



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0049928
(43) 공개일자 2025년04월14일

- | | |
|---|---|
| <p>(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) G06N 3/0464 (2023.01)
G06N 3/063 (2023.01)</p> <p>(52) CPC특허분류
G06N 3/082 (2023.01)
G06N 3/0464 (2023.01)</p> <p>(21) 출원번호 10-2024-0035340
(22) 출원일자 2024년03월13일
심사청구일자 2024년03월13일</p> <p>(30) 우선권주장
1020230132754 2023년10월05일 대한민국(KR)</p> | <p>(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)</p> <p>(72) 발명자
박중선
서울특별시 서초구 신반포로 32</p> <p>조중호
서울특별시 동대문구 약령시로5길 33-2</p> <p>(74) 대리인
특허법인주연케이알파</p> |
|---|---|

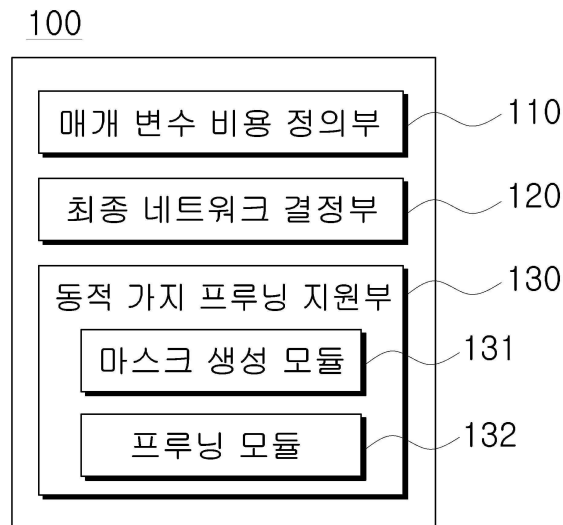
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법 및 이를 지원하는 조기 종료 가지 구성 장치

(57) 요약

본 발명은 조기 종료 가지 (Early-exit branches)의 매개 변수 비용을 정의하고, 매개 변수 비용이 가장 작은 가지의 크기, 위치 및 개수를 토대로 최적의 에너지 효율성을 가지는 네트워크 결정할 수 있으며, 특히 동적 가지 프루닝 기법을 이용하여 매개 변수 비용을 추가로 감소시킴으로써 프루닝 효율을 더욱 극대화할 수 있도록 하는 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법 및 이를 지원하는 조기 종료 가지 구성 장치에 관한 것이다.

대표도 - 도3



(52) CPC특허분류

G06N 3/063 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711182941
과제번호	2020R1A2C3014820
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	온-디바이스 개인화를 위한 에너지 효율적인 딥러닝 가속기 설계
과제수행기관명	고려대학교 산학협력단
연구기간	2023.03.01 ~ 2024.02.29

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193790
과제번호	2022-0-00266-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	PIM인공지능반도체핵심기술개발(설계)
연구과제명	초저전력 Low-Bit Precision 혼성모드 SRAM PIM 개발
과제수행기관명	서울대학교 산학협력단
연구기간	2022.04.01 ~ 2025.12.31

명세서

청구범위

청구항 1

네트워크(network) 연산에 사용되는 매개 변수 개수를 이용하여 매개 변수 비용을 정의하는 단계;
 최저의 매개 변수 비용에 해당하는 가치를 선택한 후 최종 네트워크를 결정하는 단계; 및
 동적 가지 프루닝 기법을 이용하여, 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 단계;를 포함하는 것을 특징으로 하는,
 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법.

청구항 2

제1항에 있어서,
 상기 네트워크(network) 연산에 사용되는 매개 변수 개수를 이용하여 매개 변수 비용을 정의하는 단계는,
 하기의 수학적식을 통해 산출되는 값을 매개 변수 비용으로 결정하는 단계;를 포함하는 것을 특징으로 하는,
 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법.
 [수학적식]

$$Parameter\ cost(\cdot) = \sum_{i=1}^N B_{pram.cost_i} \times BI_i$$

(여기에서, $B_{pram.cost_i}$ 는 i번째 가지에서 조기 종료할 때 필요한 매개 변수의 개수이고, BI_i 는 i번째 가지에서 조기 종료된 이미지의 개수이며, N은 가지의 총 개수이다)

청구항 3

제1항에 있어서,
 상기 최저의 매개 변수 비용에 해당하는 가치를 선택한 후 최종 네트워크를 결정하는 단계는,
 입력되는 네트워크 내 복수의 위치에 복수의 필터를 가지는 가지(branch)를 연결하고 독립적인 가지 학습(branch training)을 진행하는 단계;
 매개 변수 비용 함수를 통해 각 가지 위치 별 매개 변수 비용을 산출하는 단계;
 최저의 매개 변수 비용에 해당하는 가치를 선택한 후 공동 학습(joint training)을 진행하는 단계; 및
 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정하는 단계;를 포함하는 것을 특징으로 하는,
 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법.

청구항 4

제3항에 있어서,

상기 입력되는 네트워크 내 복수의 위치에 복수의 필터를 가지는 가지를 연결하고 독립적인 가지 학습을 진행하는 단계는,

가지 학습 진행 시, 입력되는 네트워크는 학습 대상에서 제외시키는 단계;를 포함하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법.

청구항 5

제3항에 있어서,

상기 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 공동 학습을 진행하는 단계는,

공동 학습 진행 시, 입력되는 네트워크를 학습 대상에 포함시켜 공동 학습을 진행하는 단계;를 포함하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법.

청구항 6

제3항에 있어서,

상기 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정하는 단계는,

공동 학습 진행 결과, 입력되는 네트워크의 매개 변수 비용 대비 매개 변수 비용이 감소되었는지 여부를 확인하는 단계;

네트워크의 매개 변수 비용을 감소된 매개 변수 비용으로 갱신하고, 최저의 매개 변수 비용으로 선택된 가지 보다 상대적으로 매개 변수 비용이 높은 가지를 선택한 후 공동 학습을 재진행하는 과정을 n 회 반복하여, n 번째로 선택되는 가지를 통한 공동 학습 재진행 결과에 따른 매개 변수 비용이 최종적으로 갱신된 네트워크의 매개 변수 비용을 초과하는 시점을 파악하는 단계; 및

$n-1$ 번째로 선택되는 가지를 포함하는 네트워크를 최종 네트워크로 결정하는 단계;를 포함하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법.

청구항 7

제1항에 있어서,

상기 동적 가지 프루닝 기법을 이용하여, 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 단계는,

최종 네트워크에 포함된 가지의 컨볼루션 계층에서 출력되는 결과값을 채널 별로 L1 정규화를 진행하는 단계; 및

채널 별 L1 정규화 값 중에서 가장 높은 L1 정규화 값과 기 정의된 프루닝 팩터(PF) 값을 서로 곱하여 특정 채널

널을 제거함으로써 매개 변수 비용을 추가로 감소시키는 단계;를 포함하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법.

청구항 8

네트워크(network) 연산에 사용되는 매개 변수 개수를 이용하여 매개 변수 비용을 정의하는 매개 변수 비용 정의부;

최저의 매개 변수 비용에 해당하는 가지를 선택한 후 최종 네트워크를 결정하는 최종 네트워크 결정부; 및

동적 가지 프루닝 기법을 이용하여, 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 동적 가지 프루닝 지원부;를 포함하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치.

청구항 9

제8항에 있어서,

상기 매개 변수 비용 정의부는,

하기의 수학적식을 통해 산출되는 값을 매개 변수 비용으로 결정하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치.

[수학적식]

$$Parameter\ cost(\cdot) = \sum_{i=1}^N B_{pram.cost_i} \times BI_i$$

(여기에서, $B_{pram.cost_i}$ 는 i번째 가지에서 조기 종료할 때 필요한 매개 변수의 개수이고, BI_i 는 i번째 가지에서 조기 종료된 이미지의 개수이며, N은 가지의 총 개수이다)

청구항 10

제8항에 있어서,

상기 최종 네트워크 결정부는,

입력되는 네트워크 내 복수의 위치에 복수의 필터를 가지는 가지(branch)를 연결하고 독립적인 가지 학습(branch training)을 진행하고, 매개 변수 비용 함수를 통해 각 가지 위치 별 매개 변수 비용을 산출하며, 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 공동 학습(joint training)을 진행하고, 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치.

청구항 11

제10항에 있어서,

상기 최종 네트워크 결정부는,

상기 독립적인 가지 학습 진행 시, 입력되는 네트워크는 학습 대상에서 제외시키며,

상기 공동 학습 진행 시, 입력되는 네트워크를 학습 대상에 포함시켜 공동 학습을 진행하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치.

청구항 12

제10항에 있어서,

상기 최종 네트워크 결정부는,

공동 학습 진행 결과, 입력되는 네트워크의 매개 변수 비용 대비 매개 변수 비용이 감소되었는지 여부를 확인하고, 네트워크의 매개 변수 비용을 감소된 매개 변수 비용으로 갱신하며, 최저의 매개 변수 비용으로 선택된 가지 보다 상대적으로 매개 변수 비용이 높은 가지를 선택한 후 공동 학습을 재진행하는 과정을 n회 반복하여, n 번째로 선택되는 가지를 통한 공동 학습 재진행 결과에 따른 매개 변수 비용이 최종적으로 갱신된 네트워크의 매개 변수 비용을 초과하는 시점을 파악하고, n-1번째로 선택되는 가지를 포함하는 네트워크를 최종 네트워크로 결정하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치.

