(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0037700
(43) 공개일자 2020년04월09일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2018-0117221
(22) 출원일자 2018년10월01일
심사청구일자 2018년10월01일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
강상길
서울특별시 용산구 새창로 114 (용문동) 삼성래미안 114-701
허청환
대전광역시 서구 대덕대로 150 (갈마동, 큰마을아파트) 116동 803호
(74) 대리인
이원희

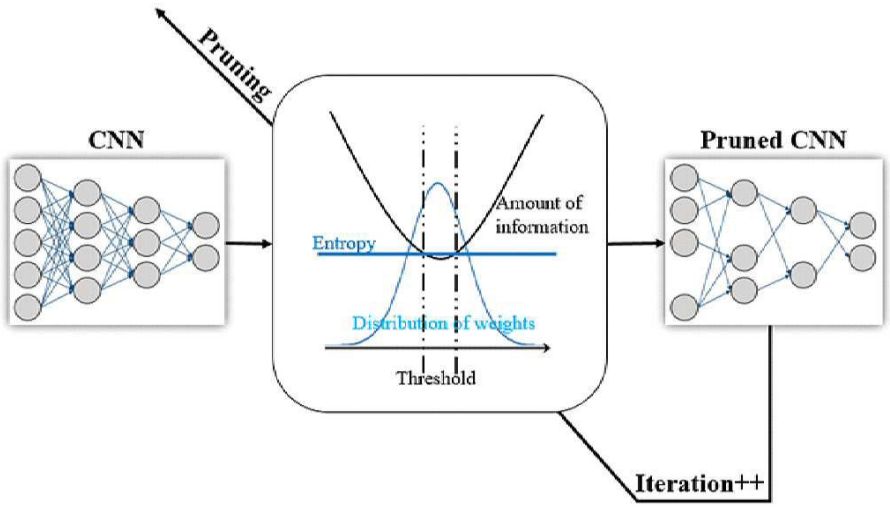
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **엔트로피 기반 신경망(Neural Networks) 가지치기 방법 및 시스템**

(57) 요 약

본 발명은 콘볼루션 신경망(Convolutional Neural Networks)을 이용한 학습에 있어서, 학습의 부하를 줄이면서도 정확도를 유지할 수 있는 방법 및 시스템에 관한 것으로, 더욱 상세하게는 엔트로피에 기반한 가중치 평가에 의한 콘볼루션 신경망의 가지치기 방법 및 시스템에 관한 것으로, NN(Neural Networks)을 이용한 학습 방법에 있어서, CNN의 복수의 가중치들의 엔트로피에 기반한 정보의 양을 연산하는 (a) 단계; 상기 복수의 가중치들 중 가지치기할 가중치를 결정하기 위한 임계값을 연산하는 (b) 단계; 및 상기 정보의 양이 임계값 이하의 값을 가지는 가중치들을 가지치기하는 (c) 단계;를 포함하는 구성을 개시한다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업
과제고유번호 57489-1
부처명 과학기술정보통신부
연구관리전문기관 정보통신기술진흥센터(IITP)
연구사업명 방송통신산업기술개발사업
연구과제명 [IITP-Ezbaro] 스마트기기를 위한 온디바이스 지능형 정보처리 가속화 SW플랫폼 기술개발(2차년도)
기 여 율 1/1
주관기관 기타공공기관
연구기간 2018.01.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

신경망(Neural Networks)을 이용한 학습 방법에 있어서,

복수의 가중치들의 엔트로피에 기반한 정보의 양을 연산하는 (a) 단계;

상기 복수의 가중치들 중 가지치기할 가중치를 결정하기 위한 임계값을 연산하는 (b) 단계; 및

상기 정보의 양이 임계값 이하의 값을 가지는 가중치들을 가지치기하는 (c) 단계;를 포함하는 엔트로피 기반 신경망(Neural Networks) 가지치기 방법.

청구항 2

제1항에 있어서,

상기 (a) 단계 내지 (c) 단계는,

상기 신경망이 학습을 수행하는 중에 수행되는 것을 특징으로 하는 엔트로피 기반 신경망(Neural Networks) 가지치기 방법.

청구항 3

제1항에 있어서,

상기 임계값은 하기 수학식 1을 연산해 얻는 엔트로피 기반 신경망(Neural Networks) 가지치기 방법.

[수학식 1]

$$ln\left(\sigma_{\omega}\sqrt{2\pi e}\right) \times \alpha$$

(여기서, σ_{ω}는 가중치의 표준 편차, α는 조정 계수이다.)

청구항 4

제3항에 있어서,

상기 정보의 양은 하기 수학식 2를 연산해 얻는 엔트로피 기반 신경망(Neural Networks) 가지치기 방법.

[수학식 2]

$$A[\omega_n] = \left(-\int_{k=n}^{n+\varepsilon} \log\left(\frac{1}{\sigma_{\omega}\sqrt{2\pi}} e^{\left(-\frac{(\omega_k-\mu_{\omega})^2}{2\sigma_{\omega}^2} \right)} \right) d\omega_k \right)$$

(여기서, σ_{ω}는 가중치의 표준 편차, ω_n은 n번째 가중치 μ_{ω}는 가중치의 평균이다.)

청구항 5

신경망(Neural Networks)을 이용한 학습 시스템에 있어서,

신경망(Neural Networks)을 이용한 학습을 수행하는 학습부;

상기 신경망의 복수의 가중치들의 엔트로피에 기반한 정보의 양 및 상기 복수의 가중치들 중 가지치기할 가중치를 결정하기 위한 임계값을 연산하는 연산부; 및

상기 정보의 양이 임계값 이하의 값을 가지는 제거 가중치들을 가지치기하는 가지치기부;를 포함하는 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템.

