 및 추론 작업을 재개하도록 하는 메모리 리소스 지원부;를 포함하는 것을 특징으로 한다.

[0019] 여기서, 복수의 GPU 중 단 하나의 GPU 만을 학습 혹은 추론 작업에 할당하여 DL 모델 학습 및 추론을 수행하는 DL 모델 학습 및 추론부를 더 포함하는 것을 특징으로 한다.

[0020] 그리고 데이터 획득부에서 획득되는 GPU의 현재 상태 데이터는, Clock frequency, Memory usage 데이터를 포함하고, 메모리 할당을 위한 GPU의 우선순위를 결정하기 위하여 각 GPU의 value ($=\text{Clock frequency}/\text{memory usage}$)으로 산출하는 것을 특징으로 한다.

[0021] 그리고 그리디-다이나믹 알고리즘 실행부는, 그리디-다이나믹 알고리즘이 실행되면 현재 GPU의 GPU의 현재 상태 데이터를 입력으로 받는 데이터 수신부와, 작업에 추가로 필요한 메모리 사용량(Threshold)을 추가로 전달하기 위해 'Greedy algorithm'을 적용하는 그리디 알고리즘 실행부와, 각 GPU의 value($=\text{Clock frequency}/\text{memory usage}$)를 비교하는 GPU 데이터 비교부와, GPU 데이터 비교 결과를 기준으로 메모리 할당을 위한 GPU의 우선순위를 결정하고, 순위에 따라 선택된 GPU로부터 메모리 리소스를 획득하는 GPU 우선순위 결정부와, 'Greedy algorithm'이 종료된 이후, 추가된 자원량이 기준치를 만족시키는지 확인하는 추가 메모리 확인부를 포함하는 것을 특징으로 한다.

[0022] 그리고 그리디-다이나믹 알고리즘 실행부는, 추가 메모리 확인부의 확인 과정에서 추가된 자원량이 기준치를 만족시키지 못하면, 이전 알고리즘에서 선택되지 않았던, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 하는 Dynamic Programming을 적용하는 다이나믹 프로그래밍 실행부를 더 포함하는 것을 특징으로 한다.

[0023] 다른 목적을 달성하기 위한 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법은 복수의 GPU 중 단 하나의 GPU 만을 학습 혹은 추론 작업에 할당되면, 메모리 리소스(Memory resource) 추가 할당에 관한 기준치를 설정하는 단계; DL 모델 학습 및 추론 작업을 수행하는 과정에서 GPU의 Memory가 부족한 상황이 발생하는지 모니터링을 하는 단계; 모니터링 과정에서 추가적인 GPU 자원이 필요한 경우, 작업에 사용되지 않은 GPU로부터 사용자가 요구하는 추가 메모리 기준치(Threshold)만큼의 메모리 리소스(Memory resource)을 추가로 가져와 작업에 할당하는 단계; 사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하고 이를 통하여 각 GPU의 현재 상태 데이터를 획득하는 단계; 획득한 GPU 데이터를 바탕으로 그리디-다이나믹 알고리즘(Greedy + Dynamic Algorithm)을 실행하고, 알고리즘이 종료되면 GPU 메모리를 기존의 작업에 추가로 공급하는 단계;를 포함하는 것을 특징으로 한다.

[0024] 여기서, GPU의 현재 상태 데이터는, Clock frequency, Memory usage 데이터를 포함하고, 메모리 할당을 위한 GPU의 우선순위를 결정하기 위하여 각 GPU의 value ($=\text{Clock frequency}/\text{memory usage}$)으로 산출하는 것을 특징으로 한다.

[0025] 그리고 그리디-다이나믹 알고리즘(Greedy + Dynamic Algorithm)을 실행하는 단계는, 현재 GPU의 현재 상태 데이

터를 입력으로 받는 단계와, 작업에 추가로 필요한 메모리 사용량(Threshold)을 추가로 전달하기 위해, 먼저 'Greedy algorithm'을 실행하는 단계와, 각 GPU의 value (=Clock frequency/memory usage)를 비교하여 메모리 할당을 위한 GPU의 우선순위를 결정하고, 순위에 따라 선택된 GPU로부터 메모리 리소스를 획득하는 단계와, 'Greedy algorithm'이 종료된 이후, 추가된 자원량이 기준치를 만족시키는지 판단하는 단계와, 추가된 자원량이 기준치를 만족시키지 못한다면, 'Dynamic Programming'을 실행하는 단계를 포함하는 것을 특징으로 한다.

