



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0050664
(43) 공개일자 2023년04월17일

(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 10-2021-0133794
(22) 출원일자 2021년10월08일
심사청구일자 2021년10월08일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
배성호
경기도 용인시 기흥구 예현로 15, 102동 501호(서천동, 서그내마을에스케이아파트)
조인천
경기도 수원시 영통구 청명로43번길 18, 204호
(74) 대리인
두호특허법인

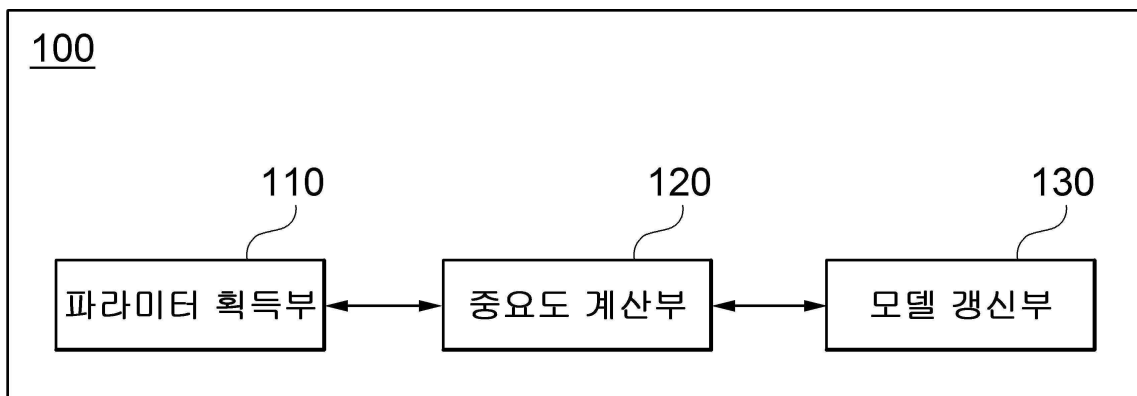
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법 및 장치

(57) 요약

심층신경망의 선형변환 파라미터를 이용한 가지치기 방법 및 장치, 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법 및 장치가 개시된다. 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법은 하나 이상의 프로세서들, 및 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 컨볼루션 레이어(Convolutional Layer) 및 정규화 레이어(Normalization Layer)를 포함하는 심층 신경망 모델에서 컨볼루션 레이어를 구성하는 복수의 필터 각각의 가중치 및 정규화 레이어의 선형변환(Linear Transform) 파라미터를 획득하는 단계, 가중치 및 선형변환 파라미터에 기초하여 복수의 필터 각각의 중요도를 계산하는 단계 및 필터 중요도에 기초하여 복수의 필터 중 일부를 제거하여 심층 신경망 모델을 갱신하는 단계를 포함한다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	1711125817
과제번호	2015-0-00742-007
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	라이프 컴패니온쉽 경험을 위한 지능형 인터랙션 융합 연구
기 여 율	1/2
과제수행기관명	성균관대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126005
과제번호	2021-0-00017-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	인공지능산업원천기술개발(R&D)
연구과제명	엣지 디바이스를 활용한 이상 행동 탐지 기술 연구
기 여 율	1/2
과제수행기관명	성균관대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명세서

청구범위

청구항 1

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서,

컨볼루션 레이어(Convolutional Layer) 및 정규화 레이어(Normalization Layer)를 포함하는 심층 신경망 모델에서 상기 컨볼루션 레이어를 구성하는 복수의 필터 각각의 가중치 및 상기 정규화 레이어의 선형변환(Linear Transform) 파라미터를 획득하는 단계;

상기 가중치 및 상기 선형변환 파라미터에 기초하여 상기 복수의 필터 각각의 중요도를 계산하는 단계; 및

상기 필터 중요도에 기초하여 상기 복수의 필터 중 일부를 제거하여 상기 심층 신경망 모델을 갱신하는 단계를 포함하는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법.

청구항 2

청구항 1에 있어서,

상기 선형 변환 파라미터는, 스케일(Scale) 값 및 바이어스(Bias) 값을 포함하는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법.

청구항 3

청구항 2에 있어서,

상기 필터 중요도는, 상기 가중치, 상기 스케일 값 및 상기 바이어스 값에 기초하여 하기 수학식에 의해 계산되는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법.

(수학식)

$$F_i = \|\gamma_i\| \|w_i\| + \alpha \text{sign}(\beta_i) \|\beta_i\|$$

F_i

: 상기 복수의 필터 중 i번째 필터의 필터 중요도

w_i :

상기 i번째 필터에 대한 가중치

γ_i

: 상기 i번째 필터에 대응하는 선형변환 파라미터의 스케일 값

β_i

: 상기 i번째 필터에 대응하는 선형변환 파라미터의 바이어스 값

α β_i

: β_i 에 대한 스케일 조정 파라미터

청구항 4

청구항 1에 있어서,

상기 심층 신경망 모델에 포함된 하나 이상의 잔여 연결(Residual Connection)에 기초하여, 상기 복수의 필터를 복수의 그룹으로 그룹화하는 단계; 및

상기 복수의 필터 중 동일한 그룹에 포함된 하나 이상의 필터 각각의 중요도에 기초하여 상기 복수의 그룹 각각에 대한 그룹 중요도를 계산하는 단계를 더 포함하고,

상기 갱신하는 단계는, 상기 필터 중요도 및 상기 그룹 중요도에 기초하여 상기 심층 신경망 모델을 갱신하는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법.