청구항 6

제5항에 있어서,

상기 연산부 및 가지치기부는,

상기 학습부가 학습을 수행하는 중에 연산 및 가지치기를 수행하는 것을 특징으로 하는 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템.

청구항 7

제5항에 있어서,

상기 임계값은 하기 수학식 1을 연산해 얻는 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템.

[수학식 1]

$$ln\left(\sigma_{\omega}\sqrt{2\pi e}\right) \times \alpha$$

(여기서, σ_{ω}는 가중치의 표준 편차, α는 조정 계수이다.)

청구항 8

제7항에 있어서,

상기 정보의 양은 하기 수학식 2를 연산해 얻는 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템.

[수학식 2]

$$A[\omega_n] = \left(-\int_{k=n}^{n+\varepsilon} \log\left(\frac{1}{\sigma_{\omega}\sqrt{2\pi}} e^{\left(-\frac{(\omega_k-\mu_{\omega})^2}{2\sigma_{\omega}^2}\right)} \right) d\omega_k \right)$$

(여기서, σ_{ω}는 가중치의 표준 편차, ω_n은 n번째 가중치 μ_{ω}는 가중치의 평균이다.)

발명의 설명

기 술 분 야

[0001] 본 발명은 콘볼루션 신경망(Convolutional Neural Networks)을 이용한 학습에 있어서, 학습의 부하를 줄이면서도 정확도를 유지할 수 있는 방법 및 시스템에 관한 것으로, 더욱 상세하게는 엔트로피에 기반한 가중치 평가에 의한 콘볼루션 신경망의 가지치기 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 깊은 신경망(DNN)이 다양한 분야에 성공적으로 적용되었다. 특히 컨벌루션뉴럴네트워크(CNN)은 폐쇄회로(CCTV)를 통해 촬영된 비디오에서 위험한 상황을 탐지하고, 콘서트를 통해 음성을 인식하며, 해커의 정보를 이용한 컴퓨터 시각 및 음향 모델링에 매우 능하다. 이러한 복잡한 작업을 실행하기 위해, DNN(CNN과 마찬가지로)은 대개 더 무거운 레이어와 엄청난 양의 학습 매개변수(또는 가중치)로 설계된다. 이 DNN은 일반적으로 큰 클라우드 서버에 있어야 작동시킬 수 있다. 그러나 원활한 운영을 위해서는 1) 리소스가 제한된 클라우드 서버와 휴대용 장치(스마트폰, 가전 제품, 군사 무기, 랩톱 등) 간의 원활한 데이터 전송을 유지하기 위한 안정적인 통신 환경과 2) 정보 전송으로 인해 취약해진 개인정보 보안이 필요하다. 컴퓨팅 및 메모리 리소스가 제한된 휴대용 장치에 DNN을 직접 포트하기 위해 일부 연구자는 DNN 구조의 압축과 높은 컴퓨팅 및 메모리 학습 복잡성을 완화하기 위해 DNN의 가중치를 제거한다.

[0003] 신경망(NN)의 연산 복잡성을 줄이기 위해 다양한 압축 접근법이 제안되었다. 대부분의 방법은 가중치 매트릭스를 분해하고 완전히 연결된 뉴럴네트워크(NN)에 기초한 주요 구성 요소 분석을 통해 중요한 가중치를 추출한 다음 추출된 가중치를 사용하여 압축 구조로 개조한다. 또한 많은 수의 가중치를 정량화하거나 클러스터링하여 NN을 소수의 가중치와 공유하는 방식으로 압축할 수 있다. 기존 기술은 입력 기능 간의 상관 관계 분석을 통해 부분적으로 연결된 NN을 모델링할 수 있는 압축 기술이 있었다. 부분적으로 연결된 NN은 상관된 입력을 위해 숨겨진 레이어에 연결을 연결하고 관련되지 않은 입력을 위해 연결을 분리하여 구조화된다. 이 방법은 정확도를 향상시켰지만 위의 다른 압축 방법처럼 압축 성능이 만족스럽지 않았다.

[0004] 또 다른 방법(정리 접근 방식)은 second derivative을 사용한 손실 균형, 테일러 확장을 사용한 손실 근사치와 같은 다양한 학문적 방법을 사용하여 NN 출력에 대한 가중치의 중요성을 판단함으로써 비중요 가중치를 제거한다. NN은 학습 및 백프로파게이션 훈련 기법을 적용할 때, 이러한 방법에서 도출된 지표는 높은 계산적 복잡성을 필요로 한다. 절대값 비교를 사용한 반복 소거 방법은 위의 방법에 비해 메트릭이 더 단순하다. 하지만, 그것은 정확성을 떨어트린다. 상기 방법은 훈련을 종료한 후 가중치 값의 크기에 따라 가중치의 중요성을 결정한다. 가중치의 값이 설정된 임계값보다 작으면 작은 값이 출력에 거의 영향을 미치지 않는다는 가정이 있기 때문에 가중치는 제거된다. 기존 기술은 훈련 중 가중치가 출력에 미치는 영향을 고려하지 않기 때문에 원래 NN의 정확도가 저하되지 않지만 네트워크 압축 성능에 제한이 있다. 또한 교육을 마친 후 가중치 값에 대한 통계적 분석을 바탕으로 임계값을 정하고, 상대적으로 중요하지 않은 가중치를 잘라내고 정확도가 저하되기 전까지 이 프로세스를 반복하는 것은 시간이 많이 걸리는 작업이다.

