(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년11월22일
(11) 등록번호 10-2329752
(24) 등록일자 2021년11월17일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)
G06N 3/063 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2019-0147532
(22) 출원일자 2019년11월18일
심사청구일자 2019년11월18일
(65) 공개번호 10-2021-0060011
(43) 공개일자 2021년05월26일
(56) 선행기술조사문헌
Structured Pruning of Deep Convolutional Neural Networks 1부*
Work-in-Progress: A SIMD-Aware Pruning Technique for Convolutional Neural Networks with Multi-Sparsity Levels 1부*
가중치행렬 재정렬 GPU를 사용하는 딥 러닝 어플리케이션의 신경망 최적화 기법 연구 1부*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
성균관대학교산학협력단
경기도 수원시 장안구 서부로 2066 (천천동, 성균관대학교내)
(72) 발명자
신동군
서울특별시 강남구 역삼로 314, 305동 1004호 (역삼동, 개나리 푸르지오)
이광배
서울특별시 강남구 영동대로 230, 6동 503호 (대치동, 우성1차아파트)
(74) 대리인
제일특허법인(유)

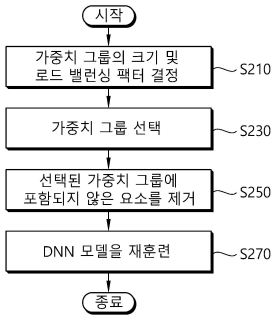
전체 청구항 수 : 총 23 항 심사관 : 양대경

(54) 발명의 명칭 **심층신경망의 그룹-레벨 프루닝 방법 및 장치**

(57) 요 약

심층신경망(Deep Neural Network, DNN)의 비정렬된(aligned) 그룹-레벨(group-level) 프루닝(pruning)에 있어서, 가중치 그룹의 크기(size) 및 로드 밸런싱 팩터(load balancing factor)를 결정하고, 상기 가중치 그룹의 크기 및 상기 로드 밸런싱 팩터에 기초하여 상기 가중치 그룹을 선택하고, 및 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들(elements)을 제거하는 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning) 방법 및 장치를 제공한다. 하드웨어의 특성을 고려한 가중치 그룹의 선택으로 심층신경망(DNN) 모델을 구동함에 있어서 높은 성능(performance)을 가지면서 동시에 정확도(accuracy) 손실을 막을 수 있다.

대 표 도



(52) CPC특허분류
G06N 3/063 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711080952
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신.방송연구개발사업(SW스타랩) 1단계 3/4
연구과제명 (SW 스타랩) 지능형 IoT 장치용 소프트웨어 프레임워크
기 여 율 1/1
과제수행기관명 성균관대학교 산학협력단
연구기간 2019.01.01 ~ 2019.12.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning) 방법에 있어서,

가중치 그룹의 크기(size) 및 로드 밸런싱 팩터(load balancing factor)를 결정하는 단계;

상기 가중치 그룹의 크기 및 상기 로드 밸런싱 팩터에 기초하여 상기 가중치 그룹을 선택하는 단계; 및

상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들(elements)을 제거하는 단계를 포함하고,

상기 로드 밸런싱 팩터는 가중치 행렬의 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수에 따라 결정되는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 2

제1항에 있어서,

상기 심층신경망(DNN) 모델을 재훈련시키는(retraining) 단계를 더 포함하는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 3

제1항에 있어서,

상기 가중치 그룹의 크기는 하드웨어의 특성에 기초하여 결정되는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 4

제3항에 있어서,

상기 하드웨어의 특성은

컴퓨팅 코어의 개수, 캐시 라인(cache line)의 크기, 및 SIMD(Single Instruction Multiple Data) 유닛의 개수 중 적어도 하나를 포함하는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 5

제1항에 있어서,

상기 가중치 그룹의 크기는 im2col 연산에 의해 행렬 곱셈을 수행하도록 재구성된 가중치 행렬에 기초하여 결정되는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 6

삭제

청구항 7

제1항에 있어서,

상기 가중치 그룹을 선택하는 단계는 고정되지 않은 임의의 위치에서 상기 가중치 그룹을 선택하는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 8

제7항에 있어서,

상기 가중치 그룹을 선택하는 단계는

목표 희소성(sparsity)을 충족시키면서 후보 가중치 그룹들 중에서 가장 큰 규모(magnitude)의 가중치 그룹을 선택하는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 9

제8항에 있어서,

이미 선택된 제1 가중치 그룹 및 상기 이미 선택된 제1 가중치 그룹과 겹치는 제2 가중치 그룹들은 상기 후보 가중치 그룹에서 제외되는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 10

제7항에 있어서,

상기 가중치 그룹을 선택하는 단계는

가중치 행렬을 1차원 가중치 행렬로 변경하고, 하기의 수학식

$$W_{s,e}^{k} = \max_{G \le i < 2G} \left(W_{s,e-i}^{k-1} + \sum_{j=e-i+1}^{e-i+G} |w_j|, W_{s,e-G}^{k} \right)$$

에 의해 계산되는 $W^{k}_{s,e}$에 대하여, $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹을 선택하는, 비정렬된 그룹-레벨 프루닝 방법.

단, $W^{k}_{s,e}$는 1차원 가중치 행렬의 s 내지 e 번째 성분으로 구성되는 하위 영역에서 선택된 k개의 겹치지 않는 가중치 그룹의 가중치 합이고, G는 그룹의 크기이고, w_j는 j번째 성분의 가중치 값이고, m은 보존될 그룹의 수이고, n은 1차원 가중치 행렬의 폭이다.

청구항 11

제10항에 있어서,

상기 $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹은 동적 프로그래밍(dynamic programming)에 의해 선택되는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 12

제1항에 있어서,

상기 가중치 요소들을 제거하는 단계는 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들을 영(0)으로 마스킹하는, 비정렬된 그룹-레벨 프루닝 방법.

청구항 13

심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning) 장치에 있어서,

가중치 그룹의 크기(size)를 결정하는 가중치 그룹 크기 결정부;

로드 밸런싱 팩터(load balancing factor)를 결정하는 로드 밸런싱 팩터 결정부;

상기 가중치 그룹의 크기 및 상기 로드 밸런싱 팩터에 기초하여 상기 가중치 그룹을 선택하는 가중치 그룹 선택부; 및

상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들(elements)을 제거하는 프루닝부를 포함하고,

상기 로드 밸런싱 팩터는 가중치 행렬의 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수에 따라 결정되는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 14

제13항에 있어서,

상기 심층신경망(DNN) 모델을 재훈련시키는(retraining) 모델 재훈련부를 더 포함하는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 15

제13항에 있어서,

상기 가중치 그룹의 크기는 하드웨어의 특성에 기초하여 결정되는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 16

제15항에 있어서,

상기 하드웨어의 특성은

컴퓨팅 코어의 개수, 캐시 라인(cache line)의 크기, 및 SIMD(Single Instruction Multiple Data) 유닛의 개수 중 적어도 하나를 포함하는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 17

제13항에 있어서,

상기 가중치 그룹의 크기는 im2col 연산에 의해 행렬 곱셈을 수행하도록 재구성된 가중치 행렬에 기초하여 결정되는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 18

삭제

청구항 19

제13항에 있어서,

상기 가중치 그룹 선택부는 고정되지 않은 임의의 위치에서 상기 가중치 그룹을 선택하는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 20

제19항에 있어서,

상기 가중치 그룹 선택부는

목표 희소성(sparsity)을 충족시키면서 후보 가중치 그룹들 중에서 가장 큰 규모(magnitude)의 가중치 그룹을 선택하는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 21

제20항에 있어서,

이미 선택된 제1 가중치 그룹 및 상기 이미 선택된 제1 가중치 그룹과 겹치는 제2 가중치 그룹들은 상기 후보 가중치 그룹에서 제외되는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 22

제19항에 있어서,

상기 가중치 그룹 선택부는

가중치 행렬을 1차원 가중치 행렬로 변경하고, 하기의 수학식

$$W^{k}_{s,e} = \max_{G \le i < 2G} \left(W^{k-1}_{s,e-i} + \sum_{j=e-i+1}^{e-i+G} |w_j|, W^{k}_{s,e-G} \right)$$

에 의해 계산되는 $W^{k}_{s,e}$에 대하여, $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹을 선택하는, 비정렬된 그룹-레벨 프루닝 장치.

단, $W^{k}_{s,e}$는 1차원 가중치 행렬의 s 내지 e 번째 성분으로 구성되는 하위 영역에서 선택된 k개의 겹치지 않는 가중치 그룹의 가중치 합이고, G는 그룹의 크기이고, w_j는 j번째 성분의 가중치 값이고, m은 보존될 그룹의 수이고, n은 1차원 가중치 행렬의 폭이다.

청구항 23

제22항에 있어서,

상기 $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹은 동적 프로그래밍(dynamic programming)에 의해 선택되는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 24

제13항에 있어서,

상기 프루닝부는 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들을 영(0)으로 마스킹하는, 비정렬된 그룹-레벨 프루닝 장치.

청구항 25

심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning)을 수행하는 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능한 기록매체에 있어서, 상기 컴퓨터 프로그램은 컴퓨팅 시스템이:

가중치 그룹의 크기(size) 및 로드 밸런싱 팩터(load balancing factor)를 결정하고;

상기 가중치 그룹의 크기 및 상기 로드 밸런싱 팩터에 기초하여 상기 가중치 그룹을 선택하고; 및

상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들(elements)을 제거하고,

상기 로드 밸런싱 팩터는 가중치 행렬의 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수에 따라 결정되도록 하는 명령을 포함하는, 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning)을 수행하는 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능한 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 심층신경망의 프루닝 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 심층신경망(Deep Neural Network, DNN)은 자연어 처리, 컴퓨터 비전 작업 등 광범위한 분야의 문제에서 뛰어난 성능을 보인다. 심층신경망(DNN)을 이용해 복잡한 문제를 해결하는 데 있어서 정확도를 높이기 위해서는 심층신경망(DNN) 모델이 더 깊어지고 커져야 한다. 그러나 대규모의 심층신경망(DNN) 모델을 구동하기 위해서는 대규모의 컴퓨팅 비용과 에너지 소비가 요구된다. 특히 컴퓨팅, 메모리, 전력자원이 제한된 모바일 환경에서 대규모의 심층신경망(DNN) 모델을 구동하는 것은 어려운 일이다. 이러한 문제를 해결하기 위해 프루닝(pruning)과 같은 압축 기법들이 제안되었다.

