(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0067311
(43) 공개일자 2021년06월08일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) *G06F 3/06* (2006.01)
(52) CPC특허분류
G06N 20/00 (2019.01)
G06F 3/061 (2013.01)
(21) 출원번호 10-2019-0156781
(22) 출원일자 2019년11월29일
심사청구일자 2019년11월29일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
전명재
울산광역시 울주군 언양읍 유니스트길 50, 401동 202호(교수아파트)
임강묵
전라북도 전주시 덕진구 오공로 71, 111동 904호(중동, 호반베르디움)
(74) 대리인
특허법인 무한

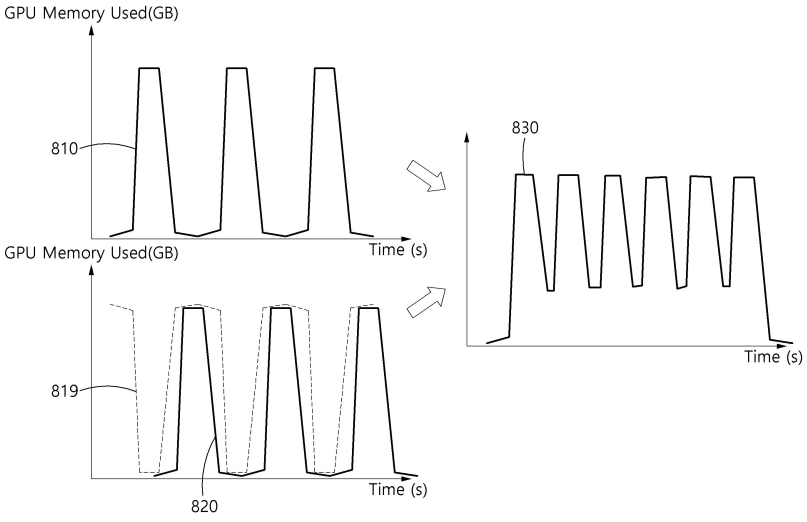
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **메모리 사용 프로필에 기초하여 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법 및 장치**

(57) 요 약

일 실시예에 따른 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법을 수행하는 컴퓨팅 장치는 기계 학습 모델의 트레이닝 작업에 대한 GPU(Graphic Processing Unit) 메모리의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득하고, 상기 기계 학습 모델에 대한 트레이닝 작업의 순전파(forward propagation) 동안 메모리 요청(memory request)에 응답하여, 상기 GPU 메모리로부터, 메모리 블록을 할당하며, 상기 순전파 이후 역전파(backpropagation)를 진행하는 동안, 상기 메모리 사용 프로필에 기초하여, 상기 할당됐던 메모리 블록의 적어도 일부를 해제하고, 다른 모델(different model)의 트레이닝 작업에서 메모리 요청에 응답하여, 상기 GPU 메모리 중 상기 역전파 동안 해제된 메모리 블록을 포함하는 잔여 메모리로부터, 메모리 블록을 할당할 수 있다.

대 표 도



(52) CPC특허분류
G06F 3/0631 (2013.01)
G06F 3/064 (2013.01)

명 세 서

청구범위

청구항 1

컴퓨팅 장치에 의해 수행되는 메모리 할당 방법에 있어서,

기계 학습 모델의 트레이닝 작업에 대한 GPU(Graphic Processing Unit) 메모리의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득하는 단계;

상기 기계 학습 모델에 대한 트레이닝 작업의 순전파(forward propagation) 동안 메모리 요청(memory request)에 응답하여, 상기 GPU 메모리로부터, 메모리 블록을 할당하는 단계;

상기 순전파 이후 역전파(backpropagation)를 진행하는 동안, 상기 메모리 사용 프로필에 기초하여, 상기 할당됐던 메모리 블록의 적어도 일부를 해제하는 단계; 및

다른 모델(different model)의 트레이닝 작업에서 메모리 요청에 응답하여, 상기 GPU 메모리 중 상기 역전파 동안 해제된 메모리 블록을 포함하는 잔여 메모리로부터, 메모리 블록을 할당하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 2

제1항에 있어서,

상기 해제하는 단계는,

현재 반복(iteration)의 역전파에서 사용 종료된(deprecated) 대상 메모리 블록과 관련된 메모리 할당이 다음 반복의 순전파에서 요청되는 다음 요청 시점을 상기 메모리 사용 프로필에 기초하여 예측하는 단계; 및

상기 대상 메모리 블록의 사용 종료 시점 및 상기 다음 요청 시점 간의 시간 차이에 기초하여 상기 대상 메모리 블록의 해제 여부를 결정하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 3

제2항에 있어서,

상기 예측하는 단계는,

상기 대상 메모리 블록과 동일한 크기에 대한 메모리 할당이 상기 다음 반복의 순전파에서 요청되는 시점을 상기 다음 요청 시점으로서 예측하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 4

제2항에 있어서,

상기 예측하는 단계는,

상기 기계 학습 모델에서 상기 대상 메모리 블록이 할당됐던 레이어와 동일한 레벨의 레이어에 대한 메모리 할당이 상기 다음 반복의 순전파에서 요청되는 시점을 상기 다음 요청 시점으로서 예측하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 5

제2항에 있어서,

상기 예측하는 단계는,

상기 대상 메모리 블록에 대한 메모리 할당이 상기 다음 반복의 순전파에서 요청되는 시점을 상기 다음 요청 시점으로서 예측하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 6

제2항에 있어서,

상기 대상 메모리 블록의 해제 여부를 결정하는 단계는,

상기 시간 차이가 미리 정한 임계 시간을 초과하는 경우에 응답하여 상기 대상 메모리 블록을 해제하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 7

제2항에 있어서,

상기 대상 메모리 블록의 해제 여부를 결정하는 단계는,

상기 시간 차이가 미리 정한 임계 시간 이하인 경우에 응답하여 상기 대상 메모리 블록에 대한 해제를 스킵하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 8

제1항에 있어서,

상기 다른 모델의 트레이닝 작업에서 할당됐던 메모리 블록의 해제에 의해 가용한(available) 잔여 메모리의 크기가, 상기 메모리 사용 프로필에 따라 상기 기계 학습 모델에 대한 트레이닝 작업의 순전파에서 요구되는 크기를 초과할 때까지, 상기 기계 학습 모델에 대한 다음 트레이닝 반복(next training iteration)을 지연시키는 단계

를 더 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 9

제1항에 있어서,

상기 잔여 메모리로부터, 메모리 블록을 할당하는 단계는,

상기 기계 학습 모델의 트레이닝 작업에서 한 반복 및 다음 반복까지 하나 이상의 다른 트레이닝 작업을 수행하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 10

제1항에 있어서,

상기 메모리 사용 프로필을 획득하는 단계는,

상기 기계 학습 모델의 트레이닝 작업에 대한 최초 반복(initial iteration) 동안 순전파에서 점유되는 메모리 사용량 및 역전파에서 점유되는 메모리 사용량을 모니터링함으로써 상기 메모리 사용 프로필을 생성하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 11

제1항에 있어서,

상기 잔여 메모리로부터, 메모리 블록을 할당하는 단계는,

상기 기계 학습 모델의 트레이닝 작업에 할당된 메모리 크기 및 상기 다른 모델의 트레이닝 작업에 할당된 메모리 크기의 합산이 상기 GPU 메모리의 크기보다 작게, 상기 다른 모델의 트레이닝 작업을 위한 메모리 블록을 할당하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 12

제1항에 있어서,

상기 잔여 메모리로부터, 메모리 블록을 할당하는 단계는,

상기 기계 학습 모델의 역전파가 완전히 완료되기 전에 상기 다른 모델에 대한 트레이닝 작업의 순전파를 위한 메모리 블록의 할당을 개시하는 단계

를 포함하는 기계 학습 모델의 트레이닝을 위한 메모리 할당 방법.

청구항 13

제1항 내지 제12항 중 어느 한 항의 방법을 수행하기 위한 명령어를 포함하는 하나 이상의 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능 기록 매체.

청구항 14

컴퓨팅 장치에 있어서,

기계 학습 모델 및 다른 모델을 저장하는 저장소;

상기 기계 학습 모델 및 상기 다른 모델의 트레이닝에 할당되는 GPU 메모리를 포함하는 GPU(Graphic Processing Unit); 및

상기 기계 학습 모델의 트레이닝 작업에 대한 GPU(Graphic Processing Unit) 메모리의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득하고, 상기 기계 학습 모델에 대한 트레이닝 작업의 순전파 동안 메모리 요청에 응답하여, 상기 GPU 메모리로부터, 메모리 블록을 할당하며, 상기 순전파 이후 역전파를 진행하는 동안, 상기 메모리 사용 프로필에 기초하여, 상기 할당됐던 메모리 블록의 적어도 일부를 해제하고, 상기 다른 모델의 트레이닝 작업에서 메모리 요청에 응답하여, 상기 GPU 메모리 중 상기 역전파 동안 해제된 메모리 블록을 포함하는

잔여 메모리로부터, 메모리 블록을 할당하는 CPU(Central Processing Unit)

을 포함하는 컴퓨팅 장치.

발명의 설명

기 술 분 야

[0001] 이하, 기계 학습 모델의 트레이닝을 위한 메모리 할당 기술이 제공된다.

배 경 기 술

[0002] 기존 메모리 관리 기법은 단일 학습 작업을 대상으로 설계되었다. 기존 메모리 관리 기법은 메모리를 공유하는 상황 및 환경을 고려하지 않는다. 따라서, 기존 메모리 관리 기법은 메모리를 최대한 미리 보유하고, 필요에 따라 제공하는 식으로 메모리를 관리해왔다.

[0003] 인공지능의 트레이닝 작업은 주기적인 메모리 사용 패턴을 보이는데, 예를 들어, 주기적으로 최소 메모리 사용량 및 최대 메모리 사용량이 반복될 수 있다. 최소 메모리 사용량 대비 최대 메모리 사용량은 약 10배 이상의 차이를 나타낼 수도 있다. 다만, 기존 메모리 관리 기법은 최대 메모리 사용량을 기준으로 메모리를 보유하고, 보유된 메모리 내에서 트레이닝을 위한 메모리를 제공했다.

[0004] 따라서 GPU같은 메모리가 제약적인 하드웨어에서 다수의 인공지능에 대한 트레이닝 작업을 수행하기 위해서는 메모리를 보다 효율적으로 관리할 필요가 있다.

발명의 내용

해결하려는 과제

[0005] 일 실시예에 따른 컴퓨팅 장치는 메모리 사용 프로필에 기초하여 복수의 트레이닝 작업들을 위한 메모리를 할당할 수 있다.

[0006] 일 실시예에 따른 컴퓨팅 장치는, 단일 GPU 메모리 내에서 복수의 트레이닝 작업들을 위한 메모리를 할당할 수 있다.

[0007] 일 실시예에 따른 컴퓨팅 장치는 각 트레이닝 작업의 순전파에 소요되는 시간 및 메모리 할당 타이밍, 및 역전파에서 소요되는 시간 및 메모리 반환 타이밍을 이용하여 복수의 트레이닝 작업들을 병렬적으로 수행할 수 있다.

[0008] 일 실시예에 따른 컴퓨팅 장치는 메모리 사용 프로필에 기초하여 순전파에서 할당됐던 메모리 블록들을 역전파에서 해제할 지 여부를 결정할 수 있다.

[0009] 일 실시예에 따른 컴퓨팅 장치는 각 작업의 주기적인 메모리 사용패턴을 검출하고, 이를 기반으로 필요한 만큼 메모리를 할당할 수 있다.

과제의 해결 수단

[0010] 일 실시예에 따른 메모리 할당 방법은, 기계 학습 모델의 트레이닝 작업에 대한 GPU(Graphic Processing Unit) 메모리의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득하는 단계; 상기 기계 학습 모델에 대한 트레이닝 작업의 순전파(forward propagation) 동안 메모리 요청(memory request)에 응답하여, 상기 GPU 메모리로부터, 메모리 블록을 할당하는 단계; 상기 순전파 이후 역전파(backpropagation)를 진행하는 동안, 상기 메모리 사용 프로필에 기초하여, 상기 할당됐던 메모리 블록의 적어도 일부를 해제하는 단계; 및 다른 모델(different model)의 트레이닝 작업에서 메모리 요청에 응답하여, 상기 GPU 메모리 중 상기 역전파 동안 해제된 메모리 블록을 포함하는 잔여 메모리로부터, 메모리 블록을 할당하는 단계를 포함할 수 있다.

[0011] 상기 해제하는 단계는, 현재 반복(iteration)의 역전파에서 사용 종료된(deprecated) 대상 메모리 블록과 관련된 메모리 할당이 다음 반복의 순전파에서 요청되는 다음 요청 시점을 상기 메모리 사용 프로필에 기초하여 예측하는 단계; 및 상기 대상 메모리 블록의 사용 종료 시점 및 상기 다음 요청 시점 간의 시간 차이에 기초하여

상기 대상 메모리 블록의 해제 여부를 결정하는 단계를 포함할 수 있다.

[0012] 상기 예측하는 단계는, 상기 대상 메모리 블록과 동일한 크기에 대한 메모리 할당이 상기 다음 반복의 순전파에서 요청되는 시점을 상기 다음 요청 시점으로서 예측하는 단계를 포함할 수 있다.

