



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0077592
(43) 공개일자 2024년06월03일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) G06N 3/04 (2023.01)
(52) CPC특허분류
G06N 3/082 (2023.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2022-0158832
(22) 출원일자 2022년11월24일
심사청구일자 2022년11월24일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
이연주
서울특별시 마포구 월드컵로11길 35, 101동 504호(망원동, 함성월드빌 2차아파트)
서재규
인천광역시 부평구 갈월동로 45, 101동 501호(갈산동, 두산아파트)
(74) 대리인
송인호, 최관락

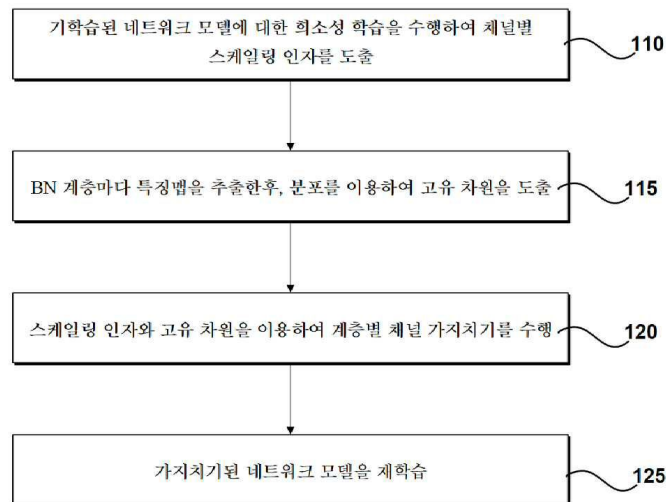
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 고유 차원 기반 계층별 채널 가지치기 방법 및 그 장치

(57) 요약

고유 차원 기반 계층별 채널 가지치기 방법 및 그 장치가 개시된다. 고유 차원 기반 계층별 채널 가지치기 방법은, (a) 기 학습된 타겟 네트워크 모델을 희소성 학습하여 스케일링 인자를 도출하는 단계; (b) 상기 타겟 네트워크 모델의 적어도 일부에서 특징맵을 추출하고, 상기 추출된 특징맵의 분포를 이용하여 고유 차원을 계산하는 단계-상기 특징맵은 학습 데이터를 상기 타겟 네트워크 모델에 적용하여 획득됨; (c) 상기 스케일링 인자와 상기 고유 차원을 이용하여 상기 타겟 네트워크 모델의 계층마다 서로 다른 비율로 가지치기(pruning)을 수행하는 단계; 및 (d) 상기 학습 데이터를 이용하여 상기 타겟 네트워크 모델을 재학습시키는 단계를 포함한다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	1345347042
과제번호	2020R1A6A1A03038540
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	이공학학술연구기반구축
연구과제명	자율지능무인비행체연구소
기 여 율	1/1
과제수행기관명	세종대학교 산학협력단
연구기간	2022.03.01 ~ 2023.02.28

명세서

청구범위

청구항 1

- (a) 기 학습된 타겟 네트워크 모델을 최소성 학습하여 스케일링 인자를 도출하는 단계;
- (b) 상기 타겟 네트워크 모델의 적어도 일부에서 특징맵을 추출하고, 상기 추출된 특징맵의 분포를 이용하여 고유 차원을 계산하는 단계-상기 특징맵은 학습 데이터를 상기 타겟 네트워크 모델에 적용하여 획득됨;
- (c) 상기 스케일링 인자와 상기 고유 차원을 이용하여 상기 타겟 네트워크 모델의 계층마다 서로 다른 비율로 가지치기(pruning)을 수행하는 단계; 및
- (d) 상기 학습 데이터를 이용하여 상기 타겟 네트워크 모델을 재학습시키는 단계를 포함하는 고유 차원 기반 계층별 채널 가지치기 방법.

청구항 2

제1 항에 있어서,

상기 최소성 학습은 스케일링 인자들의 값들을 제로(0)가 되도록 하는 학습으로, 손실 함수는 하기 수식과 같이 계산되는 것을 특징으로 하는 고유 차원 기반 계층별 채널 가지치기 방법.

$$L = \sum_{(x,y)} l(f(x, W), y) + \lambda \sum_{\gamma \in \Gamma} |\gamma|$$

여기서, (x, y) 는 학습 데이터의 입력과 타겟을 나타내고, W 는 가중치를 나타내며, γ 는 배치 정규화

계층의 스케일 파라미터를 나타내고, λ 는 밸런스 파라미터를 나타내고, $||$ 는 L1 norm을 나타내며,

$$\sum_{(x,y)} l(f(x, W), y)$$

는 상기 타겟 네트워크 모델의 손실 함수를 나타냄.

청구항 3

제1 항에 있어서,

상기 고유 차원은 상기 타겟 네트워크 모델의 백본(backbone) 부분과 넥(Neck) 부분에 포함된 배치 정규화 계층에서 추출된 특징맵들의 분포를 이용하여 계산되는 것을 특징으로 하는 고유 차원 기반 계층별 채널 가지치기 방법.

청구항 4

제1 항에 있어서,

상기 (c) 단계는,

상기 고유 차원을 이용하여 가지치기 비율을 도출하는 단계;

상기 가지치기 비율을 이용하여 스케일링 인자에 대한 임계값을 결정하는 단계; 및

상기 임계값에 기초하여 상기 타겟 네트워크 모델의 계층별 채널 가지치기를 수행하는 단계를 포함하는 고유 차원 기반 계층별 채널 가지치기 방법.

청구항 5

제4 항에 있어서,

상기 가지치기 비율은 하기 수학식을 이용하여 도출되는 것을 특징으로 하는 고유 차원 기반 계층별 채널 가지치기 방법.

$$R = \frac{N_{ch} - \alpha \sum_{i \in I} ID_i}{N_{ch}} \times 100$$

여기서, N_{ch} 는 BN 계층에 포함된 채널 개수의 합을 나타내고, ID_i 는 i번째 BN 계층의 고유 차원 값을 나타내고, α 는 상수값을 나타냄.