[0003] 프루닝은 정확도에 미치는 영향이 작은 가중치(weight) 연결부를 제거하고, 정확도를 회복하기 위해 네트워크를

재훈련(retraining)시키는 것이다. 대규모의 심층신경망(DNN)은 대체로 내부 중복성(redundancy)이 크기 때문에 정확도의 큰 손실 없이 모델의 크기를 감소시킬 수 있다.

[0004] 종래의 프루닝 방법으로서, 파인-그레인드(fine-grained) 프루닝은 모든 가중치를 고려하여 영향력이 작은 가중치를 제거하므로 정확도의 손실을 줄일 수 있지만 네트워크가 불규칙하게 되어 성능(performance)이 낮다는 문제점이 있다. 한편, 코스-그레인드(coarse-grained) 프루닝은 파인-그레인드 프루닝에 비해 구조적이고 규칙적인 네트워크를 생성하므로 성능이 높지만 가중치를 영역 단위로 제거하기 때문에 높은 희소성(sparsity)에서 정확도 손실이 크다는 문제점이 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 한국 공개특허공보 제10-2017-0128080호("신경 네트워크를 구현하는 방법 및 장치", 삼성전자주식회사, 2017.11.22.)

발명의 내용

해결하려는 과제

[0006] 본 발명의 목적은 모바일 기기와 같이 컴퓨팅 자원이 한정적인 상황에서도 대규모의 심층신경망(DNN) 모델을 구동할 수 있도록 높은 성능(performance)을 가지면서 동시에 정확도(accuracy) 손실을 막을 수 있는 그룹-레벨 프루닝 방법 및 장치를 제공하는 것이다.

과제의 해결 수단

[0007] 본 발명의 일 측면에 의하면, 심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning) 방법은, 가중치 그룹의 크기(size) 및 로드 밸런싱 팩터(load balancing factor)를 결정하는 단계, 상기 가중치 그룹의 크기 및 상기 로드 밸런싱 팩터에 기초하여 상기 가중치 그룹을 선택하는 단계, 및 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들(elements)을 제거하는 단계를 포함한다.

[0008] 상기 심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning) 방법은, 상기 심층신경망(DNN) 모델을 재훈련시키는(retraining) 단계를 더 포함할 수 있다.

[0009] 상기 가중치 그룹의 크기는 하드웨어의 특성에 기초하여 결정될 수 있다.

[0010] 상기 하드웨어의 특성은 컴퓨팅 코어의 개수, 캐시 라인(cache line)의 크기, 및 SIMD(Single Instruction Multiple Data) 유닛의 개수 중 적어도 하나를 포함할 수 있다.

[0011] 상기 가중치 그룹의 크기는 가중치 행렬의 모양에 기초하여 결정될 수 있다.

[0012] 상기 로드 밸런싱 팩터는 가중치 행렬의 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수에 따라 결정될 수 있다.

[0013] 상기 가중치 그룹을 선택하는 단계는 고정되지 않은 임의의 위치에서 상기 가중치 그룹을 선택할 수 있다.

[0014] 상기 가중치 그룹을 선택하는 단계는 목표 희소성(sparsity)을 충족시키면서 후보 가중치 그룹들 중에서 가장 큰 규모(magnitude)의 가중치 그룹을 선택할 수 있다.

[0015] 이때, 이미 선택된 제1 가중치 그룹 및 상기 이미 선택된 제1 가중치 그룹과 겹치는 제2 가중치 그룹들은 상기 후보 가중치 그룹에서 제외될 수 있다.

[0016] 상기 가중치 그룹을 선택하는 단계는 가중치 행렬을 1차원 가중치 행렬로 변경하고, 하기의 수학식

[0017] $$W_{s,e}^{k} = \max_{G \le i < 2G} \left(W_{s,e-i}^{k-1} + \sum_{j=e-i+1}^{e-i+G} |w_j|, W_{s,e-G}^{k} \right)$$

[0018] 에 의해 계산되는 $W^{k}_{s,e}$에 대하여, $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹을 선택할 수 있다.

[0019] 단, $W^{k}_{s,e}$는 1차원 가중치 행렬의 s 내지 e 번째 성분으로 구성되는 하위 영역에서 선택된 k개의 겹치지 않는 가중치 그룹의 가중치 합이고, G는 그룹의 크기이고, w_j는 j번째 성분의 가중치 값이고, m은 보존될 그룹의 수이고, n은 1차원 가중치 행렬의 폭이다.

[0020] 이때, 상기 $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹은 동적 프로그래밍(dynamic programming)에 의해 선택될 수 있다.

[0021] 상기 가중치 요소들을 제거하는 단계는 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들을 영(0)으로 마스킹하는 것일 수 있다.

[0022] 본 발명의 다른 측면에 의하면, 심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning) 장치는, 가중치 그룹의 크기(size)를 결정하는 가중치 그룹 크기 결정부, 로드 밸런싱 팩터(load balancing factor)를 결정하는 로드 밸런싱 팩터 결정부, 상기 가중치 그룹의 크기 및 상기 로드 밸런싱 팩터에 기초하여 상기 가중치 그룹을 선택하는 가중치 그룹 선택부, 및 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들(elements)을 제거하는 프루닝부를 포함한다.

[0023] 상기 심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-레벨(group-level) 프루닝(pruning) 장치는 상기 심층신경망(DNN) 모델을 재훈련시키는(retraining) 모델 재훈련부를 더 포함할 수 있다.

[0024] 상기 가중치 그룹의 크기는 하드웨어의 특성에 기초하여 결정될 수 있다.

[0025] 상기 하드웨어의 특성은 컴퓨팅 코어의 개수, 캐시 라인(cache line)의 크기, 및 SIMD(Single Instruction Multiple Data) 유닛의 개수 중 적어도 하나를 포함할 수 있다.

[0026] 상기 가중치 그룹의 크기는 가중치 행렬의 모양에 기초하여 결정될 수 있다.

[0027] 상기 로드 밸런싱 팩터는 가중치 행렬의 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수에 따라 결정될 수 있다.

[0028] 상기 가중치 그룹 선택부는 고정되지 않은 임의의 위치에서 상기 가중치 그룹을 선택할 수 있다.

[0029] 상기 가중치 그룹 선택부는 목표 희소성(sparsity)을 충족시키면서 후보 가중치 그룹들 중에서 가장 큰 규모(magnitude)의 가중치 그룹을 선택할 수 있다.

[0030] 이때, 이미 선택된 제1 가중치 그룹 및 상기 이미 선택된 제1 가중치 그룹과 겹치는 제2 가중치 그룹들은 상기 후보 가중치 그룹에서 제외될 수 있다.

[0031] 상기 가중치 그룹 선택부는 가중치 행렬을 1차원 가중치 행렬로 변경하고, 하기의 수학식

[0032] $$W^{k}_{s,e} = \max_{G \le i < 2G} \left(W^{k-1}_{s,e-i} + \sum_{j=e-i+1}^{e-i+G} |w_j|, W^{k}_{s,e-G} \right)$$

[0033] 에 의해 계산되는 $W^{k}_{s,e}$에 대하여, $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹을 선택할 수 있다.

[0034] 단, $W^{k}_{s,e}$는 1차원 가중치 행렬의 s 내지 e 번째 성분으로 구성되는 하위 영역에서 선택된 k개의 겹치지 않는 가중치 그룹의 가중치 합이고, G는 그룹의 크기이고, w_j는 j번째 성분의 가중치 값이고, m은 보존될 그룹의 수이고, n은 1차원 가중치 행렬의 폭이다.

[0035] 이때, 상기 $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹은 동적 프로그래밍(dynamic programming)에 의해 선택될 수 있다.

[0036] 상기 프루닝부는 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들을 영(0)으로 마스킹할 수 있다.

[0037] 본 발명의 또 다른 측면에 의하면, 심층신경망(Deep Neural Network, DNN) 모델의 비정렬된(unaligned) 그룹-

레벨(group-level) 프루닝(pruning)을 수행하는 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능한 기록매체에 있어서, 상기 컴퓨터 프로그램은 컴퓨팅 시스템이 가중치 그룹의 크기(size) 및 로드 밸런싱 팩터(load balancing factor)를 결정하고, 상기 가중치 그룹의 크기 및 상기 로드 밸런싱 팩터에 기초하여 상기 가중치 그룹을 선택하고, 상기 선택된 가중치 그룹에 포함되지 않은 가중치 요소들(elements)을 제거하도록 하는 명령을 포함하는 컴퓨터 프로그램을 포함한다.

발명의 효과

[0038] 본 발명의 실시예들에 따른 그룹-레벨 프루닝 방법 및 장치에 따르면, 하드웨어 특성에 기초하여 임의의 위치에서 가중치 그룹을 선택함으로써 심층신경망(DNN) 모델을 구동함에 있어서 높은 성능(performance)을 가지면서 동시에 정확도(accuracy) 손실을 막을 수 있다.

[0039] 본 발명의 실시예들에 따른 그룹-레벨 프루닝 방법 및 장치에 따르면, 로드 밸런싱 팩터를 조정하여 가중치 그룹을 선택함으로써 심층신경망 모델의 정확도와 성능 간의 트레이드-오프를 조절할 수 있다.

도면의 간단한 설명

[0040] 도 1은 종래의 프루닝 방법들을 비교하기 위한 개념도이다.

도 2는 본 발명의 일 실시예에 따른 그룹-레벨 프루닝 방법의 순서도이다.

도 3은 가중치 행렬의 재구성을 설명하기 위한 개념도이다.

도 4는 프루닝 방법에 따른 가중치 그룹의 선택을 설명하기 위한 개념도이다.

도 5는 가중치 그룹의 크기가 4일 때 $W^{m}_{1,n}$을 최대화하는 5가지 경우를 나타낸 개념도이다.