청구항 5

청구항 4에 있어서,

상기 그룹 중요도는, 상기 그룹화된 필터의 평균에 기초하여 하기 수학식에 의해 산출되는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법.

(수학식)

$$\overline{F}^g_i = \frac{1}{n(L_g)} \sum_{i \in L_g} F^l_i$$

$\overline{F}^g_{ij}$

: 상기 복수의 그룹 중 g번째 그룹에 대한 그룹 중요도

L_g : 상기 g번째 그룹에 포함된 레이어들의 인덱스 집합

$n(L_g)$: 상기 g번째 그룹에 포함된 레이어의 개수

F^l_i : 상기 g번째 그룹 내 l 번째 레이어의 에 포함된 필터의 중요도

청구항 6

컨볼루션 레이어(Convolutional Layer) 및 정규화 레이어(Normalization Layer)를 포함하는 심층 신경망 모델에서 상기 컨볼루션 레이어를 구성하는 복수의 필터 각각의 가중치 및 상기 정규화 레이어의 선형변환(Linear Transform) 파라미터를 획득하는 파라미터 획득부;

상기 가중치 및 상기 선형변환 파라미터에 기초하여 상기 복수의 필터 각각의 중요도를 계산하는 중요도 계산부; 및

상기 필터 중요도에 기초하여 상기 복수의 필터 중 일부를 제거하여 상기 심층 신경망 모델을 갱신하는 모델 갱신부를 포함하는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 장치.

청구항 7

청구항 6에 있어서,

상기 선형 변환 파라미터는, 스케일(Scale) 값 및 바이어스(Bias) 값을 포함하는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 장치.

청구항 8

청구항 7에 있어서,

상기 필터 중요도는, 상기 가중치, 상기 스케일 값 및 상기 바이어스 값에 기초하여 하기 수학식에 의해 계산되는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 장치.

(수학식)

$$F_i = \gamma_i \|w_i\| + a \operatorname{sign}(\beta_i) \|\beta_i\|$$

F_i

: 상기 복수의 필터 중 i번째 필터의 필터 중요도

w_i :

상기 i번째 필터에 대한 가중치

γ_i

: 상기 i번째 필터에 대응하는 선형변환 파라미터의 스케일 값

β_i

: 상기 i번째 필터에 대응하는 선형변환 파라미터의 바이어스 값

$a \quad \beta_i$

: β_i 에 대한 스케일 조정 파라미터

청구항 9

청구항 6에 있어서,

상기 중요도 계산부는,

상기 심층 신경망 모델에 포함된 하나 이상의 잔여 연결(Residual Connection)에 기초하여, 상기 복수의 필터를 복수의 그룹으로 그룹화하고,

상기 복수의 필터 중 동일한 그룹에 포함된 하나 이상의 필터 각각의 중요도에 기초하여 상기 복수의 그룹 각각에 대한 그룹 중요도를 계산하고,

상기 모델 갱신부는, 상기 필터 중요도 및 상기 그룹 중요도에 기초하여 상기 심층 신경망 모델을 갱신하는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 장치.

청구항 10

청구항 9에 있어서,

상기 그룹 중요도는, 상기 그룹화된 필터의 평균에 기초하여 하기 수학식에 의해 산출되는, 심층신경망의 선형변환 파라미터를 이용한 가지치기 장치.

(수학식)

$$\overline{F}_i^g = \frac{1}{n(L_g)} \sum_{l \in L_g} F^l_i$$

$\overline{F}_{ij}^g$

: 상기 복수의 그룹 중 g번째 그룹에 대한 그룹 중요도

L_g :

상기 g번째 그룹에 포함된 레이어들의 인덱스 집합

$n(L_g)$: 상기 g 번째 그룹에 포함된 레이어의 개수

F_i^l : 상기 g 번째 그룹 내 l 번째 레이어의 F_i^l 에 포함된 필터의 중요도

발명의 설명

기술 분야

[0001] 개시되는 실시예들은 심층신경망의 선형변환 파라미터를 이용한 가지치기 기술과 관련된다.

배경 기술

[0002] 심층 신경망 모델의 경량화 방법으로서, 심층신경망을 구성하는 가중치 파라미터 중 일부를 제거하는 가지치기 기술은 원본 심층 신경망 대비 동등한 또는 작은 차이가 있는 예측 정확도 성능을 가지면서 가중치가 적은 경량화된 신경망을 도출할 수 있다.