청구항 6

제4 항에 있어서,

상기 계층별 채널 가지치기를 수행하는 단계는,

계층마다 상기 임계값에 따라 프루닝 마스크(pruning mask)를 생성하는 단계-상기 프루닝 마스크는 남겨질 채널과 제거될 채널을 채널 인덱스에 따라 제1 값과 제2 값으로 포함함; 및

상기 프루닝 마스크를 이용하여 계층별로 채널 가지치기를 수행하는 단계를 포함하는 고유 차원 기반 계층별 채널 가지치기 방법.

청구항 7

제1 항에 있어서,

상기 (d) 단계 이전에,

상기 타겟 네트워크 모델이 잔차 블록을 포함하는 경우, 잔차 블록내의 합성곱 계층의 채널 수와 인덱스를 잔차 블록 전단에 위치한 합성곱 계층의 채널 수 및 인덱스와 일치시키는 단계를 더 포함하는 것을 특징으로 하는 고유 차원 기반 계층별 채널 가지치기 방법.

청구항 8

제1 항 내지 제7 항 중 어느 하나의 항에 따른 방법을 수행하기 위한 프로그램 코드를 기록한 컴퓨터로 판독 가능한 기록매체.

청구항 9

적어도 하나의 명령어를 저장하는 메모리; 및

상기 메모리에 저장된 명령어를 실행하는 프로세서를 포함하되,

상기 프로세서에 의해 실행된 명령어는,

기 학습된 타겟 네트워크 모델을 희소성 학습하여 스케일링 인자를 도출하는 단계;

상기 타겟 네트워크 모델의 적어도 일부에서 특징맵을 추출하고, 상기 추출된 특징맵의 분포를 이용하여 고유 차원을 계산하는 단계-상기 특징맵은 학습 데이터를 상기 타겟 네트워크 모델에 적용하여 획득됨;

상기 스케일링 인자와 상기 고유 차원을 이용하여 상기 타겟 네트워크 모델의 계층마다 서로 다른 비율로 가지치기(pruning)을 수행하는 단계; 및

상기 학습 데이터를 이용하여 상기 타겟 네트워크 모델을 재학습시키는 단계를 실행하는 것을 특징으로 하는 컴퓨팅 장치.

청구항 10

제9 항에 있어서,

상기 고유 차원은 상기 타겟 네트워크 모델의 백본(backbone) 부분과 넥(Neck) 부분에 포함된 배치 정규화 계층에서 추출된 특징맵들의 분포를 이용하여 계산되는 것을 특징으로 하는 컴퓨팅 장치.

청구항 11

제9 항에 있어서,

상기 계층별 채널 가지치기를 수행하는 단계는,

상기 고유 차원을 이용하여 가지치기 비율을 도출하는 단계;

상기 가지치기 비율을 이용하여 스케일링 인자에 대한 임계값을 결정하는 단계; 및

상기 임계값에 기초하여 상기 타겟 네트워크 모델의 계층별 채널 가지치기를 수행하는 단계를 더 실행하는 것을 특징으로 하는 컴퓨팅 장치.

청구항 12

제11 항에 있어서,

상기 계층별 채널 가지치기를 수행하는 단계는,

계층마다 상기 임계값에 따라 프루닝 마스크(pruning mask)를 생성하는 단계-상기 프루닝 마스크는 남겨질 채널과 제거될 채널을 채널 인덱스에 따라 제1 값과 제2 값으로 포함함; 및

상기 프루닝 마스크를 이용하여 계층별로 채널 가지치기를 수행하는 단계를 더 실행하는 것을 특징으로 하는 컴퓨팅 장치.

발명의 설명

기술 분야

[0001] 본 발명은 고유 차원 기반 계층별 채널 가지치기 방법 및 그 장치에 관한 것이다.

배경 기술

[0003] 높은 정확도를 갖는 CNN기반 객체 검출 네트워크는 리소스(resource)가 제한된 임베디드 환경에서 활용되기 어렵기 때문에 네트워크 경량화가 필요하다. 가장 널리 사용되는 경량화 방법으로는 네트워크 가지치기(network pruning)가 있으며, 구현의 편리성과 가속화의 가능성 측면에서 구조적 가지치기(structured pruning)가 선호된다.

[0004] 구조적 가지치기는 채널 혹은 필터와 같이 묶음 단위로 불필요한 것을 제거하는 방식인데, 각 묶음 단위로 중요도를 구하고 중요도가 낮은 묶음부터 순차적으로 제거한다. 제거 방식으로는 전역적 임계값(global threshold)을 사용하는 방법과 층(layer)마다 다른 임계값을 사용하는 방법이 있다.

[0005] 전역적 가지치기(global pruning)는 구현이 간단하다는 장점이 있지만, 계층에 상관없이 채널 혹은 필터에만 집중하기 때문에 층의 역할을 감소시켜 네트워크의 성능이 다소 손실될 우려가 있다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 고유 차원 기반 계층별 채널 가지치기 방법 및 그 장치를 제공하기 위한 것이다.

[0008] 또한, 본 발명은 성능이 좋은 네트워크 모델에서 불필요한 연결을 제거하여 본래 네트워크 모델의 성능이 크게 손상되지 않는 범위 내에서 네트워크 크기를 줄일 수 있는 고유 채널 기반 계층별 채널 가지치기 방법 및 그 장치를 제공하기 위한 것이다.

[0009] 또한, 본 발명은 각 채널의 희소성 학습을 통해 획득된 채널 중요도와 각 계층별 고유의 파라미터 수가 유지될 수 있도록 고유 차원을 가지치기에 활용하여 네트워크 성능 손실을 최소화하고 네트워크 경량화 효율성을 높일 수 있는 고유 채널 기반 계층별 채널 가지치기 방법 및 그 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0011] 본 발명의 일 측면에 따르면 고유 차원 기반 계층별 채널 가지치기 방법이 제공된다.

[0012] 본 발명의 일 실시예에 따르면, (a) 기 학습된 타겟 네트워크 모델을 희소성 학습하여 스케일링 인자를 도출하는 단계; (b) 상기 타겟 네트워크 모델의 적어도 일부에서 특징맵을 추출하고, 상기 추출된 특징맵의 분포를 이용하여 고유 차원을 계산하는 단계-상기 특징맵은 학습 데이터를 상기 타겟 네트워크 모델에 적용하여 획득됨; (c) 상기 스케일링 인자와 상기 고유 차원을 이용하여 상기 타겟 네트워크 모델의 계층마다 서로 다른 비율로 가지치기(pruning)를 수행하는 단계; 및 (d) 상기 학습 데이터를 이용하여 상기 타겟 네트워크 모델을 재학습시키는 단계를 포함하는 채널 가지치기 방법이 제공될 수 있다.