도 6은 로드 밸런싱 팩터에 따른 가중치 그룹 선택 결과의 차이를 설명하기 위한 개념도이다.

도 7은 본 발명의 일 실시예에 따른 그룹-레벨 프루닝 장치의 블록도이다.

도 8 내지 도 13은 본 발명의 실시예들에 따른 그룹-레벨 프루닝 방법 및 장치의 실험 결과이다.

발명을 실시하기 위한 구체적인 내용

[0041] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다.

[0042] 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0043] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는 데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 기재된 항목들의 조합 또는 복수의 기재된 항목들 중의 어느 항목을 포함한다.

[0044] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0045] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 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.

[0046] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일

반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가진 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0047] 이하에서는 첨부된 도면을 참조하여 본 발명에 따른 심층신경망(Deep Neural Network, DNN)의 그룹-레벨(group-level) 프루닝(pruning) 방법 및 장치에 대하여 본 발명이 속하는 기술분야에서 통상의 지식을 가진 사람이 본 발명을 쉽게 실시할 수 있도록 명확하고 상세하게 설명하기로 한다.

[0049] 본 명세서에서, 가중치(weight) 그룹의 크기(size)는 상기 가중치 그룹에 포함된 가중치 요소(element)의 개수 및 상기 가중치 그룹의 차원을 나타낸다. 예를 들어, 4개의 가중치 요소를 포함하는 2차원 가중치 그룹의 크기는 2×2로 표현하고, 2개의 가중치 요소를 포함하는 1차원 가중치 그룹의 크기는 2×1 또는 2로 표현한다.

[0050] 본 명세서에서, 가중치 그룹의 규모(magnitude)는 상기 가중치 그룹에 포함된 모든 가중치 요소의 가중치의 합을 나타낸다.

[0052] 도 1은 종래의 프루닝 방법들을 비교하기 위한 개념도이다.

[0053] 프루닝은 정확도에 미치는 영향이 작은 가중치(weight) 연결부를 제거하고, 정확도를 회복하기 위해 네트워크를 재훈련(retraining)시키는 것이다. 대규모의 심층신경망(DNN)은 대체로 내부 중복성(redundancy)이 크기 때문에 정확도의 큰 손실 없이 모델의 크기를 감소시킬 수 있다.

[0054] 종래의 프루닝 방법은 파인-그레인드(fine-grained) 프루닝, 코스-그레인드(coarse-grained) 프루닝, 및 미디엄-그레인드(medium-grained) 프루닝으로 나눌 수 있다.

[0055] 도 1의 (a)를 참조하면, 파인-그레인드 프루닝은 가중치 요소 단위로 가중치 그룹을 선택한다. 즉, 파인-그레인드 프루닝은 가중치 요소 단위로 가중치가 큰 요소를 선택하고, 선택되지 않은 가중치 요소를 제거하므로, 정확도의 손실을 줄일 수 있지만 네트워크를 불규칙하게 만들어 성능(즉, 실행 속도)이 낮다는 문제점이 있다.

[0056] 도 1의 (d)를 참조하면, 코스-그레인드 프루닝은 영역(region) 단위로 가중치 요소를 선택한다. 예를 들어, 영역은 필터(filter), 채널(channel), 행(row), 또는 열(column)이 될 수 있다. 코스-그레인드 프루닝은 네트워크를 규칙적이고 구조적으로 만들어 성능이 높으나 큰 가중치 요소들이 제거되어 정확도의 손실이 크다는 문제점이 있다.

[0057] 도 1의 (b) 및 (c)는 미디엄-그레인드 프루닝을 나타낸다. 도 1의 (b)는 1차원 그룹-레벨 프루닝을 나타낸 것이고, (c)는 2차원 그룹-레벨 프루닝을 나타낸 것이다. 미디엄-그레인드 프루닝은 파인-그레인드 프루닝과 코스-그레인드 프루닝의 장단점을 상호 보완하기 위한 것으로서, 그룹 단위로 가중치 요소를 선택하고, 선택되지 않은 가중치 요소를 제거한다. 여기서, 그룹은 가중치 행렬의 행 또는 열보다 작은 크기를 갖는다. 예를 들어, 가중치 행렬의 크기가 6×6인 경우 1차원 가중치 그룹의 크기는 6×1보다 작고, 2차원 가중치 그룹의 크기는 2×2보다 크고 6×6보다 작은 선에서 결정된다.

[0058] 그러나 종래의 그룹-레벨 프루닝은, 정렬된(aligned) 접근방식을 채택한 것으로서, 각 그룹은 서로 겹치지 않도록 하드웨어 특성에 따라 미리 결정되어 있으며 그 중에서 규모가 작은 그룹이 제거된다. 이미 결정되어 있는 그룹들의 규모를 비교하여 해당 가중치 그룹을 제거할지 말지를 결정하면 되므로 그 과정이 간단한 반면, 낮은 가중치 값 주변에 있는 큰 가중치 값들이 제거되어 정확도 손실이 발생할 수 있다. 이러한 정확도 손실은 높은 희소성(sparsity)에서 더 크게 발생할 수 있으며, 그룹의 크기가 커질수록 더 크게 발생할 수 있다.

[0060] 도 2는 본 발명의 일 실시예에 따른 비정렬된 그룹-레벨 프루닝 방법의 순서도이다.

[0061] 도 2를 참조하면, 단계 S210은 가중치 그룹의 크기 및 로드 밸런싱 팩터(load balancing factor)를 결정하는 단계이다.

[0062] 가중치 그룹의 크기는 하드웨어의 특성에 기초하여 결정될 수 있다. 예를 들어, 상기 하드웨어의 특성은 컴퓨팅 코어의 개수, 캐시 라인(cache line)의 길이, 및 SIMD(Single Instruction Multiple Data) 유닛의 개수 중 적어도 하나를 포함할 수 있다. 예를 들어, 병렬 컴퓨팅 장치에서 동시에 얼마나 많은 MAC(multiply-and-accumulate) 연산을 실행할 수 있는지를 고려하여 가중치 그룹의 크기를 결정할 수 있다. 예를 들어, ARM Mali-T628 GPU는 최대 8개의 셰이더(shader) 코어로 확장할 수 있으며, 각 셰이더 코어에는 두 개의 산술 파이프라인(arithmetic pipeline), 하나의 로드/스토어 파이프라인(load/store pipeline), 및 하나의 텍스처 파이프라인(texture pipeline)이 있다. 각 산술 파이프라인에는 4개의 128 비트 SIMD 유닛이 있으므로, SIMD 유닛을 최대

한 활용하려면 가중치 그룹의 크기는 4의 배수이어야 한다.

[0063] 또는 가중치 그룹의 크기는 가중치 행렬의 모양에 기초하여 결정될 수 있다. 예를 들어, 도 3을 참조하면, 컨볼루션을 구현하는 경우 입력 행렬은 컨볼루션을 행렬 곱셈으로 변환하기 위해 im2col(image-to-column) 연산에 의해 재구성된다(301). 계층 내의 컨볼루션 커널 역시 재구성된 입력 행렬로 행렬 곱셈을 수행하도록 병합 및 재구성된다. 결과적으로, 재구성된 커널 행렬은 (K^2C by F)의 형태를 갖는다. 여기서, K는 컨볼루션 커널의 크기(즉, KXK), C는 채널의 수, F는 필터의 수를 나타낸다. 2차원 커널을 1차원 배열(array)로 변환할 때 2차원 그룹-레벨 프루닝을 하기 위해서는 그룹의 가중치 값이 연속된 위치에 있어야 하고, 재구성된 커널 행렬의 너비와 높이가 그룹 크기의 배수가 아닌 경우 각 행과 열의 끝을 영(0)으로 채워야 한다.

[0064] 로드 밸런싱 팩터는 성능과 정확도의 트레이드-오프(trade-off)를 고려하기위해 스레드(thread)에 대한 로드 밸런싱의 정도를 나타내는 것으로서, 로드 밸런싱 팩터에 따라 가중치 그룹의 선택 결과가 달라진다. 이에 대해서는 단계 S230에서 후술하기로 한다.

[0065] 단계 S230은 가중치 그룹을 선택하는 단계이다. 비정렬된(unaligned) 그룹-레벨 프루닝은 종래의 정렬된 그룹-레벨 프루닝과 달리 가중치 그룹을 임의의 위치에서 선택할 수 있다. 우선, 프루닝 방법에 따라 가중치 그룹이 어떻게 선택되는지 설명하기로 한다.

[0066] 도 4는 프루닝 방법에 따른 가중치 그룹의 선택 결과의 차이를 설명하기 위한 예시이다.

[0067] 도 4의 (a)를 참조하면, 예를 들어 목표 희소성(sparsity)이 66.6%일 때, 요소-레벨 프루닝은 목표 희소성(sparsity)을 만족시키기 위해 가중치 값이 큰 상위 12개의 가중치 요소들을 선택하고, 선택되지 않은 가중치 요소를 제거한다. 이에 따라 보존된 가중치의 합(W)은 102로 다른 프루닝 방법과 비교하여 가장 높지만, 희소 행렬(sparse matrix)이 불규칙적으로 생성되므로 성능이 낮다는 문제점이 있다.

[0068] 도 4의 (b)를 참조하면, 정렬된 그룹-레벨 프루닝은 목표 희소성(sparsity)을 만족시키기 위하여 위치가 정해진 가중치 그룹 -본 예시에서, 가중치 그룹의 크기는 2×1임- 중에서 규모가 12보다 큰 가중치 그룹들을 선택하고, 선택된 가중치 그룹에 포함되지 않은 가중치 요소를 제거한다. 이에 따라 요소-레벨 프루닝보다 더 규칙적인 희소 행렬을 생성하지만, 보존된 가중치의 합(W)이 92로 감소한다. 도 4의 (a)와 비교하면 7, 8, 9와 같이 큰 값을 가지는 가중치 요소들 일부가 선택되지 않은 것을 알 수 있다.

[0069] 비정렬된 그룹-레벨 프루닝은 고정되지 않은 임의의 위치에서 가중치 그룹을 선택하므로 큰 값을 가지는 가중치 요소를 더 많이 보존할 수 있다. 비정렬된 그룹-레벨 프루닝에서 가중치 그룹의 선택 알고리즘은 탐욕(greedy) 알고리즘과 최적(optimal) 알고리즘이 있다.