청구항 13

제8항에 있어서,

상기 동적 가지 프루닝 지원부는,

상기 최종 네트워크에 포함된 가지의 컨볼루션 계층에서 출력되는 결과값에 대한 채널 별 L1 정규화를 진행하여, 제거 대상이 되는 채널을 선택하기 위한 마스크(mask)를 생성하는 마스크 생성 모듈; 및

생성된 마스크를 이용하여 가지 활성화(branch activation) 데이터를 프루닝하는 프루닝 모듈;을 포함하는 것을 특징으로 하는,

매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치.

발명의 설명

기술 분야

[0001] 본 발명은 조기 종료 가지 (Early-exit branches)의 매개 변수 비용을 정의하고, 매개 변수 비용이 가장 작은 가지의 크기, 위치 및 개수를 토대로 최적의 에너지 효율성을 가지는 네트워크 결정할 수 있도록 하는 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법 및 이를 지원하는 조기 종료 가지 구성 장치에 관한 것이다.

배경 기술

[0003] 일반적으로, Convolutional Neural Network(CNN)은 많은 파라미터와 연산량을 기반으로 이미지 인식/분류, 객체

탐지 등 이미지 처리와 관련된 여러 응용 분야에서 최고의 성능을 보이고 있다. CNN은 높은 추론 성능을 위해서는 많은 연산량이 필요하다. 많은 연산량으로 인해 발생하는 추론 시간과 추론 에너지 소모를 줄이기 위해서는 이미지 별 난이도를 고려한 추론 방법이 효과적이다.

[0004] 일반적으로 CNN 추론에서 이미지의 난이도를 고려한 추론 방법에는 대표적으로 조기 종료 기법이 있다. 조기 종료 기법이 적용된 네트워크는 쉬운 이미지는 앞쪽 가지들에서 추론이 종료되며, 어려운 이미지는 뒤쪽 가지들에서 추론이 종료된다.

[0005] 이때, 쉬운 이미지는 앞쪽 가지에서 종료되기 때문에, 기본 네트워크 보다 적은 에너지를 사용하여 추론이 완료될 수 있다. 하지만 어려운 이미지의 경우에는 기본 네트워크 뿐만 아니라 가지 네트워크도 연산을 진행해야 하므로, 조기 종료 기법이 적용되지 않았을 경우보다 오히려 더 많은 에너지를 필요로 하는 문제점을 가지게 된다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국공개특허 제10-2024-0008143호

발명의 내용

해결하려는 과제

[0008] 본 발명은 전술한 문제점을 해결하기 위한 것으로, 조기 종료 가지 (Early-exit branches)의 매개 변수 비용을 정의하고, 매개 변수 비용이 가장 작은 가지의 크기, 위치 및 개수를 토대로 최적의 에너지 효율성을 가지는 네트워크 결정할 수 있으며, 특히 동적 가지 프루닝 기법을 이용하여 매개 변수 비용을 추가로 감소시킴으로써 프루닝 효율을 더욱 극대화할 수 있도록 하는 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법 및 이를 지원하는 조기 종료 가지 구성 장치를 제공하고자 한다.

과제의 해결 수단

[0010] 본 발명의 일 실시예에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 방법은 네트워크(network) 연산에 사용되는 매개 변수 개수를 이용하여 매개 변수 비용을 정의하는 단계, 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 최종 네트워크를 결정하는 단계 및 동적 가지 프루닝 기법을 이용하여, 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 단계를 포함할 수 있다.

[0011] 일 실시예에서, 상기 네트워크(network) 연산에 사용되는 매개 변수 개수를 이용하여 매개 변수 비용을 정의하는 단계는 하기의 수학적식을 통해 산출되는 값을 매개 변수 비용으로 결정하는 단계를 포함할 수 있다.

[0012] [수학식]

$$Parameter\ cost(\cdot) = \sum_{i=1}^N B_{pram.cost_i} \times BI_i$$

[0013]

[0014] (여기에서, $B_{pram.cost_i}$ 는 i번째 가지에서 조기 종료할 때 필요한 매개 변수의 개수이고, BI_i 는 i번째 가지에서 조기 종료된 이미지의 개수이며, N은 가지의 총 개수이다)

[0015] 일 실시예에서, 상기 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 최종 네트워크를 결정하는 단계는 입

력되는 네트워크 내 복수의 위치에 복수의 필터를 가지는 가지(branch)를 연결하고 독립적인 가지 학습(branch training)을 진행하는 단계, 매개 변수 비용 함수를 통해 각 가지 위치 별 매개 변수 비용을 산출하는 단계, 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 공동 학습(joint training)을 진행하는 단계 및 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정하는 단계를 포함할 수 있다.

[0016] 일 실시예에서, 상기 입력되는 네트워크 내 복수의 위치에 복수의 필터를 가지는 가지를 연결하고 독립적인 가지 학습을 진행하는 단계는 가지 학습 진행 시, 입력되는 네트워크는 학습 대상에서 제외시키는 단계를 포함할 수 있다.

[0017] 일 실시예에서, 상기 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 공동 학습을 진행하는 단계는 공동 학습 진행 시, 입력되는 네트워크를 학습 대상에 포함시켜 공동 학습을 진행하는 단계를 포함할 수 있다.

[0018] 일 실시예에서, 입력되는 네트워크의 매개 변수 비용 대비 매개 변수 비용이 감소되었는지 여부를 확인하는 단계, 네트워크의 매개 변수 비용을 감소된 매개 변수 비용으로 갱신하고, 최저의 매개 변수 비용으로 선택된 가지 보다 상대적으로 매개 변수 비용이 높은 가지를 선택한 후 공동 학습을 재진행하는 과정을 n회 반복하여, n 번째로 선택되는 가지를 통한 공동 학습 재진행 결과에 따른 매개 변수 비용이 최종적으로 갱신된 네트워크의 매개 변수 비용을 초과하는 시점을 파악하는 단계 및 n-1번째로 선택되는 가지를 포함하는 네트워크를 최종 네트워크로 결정하는 단계를 포함할 수 있다.

[0019] 일 실시예에서, 상기 동적 가지 프루닝 기법을 이용하여, 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 단계는,

[0020] 최종 네트워크에 포함된 가지의 컨볼루션 계층에서 출력되는 결과값을 채널 별로 L1 정규화를 진행하는 단계 및 채널 별 L1 정규화 값 중에서 가장 높은 L1 정규화 값과 기 정의된 프루닝 팩터(PF) 값을 서로 곱하여 특정 채널을 제거함으로써 매개 변수 비용을 추가로 감소시키는 단계를 포함할 수 있다.

[0022] 본 발명의 다른 실시예에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 초기 종료 가지 구성 장치(100)는 네트워크(network) 연산에 사용되는 매개 변수 개수를 이용하여 매개 변수 비용을 정의하는 매개 변수 비용 정의부(110), 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 최종 네트워크를 결정하는 최종 네트워크 결정부(120) 및 동적 가지 프루닝 기법을 이용하여, 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 동적 가지 프루닝 지원부(130)를 포함할 수 있다.

[0023] 일 실시예에서, 상기 매개 변수 비용 정의부(110)는 하기의 수학적식을 통해 산출되는 값을 매개 변수 비용으로 결정할 수 있다.

[0024] [수학적식]

$$Parameter\ cost(\cdot) = \sum_{i=1}^N B_{pram.cost_i} \times BI_i$$

[0025]

[0026] (여기에서, $B_{pram.cost_i}$ 는 i번째 가지에서 초기 종료할 때 필요한 매개 변수의 개수이고, BI_i 는 i번째 가지에서 종기 종료된 이미지의 개수이며, N은 가지의 총 개수이다)

[0027] 일 실시예에서, 상기 최종 네트워크 결정부(120)는 입력되는 네트워크 내 복수의 위치에 복수의 필터를 가지는 가지(branch)를 연결하고 독립적인 가지 학습(branch training)을 진행하고, 매개 변수 비용 함수를 통해 각 가지 위치 별 매개 변수 비용을 산출하며, 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 공동 학습(joint training)을 진행하고, 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정할 수 있다.

[0028] 일 실시예에서, 상기 최종 네트워크 결정부(120)는 상기 독립적인 가지 학습 진행 시, 입력되는 네트워크는 학습 대상에서 제외시키며, 상기 공동 학습 진행 시, 입력되는 네트워크를 학습 대상에 포함시켜 공동 학습을 진

행할 수 있다.

[0029] 일 실시예에서, 상기 최종 네트워크 결정부(120)는 공동 학습 진행 결과, 입력되는 네트워크의 매개 변수 비용 대비 매개 변수 비용이 감소되었는지 여부를 확인하고, 네트워크의 매개 변수 비용을 감소된 매개 변수 비용으로 갱신하며, 최저의 매개 변수 비용으로 선택된 가지 보다 상대적으로 매개 변수 비용이 높은 가지를 선택한 후 공동 학습을 재진행하는 과정을 n회 반복하여, n번째로 선택되는 가지를 통한 공동 학습 재진행 결과에 따른 매개 변수 비용이 최종적으로 갱신된 네트워크의 매개 변수 비용을 초과하는 시점을 파악하고, n-1번째로 선택되는 가지를 포함하는 네트워크를 최종 네트워크로 결정할 수 있다.