[0013] 상기 예측하는 단계는, 상기 기계 학습 모델에서 상기 대상 메모리 블록이 할당됐던 레이어와 동일한 레벨의 레이어에 대한 메모리 할당이 상기 다음 반복의 순전파에서 요청되는 시점을 상기 다음 요청 시점으로서 예측하는 단계를 포함할 수 있다.

[0014] 상기 예측하는 단계는, 상기 대상 메모리 블록에 대한 메모리 할당이 상기 다음 반복의 순전파에서 요청되는 시점을 상기 다음 요청 시점으로서 예측하는 단계를 포함할 수 있다.

[0015] 상기 대상 메모리 블록의 해제 여부를 결정하는 단계는, 상기 시간 차이가 미리 정한 임계 시간을 초과하는 경우에 응답하여 상기 대상 메모리 블록을 해제하는 단계를 포함할 수 있다.

[0016] 상기 대상 메모리 블록의 해제 여부를 결정하는 단계는, 상기 시간 차이가 미리 정한 임계 시간 이하인 경우에 응답하여 상기 대상 메모리 블록에 대한 해제를 스킵하는 단계를 포함할 수 있다.

[0017] 메모리 할당 방법은 상기 다른 모델의 트레이닝 작업에서 할당됐던 메모리 블록의 해제에 의해 가용한(available) 잔여 메모리의 크기가, 상기 메모리 사용 프로필에 따라 상기 기계 학습 모델에 대한 트레이닝 작업의 순전파에서 요구되는 크기를 초과할 때까지, 상기 기계 학습 모델에 대한 다음 트레이닝 반복(next training iteration)을 지연시키는 단계를 더 포함할 수 있다.

[0018] 상기 잔여 메모리로부터, 메모리 블록을 할당하는 단계는, 상기 기계 학습 모델의 트레이닝 작업에서 한 반복 및 다음 반복까지 하나 이상의 다른 트레이닝 작업을 수행하는 단계를 포함할 수 있다.

[0019] 상기 메모리 사용 프로필을 획득하는 단계는, 상기 기계 학습 모델의 트레이닝 작업에 대한 최초 반복(initial iteration) 동안 순전파에서 점유되는 메모리 사용량 및 역전파에서 점유되는 메모리 사용량을 모니터링함으로써 상기 메모리 사용 프로필을 생성하는 단계를 포함할 수 있다.

[0020] 상기 잔여 메모리로부터, 메모리 블록을 할당하는 단계는, 상기 기계 학습 모델의 트레이닝 작업에 할당된 메모리 크기 및 상기 다른 모델의 트레이닝 작업에 할당된 메모리 크기의 합산이 상기 GPU 메모리의 크기보다 작게, 상기 다른 모델의 트레이닝 작업을 위한 메모리 블록을 할당하는 단계를 포함할 수 있다.

[0021] 상기 잔여 메모리로부터, 메모리 블록을 할당하는 단계는, 상기 기계 학습 모델의 역전파가 완전히 완료되기 전에 상기 다른 모델에 대한 트레이닝 작업의 순전파를 위한 메모리 블록의 할당을 개시하는 단계를 포함할 수 있다.

[0022] 일 실시예에 따른 컴퓨팅 장치는, 기계 학습 모델 및 다른 모델을 저장하는 저장소; 상기 기계 학습 모델 및 상기 다른 모델의 트레이닝에 할당되는 GPU 메모리를 포함하는 GPU(Graphic Processing Unit); 및 상기 기계 학습 모델의 트레이닝 작업에 대한 GPU(Graphic Processing Unit) 메모리의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득하고, 상기 기계 학습 모델에 대한 트레이닝 작업의 순전파 동안 메모리 요청에 응답하여, 상기 GPU 메모리로부터, 메모리 블록을 할당하며, 상기 순전파 이후 역전파를 진행하는 동안, 상기 메모리 사용 프로필에 기초하여, 상기 할당됐던 메모리 블록의 적어도 일부를 해제하고, 상기 다른 모델의 트레이닝 작업에서 메모리 요청에 응답하여, 상기 GPU 메모리 중 상기 역전파 동안 해제된 메모리 블록을 포함하는 잔여 메모리로부터, 메모리 블록을 할당하는 CPU(Central Processing Unit)를 포함할 수 있다.

발명의 효과

[0023] 일 실시예에 따른 컴퓨팅 장치는 메모리 사용 프로필에 기초하여 복수의 트레이닝 작업들을 위한 메모리를 할당하므로, 복수의 트레이닝 작업들을 병렬적으로 수행하는 동안 요구되는 최대 메모리 요구량을 최소화할 수 있다.

[0024] 일 실시예에 따른 컴퓨팅 장치는, 단일 GPU 메모리 내에서 복수의 트레이닝 작업들을 위한 메모리를 할당하므로, DRAM(dynamical random access memory) 사용 없이, 트레이닝 작업의 성능(performance)을 최대화할 수 있다.

[0025] 일 실시예에 따른 컴퓨팅 장치는 각 트레이닝 작업의 순전파에 소요되는 시간 및 메모리 할당 타이밍, 및 역전파에서 소요되는 시간 및 메모리 반환 타이밍을 이용하여 복수의 트레이닝 작업들을 병렬적으로 수행하는데 소

요되는 시간을 최소화할 수 있다.

[0026] 일 실시예에 따른 컴퓨팅 장치는 메모리 사용 프로필에 기초하여 순전파에서 할당됐던 메모리 블록들을 역전파에서 해제할 지 여부를 결정함으로써, 메모리 해제 및 할당에 소요되는 시간을 최소화할 수 있다.

[0027] 일 실시예에 따른 컴퓨팅 장치는 전체 GPU 메모리를 복수의 트레이닝 작업들에게 다양한 환경(예를 들어, 단일 머신 및 클라우드)에서 효율적으로 공유할 수 있다.

도면의 간단한 설명

[0028] 도 1은 기계 학습 모델의 예시로서 뉴럴 네트워크 모델을 설명한다.

도 2는 도 1에 도시된 기계 학습 모델의 메모리 사용 패턴을 설명한다.

도 3a 및 도 3b는 메모리 사용 프로필의 예시를 설명한다.

도 4는 종래의 메모리 할당 패턴을 설명한다.

도 5는 일 실시예에 따른 메모리 할당 방법을 설명하는 흐름도이다.

도 6 및 도 7은 일 실시예에 따른 트레이닝 작업(training task)의 한 반복(iteration) 동안의 메모리 할당 및 해제 동작을 설명한다.

도 8은 일 실시예에 따른 메모리 사용 프로필에 기초하여 복수의 트레이닝 작업들을 교대로 수행하는 예시를 설명한다.

도 9는 일 실시예에 따른 메모리 사용 프로필에 기초한 메모리 해제 스킵의 예시를 설명한다.

도 10은 일 실시예에 따른 메모리 사용 프로필에 기초하여 한 트레이닝 작업의 한 반복이 완료된 이후에 다른 트레이닝 작업의 반복을 개시하는 예시를 설명한다.

도 11은 일 실시예에 따른 메모리 사용 프로필에 기초하여 한 트레이닝 작업의 반복들 사이에 다른 트레이닝 작업의 복수의 반복들을 수행하는 예시를 설명한다.

도 12는 일 실시예에 따른 컴퓨팅 장치의 구성을 설명하는 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0029] 이하에서, 첨부된 도면을 참조하여 실시예들을 상세하게 설명한다. 그러나, 실시예들에는 다양한 변경이 가해질 수 있어서 특허출원의 권리 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 실시예들에 대한 모든 변경, 균등물 내지 대체물이 권리 범위에 포함되는 것으로 이해되어야 한다.

[0030] 실시예에서 사용한 용어는 단지 설명을 목적으로 사용된 것으로, 한정하려는 의도로 해석되어서는 안된다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0031] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0032] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0033] 도 1은 기계 학습 모델의 예시로서 뉴럴 네트워크 모델을 설명한다.

[0034] 기계 학습 모델은 특정 입력에 대해 특정 출력이 출력되도록 학습된 모델로서, 특정 입력에 대해 미리 트레이닝된 파라미터에 기초하여 임의의 출력이 생성되는 블랙박스 함수로 표현될 수 있다. 도 1에서는 기계 학습 모델

로서 뉴럴 네트워크 모델(neural network model)(100)을 설명한다.

[0035] 뉴럴 네트워크 모델(100)은 딥 뉴럴 네트워크(DNN, deep neural network)의 예시에 해당할 수 있다. DNN은 완전 연결 네트워크(fully connected network), 딥 컨볼루셔널 네트워크(deep convolutional network) 및 리커런트 뉴럴 네트워크(recurrent neural network) 등을 포함할 수 있다. 뉴럴 네트워크 모델(100)은 딥 러닝에 기반하여 비선형적 관계에 있는 입력 데이터 및 출력 데이터를 서로 매핑함으로써 객체 분류, 객체 인식, 음성 인식 및 이미지 인식 등을 수행할 수 있다. 딥 러닝은 빅 데이터 세트로부터 이미지 또는 음성 인식과 같은 문제를 해결하기 위한 기계 학습 기법으로 지도식(supervised) 또는 비지도식(unsupervised) 학습을 통해 입력 데이터 및 출력 데이터를 서로 매핑할 수 있다.

[0036] 도 1을 참조하면, 뉴럴 네트워크 모델(100)은 입력 레이어(110), 히든 레이어(120) 및 출력 레이어(130)를 포함한다. 입력 레이어(110), 히든 레이어(120) 및 출력 레이어(130)는 각각 복수의 인공 노드들을 포함한다.

[0037] 도 1에는 설명의 편의를 위해 히든 레이어(120)가 3개의 레이어를 포함하는 것으로 도시되었으나, 히든 레이어(120)는 다양한 수의 레이어들을 포함할 수 있다. 또한, 도 1에서 뉴럴 네트워크(100)는 입력 데이터를 수신하기 위한 별도의 입력 레이어를 포함하는 것으로 도시되었으나, 입력 데이터가 히든 레이어(120)에 직접 입력될 수 있다. 뉴럴 네트워크(100)에서 출력 레이어(130)를 제외한 레이어들의 인공 노드들은 출력 신호를 전송하기 위한 링크들을 통해 다음 레이어의 인공 노드들과 연결될 수 있다. 링크들의 수는 다음 레이어에 포함된 인공 노드들의 수에 대응할 수 있다.

[0038] 히든 레이어(120)에 포함된 각각의 인공 노드에는 이전 레이어에 포함된 인공 노드들의 가중된 입력들(weighted inputs)에 관한 활성 함수(activation function)의 출력이 입력될 수 있다. 가중된 입력은 이전 레이어에 포함된 인공 노드들의 입력에 가중치(weight)가 곱해진 것이다. 가중치는 뉴럴 네트워크(100)의 파라미터로 지칭될 수 있다. 활성 함수는 시그모이드(sigmoid), 하이퍼볼릭 탄젠트(hyperbolic tangent; tanh) 및 렐루(rectified linear unit; ReLU)를 포함할 수 있고, 활성 함수에 의해 뉴럴 네트워크(100)에 비선형성이 형성될 수 있다. 출력 레이어(130)에 포함된 각각의 인공 노드에는 이전 레이어에 포함된 인공 노드들의 가중된 입력들이 입력될 수 있다.

[0039] 또한, 히든 레이어(120)는 예시적으로, 컨볼루션 레이어(convolution layer), 풀링 레이어(pooling layer), 및 완전연결 레이어(fully connected layer) 등을 포함할 수 있다. 컨볼루션 레이어는 이전 레이어에서 출력된 정보에 대해 미리 정한 크기의 커널 필터(kernel filter)를 이용하여 컨볼루션 필터링이 적용된 결과를 출력하는 레이어를 나타낼 수 있다. 풀링 레이어는, 이전 레이어의 정보로부터 대표 값을 풀링(pooling)한 결과를 출력하는 레이어를 나타낼 수 있다. 정규화 레이어는 이전 레이어의 값이 정규화된 결과를 출력하는 레이어를 나타낼 수 있다. 완전연결 레이어는 이전 레이어의 모든 노드들과 연결되는 레이어를 나타낼 수 있다. 참고로, 컨볼루션 레이어의 출력은 특징 데이터로서, 컨볼루션 레이어에 입력된 정보가 추상화된 데이터를 나타낼 수 있다. 특징 데이터는 벡터 형태의 특징 벡터 또는 맵 형태의 특징 맵일 수 있다.

[0040] 뉴럴 네트워크 모델(100)의 폭과 깊이가 충분히 크면 임의의 함수를 구현할 수 있을 만큼의 용량(capacity)을 가질 수 있다. 뉴럴 네트워크 모델(100)이 적절한 트레이닝 과정을 통해 충분히 많은 트레이닝 데이터를 학습하면, 최적의 인식 성능을 달성할 수 있다.

[0041] 본 명세서에서는 기계 학습 모델의 트레이닝을 위해 메모리를 효율적으로 할당 및 해제하는 동작을 설명한다. 이하, 예시적인 기계 학습 모델로서 뉴럴 네트워크 모델(100)의 트레이닝 동안의 메모리 할당 및 해제를 설명하지만, 이로 한정하는 것은 아니고, 다른 유형의 기계 학습 모델에도 적용될 수 있다.