[0070] 도 4의 (c)를 참조하면, 탐욕 알고리즘에서는 목표 희소성(sparsity)을 충족시키면서 후보 가중치 그룹들 중에서 가장 큰 규모의 가중치 그룹을 선택한다. 가중치 그룹의 각 선택 단계에서, 이전 단계에서 이미 선택된 가중치 그룹은 후보 가중치 그룹에서 제외되고, 이전 단계에서 이미 선택된 가중치 그룹과 겹치는 다른 가중치 그룹들도 후보 가중치 그룹에서 제외된다. 예를 들어, 도 4의 (c)의 예시에서, 6번째 행의 (11, 8) 그룹이 가장 먼저 선택되고, 추가적으로 5개의 그룹이 (8, 9), (7, 9), (9, 6), (5, 9), 및 (5, 8)의 순서로 선택된다. 선택 결과를 도 4의 (b)와 비교하면, 동일한 목표 희소성(sparsity)을 달성하면서도 보존된 가중치의 합(W)이 94로 더 큰 값을 가진다.

[0071] 도 4의 (d)를 참조하면, 최적 알고리즘에서는 탐욕 알고리즘보다 보존된 가중치의 합(W)이 97로 더 큰 값을 가진다. 이하에서, 최적 알고리즘에 대해 설명한다.

[0072] 최적 알고리즘에서는 1차원 공간에서 가중치 그룹을 선택하기 위해, 도 3과 같이, 목표 가중치 행렬을 (K^2C by F)에서 (K^2CF by 1)의 형태로 변환한다(303). 여기서, 보존될 그룹의 수를 m, 변환된 행렬의 폭을 n, 목표 희소성(sparsity)을 S라고 하면 아래의 관계를 만족한다.

수학식 1

[0073] $$n = K^2CF$$

수학식 2

[0074] $$m = n(1 - S)$$

[0076] 최적 알고리즘에서 가중치 그룹의 선택은 수학식 3과 같이 정의되는 $W^{k}_{s,e}$에 대하여 $W^{m}_{1,n}$이 최대가 되는 m개의 가중치 그룹을 선택하는 것이다. 이때 동적 프로그래밍(dynamic programming)을 이용해 해를 구할 수 있다. 예를 들어, 도 5에 도시된 바와 같이, 그룹의 크기가 4인 경우 5개의 해가 존재할 수 있다.

수학식 3

[0077] $$W^{k}_{s,e} = \max_{G \le i < 2G} (W^{k-1}_{s,e-i} + \sum_{j=e-i+1}^{e-i+G} |w_j|, W^{k}_{s,e-G})$$

[0078] 여기서, $W^{k}_{s,e}$는 1차원 가중치 행렬의 s 내지 e 번째 성분으로 구성되는 하위 영역에서 선택된 k개의 겹치지 않는 가중치 그룹의 가중치의 합이고, G는 가중치 그룹의 크기이고, w_j는 j번째 성분의 가중치 값이다.

[0079] 한편, 가중치 그룹의 선택 결과는 로드 밸런싱 팩터에 따라 달라질 수 있다. 가중치 행렬에서 SpMV(Sparse Matrix-Vector) 또는 SpMM(Sparse Matrix-Matrix) 연산은 일반적으로 여러 스레드에 의해 수행되며, 각 스레드는 하나의 행을 처리한다. 이때 모든 행에 비슷한 개수의 영(0)이 아닌 수(non-zeros)가 있으면 여러 스레드에 대한 로드가 균형되어 더 좋은 성능을 얻을 수 있다. 그러므로 프루닝에 있어서 균형된 희소 행렬(balanced sparse matrix)을 생성할 필요가 있다. 그러나 모든 행이 로드 밸런스를 위해 동일한 개수의 가중치 요소를 갖도록 제한하면 큰 값의 가중치 요소들이 제거되어 정확도가 낮아진다. 이러한 트레이드-오프(trade-off)를 고려하기 위해 스레드의 로드 밸런싱 정도를 나타내는 로드 밸런싱 팩터를 사용한다.

[0080] 로드 밸런싱 팩터(L)는 $0 \le L \le 1$의 값을 가지며, 수학식 4와 같이 각 행의 희소성(sparsity)의 하한을 결정하고, 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수를 결정할 수 있다.

수학식 4

[0081] $$S_{l,r} \ge S_l * L$$

[0082] 여기서, $S_{l,r}$은 목표 신경망의 l번째 계층(layer)에서 r번째 행의 희소성(sparsity)을 나타내고, S_l은 l번째 계층의 목표 희소성(sparsity)을 나타낸다.

[0083] 도 6은 목표 희소성(sparsity)이 61.6%인 경우 로드 밸런싱 팩터에 따른 가중치 그룹의 선택 결과를 비교하기 위한 예시이다. 도 6의 (a)와 같이 L = 0인 경우에는 각 행에서 보존될 수 있는 그룹의 수에 제한이 없다. 즉, 로드 밸런싱을 하지 않는다. 그 결과 목표 희소성(sparsity)을 만족하면서 정확도가 가장 큰 프루닝 행렬을 얻을 수 있다. 그러나, 도 6의 (c)와 같이 L = 1이면 각 행에서 동일한 수의 가중치 그룹을 선택하여 목표 희소성(sparsity)을 만족하면서 성능이 가장 높은 프루닝 행렬을 얻을 수 있다. 한편, 도 6의 (b)와 같이 L = 0.67이면 각 행은 41.2%(= 61.6% × 0.67)보다 높은 희소성(sparsity)을 가질 수 있으므로 각 행에서 최대 2개의 가중치 그룹을 보존할 수 있다.

[0084] 다시 도 2를 참조하면, 단계 S250은 선택된 가중치 그룹에 포함되지 않은 가중치 요소를 제거하는 단계이다. 즉, 선택된 가중치 그룹에 포함되지 않은 가중치 요소는 영(0)으로 마스킹되며, 향후 연산의 대상에서 제외된다.

[0085] 단계 S270은 선택된 가중치 그룹에 포함되지 않은 가중치 요소가 제거된 신층신경망(DNN) 모델을 재훈련(retraining)시키는 단계이다. 이후 미리 설정된 반복횟수만큼 단계 S210 내지 S270을 반복한 후 프루닝 과정이

종료된다.

[0087] 도 7은 본 발명의 일 실시예에 따른 비정렬된 그룹-레벨 프루닝 장치의 구성도이다.

[0088] 비정렬된 그룹-레벨 프루닝 장치(700)는 가중치 그룹 크기 결정부(710), 로드 밸런싱 팩터 결정부(730), 가중치 그룹 선택부(750), 프루닝부(770), 및 모델 재훈련부(790)를 포함할 수 있으며, 상술한 비정렬된 그룹-레벨 프루닝 방법을 수행할 수 있다.

[0089] 가중치 그룹 크기 결정부(710)는 가중치 그룹 선택부(750)에서 선택될 가중치 그룹의 크기를 결정한다. 가중치 그룹의 크기는 하드웨어의 특성에 기초하여 결정될 수 있다. 예를 들어, 상기 하드웨어의 특성은 컴퓨팅 코어의 개수, 캐시 라인(cache line)의 길이, 및 SIMD(Single Instruction Multiple Data) 유닛의 개수 중 적어도 하나를 포함할 수 있다. 또는 가중치 그룹의 크기는 가중치 행렬의 모양에 기초하여 결정될 수 있다.

[0090] 로드 밸런싱 팩터 결정부(730)는 스레드의 로드 밸런싱 정도를 나타내는 로드 밸런싱 팩터를 결정한다. 로드 밸런싱 팩터(L)는 $0 \leq L \leq 1$의 값을 가지며, 상기 수학식 4와 같이 각 행의 희소성(sparsity)의 하한을 결정하고, 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수를 결정할 수 있다.

[0091] 가중치 그룹 선택부(750)는 가중치 그룹의 크기 및 로드 밸런싱 팩터에 기초하여 가중치 그룹을 선택한다. 가중치 그룹 선택부(750)는 탐욕(greedy) 알고리즘 또는 최적(optimal) 알고리즘에 따라 가중치 그룹을 선택할 수 있다.

[0092] 탐욕 알고리즘에서는 목표 희소성(sparsity)을 충족시키면서 후보 가중치 그룹들 중에서 가장 큰 규모의 가중치 그룹을 선택한다. 가중치 그룹의 각 선택 단계에서, 이전 단계에서 이미 선택된 가중치 그룹은 후보 가중치 그룹에서 제외되고, 이전 단계에서 이미 선택된 가중치 그룹과 겹치는 다른 가중치 그룹들도 후보 가중치 그룹에서 제외된다.

[0093] 최적 알고리즘에서는 1차원 공간에서 가중치 그룹을 선택하기 위해, 도 3과 같이, 목표 가중치 행렬을 (K^2C by F)에서 (K^2CF by 1)의 형태로 변환한다. 또한 최적 알고리즘에서는, 보존될 그룹의 수를 m, 변환된 행렬의 폭을 n, 목표 희소성(sparsity)을 S, 가중치 그룹의 크기를 G라고 할 때, 상기 수학식 3과 같이 정의되는 $W^k_{s,e}$에 대하여 $W^m_{1,n}$이 최대가 되는 m개의 가중치 그룹을 선택한다. 이때 동적 프로그래밍(dynamic programming)을 이용해 해를 구할 수 있다.

[0094] 가중치 그룹 선택부(750)는 로드 밸런싱 팩터(L)에 따라 상기 수학식 4에 의해 계산되는 각 행의 희소성(sparsity)의 하한에 기초하여 각 행에서 최대로 보존될 수 있는 가중치 그룹의 수만큼 각 행에서 가중치 그룹을 선택할 수 있다.

[0095] 프루닝부(770)는 선택된 가중치 그룹에 포함되지 않은 가중치 요소를 제거한다. 즉, 선택된 가중치 그룹에 포함되지 않은 가중치 요소는 영(0)으로 마스킹되며, 향후 연산의 대상에서 제외된다

[0096] 모델 재훈련부(790)는 선택된 가중치 그룹에 포함되지 않은 가중치 요소가 제거된 신층신경망(DNN) 모델을 재훈련시킨다.