[0026] 그리고 'Dynamic Programming'은 이전 알고리즘에서 선택되지 않았던, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 하고, Dynamic Programming을 통하여 기준치만큼의 추가 메모리 (GPU memory resource)를 획득하면, 알고리즘을 종료하는 것을 특징으로 한다.

[0027] 그리고 Value가 동일한 경우 우선순위는 GPU 장치의 순서(order)를 통해 순위를 결정하고, 여러 개의 GPU가 존재하는 시스템에서 각 GPU에 순서가 매겨지는 순서(order)의 더 작은 번호를 갖는 GPU가 높은 우선순위를 갖도록 하는 것을 특징으로 한다.

발명의 효과

[0028] 이상에서 설명한 바와 같은 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법은 다음과 같은 효과가 있다.

[0029] 첫째, 복수의 GPU가 존재하는 컴퓨팅 환경에서 DL 모델의 학습(Training) 및 추론(Inference) 작업에 요구되는 GPU 자원을 효율적으로 작업에 할당할 수 있도록 한다.

[0030] 둘째, 작업을 수행하는 과정에서 GPU의 Memory가 부족한 상황이 발생하여, 추가적인 GPU 자원이 필요한 경우, 작업에 사용되지 않은 GPU로부터 사용자가 요구하는 추가 메모리 기준치(Threshold)만큼의 메모리 리소스 (Memory resource)을 추가로 가져와 할당하여 전체 학습에서 사용되는 자원(Hardware resource)을 효과적으로 줄일 수 있도록 한다.

[0031] 셋째, 사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하여 각 GPU의 현재 Clock frequency, Memory usage 데이터를 획득하고, 이를 바탕으로 'Greedy + Dynamic Algorithm'을 실행하고 알고리즘이 종료되면 GPU memory를 기존의 작업에 추가로 공급하여 효율적인 자원 관리가 가능하도록 한다.

도면의 간단한 설명

[0032] 도 1은 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치의 전체 구성도

도 2는 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치의 상세 구성도

도 3은 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법을 나타낸 플로우 차트

도 4는 그리디-다이나믹 알고리즘 실행부의 상세 구성도

도 5는 그리디-다이나믹 알고리즘 실행 동작을 나타낸 플로우 차트

발명을 실시하기 위한 구체적인 내용

[0033] 이하, 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법의 바람직한 실시 예에 관하여 상세히 설명하면 다음과 같다.

[0034] 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법의 특징 및 이점들은 이하에서의 각 실시 예에 대한 상세한 설명을 통해 명백해질 것이다.

[0035] 도 1은 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치의 전체 구성도이다.

[0036] 본 개시에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 관례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세

히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

- [0037] 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.
- [0038] 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법은 복수의 GPU가 존재하는 컴퓨팅 환경에서 DL 모델의 학습(Training) 및 추론(Inference) 작업에 요구되는 GPU 자원을 효율적으로 작업에 할당할 수 있도록 한 것이다.
- [0039] 이를 위하여, 본 발명은 작업을 수행하는 과정에서 GPU의 Memory가 부족한 상황이 발생하여, 추가적인 GPU 자원이 필요한 경우, 작업에 사용되지 않은 GPU로부터 사용자가 요구하는 추가 메모리 기준치(Threshold)만큼의 메모리 리소스(Memory resource)를 추가로 가져와 할당하여 전체 학습에서 사용되는 자원(Hardware resource)을 효과적으로 줄일 수 있도록 하는 구성을 포함할 수 있다.
- [0040] 본 발명은 사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하여 각 GPU의 현재 Clock frequency, Memory usage 데이터를 획득하고, 이를 바탕으로 'Greedy + Dynamic Algorithm'을 실행하고 알고리즘이 종료되면 GPU memory를 기존의 작업에 추가로 공급하여 효율적인 자원 관리가 가능하도록 하는 구성을 포함할 수 있다.
- [0041] 도 2는 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치의 상세 구성도이다.
- [0042] 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치는 도 2에서와 같이, 메모리 리소스(Memory resource) 추가 할당에 관한 기준치를 설정하는 추가 메모리 기준치 설정부(10)와, 사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하는 실시간 자원 모니터링부(20)와, 실시간 자원 모니터링을 통하여 각 GPU의 현재 Clock frequency, Memory usage 데이터를 획득하는 GPU 데이터 획득부(30)와, 획득한 GPU 데이터를 바탕으로 'Greedy + Dynamic Algorithm'을 실행하는 그리디-다이나믹 알고리즘 실행부(40)와, 그리디-다이나믹 알고리즘 실행이 종료되면 GPU memory를 기존의 작업에 추가로 공급하여 DL 모델 학습 및 추론 작업을 재개하도록 하는 메모리 리소스 지원부(50)와, 복수의 GPU 중 단 하나의 GPU 만을 학습 혹은 추론 작업에 할당하여 DL 모델 학습 및 추론을 수행하는 DL 모델 학습 및 추론부(60)를 포함한다.
- [0043] 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법을 구체적으로 설명하면 다음과 같다.
- [0044] 도 3은 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 방법을 나타낸 플로우 차트이다.
- [0045] 먼저, 복수의 GPU 중 단 하나의 GPU 만을 학습 혹은 추론 작업에 할당되면, 메모리 리소스(Memory resource) 추가 할당에 관한 기준치를 설정한다.(S301)
- [0046] 이어, DL 모델 학습 및 추론 작업을 수행하는 과정에서 GPU의 Memory가 부족한 상황이 발생하는지 모니터링을 한다.(S302)
- [0047] 모니터링 과정에서 추가적인 GPU 자원이 필요한 경우, 작업에 사용되지 않은 GPU로부터 사용자가 요구하는 추가 메모리 기준치(Threshold)만큼의 메모리 리소스 (Memory resource)를 추가로 가져와 작업에 할당한다.
- [0048] 그리고 사용되지 않은 GPU들에 대해 실시간 자원 모니터링을 수행하고 이를 통하여 각 GPU의 현재 Clock frequency, Memory usage 데이터를 획득한다.(S303)
- [0049] 이어, 획득한 GPU 데이터를 바탕으로 'Greedy + Dynamic Algorithm'을 실행하고, 알고리즘이 종료되면 GPU memory를 기존의 작업에 추가로 공급한다.
- [0050] GPU memory를 기존의 작업에 추가로 공급하는 작업이 종료되면, DL 모델 학습 및 추론 작업을 재개하고 추가적인 자원 요청이 발생하는지 모니터링을 한다.
- [0051] 그리디-다이나믹 알고리즘 실행부(40)의 상세 구성은 다음과 같다.