발명의 내용

해결하려는 과제

[0006] 따라서, 본 발명은 상기한 바와 같은 문제점을 해결하기 위한 것으로서, 훈련하는 동안 가중치가 출력에 기여하는 정보의 평균 양을 고려하여 임계치를 결정한다. NN은 블랙박스 방식으로 훈련되므로 훈련 중 가중치의 기여도에 대한 정보를 엄격하게 식별하기는 어렵다. 따라서 본 발명에서 정보는 정보 이론에 일반적으로 사용되는 엔트로피 기법을 사용함으로써 확률적으로 측정된다.

[0007] 본 발명은 엔트로피 기반 정보를 일반적으로 사용되는 가우스 분포로 가정한 가중치의 통계 분포에 매핑한다. 매핑에서 임계값의 메트릭을 도출하여 훈련 중 가중치의 중요성을 정량화한다. 임계값은 교육이 종료될 때까지 모든 반복 중에 반복적으로 변경될 수 있다.

과제의 해결 수단

[0009] 상기한 문제를 해결하기 위한 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법은 신경망(Neural Networks)을 이용한 학습 방법에 있어서, CNN의 복수의 가중치들의 엔트로피에 기반한 정보의 양을 연산하는 (a) 단계; 상기 복수의 가중치들 중 가지치기할 가중치를 결정하기 위한 임계값을 연산하는 (b) 단계; 및 상기 정보의 양이 임계값 이하의 값을 가지는 가중치들을 가지치기하는 (c) 단계;를 포함할 수 있다.

[0010] 본 발명의 일 실시 예에 따르면, 상기 (a) 단계 내지 (c) 단계는 상기 신경망이 학습을 수행하는 중에 수행되는 것일 수 있다.

[0011] 본 발명의 일 실시 예에 따르면, 상기 임계값은 하기 수학식 1을 연산해 얻을 수 있다.

[0012] [수학식 1]

[0013] $$ln(\sigma_\omega\sqrt{2\pi e}) \times \alpha$$

[0014] (여기서, σ_ω는 가중치의 표준 편차, α는 조정 계수이다.)

[0015] 본 발명의 일 실시 예에 따르면, 상기 정보의 양은 하기 수학식 2를 연산해 얻을 수 있다.

[0016] [수학식 2]

[0017] $$A[\omega_n] = \left(-\int_{k=n}^{n+\varepsilon} \log\left(\frac{1}{\sigma_\omega\sqrt{2\pi}} e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega^2} \right)} \right) d\omega_k \right)$$

[0018] (여기서, σ_ω는 가중치의 표준 편차, ω_n은 n번째 가중치 μ_ω는 가중치의 평균이다.)

[0019] 상기한 문제를 해결하기 위한 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템은 신경망(Neural Networks)을 이용한 학습 시스템에 있어서, 신경망(Neural Networks)을 이용한 학습을 수행하는 학습부; CNN의 복수의 가중치들의 엔트로피에 기반한 정보의 양 및 상기 복수의 가중치들 중 가지치기 할 가중치를 결정하기 위한 임계값을 연산하는 연산부; 및 상기 정보의 양이 임계값 이하의 값을 가지는 제거 가중치들을 가지치기하는 가지치기부;를 포함할 수 있다.

[0020] 본 발명의 일 실시 예에 따르면, 상기 연산부 및 가지치기부는, 상기 학습부가 학습을 수행하는 중에 연산 및 가지치기를 수행할 수 있다.

[0021] 본 발명의 일 실시 예에 따르면, 상기 임계값은 하기 수학식 1을 연산해 얻을 수 있다.

[0022] [수학식 1]

[0023] $$ln(\sigma_\omega\sqrt{2\pi e}) \times \alpha$$

[0024] (여기서, σ_ω는 가중치의 표준 편차, α는 조정 계수이다.)

[0025] 본 발명의 일 실시 예에 따르면, 상기 정보의 양은 하기 수학식 2를 연산해 얻을 수 있다.

[0026] [수학식 2]

[0027] $$A[\omega_n] = \left(-\int_{k=n}^{n+\varepsilon} \log\left(\frac{1}{\sigma_\omega\sqrt{2\pi}} e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega^2} \right)} \right) d\omega_k \right)$$

[0028] (여기서, σ_ω는 가중치의 표준 편차, ω_n은 n번째 가중치 μ_ω는 가중치의 평균이다.)

발명의 효과

[0030] 본 발명에 따르면, 기존의 방법보다 압축보다 성능이 뛰어나며, 기존 압축 방법인 MNIST(Mixed National Institute of Standards and Technology) 데이터를 사용하여 처리 시간의 오버헤드를 줄여준다.

[0031] 또한, 학습 중에 가중치에 따른 가지치기를 수행하여 연산 중간에도 시스템의 부하를 감소시킬 수 있다.

[0032] 한편, 본 발명의 효과는 이상에서 언급한 효과들로 제한되지 않으며, 이하에서 설명할 내용으로부터 통상의 기술자에게 자명한 범위 내에서 다양한 효과들이 포함될 수 있다.

도면의 간단한 설명

[0033] 도 1은 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법의 개념도이다.

도 2는 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템의 블록도이다.

도 3은 신경망(Neural Networks)의 학습 곡선이다.

도 4는 본 발명의 일 실시 예에 따른 문턱, 엔트로피 및 반복 8000에서 각 레이어의 작은 값의 가중치 정보를 도시한 그래프이다.

도 5는 본 발명의 일 실시 예에 따른 각 레이어에 대한 α_k의 값에 따른 훈련 정확도이다.

도 6는 본 발명의 일 실시 예에 따른 훈련 중 $\alpha_1 = 0.5$, $\alpha_2 = 0.7$, $\alpha_3 = 1.1$, $\alpha_4 = 0.6$ 인 각 레이어의 압축률이다.