[0013] 상기 희소성 학습은 스케일링 인자들의 값들을 제로(0)가 되도록 하는 학습으로, 손실 함수는 하기 수식식과 같이 계산되며,

$$L = \sum_{(x,y)} l(f(x, W), y) + \lambda \sum_{\gamma \in \Gamma} |\gamma|$$

[0014]

[0015] 여기서, (x, y) 는 학습 데이터의 입력과 타겟을 나타내고, W 는 가중치를 나타내며, γ 는 배치 정규화

계층의 스케일 파라미터를 나타내고, λ 는 밸런스 파라미터를 나타내고, $||$ 는 L1 norm을 나타낸다.

[0016] 상기 고유 차원은 상기 타겟 네트워크 모델의 백본(backbone) 부분과 넥(Neck) 부분에 포함된 배치 정규화 계층에서 추출된 특징맵들의 분포를 이용하여 계산될 수 있다.

[0017] 상기 (c) 단계는, 상기 고유 차원을 이용하여 가지치기 비율을 도출하는 단계; 상기 가지치기 비율을 이용하여 스케일링 인자에 대한 임계값을 결정하는 단계; 및 상기 임계값에 기초하여 상기 타겟 네트워크 모델의 계층별 채널 가지치기를 수행하는 단계를 더 포함할 수 있다.

[0018] 상기 가지치기 비율은 하기 수학식을 이용하여 도출되되,

$$R = \frac{N_{ch} - \alpha \sum_{i \in I} ID_i}{N_{ch}} \times 100$$

[0019]

[0020] 여기서, N_{ch} 는 BN 계층에 포함된 채널 개수의 합을 나타내고, ID_i 는 i번째 BN 계층의 고유 차원 값을 나타내고, α 는 상수값을 나타낸다.

[0021] 상기 계층별 채널 가지치기를 수행하는 단계는, 계층마다 상기 임계값에 따라 프루닝 마스크(pruning mask)를 생성하는 단계-상기 프루닝 마스크는 남겨질 채널과 제거될 채널을 채널 인덱스에 따라 제1 값과 제2 값으로 포함함; 및 상기 프루닝 마스크를 이용하여 계층별로 채널 가지치기를 수행하는 단계를 더 포함할 수 있다.

[0023] 본 발명의 다른 측면에 따르면, 고유 차원 기반 계층별 채널 가지치기 방법을 실행하는 장치가 제공될 수 있다.

[0024] 본 발명의 일 실시예에 따르면, 적어도 하나의 명령어를 저장하는 메모리; 및 상기 메모리에 저장된 명령어를 실행하는 프로세서를 포함하되, 상기 프로세서에 의해 실행된 명령어는, 기 학습된 타겟 네트워크 모델을 희소성 학습하여 스케일링 인자를 도출하는 단계; 상기 타겟 네트워크 모델의 적어도 일부에서 특징맵을 추출하고, 상기 추출된 특징맵의 분포를 이용하여 고유 차원을 계산하는 단계-상기 특징맵은 학습 데이터를 상기 타겟 네트워크 모델에 적용하여 획득됨; 상기 스케일링 인자와 상기 고유 차원을 이용하여 상기 타겟 네트워크 모델의 계층마다 서로 다른 비율로 가지치기(pruning)을 수행하는 단계; 및 상기 학습 데이터를 이용하여 상기 타겟 네트워크 모델을 재학습시키는 단계를 실행하는 것을 특징으로 하는 컴퓨팅 장치가 제공될 수 있다.

발명의 효과

[0026] 본 발명의 일 실시예에 따른 고유 차원 기반 계층별 채널 가지치기 방법 및 그 장치를 제공함으로써, 성능이 좋은 네트워크 모델에서 불필요한 연결을 제거하여 본래 네트워크 모델의 성능이 크게 손상되지 않는 범위 내에서 네트워크 크기를 줄일 수 있다.

[0027] 또한, 본 발명은 각 채널의 희소성 학습을 통해 획득된 채널 중요도와 각 계층별 고유의 파라미터 수가 유지될 수 있도록 고유 차원을 가지치기에 활용하여 네트워크 성능 손실을 최소화하고 네트워크 경량화 효율성을 높일 수 있는 이점이 있다.

도면의 간단한 설명

[0029] 도 1은 본 발명의 일 실시예에 따른 고유 차원 기반 계층별 채널 가지치기 방법을 나타낸 순서도.

도 2는 본 발명의 일 실시예에 따른 계층별 채널 가지치기 일 예를 도시한 도면.

도 3은 본 발명의 일 실시예에 따른 잔차 블록에서의 가지치기 과정을 설명하기 위해 도시한 도면.

도 4는 본 발명의 일 실시예에 따른 VOC 데이터에 대한 평가 결과(mAP)를 나타낸 도면.

도 5는 본 발명의 일 실시예에 따른 VOC 데이터에서 AND 방법을 적용하여 가지치기한 경우 성능을 나타낸 도면.

도 6은 본 발명의 일 실시예에 따른 VisDrone 데이터 셋에 대한 평가 결과(mAP)를 나타낸 도면.

도 7은 본 발명의 일 실시예에 따른 VisDrone 데이터에서 AND 방법을 적용하여 가지치기한 경우 성능을 나타낸 도면.

도 8은 종래와 본 발명의 일 실시예에 따른 성능을 비교한 결과를 도시한 도면.

도 9는 본 발명의 일 실시예에 따른 고유 차원 기반 채널 가지치기 방법을 수행하기 위한 컴퓨팅 장치의 내부 구성을 개략적으로 도시한 블록도.

발명을 실시하기 위한 구체적인 내용

[0030] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 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.

[0031] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0033] 도 1은 본 발명의 일 실시예에 따른 고유 차원 기반 계층별 채널 가지치기 방법을 나타낸 순서도이고, 도 2는 본 발명의 일 실시예에 따른 계층별 채널 가지치기 일 예를 도시한 도면이고, 도 3은 본 발명의 일 실시예에 따른 잔차 블록에서의 가지치기 과정을 설명하기 위해 도시한 도면이다.

