



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0094056
(43) 공개일자 2020년08월06일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2006.01) G06N 3/063 (2006.01)
(52) CPC특허분류
G06N 3/04 (2013.01)
G06N 3/063 (2013.01)
(21) 출원번호 10-2019-0058453(분할)
(22) 출원일자 2019년05월18일
심사청구일자 2019년05월18일
(62) 원출원 특허 10-2019-0011516
원출원일자 2019년01월29일
심사청구일자 2019년01월29일

(71) 출원인
주식회사 디퍼아이
서울특별시 서초구 남부순환로347길 42-3, 3층 (서초동, 송인빌딩)
(72) 발명자
이상현
경기도 용인시 기흥구 용구대로2469번길 20, 819호
김명겸
경기도 수원시 영통구 인계로189번길 14
매탄주공4단지 419동 501호
김주혁
서울특별시 중랑구 면목로79길 8-9
(74) 대리인
김봉조

전체 청구항 수 : 총 18 항

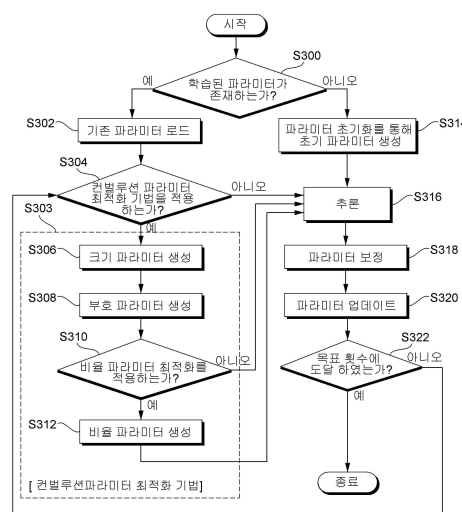
(54) 발명의 명칭 컨벌루션 신경망 파라미터 최적화 방법, 컨벌루션 신경망 연산방법 및 그 장치

(57) 요약

본 발명은 컨벌루션 신경망 파라미터 최적화 방법, 컨벌루션 신경망 연산방법 및 그 장치에 관한 것으로서, 본 발명에 따른 하드웨어 구현에 적합한 신경망 파라미터 최적화 방법은, 신경망의 기존 파라미터를 부호 파라미터 및 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환하는 단계; 및 상기 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성하는 단계;를 포함할 수 있다.

따라서, 본 발명은, 컨벌루션 신경망이 가지는 많은 연산량과 파라미터를 하드웨어 구현에 효과적으로 최적화시켜 최소한의 정확도 손실 및 최대한의 연산속도를 얻을 수 있는 신경망 파라미터 최적화 방법, 신경망 연산방법 및 그 장치를 제공하는 효과가 있다.

대표도 - 도3



명세서

청구범위

청구항 1

파라미터 최적화부를 구성하는 부호 파라미터 변환부와 크기 파라미터 변환부를 이용해 신경망의 기존 파라미터를 부호 파라미터 및 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환하는 단계; 및

파라미터 최적화부를 구성하는 파라미터 가지치기부를 이용해 상기 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성하는 단계;를 포함하는 것을 특징으로 하는 신경망 파라미터 최적화 방법.

청구항 2

제 1 항에 있어서,

상기 부호 파라미터는 기존 파라미터의 채널 별 원소들의 방향을 결정하고,

상기 크기 파라미터는 기존 파라미터의 채널 당 단일의 대표 값으로 가중치들을 최적화한 것을 특징으로 하는 신경망 파라미터 최적화 방법.

청구항 3

제 1 항에 있어서, 상기 형태 변환하는 단계는,

기존 파라미터의 채널 별 원소들의 절대값의 평균을 연산하여 크기 파라미터를 생성하는 것을 특징으로 하는 신경망 파라미터 최적화 방법.

청구항 4

제 1 항에 있어서, 상기 형태 변환하는 단계는,

기존 파라미터의 채널을 구성하는 소정의 원소가 0보다 작으면 대응하는 부호 파라미터의 원소 값을 0으로, 0보다 크거나 같으면 대응하는 부호 파라미터의 원소 값을 1로 표현하여 부호 파라미터를 생성하는 것을 특징으로 하는 신경망 파라미터 최적화 방법.

청구항 5

제 1 항에 있어서, 상기 가지치기하는 단계는,

입력채널 별 크기 파라미터의 평균 값 및 크기 분포 또는 입력 및 출력채널 별 크기 파라미터의 평균 값 및 크기 분포를 이용하여 기준 값을 계산하고, 계산된 기준 값보다 적은 값을 가진 크기 파라미터 값을 0으로 만들어 해당 채널의 컨볼루션 연산을 생략하도록 하는 것을 특징으로 하는 신경망 파라미터 최적화 방법.

청구항 6

제 1 항에 있어서, 상기 신경망 파라미터 최적화 방법은,

파라미터 최적화부를 구성하는 비율 파라미터 변환부를 이용해 신경망의 기존 파라미터를 비율 파라미터로 형태 변환하는 단계;

를 더 포함하는 것을 특징으로 하는 신경망 파라미터 최적화 방법.

청구항 7

제 6 항에 있어서, 상기 비율 파라미터로 형태 변환하는 단계는,
비율 파라미터의 비트를 가변적으로 할당하는 단계; 및
비율 파라미터 원소의 값의 범위 및 가중치를 사용자 선택에 의해 양자화되는 단계;
를 포함하는 것을 특징으로 하는 신경망 파라미터 최적화 방법.

청구항 8

신경망 연산장치를 구성하는 프로세서를 이용해 신경망의 기존 파라미터 및 입력채널 데이터를 메모리에 로드하는 단계;
상기 프로세서에 포함되는 파라미터 최적화부를 구성하는 부호 파라미터 변환부 및 크기 파라미터 변환부를 이용해 신경망의 기존 파라미터를 부호 파라미터 및 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환하고, 상기 파라미터 최적화부를 구성하는 파라미터 가지치기부를 이용해 상기 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성하는 단계;
상기 프로세서에 포함되는 추론부를 이용해 상기 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계;
상기 프로세서에 포함되는 보정부부를 이용해 상기 최적화된 파라미터를 보정하는 단계; 및
상기 프로세서에 포함되는 업데이트부를 이용해 기존 파라미터를 보정된 최적화 파라미터로 업데이트하는 단계;
를 포함하는 것을 특징으로 하는 신경망 연산방법.

청구항 9

제 8 항에 있어서, 상기 신경망 연산방법은,
학습된 파라미터가 존재하는지 판단하는 단계;
상기 학습된 파라미터가 없으면 파라미터 초기화를 통해 초기 파라미터를 생성하는 단계;
초기 파라미터를 대상으로 최적화된 파라미터를 생성하는 단계; 및
상기 학습된 파라미터가 있으면 기존 파라미터를 로드하는 단계;
를 더 포함하는 것을 특징으로 하는 신경망 연산방법.

청구항 10

제 8 항에 있어서, 상기 추론부를 이용해 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계는,
상기 최적화된 파라미터를 메모리에 로드하는 단계;
상기 로드된 최적화된 파라미터에 포함되는 크기 파라미터의 값이 0인지 판단하는 단계; 및
상기 크기 파라미터의 값이 0이면 컨볼루션 연산 과정을 생략하는 단계;
를 포함하는 것을 특징으로 하는 신경망 연산방법.

청구항 11

제 8 항에 있어서, 상기 추론부를 이용해 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계는,

상기 크기 파라미터의 값이 0이 아니면 부호 파라미터와 입력채널 데이터를 비트 연산하여 데이터의 방향을 결정하는 단계;

컨볼루션 파라미터 필터 크기만큼 입력채널 데이터 간 합 연산하는 단계; 및

상기 크기 파라미터와 입력채널 데이터를 대상으로 단일의 곱 연산을 수행하는 단계;

를 포함하는 것을 특징으로 하는 신경망 연산방법.

청구항 12

제 10 항에 있어서, 상기 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계는,

비율 파라미터가 존재하면 비율 파라미터를 이용하여 크기 파라미터에 가중치를 차등 반영함에 따라 연산결과의 오차를 줄이는 단계;

를 더 포함하는 것을 특징으로 하는 신경망 연산방법.

청구항 13

기존 파라미터를 부호 파라미터로 형태 변환하는 부호 파라미터 변환부;

상기 기존 파라미터를 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환하는 크기 파라미터 변환부; 및

상기 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성하는 파라미터 가지치기부;

를 포함하는 것을 특징으로 하는 신경망 연산장치.

청구항 14

제 13 항에 있어서, 상기 파라미터 가지치기부는

입력채널 별 크기 파라미터의 평균 값 및 크기 분포 또는 입력 및 출력채널 별 크기 파라미터의 평균 값 및 크기 분포를 이용하여 기준 값을 계산하고, 계산된 기준 값보다 적은 값을 가진 크기 파라미터 값을 0으로 만들어 해당 채널의 컨볼루션 연산을 생략하도록 하는 것을 특징으로 하는 신경망 연산장치.

청구항 15

제 13 항에 있어서, 상기 신경망 연산장치는

기존 파라미터를 비율 파라미터로 형태 변환하는 비율 파라미터 변환부;

를 더 포함하는 것을 특징으로 하는 신경망 연산장치.

청구항 16

제 13 항에 있어서, 상기 신경망 연산장치는

최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 추론부;

를 더 포함하는 것을 특징으로 하는 신경망 연산장치.

청구항 17

제 16 항에 있어서, 상기 추론부는

최적화 파라미터에 포함된 크기 파라미터의 값이 0인지 판단하고 크기 파라미터의 값이 0이면 컨볼루션 연산 과정을 생략하는 것을 특징으로 하는 신경망 연산장치.

청구항 18

제 16 항에 있어서, 상기 추론부는

비율 파라미터가 존재하면 비율 파라미터를 이용하여 크기 파라미터에 가중치를 차등 반영함에 따라 연산결과의 오차를 줄이는 것을 특징으로 하는 신경망 연산장치.

발명의 설명

기술 분야

[0001] 본 발명은 인공 신경망에서의 연산기술에 관한 것으로, 더욱 상세하게는, 하드웨어 구현에 적합한 신경망 파라미터 최적화 방법, 신경망 연산방법 및 그 장치에 관한 것이다.