[0042] 도 2는 도 1에 도시된 기계 학습 모델의 메모리 사용 패턴을 설명한다.

[0043] 일실시예에 따르면 컴퓨팅 장치는 GPU(Graphic Processing Unit)를 이용하여 뉴럴 네트워크 모델(100)의 트레이닝을 위해 요구되는 연산을 수행할 수 있다. 컴퓨팅 장치는 트레이닝 작업(training task)을 위해 GPU의 GPU 메모리를 할당 및 사용할 수 있다.

[0044] 예를 들어, 기계 학습 모델의 트레이닝 작업은 주기적인 복수의 반복들(iteration)을 포함할 수 있고, 각 반복은 입력 데이터를 기계 학습 모델의 입력 레이어로부터 출력 레이어까지 순방향으로 전파(이하, 순전파)하는 연산 및 순전파에서 산출된 손실을 기계 학습 모델의 출력 레이어로부터 입력 레이어까지 역방향으로 전파(이하, 역전파)하는 연산으로 구성될 수 있다. 이때, 각 반복은 입력 데이터를 배치(batch) 단위로 한번에 순전파 및 역전파시키는 과정을 나타낼 수 있다.

[0045] 도 2를 참조하면, 순전파 동안 각 레이어에 대응하는 연산을 처리하기 위해, 여러 트레이닝 관련 정보(예를 들어, 입력 데이터, 특징 맵, 출력 데이터, 손실 등)가 저장될 필요가 있다. 예를 들어, 뉴럴 네트워크 모델(100)의 입력 레이어(110)로 입력 데이터를 전파하기 위해 입력 데이터를 위한 크기(size)의 메모리가 요구될 수 있다. 컴퓨팅 장치는 입력 데이터를 위한 크기의 메모리 블록을 GPU 메모리로부터 할당할 수 있다. 입력 레이어(110) 이후 히든 레이어(120)에 대응하는 연산(예를 들어, 특징 맵 추출)을 처리하기 위해, 특징 맵을 위한 크기의 메모리가 추가로 더 요구될 수 있다. 컴퓨팅 장치는 특징 맵을 위한 크기의 메모리 블록을 GPU 메모리로부터 추가로 더 할당할 수 있다. 컴퓨팅 장치는 순전파 동안 산출된 각 레이어의 출력들을 역전파에서 사용하기 위해, 역전파를 개시하기 전까지 해당 출력들에 대해 할당된 메모리 블록을 할당 및 사용할 수 있다. 역전파를 개시한 이후 역전파 동안 각 레이어의 순전파 출력들이 상위 레이어부터 하위 레이어의 순서로 사용될 수 있다. 순전파 출력들에 할당됐던 메모리 블록은, 역전파 동안 각 레이어의 연산이 종료된 이후 해제될 수 있다. 따라서 도 2에 도시된 바와 같이, 메모리 사용 패턴(200)은 순전파 동안에는 GPU 메모리 사용량(210)이 증가하고, 역전파 동안에는 GPU 메모리 사용량(210)이 감소하는 웨이브 패턴을 나타낼 수 있다.

[0046] 참고로, 본 명세서에서 메모리 블록은 GPU 메모리가 분할되어 사용되는 블록을 나타낼 수 있으나, 이로 한정하는 것은 아니다. 예를 들어, 메모리 블록은 하나 이상의 서브 블록(sub-block)을 포함할 수도 있다.

[0047] 도 3a 및 도 3b는 메모리 사용 프로필의 예시를 설명한다.

[0048] 도 3a는 ResNet50/Imagenet 모델에서의 메모리 사용 패턴(310)을 나타내고, 도 3b는 GNMT/WMT'14 En-De 모델에서의 메모리 사용 패턴(320)을 나타낼 수 있다. 메모리 사용 패턴들(310, 320)의 세로 축은 사용된 GPU 메모리 크기로서 GB (Giga Byte) 단위이고, 가로 축은 시간 축으로서 초 단위일 수 있다. 기계 학습 모델의 유형에 따라 메모리 사용 패턴의 형태는 다를 수 있으나, 전반적으로 트레이닝 작업의 각 반복에서 순전파 동안 메모리 사용량이 증가하고 역전파 동안 메모리 사용량이 감소할 수 있다.

[0049] 도 4는 종래의 메모리 할당 패턴을 설명한다.

[0050] 종래에는 단일 트레이닝 작업의 메모리를 관리하는 메모리 관리자 어플리케이션이 상술한 메모리 사용 패턴에 따른 메모리 할당 및 해제를 처리했다. 메모리 관리자 어플리케이션은 스왑 아웃 및 스왑인을 처리할 수도 있다. 이러한 메모리 관리자 어플리케이션은 GPU를 혼자 점유해서 사용하는 단일 트레이닝을 위해 설계된다. 따라서 트레이닝 작업 동안 메모리 사용량이 감소하더라도, 사용종료된(deprecated) 메모리가 해당 트레이닝 작업을 위한 메모리 관리자 어플리케이션에게 반환(return)될 뿐이었다. 메모리 관리자 어플리케이션은 사용종료된 메모리를 실제 GPU 메모리로 반환하지 않았다.

[0051] 도 4에 도시된 메모리 할당 패턴(400)은, 메모리 사용 패턴(410)과 달리, 항상 최대 크기(420)(예를 들어, 약 23GB)의 GPU 메모리가 할당되어 사용되는 패턴을 나타낸다. 메모리 사용 패턴(410)에 따르면 최대 메모리 사용량인 23GB보다 적은 메모리가 사용되는 시간이 존재한다. 다만, 실제 메모리 할당 패턴(400)에서는 트레이닝 동안 항상 최대 크기의 메모리가 할당되므로 GPU 메모리의 활용성이 저하된다.

[0052] 예를 들어, 종래의 메모리 관리자 어플리케이션은 개별 트레이닝 작업마다 최대 크기(420)의 메모리를 항상 할당한다. 이러한 종래의 메모리 관리자 어플리케이션은, 2개의 트레이닝 작업들의 수행을 위해, 약 46 GB의 메모리를 요구한다. 다만, 현재 판매 중인 제품군의 최대 GPU 메모리 크기는 32GB이므로, 종래의 메모리 관리자 어플리케이션은 단일 GPU를 이용한 복수의 트레이닝 작업들의 병렬 수행을 지원할 수 없다. 특히, 단일 GPU의 최대 지원 크기를 초과하는 메모리가 요구되는 경우, 종래의 메모리 관리자 어플리케이션은 나머지 초과된 크기의 메모리를 DRAM(dynamic random access memory)을 통해 할당했다. DRAM과 GPU 간의 데이터 통신 대역폭은 작은 편이므로, DRAM 및 GPU 간의 데이터 통신에 소요되는 시간으로 인해 종래의 메모리 관리자 어플리케이션을 이용한 트레이닝 시간이 증가되었다.

[0053] 상술한 GPU 메모리의 활용성 문제를 해결하기 위해, 일 실시예에 따른 컴퓨팅 장치는 여러 트레이닝 작업들을 위해 GPU 메모리를 효율적으로 공유할 수 있다. 컴퓨팅 장치에 의한 메모리 할당 및 해제는 아래에서 설명한다.

[0054] 도 5는 일 실시예에 따른 메모리 할당 방법을 설명하는 흐름도이다.

[0055] 일 실시예에 따르면 컴퓨팅 장치는 복수의 트레이닝 작업들의 메모리 요청 및 할당을 처리할 수 있는 메모리 할당 어플리케이션(memory allocation application)을 실행할 수 있다.

[0056] 우선, 단계(510)에서 컴퓨팅 장치는 기계 학습 모델의 트레이닝 작업에 대한 GPU(Graphic Processing Unit) 메

모리의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득할 수 있다. 트레이닝 작업은 기계 학습 모델을 트레이닝시키는 작업으로서, 예를 들어, 손실 함수에 따른 손실 값이 목표 값에 도달할 때까지 기계 학습 모델의 파라미터(예를 들어, 연결 가중치 등)를 반복적으로 업데이트하는 작업을 나타낼 수 있다. 트레이닝 작업은 순전파 및 역전파를 통해 파라미터를 업데이트하는 반복들(iterations)을 포함할 수 있다.

[0057] 메모리 사용 프로필은 임의의 기계 학습 모델의 트레이닝 작업에서 한 반복(iteration) 동안 순전파에서 요구되는 메모리 사용량 및 역전파에서 요구되는 메모리 사용량을 포함할 수 있다. 메모리 사용 프로필에서 순전파 동안 요구되는 메모리 사용량은 시간 경과에 따라 순전파가 완료될 때까지 증가하고, 역전파 동안 요구되는 메모리 사용량은 시간 경과에 따라 역전파가 완료될 때까지 감소할 수 있다. 또한, 임의의 단일 기계 학습 모델이 복수의 GPU를 통해 분산 트레이닝되는 경우, 메모리 사용 프로필은 역전파 이후 각 GPU에서 업데이트된 파라미터를 동기화하는데 요구되는 동기화 시간도 포함할 수 있다.

[0058] 그리고 단계(520)에서 컴퓨팅 장치는 기계 학습 모델에 대한 트레이닝 작업의 순전파(forward propagation) 동안 메모리 요청(memory request)에 응답하여, GPU 메모리로부터, 메모리 블록을 할당할 수 있다. 예를 들어, 컴퓨팅 장치는 순전파 동안 입력 데이터의 저장을 위한 메모리 블록, 각 레이어에서 특징 데이터의 산출에 사용되는 메모리 블록, 및 각 레이어에서 산출된 특징 데이터를 저장하기 위한 메모리 블록 등을 할당할 수 있다. 컴퓨팅 장치는 기계 학습 모델의 하위 레이어로부터 상위 레이어로(예를 들어, 입력 레이어로부터 출력 레이어까지) 순전파가 진행되는 동안, 순차적인 메모리 요청에 따라 할당되는 메모리 블록들을 추가할 수 있다.

[0059] 이어서 단계(530)에서 컴퓨팅 장치는 순전파 이후 역전파(backpropagation)를 진행하는 동안, 메모리 사용 프로필에 기초하여, 할당됐던 메모리 블록의 적어도 일부를 해제할 수 있다. 예를 들어, 컴퓨팅 장치는 역전파 동안 메모리 블록을 역전파 순서에 따라 순차적으로 해제할 수 있다. 컴퓨팅 장치는 역전파에서 사용종료된(deprecated) 순서에 따라 순전파에서 할당됐던 메모리 블록을 해제할 수 있다. 컴퓨팅 장치는 해제된 메모리 블록을 GPU 메모리로 반환할 수 있다. 컴퓨팅 장치는 역전파의 진행에 따라 메모리 블록을 해제하거나 메모리 사용 프로필에 기초하여 일부 메모리 블록을 유지할 수도 있다. 메모리 블록의 할당 및 해제는 하기 도 6 및 도 7에서 설명한다.

[0060] 그리고 단계(540)에서 컴퓨팅 장치는 다른 모델(different model)의 트레이닝 작업에서 메모리 요청에 응답하여, GPU 메모리 중 역전파 동안 해제된 메모리 블록을 포함하는 잔여 메모리로부터, 메모리 블록을 할당할 수 있다. 컴퓨팅 장치는 기계 학습 모델의 한 트레이닝 반복이 완료된 후 다른 모델의 순전파를 위한 메모리 할당을 개시할 수 있다. 다만, 이로 한정하는 것은 아니고, 컴퓨팅 장치는 기계 학습 모델의 한 트레이닝 반복이 완료되기 전이더라도, 해당 트레이닝 반복의 역전파에서 일부 메모리 블록이 해제된 경우 가용한 GPU 메모리 내에서 다른 모델의 순전파를 위한 메모리 할당을 개시할 수 있다. 참고로, 다른 모델은, 예시적으로 기계 학습 모델과 다른 구조의 모델일 수 있으나, 이로 한정하는 것은 아니다.

[0061] 따라서 컴퓨팅 장치는 임의의 기계 학습 모델에 대한 트레이닝 작업의 반복들 사이에 다른 모델에 대한 트레이닝 작업의 반복을 수행할 수 있다. 일 실시예에 따르면 상대적으로 적은 크기의 GPU 메모리를 이용하여 복수의 트레이닝 작업들이 수행될 수 있다.

[0062] 도 6 및 도 7은 일 실시예에 따른 트레이닝 작업(training task)의 한 반복(iteration) 동안의 메모리 할당 및 해제 동작을 설명한다.

[0063] 일 실시예에 따르면 컴퓨팅 장치는 기계 학습 모델에 대한 트레이닝 작업의 한 반복 동안 메모리 요청에 따라 메모리 블록을 할당하고, 사용종료된 메모리 블록에 대한 할당을 해제할 수 있다. 도 6은 메모리 할당 및 해제 따른 GPU 사용 그래프로서, 가로축은 사용된 GPU 메모리 크기를 GB (gigabyte) 단위로 나타내고, 세로축은 시간을 초 단위로 나타낼 수 있다.