도 7은 테스트 데이터의 정확도와 8000에서 시작하는 1000 회 반복마다의 전체 압축률을 도시한 것이다.

도 8은 본 발명의 일 실시 예에 따른 방법과 기존의 방법의 압축 성능을 비교한 것이다.

도 9은 본 발명의 일 실시 예에 방법과 기존의 방법의 가지치기 속도를 비교한 것이다.

도 10은 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0034] 이하, 첨부된 도면들을 참조하여 본 발명에 따른 '엔트로피 기반 신경망(Neural Networks) 가지치기 방법 및 시스템'을 상세하게 설명한다. 설명하는 실시 예들은 본 발명의 기술 사상을 당업자가 용이하게 이해할 수 있도록 제공되는 것으로 이에 의해 본 발명이 한정되지 않는다. 또한, 첨부된 도면에 표현된 사항들은 본 발명의 실시 예들을 쉽게 설명하기 위해 도식화된 도면으로 실제로 구현되는 형태와 상이할 수 있다.

[0035] 한편, 이하에서 표현되는 각 구성부는 본 발명을 구현하기 위한 예일 뿐이다. 따라서, 본 발명의 다른 구현에서는 본 발명의 사상 및 범위를 벗어나지 않는 범위에서 다른 구성부가 사용될 수 있다.

[0036] 또한, 각 구성부는 순전히 하드웨어 또는 소프트웨어의 구성만으로 구현될 수도 있지만, 동일 기능을 수행하는 다양한 하드웨어 및 소프트웨어 구성들의 조합으로 구현될 수도 있다. 또한, 하나의 하드웨어 또는 소프트웨어에 의해 둘 이상의 구성부들이 함께 구현될 수도 있다.

[0037] 또한, 어떤 구성요소들을 '포함'한다는 표현은, '개방형'의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.

[0038] 도 1은 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법의 개념도이다.

[0039] 도 1을 참조하면, 엔트로피 기반 가중치 제거 방법의 전체 과정을 개략적으로 도시 한 것으로, CNN을 학습하는 동안 가중치의 엔트로피를 통계적 가중치로 매핑하는 가지치기를 위한 임계값을 결정할 수 있다. 훈련을 반복하는 동안 임계값은 엔트로피 및 가중치의 분포에 따라 동적으로 결정될 수 있다. 이러한 프로세스는 훈련이 끝날 때까지 반복적으로 계속 진행되므로 정리된 CNN이 도 1의 오른쪽에 표시된 것과 같이 가중치가 일부 제거된 형태를 가질 수 있다. 엔트로피는 하기 수학식 1에서 볼 수 있듯이 가중치의 확률 정보를 기대하여 얻을 수 있다.

[0040] [수학식 1]

[0041] $$E[W] = \sum_k Pr(\omega_k) \mathrm{A}[\omega_k]$$

[0042] 여기서 E[W]는 가중치의 엔트로피이고, ω_k는 CNN의 k번째 가중치이고, $Pr(\omega_k)$은 가중치가 발생할 확률이며, $A[\omega_k]$는 정보의 양이다. $A[\omega_k]$ 및 $Pr(\omega_k)$은 반비례 관계이다. $A[\omega_k]$는 하기 수학식 2에서와 같이 가중치의 확률론적 사건 로그로 추정할 수 있다.

[0043] [수학식 2]

[0044] $$\mathrm{A}[\omega_k] = -\log Pr(\omega_k)$$

[0045] CNN의 불확실성이 크다는 것은 훈련 중 가중치의 출력 기여도에 대한 정보를 얻을 수 있다는 것을 의미한다. 다시 말해서 엔트로피가 높으면 더 많은 정보를 얻을 수 있다. 반대로 엔트로피가 낮으면, CNN은 이미 안정된 상태에 있고, 이용할 수 있는 정보가 거의 없다. 엔트로피가 높을 때 가지를 치면 가중치와 관련된 정보를 더 쉽게 얻을 수 있다. 따라서 현재 엔트로피보다 낮은 정보를 가진 가중치는 중요하지 않은 것으로 간주될 수 있다. 따라서 현재 엔트로피보다 낮은 정보를 가진 가중치가 더 잘 제거될 수 있다. 수학식 2와 같이 정보는 확률론적 분포를 사용하여 계산한다. 데이터 크기가 클 경우 확률 분포는 중심 극한 정리에 따라 가우스 분포로 가정할 수 있다. 일반적으로 CNN의 가중치는 매우 많기 때문에, 본 발명은 다음과 같은 수학식 3과 같이 가중치가 가우스 분포를 가지고 있다는 가정 하에 가우스 분포에 대한 수식을 도출한다.

[0046] [수학식 3]

$$G(\omega_k|\mu_\omega,\sigma_\omega) = \frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}$$

[0047]

[0048] 여기서 μ_ω는 가중치 평균이고, σ_ω는 가중치 표준 편차이며, $g(\omega_k|\mu_\omega, \sigma_\omega)$는 가중치의 분포다. 일반적으로, 가중치의 값이 매우 작으면, 가중치의 산출 기여도는 중요하지 않은 것으로 간주되므로, 훈련 중에 가중치를 줄일 수 있다. 이러한 관점에서, 본 발명에서는 그 가중치의 값이 작다는 이유만으로 안정적이지 않다고 판단할 수 없기 때문에 작은 가중치 값에 대한 잠재성을 분석할 필요가 있다. 훈련 반복 중에는 훈련 성과가 수렴될 때까지 작은 값의 가중치의 기여도가 증가될 수도 있다. 작은 가중치의 잠재력은 엔트로피를 사용하여 추정할 수 있다. 수학식 2에서 볼 수 있듯이 가중치의 값이 작고 μ_ω에 가까울 경우 가우스 분포가 평균에 및 근처에 있을 확률이 가장 높기 때문에 가중치의 정보 양이 작다. 그것은 다음 반복에 거의 기여하지 않는 것으로 간주될 수 있다. 따라서 가지치기를 하는 것은 훈련 성능을 해치지 않는다. 가중치 및 엔트로피 공식의 가우스 분포를 사용하면 수학식 3을 수학식 1에 대입시켜 다음과 같은 수학식 4를 얻을 수 있다.