[0003] 전형적으로 네트워크 가지치기 방법은 크게 가중치를 제거하여 네트워크를 가지치기하는 웨이트 가지치기 기법과 필터를 구성하고 있는 모든 가중치를 제거하여 네트워크를 가지치기하는 필터 가지치기 기법의 두가지로 분류할 수 있다. 특히, 웨이트 가지치기 기법은 특화된 소프트웨어와 하드웨어가 없는 경우 연산 속도가 느린 문제가 있었다.

[0004] 이에, 최근에는 필터 자체를 제거하여 행렬 연산 수를 감소시키는 필터 가지치기에 대한 연구가 활발히 진행되고 있으며, 그 중에서도 제거된 가중치에 다시 학습의 기회를 부여하는 동적 필터 가지치기 기술에 대한 연구가 이루어지고 있다.

발명의 내용

해결하려는 과제

[0005] 개시되는 실시예들은 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법 및 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0006] 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법은 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 컨볼루션 레이어(Convolutional Layer) 및 정규화 레이어(Normalization Layer)를 포함하는 심층 신경망 모델에서 상기 컨볼루션 레이어를 구성하는 복수의 필터 각각의 가중치 및 상기 정규화 레이어의 선형변환(Linear Transform) 파라미터를 획득하는 단계, 상기 가중치 및 상기 선형변환 파라미터에 기초하여 상기 복수의 필터 각각의 중요도를 계산하는 단계 및 상기 필터 중요도에 기초하여 상기 복수의 필터 중 일부를 제거하여 상기 심층 신경망 모델을 갱신하는 단계를 포함한다.

[0007] 상기 선형 변환 파라미터는, 스케일(Scale) 값 및 바이어스(Bias) 값을 포함할 수 있다.

[0008] 상기 필터 중요도는, 상기 가중치, 상기 스케일 값 및 상기 바이어스 값에 기초하여 하기 수학식에 의해 계산될 수 있다.

[0009] (수학식)

$$F_i = \gamma_i \|w_i\| + \alpha \text{sign}(\beta_i) \|\beta_i\|$$

[0010]

F_i

[0011] : 상기 복수의 필터 중 i 번째 필터의 필터 중요도

w_i :

[0012] 상기 i 번째 필터에 대한 가중치

γ_i

[0013] : 상기 i번째 필터에 대응하는 선형변환 파라미터의 스케일 값

β_i

[0014] : 상기 i번째 필터에 대응하는 선형변환 파라미터의 바이어스 값

$\alpha \quad \beta_i$

[0015] : α 에 대한 스케일 조정 파라미터

[0016] 상기 심층 신경망 모델에 포함된 하나 이상의 잔여 연결(Residual Connection)에 기초하여, 상기 복수의 필터를 복수의 그룹으로 그룹화하는 단계, 및 상기 복수의 필터 중 동일한 그룹에 포함된 하나 이상의 필터 각각의 중요도에 기초하여 상기 복수의 그룹 각각에 대한 그룹 중요도를 계산하는 단계를 더 포함하고, 상기 갱신하는 단계는, 상기 필터 중요도 및 상기 그룹 중요도에 기초하여 상기 심층 신경망 모델을 갱신할 수 있다.

[0017] 상기 그룹 중요도는, 상기 그룹화된 필터의 평균에 기초하여 하기 수학식에 의해 산출될 수 있다.

[0018] (수학식)

$$\overline{F}_i^g = \frac{1}{n(L_g)} \sum_{i \in L_g} F_i^l$$

[0019]

$\overline{F}_{ij}^g$

[0020] : 상기 복수의 그룹 중 g번째 그룹에 대한 그룹 중요도

[0021] L_g : 상기 g번째 그룹에 포함된 레이어들의 인덱스 집합

$n(L_g)$

[0022] : 상기 g번째 그룹에 포함된 레이어의 개수

F_i^l

[0023] : 상기 g번째 그룹 내 l번째 레이어의 α 에 포함된 필터의 중요도

[0024] 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기 장치는 컨볼루션 레이어(Convolutional Layer) 및 정규화 레이어(Normalization Layer)를 포함하는 심층 신경망 모델에서 상기 컨볼루션 레이어를 구성하는 복수의 필터 각각의 가중치 및 상기 정규화 레이어의 선형변환(Linear Transform) 파라미터를 획득하는 파라미터 획득부, 상기 가중치 및 상기 선형변환 파라미터에 기초하여 상기 복수의 필터 각각의 중요도를 계산하는 중요도 계산부 및 상기 필터 중요도에 기초하여 상기 복수의 필터 중 일부를 제거하여 상기 심층 신경망 모델을 갱신하는 모델 갱신부를 포함한다.

[0025] 상기 선형 변환 파라미터는, 스케일(Scale) 값 및 바이어스(Bias) 값을 포함할 수 있다.

[0026] 상기 필터 중요도는, 상기 가중치, 상기 스케일 값 및 상기 바이어스 값에 기초하여 하기 수학식에 의해 계산될 수 있다.