[0098] 본 발명의 실시예들에 따른 그룹-레벨 프루닝 방법 및 장치의 효과를 검증하기 위해 실험을 수행하였다. 대상 하드웨어는 ARM Mali-T628 GPU가 장착된 ODROID-XU4 보드이고, 심층신경망(DNN) 모델은 LeNet5, VGG-13, NIN(Network-In-Network), AlexNet, 및 ConvNet을 대상으로 하였다. 각 모델의 첫 번째 층(layer)과 마지막 층(layer)에 대해서는 프루닝을 수행하지 않았다. ARM Mali-T628 GPU의 SIMD 유닛을 최대한 사용하기 위해 1차원 가중치 그룹의 크기는 4로 결정하였고, 2차원 가중치 그룹의 크기는 4*4로 결정하였다. 각 심층신경망(DNN) 모델에 대하여, 프루닝되지 않은 모델의 정확도 및 성능을 기준으로 하여 프루닝된 모델의 정확도 및 성능을 비교하였다. 로드 밸런싱 팩터는 다른 언급이 없으면 0으로 결정하였다.

[0099] 도 8은 MNIST 데이터에 대한 LeNet5 모델에 관한 실험결과를 나타낸 그래프이다.

[0100] 도 8을 참조하면, 정확도는 1차원 그룹-레벨 프루닝이 2차원 그룹-레벨 프루닝보다 더 높다. 이는 1차원 그룹-레벨 프루닝이 2차원 그룹-레벨 프루닝보다 더 파인-그레인드 프루닝에 가깝기 때문이다. 또한, 비정렬된 프루닝은 정렬된 프루닝에 비해 정확도가 최대 0.8% 높으며, 특히 희소성(sparsity)이 98.4%일 때 1차원 비정렬된 프루닝과 2차원 비정렬된 프루닝은 각각 1차원 정렬된 프루닝과 2차원 정렬된 프루닝에 비해 0.59%, 0.72% 높은

정확도를 가진다. 한편, 1차원 그룹-레벨 프루닝보다 2차원 그룹-레벨 프루닝에서 더 큰 정확도 차이가 생기므로, 정렬 제한은 가중치 그룹의 크기가 클수록 더 큰 정확도 손실을 불러온다는 것을 알 수 있다.

[0101] 또한 비정렬된 프루닝은 정렬된 프루닝보다 희소성(sparsity)이 더 높더라도 비슷하거나 더 높은 정확도를 보이는 것을 알 수 있다. 예를 들어, 98.3%의 희소성(sparsity)에서 2차원 비정렬된 프루닝의 정확도는 98%의 희소성(sparsity)에서 2차원 정렬된 프루닝의 정확도와 비슷하고, 98.3%의 희소성(sparsity)에서 1차원 비정렬된 프루닝의 정확도는 98%의 희소성(sparsity)에서 1차원 정렬된 프루닝의 정확도보다 높다. 특히, 1차원 비정렬된 그룹-레벨 프루닝은 98%의 희소성(sparsity)에서 프루닝 되지 않은 모델의 기준 정확도 99.38%와 비교하여 정확도 손실이 없는 것을 알 수 있다.

[0102] 도 9는 CIFAR-10 데이터에 대한 VGG-13 모델에 관한 실험결과를 나타낸 그래프이다.

[0103] 도 9의 (a)를 참조하면, 비정렬된 프루닝의 최적 알고리즘은 정렬된 프루닝 및 비정렬된 프루닝의 탐욕 알고리즘보다 더 큰 정확도를 얻는 것을 알 수 있으며, 희소성(sparsity)이 높을수록 그 차이는 더욱 커지는 것을 알 수 있다. 특히, 비정렬된 프루닝의 최적 알고리즘은 80.2%의 희소성(sparsity)에서 기준 정확도 93.57%와 비교하여 정확도 손실이 없는 것을 알 수 있다.

[0104] 도 9의 (b)는 프루닝된 모델의 추론 지연시간(inference latency)을 프루닝되지 않은 모델의 추론 지연시간(inference latency)으로 정규화한 그래프이다.

[0105] 도 9의 (b)를 참조하면, 정렬-프루닝된 모델과 비정렬-프루닝된 모델 모두 프루닝되지 않은 모델에 비하여 약 2배 빠른 성능을 나타낸다. 정렬된 프루닝과 비정렬된 프루닝 사이에 3-5%의 성능 차이가 있으나, 이는 로드 밸런싱 팩터를 고려하지 않고 가중치 그룹을 선택함에 따른 결과이며 로드 밸런싱 팩터를 고려한 경우의 결과는 도 13을 참조하여 후술한다.

[0106] 도 10은 CIFAR-10 데이터에 대한 NIN 모델에 관한 실험결과를 나타낸 그래프이다.

[0107] 도 10의 (a)를 참조하면, 1차원 프루닝은 파인-그레인드 프루닝에 가까우므로, 프루닝 방법에 따른 정확도의 차이가 크지는 않으나, 비정렬된 프루닝의 최적 알고리즘이 비정렬된 프루닝의 탐욕 알고리즘 및 정렬된 프루닝보다 더 높은 정확도를 얻는 것을 알 수 있다.

[0108] 도 10의 (b)를 참조하면, 정렬된 프루닝과 비정렬된 프루닝의 성능 차이는 크지 않은 것을 알 수 있다.

[0109] 도 11은 ImageNet 데이터에 대한 AlexNet 모델에 관한 실험결과를 나타낸 그래프이다.

[0110] AlexNet 모델은 상술한 다른 모델들보다 마지막 층(layer)의 파라미터가 많고, 본 실험에서 마지막 층(layer)에 대해서는 프루닝을 하지 않으므로 높은 희소성(sparsity)과 높은 정확도를 얻기가 어렵다. 그럼에도 불구하고, 도 11의 (a)를 참조하면, 비정렬된 프루닝이 정렬된 프루닝보다 더 높은 정확도를 얻는 것을 알 수 있다. 한편 도 11의 (b)를 참조하면, 정렬된 프루닝과 비정렬된 프루닝의 성능 차이는 2%로 크지 않은 것을 알 수 있다.

[0111] 도 12는 정렬된 프루닝과 비정렬된 프루닝의 지연시간(latency)과 캐시 미스 비율(cache miss ratio)을 비교한 그래프이다.

[0112] 본 실험은 비정렬로 인한 성능 저하를 분석하기 위한 것으로서, 도 12의 (a)를 참조하면, 비정렬된 프루닝은 정렬된 프루닝보다 약 9% 증가한 지연시간을 나타내는 것을 알 수 있다. 한편 도 12의 (b)를 참조하면, 비정렬된 프루닝은 정렬된 프루닝보다 1-5% 더 높은 캐시 미스 비율(cache miss ratio)을 나타내는 것을 알 수 있다. 이는 비정렬된 프루닝에서는 가중치 그룹의 위치가 제한되지 않으므로 해당 데이터를 로드(load)하기 위해 두 개의 캐시 라인이 소모되지 때문이다. 그러나 희소성(sparsity)이 높아질수록 캐시 미스 비율의 차이는 줄어드는 것을 알 수 있다. 즉, 희소성(sparsity)이 60%에서 90%로 변하는 동안 성능차이는 2.3ms에서 0.8ms로 줄어든다.

[0113] 도 13은 CIFAR-10 데이터에 대한 ConvNet 모델에 관하여 로드 밸런싱 팩터에 따른 정확도와 지연시간의 변화를 나타낸 그래프이다.

[0114] 본 실험에서, 기준 정확도는 78.8%이고, 2차원 비정렬된 프루닝 방법을 사용하였고, 목표 희소성은 81.6%로 설정하였다. 도 13을 참조하면, 로드 밸런싱 팩터가 증가함에 따라 정확도가 감소하고 성능이 증가하는 것을 알 수 있다. 로드 밸런싱 팩터가 1일 때 로드 밸런싱 팩터가 0인 경우에 비해 2.3%의 정확도 손실이 있다. 그러나 로드 밸런핑 팩터가 1인 경우에도 비정렬된 프루닝은 정렬된 프루닝에 비해서 더 높은 정확도를 얻을 수 있으며, 동일한 수준의 지연시간을 갖는다.

[0116] 전술한 본 발명의 일 실시예에 따른 비정렬된 그룹-레벨 프루닝 방법은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현되는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체로는 컴퓨터 시스템에 의하여 해독될 수 있는 데이터가 저장된 모든 종류의 기록매체를 포함한다. 예를 들어, ROM(Read Only Memory), RAM(Random Access Memory), 자기 테이프, 자기 디스크, 플래시 메모리, 광 데이터 저장장치 등이 있을 수 있다. 또한 컴퓨터로 판독 가능한 기록매체는 컴퓨터 통신망으로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 읽을 수 있는 코드로서 저장되고 실행될 수 있다.