배경 기술

[0002] 최근 인공지능(artificial intelligence) 기술이 발달함에 따라 다양한 산업에서 인공지능 기술을 응용 및 도입하고 있다.

[0003] 이러한 추세에 따라 컨볼루션 신경망(Convolutional Neural Network)과 같은 인공 신경망(artificial neural network)을 실시간 하드웨어로 구현하고자 하는 수요도 증가하고 있다.

[0004] 그러나, 종래기술에 따른 인공신경망 관련 연산방법은, 컨볼루션 신경망이 가지는 많은 파라미터와 연산량 때문에 실제 구현은 쉽지 않다.

[0005] 이를 해결하기 위해 학습을 통해 불필요한 파라미터를 생략하는 방법과 파라미터의 비트 수를 감소시키는 방법 등, 정확도 대비 연산량을 감소시키기 위한 파라미터 최적화 방법들이 제안되고 있다.

[0006] 하지만, 종래기술에 따른 파라미터 최적화 방법들은 하드웨어 구현을 고려하지 않은 소프트웨어 연산방법에 한정 되기 때문에, 이러한 방법만을 적용하여 컨볼루션 신경망을 실시간 하드웨어로 구현하는 것은 한계가 있다.

[0007] 따라서, 하드웨어 구현에 적합한 인공신경망 관련 연산기술에 관한 현실적이고도 적용이 가능한 기술이 절실히 요구되고 있는 실정이다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 공개특허공보 KR 10-2011-0027916호(공개일자 2011.03.17)

발명의 내용

해결하려는 과제

[0009] 본 발명은 상기 문제점을 해결하기 위하여 안출된 것으로서, 본 발명은, 컨볼루션 신경망이 가지는 많은 연산량과 파라미터를 하드웨어 구현에 효과적으로 최적화시켜 최소한의 정확도 손실 및 최대한의 연산속도를 얻기 위한 신경망 파라미터 최적화 방법, 신경망 연산방법 및 그 장치를 제공하는데 있다.

[0010] 또한, 최소한의 연산속도 손실을 통해 정확도를 보정할 수 있는 신경망 파라미터 최적화 방법, 신경망 연산방법 및 그 장치를 제공하는데 있다.

과제의 해결 수단

[0011] 본 발명의 실시예에 따른 신경망 파라미터 최적화 방법은, 신경망의 기존 파라미터를 부호 파라미터 및 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환하는 단계와, 상기 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성하는 단계를 포함한다.

[0012] 상기 부호 파라미터는 기존 파라미터의 채널 별 원소들의 방향을 결정하고, 상기 크기 파라미터는 기존 파라미터의 채널 당 단일의 대표 값으로 가중치들을 최적화한 것이다.

[0013] 상기 형태 변환하는 단계에서, 기존 파라미터의 채널 별 원소들의 절대값의 평균을 연산하여 크기 파라미터를 생성할 수 있다.

[0014] 상기 형태 변환하는 단계에서, 기존 파라미터의 채널을 구성하는 소정의 원소가 0보다 작으면 대응하는 부호 파라미터의 원소 값을 0으로, 0보다 크거나 같으면 대응하는 부호 파라미터의 원소 값을 1로 표현하여 부호 파라미터를 생성할 수 있다.

[0015] 상기 가지치기하는 단계에서, 입력채널 별 크기 파라미터의 평균 값 및 크기 분포 또는 입력 및 출력채널 별 크기 파라미터의 평균 값 및 크기 분포를 이용하여 기준 값을 계산하고, 계산된 기준 값보다 적은 값을 가진 크기 파라미터 값을 0으로 만들어 해당 채널의 컨볼루션 연산을 생략하도록 할 수 있다.

[0016] 소정의 레이어를 구성하는 입력채널의 크기 파라미터의 평균 값 및 크기 분포를 이용하여 가지치기하는 경우, 입력채널들의 크기 파라미터의 평균 값에 레이어 별 상수를 곱한 값으로 기준 값을 계산하고, 소정의 입력채널의 크기 파라미터 값이 기준 값보다 작으면 해당 채널의 크기 파라미터 값을 0으로 변경하여 가지치기하며, 소정의 레이어를 구성하는 입력 및 출력채널의 크기 파라미터의 평균 값 및 크기 분포를 이용하여 가지치기하는 경우, 입력채널들 및 출력채널들의 크기 파라미터의 평균 값에 레이어 별 상수를 곱한 값으로 기준 값을 계산하고, 소정의 입력채널의 크기 파라미터 값이 기준 값보다 작으면 해당 채널의 크기 파라미터 값을 0으로 변경하여 가지치기할 수 있다.

[0017] 상기 레이어 별 상수 값은 레이어 별 컨볼루션 파라미터 분포에 따라 결정될 수 있다.

[0018] 상기 신경망 파라미터 최적화 방법은, 신경망의 기존 파라미터를 비율 파라미터로 형태 변환하는 단계를 더 포함할 수 있다.

[0019] 상기 비율 파라미터로 형태 변환하는 단계는, 비율 파라미터의 비트를 가변적으로 할당하는 단계와, 비율 파라미터 원소의 값의 범위 및 가중치를 사용자 선택에 의해 양자화되는 단계를 포함할 수 있다.

[0020] 본 발명의 다른 실시 예에 따른 신경망 연산방법은, 신경망의 기존 파라미터 및 입력채널 데이터를 메모리에 로드하는 단계와, 기존 파라미터를 부호 파라미터 및 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환하고 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성하는 단계와, 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계와, 최적화된 파라미터를 보정하는 단계와, 기존 파라미터를 보정된 최적화 파라미터로 업데이트하는 단계를 포함할 수 있다.

[0021] 상기 신경망 연산방법은, 학습된 파라미터가 존재하는지 판단하는 단계와, 학습된 파라미터가 없으면 파라미터 초기화를 통해 초기 파라미터를 생성하는 단계와, 초기 파라미터를 대상으로 최적화된 파라미터를 생성하는 단계, 및 학습된 파라미터가 있으면 기존 파라미터를 로드하는 단계를 더 포함할 수 있다.

[0022] 상기 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계는, 최적화된 파라미터를 메모리에 로드하는 단계와, 상기 메모리에 로드된 최적화된 파라미터에 포함되는 크기 파라미터의 값이 0인지 판단하는 단계와, 상기 크기 파라미터의 값이 0이면 컨볼루션 연산 과정을 생략하는 단계를 포함할 수 있다.

[0023] 상기 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계는, 상기 크기 파라미터의 값이 0이 아니면 부호 파라미터와 입력채널 데이터를 비트 연산하여 데이터의 방향을 결정하는 단계와, 컨볼루션 파라미터 필터 크기만큼 입력채널 데이터 간 합 연산하는 단계와, 크기 파라미터와 입력채널 데이터를 대상으로 단일의 곱 연산을 수행하는 단계를 포함할 수 있다.

[0024] 상기 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 단계는, 비율 파라미터가 존재하면

비율 파라미터를 이용하여 크기 파라미터에 가중치를 차등 반영함에 따라 연산결과의 오차를 줄이는 단계를 더 포함할 수 있다.

[0025] 본 발명의 또 다른 실시 예에 따른 신경망 연산장치는, 기존 파라미터를 부호 파라미터로 형태 변환하는 부호 파라미터 변환부와, 기존 파라미터를 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환하는 크기 파라미터 변환부, 및 상기 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성하는 파라미터 가지치기부를 포함할 수 있다.

[0026] 상기 파라미터 가지치기부는, 입력채널 별 크기 파라미터의 평균 값 및 크기 분포 또는 입력 및 출력채널 별 크기 파라미터의 평균 값 및 크기 분포를 이용하여 기준 값을 계산하고, 계산된 기준 값보다 적은 값을 가진 크기 파라미터 값을 0으로 만들어 해당 채널의 컨볼루션 연산을 생략하도록 할 수 있다.

[0027] 상기 파라미터 가지치기부는, 소정의 레이어를 구성하는 입력채널의 크기 파라미터의 평균 값 및 크기 분포를 이용하여 가지치기하는 경우, 입력채널들의 크기 파라미터의 평균 값에 레이어 별 상수를 곱한 값으로 기준 값을 계산하고, 소정의 입력채널의 크기 파라미터 값이 기준 값보다 작으면 해당 채널의 크기 파라미터 값을 0으로 변경하여 가지치기하며, 소정의 레이어를 구성하는 입력 및 출력채널의 크기 파라미터의 평균 값 및 크기 분포를 이용하여 가지치기하는 경우, 입력채널들 및 출력채널들의 크기 파라미터의 평균 값에 레이어 별 상수를 곱한 값으로 기준 값을 계산하고, 소정의 입력채널의 크기 파라미터 값이 기준 값보다 작으면 해당 채널의 크기 파라미터 값을 0으로 변경하여 가지치기할 수 있다.

[0028] 상기 신경망 연산장치는, 기존 파라미터를 비율 파라미터로 형태 변환하는 비율 파라미터 변환부를 더 포함할 수 있다.

[0029] 상기 신경망 연산장치는, 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론하는 추론부를 더 포함할 수 있다.

[0030] 상기 추론부는 최적화 파라미터에 포함된 크기 파라미터의 값이 0인지 판단하고 크기 파라미터의 값이 0이면 컨볼루션 연산 과정을 생략할 수 있다.

[0031] 상기 추론부는 비율 파라미터가 존재하면 비율 파라미터를 이용하여 크기 파라미터에 가중치를 차등 반영함에 따라 연산결과의 오차를 줄일 수 있다.

발명의 효과

[0032] 이상에서 설명한 바와 같이, 본 발명은, 컨볼루션 신경망이 가지는 많은 연산량과 파라미터를 하드웨어 구현에 효과적으로 최적화시켜 최소한의 정확도 손실 및 최대한의 연산속도를 얻을 수 있는 신경망 파라미터 최적화 방법, 신경망 연산방법 및 그 장치를 제공하는 효과가 있다.

[0033] 또한, 본 발명은, 신경망의 파라미터를 최적화함에 따라, 실시간으로 동작하는 컨볼루션 신경망을 구현한 하드웨어에서 요구되는 저전력, 고성능을 만족시킬 수 있다. 일 실시 예에 따른 신경망 파라미터 최적화 기법을 적용하면, 컨볼루션 연산에서의 채널의 크기에 따라 채널 별로 연산을 생략할 수 있다. 게다가, 하드웨어에서 채널 내 각각의 원소에 대해 연산을 생략할 것인지 수행할 것인지 판별하는 것이 아니라, 각 채널 별로 연산 생략 여부를 판별하므로 채널 원소 개수의 배수로 연산이 줄어드는 효과가 있다.