[0027] (수학식)

$$F_i = \|\gamma_i\| \|w_i\| + \alpha \operatorname{sign}(\beta_i) \|\beta_i\|$$

[0028]

F_i

[0029] : 상기 복수의 필터 중 i번째 필터의 필터 중요도

w_i :

[0030] 상기 i번째 필터에 대한 가중치

γ_i

[0031] : 상기 i번째 필터에 대응하는 선형변환 파라미터의 스케일 값

β_i

[0032] : 상기 i 번째 필터에 대응하는 선형변환 파라미터의 바이어스 값

$a \quad \beta_i$

[0033] : a 에 대한 스케일 조정 파라미터

[0034] 상기 중요도 계산부는, 상기 심층 신경망 모델에 포함된 하나 이상의 잔여 연결(Residual Connection)에 기초하여, 상기 복수의 필터를 복수의 그룹으로 그룹화하고, 상기 복수의 필터 중 동일한 그룹에 포함된 하나 이상의 필터 각각의 중요도에 기초하여 상기 복수의 그룹 각각에 대한 그룹 중요도를 계산하고, 상기 모델 갱신부는, 상기 필터 중요도 및 상기 그룹 중요도에 기초하여 상기 심층 신경망 모델을 갱신할 수 있다.

[0035] 상기 그룹 중요도는, 상기 그룹화된 필터의 평균에 기초하여 하기 수학식에 의해 산출될 수 있다.

[0036] (수학식)

$$\overline{F}_i^g = \frac{1}{n(L_g)} \sum_{i \in L_g} F_i^l$$

[0037]

$\overline{F}_{ij}^g$

[0038] : 상기 복수의 그룹 중 g 번째 그룹에 대한 그룹 중요도

[0039] L_g : 상기 g 번째 그룹에 포함된 레이어들의 인덱스 집합

$n(L_g)$

[0040] : 상기 g 번째 그룹에 포함된 레이어의 개수

F_{ij}^l

[0041] : 상기 g 번째 그룹 내 l 번째 레이어의 i 에 포함된 필터의 중요도

발명의 효과

[0042] 개시되는 실시예들에 따르면, 심층신경망의 선형변환 파라미터를 이용하여 필터의 중요도에 기초하여 가지치기를 수행함으로써, 최적의 정확도를 가진 모델을 생성할 수 있다.

[0043] 또한, 개시되는 실시예들에 따르면, 잔여 연결에 속한 필터를 그룹화하고, 그룹화된 필터에 대한 그룹 중요도에 기초하여 가지치기를 수행함으로써, 잔여 연결로 인한 인덱스 불일치 문제를 해결할 수 있다.

도면의 간단한 설명

[0044] 도 1은 일 실시예에 따른 심층신경망의 선형변환 파라미터 이용한 가지치기 장치를 설명하기 위한 블록도이다.

도 2는 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기 대상인 심층 신경망(Deep Neural Network) 모델을 개략적으로 나타낸 도면이다.

도 3은 일 실시예에 따른 필터 중요도에 기초하여 심층 신경망 모델을 가지치기하는 방법을 나타낸 도면이다.

도 4는 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법을 나타낸 흐름도이다.

도 5는 일 실시예에 따른 그룹 중요도 계산 방법을 나타낸 흐름도이다.

도 6은 일 실시예에 따른 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0045] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0046] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를

불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 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. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.

- [0047] 도 1은 일 실시예에 따른 심층신경망의 선형변환 파라미터 이용한 가지치기 장치를 설명하기 위한 블록도이다.
- [0048] 도시된 바와 같이, 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기 장치(100)(이하, 가지치기 장치)는 파라미터 획득부(110), 중요도 계산부(120) 및 모델 갱신부(140)를 포함한다.
- [0049] 파라미터 획득부(110)는 컨볼루션 레이어(Convolutional Layer) 및 정규화 레이어(Normalization Layer)를 포함하는 심층 신경망 모델에서 컨볼루션 레이어를 구성하는 복수의 필터 각각의 가중치 및 정규화 레이어의 선형변환(Linear Transform) 파라미터를 획득한다.
- [0050] 일 실시예에 따르면, 선형 변환 파라미터는, 스케일(Scale) 값 및 바이어스(Bias) 값을 포함할 수 있다.
- [0051] 도 2는 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기의 대상인 심층 신경망(Deep Neural Network) 모델을 개략적으로 나타낸 도면이다.
- [0052] 도 2를 참조하면, 심층 신경망 모델(200)은 컨볼루션 레이어(Convolutional Layer)(210) 및 정규화 레이어(Normalization Layer)(220)를 포함할 수 있다. 이때, 입력 데이터(x)는 2차원 행렬(1개의 채널)로 가정한다.
- [0053] 컨볼루션 레이어(210)는 입력 데이터(x)를 입력 받고, 입력 데이터(x)로부터 특징 맵(feature map)을 생성하기 위한 하나 이상의 필터로 이루어질 수 있다. 구체적으로, 컨볼루션 레이어는 w_1, w_2, w_3, w_4 의 가중치 행렬을 가질 수 있다. 일 실시예에 따라, 컨볼루션 레이어(210)는 입력 데이터(x)에 하나 이상의 필터를 합성곱하여 각각의 필터에 대응하는 y_1, y_2, y_3, y_4 을 출력할 수 있다. 여기서, y_i 는 1개의 채널을 가진 행렬인 것으로 가정한다.
- [0054] 정규화 레이어(220)는 컨볼루션 레이어(210)의 출력값 y_1, y_2, y_3, y_4 을 정규화할 수 있다. 구체적으로, 컨볼루션 레이어(210)의 필터 각각을 통과한 출력값 y_1, y_2, y_3, y_4 에 평균(mean)과 표준편차(std)를 적용하는 배치 정규화(Batch Normalization) 연산을 통해 정규화 할 수 있다.
- [0055] 중요도 계산부(120)는 가중치 및 선형변환 파라미터에 기초하여 복수의 필터 각각의 중요도를 계산한다.
- [0056] 일 실시예에 따르면, 필터 중요도는, 가중치, 스케일 값 및 바이어스 값에 기초하여 계산될 수 있다.
- [0057] 일 실시예에 따르면, 필터 중요도는 컨볼루션 레이어 및 배치 정규화 레이어를 포함하는 심층 신경망 모델의 출력값은 하기의 수학적 식 1로 표현될 수 있다.