[0049] [수학식 4]

$$E[W] = -\int\left(\frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}\right)\log\left(\frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}\right)d\omega_k$$

$$= \frac{1}{2}(1 + ln(2\pi\sigma_\omega{}^2))$$

$$= ln(\sigma_\omega\sqrt{2\pi e})$$

[0050]

[0051] 여기서 σ_ω는 가중치의 표준 편차다. 가중치는 가우스 분포로 가정되기 때문에 다음과 같이 수학식 5에 가중치 분포를 간단한 확률 간격으로 통합함으로써 수학식 2을 사용하여 n번째 중량, ω_n의 정보를 얻을 수 있다.

[0052] [수학식 5]

$$A[\omega_n] = \left(-\int_{k=n}^{n+\varepsilon}\log\left(\frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}\right)d\omega_k\right)$$

[0053]

[0054] 수학식 5을 사용하여 계산된 ω_n, $A[\omega_n]$에 대한 정보량이 수학식 4을 사용하여 계산된 임계값보다 작으면 훈련 중 중요하지 않은 가중치로 판단하여 상기 ω_n을 삭제할 수 있다. 그러나 가중치의 분포가 일반적으로 가우스 분포에 완전히 맞지 않기 때문에 수학식 5와 수학식 4를 직접 비교하면 무게의 중요성이 결정될 수 없다. 가지치기 결정은 수학식 6과 같이 조정해야 한다.

[0055] [수학식 6]

[0056] $$A[\omega_n] \leq ln(\sigma_\omega \sqrt{2\pi e}) \times \alpha$$

[0057] 여기서 α는 조정 계수이다.

[0058] 일반적으로, 가중치의 수는 완전히 연결된 레이어가 CNN의 커널을 사용하는 컨벌루션레이어의 수보다 훨씬 더 크다. 이것은 완전히 연결된 레이어의 가중치 분포가 가우스 분포에 가깝다는 것을 의미한다. 상기 수학식 6은 완전히 연결된 레이어의 가중치 가지치기에 잘 맞는다. 컨벌루션 레이어의 임계치는 완전히 연결된 층보다 약간 큰 α의 값으로 설정해야 한다. 또한, 훈련 초기에 임계값을 결정하고 가중치를 줄이기 시작할 때 문제가 된다. 훈련을 시작할 때부터 가지치기 작업을 수행하면 가중치가 임의로 초기화되기 때문에 가중치에 대한 엔트로피 정보가 신뢰할 수 없다. 따라서, 가지치기 시작점을 찾기 위해 학습 비용과 학습 중 결과 노드 사이의 비교에서 안정적인 상태를 제공하는 학습 곡선을 사용할 수 있다. 일단 시작점이 결정되면, 본 발명의 일 실시 예에 따른 가지치기 프로세스는 정확도가 저하되지 않는 한 반복적으로 계속될 수 있다.

[0059] 도 2는 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템의 블록도이다.

[0060] 도 2를 참조하면, 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 시스템은 연산부(210), 가지치기부(220) 및 학습부(230)를 포함할 수 있다.

[0061] 상기 연산부(210)는 시스템에 포함되는 CNN의 가중치의 정보의 양을 계산할 수 있다. 상기 연산부(210)는 시스템에 포함되는 CNN의 가중치를 가지치기 하기 위한 임계값을 연산할 수 있다.

[0062] 상기 연산부(210)는 엔트로피는 하기 수학식 1에서 볼 수 있듯이 가중치의 확률 정보를 기대하여 얻을 수 있다.

[0063] [수학식 1]

[0064] $$E[W] = \sum_k Pr(\omega_k) A[\omega_k]$$

[0065] 여기서 E[W]는 가중치의 엔트로피이고, ω_k는 CNN의 k번째 가중치이고, $Pr(\omega_k)$은 가중치가 발생할 확률이며, $A[\omega_k]$는 정보의 양이다. $A[\omega_k]$ 및 $Pr(\omega_k)$은 반비례 관계이다.

[0066] 상기 연산부(210)는 $A[\omega_k]$를 하기 수학식 2에서와 같이 가중치의 확률론적 사건 로그로 추정할 수 있다.

[0067] [수학식 2]

[0068] $$A[\omega_k] = -\log Pr(\omega_k)$$

[0069] CNN의 불확실성이 크다는 것은 훈련 중 가중치의 출력 기여도에 대한 정보를 얻을 수 있다는 것을 의미한다. 다시 말해서 엔트로피가 높으면 더 많은 정보를 얻을 수 있다. 반대로 엔트로피가 낮으면, CNN은 이미 안정된 상태에 있고, 이용할 수 있는 정보가 거의 없다. 엔트로피가 높을 때 가지를 치면 가중치와 관련된 정보를 더 쉽게 얻을 수 있다. 따라서 현재 엔트로피보다 낮은 정보를 가진 가중치는 중요하지 않은 것으로 간주될 수 있다. 따라서 현재 엔트로피보다 낮은 정보를 가진 가중치가 더 잘 제거될 수 있다. 수학식 2와 같이 정보는 확률론적 분포를 사용하여 계산한다. 데이터 크기가 클 경우 확률 분포는 중심 극한 정리에 따라 가우스 분포로 가정할 수 있다. 일반적으로 CNN의 가중치는 매우 많기 때문에, 상기 연산부(210)는 다음과 같은 수학식 3과 같이 가중치가 가우스 분포를 가지고 있다는 가정 하에 가우스 분포를 추정할 수 있다.