[0034] 또한, 본 발명은, 최소한의 연산속도 손실을 통해 정확도를 보정할 수 있다. 예를 들어, 비율 파라미터(scale parameter)를 별도로 분리해 값의 범위에 따라 다른 가중치를 효과적으로 적용할 수 있어 하드웨어 구현 시 성능을 효율적으로 높일 수 있다.

도면의 간단한 설명

[0035] 도 1은 본 발명이 일 실시 예에 따른 신경망 컨볼루션 연산에 사용되는 컨볼루션 파라미터(convolution parameter)를 정의한 도면,

도 2는 본 발명의 일 실시 예에 따른 컨볼루션 신경망(Convolution Neural Network: CNN)에서 최적화 파라미터 생성을 위한 학습 프로세스를 도시한 도면,

도 3은 본 발명의 일 실시 예에 따른 도 2의 최적화 파라미터 생성을 위한 학습 프로세스를 세부화하여 도시한 도면,

- 도 4는 본 발명의 일 실시 예에 따른 추론 프로세스를 세부화하여 도시한 도면,
 도 5는 본 발명의 일 실시 예에 따른 컨볼루션 파라미터 최적화 프로세스의 적용 예를 보여주는 도면,
 도 6은 일반적인 컨볼루션 연산과 본 발명의 일 실시 예에 따른 컨볼루션 연산을 비교한 도면,
 도 7은 본 발명의 일 실시 예에 따른 파라미터 최적화 이후 컨볼루션 연산의 이점을 보여주는 도면,
 도 8은 본 발명의 다른 실시 예에 따른 컨볼루션 파라미터 최적화 프로세스의 적용 예를 보여주는 도면,
 도 9는 비율 파라미터 최적화를 설명하기 위해 크기 파라미터만을 적용한 파라미터 분포와 비율 파라미터를 추가 적용한 파라미터 분포를 비교한 도면,
 도 10은 본 발명의 일 실시 예에 따른 최적화된 파라미터를 이용한 컨볼루션 연산의 예를 도시한 도면,
 도 11은 본 발명의 일 실시 예에 따른 신경망 연산장치의 구성을 도시한 도면,
 도 12는 본 발명의 일 실시 예에 따른 도 11의 파라미터 최적화부의 세부 구성을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0036] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.
- [0037] 본 발명의 실시예들을 설명함에 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이며, 후술되는 용어들은 본 발명의 실시예에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [0038] 첨부된 블록도의 각 블록과 흐름도의 각 단계의 조합들은 컴퓨터 프로그램인스트럭션들(실행 엔진)에 의해 수행될 수도 있으며, 이들 컴퓨터 프로그램 인스트럭션들은 범용 컴퓨터, 특수용 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장치의 프로세서에 탑재될 수 있으므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장치의 프로세서를 통해 수행되는 그 인스트럭션들이 블록도의 각 블록 또는 흐름도의 각 단계에서 설명된 기능들을 수행하는 수단을 생성하게 된다.
- [0039] 이들 컴퓨터 프로그램 인스트럭션들은 특정 방식으로 기능을 구현하기 위해 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장치를 지향할 수 있는 컴퓨터 이용가능 또는 컴퓨터 판독 가능 메모리에 저장되는 것도 가능하므로, 그 컴퓨터 이용가능 또는 컴퓨터 판독 가능 메모리에 저장된 인스트럭션들은 블록도의 각 블록 또는 흐름도의 각 단계에서 설명된 기능을 수행하는 인스트럭션 수단을 내포하는 제조 품목을 생산하는 것도 가능하다.
- [0040] 그리고 컴퓨터 프로그램 인스트럭션들은 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장치 상에 탑재되는 것도 가능하므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장치 상에서 일련의 동작 단계들이 수행되어 컴퓨터로 실행되는 프로세스를 생성해서 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장치를 수행하는 인스트럭션들은 블록도의 각 블록 및 흐름도의 각 단계에서 설명되는 기능들을 실행하기 위한 단계들을 제공하는 것도 가능하다.
- [0041] 또한, 각 블록 또는 각 단계는 특정된 논리적 기능들을 실행하기 위한 하나 이상의 실행 가능한 인스트럭션들을 포함하는 모듈, 세그먼트 또는 코드의 일부를 나타낼 수 있으며, 몇 가지 대체 실시예들에서는 블록들 또는 단계들에서 언급된 기능들이 순서를 벗어나서 발생하는 것도 가능함을 주목해야 한다. 예컨대, 잇달아 도시되어 있는 두 개의 블록들 또는 단계들은 사실 실질적으로 동시에 수행되는 것도 가능하며, 또한 그 블록들 또는 단계들이 필요에 따라 해당하는 기능의 역순으로 수행되는 것도 가능하다.
- [0042] 이하, 첨부 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다. 그러나 다음에 예시하는 본 발명의 실시예는 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 다음에 상술하는 실시예에 한정되는 것은 아니다. 본 발명의 실시예는 이 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 본 발명을 보다 완전하

게 설명하기 위하여 제공된다.

- [0043] 도 1은 본 발명의 일 실시 예에 따른 신경망 컨볼루션 연산에 사용되는 컨볼루션 파라미터(convolution parameter)를 정의한 도면이다.
- [0044] 도면에 도시된 바와 같이, 도 1을 참조하면, 신경망의 전체 컨볼루션 레이어 개수를 l 이라 하면, 각 레이어마다 k 개의 출력채널(Output Channel)($O_1, \dots, O_k$)이 존재하고, 각 출력채널은 j 개의 입력채널(Input Channel)($I_1, I_2, \dots, O_j$)로 구성된다. 각각의 입력채널은 i 개의 가중치 파라미터($W_1, W_2, \dots, W_i$)를 가진다. 본 발명에서는 위에서 정의한 컨볼루션 파라미터의 정의를 기반으로 최적화 기법을 수식적으로 설명하였다.
- [0045] 도 2는 본 발명의 일 실시 예에 따른 컨볼루션 신경망(Convolution Neural Network: CNN)에서 최적화 파라미터 생성을 위한 학습 프로세스를 도시한 도면이다.
- [0046] 도 2를 참조하면, 학습 프로세스는 기존 파라미터 및 입력 데이터 로드 단계(S200), 파라미터 최적화 단계(S210), 추론 단계(S220), 보정 단계(S230) 및 파라미터 업데이트 단계(S240)를 포함할 수 있다.
- [0047] 보다 상세하게는, 상기 파라미터 최적화 단계(S210)를 생략할 경우 일반적인 딥러닝 학습 과정과 동일하며, 이는 일반적인 딥러닝 학습환경에서 본 발명을 쉽게 적용할 수 있음을 의미한다.
- [0048] 본 발명의 실시예에 따른 컨볼루션 신경망 최적화 기법은 기존의 학습된 파라미터가 존재할 경우, 기존의 학습 환경에 파라미터 최적화 단계(S210)만 추가하여 수행할 수 있다. 이때, 기존의 학습된 파라미터가 존재하지 않을 경우, 파라미터 최적화 단계(S210)를 생략한 학습을 통해 초기 학습을 진행한 후 파라미터 최적화 단계(S210)를 적용하도록 할 수 있다. 한편, 처음부터 파라미터 최적화 단계(S210)를 적용하여 학습을 진행할 경우 동일한 연산감소량에 비해 결과 정확도가 낮아질 수 있기 때문에, 기존의 학습된 컨볼루션 파라미터와 입력 데이터를 메모리에 로드(S200)한 후, 로드된 기존 컨볼루션 파라미터에 본 발명에서 제안하는 컨볼루션 파라미터 최적화 기법을 적용한다(S210).
- [0049] 다음으로, 상기 최적화가 적용된 파라미터와 입력 데이터를 컨볼루션 연산하여 추론(S220)을 실행하고, 파라미터 보정 단계(S230)를 거쳐 기존 파라미터를 업데이트한다(S240). 상기 파라미터 보정 단계(S230)에서는 역전파를 구하는 과정이 포함된다. 학습이 종료된 이후, 컨볼루션 파라미터 최적화 기법을 적용하여 파라미터를 저장하도록 한다. 파라미터 최적화 단계(S210), 추론 단계(S220), 파라미터 보정 단계(S230) 및 업데이트 단계(S240)는 목표 횟수에 도달할 때까지 반복적으로 수행될 수 있다.
- [0050] 도 3은 본 발명의 일 실시 예에 따른 도 2의 최적화 파라미터 생성을 위한 학습 프로세스를 세부화하여 도시한 도면이다.
- [0051] 도 3을 참조하면, 본 발명의 실시예에 따른 신경망 연산장치는, 먼저, 학습된 파라미터가 존재하는지 판단한다(S300). 다음으로, 학습된 파라미터가 존재하면 이를 메모리에 로드한다(S302). 이어서, 본 발명의 컨볼루션 파라미터 최적화(S303)를 적용할지 여부를 판단한다(S304). 컨볼루션 파라미터 최적화(S303)를 적용하지 않는 경우는 일반적인 딥러닝 학습과정과 동일하다.
- [0052] 상기 컨볼루션 파라미터 최적화 단계(S303)는 크기 파라미터 생성 단계(S306), 부호 파라미터 생성 단계(S308)를 포함하며, 비율 파라미터 최적화를 적용할지 여부를 판단하는 단계(S310) 및 비율 파라미터 최적화를 적용하는 것으로 판단한 경우 비율 파라미터를 생성하는 단계(S312)를 더 포함할 수 있다.
- [0053] 본 발명의 실시예에 따르면, 상기 컨볼루션 파라미터 최적화 단계(S303)가 완료되면, 신경망 연산장치는 추론 단계(S316)를 통해 최적화가 적용된 파라미터 및 입력 데이터로 추론을 실행하고, 보정 단계(S318)를 거쳐 기존 파라미터를 최적화 및 보정된 파라미터로 업데이트한다(S320). 이때, 미리 설정된 목표 횟수에 도달하였는지 판단(S322)하여, 그렇지 않은 경우 최적화 파라미터 최적화 단계(S303), 추론 단계(S316), 보정 단계(S318) 및 파라미터 업데이트 단계(S320)를 반복 수행한다.
- [0054] 도 4는 본 발명의 일 실시 예에 따른 추론 프로세스를 세부화하여 도시한 도면이다.
- [0055] 도 4를 참조하면, 본 발명의 실시예에 따른 신경망 연산장치는 최적화 파라미터를 메모리에 로드(S400)하고, 크기 파라미터 값이 존재하는지, 즉, 크기 파라미터 값이 0(zero)인지 아닌지를 판단한다(S410). 크기 파라미터 값이 0이면 이후 과정이 생략됨에 따라 추론 연산량을 감소시킬 수 있다.
- [0056] 다음으로, 크기 파라미터 값이 0이 아니면, 입력 데이터를 메모리에 로드(S420)하고, 부호 파라미터와 입력 데

이터를 비트 연산한다(S430). 그리고 입력 데이터 간 합 연산을 적용(S460)하고 크기 파라미터와 입력 데이터의 곱 연산(S470)을 수행한다. 비트 연산(S430) 이후, 비율 파라미터를 적용하는 경우(S440), 비율 파라미터를 참조해 비율 파라미터 대응상수를 크기 파라미터에 적용하는 단계(S470)를 더 포함할 수 있다.