수학적 식 1

$$\overline{x_{out}} = \gamma_i \circ \frac{w * x_{in} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta_i$$

[0059]

$\overline{x_{out}}$

[0061] : 컨볼루션 레이어 및 배치 정규화 레이어를 포함하는 심층 신경망 모델의 출력값

$\mathbf{x}_{in}$

[0062] : 입력값

w_i :

[0063] i번째 필터에 대한 가중치

γ_i

[0064] : i번째 필터에 대응하는 선형변환 파라미터의 스케일 값

β_i

[0065] : i번째 필터에 대응하는 선형변환 파라미터의 바이어스 값

[0066] 한편, 필터 중요도는 가지치기에 의해 발생될 왜곡도에 기초하여 필터 중요도를 계산할 수 있으며, 가지치기 필터의 왜곡도는 다음의 수학적 식 2에 의해 표현될 수 있다.

수학적 식 2

$$D_i = \|\overline{\mathbf{x}_{out}} - m^o \mathbf{x}_{out}\| = \left\| \gamma_i^o \frac{\mathbf{w} * \mathbf{x}_{in} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta_i \right\|$$

[0067]

D_i

[0069] : i번째 필터의 가지치기에 의해 발생될 왜곡도

m

[0070] : i번째 원소는 0으로, i번째 원소를 제외한 모든 원소는 1로 구성된 벡터 ($m \in R^{c_{out}}$, 여기서 c_{out} 컨볼루션 레이어를 통과한 출력값의 차원)

[0071] 일 실시예에 따르면, 수학적 식 2는 코시-슈바르츠 부등식(Cauchy-Schwarz inequality) 및 영의 컨볼루션 부등식(Young's convolution inequality)에 기초하여 다음의 수학적 식 3으로 변환할 수 있다.

수학적 식 3

$$D_i \leq \|\gamma_i\| \|\mathbf{w}_i\| \|\mathbf{x}_{in}\| + \|\beta_i\|$$

[0072]

[0074] 필터 중요도는 수학적 식 3에 기초하여 바이아스의 부호(sign) 및 바이아스에 대한 스케일 조정 파라미터에 기초하여 다음의 수학적 식 4를 통해 계산될 수 있다.

수학적 식 4

$$F_i = \|\gamma_i\| \|\mathbf{w}_i\| + \alpha \text{sign}(\beta_i) \|\beta_i\|$$

[0075]

F_i

[0077] : 복수의 필터 중 i번째 필터의 필터 중요도

α β_i

[0078] : β_i 에 대한 스케일 조정 파라미터

$$\alpha \qquad \beta_i \qquad \beta_i$$

[0079] 여기서 α 는 배치 정규화 바이아스(β_i)의 신뢰도를 반영하는 스케일을 의미하며, 배치 정규화 바이아스(β_i)의 부호가 음수일 경우 필터 중요도 값이 감소할 수 있다.

[0080] 일 실시예에 따르면, 파라미터의 수에 증가에 따르는 검색 비용, 계산 증가 문제로 인한 불안정성을 고려하여 $\alpha = 0$ 을 대입하여 필터 중요도를 계산하기 위한 수학식 4를 다음의 수학식 5와 같이 단순화할 수 있다.

수학식 5

$$F_i = \|\gamma_i\| \|w_i\|$$

[0081]

[0082] 모델 갱신부(130)는 필터 중요도에 기초하여 복수의 필터 중 일부를 제거하여 심층 신경망 모델을 갱신한다.