[0034] 단계 110에서 컴퓨팅 장치(100)는 기학습된 네트워크 모델에 대한 희소성 학습을 수행하여 각 채널에 대한 스케일링 인자를 획득한다.

[0035] 본 발명의 일 실시예에 따르면, 컴퓨팅 장치(100)는 네트워크 모델의 각 채널에 곱해지는 스케일링 인자(scaling factor)를 학습하여 채널의 중요도를 식별한다. 여기서, 네트워크 모델은 합성곱 기반 네트워크 모델인 것을 가정하기로 한다. 예를 들어, 네트워크 모델은 YOLOv4일 수 있다. 이해와 설명의 편의를 도모하기 위해 네트워크 모델이 YOLOv4인 것을 가정하여 이를 중심으로 설명하나 이외의 합성곱 기반 모델인 경우에도 동일하게 적용될 수 있음은 당연하다.

[0036] 배치 정규화 계층(batch normalization layer, 이하 BN 계층이라 칭하기로 함)의 스케일(scale) 파라미터인

γ 를 스케일링 인자로 사용할 수 있다.

[0037] CNN 모델들은 통상적으로 BN 계층을 포함하며, BN 계층은 수렴 속도와 일반화(generalization) 성능을 향상시키는 것으로 알려져 있다. BN 계층은 수학식 1과 같이 작은 배치(mini-batch) 단위의 통계를 이용하여 입력되는 값들을 정규화하고, 이를 선형 변환하여 출력 값을 도출할 수 있다.

수학식 1

$$z_o = \gamma \frac{z_i - \mu_B}{\sqrt{\sigma_B^2 + c}} + \beta$$

[0038]

[0039] 여기서, z_i 는 입력 값을 나타내고, z_o 는 출력 값을 나타내며, μ_B 와 σ_B 는 각각 작은 배치에서 얻은 입력 값들의 평균과 표준 편차를 나타낸다. 또한, γ 과 β 는 각각 스케일과 이동을 담당하는 학습 파라미터이다. 여기서, γ 는 특정 채널과 대응되어 출력의 크기에 영향을 주기 때문에 채널의 중요도를 식별하는데 유용하다. 즉, 작은 스케일링 인자를 갖는 채널들은 네트워크에서 의미가 없는 채널로 선별되고 가지치기 단계에서 제거될 수 있다.

[0040] 이러한, 스케일링 인자들은 네트워크 모델의 가중치들과 함께 학습되는데, 이를 희소성 학습이라고 한다. 이러한 희소성 학습에 대해서는 'Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huan, Shoumeng Yan, and Changshui Zhang, Learning Efficient Convolutional Networks through Network Slimming, 2017 IEEE International Conference on Computer Vision, 2017'에 공지되어 있으며, 본 발명의 일 실시예에서도 이와 같이 공지된 방법을 이용하여 기학습된 네트워크 모델에 대한 희소성 학습을 수행할 수 있다. 물론, 이외에도 공지된 희소성 학습을 위한 방법들은 모두 적용될 수 있음은 당연하다.

[0041] 컴퓨팅 장치(100)는 희소성 학습을 수행함에 있어 네트워크 모델의 배치 정규화 계층(batch normalization layer)에 있는 스케일링 인자(scaling factor)들의 값을 '0'이 되도록 만든다. 배치 정규화 계층의 각 스케일링 인자는 전단에 위치한 합성곱 계층의 한 채널에 대응되며, 해당 채널의 특징값에 곱하여 출력값을 도출한다.

[0042] 따라서, 컴퓨팅 장치(100)는 희소성 학습을 통해 획득된 스케일링 인자의 값이 크면 대응되는 채널이 의미가 있는 것으로 식별할 수 있다. 반면, 컴퓨팅 장치(100)는 희소성 학습을 통해 획득된 스케일링 인자의 값이 '0'에 가까울수록 대응되는 채널이 의미가 없는 것으로 식별할 수 있다.

[0043] 컴퓨팅 장치(100)는 네트워크 모델의 손실 함수 뒤에 스케일링 인자에 대한 L1 정규화(regularization)를 더한 손실 함수로 네트워크 모델에 대한 희소성 학습을 수행할 수 있다. 이를 수학식으로 나타내면 수학식 2와 같다.

수학식 2

$$L = \sum_{(x,y)} l(f(x, W), y) + \lambda \sum_{\gamma \in \Gamma} |\gamma|$$

[0044]

$$\sum_{(x,y)} l(f(x, W), y)$$

[0045] 여기서, 첫번째 항($\sum_{(x,y)} l(f(x, W), y)$)은 네트워크 모델의 손실 함수를 나타내고, 두번째 항은

스케일링 인자에 대한 희소성을 유도하기 위한 패널티(penalty)이다. 또한, (x, y) 는 학습 데이터의 입력

과 타겟을 나타내고, W 는 가중치를 나타내며, λ 는 두 부분의 균형을 맞춰주는 파라미터를 나타낸다.

[0046] 상술한 바와 같이, 희소성 학습과정에서 배치 정규화 계층의 파라미터인 스케일링 인자는 네트워크의 가중치들과 함께 최적화되며, 네트워크의 일반화 성능에 큰 영향을 주지 않으며 자동적으로 불필요한 의미 없는 채널들을 식별해낼 수 있다.

[0047] 단계 115에서 컴퓨팅 장치(100)는 BN 계층마다 특징맵을 추출하고, 추출된 특징맵의 분포를 이용하여 고유 차원(ID)를 도출한다. 지면 설명의 한계로 인해, 단계 115가 단계 110 이후에 수행되는 것으로 기술되고 있으나, 단계 110과 단계 115는 병렬로 수행될 수도 있다.

[0048] YOLOv4는 backbone-neck-head로 구성되는데, 가장치기는 head를 제외한 나머지 영역에서 수행될 수 있다. 컴퓨

팅 장치(100)는 네트워크 모델의 backbone과 neck 부분에 속해 있는 BN 계층마다 특징맵을 추출하고, 특징 맵의 분포를 계산하여 고유차원을 계산할 수 있다.