[0057] 도 5는 본 발명의 일 실시 예에 따른 컨볼루션 파라미터 최적화 프로세스의 적용 예를 보여주는 도면이다.

[0058] 도 5를 참조하면, 일 실시 예에 따른 파라미터 최적화 프로세스는 크게 형태변환 단계(510)와 가지치기(pruning) 단계(520) 두 부분으로 구성된다. 파라미터 형태변환(510)은 기존의 컨볼루션 파라미터의 연산량을 줄이고 가지치기 용이한 형태로 변환하는 과정을 의미한다. 예를 들어, 도 5에 도시된 바와 같이 기존 파라미터(50)가 크기를 나타내는 크기 파라미터(Magnitude Parameter)(52)와 방향을 나타내는 부호 파라미터(Signed Parameter)(54)로 분리된다. 도 5에서는 기존 파라미터가 3개의 입력채널(50-1, 50-2, 50-3)을 가지는 예를 들어 도시하고 있다. 각 입력채널(50-1, 50-2, 50-3)은 다수의 파라미터 원소들(elements)로 이루어진다. 부호 파라미터(54)는 기존 파라미터(50)의 채널 별 원소들의 방향을 결정한 것이고, 크기 파라미터(52)는 기존 파라미터(50)의 채널 당 단일의 대표 값으로 가중치들을 최적화한 것이다.

[0059] 사용자는 크기 파라미터(52)의 각 구성원소들의 비트 수를 임의로 설정할 수 있으며, 정수(Integer)와 소수(Floating) 등으로 표현될 수 있다. 비트 수가 많아질수록 원래의 컨볼루션 연산 결과와 동일한 결과를 가지게 된다. 부호 파라미터(54) 내 각 구성원소의 비트 수는 1비트(Bit)이며, 이는 양(+)의 방향과 음(-)의 방향만을 나타낸다. 기존 파라미터(50)에서 크기 파라미터(52) 및 부호 파라미터(54)를 분리하기 위해 다양한 방법을 적용할 수 있다. 예를 들어, 기존 파라미터(50)에서 컨볼루션 레이어의 입력채널의 평균과 각 파라미터의 최상위 비트를 통해 형태변환을 수행할 수 있다.

[0060] 크기 파라미터(52)로의 형태 변환 방법 중 하나는 입력채널의 구성원소들(가중치 파라미터들)의 평균 값을 이용하는 것이다. 예를 들어, i 개의 구성원소 W_i , j 개의 입력채널 I_j , k 개의 출력채널 O_k 이고, 크기 파라미터 $M = \{M_1, M_2, \dots, M_j\}$ 이면, M_j 는 수학식 1과 같다.

수학식 1

$$M_j = \frac{1}{i} \sum_{n=1}^i |W_n|$$

[0061]

[0062] 부호 파라미터(54)로의 형태 변환 방법 중 하나는 채널의 구성원소(가중치 파라미터들)가 0보다 크거나 같은 값이면 0 값을 할당하고, 0보다 작으면 1 값을 부여하는 방식이다. 즉, 부호 파라미터 $S = \{S_1, S_2, \dots, S_i\}$ 이면, S_i 는 수학식 2와 같다.

수학식 2

$$S_i = \begin{cases} 0, & \text{if } W_i \geq 0 \\ 1, & \text{if } W_i < 0 \end{cases}$$

[0063]

[0064] 형태변환 과정을 수행한 후 불필요한 파라미터들을 제거하는 가지치기 단계(520)를 진행한다. 가지치기 단계(520)에서는 중요도가 낮은 파라미터를 선별하여 그 값을 0(zero)로 만들어 연산을 생략할 수 있게 하여 전체 연산량을 줄이는 효과를 얻도록 한다. 이때, 크기 파라미터를 크기에 따라 가지치기하여 최적화된 파라미터를 생성한다. 예를 들어, 소정의 크기 파라미터 값이 미리 설정된 기준 값보다 적으면 해당 채널의 크기 파라미터를 가지치기한다.

[0065] 여기서 중요한 점은 이 가지치기 과정을 컨볼루션 레이어 전체 파라미터에 적용하는 것이 아니라, 형태변환된 크기 파라미터(52)에만 적용하도록 하는 것이다. 크기 파라미터(52)만 가지치기(520)를 적용함으로써 생략된 파라미터가 컨볼루션 레이어의 각 입력채널의 모든 파라미터에 영향을 끼치게 되어 가지치기(520)의 효과를 극대

화할 수 있다.

[0066] 불필요한 파라미터를 생략하기 위해 다양한 방법들이 적용될 수 있다. 예를 들어, 컨볼루션 레이어 입력채널 및 출력채널 별 평균값과 분포를 이용해서 가지치기를 수행한다. 각 레이어 별로 사용자가 임의로 정한 상수가 존재하며, 이 상수는 레이어 별 컨볼루션 파라미터 분포에 따라 정해진다. 값을 생략하기 위한 기준 값으로 입력채널의 크기 파라미터들의 평균값 혹은 입력채널 및 출력채널의 크기 파라미터의 평균에 가중치를 곱한 값을 사용할 수 있다. 레이어 상수는 0~1 값을 가지는데, 레이어 별 컨볼루션 파라미터를 연속균등분포 혹은 정규분포라 가정할 때, 50%의 크기 파라미터를 생략하기 위해서 레이어 상수는 0.5~0.7일 수 있다. 이 값을 참조하여 실제 파라미터의 분포를 고려해 레이어 상수를 정하도록 한다.

[0067] $M_{LOI} = \{M_{111}, M_{112}, \dots, M_{lkj}\}$, $L = \{C_1, C_2, \dots, C_l\}$ 이고, C는 레이어 상수인 경우, 입력채널 별 기준을 적용한 크기 파라미터 가지치기 방식은 다음 수학적 식 3과 같다.

수학적 식 3

$$M'_{LOI} = \begin{cases} M_{lkj}, \text{if } M_{lkj} \geq \frac{1}{j} \sum_{n=1}^j M_{lkn} \in L_l \\ 0, \text{if } M_{lkj} < \frac{1}{j} \sum_{n=1}^j M_{lkn} \in L_l \end{cases}$$

[0068]

[0069] 출력채널 별 기준을 적용하나 크기 파라미터 가지치기 방식은 다음 수학적 식 4와 같다.

수학적 식 4

$$M'_{LOI} = \begin{cases} M_{lkj}, \text{if } M_{lkj} \geq \frac{1}{k} \frac{1}{j} \sum_{m=1}^k \sum_{n=1}^j M_{lmn} \in L_l \\ 0, \text{if } M_{lkj} < \frac{1}{k} \frac{1}{j} \sum_{m=1}^k \sum_{n=1}^j M_{lmn} \in L_l \end{cases}$$

[0070]

[0071] 이하, 전술한 수학적 식 1 내지 4를 적용한 컨볼루션 파라미터 최적화 계산 예시를 설명한다.

[0072] i=4, j=2, k=2인 기존 컨볼루션 파라미터를 대상으로 크기 파라미터 및 부호 파라미터를 계산하고, 크기 파라미터를 대상으로 가지치기를 한다고 가정한다. 이때, $O_1 = \{I_{11}, I_{12}\} = \{[W_{111}, W_{112}, W_{113}, W_{114}], [W_{121}, W_{122}, W_{123}, W_{124}]\} = \{[-0.5, 1.0, 0.5, -1.0], [1.0, -1.5, 1.0, -1.5]\}$, $O_2 = \{I_{21}, I_{22}\} = \{[W_{211}, W_{212}, W_{213}, W_{214}], [W_{231}, W_{222}, W_{223}, W_{224}]\} = \{[1.5, -2.0, -1.5, 2.0], [-2.0, -2.5, 2.0, 2.5]\}$ 라고 가정한다.

[0073] 수학적 식 1을 적용하여 계산한 크기 파라미터 $M = \{M_{11}, M_{12}, M_{21}, M_{22}\} = \{0.75, 1.25, 1.75, 2.25\}$ 이다. 즉, $M_{11} = 1/4 (0.5 + 1.0 + 0.5 + 1.0) = 0.75$, $M_{12} = 1/4 (1.0 + 1.5 + 1.0 + 1.5) = 1.25$, $M_{21} = 1/4 (1.5 + 2.0 + 1.5 + 2.0) = 1.75$, $M_{22} = 1/4 (2.0 + 2.5 + 2.0 + 2.5) = 2.25$ 이다.

[0074] 수학적 식 2를 적용하여 계산한 부호 파라미터 $S = \{S_1, S_2, S_3, S_4\} = \{[1, 0, 0, 1], [0, 1, 0, 1], [0, 1, 1, 0], [1, 1, 0, 0]\}$ 이다.