[0030] 일 실시예에서, 상기 동적 가지 프루닝 지원부(130)는 상기 최종 네트워크에 포함된 가지의 컨볼루션 계층에서 출력되는 결과값에 대한 채널 별 L1 정규화를 진행하여, 제거 대상이 되는 채널을 선택하기 위한 마스크(mask)를 생성하는 마스크 생성 모듈(131) 및 생성된 마스크를 이용하여 가지 활성화(branch activation) 데이터를 프루닝하는 프루닝 모듈(132)을 포함할 수 있다.

발명의 효과

[0032] 본 발명에 따르면, 매개 변수 비용이 가장 작은 가지의 크기, 위치 및 개수를 토대로 최적의 에너지 효율성을 가지는 네트워크 결정할 수 있으며, 특히 동적 가지 프루닝 기법을 이용하여 매개 변수 비용을 추가로 감소시킴으로써 프루닝 효율을 더욱 극대화할 수 있는 이점을 가진다.

[0033] 또한 본 발명에 따르면, CIFAR-10, CIFAR-100, Tiny-ImageNet 3개의 데이터셋과 VGG-16, ResNet-56, MobileNet-V2 3개의 네트워크에 대해 높은 에너지 감소 효율성을 가질 수 있다.

[0034] 특히 본 발명에 따르면, VGG-16 네트워크에서 CIFAR-10 데이터를 사용할 경우, 최대 93%까지 에너지 감소 효과가 극대화될 수 있는 이점을 가질 수 있다.

[0035] 또한 본 발명에 따르면, 하드웨어 구성에 있어서 종래의 baseline 대비 모듈이 추가되어 면적 및 소비 전력이 일부 증가한 반면, 에너지 효율(TOPS/W)은 오히려 증가되는 이점을 가질 수 있다.

도면의 간단한 설명

[0037] 도 1은 종래의 CNN이 컬러 이미지 처리에 필요로 하는 연산량을 나타낸 도면이다.

도 2는 종래의 일반적인 조기 종료 기법의 문제점을 나타낸 도면이다.

도 3은 본 발명의 일 실시예에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치(100)의 구성을 개략적으로 나타낸 도면이다.

도 4는 도 3에 도시된 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치(100)를 이용하여 하드웨어 친화적인 조기 종료 가지를 구성하고 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 전 과정을 일련의 순서대로 나타낸 순서도이다.

도 5는 최종 네트워크 결정부(120)에서 최종 네트워크를 결정하는 전 과정을 일련의 순서대로 나타낸 순서도이다.

도 6은 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과, 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정하는 과정을 보다 구체적으로 나타낸 순서도이다.

도 7은 도 6에 도시된 최종 네트워크 결정부(120)에서 최종 네트워크를 결정하는 전 과정의 일 예를 나타낸 도면이다.

도 8은 동적 가지 프루닝 지원부(130)의 하드웨어 구성을 개략적으로 나타낸 도면이다.

도 9는 도 8에 도시된 동적 가지 프루닝 지원부(130)를 통해 동적 가지 프루닝을 진행하는 개념을 나타낸 도면이다.

도 10은 본 발명에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적

인 조기 종료 가지 구성 장치(100)에 따른, CIFAR-10, CIFAR-100, Tiny-ImageNet 3개의 데이터셋과 VGG-16, ResNet-56, MobileNet-V2 각각의 에너지 감소량을 비교하여 나타낸 그래프이다.

도 11은 본 발명에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치(100)에 VGG-16 네트워크 적용 시, CIFAR-10 데이터에 대한 속도 향상도를 비교하여 나타낸 그래프이다.

발명을 실시하기 위한 구체적인 내용

- [0038] 이하, 첨부한 도면을 참고로 하여 본 발명의 실시예들에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예들에 한정되지 않는다.
- [0039] 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 동일 또는 유사한 구성요소에 대해서는 동일한 참조 부호를 붙이도록 한다.
- [0040] 또한, 여러 실시예들에 있어서, 동일한 구성을 가지는 구성요소에 대해서는 동일한 부호를 사용하여 대표적인 실시예에서만 설명하고, 그 외의 다른 실시예에서는 대표적인 실시예와 다른 구성에 대해서만 설명하기로 한다.
- [0041] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐만 아니라, 다른 부재를 사이에 두고 "간접적으로 연결"된 것도 포함한다. 또한, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함하는 것을 의미할 수 있다.
- [0043] 도 1은 종래의 CNN이 컬러 이미지 처리에 필요로 하는 연산량을 나타낸 도면이고, 도 2는 종래의 일반적인 조기 종료 기법의 문제점을 나타낸 도면이다.
- [0044] 도 1을 살펴보면, 일반적으로 컨볼루션 뉴럴 네트워크(Convolutional neural network, CNN)는 computer vision 분야에서 많이 사용되는 뉴럴 네트워크이다. CNN은 convolution 연산을 통해 가까운 픽셀 간 연관관계를 활용하며, 수많은 곱셈과 덧셈 연산으로 구성된다.
- [0045] 이때, CNN은 224*224 컬러 이미지 1장을 처리하는데 수십억 번의 곱셈을 필요로 한다. 예를 들어, ResNet-50 네트워크의 경우 224*224 컬러 이미지 1장을 처리하는데 약 40억번의 곱셈을 필요로 하고, Inception-v4 네트워크의 경우 약 120억번의 곱셈을 필요로 한다.
- [0046] 만약 224*224 크기가 아닌 HD 해상도(1280*720)나, FHD 해상도(1920*1080) 혹은 UHD(4K) 해상도(3840*2160)의 경우에는 필요로 하는 곱셈수가 18배에서 165배 이상으로 기하급수적으로 증가하기 때문에, Edge 기기(예를 들어, IoT 게이트웨이, 산업용 컨트롤러, 스마트 디스플레이, POS 터미널, 자동판매기, 로봇, 드론 등)와 같이 자원이 제한적인 기기에서 동작하기에는 과도한 연산수로 인해 데이터 처리에 어려움을 겪고 있는 실정이다.
- [0047] 도 2를 살펴보면, 조기 종료 기법은 네트워크 중간에 가지(branch)를 추가로 연결하여, 이미지의 난이도에 따라 추론 연산을 조기에 종료할 수 있도록 함으로써 연산량을 감소시킬 수 있는 기법이다. 이때 가지의 결과가 특정 임계값(threshold) 보다 높으면 해당 가지에서 이미지 추론이 종료되는 방식이다. 이때, 어려운 이미지는 뒤쪽에 위치한 가지에서 추론이 종료되고, 쉬운 이미지일수록 앞쪽에 위치한 가지에서 추론이 종료되면서 전체적인 연산량이 감소되는 것이다.
- [0048] 하지만, 도 2와 같이 조기 종료 기법이 적용되지 않은 일반적인 네트워크에서는 어려운 이미지의 연산 경로가 'a'와 같이 순차적으로 진행되는 반면, 조기 종료 기법 적용 시 어려운 이미지가 모든 가지 별로 연산되어야 하기 때문에, 오히려 훨씬 많은 연산을 수행하여야 연산량이 증가하고 속도가 저하되는 문제가 발생된다. 따라서, 효율적인 에너지 감소 및 처리 속도 증가를 위해서는 적절한 가지의 개수, 위치 및 크기를 고려하여야 한다.
- [0049] 이에, 본 발명에서는 후술되는 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치(100)와, 이를 통한 하드웨어 친화적인 조기 종료 가지 구성 방법을 통해 효율적으로 에너지를 절감하고 처리 속도를 향상시킬 수 있도록 하는 기술에 대해 살펴보기로 한다.
- [0051] 도 3은 본 발명의 일 실시예에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드

웨어 친화적인 조기 종료 가지 구성 장치(100)의 구성을 개략적으로 나타낸 도면이고, 도 4는 도 3에 도시된 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치(100)를 이용하여 하드웨어 친화적인 조기 종료 가지를 구성하고 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시키는 전 과정을 일련의 순서대로 나타낸 순서도이다.

[0052] 도 3과 도 4를 함께 살펴보면, 본 발명의 일 실시예에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치(100)는 크게 매개 변수 비용 정의부(110), 최종 네트워크 결정부(120) 및 동적 가지 프루닝 지원부(130)로 구성될 수 있다.

[0053] 먼저 매개 변수 비용 정의부(110)는 네트워크(network) 연산에 사용되는 매개 변수 개수를 이용하여 매개 변수 비용을 정의한다(S401). 보다 구체적으로, 매개 변수 비용 정의부(110)는 하기 수학적식을 통해 산출되는 값을 매개 변수 비용으로 결정한다.