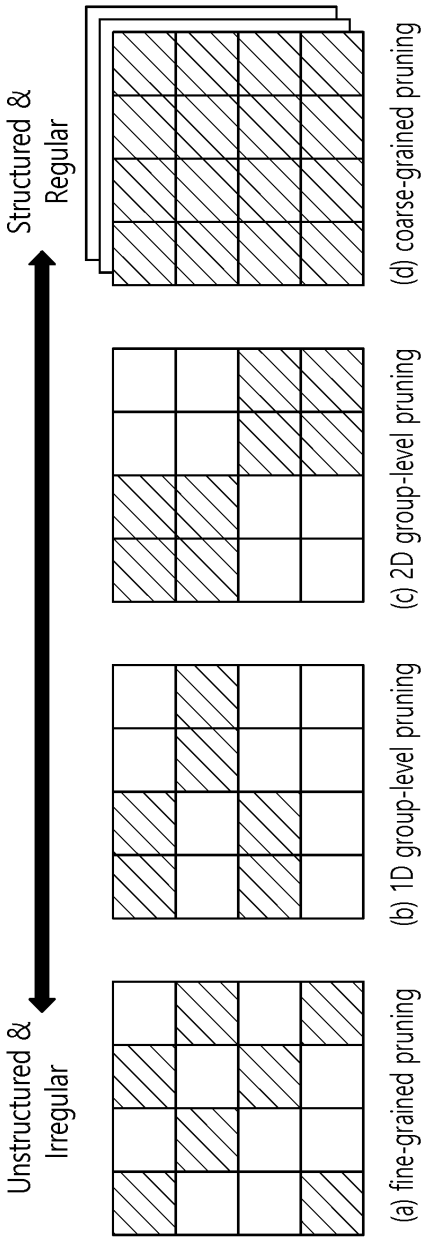
[0118] 이상에서 도면 및 실시예를 참조하여 설명하였지만, 본 발명의 보호범위가 상기 도면 또는 실시예에 의해 한정되는 것을 의미하지는 않으며 해당 기술 분야의 숙련된 당업자는 하기의 청구범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

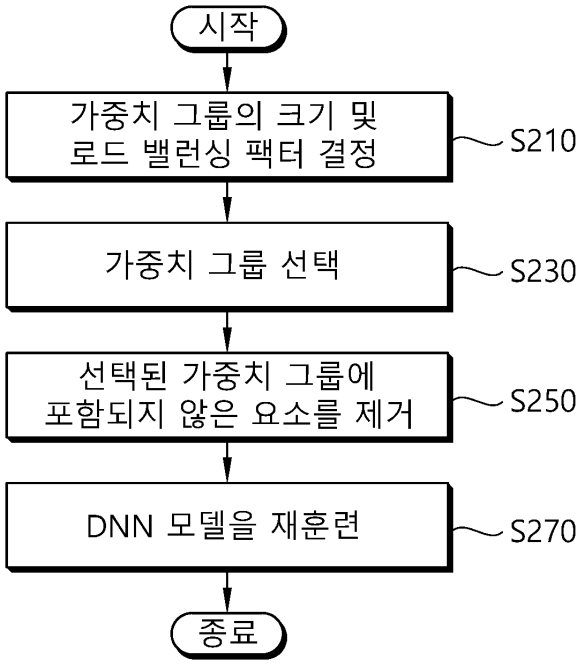
[0119] 700: 비정렬된 그룹-레벨 프루닝 장치

도면

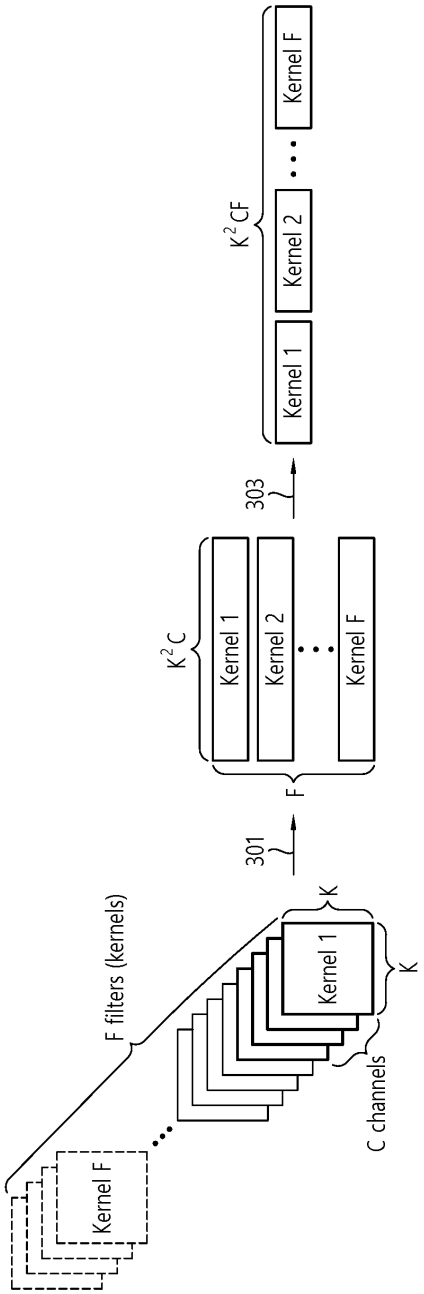
도면1


Structured &
Regular
Unstructured &
Irregular
(a) fine-grained pruning
(b) 1D group-level pruning
(c) 2D group-level pruning
(d) coarse-grained pruning

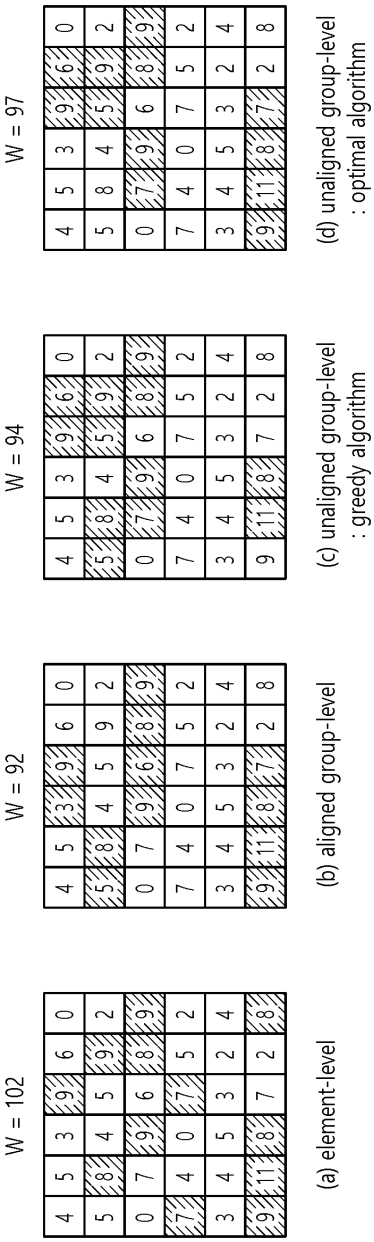
도면2


시작
가중치 그룹의 크기 및
로드 밸런싱 팩터 결정
S210
가중치 그룹 선택
S230
선택된 가중치 그룹에
포함되지 않은 요소를 제거
S250
DNN 모델을 재훈련
S270
종료

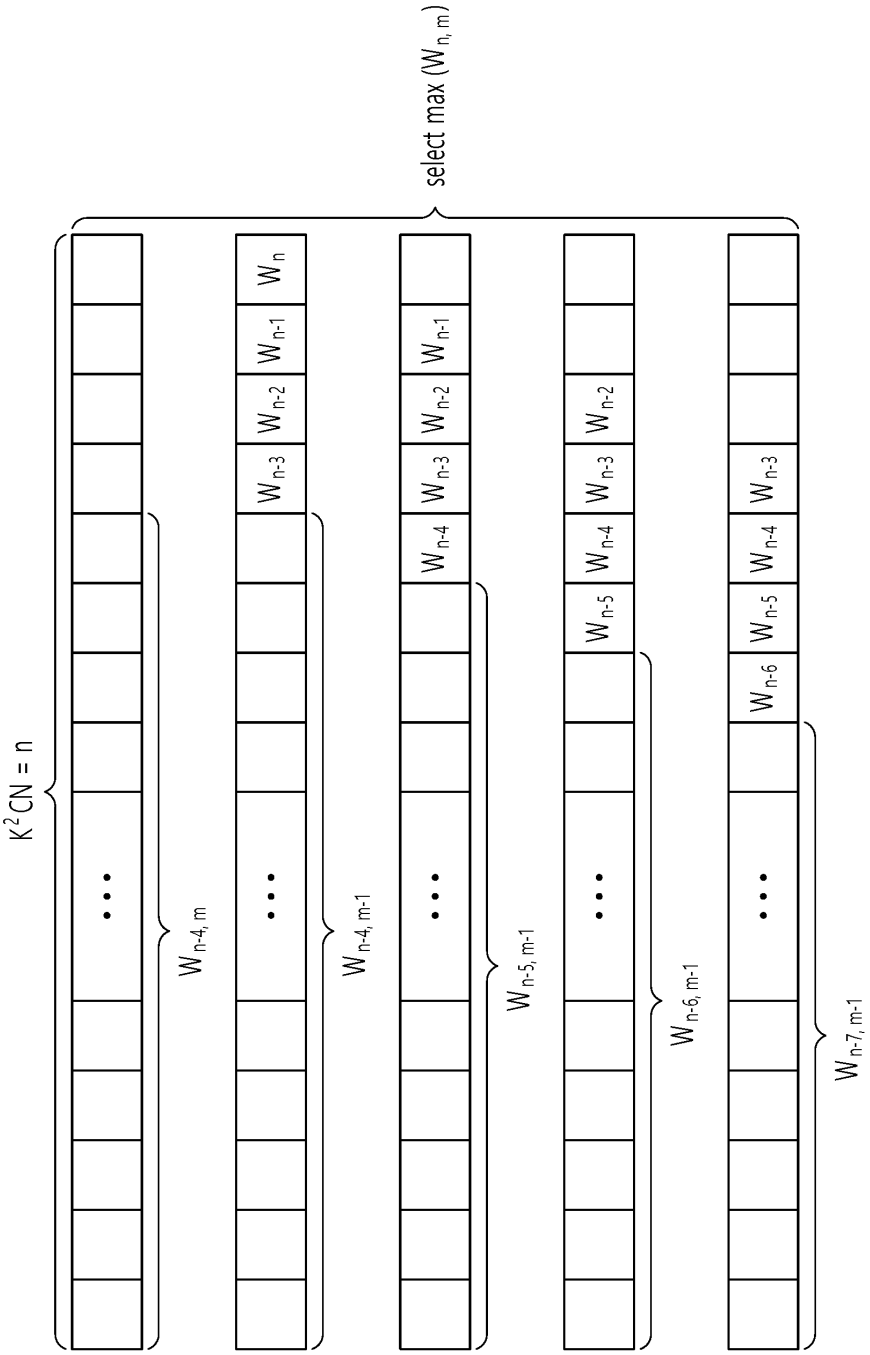
도면3


F filters (kernels)
Kernel F
C channels
K
K
Kernel 1
301
F
K²C
Kernel 1
Kernel 2
Kernel F
303
K²CF
Kernel 1
Kernel 2
Kernel F