[0083] 일 실시예에 따르면, 모델 갱신부(130)는 복수의 필터 중 필터 중요도가 사전 설정된 임계값 이하인 필터를 제거하여 심층 신경망 모델을 갱신할 수 있다.

[0084] 도 3은 일 실시예에 따른 필터 중요도 기초하여 심층 신경망 모델을 가지치기하는 방법을 나타낸 도면이다.

[0085] 도 3 (a)은 파라미터의 수에 증가에 따르는 검색 비용, 계산 증가 문제로 인한 불안정성을 고려하여 $\alpha = 0$ 을 대입한 필터 중요도 계산식($F_i = \|\gamma_i\| \|w_i\|$)에 기초하여 각각의 필터의 중요도($f(w_i, \gamma_i)$)를 나타낸 것이다.

[0086] 한편, 일 실시예에 따르면, 모델 갱신부(130)는 복수의 필터 중 필터 중요도가 사전 설정된 임계값 이하의 필터를 제거하여 심층 신경망 모델을 갱신할 수 있다.

[0087] 도 3 (b)는 도 (a)의 1 내지 4 필터에 대한 필터 중요도에 기초하여 일부 필터를 가지치기하는 과정을 나타낸 도면이다.

[0088] 이때, 도 2와 같이, 심층 신경망 모델은 컨볼루션 레이어(Convolutional Layer)(210) 및 정규화 레이어(Normalization Layer)(220)를 포함할 수 있다. 이때, 입력 데이터(x)는 2차원 행렬(1개의 채널)로 가정한다.

[0089] 컨볼루션 레이어(210)는 입력 데이터(x)를 입력 받고, 입력 데이터(x)로부터 특징 맵(feature map)을 생성하기 위한 하나 이상의 필터로 이루어질 수 있다. 구체적으로, 컨볼루션 레이어는 w_1, w_2, w_3, w_4 의 가중치 행렬을 가질 수 있다. 일 실시예에 따라, 컨볼루션 레이어(210)는 입력 데이터(x)에 하나 이상의 필터를 합성곱하여 각각의 필터에 대응하는 y_1, y_2, y_3, y_4 을 출력할 수 있다. 여기서, y_i 는 1개의 채널을 가진 행렬인 것으로 가정한다.

$$w_1, \gamma_i$$

[0090] 일 실시예에 따라, 도 (b)는 도 3 (a) 중 1 내지 4 필터 중 필터 중요도 $f(w_1, \gamma_i) < 0.4$ 인 1, 4 필터를 가지치기한 것을 나타낸 것이다.

[0091] 일 실시예에 따르면, 중요도 계산부(120)는 심층 신경망 모델에 포함된 하나 이상의 잔여 연결(Residual Connection)에 기초하여, 복수의 필터를 복수의 그룹으로 그룹화할 수 있다. 이후, 중요도 계산부(120)는 복수의 필터 중 동일한 그룹에 포함된 하나 이상의 필터 각각의 중요도에 기초하여 복수의 그룹 각각에 대한 그룹 중요도를 계산할 수 있다.

[0092] 한편, 일 실시예에 따르면, 모델 갱신부(130)는 필터 중요도 및 그룹 중요도에 기초하여 심층 신경망 모델을 갱신할 수 있다.

[0093] 일 실시예에 따르면, 중요도 계산부(120)는 그룹화된 필터의 평균에 기초하여 하기 수학식에 의해 그룹 중요도를 산출할 수 있다.

수학식 6

$$\overline{F}_i^g = \frac{1}{n(L_g)} \sum_{i \in L_g} F_i^l$$

[0094]

$$\overline{F}_{ij}^g$$

[0095]

: 상기 복수의 그룹 중 g번째 그룹에 대한 그룹 중요도

[0096]

L_g : 상기 g번째 그룹에 포함된 레이어들의 인덱스 집합

[0097]

$n(L_g)$: 상기 g번째 그룹에 포함된 레이어의 개수

[0098]

F_{ij}^l : 상기 g번째 그룹 내 l 번째 레이어의 e에 포함된 필터의 중요도

[0099]

중요도 계산부(120)는 하나 이상의 잔여 연결에 기초하여, 복수의 필터를 복수의 그룹으로 그룹화하고, 복수의 필터 중 동일한 그룹에 포함된 하나 이상의 필터 각각의 중요도에 기초하여 계산된 그룹 중요도에 기초하여 심층 신경망 모델을 갱신함으로써, 기존의 동일한 잔여 연결에 속하는 필터의 모든 중요도를 추가하는 과정에서 발생하는 병목 현상 및 모든 중요도 계산에 따른 시간을 절약할 수 있다.

[0100]

도 4는 일 실시예에 따른 심층신경망의 선형변환 파라미터를 이용한 가지치기 방법을 나타낸 흐름도이다. 도 4에 도시된 방법은 예를 들어, 도 1에 도시된 가지치기 장치(100)에 의해 수행될 수 있다.