[0070] [수학식 3]

[0071] $$G(\omega_k|\mu_\omega, \sigma_\omega) = \frac{1}{\sigma_\omega \sqrt{2\pi}} e^{\left(-\frac{(\omega_k - \mu_\omega)^2}{2\sigma_\omega^2}\right)}$$

[0072] 여기서 μ_ω는 가중치 평균이고, σ_ω는 가중치 표준 편차이며, $g(\omega_k|\mu_\omega, \sigma_\omega)$는 가중치의 분포다. 일반적으로, 가중치의 값이 매우 작으면, 가중치의 산출 기여도는 중요하지 않은 것으로 간주되므로, 훈련 중에 가중치를 줄일 수 있다. 이러한 관점에서, 본 발명에서는 그 가중치의 값이 작다는 이유만으로 안정적이지 않다고 판단할 수 없기 때문에 작은 가중치 값에 대한 잠재성을 분석할 수 있다. 훈련 반복 중에는 훈련 성과가 수렴될 때까지

작은 값의 가중치의 기여도가 증가될 수도 있다. 작은 가중치의 잠재력은 엔트로피를 사용하여 추정할 수 있다. 수학식 2에서 볼 수 있듯이 가중치의 값이 작고 μ_ω에 가까울 경우 가우스 분포가 평균에 및 근처에 있을 확률이 가장 높기 때문에 가중치의 정보 양이 작다. 그것은 다음 반복에 거의 기여하지 않는 것으로 간주될 수 있다. 따라서 가지치기를 하는 것은 훈련 성능을 해치지 않는다. 상기 연산부(210)는 다음과 같은 수학식 4를 연산해 엔트로피의 값을 얻을 수 있다.

[0073] [수학식 4]

$$E[W] = -\int \left(\frac{1}{\sigma_\omega \sqrt{2\pi}} e^{\left(-\frac{(\omega_k - \mu_\omega)^2}{2\sigma_\omega{}^2} \right)} \right) \log \left(\frac{1}{\sigma_\omega \sqrt{2\pi}} e^{\left(-\frac{(\omega_k - \mu_\omega)^2}{2\sigma_\omega{}^2} \right)} \right) d\omega_k$$

$$= \frac{1}{2}(1 + ln(2\pi\sigma_\omega{}^2))$$

$$= ln(\sigma_\omega \sqrt{2\pi e})$$

[0074]

[0075] 여기서 σ_ω는 가중치의 표준 편차다. 가중치는 가우스 분포로 가정되기 때문에 상기 연산부(210)는 다음과 같이 수학식 5에 가중치 분포를 간단한 확률 간격으로 통합함으로써 수학식 2을 사용하여 n번째 중량, ω_n의 정보를 얻을 수 있다.

[0076] [수학식 5]

$$A[\omega_n] = \left(-\int_{k=n}^{n+\varepsilon} \log \left(\frac{1}{\sigma_\omega \sqrt{2\pi}} e^{\left(-\frac{(\omega_k - \mu_\omega)^2}{2\sigma_\omega{}^2} \right)} \right) d\omega_k \right)$$

[0077]

[0078] 상기 가지치기부(220)는 상기 연산부(210)의 연산 결과를 이용해 가지치기할 가중치를 결정하고 제거할 수 있다. 상기 가지치기부(220)는 상기 CNN의 복수의 가중치 중 정보의 양이 상기 임계값 이하인 가중치를 제거할 수 있다. 상기 임계값은 $ln(\sigma_\omega \sqrt{2\pi e}) \times \alpha$일 수 있다.

[0079] 상기 가지치기부(220)는 수학식 5을 사용하여 계산된 ω_n, $A[\omega_n]$에 대한 정보량이 수학식 4을 사용하여 계산된 임계값보다 작으면 훈련 중 중요하지 않은 가중치로 판단하여 상기 ω_n을 삭제할 수 있다. 그러나 가중치의 분포가 일반적으로 가우스 분포에 완전히 맞지 않기 때문에 수학식 5와 수학식 4를 직접 비교하면 무게의 중요성이 결정될 수 없다. 가지치기 결정은 수학식 6과 같이 조정해야 한다.

[0080] [수학식 6]

$$A[\omega_n] \leq ln(\sigma_\omega \sqrt{2\pi e}) \times \alpha$$

[0081]

[0082] 여기서 α는 조정 계수이다.

[0083] 상기 가지치기부(220)는 컨벌루션 레이어의 임계치는 완전히 연결된 층보다 약간 큰 α의 값으로 설정할 수 있다. 또한, 상기 가지치기부(220)는 훈련 초기에 임계값을 결정하고 가중치를 줄이기 시작할 때 문제가 된다. 훈련을 시작할 때부터 가지치기 작업을 수행하면 가중치가 임의로 초기화되기 때문에 가중치에 대한 엔트로피 정보가 신뢰할 수 없다. 따라서, 가지치기 시작점을 찾기 위해 학습 비용과 학습 중 결과 노드 사이의 비교에서 안정적인 상태를 제공하는 학습 곡선을 사용할 수 있다. 일단 시작점이 결정되면, 본 발명의 일 실시 예에 따른 가지치기 프로세스는 정확도가 저하되지 않는 한 반복적으로 계속될 수 있다.