[0075] 크기 파라미터에 대한 가지치기 적용 예를 들면 다음과 같다. j=2, k=2, l=1인 기존 파라미터의 경우, 기존 파라미터로부터 분리된 크기 파라미터 $M = \{M_{111}, M_{112}, M_{121}, M_{122}\} = \{0.75, 1.25, 1.75, 2.25\}$ 이고, 레이어 상수 $L_1=0.9$ 라 하면, 수학적 식 3의 입력채널 별로 적용하면, 가지치기된 크기 파라미터 $M' = \{M_{111}, M_{112}, M_{121}, M_{122}\} = \{0, 1.25, 0, 2.25\}$ 이다. 예를 들어, M_{112} 의 경우, $1.25 < 1/2(0.75+1.25) \cdot 0.9$ 이므로, 수학적 식 3에 의해 $M_{112} = 0$ 이다.

- [0076] 수학적식 4의 입력채널 및 출력채널 별로 적용하면, 가지치기된 크기 파라미터 $M' = \{M_{111}, M_{112}, M_{121}, M_{122}\} = \{0, 0, 1.75, 2.25\}$ 이다. M_{112} 의 경우, $1.25 < 1/4(0.75+1.25+1.75+2.25) \cdot 0.9$ 이므로, 수학적식 4에 의해 $M_{112} = 0$ 이다. 입력채널 별로 적용할지 입력 및 출력채널 별로 적용할지는 분포를 따라 결정될 수 있다.
- [0077] 도 6은 일반적인 컨볼루션 연산과 본 발명의 일 실시 예에 따른 컨볼루션 연산을 비교한 도면이다.
- [0078] 도 6을 참조하면, 일반적인 컨볼루션 연산(60)은 곱 연산(600)과 합 연산(602)이 컨볼루션 파라미터 필터 크기 만큼 존재한다. 하지만 일 실시 예에 따른 파라미터 최적화를 적용한 컨볼루션 연산(62)은 비트 연산(620)을 통해 데이터의 방향을 정하고, 필터 크기만큼 모두 합 연산(622)한 후, 단 한 번의 곱 연산(624)으로 결과를 도출할 수 있다. 예를 들어, 도 6에 도시된 바와 같이 일반적인 컨볼루션 연산(60)은 곱 연산(600)은 9번이나, 일 실시 예에 따른 파라미터 최적화를 적용한 컨볼루션 연산(62)은 단 1번임을 확인할 수 있다. 이는 필터 크기만큼의 곱 연산 이득이 있음을 의미한다.
- [0079] 또한, 컨볼루션 파라미터를 형태변환 하지 않고 가지치기를 적용하게 되면 생략해야할 연산인지 아닌지를 판단하는데에 필터 크기만큼의 비교연산이 필요하다. 도 6의 예시의 일반적인 컨볼루션 연산(60) 경우는 총 9번의 비교연산이 필요하다. 그러나 본 발명의 파라미터 최적화 기법을 적용한 컨볼루션 연산(62)의 경우, 컨볼루션 레이어 입력채널 당 단 한 번의 비교만으로 연산을 생략할 것인지 여부를 판단할 수 있다. 이러한 연산량의 감소는 기존의 컨볼루션 연산(60)보다 적은 연산량을 요구한다. 따라서 효율적으로 하드웨어를 설계할 수 있다.
- [0080] 도 7은 본 발명의 일 실시 예에 따른 파라미터 최적화 이후 컨볼루션 연산의 이점을 보여주는 도면이다.
- [0081] 도 7을 참조하면, 입력 데이터와 크기 파라미터 및 부호 파라미터를 컨볼루션 연산하여 출력 데이터를 생성하는데, 크기 파라미터 값이 0이면 연산을 수행할 필요가 없게 되어 채널 단위로 연산량을 크게 줄일 수 있다.
- [0082] 도 8은 본 발명의 다른 실시 예에 따른 컨볼루션 파라미터 최적화 프로세스의 적용 예를 보여주는 도면이다.
- [0083] 컨볼루션 파라미터 최적화의 목표는 최대한의 최적화(연산감소)를 가지고 최대한의 정확도를 얻는 것이다. 연산 감소와 정확도는 보편적으로 반비례 관계이다. 실제 응용단계에서는 만족해야할 성능(속도 및 정확도)이 존재하는데, 이를 사용자가 조절하기는 쉽지 않다. 이를 해결하기 위해 본 발명은 비율 파라미터(Scale Parameter)를 이용한 최적화 방식을 제안한다. 이는 도 5를 참조로 하여 전술한 최적화 방식에 비율 파라미터가 추가된 것이다.
- [0084] 형태변환 과정(510)에서 기존의 파라미터(50)의 형태를 크기 파라미터(52), 부호 파라미터(54) 및 비율 파라미터(58)로 변환하고, 크기 파라미터(52)에만 가지치기(520)를 적용한다. 추론 시에는 비율 파라미터(58)를 참고하여 크기 파라미터에 가중치를 적용해 기존 연산결과의 오차를 줄이도록 한다.
- [0085] 비율 파라미터(58)는 크기 파라미터(52)가 얼마나 중요한지를 나타낸다. 즉, 컨볼루션 레이어 입력채널 파라미터 간의 중요도를 비교하여 상수를 차등 적용하는 방식이다. 비율 파라미터(58)의 하나의 구성원소에 대해 사용자는 원하는 비트를 할당할 수 있으며, 사용자는 이를 통해 최적화 정도를 조절할 수 있다. 비율 파라미터(58)는 가중치를 의미하지 않으며, 비율 파라미터(58)를 참조하는 상수 값들이 존재한다. 이 값들은 하드웨어로 구현되기 쉬운 값이 할당되어 있어 비트 연산만으로 구현 가능하게 한다. 크기 파라미터(52)의 값을 바로 적용하지 않고 비율 파라미터(58)를 통한 가중치를 적용하여 최적화를 진행하게 되면, 기존의 최적화 전 파라미터와의 유사도가 높아져 정확도가 올라가게 된다.
- [0086] 도 9는 비율 파라미터 최적화를 설명하기 위해 크기 파라미터만을 적용한 파라미터 분포와 비율 파라미터를 추가 적용한 파라미터 분포를 비교한 도면이다.
- [0087] 도 8 및 도 9를 참조하면, 비율 파라미터(58)를 생성하기 위해 다양한 방법이 적용될 수 있는데, 본 발명에서는 컨볼루션 파라미터의 분포를 통해 비율 파라미터(58)를 생성한다. 비율 파라미터 비트 수를 b 라고 가정하면, 2^b 개의 사용자가 정의한 임의의 상수가 존재하며, 이를 비율 파라미터 대응상수라 한다. 비율 파라미터 대응상수는 컨볼루션 파라미터 분포에 근거하여 사용자가 임의로 정할 수 있으며, 비율 파라미터에 대응되는 상수이다.
- [0088] 비율 파라미터(58)를 연산하는 예는 다음과 같다.
- [0089] 비율 파라미터 비트 수 = b ,
- [0090] $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_t\}$, $t = \{1, 2, \dots, 2^b - 1\}$, $\lambda_t > \lambda_{t+1}$ 이고, 비율 파라미터 $SC = \{SC_1, SC_2, \dots, SC_t\}$ 이

면, SC_i 는 다음 수학적 식 5와 같다.

수학적 식 5

$$SC_i = \begin{cases} 8, & \text{if } \frac{1}{i} \sum_{n=1}^i |W_n| E \lambda_t \mid W_n \mid E \frac{1}{i} \sum_{n=1}^i |W_n| E \lambda_{t+1}, kt < 2^b - 1 \\ 0, & \text{if } |W_n| > \frac{1}{i} \sum_{n=1}^i |W_n| E \lambda_t, t=1 \\ 2^b - 1, & \text{if } |W_n| \leq \frac{1}{i} \sum_{n=1}^i |W_n| E \lambda_t, t=1 \end{cases}$$

[0091]

[0092] 추론 시에는 비율 파라미터 대응상수가 크기 파라미터에 곱해진 값이 실제 연산 값으로 사용된다.

[0093] 이하, 전술한 수학적 식 5를 적용한 비율 파라미터 최적화 계산 예시를 설명한다.

[0094] $i=4$, $j=2$, $k=2$ 인 기준 컨볼루션 파라미터를 대상으로 비율 파라미터를 계산한다고 가정한다. 비율 파라미터 비트 수 $b=2$, $t=\{1, 2, 3\}$, $\lambda=\{0.9, 0.7, 0.5\}$ 라고 가정한다. 이때, $O_1=\{I_{11}, I_{12}\}=\{W_{111}, W_{112}, W_{113}, W_{114}, W_{121}, W_{122}, W_{123}, W_{124}\}=\{-0.5, 1.0, 0.5, -1.0, 1.0, -1.5, 1.0, -1.5\}$, $O_2=\{I_{21}, I_{22}\}=\{W_{211}, W_{212}, W_{213}, W_{214}, W_{231}, W_{222}, W_{223}, W_{224}\}=\{1.5, -2.0, -1.5, 2.0, -2.0, -2.5, 2.0, 2.5\}$ 라고 가정한다.

[0095] 수학적 식 5에 의해, $SC_{111}=2$ 이다. $1/4(0.5+1.0+0.5+1.0) \cdot 0.7 > |-0.5| > 1/4(0.5+1.0+0.5+1.0) \cdot 0.5$ 이기 때문이다. 이런 식으로 계산된 비율 파라미터 $SC_i=(SC_{111}, SC_{112}, SC_{113}, SC_{114}, SC_{121}, SC_{122}, SC_{123}, SC_{124}, SC_{211}, SC_{212}, SC_{213}, SC_{214}, SC_{221}, SC_{222}, SC_{223}, SC_{224})=\{2, 0, 2, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0\}$ 이다.

[0096] 도 10은 본 발명의 일 실시 예에 따른 최적화된 파라미터를 이용한 컨볼루션 연산의 예를 도시한 도면이다.

[0097] 도면에 도시된 바와 같이, 크기 파라미터(M), 부호 파라미터(S), 비율 파라미터(SC), 그리고, 비율 파라미터 대응상수(C)가 존재하고, 입력 데이터(In)가 입력될 때의 컨볼루션 연산은 도 10에 도시된 바와 같다. 이때, $M=\{0.75\}$, $S=\{0, 0, 1, 1\}$, $SC=\{0, 0, 1, 1\}$, $C=\{1.0, 0.5\}$ 이고, 입력 데이터 $In=\{1.2, -2.0, 0.5, 3.1\}$ 인 경우를 예로 들었다.

[0098] 도 10을 참조한 예시에서는 비율 파라미터 대응상수(C)는 2의 배수로 정하여 일반적인 산술연산이 아닌 쉬프트(Shift) 비트연산으로 결과를 도출할 수 있도록 한다. 이 경우, 하드웨어 구현 시 이점을 가진다.