[0064] 예를 들어, 기계 학습 모델의 최하위 레이어를 통해 데이터를 전파시키기 위해 일정 크기의 메모리 요청이 발생(occur)할 수 있다. 컴퓨팅 장치는 최하위 레이어의 전파를 위해 제1 메모리 블록(621)의 할당(621a)을 수행할 수 있다. 이후, 다음 레이어의 전파를 위해, 컴퓨팅 장치는 제2 메모리 블록(622)의 할당(622a)을 수행할 수 있다. 그 다음 레이어의 전파를 위해 컴퓨팅 장치는 제3 메모리 블록(623)의 할당(623a) 내지 제k-1 메모리 블록(624)의 할당(624a)을 수행할 수 있다. 마지막 레이어의 전파를 위해 컴퓨팅 장치는 제k 메모리 블록(625)의 할당(625a)을 수행할 수 있다. 여기서, k는 2이상의 정수일 수 있다. 순전파 동안, 컴퓨팅 장치는 제1 메모리 블록 내지 제k 메모리 블록을 순차적으로 추가 할당할 수 있다. 따라서 메모리 사용 프로필(610)에서 트레이닝 작업의 각 반복에서 순전파 동안 컴퓨팅 장치에 의해 할당되는 메모리 사용량은 증가될 수 있다.

[0065] 순전파가 완료된 후 컴퓨팅 장치는 손실 함수에 기초하여 손실을 산출할 수 있다. 컴퓨팅 장치는 최상위 레이어로부터 최하위 레이어까지 산출된 손실을 역전파시킬 수 있다. 컴퓨팅 장치는 이전에 순전파에서 할당됐던 제k 메모리 블록(625)을 이용하여 최상위 레이어의 역전파를 진행할 수 있다. 최상위 레이어에서의 역전파가 완료되면 제k 메모리 블록(625)은 사용종료(deprecated)될 수 있다. 현재 반복에서는 제k 메모리 블록(625)이 더 이상 필요없을 수 있다. 따라서 컴퓨팅 장치는 제k 메모리 블록(625)의 할당 해제(625b)(deallocation)를 수행할 수 있다. 할당 해제된 제k 메모리 블록(625)은 GPU 메모리로 반환될 수 있다. 이후, 컴퓨팅 장치는 역전파 순서에 따라 제k-1 메모리 블록(624)의 할당 해제(624b) 내지 제3 메모리 블록(623)의 할당 해제(623b), 제2 메모리 블록(622)의 할당 해제(622b), 및 제1 메모리 블록(621)의 할당 해제(621b)를 순차적으로 수행할 수 있다. 따라서 메모리 사용 프로필(610)에서 트레이닝 작업의 각 반복에서 역전파 동안 컴퓨팅 장치에 의해 할당됐던 메모리 사용량이 감소될 수 있다.

[0066] 도 7은 예시적인 뉴럴 네트워크 모델(700)에서의 메모리 할당 및 해제를 설명한다.

[0067] 일 실시예에 따르면 컴퓨팅 장치는 뉴럴 네트워크 모델(700)에 대한 트레이닝 작업의 각 반복에 있어서, 제1 레이어($Layer_1$)로부터 제n 레이어($Layer_n$)까지 순전파를 순차적으로 수행하고, 이어서, 제n 레이어($Layer_n$)로부터 제1 레이어($Layer_1$)까지 역전파를 순차적으로 수행할 수 있다.

[0068] 예를 들어, 컴퓨팅 장치는 입력 데이터로부터 제1 레이어($Layer_1$)의 순전파(fw_1)를 이용하여 제1 특징 맵(Feature maps 1)을 추출할 수 있다. 컴퓨팅 장치는 추출된 제1 특징 맵(Feature maps 1)을 저장하기 위해 제1 메모리 블록(721)을 할당할 수 있다. 여기서, 제1 레이어($Layer_1$)의 순전파(fw_1)는 제1 가중치 행렬(weight matrix 1)를 이용한 연산으로서, 예를 들어, 컨볼루션 연산일 수 있으나, 이로 한정하는 것은 아니다. 제1 가중치 행렬(weight matrix 1)는 뉴럴 네트워크 모델(700)의 파라미터일 수 있다. 컴퓨팅 장치는 제1 특징 맵(Feature maps 1)으로부터 제2 레이어($Layer_2$)의 순전파(fw_2)를 이용하여 제2 특징 맵(Feature maps 2)을 추출할 수 있다. 컴퓨팅 장치는 추출된 제2 특징 맵(Feature maps 2)을 저장하기 위해 제2 메모리 블록(722)을 할당할 수 있다. 나머지 상위 레이어에 대해서도 컴퓨팅 장치는 유사하게 순전파를 수행할 수 있다. 컴퓨팅 장치는 제n-1 특징 맵을 저장하기 위해 제k-1 메모리 블록(724)을 할당할 수 있다. 컴퓨팅 장치는 제n-1 특징 맵(Feature maps n-1)으로부터 제n 레이어($Layer_n$)의 순전파(fw_n)를 이용하여 출력 데이터(output)을 산출할 수 있다. 컴퓨팅 장치는 출력 데이터(output)을 저장하기 위해 제k 메모리 블록을 할당할 수 있다. 여기서, n, k는 각각 2이상의 정수일 수 있다.

[0069] 컴퓨팅 장치는 산출된 출력 데이터 및 참값 데이터(Ground Truth)로부터 손실 함수(loss function)를 이용하여 트레이닝 오차(Error)를 산출할 수 있다. 컴퓨팅 장치는 트레이닝 오차가 감소하도록 역전파를 통해 파라미터들(예를 들어, 가중치 행렬)을 업데이트할 수 있다.

[0070] 예를 들어, 컴퓨팅 장치는 트레이닝 오차(Error)로부터 제n 레이어($Layer_n$)의 역전파(bw_n)를 이용하여 제n-1 경사 맵(Gradient Maps n-1)을 산출할 수 있다. 여기서, 제n 레이어($Layer_n$)의 역전파(bw_n)에는 제n-1 특징 맵, 제n 가중치 행렬, 및 트레이닝 오차가 사용될 수 있다. 역전파(bw_n)의 결과를 이용하여, 트레이닝 오차가 감소하도록, 제n 가중치 행렬이 업데이트될 수 있다. 현재 반복의 역전파 동안 제n-1 특징 맵은 더 이상 필요 없으므로, 컴퓨팅 장치는, 이전 순전파에서 제n-1 특징 맵을 위해 할당됐던, 제k-1 메모리 블록(724)에 제n-1 경사 맵을 저장할 수 있다. 출력 데이터 및 트레이닝 오차를 위해 사용됐던 제k 메모리 블록(725)은 더 이상 필요 없으므로, 컴퓨팅 장치는 제k 메모리 블록(725)의 할당을 해제할 수 있다. 해제된 제k 메모리 블록(725)은 GPU 메모리(720)로 반환될 수 있다. 역전파 순서대로 레이어들의 경사 맵들이 전파되면서, 각 레이어의 가중치 행렬이 업데이트되고, 해당 레이어의 출력(예를 들어, 특징 맵 및 경사 맵)을 위해 할당됐던 메모리 블록들이 사용종료될 수 있다. 따라서 컴퓨팅 장치는 사용종료된 순서에 따라 메모리 블록들을 해제하고, GPU 메모리(720)로 반환할 수 있다.

[0071] 도 8은 일 실시예에 따른 메모리 사용 프로필에 기초하여 복수의 트레이닝 작업들을 교대로 수행하는 예시를 설명한다.

[0072] 일 실시예에 따르면 컴퓨팅 장치는, 메모리의 공유를 위해, 동시에 병렬적으로 수행되는 트레이닝 작업들의 순전파 및 역전파를 각 메모리 사용 프로필에 기초하여 타임라인에 정렬할 수 있다. 참고로, 도 8에 도시된 그래프들의 세로 축은 GPU 메모리 중 사용된 GPU 메모리 크기를 나타내고, 가로 축은 시간을 나타낼 수 있다.

[0073] 예를 들어, 컴퓨팅 장치는 제1 트레이닝 작업의 역전파 과정 동안 메모리를 반환할 수 있고, 메모리 반환 타이밍에 맞춰 제2 트레이닝 작업의 순전파를 위한 메모리를 할당할 수 있다. 다시 말해, 컴퓨팅 장치는 기계 학습 모델에 대한 제1 트레이닝 작업의 역전파가 완전히 완료되기 전에 다른 모델에 대한 제2 트레이닝 작업의 순전파를 위한 메모리 블록의 할당을 개시할 수 있다. 따라서 컴퓨팅 장치는 제1 트레이닝 작업의 반복에서 역전파 동안 반환된 메모리를 제2 트레이닝 작업의 반복에서 순전파를 위해 즉시 할당할 수 있다.

[0074] 컴퓨팅 장치는 제2 트레이닝 작업의 반복에서 순전파 동안 메모리를 할당하고, 이후 역전파 동안 할당됐던 메모리를 해제할 수 있다. 컴퓨팅 장치는 제2 트레이닝 작업의 역전파에서 해제된 메모리를, 제1 트레이닝 작업의 순전파를 위해 할당할 수 있다. 따라서, 컴퓨팅 장치는 제1 트레이닝 작업의 반복 및 제2 트레이닝 작업의 반복을 위해 교대로 GPU 메모리의 메모리 블록을 할당할 수 있다.

[0075] 여기서, 제2 트레이닝 작업의 메모리 사용 프로필(820)은 잔여 메모리 프로필(819)에 포함될 수 있다. 잔여 메모리 프로필(819)은 GPU 메모리에서 제1 트레이닝 작업을 위해 할당된 메모리를 제외하고 남을 것으로 예측되는 메모리에 관한 프로필로서, GPU 메모리의 전체 크기로부터 제1 트레이닝 작업의 메모리 사용 프로필(810)이 차감된 프로필일 수 있다. 예를 들어, 제2 트레이닝 작업의 메모리 사용 프로필(820)에 따른 최대 메모리 사용량은 잔여 메모리 프로필(819)의 최대 크기보다 작을 수 있다. 더 나아가, 시간축이 조정된 제2 트레이닝 작업의 메모리 사용 프로필(820)은, 전 시간 구간에서 잔여 메모리 프로필(819)보다 작은 메모리 크기를 나타낼 수 있다.

[0076] 일 실시예에 따른 컴퓨팅 장치는 기계 학습 모델의 트레이닝 작업에 할당된 메모리 크기 및 다른 모델의 트레이닝 작업에 할당된 메모리 크기의 합산이 GPU 메모리의 크기보다 작게, 다른 모델의 트레이닝 작업을 위한 메모리 블록을 할당할 수 있다.

[0077] 따라서, 동일 또는 유사한 메모리 사용 패턴의 트레이닝 작업들이 병렬적으로 수행되는 경우, 컴퓨팅 장치는 기존의 46 GB 가 아닌, 약 23GB로 제1 트레이닝 작업 및 제2 트레이닝 작업을 병렬적으로 수행할 수 있다. 도 8에 제1 트레이닝 작업에 대응하는 메모리 사용 프로필(810) 및 제2 트레이닝 작업에 대응하는 메모리 사용 프로필의 합산에 따른 전체 메모리 사용량(830)이 도시된다. 전체 메모리 사용량(830)의 최대 메모리 사용량은, 제1 트레이닝 작업을 위한 메모리 사용 프로필(810)에서의 최대 메모리 크기 및 제2 트레이닝 작업을 위한 메모리 사용 프로필(820)에서의 최대 메모리 크기와 유사할 수 있다.

[0078] 일 실시예에 따른 컴퓨팅 장치는, 단일 GPU 메모리 내에서 복수의 트레이닝 작업들을 위한 메모리를 할당하므로, 트레이닝 작업의 성능(performance)가 최대화될 수 있다.

[0079] 참고로, 컴퓨팅 장치는 기계 학습 모델의 트레이닝 작업에 대한 최초 반복(initial iteration) 동안 순전파에서 점유되는 메모리 사용량 및 역전파에서 점유되는 메모리 사용량을 모니터링함으로써 메모리 사용 프로필(810, 820)을 생성할 수 있다. 다만, 이로 한정하는 것은 아니고, 컴퓨팅 장치는 개별 기계 학습 모델의 트레이닝 작업에 대해 미리 주어지는 메모리 사용 프로필(810, 820)을 획득할 수 있다. 기계 학습 모델의 구조(예를 들어, 레이어에 포함되는 노드의 개수, 레이어의 개수, 및 레이어 별 연산 타입 등)에 따라 대략적인 메모리 사용 프로필(810, 820)이 추정될 수 있다. 컴퓨팅 장치는 사용자로부터 상술한 메모리 사용 프로필(810, 820)을 입력받거나, 기계 학습 모델의 구조로부터 메모리 사용 프로필(810, 820)을 추정할 수 있다.

[0080] 아울러, 도 8에서는 병렬적으로 수행되는 트레이닝 작업들이 미리 선택된 것을 설명하였으나, 이로 한정하는 것은 아니다. 컴퓨팅 장치는 복수의 트레이닝 작업들 중 서로 병렬적으로 수행 가능한 트레이닝 작업들을 선택할 수도 있다. 예를 들어, 컴퓨팅 장치는 개별 트레이닝 작업의 메모리 사용 프로필(810, 820)에 기초하여 해당 트레이닝 작업과 병렬적으로 수행될 수 있는 다른 트레이닝 작업을 검색할 수 있다. 제1 트레이닝 작업의 메모리 사용 프로필(810)에 기초하여, 컴퓨팅 장치는 잔여 메모리 프로필을 산출할 수 있다. 컴퓨팅 장치는 잔여 메모리 프로필에 매칭하는 메모리 사용 프로필을 나타내는 다른 트레이닝 작업을 검색할 수 있다. 예를 들어, 컴퓨팅 장치는 다른 트레이닝 작업에 대응하는 메모리 사용 프로필(820)의 시간축을 조정함으로써, 시간축이 조정된 메모리 사용 프로필(820)이 잔여 메모리 프로필에 포함되는지 여부를 판단할 수 있다. 잔여 메모리 프로필에 포함되는 메모리 사용 프로필이, 잔여 메모리 프로필에 매칭되는 프로필일 수 있다.