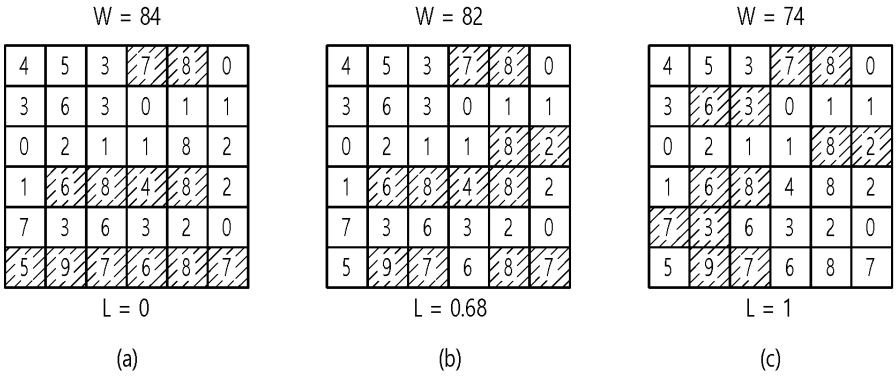
도면4


W = 102
W = 92
W = 94
W = 97
(a) element-level
(b) aligned group-level
(c) unaligned group-level
: greedy algorithm
(d) unaligned group-level
: optimal algorithm

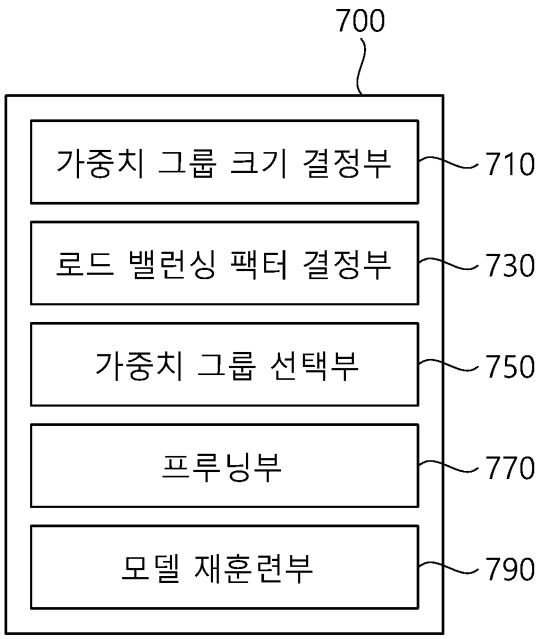
도면5


select max ($W_{n, m}$)
$K^2CN = n$
$W_{n-4, m}$
$W_{n-4, m-1}$
$W_{n-5, m-1}$
$W_{n-6, m-1}$
$W_{n-7, m-1}$
W_n
W_{n-1}
W_{n-2}
W_{n-3}
W_{n-4}
W_{n-5}
W_{n-6}

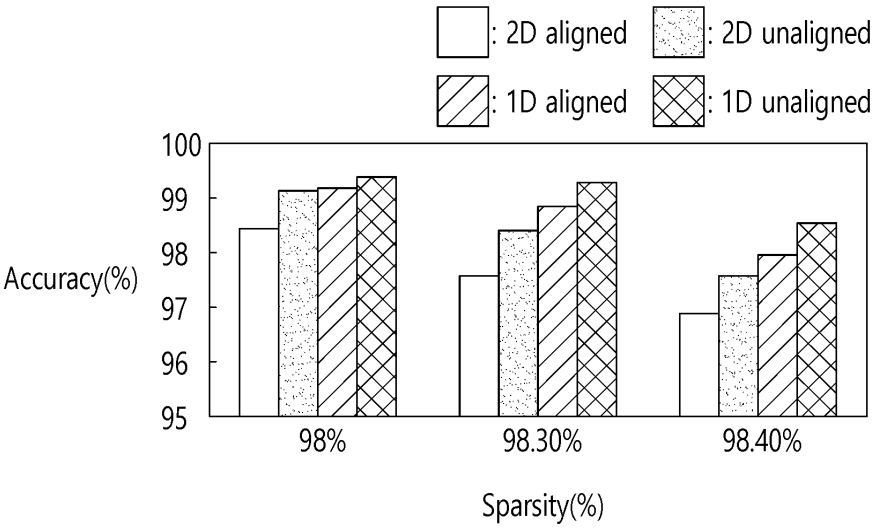
도면6


W = 84
4 5 3 7 8 0
3 6 3 0 1 1
0 2 1 1 8 2
1 6 8 4 8 2
7 3 6 3 2 0
5 9 7 6 8 7
L = 0
(a)
W = 82
4 5 3 7 8 0
3 6 3 0 1 1
0 2 1 1 8 2
1 6 8 4 8 2
7 3 6 3 2 0
5 9 7 6 8 7
L = 0.68
(b)
W = 74
4 5 3 7 8 0
3 6 3 0 1 1
0 2 1 1 8 2
1 6 8 4 8 2
7 3 6 3 2 0
5 9 7 6 8 7
L = 1
(c)

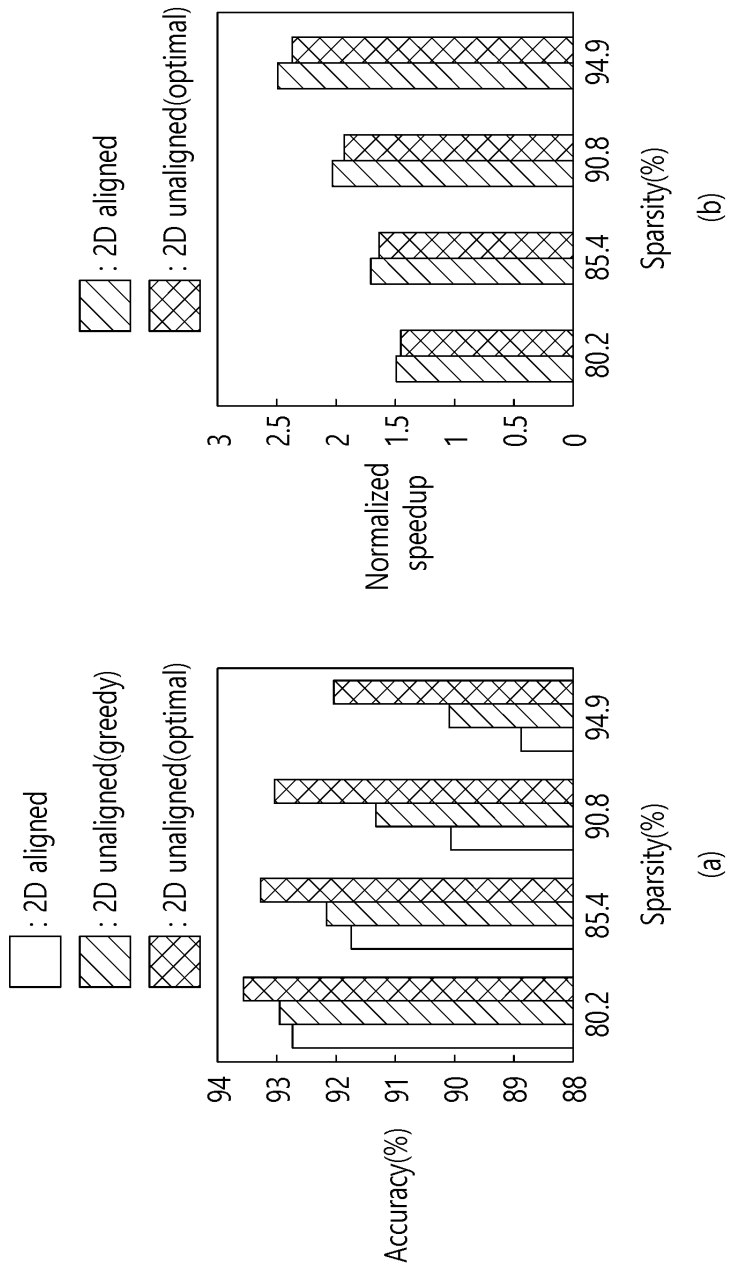
도면7


700
가중치 그룹 크기 결정부
710
로드 밸런싱 팩터 결정부
730
가중치 그룹 선택부
750
프루닝부
770
모델 재훈련부
790

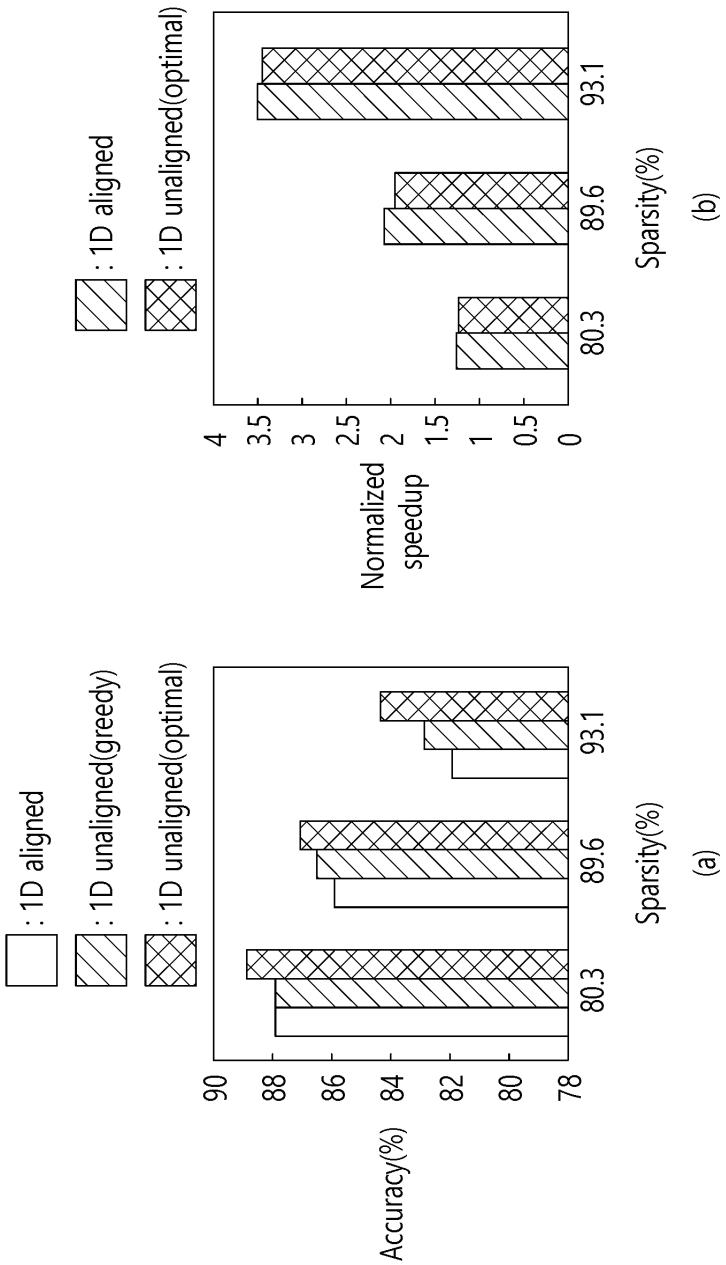
도면8


: 2D aligned
: 2D unaligned
: 1D aligned
: 1D unaligned
Accuracy(%)
100
99
98
97
96
95
98%
98.30%
98.40%
Sparsity(%)