[0055] [수학식]

$$Parameter\ cost(\cdot) = \sum_{i=1}^N B_{pram.cost_i} \times BI_i$$

[0056] (여기에서, $B_{pram.cost_i}$ 는 i번째 가지에서 조기 종료할 때 필요한 매개 변수의 개수이고, BI_i 는 i번째 가지에서 조기 종료된 이미지의 개수이며, N은 가지의 총 개수이다)

[0058] 이때, 매개 변수 비용은 추론에 사용되는 전체 파라미터 개수를 의미할 수 있으며, 매개 변수 비용이 작다는 것은 그만큼 추론 연산에 필요한 에너지 소모량이 적다는 것을 의미할 수 있다. 후술되는 최종 네트워크 결정부(120)에서는 네트워크와 데이터셋이 입력된 상태에서 매개 변수 비용 정의부(110)를 통해 정의된 매개 변수 비용 중에서 최저의 매개 변수 비용에 해당하는 가지를 선택할 수 있다. 이에 대해 살펴보면 다음과 같다.

[0060] 최종 네트워크 결정부(120)는 매개 변수 비용 정의부(110)를 통해 정의된 매개 변수 비용을 바탕으로, 매개 변수 비용이 가장 작은 가지의 크기, 위치 및 개수를 선택하고, 이 중에서 최저의 매개 변수 비용에 해당하는 가지를 선택한 후 이를 네트워크에 반영하여 최종 네트워크로 결정한다(S402). 이에 대해 보다 구체적으로 살펴보면 다음과 같다.

[0061] 도 5는 최종 네트워크 결정부(120)에서 최종 네트워크를 결정하는 전 과정을 일련의 순서대로 나타낸 순서도이고, 도 6은 도 5에 도시된 최종 네트워크 결정부(120)에서 최종 네트워크를 결정하는 전 과정의 일 예를 나타낸 도면이다.

[0062] 도 5를 살펴보면, 최종 네트워크 결정부(120)는 먼저 입력되는 네트워크 내에서 우선 가지를 연결할 수 있는 모든 위치에 대해 복수의 필터를 가지는 가지(branch)를 연결하고 독립적인 가지 학습(branch training)을 진행한다(S501). 여기에서, 복수의 필터라 함은 1*1, 3*3, 5*5와 같이 3가지 종류의 필터를 의미할 수 있다. 이때, S501 단계에서 최종 네트워크 결정부(120)는 독립적인 가지 학습 과정에서 입력된 네트워크는 학습 대상에서 제외시키고, 가지에 대한 필터들을 가지 학습한다.

[0063] 독립적인 가지 학습이 완료되면, 최종 네트워크 결정부(120)는 매개 변수 비용 함수를 바탕으로, 각 가지 위치별로 매개 변수 비용을 산출한다(S502). 이 후, 최종 네트워크 결정부(120)는 산출된 매개 변수 비용 중에서 가장 작은 크기를 가지는 매개 변수 비용에 해당되는 가지를 선택하고, 이를 토대로 공동 학습(joint trainige)을 진행한다(S503). 이때, 최종 네트워크 결정부(120)는 공동 학습 진행 과정에서 앞서 S501 단계와는 달리 입력되는 네트워크를 학습 대상에 포함시켜 공동 학습을 진행한다.

[0064] 다음으로, 최종 네트워크 결정부(120)는 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과, 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정한다(S504). 이러한 S504 단계에 대해 보다 구체적으로 살펴보면 다음과 같다.

- [0065] 도 6은 공동 학습 진행 후 입력되는 네트워크의 매개 변수 비용과, 공동 학습 결과에 따른 매개 변수 비용을 서로 비교하는 과정을 반복함으로써, 매개 변수 비용이 가장 작은 최종 네트워크를 결정하는 과정을 보다 구체적으로 나타낸 순서도이다.
- [0066] 도 6을 살펴보면, 최종 네트워크 결정부(120)는 먼저, 입력된 네트워크의 매개 변수 비용과 앞서 공동 학습 결과에 따라 발생하는 매개 변수 비용을 서로 비교하여, 매개 변수 비용이 감소되었는지 여부를 확인한다(S601). 이때, 공동 학습 결과에 따라 발생하는 매개 변수 비용이 감소된 경우, 최종 네트워크 결정부(120)는 네트워크의 매개 변수 비용을 감소된 매개 변수 비용으로 갱신한다(S602).
- [0067] 네트워크의 매개 변수 비용이 감소된 매개 변수 비용으로 갱신된 상태에서, 최종 네트워크 결정부(120)는 최저의 매개 변수 비용으로 선택된 가지 보다 상대적으로 매개 변수 비용이 더 높은 가지를 선택한다(S603). 여기에서, 가지의 위치는 순차적이지 않을 수 있다. 하지만, 선택되는 가지는 앞서 최저의 매개 변수 비용보다는 크되, 남은 모든 가지들 중에서는 최저에 해당한다.
- [0068] 최종 네트워크 결정부(120)는 새로 선택된 가지를 토대로 S503 단계인 공동 학습을 재진행하고, 마찬가지로 공동 학습 결과에 따라 발생하는 매개 변수 비용이 재감소된 경우, 최종 네트워크 결정부(120)는 네트워크의 매개 변수 비용을 재감소된 매개 변수 비용으로 재갱신한다(S604). 이러한 S604 단계의 과정들은 공동 학습을 재진행하여 나오는 매개 변수 비용이 갱신된 네트워크의 매개 변수 비용보다 커지는 시점까지 계속해서 반복될 수 있다.
- [0069] 보다 구체적으로, 최종 네트워크 결정부(120)는 공동 재학습을 재진행하는 과정을 n 회 반복하여, n 번째로 선택되는 가지를 통한 공동 학습 재진행 결과에 따른 매개 변수 비용이 최종적으로 갱신된 네트워크의 매개 변수 비용을 초과하는 시점을 파악한다(S605). 만약, n 번째로 선택되는 가지를 통한 공동 학습 재진행 결과에 따른 매개 변수 비용이 최종적으로 갱신된 네트워크의 매개 변수 비용을 초과한 경우, 최종 네트워크 결정부(120)는 $n-1$ 번째로 선택된 가지를 최종적인 가지로 선택하고, 해당 가지를 포함하는 네트워크를 최종 네트워크로 결정하는 것이다(S606).
- [0070] 이하에서는, 도 7을 통해 살펴본 과정을 일 예로서 구체적으로 살펴보기로 한다.
- [0072] 도 7은 도 6에 도시된 최종 네트워크 결정부(120)에서 최종 네트워크를 결정하는 전 과정의 일 예를 나타낸 도면이다.
- [0073] 도 7을 살펴보면, 먼저 최종 네트워크 결정부(120)는 도 7(a)와 같이 네트워크의 모든 위치에 대하여 다양한 크기(1*1, 3*3, 5*5)의 필터를 사용하는 가지를 연결하여 가지 학습을 진행한다. 이때, 최종 네트워크 결정부(120)는 가지 학습 과정에서 네트워크는 학습 대상에 반영하지 않는다.
- [0074] 다음으로, 도 7(b)와 같이 최종 네트워크 결정부(120)는 매개 변수 비용 함수를 통해 모든 가지 별 매개 변수 비용을 산출한다. 이때, 각 매개 변수 비용은 가지의 위치 별로 필터 크기에 따라 서로 다르게 산출된다.
- [0075] 다음으로, 도 7(c)와 같이 최종 네트워크 결정부(120)는 모든 매개 변수 비용 중에서 우선 가장 작은 매개 변수 비용에 해당하는 가지를 선택하고 이를 토대로 공동 학습을 진행한다. 이때 공동 학습은 입력된 네트워크도 학습 과정에 포함시켜 진행된다. 예를 들어, 도 7(c)에서는 '39'에 해당하는 매개 변수 비용을 가지는 가지가 최저의 가지에 해당한다.
- [0076] 따라서, 최종 네트워크 결정부(120)는 이러한 '39'의 매개 변수 비용을 가지는 가지를 이용하여 공동 학습을 진행한다. 이 결과, 도 7(d)와 같이 입력된 네트워크의 매개 변수 비용은 '100'에서 '73'으로 감소되는 것을 알 수 있다.
- [0077] 최종 네트워크 결정부(120)는 네트워크의 매개 변수 비용을 '100'이 아닌 '73'으로 갱신하고, 도 7(e)와 같이 다음 가지를 선택한다. 이때 선택되는 가지는 '39' 보다 다음으로 큰 '45'의 매개 변수 비용에 해당되는 가지를 의미하며, 최종 네트워크 결정부(120)는 해당 가지를 이용하여 공동 학습을 재진행한다. 이 결과, 도 7(f)와 같이 입력된 네트워크의 매개 변수 비용은 다시 '73'에서 '64'로 더욱 감소된 것을 알 수 있다.
- [0078] 최종 네트워크 결정부(120)는 네트워크의 매개 변수 비용을 다시 '64'로 재 갱신하며, 이러한 과정은 갱신된 네트워크의 매개 변수 비용 보다 n 번째 선택된 가지를 이용하여 공동 학습을 재진행하여 발생하는 매개 변수 비용이 클 시점까지 반복된다.