[0081] 도 9는 일 실시예에 따른 메모리 사용 프로필에 기초한 메모리 해제 스킵의 예시를 설명한다.

[0082] 일 실시예에 따른 컴퓨팅 장치는 현재 반복(iteration)의 역전파에서 사용 종료된(deprecated) 대상 메모리 블록과 관련된 메모리 할당이 다음 반복의 순전파에서 요청되는 다음 요청 시점을 메모리 사용 프로필에 기초하여 예측할 수 있다. 도 9에 도시된 메모리 사용 프로필(900)의 세로 축은 사용된 GPU 메모리 크기를 GB 단위로 나

타내고, 가로축은 시간을 초 단위로 나타낼 수 있다.

[0083] 예를 들어, 컴퓨팅 장치는 대상 메모리 블록과 동일한 크기에 대한 메모리 할당이 다음 반복의 순전파에서 요청되는 시점을 다음 요청 시점으로서 예측할 수 있다. 다른 예를 들어, 컴퓨팅 장치는 기계 학습 모델에서 대상 메모리 블록이 할당됐던 레이어와 동일한 레벨의 레이어에 대한 메모리 할당이 다음 반복의 순전파에서 요청되는 시점을 다음 요청 시점으로서 예측할 수 있다. 또 다른 예를 들어, 컴퓨팅 장치는 대상 메모리 블록에 대한 메모리 할당이 다음 반복의 순전파에서 요청되는 시점을 다음 요청 시점으로서 예측할 수 있다.

[0084] 컴퓨팅 장치는 대상 메모리 블록의 사용 종료 시점 및 다음 요청 시점 간의 시간 차이에 기초하여 대상 메모리 블록의 해제 여부를 결정할 수 있다. 예를 들어, 컴퓨팅 장치는 시간 차이가 미리 정한 임계 시간을 초과하는 경우에 응답하여 대상 메모리 블록을 해제할 수 있다. 도 9에 도시된 메모리 사용 프로필(900)에서, 컴퓨팅 장치는 제1 대상 블록의 사용 종료 시점(911) 및 다음 요청 시점(912) 간의 시간 차이(919)가 임계 시간을 초과하는 지 여부를 판단할 수 있다. 컴퓨팅 장치는 시간 차이(919)가 임계 시간을 초과하는 경우에 응답하여, 제1 대상 블록을 해제할 수 있다. 제1 대상 블록 또는 제1 대상 블록에 대응하는 크기의 메모리가 다음 반복에서 사용될 때까지 상당한 시간이 소요되므로, 컴퓨팅 장치는 제1 대상 블록을 해제(release)시킴으로써 제1 대상 블록에 대응하는 크기의 잔여 메모리를 다른 작업(예를 들어, 다른 트레이닝 작업의 반복에서의 순전파 동작)을 위해 준비할 수 있다.

[0085] 다른 예를 들어, 컴퓨팅 장치는 시간 차이가 미리 정한 임계 시간 이하인 경우에 응답하여 대상 메모리 블록에 대한 해제를 스킵할 수 있다. 도 9에 도시된 메모리 사용 프로필(900)에서 컴퓨팅 장치는 제2 대상 블록의 사용 종료 시점(921) 및 다음 요청 시점(922) 간의 시간 차이(929)가 임계 시간을 초과하는 지 여부를 판단할 수 있다. 컴퓨팅 장치는 시간 차이(929)가 임계 시간 이하인 경우에 응답하여, 제2 대상 블록의 할당을 유지할 수 있다. 제2 대상 블록 또는 제2 대상 블록에 대응하는 크기의 메모리가 다음 반복에서 금방 사용되므로, 컴퓨팅 장치는 제2 대상 블록을 유지함으로써 제2 대상 블록에 대응하는 크기를 재할당하는데 소요되는 처리 시간을 절감할 수 있다.

[0086] 도 10은 일 실시예에 따른 메모리 사용 프로필에 기초하여 한 트레이닝 작업의 한 반복이 완료된 이후에 다른 트레이닝 작업의 반복을 개시하는 예시를 설명한다.

[0087] 일 실시예에 따르면 임의의 단일 기계 학습 모델이 복수의 GPU를 통해 분산 트레이닝되는 경우, 컴퓨팅 장치는 역전파 이후 각 GPU에서 업데이트된 파라미터를 동기화할 수 있다. 데이터 통신 채널 및 네트워트 채널의 상태에 따라, 기계 학습 모델을 동기화하기 위한 동기화 시간(1015)이 요구될 수 있다. 도 10에 도시된 메모리 사용 그래프들의 세로축은 사용된 GPU 메모리 크기를 GB 단위로 나타내고, 가로축은 시간을 초 단위로 나타낼 수 있다.

[0088] 예를 들어, 컴퓨팅 장치는, 기계 학습 모델에 대한 트레이닝 작업(1010)의 메모리 사용 프로필에 기초하여, 트레이닝 작업(1010)의 동기화 시간(1015) 동안 다른 모델에 대한 트레이닝 작업(1020)의 반복(iteration)을 수행할 수 있다. 동기화 시간(1015) 동안에는 상대적으로 적은 양의 GPU 메모리가 요구되므로, 기계 학습 모델 및 다른 모델에 대한 트레이닝 작업들의 병렬 수행에 요구되는 최대 메모리 사용량이 감소될 수 있다.

[0089] 다른 일 실시예에 따르면 컴퓨팅 장치는 다른 모델의 트레이닝 작업(1020)에서 할당됐던 메모리 블록의 해제에 의해 가용한(available) 잔여 메모리의 크기가, 메모리 사용 프로필에 따라 기계 학습 모델에 대한 트레이닝 작업(1010)의 순전파에서 요구되는 크기를 초과할 때까지, 기계 학습 모델에 대한 다음 트레이닝 반복(next training iteration)을 지연시킬 수도 있다. 따라서 컴퓨팅 장치는 기계 학습 모델의 트레이닝 작업(1010) 및 다른 모델의 트레이닝 작업(1020)에서 요구되는 최대 메모리 크기가 단일 GPU 메모리의 크기를 초과하지 않도록, 트레이닝 반복의 개시 시점을 조절할 수 있다.

[0090] 도 11은 일 실시예에 따른 메모리 사용 프로필에 기초하여 한 트레이닝 작업의 반복들 사이에 다른 트레이닝 작업의 복수의 반복들을 수행하는 예시를 설명한다.

[0091] 일 실시예에 따르면 컴퓨팅 장치는 기계 학습 모델의 트레이닝 작업에서 한 반복 및 다음 반복까지 하나 이상의 다른 트레이닝 작업을 수행할 수 있다. 도 11에 도시된 메모리 사용 그래프의 세로축은 사용된 GPU 메모리 크기를 GB 단위로 나타내고, 가로축은 시간을 초 단위로 나타낼 수 있다.

[0092] 예를 들어, 도 11에서 컴퓨팅 장치는 제1 트레이닝 작업(1110)의 반복들 사이에, 제2 트레이닝 작업(1120)의 반복들(1125)을 위한 메모리를 할당할 수 있다. 또한, 반복들(1125)이 동일한 제2 트레이닝 작업(1120)에 속하는 것으로 설명하였으나, 이로 한정하는 것은 아니고, 반복들(1125) 중 적어도 일부는 또 다른 트레이닝 작업에 속

할 수도 있다. 따라서, 컴퓨팅 장치는 2개 이상의 트레이닝 작업들을 병렬적으로 수행하기 위해 메모리를 할당 및 해제할 수 있다.

[0093] 도 12는 일 실시예에 따른 컴퓨팅 장치의 구성을 설명하는 블록도이다.

[0094] 일 실시예에 따른 컴퓨팅 장치(1200)는 저장소(1210), 그래픽 처리부(GPU, Graphic Processing Unit)(1220), 중앙 처리부(CPU, Central Processing Unit)(1230), 및 데이터 통신 채널(1250)을 포함할 수 있다.

[0095] 저장소(1210)는 기계 학습 모델(1221) 및 다른 모델을 저장할 수 있다. 예를 들어, 저장소(1210)는 기계 학습 모델(1221) 및 다른 모델의 구조 및 해당 구조에 따른 파라미터(예를 들어, 연결 가중치 등)를 저장할 수 있다. 더 나아가, 저장소(1210)는 입력 데이터(1222)를 저장할 수도 있다.

[0096] GPU(1220)는 기계 학습 모델(1221) 및 다른 모델의 트레이닝에 할당되는 GPU 메모리(1225)를 포함할 수 있다. GPU 메모리(1225)는 기계 학습 모델(1221)의 트레이닝에 요구되는 데이터를 임시적으로 저장할 수 있다. CPU(1230)는 GPU 메모리(1225)에서 필요한 메모리 블록을 할당하거나 해제할 수 있다. 예를 들어, GPU 메모리(1225)는 기계 학습 모델(1221), 입력 데이터(1222), 특징 맵(1223), 및 경사 맵 등을 저장할 수 있다. GPU 메모리(1225)는 트레이닝 작업 동안 기계 학습 모델(1221)을 저장할 수 있다. 따라서 기계 학습 모델(1221)을 위한 메모리 블록은 트레이닝 작업 동안 유지될 수 있다. GPU 메모리(1225)는 입력 데이터(1222)를 저장소(1210)로부터 수신하여 저장할 수 있다. 입력 데이터(1222)는 순전파를 위해 사용될 수 있다. GPU 메모리(1225)는 순전파 및 역전파 동안 특징 맵(1223)을 저장할 수 있다. GPU 메모리(1225)에서 특징 맵(1223)을 위한 메모리 블록은 순전파 동안 순차적으로 할당되고, 역전파 동안 순차적으로 해제될 수 있다.

[0097] CPU(1230)는 기계 학습 모델(1221)의 트레이닝 작업에 대한 GPU 메모리(1225)의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득할 수 있다. CPU(1230)는 기계 학습 모델(1221)에 대한 트레이닝 작업의 순전파 동안 메모리 요청에 응답하여, GPU 메모리(1225)로부터, 메모리 블록을 할당할 수 있다. CPU(1230)는 순전파 이후 역전파를 진행하는 동안, 메모리 사용 프로필에 기초하여, 할당됐던 메모리 블록의 적어도 일부를 해제할 수 있다. CPU(1230)는 다른 모델의 트레이닝 작업에서 메모리 요청에 응답하여, GPU 메모리(1225) 중 역전파 동안 해제된 메모리 블록을 포함하는 잔여 메모리로부터, 메모리 블록을 할당할 수 있다.

[0098] 데이터 통신 채널(1250)은 CPU(1230) 및 GPU(1220) 간의 데이터 통신을 지원할 수 있다. 예를 들어, 데이터 통신 채널(1250)은 PCI 버스(Peripheral Component Interconnect Bus)로 구현될 수 있다.

[0099] 다만, 컴퓨팅 장치(1200)의 동작을 도 12에서 설명한 바로 한정하는 것은 아니고, 도 1 내지 도 11에서 상술한 동작들 중 적어도 일부와 병렬적으로 또는 순차적으로 수행될 수도 있다.

[0100] 실시예에 따른 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 실시예를 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 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0101] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0102] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면

상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0103] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속한다.