[0049] 특징맵들은 학습 데이터로부터 추출되는데, 고유차원 값이 학습 데이터의 특성에 의존적이기 때문에 여러 차례 고유차원을 도출한 후 이들의 평균값을 사용하여 계산될 수 있다. 이와 같이 도출된 고유차원은 각 계층에서 추출된 특징 맵의 분포를 표현하는데 필요한 최소의 파라미터수가 된다. 본 발명의 일 실시예에서는 고유 차원을 계층에서 필요한 채널의 수로 사용할 수 있다.

[0050] 단계 120에서 컴퓨팅 장치(100)는 스케일링 인자와 고유 차원을 이용하여 계층별 채널 가지치기를 수행한다.

[0051] 컴퓨팅 장치(100)는 고유 차원을 그대로 사용하지 않고, 고유 차원을 이용하여 가지치기 비율을 도출한 후 네트워크 가지치기를 수행할 수 있다. 즉, 컴퓨팅 장치(100)는 가지치기 비율을 고려하여 고유 차원에 상수(α)를 곱한 값을 계층에서 남길 채널의 개수로 결정할 수 있다.

[0052] 따라서, 네트워크의 가지치기 비율은 가지치기할 계층의 고유 차원 값들과 α 값에 따라 결정될 수 있다. 이를 수학적식으로 나타내면 수학적식 3과 같다.

수학적식 3

$$R = \frac{N_{ch} - \alpha \sum_{i \in I} ID_i}{N_{ch}} \times 100$$

[0053]

[0054] 여기서, N_{ch} 는 BN 계층에 포함된 채널 개수의 합을 나타내고, ID_i 는 i번째 BN 계층의 고유 차원 값을 나타내고, α 는 상수값이다. 즉, α 값이 크면 가지치기 비율이 낮아지고, 작으면 가지치기 비율이 높아진다.

[0055] α 가 결정되면, 남겨질 채널의 수가 결정되고 이를 토대로 불필요한 채널을 판단하기 위해 스케일링 인자의 임계값(θ)을 결정할 수 있다. 컴퓨팅 장치(100)는 각 계층마다 임계값에 따라 제거할 채널과 남겨질 채널을 채널 인덱스에 따라 0과 1로 나타낸 프루닝 마스크(pruning mask)를 생성한다.

[0056] 컴퓨팅 장치(100)는 각 계층마다 생성된 프루닝 마스크를 적용하여 계층별 채널 가지치기를 수행한다.

[0057] 도 2에는 계층별 채널 가지치기를 수행하는 일련의 과정을 도시한 도면이다. 도 2를 참조하여 이에 대해 설명하기로 한다.

[0058] 도 2에 도시된 바와 같이, 회소성 학습을 통해 각 채널에 대한 스케일링 인자가 도출되면, 고유 차원(ID)를 활용하여 프루닝 마스크를 생성한 후 계층에 대한 채널 가지치기를 수행할 수 있다. 즉, 임계값 이하의 스케일링 인자를 가지는 채널들을 가지치기할 수 있다.

[0059] 다만, YOLOv4와 같이 잔차 블록(residual block)을 포함하는 네트워크는 가지치기시 주의를 필요로 한다. 잔차 블록은 숏컷 연결(shortcut connection)과의 add 연산을 포함한다. 이러한, add 연산을 수행하기 위해서는 입력

되는 두 특징맵의 차원이 일치되어야 한다.