[0101]

도 4를 참조하면, 가지치기 장치(100)는 컨볼루션 레이어(Convolutional Layer) 및 정규화 레이어(Normalization Layer)를 포함하는 심층 신경망 모델에서 컨볼루션 레이어를 구성하는 복수의 필터 각각의 가중치 및 정규화 레이어의 선형변환(Linear Transform) 파라미터를 획득한다(410).

[0102]

이후, 가지치기 장치(100)는 가중치 및 선형변환 파라미터에 기초하여 복수의 필터 각각의 중요도를 계산한다(420).

[0103]

이후, 가지치기 장치(100)는 필터 중요도에 기초하여 복수의 필터 중 일부를 제거하여 심층 신경망 모델을 갱신한다(430).

[0104]

한편, 도 4에 도시된 순서도에서 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 도시되지 않은 하나 이상의 단계가 추가되어 수행될 수 있다.

[0105]

도 5는 일 실시예에 따른 그룹 중요도 계산 방법을 나타낸 흐름도이다.

[0106]

도 5에 도시된 방법은 예를 들어, 도 1에 도시된 중요도 계산부(120)에 의해 수행될 수 있다.

[0107]

도 5를 참조하면, 중요도 계산부(120)는 심층 신경망 모델에 포함된 하나 이상의 잔여 연결(Residual Connection)에 기초하여, 복수의 필터를 복수의 그룹으로 그룹화한다(510).

[0108]

이후, 중요도 계산부(120)는 복수의 필터 중 동일한 그룹에 포함된 하나 이상의 필터 각각의 중요도에 기초하여 복수의 그룹 각각에 대한 그룹 중요도를 계산한다(520).

[0109]

한편, 도 5에 도시된 순서도에서 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 도시되지 않은 하나 이상의 단계가 추가되어 수행될 수 있다.

[0110]

도 6은 일 실시예에 따른 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다.

[0111]

도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0112]

도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 컴퓨팅 장치(12)는 일 실시예에 따른 문서 구조 분석

학습 장치(100, 600, 700)에 포함된 하나 이상의 컴포넌트일 수 있다.

- [0113] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.
- [0114] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 장치(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.
- [0115] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.
- [0116] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.
- [0117] 이상에서 대표적인 실시예를 통하여 본 발명에 대하여 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 전술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

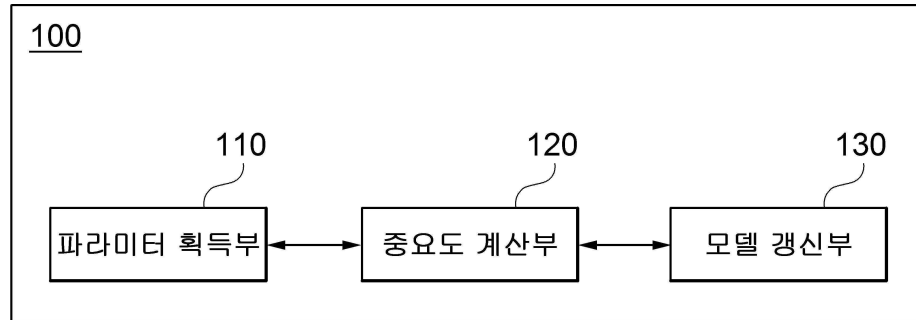
부호의 설명

- [0119] 10: 컴퓨팅 환경
12: 컴퓨팅 장치
14: 프로세서
16: 컴퓨터 판독 가능 저장 매체
18: 통신 버스
20: 프로그램
22: 입출력 인터페이스
24: 입출력 장치
26: 네트워크 통신 인터페이스
100: 가지치기 장치
110: 파라미터 획득부