부호의 설명

[0104] 1200: 컴퓨팅 장치

1210: 저장소

1220: GPU

1221: 기계 학습 모델

1222: 입력 데이터

1223: 특징 맵

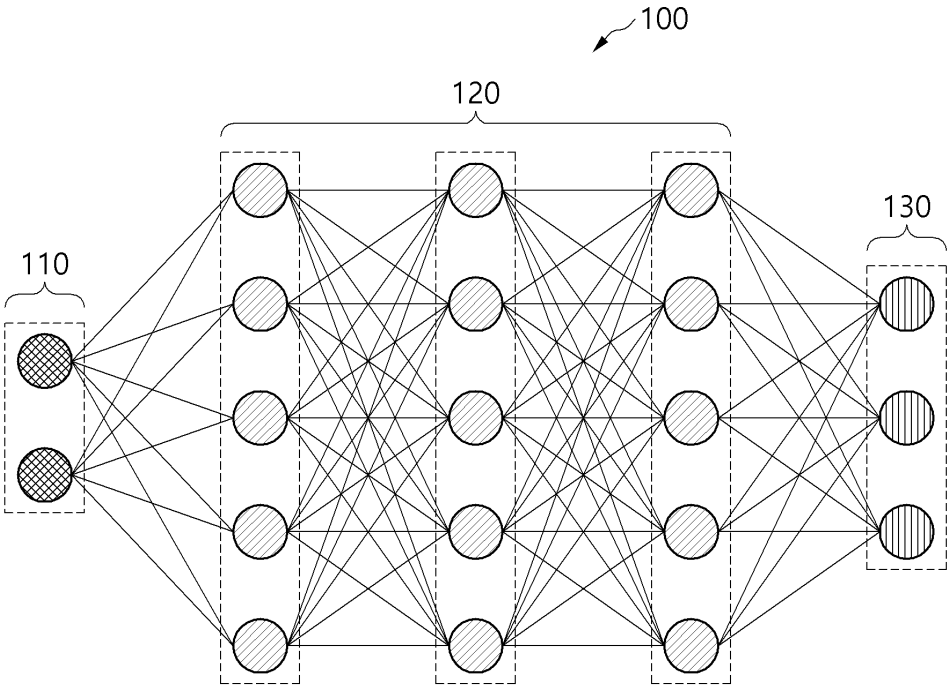
1225: GPU 메모리

1230: CPU

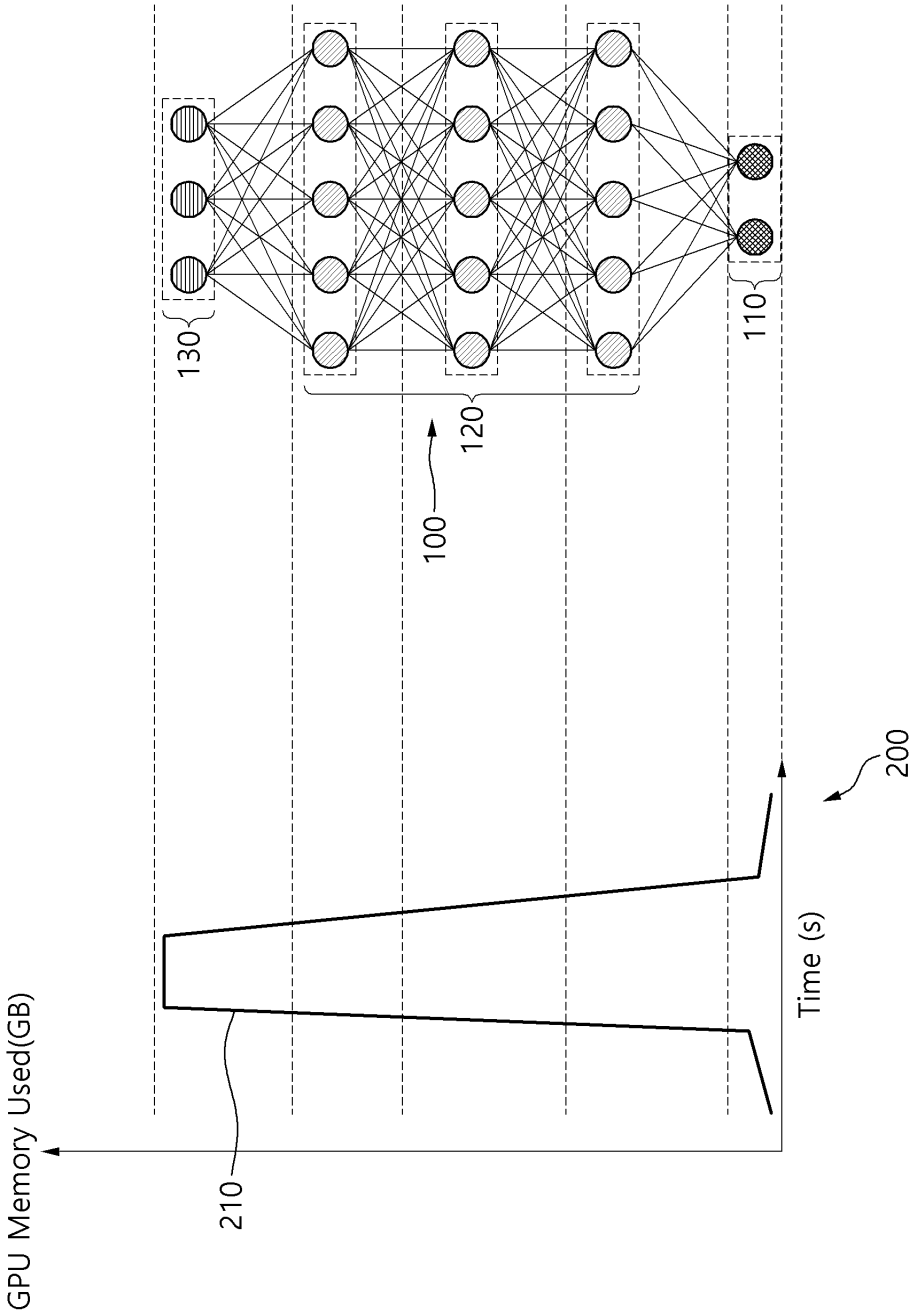
1250: 데이터 통신 채널

도면

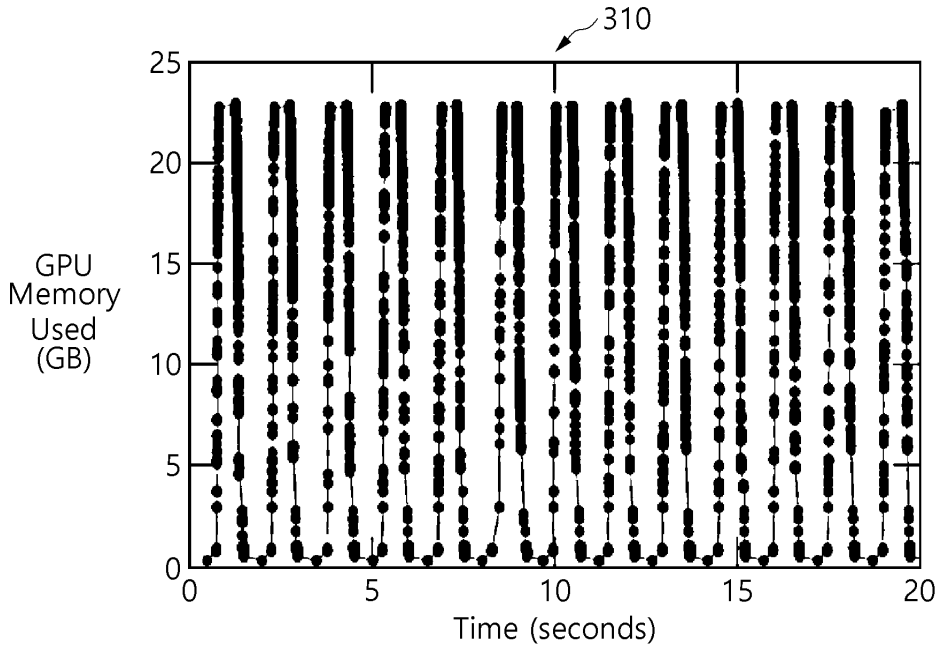
도면1



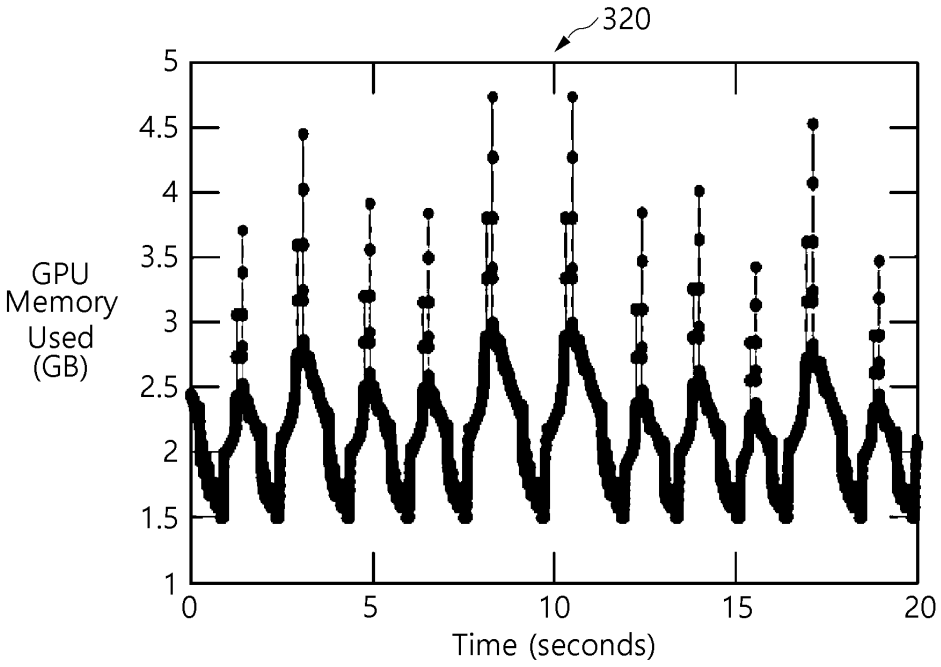
도면2



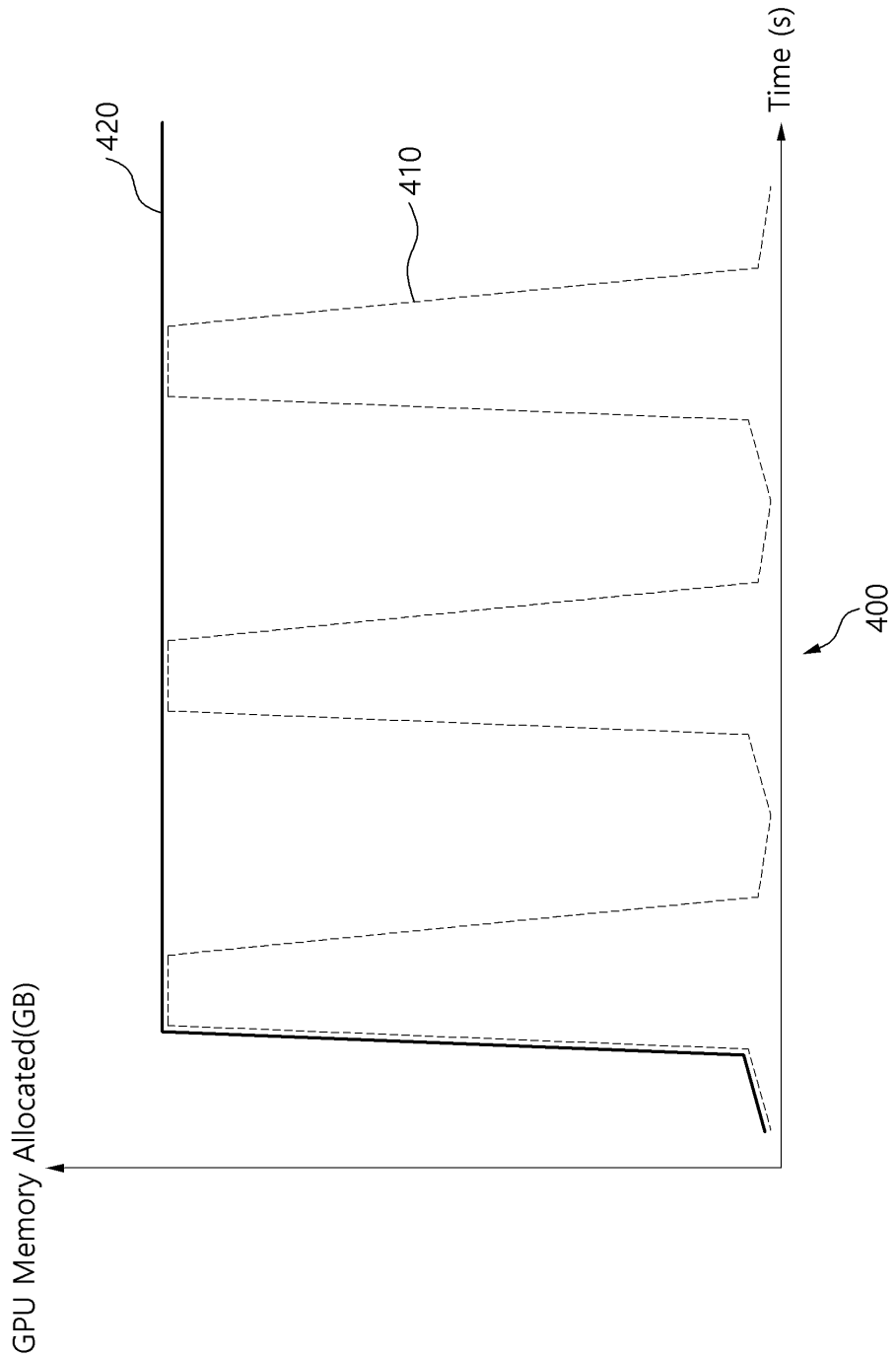
도면3a


310
25
20
15
10
5
0
GPU
Memory
Used
(GB)
0
5
10
15
20
Time (seconds)