- [0060] 도 3을 참조하기로 한다. 도 3에서 310의 합성곱 블록들과 같이 잔차 블록에 입력되기 전의 합성곱 계층과 잔차 블록내의 add 연산 바로 전에 위치한 합성곱 계층에서 가지치기되는 채널의 수와 인덱스가 동일해야 한다.
- [0061] 이에, 본 발명의 일 실시예에 따르면 컴퓨팅 장치(100)는 잔차 블록을 포함하는 경우, 채널 수와 인덱스를 일치시키는 과정을 수행할 수 있다. 채널 수와 인덱스를 일치시키는 방법은 다음과 같다. 첫째, 가장 많이 사용되는 방법으로 일치시켜야 하는 계층들의 프루닝 마스크를 OR 연산하여 일치되어야 하는 계층들에 동일하게 적용하는 방법이 있다. 또한, 두번째, AND 연산을 수행하는 방법이 있으며, 마지막으로 FR(Front) 방법으로, 맨 앞에 위치한 숏컷 계층의 프루닝 마스크를 그대로 적용하는 방법이 존재한다. 이외에도, 채널 수와 인덱스를 일치시키는 공지된 방법들을 모두 적용될 수 있다.
- [0062] 단계 125에서 컴퓨팅 장치(100)는 가지치기된 네트워크 모델을 재학습시킨다.
- [0063] 채널 가지치기에 의해 기본 네트워크 모델에서 연결들이 임의로 가지치기 된다. 가지치기 비율이 작으면 끊어진 연결이 전체 네트워크 모델에 영향을 주지 않아 성능 저하가 작지만, 가지치기 비율이 큰 경우 끊어진 연결이 더 많아지게 되고 이로 인해 성능이 크게 저하된다.
- [0064] 따라서, 컴퓨팅 장치(100)는 학습 데이터를 이용하여 가지치기된 네트워크 모델을 재학습시킬 수 있다.
- [0065] 도 4는 본 발명의 일 실시예에 따른 VOC 데이터에 대한 평가 결과(mAP)를 나타낸 도면이고, 도 5는 본 발명의 일 실시예에 따른 VOC 데이터에서 AND 방법을 적용하여 가지치기 한 경우 성능을 나타낸 도면이고, 도 6은 본 발명의 일 실시예에 따른 VisDrone 데이터 셋에 대한 평가 결과(mAP)를 나타낸 도면이며, 도 7은 본 발명의 일 실시예에 따른 VisDrone 데이터에서 AND 방법을 적용하여 가지치기 한 경우 성능을 나타낸 도면이고, 도 8은 종래와 본 발명의 일 실시예에 따른 성능을 비교한 결과를 도시한 도면이다.
- [0066] 본 발명의 일 실시예에 따른 고유 차원 기반 계층별 채널 가지치기 방법의 성능을 평가하기 위해, PASCAL VOC 2012 데이터 셋과 VisDrone2019-DET 데이터 셋을 이용하였다. VOC 데이터 셋은 탈 것, 가정용품, 동물 등을 포함한 20개 종류의 객체를 가지며, 16,551개의 학습 영상과 4,952개의 테스트 영상으로 구성된다. VisDrone 데이터 셋은 드론에 탑재된 카메라에서 촬영한 영상으로 도로환경에서 주로 관찰되는 보행자, 자동차, 자전거 등을 포함하는 10개 종류의 객체를 가지며, 6,471개의 학습 영상과 548개의 테스트 영상으로 구성된다. 모든 모델의 입력 크기는 416x416으로 고정하였으며, 배치 크기는 모든 학습에서 16으로 동일하게 설정하였고, 50 epoch 동안 훈련을 진행하였다. 희소성 학습은 100 epoch 동안 진행되었으며, penalty factor의 값은 0.01에서 0.0001까지 5개의 값 중에 mAP가 본래의 YOLOv4 모델과 거의 유사하면서 가장 큰 값을 선택하였다. VOC 데이터 셋의 penalty factor는 0.0001로, VisDrone 데이터 셋은 0.01로 설정하였다. 가지치기 비율은 50%, 60%, 70%로 설정하고, 잔차 블록에서의 가지치기 방법인 OR, AND, FR 세 방법을 적용하여 본 발명의 일 실시예에 따른 고유 차원 기반 채널 가지치기 방법의 mAP를 결과로 획득하였다. 또한, OR, AND, FR 중 성능이 가장 좋은 경우에 대해서 파라미터 수와 FLOPs, 모델의 크기를 가지치기 비율에 따라 기본 YOLOv4 모델과 비교하였으며, 마지막으로 기존의 전역적 방법과 본 발명에서 제시된 방법의 성능을 비교하였다.
- [0067] 도 4와 도 6은 VOC 데이터 셋과 VisDrone 데이터 셋에 대한 mAP성능을 보여준다. 두 데이터 셋의 결과를 보면 모두 AND 방법을 적용했을 때 가지치기 비율에 상관없이 mAP가 가장 높다. 도 5와 도 7은 VOC 데이터 셋과 VisDrone 데이터 셋에서 AND 방법을 적용하였을 때 모델의 크기와 계산량이 얼마나 줄어드는지를 보여준다.
- [0068] 도 5에서 보면 VOC 데이터 셋에서는 가지치기 비율이 증가함에 따라 mAP는 7.24%, 13.85%, 27.66% 감소하였으나, 파라미터 수는 기본모델에 비해 81.24%, 89.28%, 94.10% 감소하였고, FLOPs는 43.38%, 54.77%, 66.00% 감소하였으며, 모델의 크기는 78.03%, 85.83%, 90.47% 감소한 것을 알 수 있다. 도 7에서 보면 VisDrone 데이터 셋에서는 mAP가 가지치기 비율이 70%일 때만 11.24% 감소하였고, 파라미터 수는 80.08%, 89.35%, 93.97%, FLOPs는 42.45%, 55.37%, 63.76%, 모델의 크기는 각각 76.92%, 85.90%, 90.35% 감소한 것을 알 수 있다.
- [0069] 도 8의 (a)와 같이 VOC 데이터의 경우는 전역적 방법의 OR, AND, FR 방법의 결과와 모두 비교하였을 때 본 발명의 일 실시예에 따른 고유 차원 기반 채널 가지치기 방법의 성능이 가장 높은 것을 확인할 수 있다. 또한, 도 8의 (b)에서와 같이 작은 크기의 객체가 상대적으로 많이 포함되어 있는 VisDrone 데이터의 경우는 가지치기 비율이 70% 일 때를 제외한 나머지 비율에서는 본 발명의 일 실시예에 따른 고유 차원 기반 채널 가지치기 방법의 성능이 가장 높은 것을 알 수 있다.