[0099] 도 11은 본 발명의 일 실시 예에 따른 신경망 연산장치의 구성을 도시한 도면이다.

[0100] 도면에 도시된 바와 같이, 본 발명의 실시예에 따른 신경망 연산장치(1)는 입력부(10), 프로세서(12), 메모리(14) 및 출력부(16)를 포함할 수 있다.

[0101] 상기 입력부(10)는 사용자 명령을 포함한 데이터를 입력받고, 입력된 데이터는 메모리(18)에 저장된다. 상기 프로세서(12)는 신경망 연산을 수행하고 연산 결과를 메모리(18)에 저장하며, 출력부(16)를 통해 출력할 수 있다. 상기 메모리(14)에는 입력 데이터 및 파라미터가 로드될 수 있다. 일 실시 예에 따른 프로세서(12)는 파라미터 최적화부(120), 추론부(122), 보정부(124) 및 업데이트부(126)를 포함한다.

[0102] 상기 파라미터 최적화부(120)는 기존 파라미터를 형태 변환하고, 형태 변환된 파라미터들 중에 크기 파라미터를 가지치기한다. 기존 파라미터는 이미 학습된 파라미터일 수 있고, 연산을 위해 메모리에 로드된다.

[0103] 본 발명의 실시 예에 따른 파라미터 최적화부(120)는 기존 파라미터를 크기 파라미터 및 부호 파라미터로 형태 변환하거나, 크기 파라미터, 부호 파라미터 및 비율 파라미터로 형태 변환한다. 부호 파라미터는 기존 파라미터의 채널 별 원소들의 방향을 결정하는 것이고, 크기 파라미터는 기존 파라미터의 채널 당 단일의 대표 값으로 가중치들을 최적화한 것이다. 비율 파라미터는 크기 파라미터가 얼마나 중요한지를 나타내는 것으로, 추론 시에 비율 파라미터를 참고하여 크기 파라미터에 차등적 가중치를 적용할 수 있다.

[0104] 본 발명의 실시예에서, 상기 가지치기는 형태 변환된 파라미터들 중에서 크기 파라미터만을 대상으로 수행된다.

가지치기는 중요도가 낮은 크기 파라미터를 선별하여 그 값을 0(zero)로 만드는 것을 의미한다. 가지치기된 파라미터는 채널 단위로 컨볼루션 연산을 생략할 수 있어서 전체 연산량을 줄이는 효과를 얻는다. 중요한 점은 이 가지치기 과정을 컨볼루션 레이어의 각 입력채널을 구성하는 모든 파라미터 원소에 영향을 끼치게 되어 가지치기의 효과를 극대화할 수 있다.

[0105] 또한, 상기 추론부(122)는 최적화된 파라미터 및 입력채널 데이터를 컨볼루션 연산하여 추론한다. 여기서, 상기 추론부(122)는 최적화 파라미터에 포함된 크기 파라미터의 값이 0인지 판단하고 크기 파라미터의 값이 0이면 컨볼루션 연산 과정을 생략할 수 있다. 또한, 상기 추론부(122)는 비율 파라미터가 존재하면 비율 파라미터를 이용하여 크기 파라미터에 가중치를 차등 반영함에 따라 연산결과와 오차를 줄일 수 있다. 본 발명의 실시예에서, 상기 보정부(124)는 최적화된 파라미터를 보정하고, 업데이트부(126)는 기존 파라미터를 보정된 최적화 파라미터로 업데이트한다.

[0106] 도 12는 본 발명의 일 실시 예에 따른 도 11의 파라미터 최적화부의 세부 구성을 도시한 도면이다.

[0107] 도 11 및 도 12를 참조하면, 본 발명의 실시예에 따른 파라미터 최적화부(120)는 크기 파라미터 변환부(1200), 부호 파라미터 변환부(1202), 파라미터 가지치기부(1204) 및 비율 파라미터 변환부(1206)를 포함할 수 있다.

[0108] 상기 부호 파라미터 변환부(1202)는 기존 파라미터를 부호 파라미터로 형태 변환한다. 상기 크기 파라미터 변환부(1200)는 기존 파라미터를 채널 당 단일의 값을 가지는 크기 파라미터로 형태 변환한다. 상기 비율 파라미터 변환부(1206)는 기존 파라미터를 비율 파라미터로 형태 변환한다.

[0109] 또한, 상기 파라미터 가지치기부(1204)는 형태 변환된 크기 파라미터를 가지치기하여 최적화된 파라미터를 생성한다. 예를 들어, 소정의 크기 파라미터 값이 미리 설정된 기준 값보다 적으면 해당 채널의 크기 파라미터를 가지치기한다.

[0110] 본 발명의 실시 예에 따른 파라미터 가지치기부(1204)는 입력채널 별 크기 파라미터의 평균 값 및 크기 분포 또는 입력 및 출력채널 별 크기 파라미터의 평균 값 및 크기 분포를 이용하여 기준 값을 계산하고, 계산된 기준 값보다 적은 값을 가진 크기 파라미터 값을 0으로 만들어 해당 채널의 컨볼루션 연산을 생략한다.

[0111] 또한, 상기 파라미터 가지치기부(1204)는 소정의 레이어를 구성하는 입력채널의 크기 파라미터의 평균 값 및 크기 분포를 이용하여 가지치기하는 경우, 입력채널들의 크기 파라미터의 평균 값에 레이어 별 상수를 곱한 값으로 기준 값을 계산하고, 소정의 입력채널의 크기 파라미터 값이 기준 값보다 작으면 해당 채널의 크기 파라미터 값을 0으로 변경하여 가지치기할 수 있다.

[0112] 또한, 상기 파라미터 가지치기부(1204)는 소정의 레이어를 구성하는 입력 및 출력채널의 크기 파라미터의 평균 값 및 크기 분포를 이용하여 가지치기하는 경우, 입력채널들 및 출력채널들의 크기 파라미터의 평균 값에 레이어 별 상수를 곱한 값으로 기준 값을 계산하고, 소정의 입력채널의 크기 파라미터 값이 기준 값보다 작으면 해당 채널의 크기 파라미터 값을 0으로 변경하여 가지치기할 수 있다.

[0113] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술분야에서 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