- [0079] 예를 들어, 도 7(g)와 같이 4번째 순서로 선택된 위치의 가치를 통해 공동 학습을 재진행하여 발생하는 매개 변수 비용이 '56'인데 반해, 그 이전에 3번째로 갱신된 네트워크의 매개 변수 비용이 '55'일 경우, 최종 네트워크 결정부(120)는 '56'의 매개 변수 비용을 네트워크에 대한 매개 변수 비용으로 갱신하는 것이 아니라, 도 7(h)와 같이 n-1번째에 해당하는 3번째 순서로 선택된 위치의 가치를 최종적인 가치로 선택하고, 이를 포함하는 네트워크를 최종 네트워크로 결정하게 되는 것이다.
- [0080] 한편, 이렇게 최종 네트워크가 결정되고 나면, 매개 변수 비용을 더욱 추가적으로 감소시켜 에너지 효율성을 높일 수 있도록, 동적 가지 프루닝 지원부(130)에서는 동적 가지 프루닝 기법을 이용하여 최종 네트워크의 완전 연결 계층에 대한 매개 변수 비용을 추가로 감소시킨다(S403). 여기에서 동적 가지 프루닝 지원부(130)에 대해 보다 구체적으로 살펴보면 다음과 같다.
- [0081] 도 8은 동적 가지 프루닝 지원부(130)의 하드웨어 구성을 개략적으로 나타낸 도면이고, 도 9는 도 8에 도시된 동적 가지 프루닝 지원부(130)를 통해 동적 가지 프루닝을 진행하는 개념을 나타낸 도면이다.
- [0082] 도 8 및 도 9를 함께 살펴보면, 본 발명에 따른 동적 가지 프루닝 지원부(130)는 크게 마스크 생성 모듈(131)과 프루닝 모듈(132)로 구성될 수 있다.
- [0083] 마스크 생성 모듈(131)은 최종 네트워크에 포함된 가지의 컨볼루션 계층에서 출력되는 결과값에 대한 채널 별 L1 정규화(norm)를 진행함으로써, 제거 대상이 되는 채널을 선택하기 위한 마스크(mask)를 생성한다.
- [0084] 마스크가 생성되고 나면, 프루닝 모듈(132)은 생성된 마스크를 이용하여 가지 활성화(branch activation) 데이터를 프루닝하여 최종 네트워크의 매개 변수 비용을 더욱 감소시키는 것이다.
- [0085] 이때, 입력 이미지의 난이도에 따라 동적 가지 프루닝 지원부(130)에서 프루닝하는 프루닝 위치와 개수가 달라질 수 있다.
- [0086] 동적 가지 프루닝이 진행되는 과정을 살펴보면, 일반적으로 가지는 Feature Reduction 계층과, 컨볼루션 계층과 FC 계층의 3단계로 이루어진다. 이때, 컨볼루션 계층을 통해 출력된 데이터에 대해 마스크 생성 모듈(131)은 채널 별 L1 정규화를 진행하고, 전체 L1 정규화 값 중에서 가장 큰 값과 사전에 미리 정의된 프루닝 팩터(PF) 값을 서로 곱하여 산출되는 값을 이용하여, 불필요한 채널을 제거할 수 있는 것이다.
- [0087] 이 경우, 쉬운 이미지에서는 L1 정규화 값이 크게 나타나기 때문에 상대적으로 더 많은 채널이 스킵될 수 있고, 어려운 이미지에서는 L1 정규화 값이 작게 나타나기 때문에 상대적으로 적은 채널이 스킵될 수 있다.
- [0088] 그 결과, 전체 정확도에는 크게 영향을 주지 않으면서도 전체적인 연산량을 감소시킬 수 있는 것이다.
- [0090] 다음으로는, 본 발명에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 초기 종료 가지 구성 장치(100)에 적용된 다양한 네트워크 및 데이터셋의 처리에 필요한 에너지 소모량을 비교하여 살펴보기로 한다.
- [0091] 도 10은 본 발명에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 초기 종료 가지 구성 장치(100)에 따른, CIFAR-10, CIFAR-100, Tiny-ImageNet 3개의 데이터셋과 VGG-16, ResNet-56, MobileNet-V2 각각의 에너지 감소량을 비교하여 나타낸 그래프이고, 도 11은 본 발명에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 초기 종료 가지 구성 장치(100)에 VGG-16 네트워크 적용 시, CIFAR-10 데이터에 대한 속도 향상도를 비교하여 나타낸 그래프이다.
- [0092] 도 10을 살펴보면, 본 발명에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 초기 종료 가지 구성 장치(100)에 대하여 CIFAR-10, CIFAR-100, Tiny-ImageNet 3개의 데이터셋과 VGG-16, ResNet-56, MobileNet-V2 3개의 네트워크를 적용한 결과, 비교군인 BN(1), SDN(2) 및 HAPI(3)의 초기 종료 기법 대비 높은 에너지 감소량을 보이는 것을 알 수 있다. 특히, VGG-16 네트워크에서 CIFAR-10 데이터를 사용할 경우, 최대 93%의 에너지 감소 효과를 보이는 것을 확인할 수 있다. 또한, MobileNet-V2에서 Tiny-ImageNet을 사용한 경우를 제외하고, 비교군인 BN(1), SDN(2) 및 HAPI(3)의 초기 종료 기법과 비교하였을 때 낮은 에너지 소모를 보이는 것을 알 수 있다.
- [0093] 또한, 속도 및 에너지 효율 관점에서 봤을 때, 도 11을 살펴보면, 본 발명에 따른 매개 변수 비용에 기반한 저복잡도 컨볼루션 뉴럴 네트워크를 위한 하드웨어 친화적인 초기 종료 가지 구성 장치(100)는 VGG-16 네트워크에서 CIFAR-10 데이터를 사용할 경우, 최대 4.46배의 속도 향상을 보이는 것을 알 수 있다. 또한, 종래의

Baseline 하드웨어 대비 모듈이 더욱 추가되면서 면적과 소비 전력은 증가하였지만, 본 발명에 의해 종래의 Baseline 하드웨어 대비 에너지 효율(TOPS/W)이 1.5배에서 최대 3.85배까지 증가되는 것을 알 수 있다.

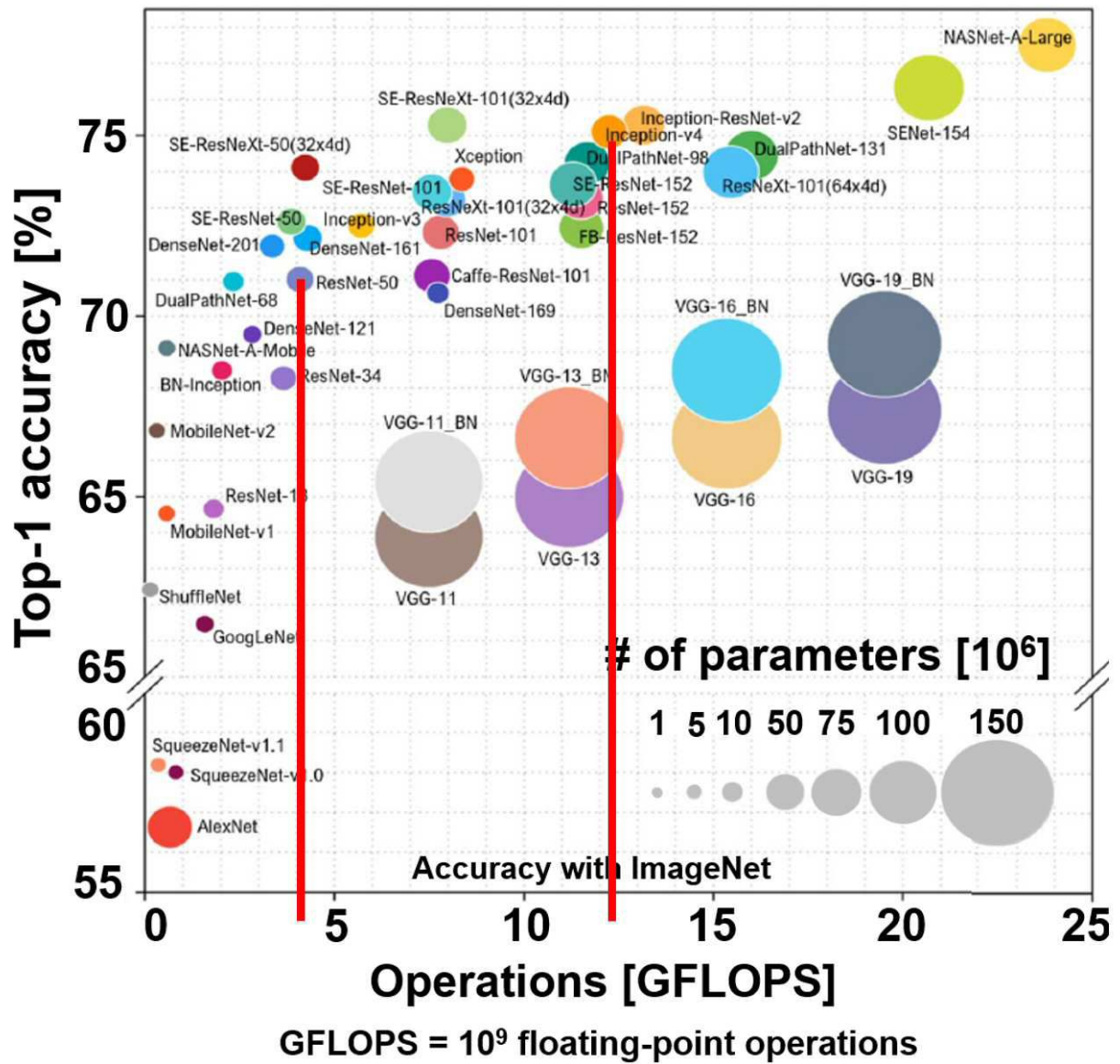
[0095] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