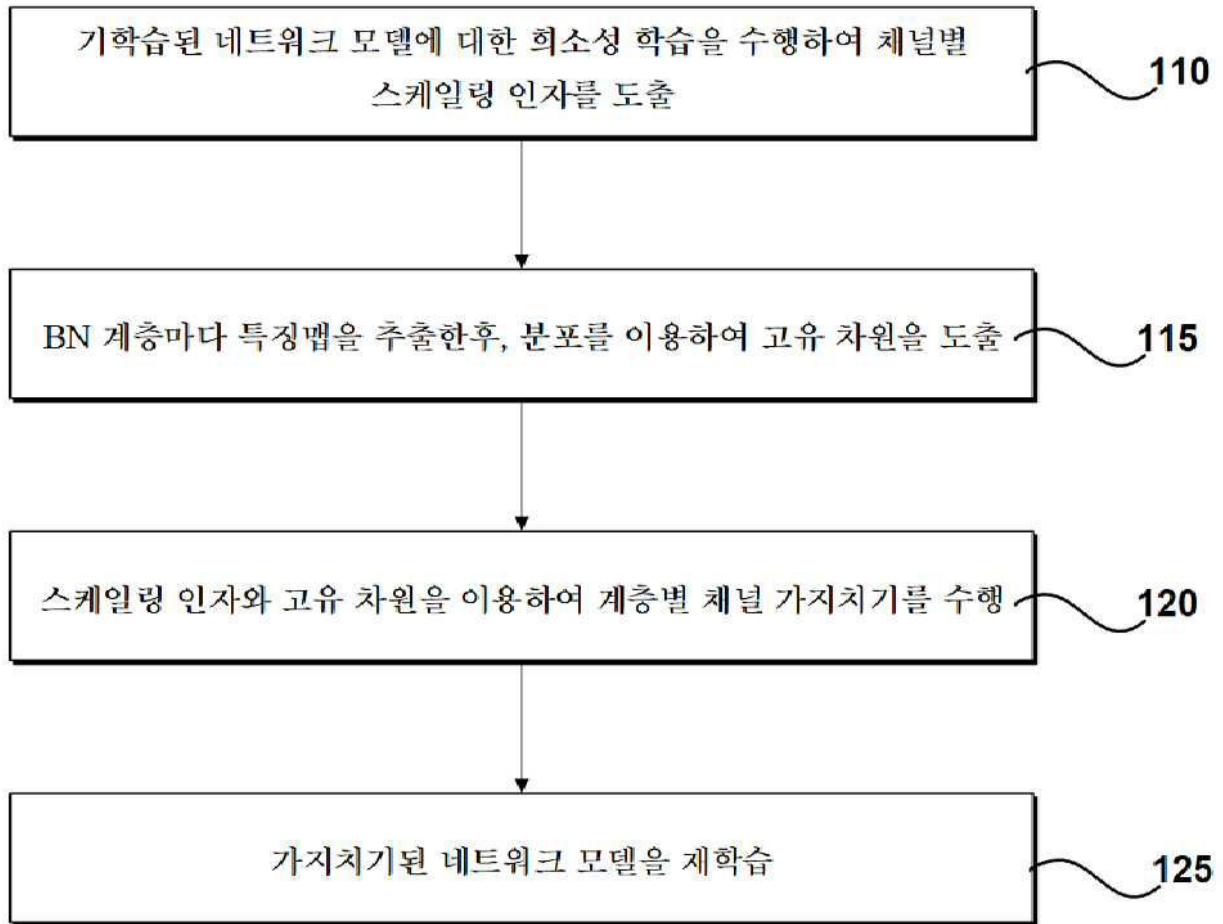
- [0070] 결과적으로 동일한 비율로 가지치기하였을 때 본 발명의 일 실시예에 따른 고유 차원 기반 채널 가지치기 방법이 기존의 전역적 방법보다 성능저하가 더 낮은 것을 확인할 수 있다. 또한, 도 4 내지 도 8에서 보여지는 바와 같이, 각 계층에서 얻은 고유차원은 계층마다 보존할 채널의 양을 추정하는데 효과적이고 이를 기반으로 계층별 채널 가지치기를 수행하면 성능 저하를 최소화하는 것이 가능함을 알 수 있다.
- [0072] 도 9는 본 발명의 일 실시예에 따른 고유 차원 기반 채널 가지치기 방법을 수행하기 위한 컴퓨팅 장치의 내부 구성을 개략적으로 도시한 블록도이다.
- [0073] 도 9를 참조하면, 본 발명의 일 실시예에 따른 컴퓨팅 장치(100)는 메모리(910) 및 프로세서(920)를 포함하여 구성된다.
- [0074] 메모리(910)는 본 발명의 일 실시예에 따른 고유 차원 기반 계층별 채널 가지치기 방법을 수행하기 위해 필요한 다양한 명령어를 저장한다.
- [0075] 프로세서(920)는 본 발명의 일 실시예에 따른 컴퓨팅 장치(100)의 내부 구성 요소들을 제어하기 위한 수단이다. 또한, 프로세서(920)는 메모리(910)에 저장된 명령어를 실행할 수 있다. 프로세서(920)에 의해 실행된 명령어는 도 1에서 설명된 각각의 단계를 수행할 수 있다. 이는 전술한 바와 동일하므로 중복되는 설명은 생략하기로 한다.
- [0077] 본 발명의 실시 예에 따른 장치 및 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야 통상의 기술자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기매체(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.
- [0078] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0079] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