- [0052] 도 4는 그리디-다이나믹 알고리즘 실행부의 상세 구성도이다.
- [0053] 그리디-다이나믹 알고리즘 실행부(40)는 도 4에서와 같이, 그리디-다이나믹 알고리즘 실행되면 현재 GPU의 clock frequency와 memory usage 데이터를 입력으로 받는 데이터 수신부(41)와, 작업에 추가로 필요한 메모리 사용량(Threshold)을 추가로 전달하기 위해 'Greedy algorithm'을 적용하는 그리디 알고리즘 실행부(42)와, 각 GPU의 value (=Clock frequency/memory usage)를 비교하는 GPU 데이터 비교부(43)와, GPU 데이터 비교 결과를 기준으로 메모리 할당을 위한 GPU의 우선순위를 결정하고, 순위에 따라 선택된 GPU로부터 메모리 리소스를 획득하는 GPU 우선순위 결정부(44)와, 'Greedy algorithm'이 종료된 이후, 추가된 자원량이 기준치를 만족시키는지 확인하는 추가 메모리 확인부(45)와, 추가된 자원량이 기준치를 만족시키지 못하면, 이전 알고리즘에서 선택되지 않았던, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 하는 Dynamic Programming을 적용하는 다이나믹 프로그래밍 실행부(46)를 포함한다.
- [0054] 여기서, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 하는 Dynamic Programming을 통하여 기준치만큼의 추가 메모리(GPU memory resource)를 획득하면, 알고리즘을 종료한다.
- [0055] 그리디-다이나믹 알고리즘 실행 동작을 구체적으로 설명하면 다음과 같다.
- [0056] 도 5는 그리디-다이나믹 알고리즘 실행 동작을 나타낸 플로우 차트이다.
- [0057] 먼저, 알고리즘이 시작되면 현재 GPU의 clock frequency와 memory usage 데이터를 입력으로 받는다.(S501)
- [0058] 이어, 작업에 추가로 필요한 메모리 사용량(Threshold)을 추가로 전달하기 위해, 먼저 'Greedy algorithm'을 실행한다.(S502)
- [0059] 이때, 각 GPU의 value (=Clock frequency/memory usage)를 비교하여 메모리 할당을 위한 GPU의 우선순위를 결정하고, 순위에 따라 선택된 GPU로부터 메모리 리소스를 획득한다.
- [0060] 이어, 'Greedy algorithm'이 종료된 이후, 추가된 자원량이 기준치를 만족시키는지 판단한다.(S503)
- [0061] 그리고 추가된 자원량이 기준치를 만족시키지 못한다면, 'Dynamic Programming'을 실행한다.(S504)
- [0062] 'Dynamic Programming'은 이전 알고리즘에서 선택되지 않았던, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 한다.
- [0063] 이어, GPU들에 대해 Value를 계산하고 잔여 메모리 요구량을 충족시킬 때까지 반복 동작을 하는 Dynamic Programming을 통하여 기준치만큼의 추가 메모리(GPU memory resource)를 획득하면, 알고리즘을 종료한다.(S505)
- [0064] Memory usage가 동일한 경우, 'Value = Memory usage / Clock frequency'를 적용하는 경우에 Clock frequency가 크면 Value가 작아지는데, 본 발명에서는 'Value = Clock frequency / Memory usage'의 형태로 변경하여 같은 Memory usage에서 Clock frequency가 높은 GPU가 더 큰 Value를 갖게 한다.
- [0065] 그리고 Value가 동일한 경우 우선순위는 GPU 장치의 순서(order)를 통해 순위를 결정한다.
- [0066] 여러 개의 GPU가 존재하는 시스템에서는 각 GPU에 순서가 매겨지기 때문에, 해당 정보를 활용하여, 더 작은 번호를 갖는 GPU가 높은 우선순위를 갖도록 한다.
- [0067] 이상에서 설명한 본 발명에 따른 DL 모델 학습 및 추론의 효율성 향상을 위한 GPU 자원 할당 최적화를 위한 장치 및 방법은 복수의 GPU가 존재하는 컴퓨팅 환경에서 DL 모델의 학습(Training) 및 추론(Inference) 작업에 요구되는 GPU 자원을 효율적으로 작업에 할당할 수 있도록 한 것이다.
- [0068] 이상에서의 설명에서와 같이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 본 발명이 구현되어 있음을 이해할 수 있을 것이다.
- [0069] 그러므로 명시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 하고, 본 발명의 범위는 전술한 설명이 아니라 특허청구 범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