[0114] 10 : 입력부
12 : 프로세서
14 : 메모리
16 : 출력부
120 : 파라미터 최적화부
122 : 추론부
124 : 보정부

126 : 업데이트부

1200 : 크기 파라미터 변환부

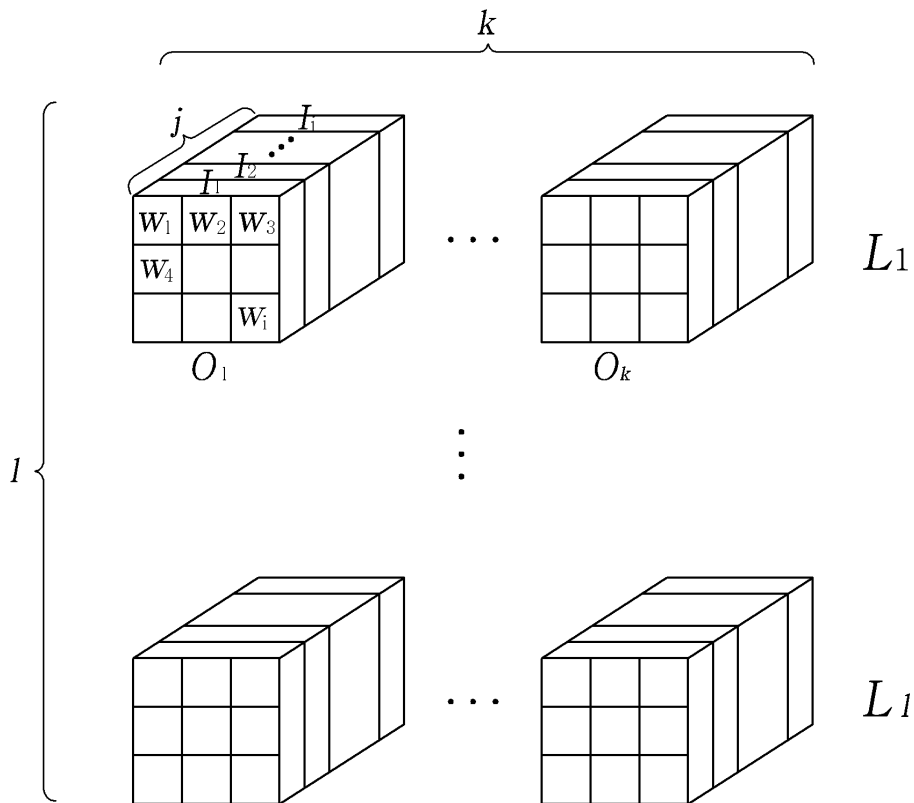
1202 : 부호 파라미터 변환부

1204 : 파라미터 가지치기부

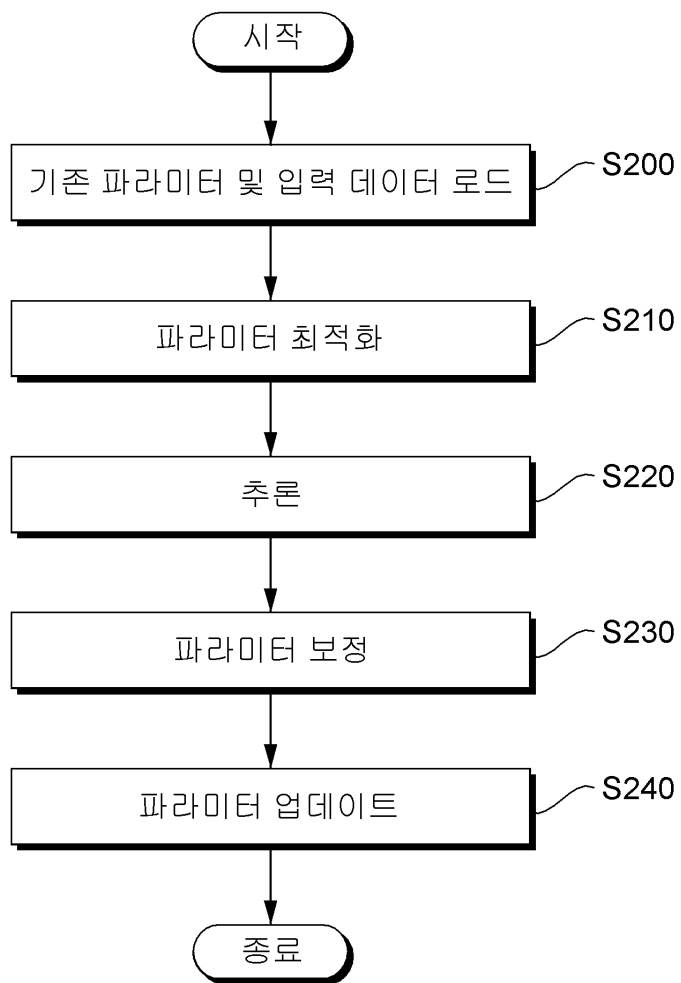
1206 : 비율 파라미터 변환부

도면

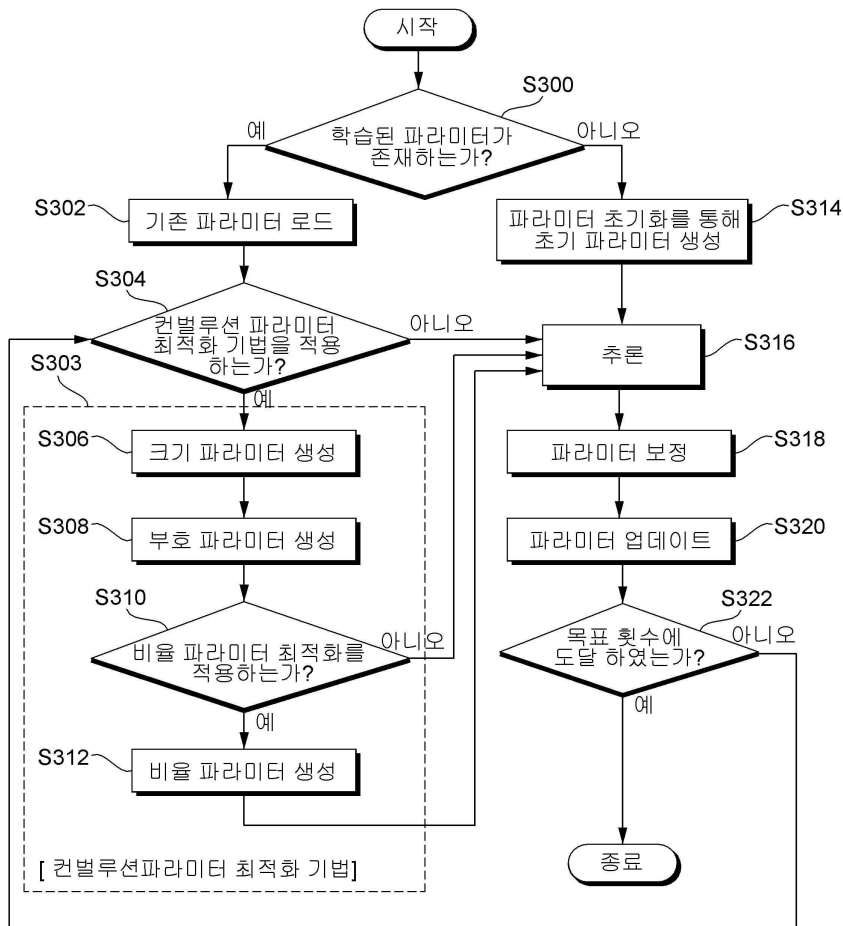
도면1



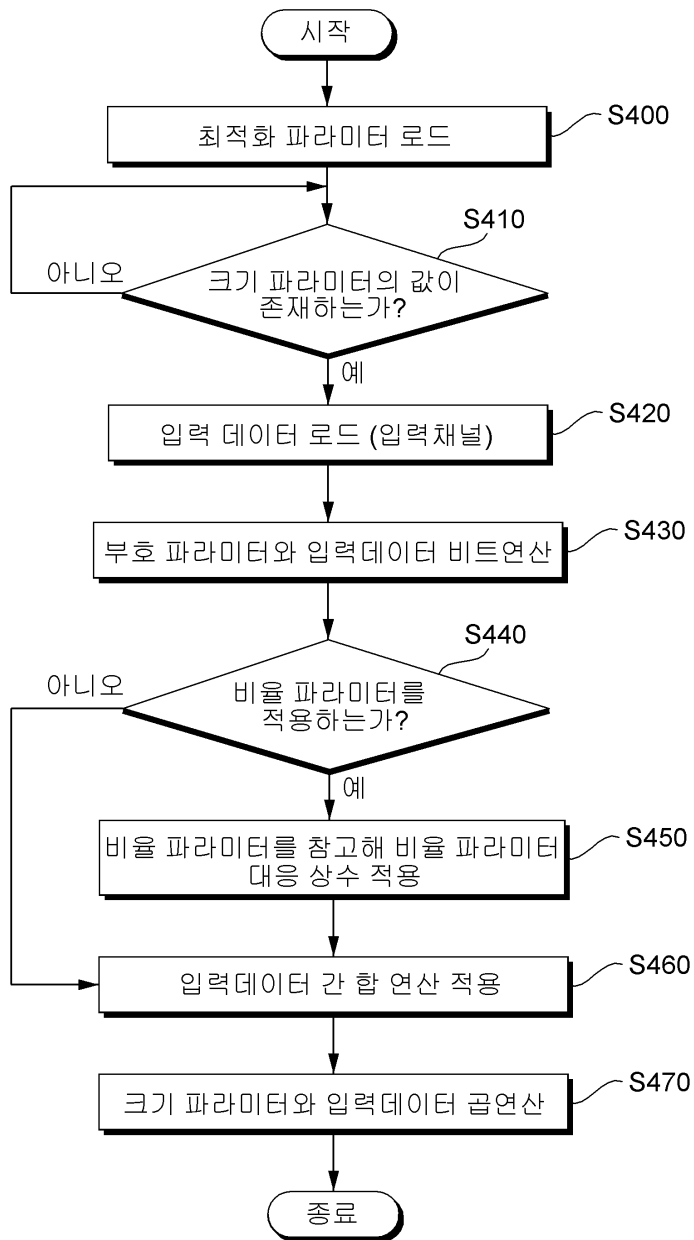
도면2