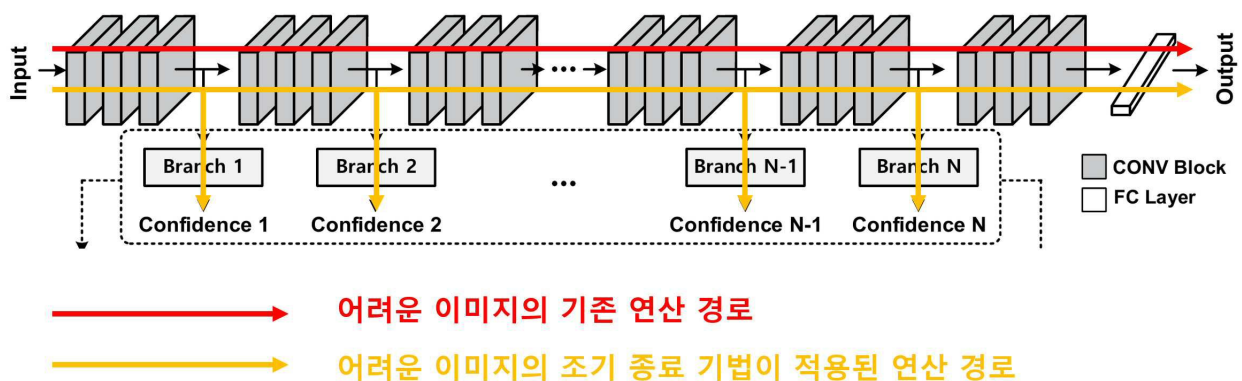
[0097] 100: 매개 변수 비용에 기반한 저복잡도 컨볼루셔널 뉴럴 네트워크를 위한 하드웨어 친화적인 조기 종료 가지 구성 장치
110: 매개 변수 비용 정의부
120: 최종 네트워크 결정부
130: 동적 가지 프루닝 지원부
131: 마스크 생성 모듈
132: 프루닝 모듈