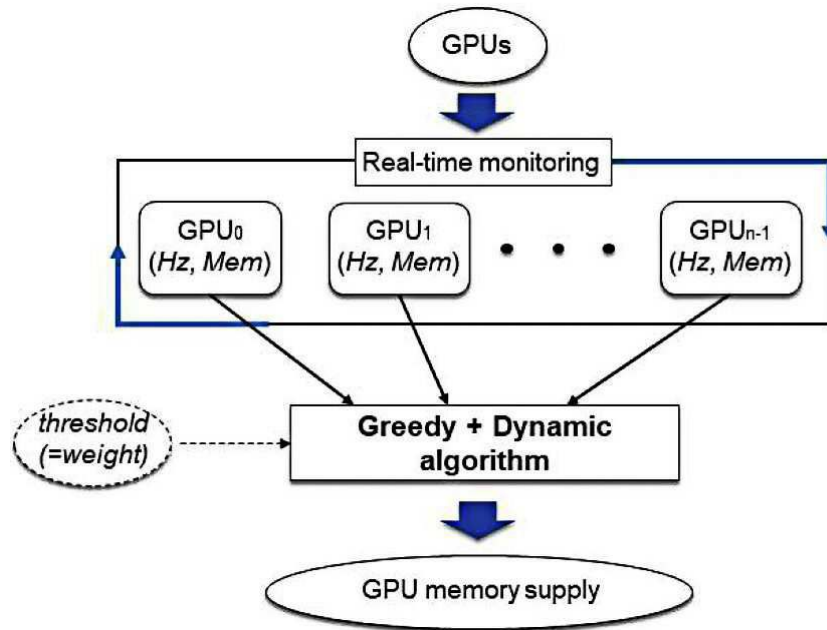
부호의 설명

- [0070] 10. 추가 메모리 기준치 설정부

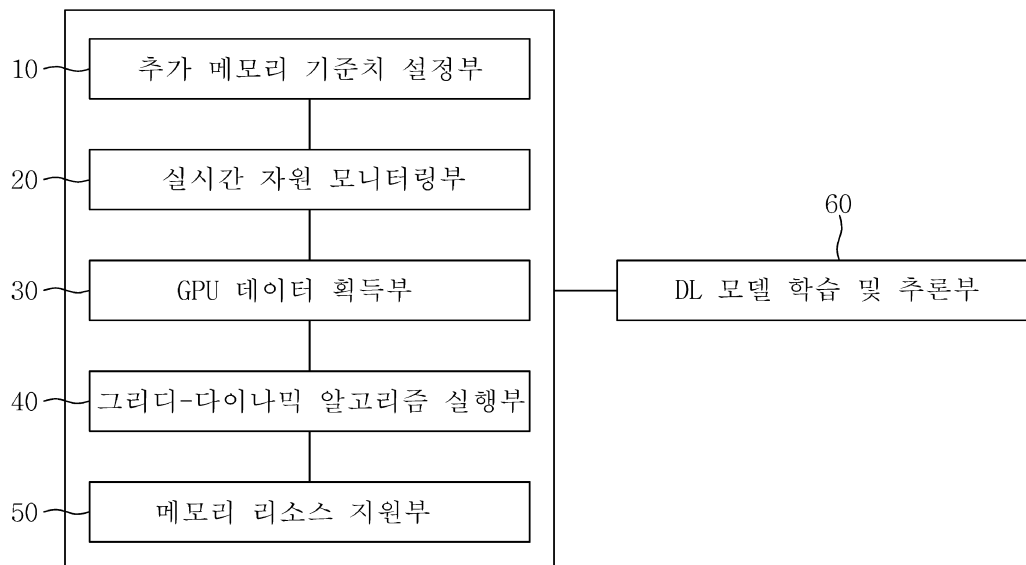
- 20. 실시간 자원 모니터링부
- 30. GPU 데이터 획득부
- 40. 그리디-다이나믹 알고리즘 실행부