[0084] 상기 α는 상기 시스템에 포함되는 복수의 컨벌루션 레이어 및 복수의 완전 연결 레이어가 각각의 값을 가질 수 있다.

[0085] 상기 학습부(230)는 CNN을 이용한 학습 또는 트레이닝을 수행할 수 있다. 상기 학습부(230)가 학습 또는 트레이

닝을 수행하는 동안 상기 연산부(210) 및 상기 가지치기부(220)는 상기 학습부(230)의 CNN의 가중치를 가지치기 할 수 있다. 상기 학습부(230)는 상기 임계값 기준에 따라 일부 가중치를 제거하여 가지치기된 CNN을 이용해 계속해서 학습 또는 트레이닝을 수행할 수 있다.

[0086] 상기 연산부(210) 및 상기 가지치기부(220)는 상기 학습부(230)가 학습을 수행하는 동안 반복적으로 가지치기를 수행할 수 있다. 예를 들어, 가지치기 후 학습을 통해 안정화되는 가중치가 생기면 추가적 작업을 통해 추가적 가지치기를 수행해 학습의 부하를 감소시켜 학습 속도를 점진적으로 증가시킬 수 있다.

[0087] 도 3은 신경망(Neural Networks)의 학습 곡선이다.

[0088] 도 3을 참조하면, 본 바발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법은 2 개의 컨벌루션 레이어와 2 개의 완전히 연결된 레이어로 구성된 CNN으로 구성된 LeNet-5 네트워크를 포함할 수 있다. 입력 레이어의 다음 레이어인 첫 번째 콘볼루션 레이어 (C1)에는 500 개의 가중치가 포함될 수 있다. 제 1 콘볼 루션 레이어의 다음 층인 제 2 콘볼 루션 레이어(C2)는 25,000 개의 가중치를 가지며; 제 1 완전 연결 레이어(F1)는 400,000 가중치를 가지며, 제 2 완전 연결 레이어(F2)는 5000 개의 가중치를 갖는다. CNN은 TensorFlow 프레임 워크를 사용하여 50,000 개의 이미지로 학습을 하고 MNIST 데이터 세트에서 10,000 개의 이미지로 테스트 했다. 본 발명의 일 실시 예에서 컴퓨터의 사양에는 NVIDIA Titan X Pascal 그래픽 처리 장치가 포함될 수 있다. 본 발명은 TensorFlow를 수정하여 훈련 중에 잘라내 진 체중을 무시하는 마스크를 추가할 수 있다. 도 3은 약 8000 회 반복에서 학습 곡선이 채도 방향으로 수렴한다는 것을 보여준다. 본 발명의 일 실시 예에서 반복 8000회에서 가지치기를 시작할 수 있다.

[0089] 도 4는 본 발명의 일 실시 예에 따른 문턱, 엔트로피 및 반복 8000에서 각 레이어의 작은 값의 가중치 정보를 도시한 그래프이다.

[0090] 도 4를 참조하면, 도 4 (a), 도 4 (b), 도 4 (c), 도 4 (d)는 각각 C1, C2, F1, F2의 가중치 정보를 도시한 것이다. 엔트로피와 각 레이어에 대한 작은 가중치 정보의 양에서 결정된 임계 값에 의한 가지 치기 프로세스를 확인할 수 있다. 본 발명의 일 실시 예에서 반복 8000회에서 시작하여 매 100 회 반복 한 후에 가지치기를 수행할 수 있다. 반복 횟수 8000에서 도 4의 위 그래프는 도 4의 아래 그래프인 가중치 분포에서 상기 수학식 5를 사용하여 계산 된 각 가중치의 플롯이다. 본 발명읠 일 실시 예에서 이러한 분포를 바탕으로 수학식 4를 이용하여 계산 된 C1, C2, F1, F2의 엔트로피는 각각 1.66, 1.52, 1.16 및 1.71이다. 도 4 아래 그래프에서와 같이 가중치 분포가 완벽하게 가우스가 아니기 때문에 엔트로피를 직접 적용 할 수 없다.

[0091] 도 5는 본 발명의 일 실시 예에 따른 각 레이어에 대한 α_k의 값에 따른 훈련 정확도이다.

[0092] 도 5를 참조하면, 수학식 6에서 확인할 수 있는 바와 같이, 각 레이어에 대한 α_k의 값에 따른 훈련 정확도이다. α_k의 값은 0.1에서 1.2까지 0.1씩 증가할 수 있다. C1의 경우, $\alpha_1 = 0.1$에 대한 훈련 정확도는 98.89 %이며, α_1의 값을 증가 시키면 $\alpha_1 = 0.5$까지 눈에 띄는 성능 저하는 없다. 그러나 정확도는 $\alpha_1 = 0.6$에서 현저하게 떨어진다. 따라서, 0.5는 제 1 콘볼 루션 레이어의 최적 조정 인자일 수 있다. 이는 C1 (1.66)에 대한 엔트로피의 값을 가지 치기 후에 정확도를 떨어 뜨리지 않기 위해 0.83으로 조정해야 함을 의미한다. 같은 방식으로, C2, F1 및 F2에 대한 최적 조정 계수는 각각 $\alpha_2 = 0.7$, $\alpha_3 = 1.1$ 및 $\alpha_4 = 0.6$으로 설정될 수 있다. 따라서, C2, F1 및 F2의 엔트로피는 각각 1.52, 1.16 및 1.71에서 1.05, 1.27 및 1.02로 조정된다. 상기 결과에서, C1은 가장 많이 조정되었지만 F1은 가장 작다. 왜냐하면 C1은 이전 섹션에서 설명한 것처럼 F1 (400,000)에 비해 매우 적은 가중치 (500)를 갖기 때문이다.