도면

도면1

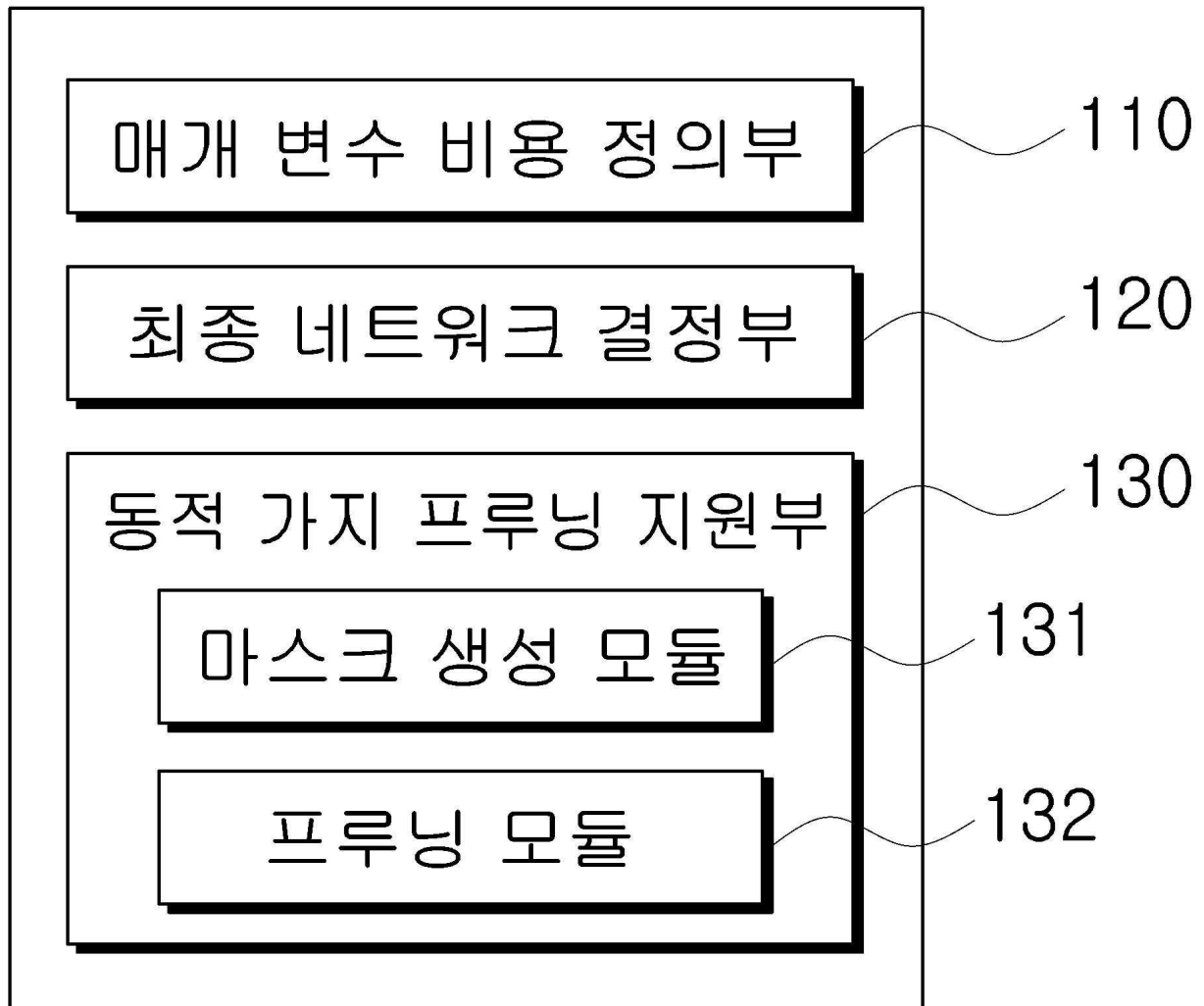


도면2

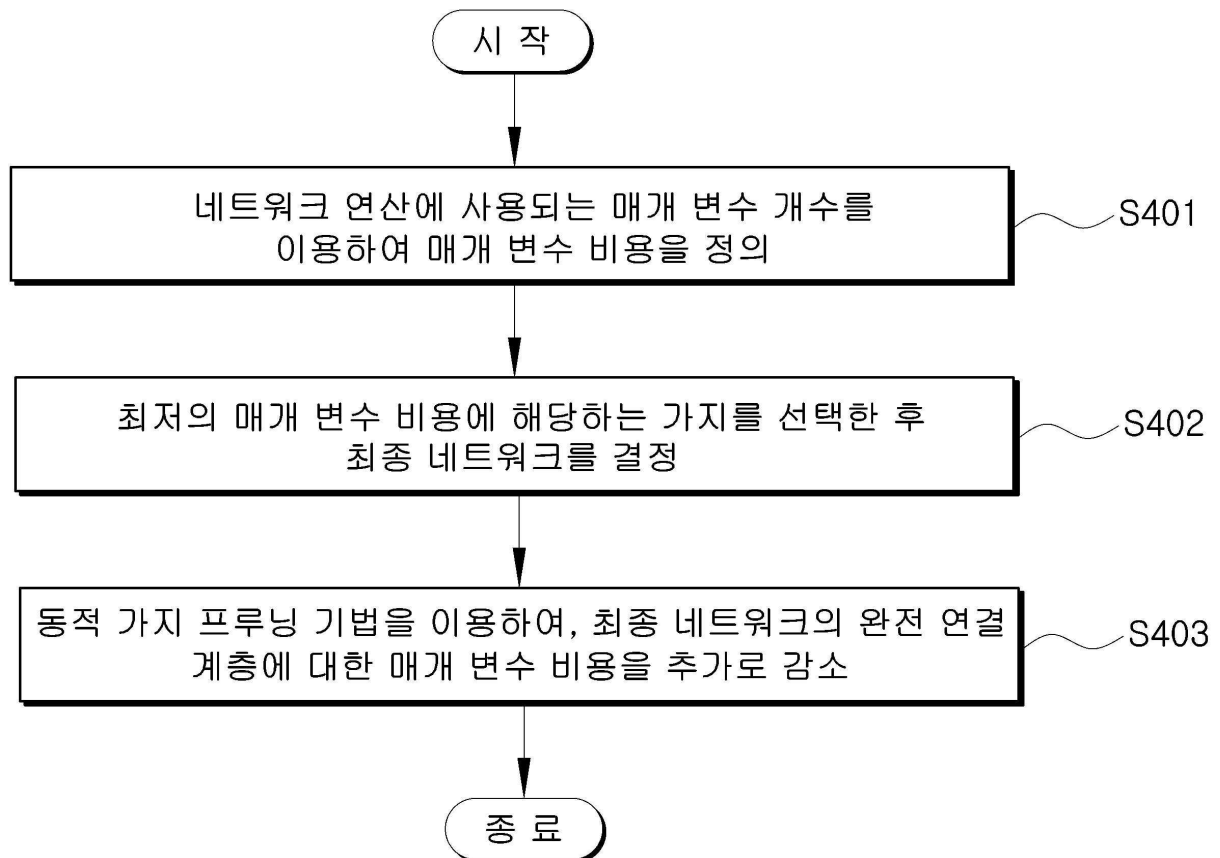


도면3

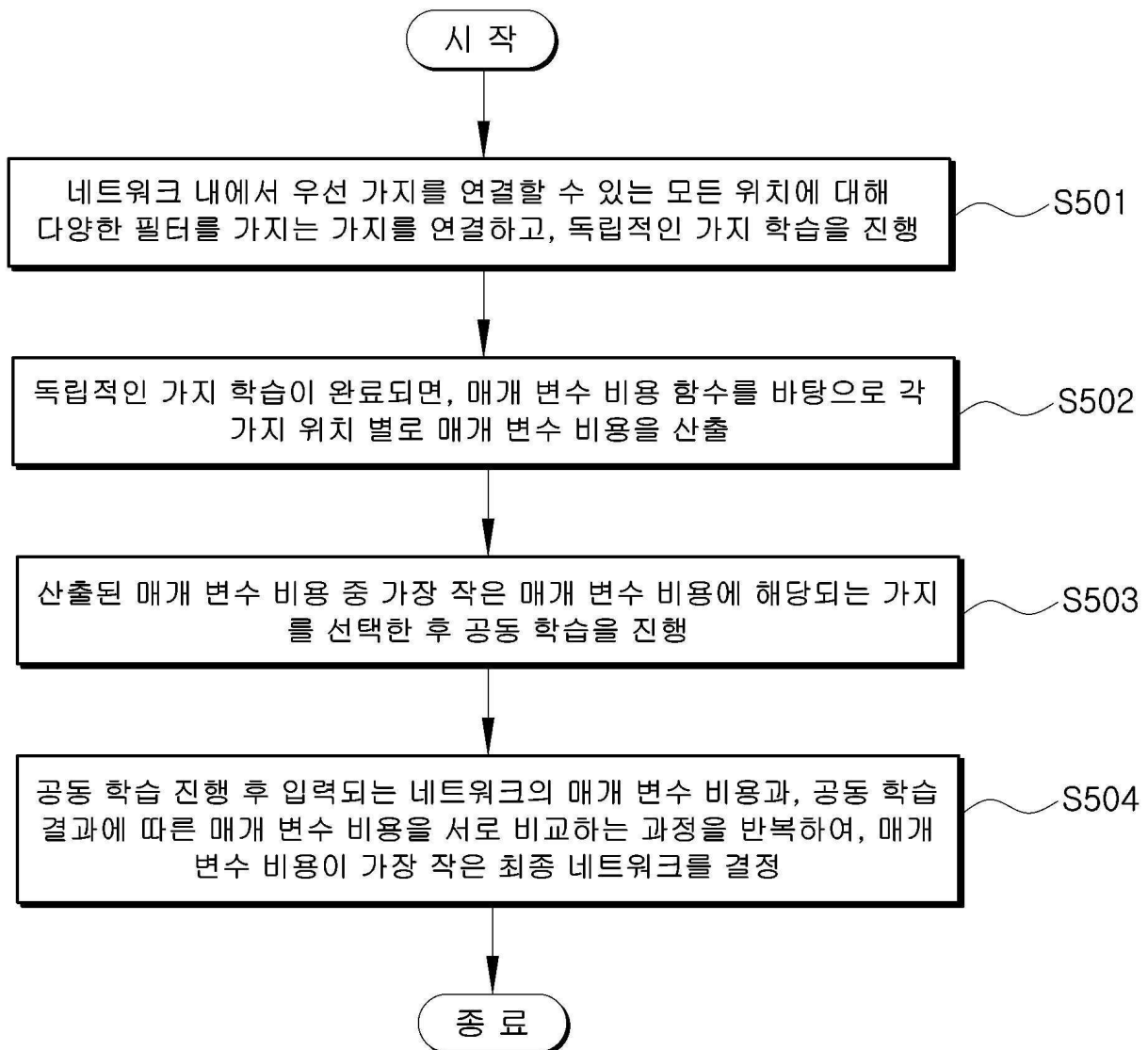
100



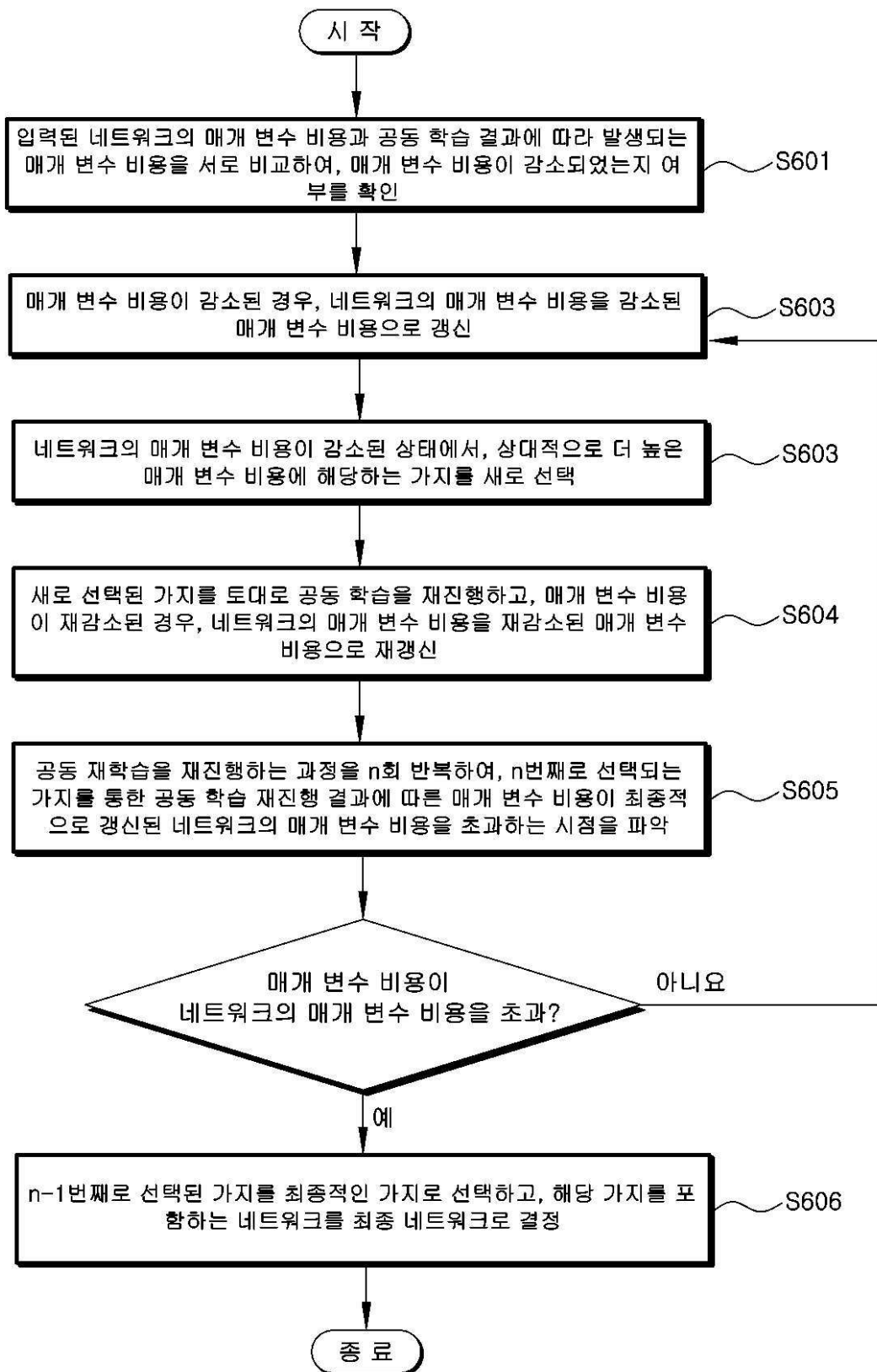
도면4