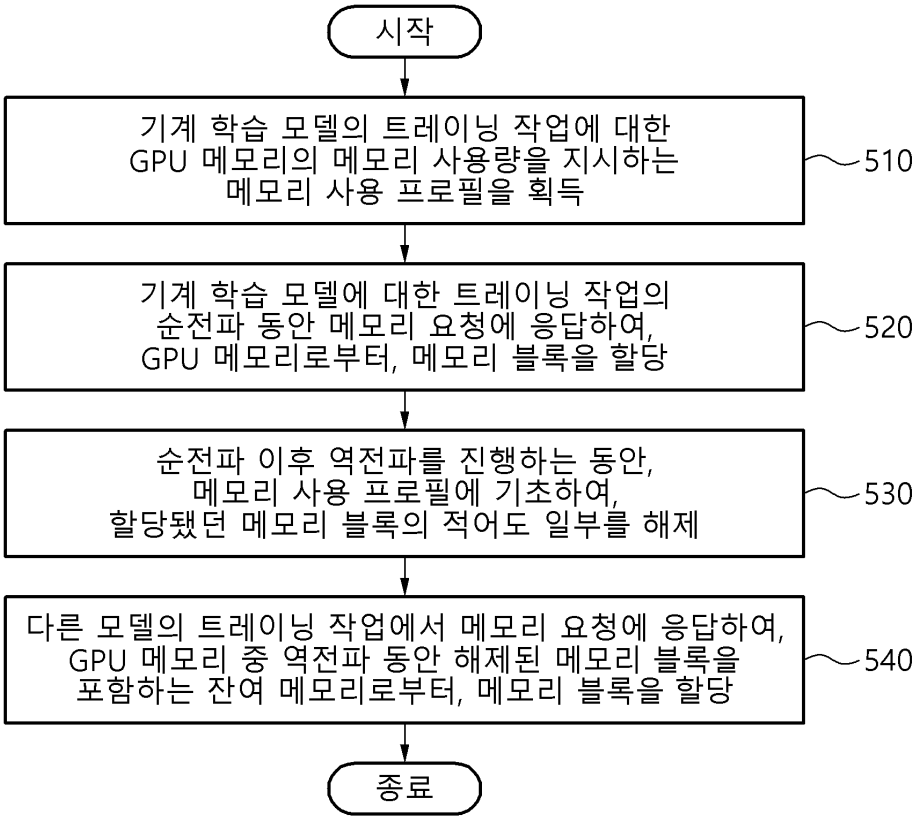
도면3b


320
5
4.5
4
3.5
3
2.5
2
1.5
1
GPU
Memory
Used
(GB)
0
5
10
15
20
Time (seconds)

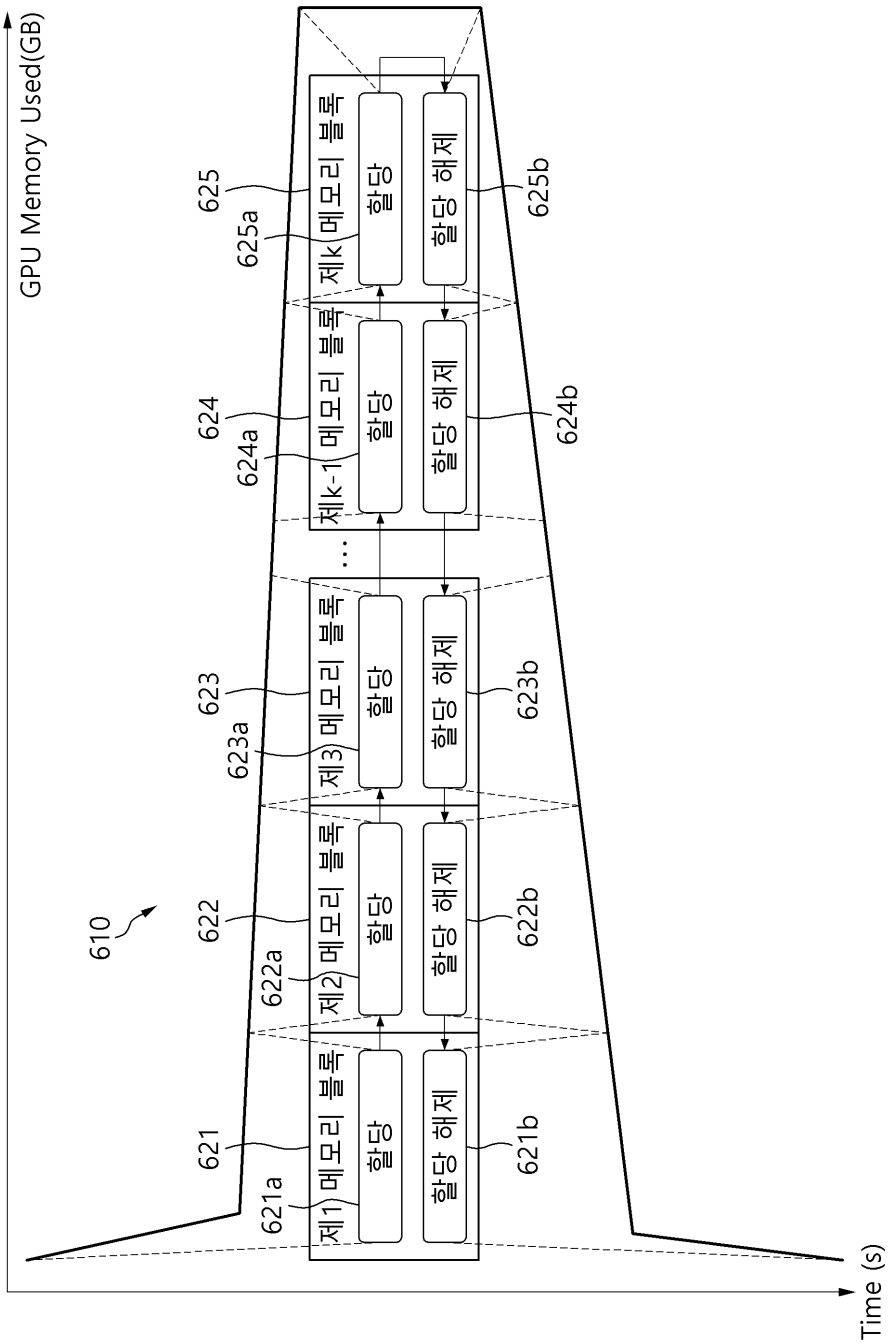
도면4



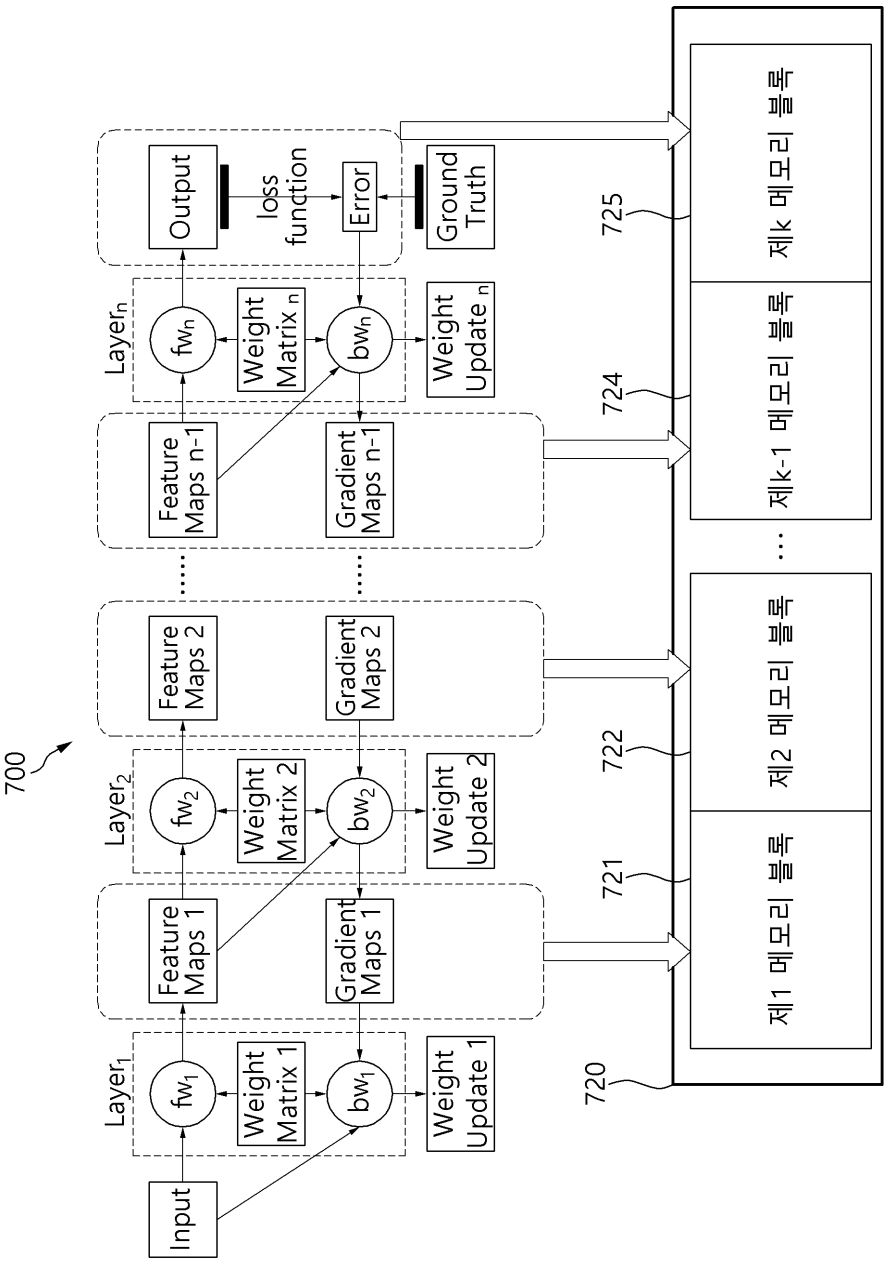
도면5


시작
기계 학습 모델의 트레이닝 작업에 대한 GPU 메모리의 메모리 사용량을 지시하는 메모리 사용 프로필을 획득
510
기계 학습 모델에 대한 트레이닝 작업의 순전파 동안 메모리 요청에 응답하여, GPU 메모리로부터, 메모리 블록을 할당
520
순전파 이후 역전파를 진행하는 동안, 메모리 사용 프로필에 기초하여, 할당됐던 메모리 블록의 적어도 일부를 해제
530
다른 모델의 트레이닝 작업에서 메모리 요청에 응답하여, GPU 메모리 중 역전파 동안 해제된 메모리 블록을 포함하는 잔여 메모리로부터, 메모리 블록을 할당
540
종료

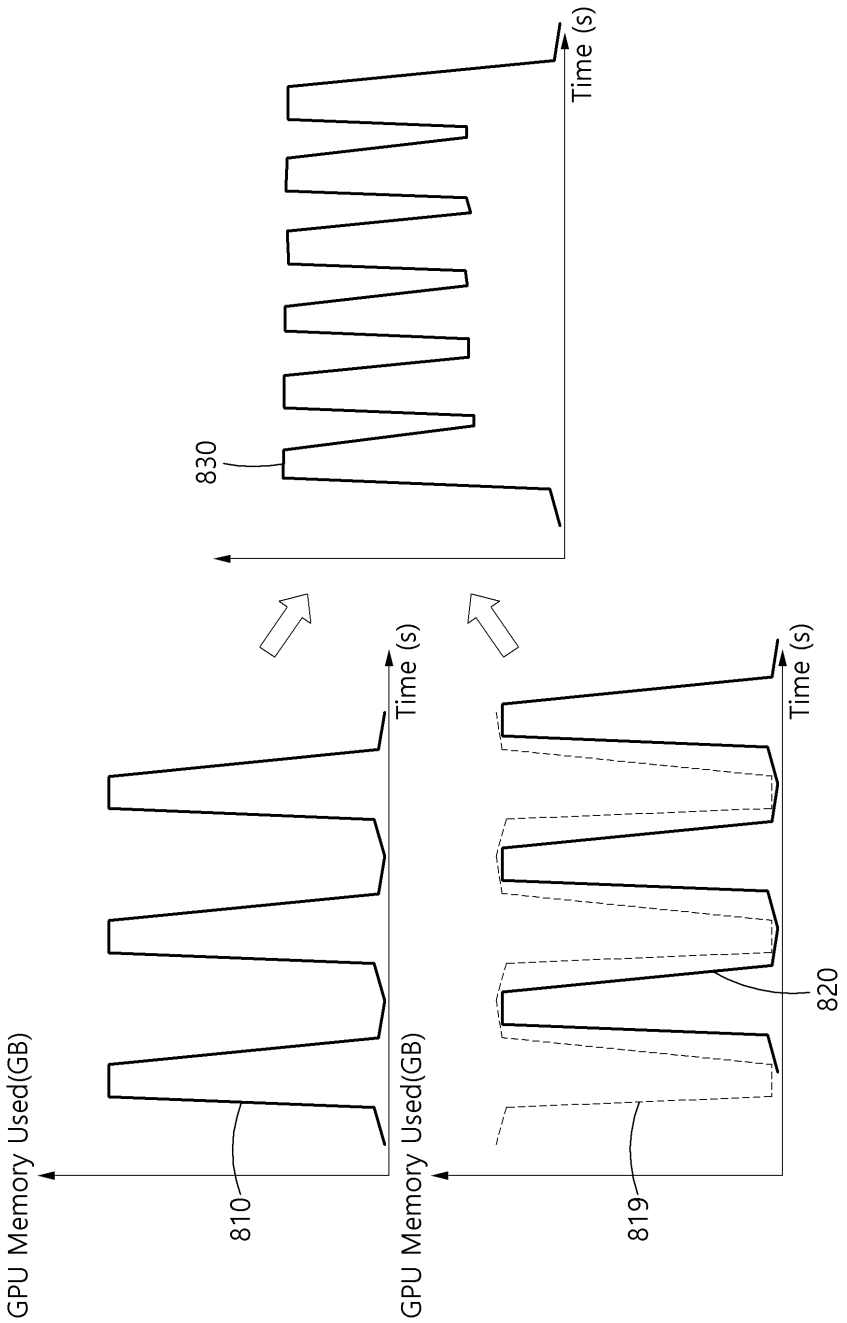
도면6


GPU Memory Used(GB)
Time (s)
610
621
621a
621b
622
622a
622b
623
623a
623b
624
624a
624b
625
625a
625b
제1) 메모리 블록
제2) 메모리 블록
제3) 메모리 블록
제k-1) 메모리 블록
제k) 메모리 블록
할당
할당 해제
...