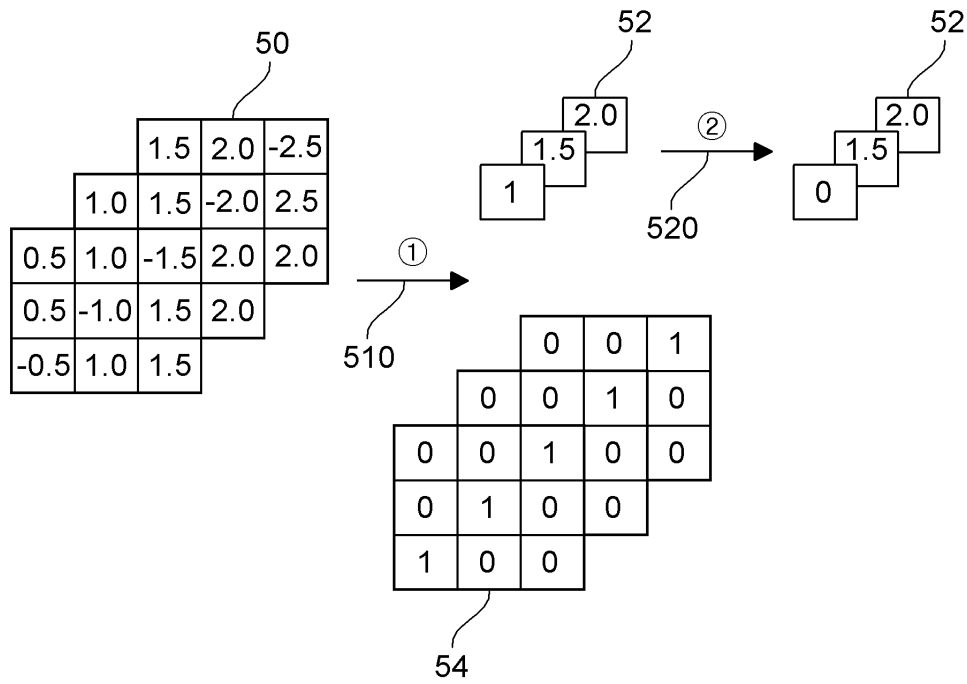
도면3



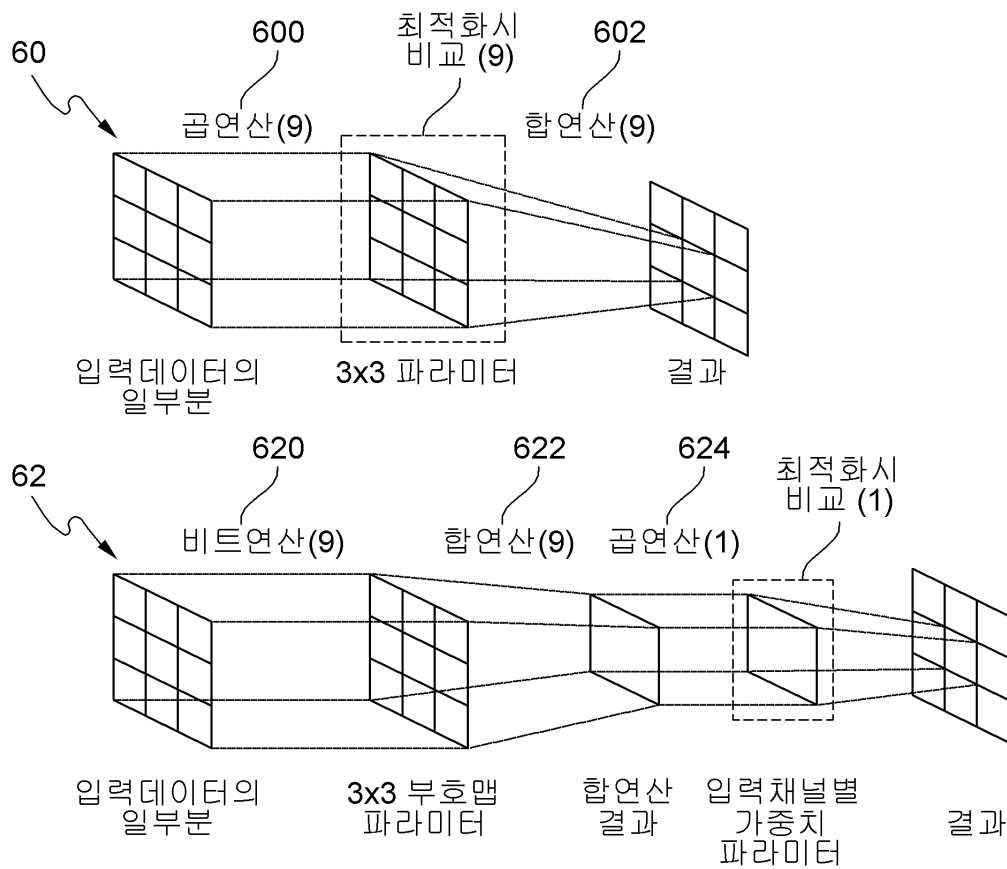
도면4



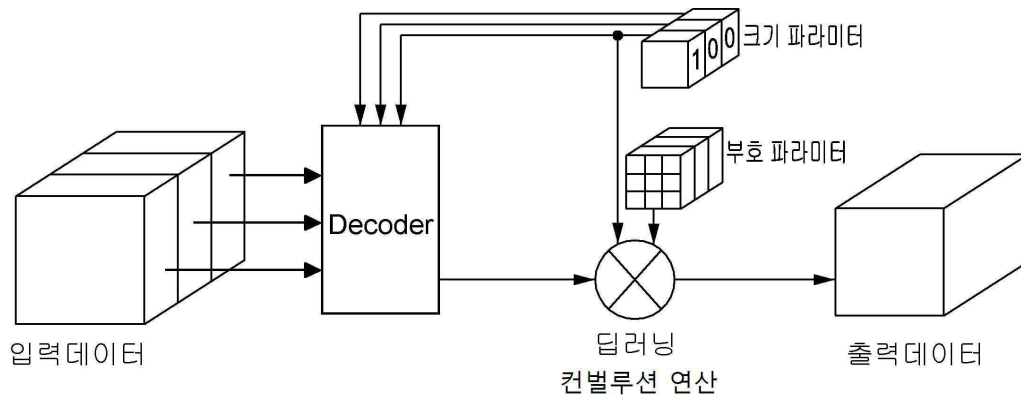
도면5



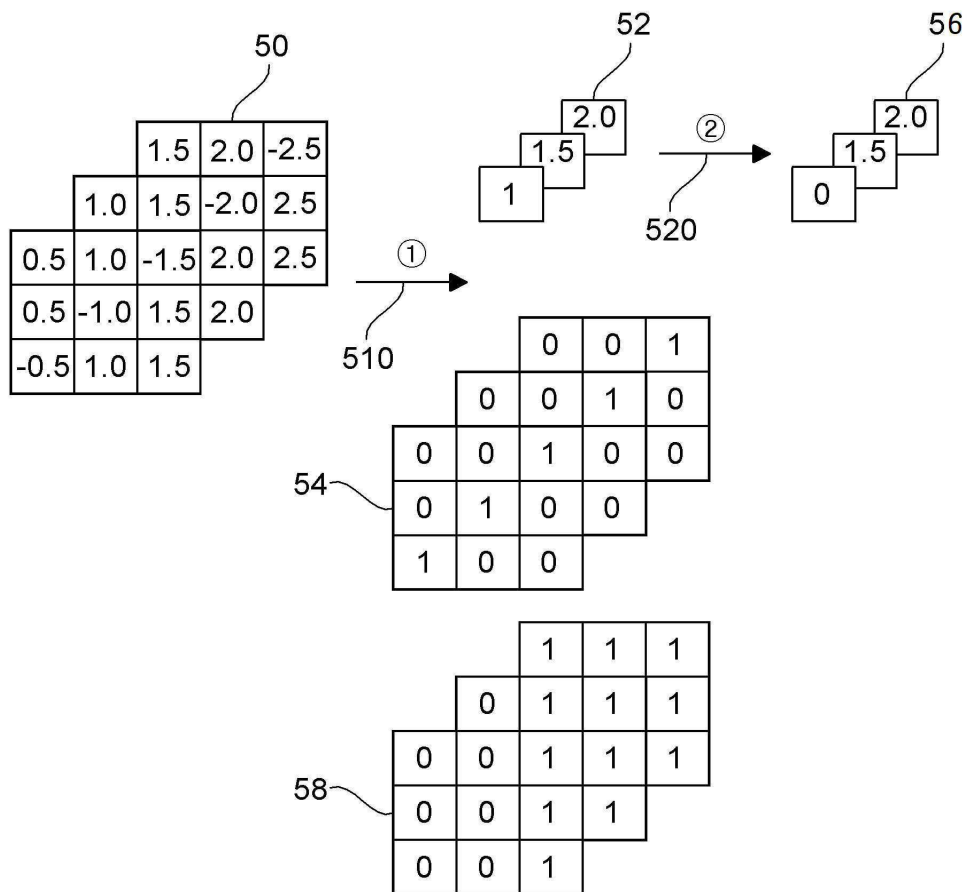
도면6



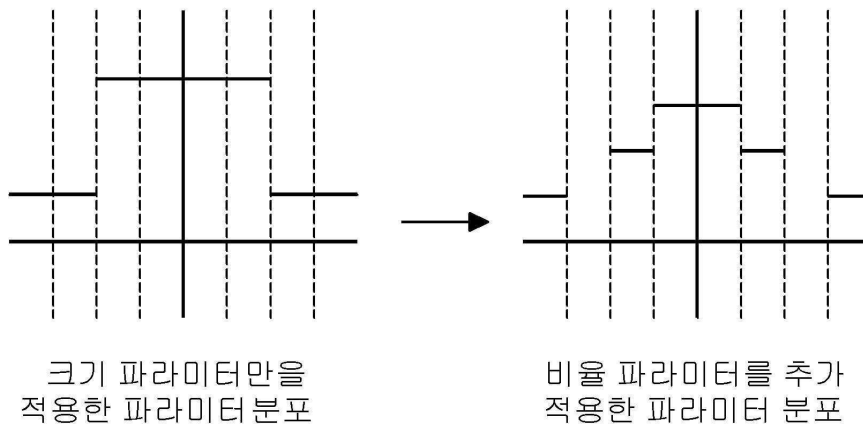
도면7



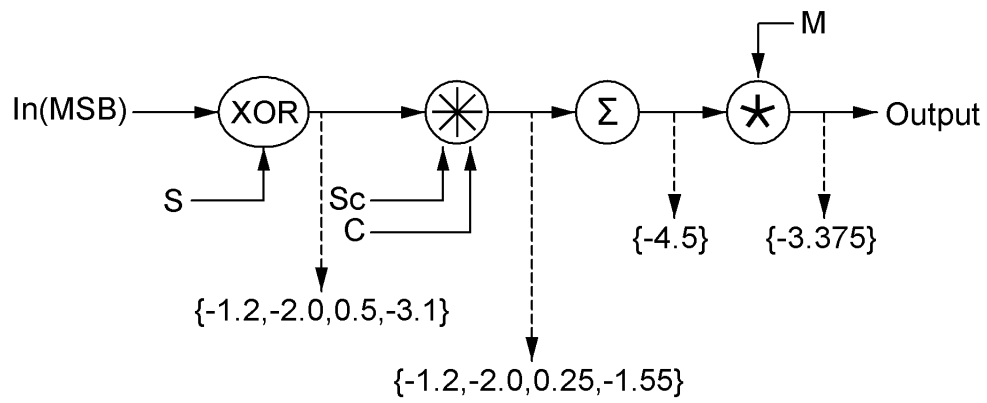
도면8



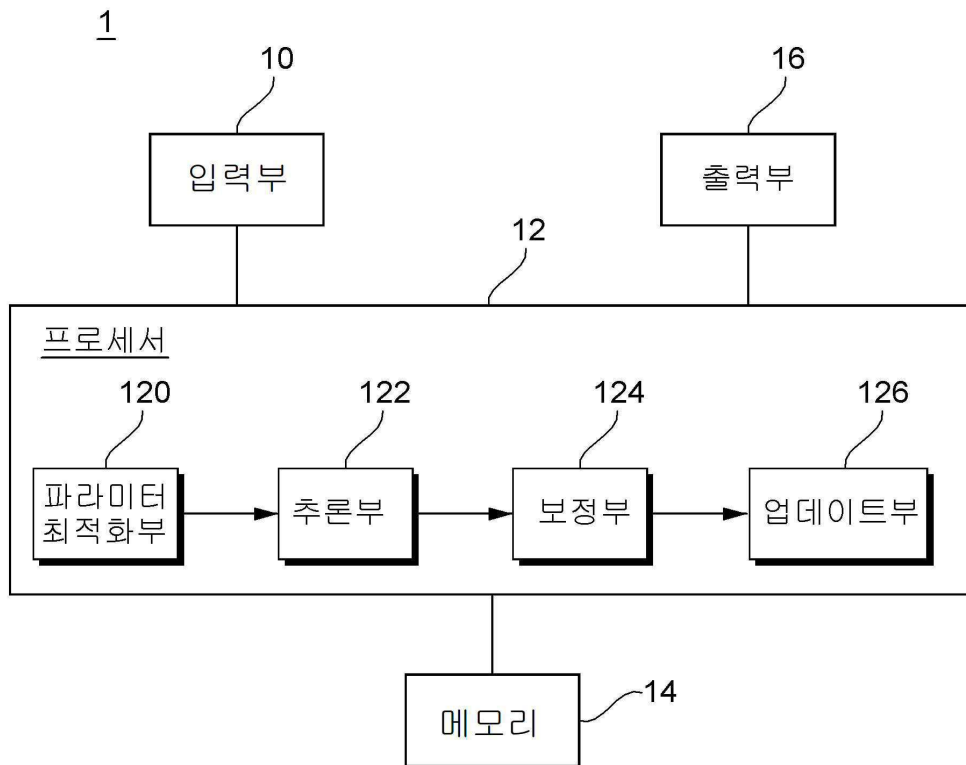
도면9



도면10



도면11



도면12