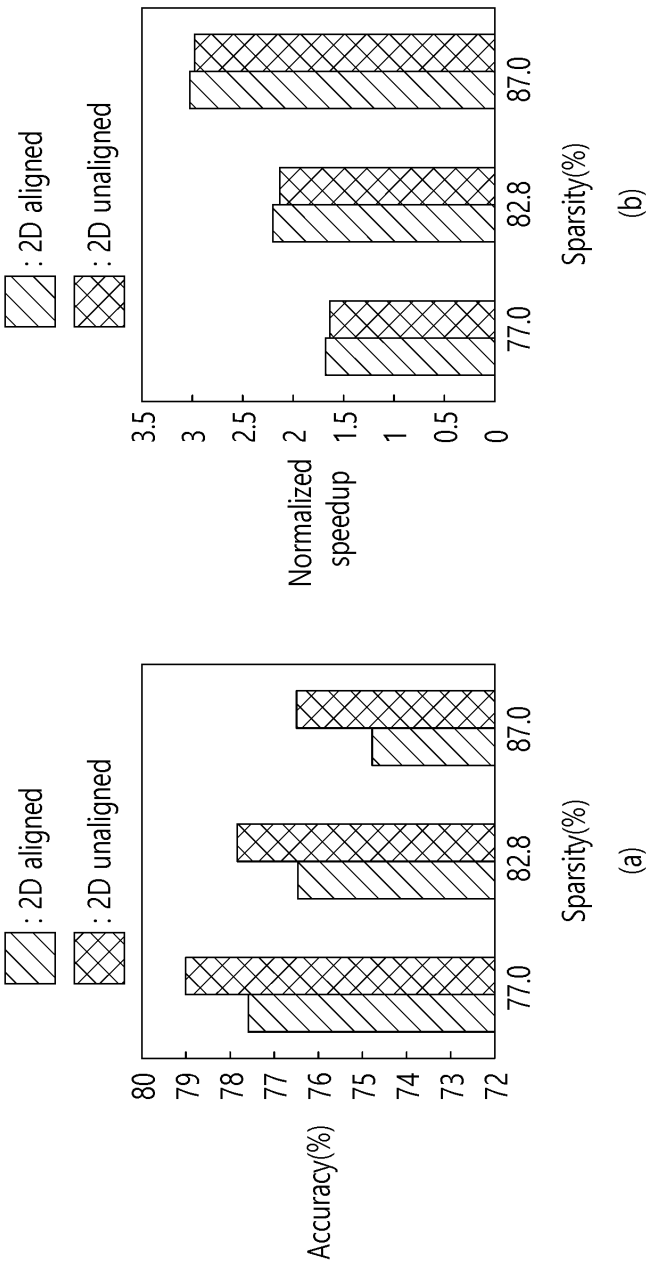
도면9


: 2D aligned
: 2D unaligned(greedy)
: 2D unaligned(optimal)
Accuracy(%)
94
93
92
91
90
89
88
80.2
85.4
90.8
94.9
Sparsity(%)
(a)
: 2D aligned
: 2D unaligned(optimal)
Normalized speedup
3
2.5
2
1.5
1
0.5
0
80.2
85.4
90.8
94.9
Sparsity(%)
(b)

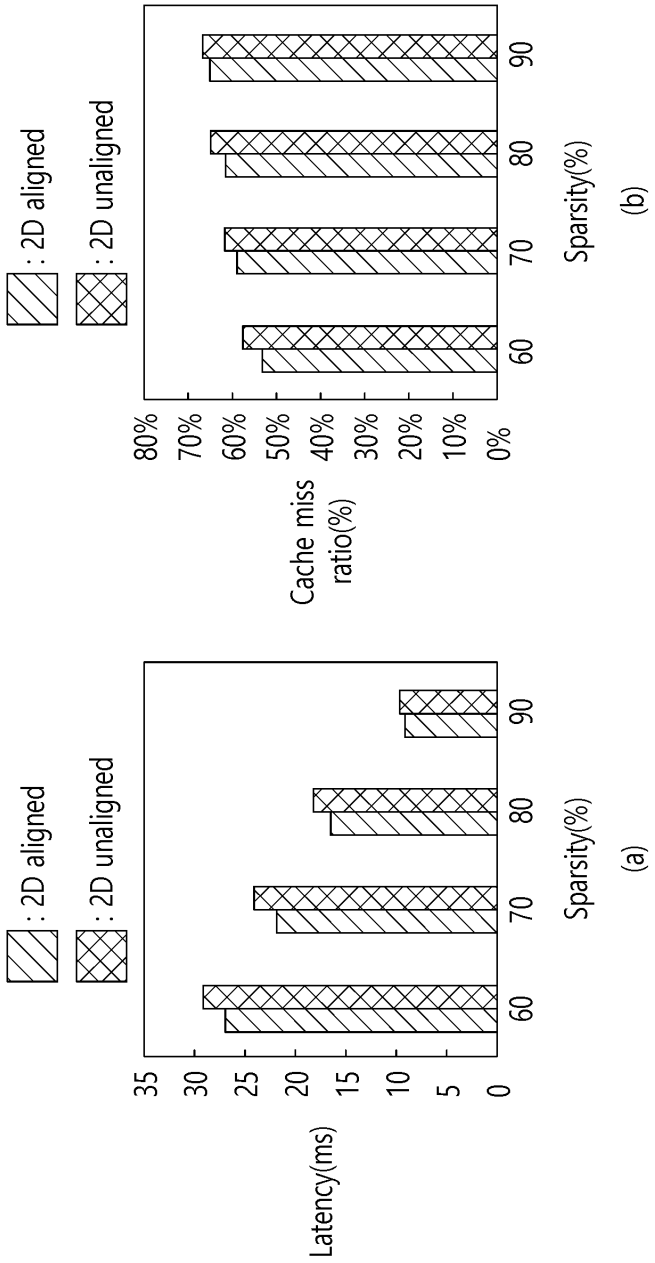
도면10


: 1D aligned
: 1D unaligned(greedy)
: 1D unaligned(optimal)
Accuracy(%)
90
88
86
84
82
80
78
80.3
89.6
93.1
Sparsity(%)
(a)
: 1D aligned
: 1D unaligned(optimal)
Normalized speedup
4
3.5
3
2.5
2
1.5
1
0.5
0
80.3
89.6
93.1
Sparsity(%)
(b)

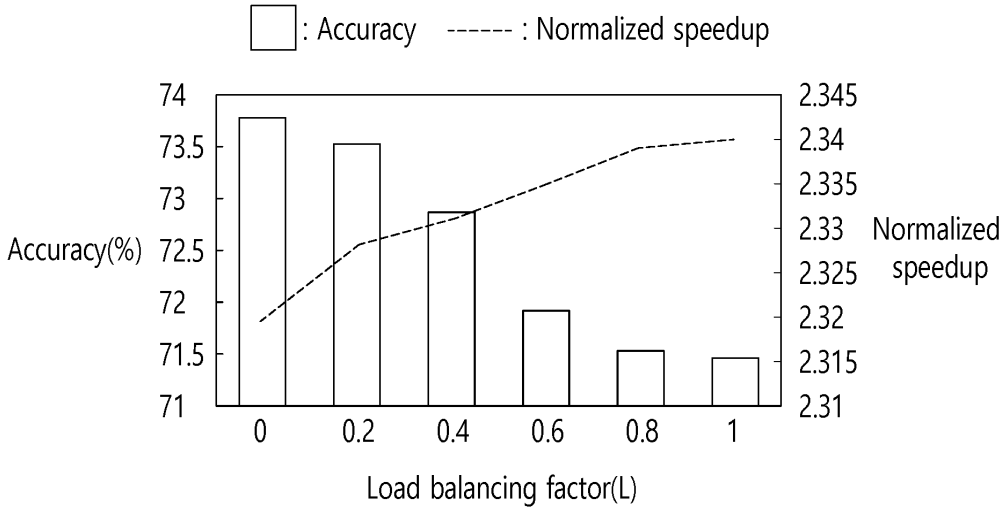
도면11


: 2D aligned
: 2D unaligned
Accuracy(%)
80
79
78
77
76
75
74
73
72
77.0
82.8
87.0
Sparsity(%)
(a)
Normalized speedup
3.5
3
2.5
2
1.5
1
0.5
0
77.0
82.8
87.0
Sparsity(%)
(b)

도면12


: 2D aligned
: 2D unaligned
Latency(ms)
35
30
25
20
15
10
5
0
60
70
80
90
Sparsity(%)
(a)
: 2D aligned
: 2D unaligned
Cache miss ratio(%)
80%
70%
60%
50%
40%
30%
20%
10%
0%
60
70
80
90
Sparsity(%)
(b)

도면13


: Accuracy
------ : Normalized speedup
Accuracy(%)
74
73.5
73
72.5
72
71.5
71
2.345
2.34
2.335
2.33
2.325
2.32
2.315
2.31
Normalized speedup
0
0.2
0.4
0.6
0.8
1
Load balancing factor(L)