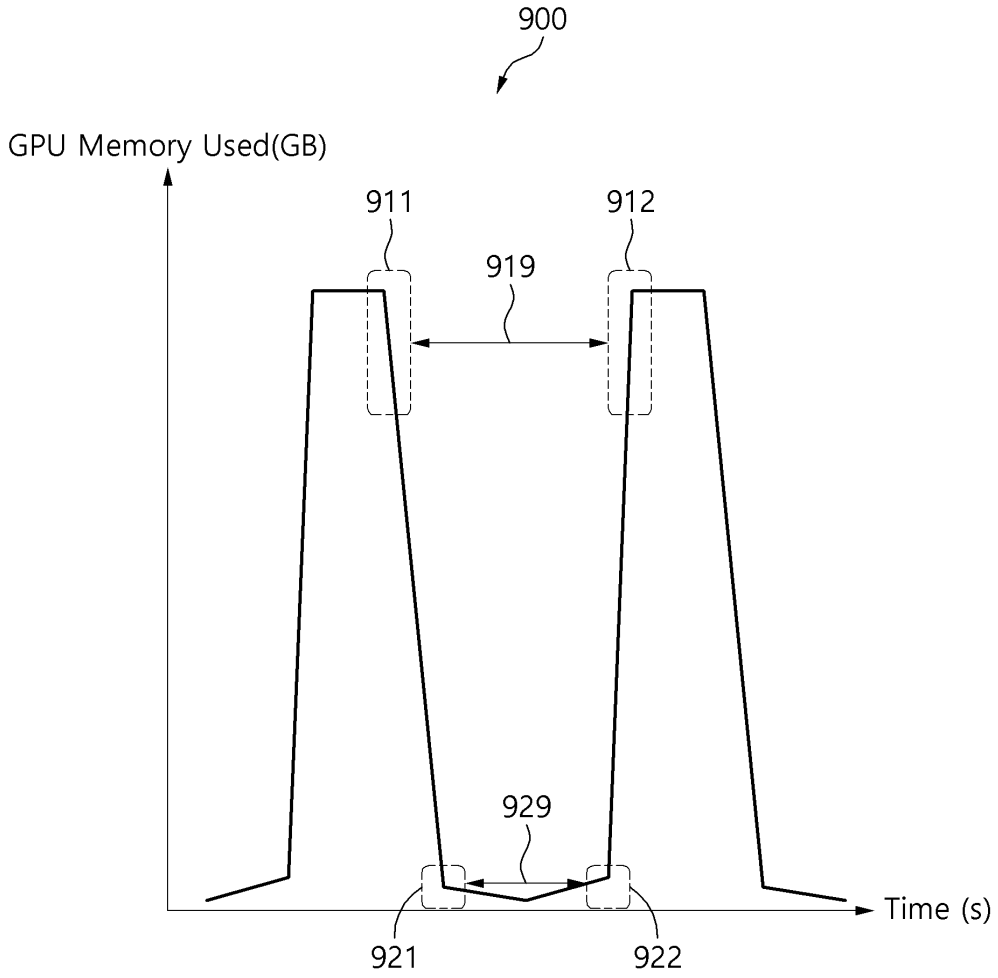
도면7


700
Input
Layer1
fw1
Weight Matrix 1
bw1
Weight Update 1
Feature Maps 1
Gradient Maps 1
Layer2
fw2
Weight Matrix 2
bw2
Weight Update 2
Feature Maps 2
Gradient Maps 2
Layern
fwn
Weight Matrix n
bwn
Weight Update n
Feature Maps n-1
Gradient Maps n-1
Output
loss function
Error
Ground Truth
720
721
722
724
725
제1 메모리 블록
제2 메모리 블록
제k-1 메모리 블록
제k 메모리 블록

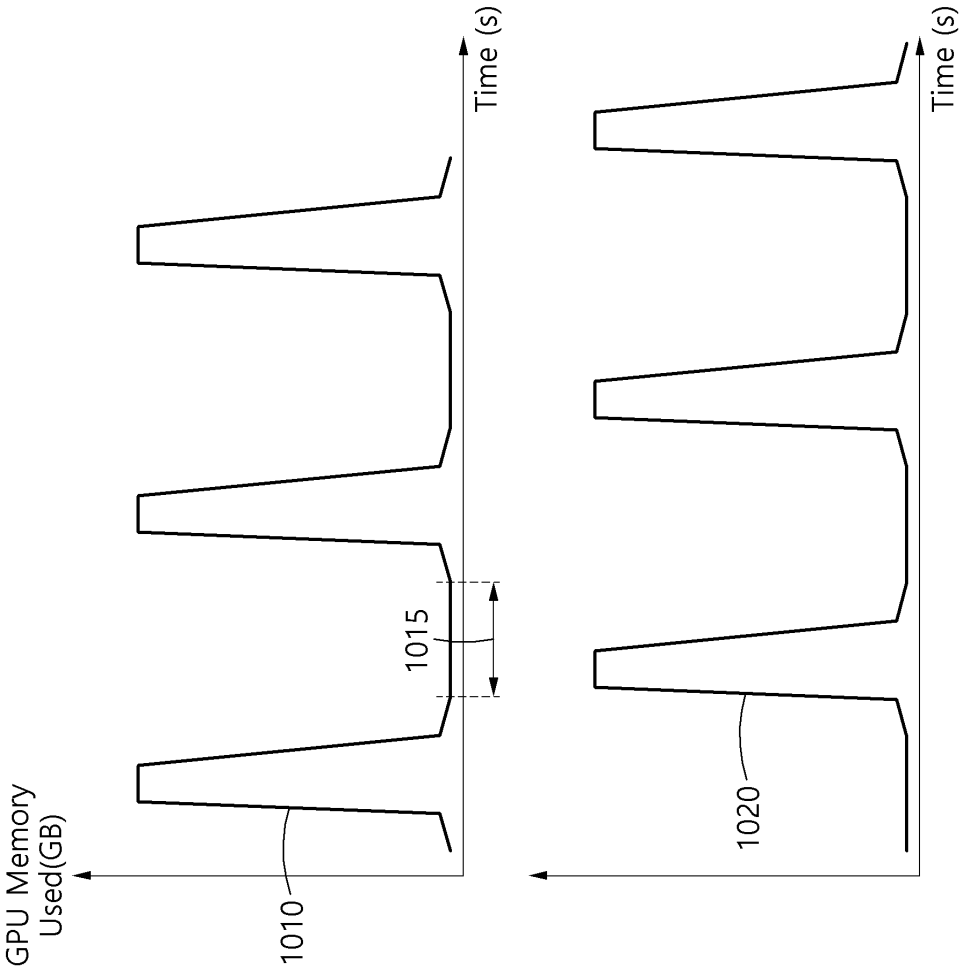
도면8


Time (s)
830
Time (s)
GPU Memory Used(GB)
810
Time (s)
GPU Memory Used(GB)
820
819

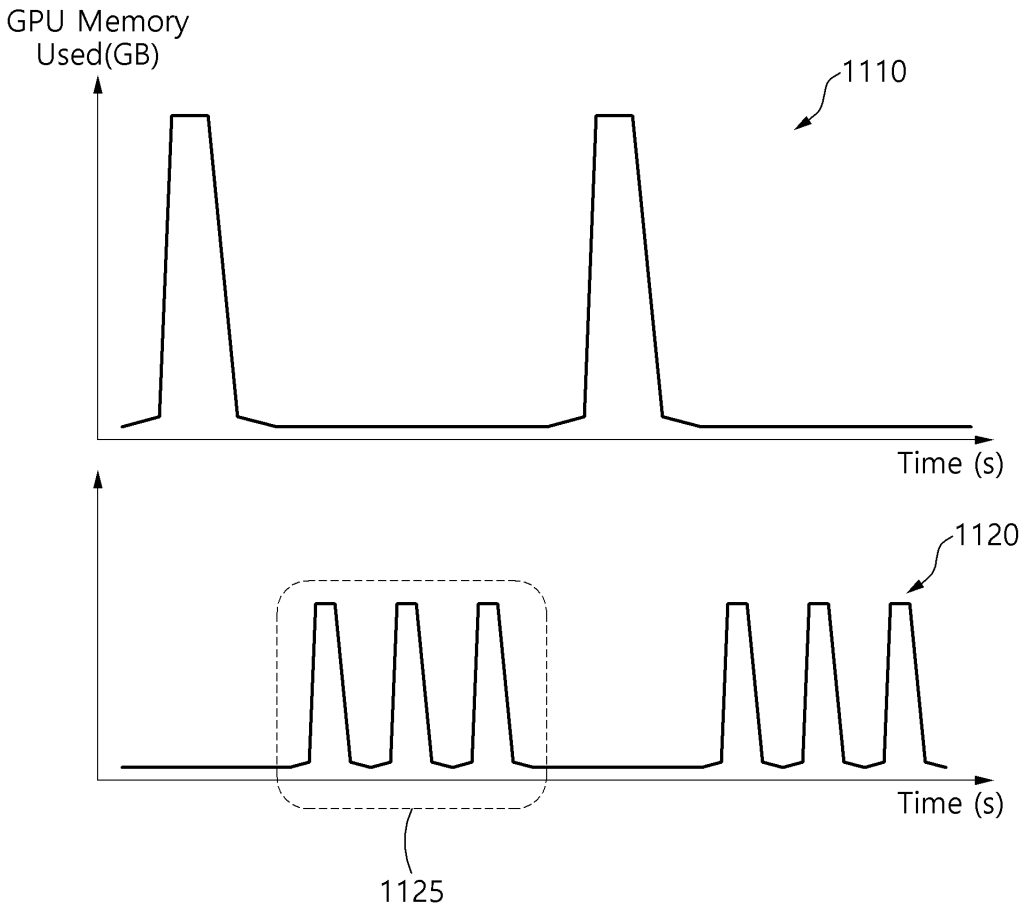
도면9



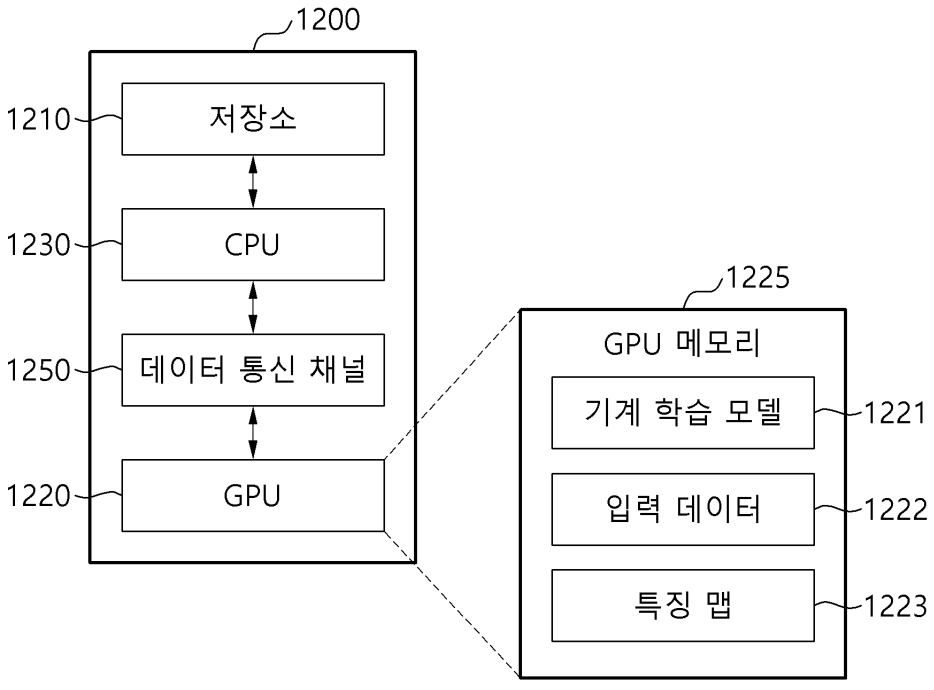
도면10



도면11


GPU Memory
Used(GB)
1110
Time (s)
1120
Time (s)
1125

도면12


1200
1210
저장소
1230
CPU
1250
데이터 통신 채널
1220
GPU
1225
GPU 메모리
기계 학습 모델
1221
입력 데이터
1222
특징 맵
1223