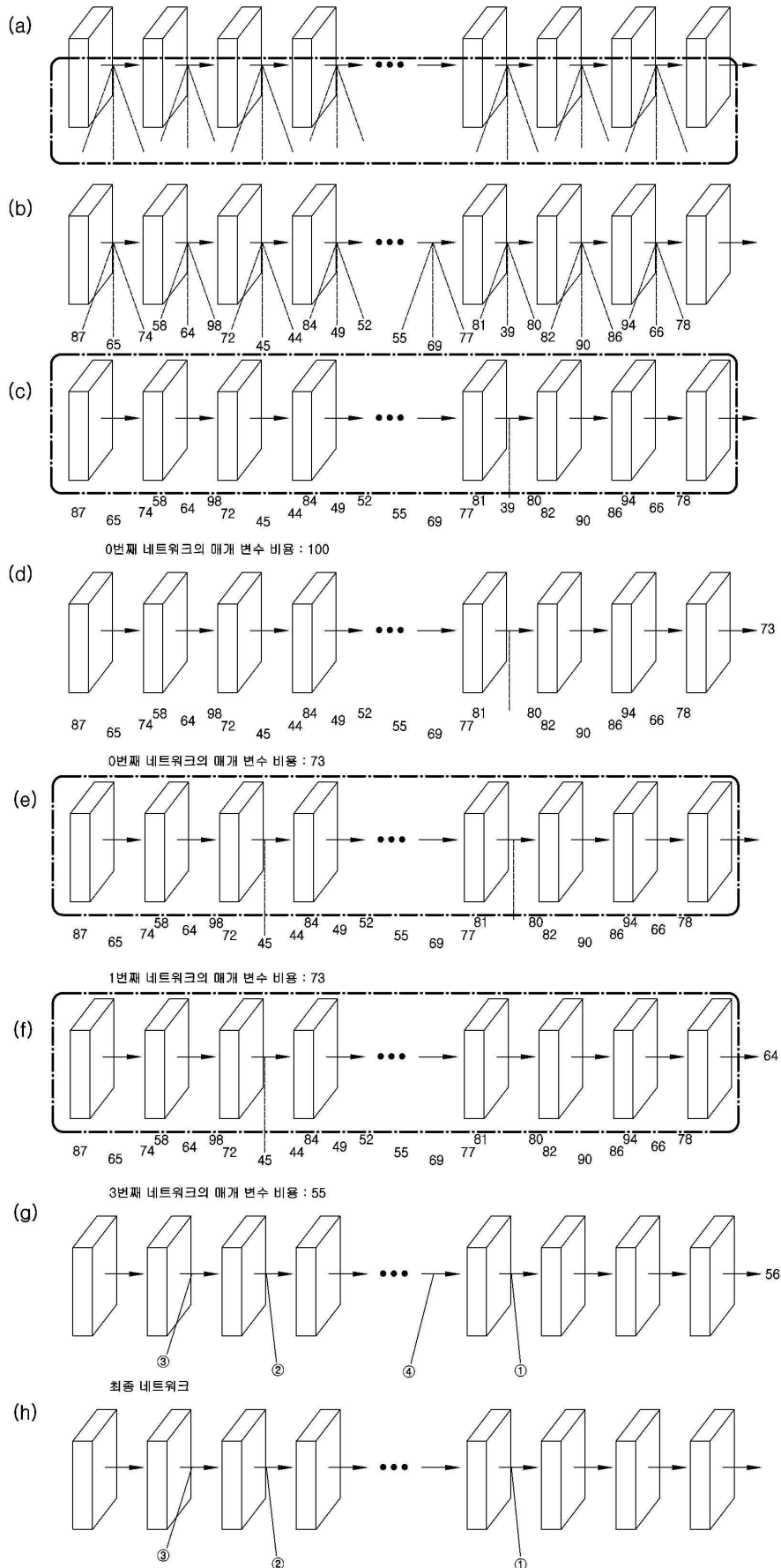
도면5



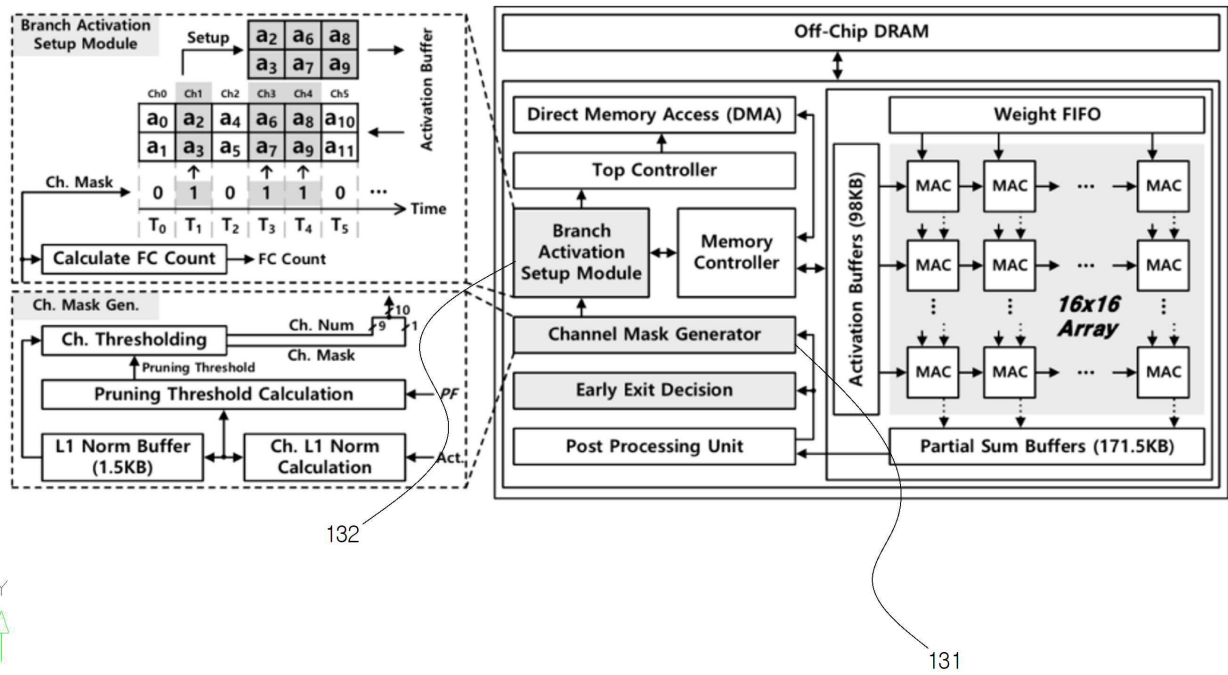
도면6



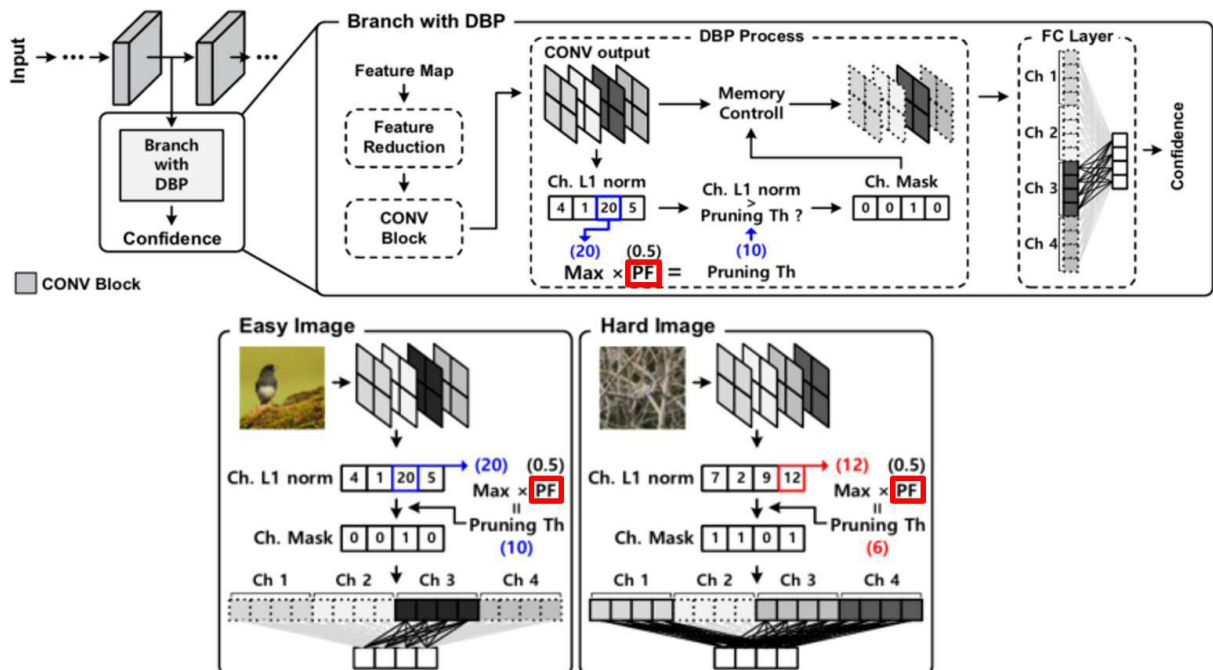
도면7



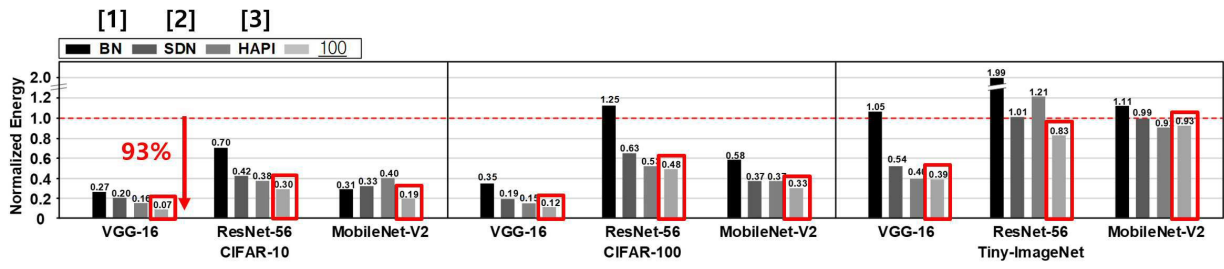
도면8



도면9



도면10



도면11