- 120: 중요도 계산부
- 130: 모델 갱신부
- 200: 심층신경망 모델
- 210: 컨볼루션 레이어
- 220: 정규화 레이어

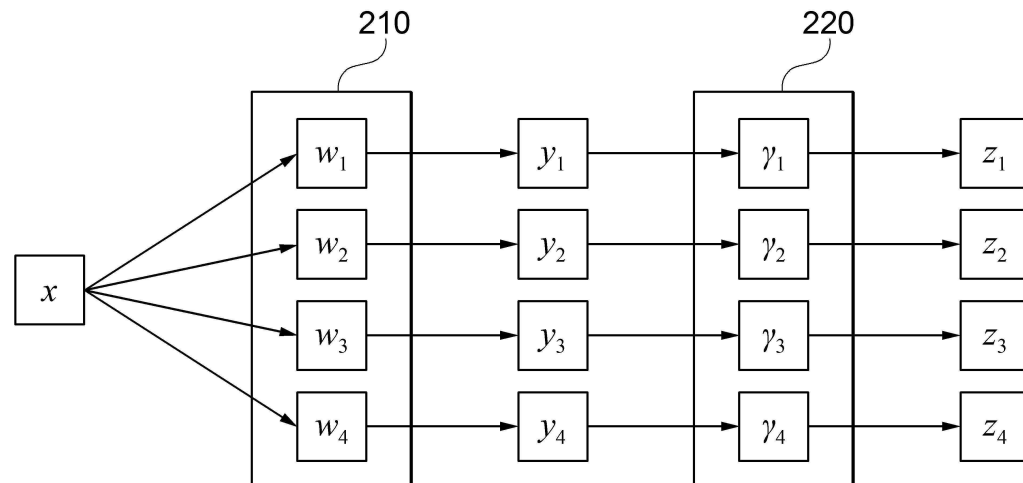
도면

도면1



도면2

200



도면3

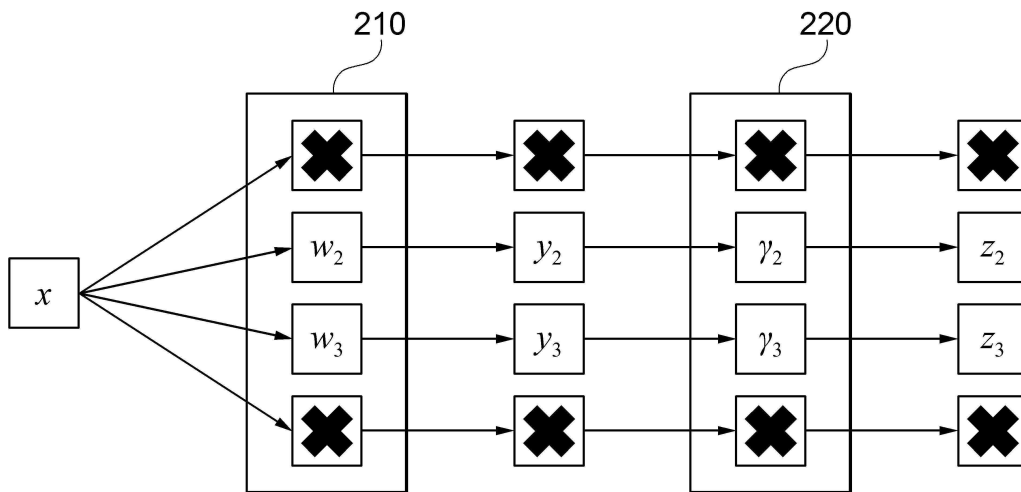
$$0.1 = f(w_1, \gamma_1)$$

$$0.5 = f(w_2, \gamma_2)$$

$$0.7 = f(w_3, \gamma_3)$$

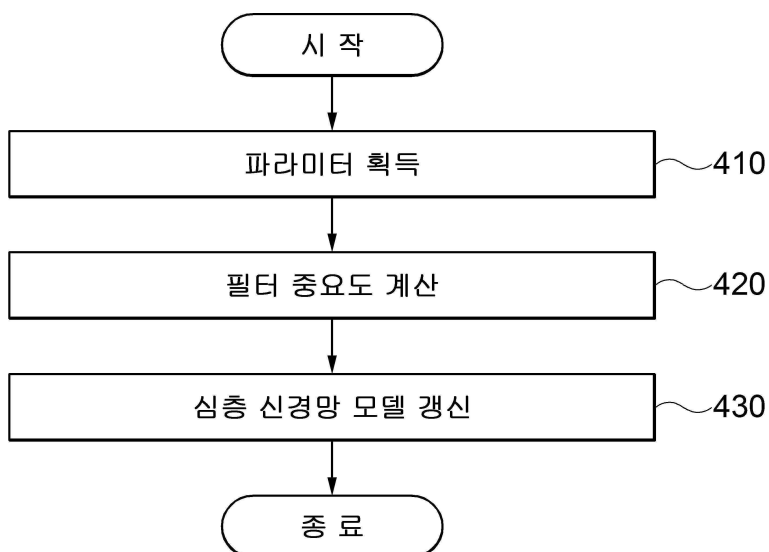
$$0.3 = f(w_4, \gamma_4)$$

(b)

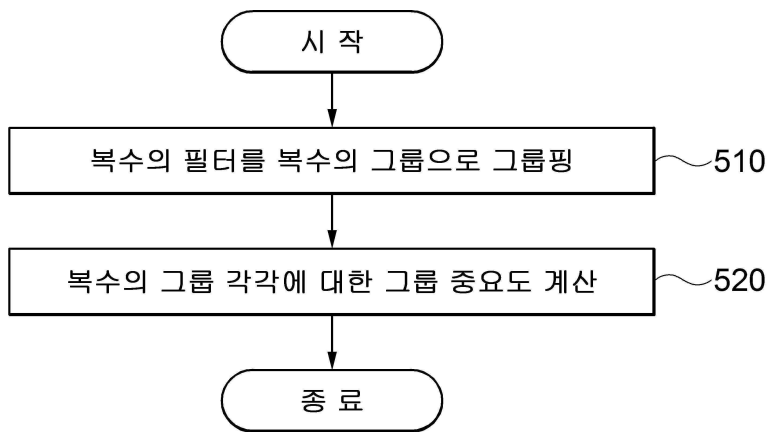


(b)

도면4



도면5



도면6

10