- 50. 메모리 리소스 지원부
- 60. DL 모델 학습 및 추론부

도면

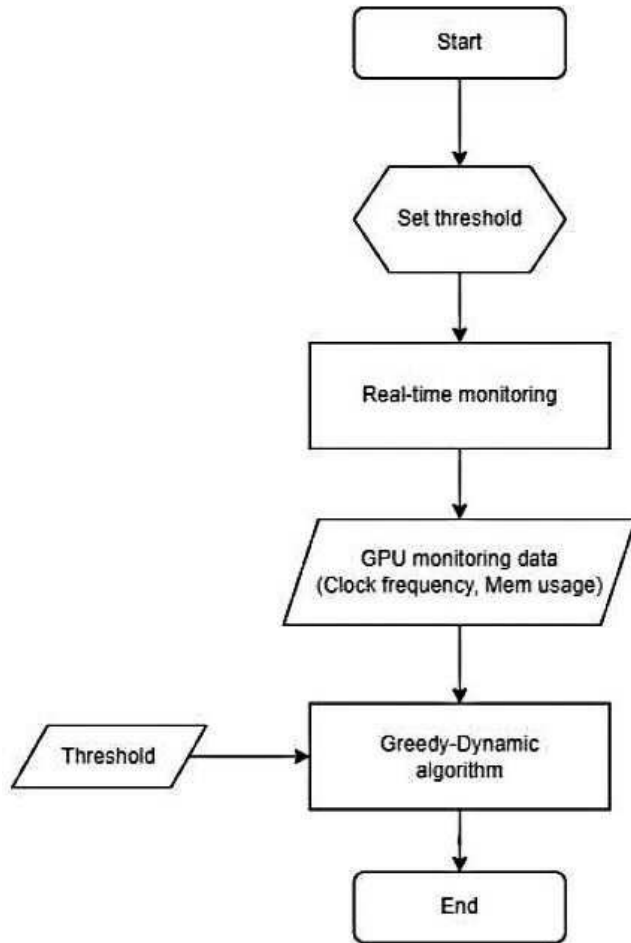
도면1



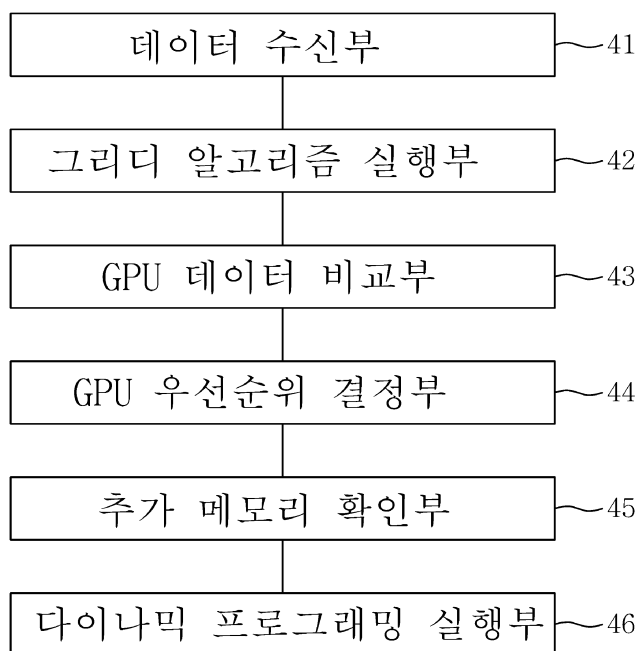
도면2



도면3



도면4



도면5