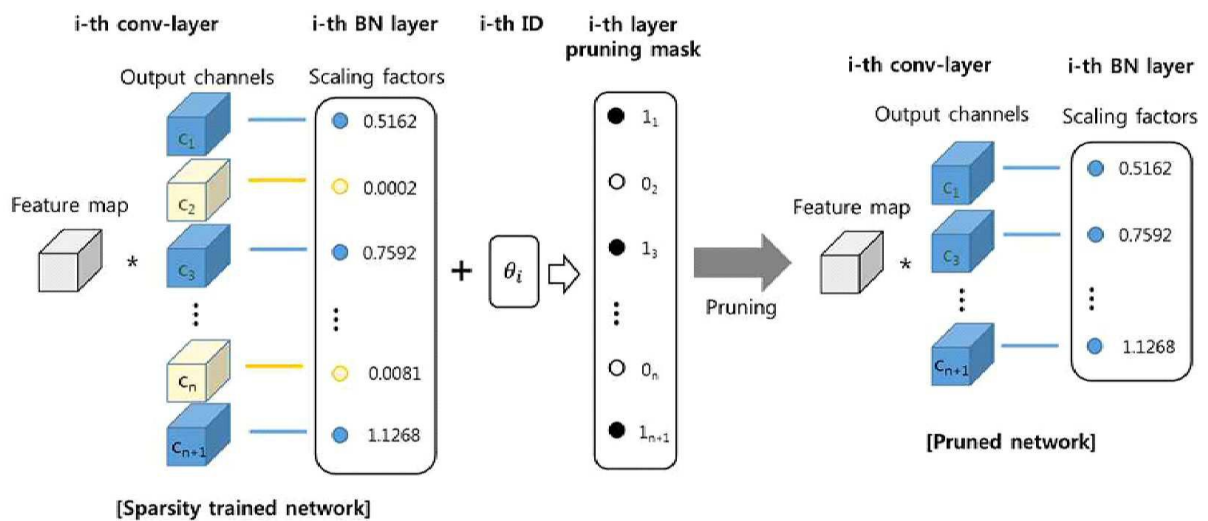
- [0081] 100: 컴퓨팅 장치
910: 메모리
920: 프로세서

도면

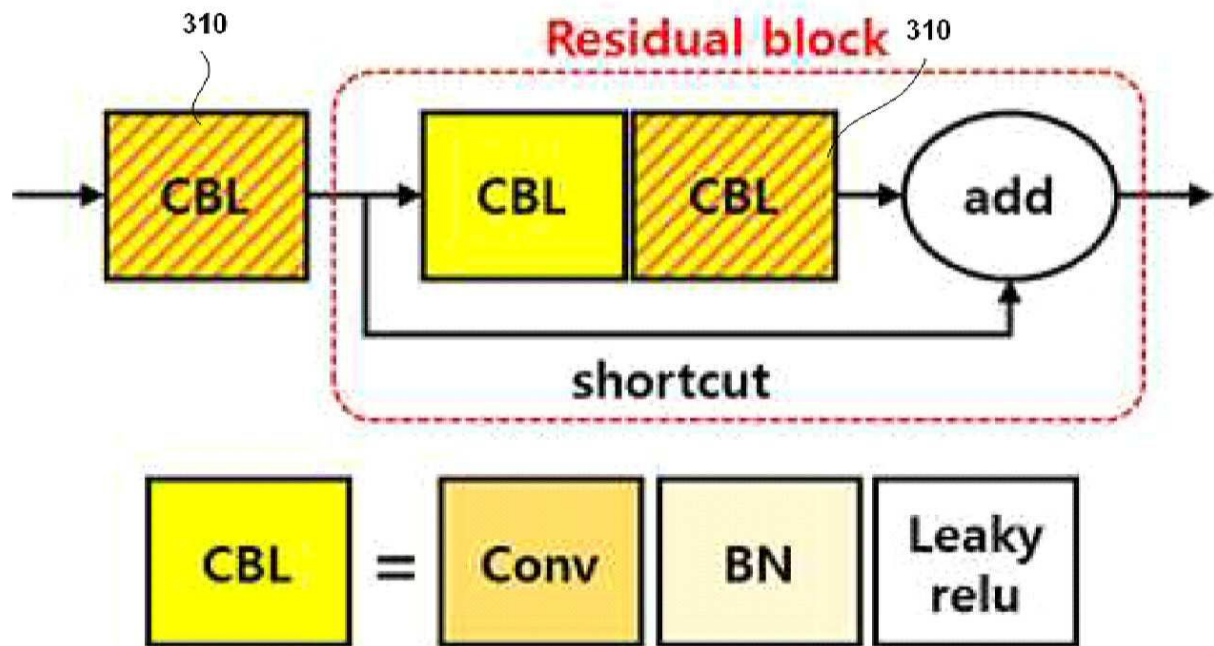
도면1



도면2



도면3



도면4

mode ratio(%)	OR	FR	AND
0	87.61	87.61	87.61
50	78.58	80.23	81.27
60	68.30	73.99	75.48
70	40.55	58.83	63.38

도면5

Ratio(%)	0	50	60	70
mAP(%)	87.61	81.27	75.48	63.38
Param. size(M)	64.11	12.03	6.87	3.78
FLOPs (G)	59.7	33.8	27.0	20.3
Model size(MB)	265.4	58.3	37.6	25.3
Input size	416×416	416×416	416×416	416×416

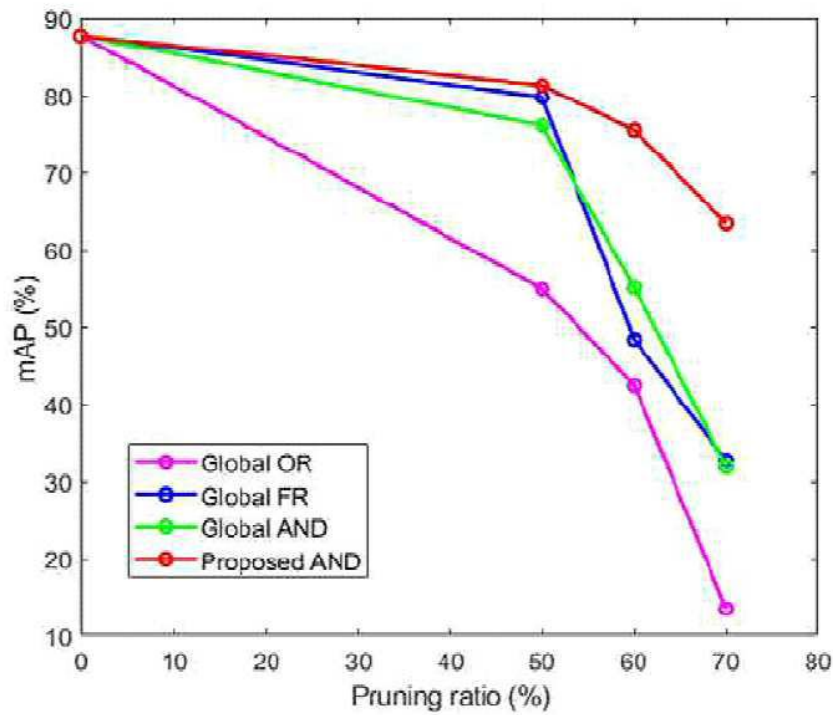
도면6

mode ratio(%)	OR	FR	AND
0	19.13	19.13	19.13
50	19.13	19.21	20.12
60	17.43	18.79	20.03
70	10.42	15.35	16.98

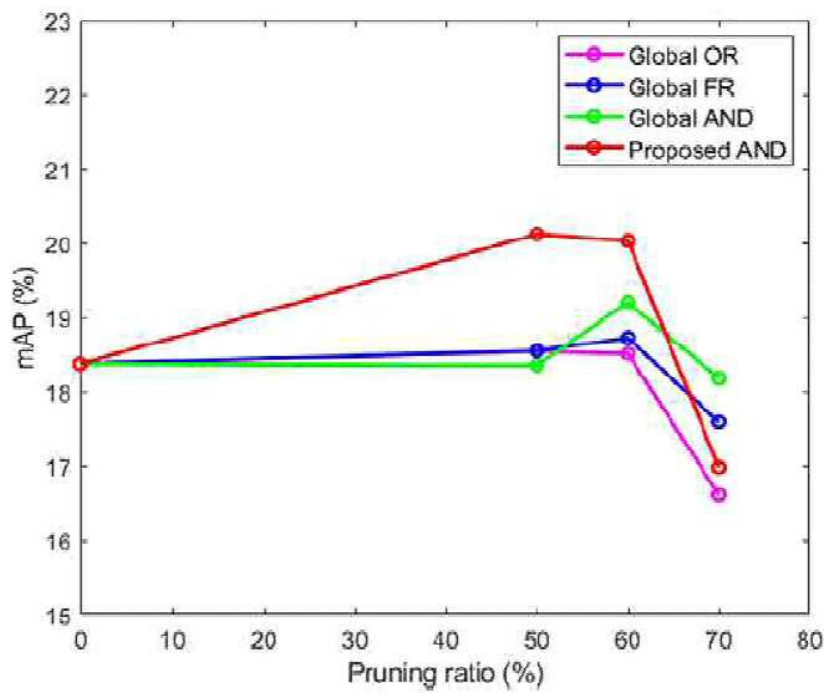
도면7

Ratio	0	50	60	70
mAP	19.13	20.12	20.03	16.98
Param. size(M)	64.05	12.76	6.82	3.86
FLOPs(G)	59.6	34.3	26.6	21.6
Model size(MB)	265.2	61.2	37.4	25.6
Input size	416×416	416×416	416×416	416×416

도면8



(a)



(b)

도면9