[0093] 도 6는 본 발명의 일 실시 예에 따른 훈련 중 $\alpha_1 = 0.5$, $\alpha_2 = 0.7$, $\alpha_3 = 1.1$, $\alpha_4 = 0.6$ 인 각 레이어의 압축률이다.

[0094] 도 6을 참조하면, 훈련 중 $\alpha_1 = 0.5$, $\alpha_2 = 0.7$, $\alpha_3 = 1.1$, $\alpha_4 = 0.6$ 인 각 레이어의 압축률을 보여준다. C1의 압축률은 C2, F1 및 F2에 비해 점차적으로 증가한다. 이는 가중치의 제거가 입력 레이어에 가까울수록 성능에 더 민감하기 때문이다.

[0095] 도 7은 테스트 데이터의 정확도와 8000에서 시작하는 1000 회 반복마다의 전체 압축률을 도시한 것이다.

[0096] 도 7을 참조하면, 테스트 데이터의 정확도와 8000에서 시작하는 1000 회 반복마다의 전체 압축률을 보여준다. 결과에서 볼 수 있듯이 테스트 데이터의 정확도는 잘라낸 CNN에 대해 허용 가능한 수준에 도달하며 전체 압축률

은 94 %이다.

[0097] 도 8은 본 발명의 일 실시 예에 따른 방법과 기존의 방법의 압축 성능을 비교한 것이고, 도 9는 본 발명의 일 실시 예에 방법과 기존의 방법의 가지치기 속도를 비교한 것이다.

[0098] 도 8 및 도 9를 참조하면, 본 발명의 일 실시 예에 따른 방법을 압축 방법과 가지치키 실행 시간과 관련하여 Han의 방법(HAN, Song, et al. Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems. 2015. p. 1135-1143.)과 비교했다. Han의 방법은 가중치의 평균 및 분산을 기반으로 결정된 특정 임계 값보다 작은 값의 가중치를 잘라낸다. 정확도 성능에 대해서도 비슷한 비교를 하였다. 첫 번째 컨볼루션 레이어(C1)의 경우, 본 발명의 방법과 관련 방법은 각각 500 개의 가중치 중 235와 170을 잘라 냈다. 두 번째 컨볼 루션 레이어 (C2)의 경우 25,000 개의 가중치 중 23,000 및 22,000 개다. 첫 번째 완전히 연결된 레이어 (F1)의 경우 400,000 개의 가중치 중 376,000 및 368,000 개다. 제 2 완전 연결 레이어 (F2)의 경우, 5000 가중치 중 4250 및 4050 개다. 결과는 본 발명의 방법이 기존의 기술에 비해 C1, C2, F1 및 F2에 대해 프로세스를 20 %, 33 %, 25 % 및 31 % 향상시키는 것을 보여준다. 전체적으로 우리의 방법은 관련 방법에 비해 압축률을 28.25 % 향상시킨다.

[0099] 도 9를 참조하면, 기존 기술에 대비한 본 발명의 방법의 가지 치기 속도는 각각 395 및 459 초다. 본 발명의 방법은 기존 기술의 방법보다 14 % 빠르다. 이는 본 발명의 방법이 하나의 완전한 트레이닝 기간 동안에 재귀적으로 제거 프로세스를 실행하기 때문이다. 그러나, 기존 기술은 잘라낸 CNN 모델을 미세 조정 한 후에 반복적으로 제거 프로세스를 실행한다.

[0100] 본 발명은 CNN을 훈련하는 동안 가중치에서 나온 평균 정보량을 바탕으로 엔트로피 기반의 가지치기 방법을 수행한다. 정보는 엔트로피 기법을 사용하여 확률적 방식으로 측정한다. 임계치에 대한 수학 공식을 도출하기 위해, 본 발명은 가중치의 통계적 분포가 가우스라고 가정했다. 기존의 가지치기 방법은 훈련이 완전히 종료된 후 무게 또는 입력 기능의 중요성을 파악한다. 이러한 제거 프로세스는 정확도가 저하될 때까지 반복된다. 하지만, 본 발명의 방법은 훈련하는 동안 가중치의 중요성을 고려한다. 그렇기 때문에 본 발명의 방법은 훈련 중 가중치가 결과 노드에 미치는 영향을 고려할 수 있다. 이 실험에서 우리는 우리의 방법이 압축률과 제거 속도에 관해서 잘 알려진 제거 기술을 어떻게 능가하는지를 보여주었다.

[0101] 도 10은 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법의 흐름도이다.

[0102] 도 10을 참조하면, 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법은 CNN의 복수의 가중치들의 엔트로피에 기반한 정보의 양을 연산하는 (a) 단계(S1310)를 포함할 수 있다.

[0103] S1310 단계에서, 상기 연산부(210)는 시스템에 포함되는 CNN의 가중치의 정보의 양을 계산할 수 있다. 상기 연산부(210)는 시스템에 포함되는 CNN의 가중치를 가지치기 하기 위한 임계값을 연산할 수 있다.

[0104] 상기 연산부(210)는 엔트로피는 하기 수학식 1에서 볼 수 있듯이 가중치의 확률 정보를 기대하여 얻을 수 있다.

[0105] [수학식 1]

[0106] $$E[W] = \sum_k Pr(\omega_k)\mathrm{A}[\omega_k]$$

[0107] 여기서 E[W]는 가중치의 엔트로피이고, ω_k는 CNN의 k번째 가중치이고, Pr(ω_k)은 가중치가 발생할 확률이며, A[ω_k]는 정보의 양이다. A[ω_k] 및 Pr(ω_k)은 반비례 관계이다.

[0108] 상기 연산부(210)는 A[ω_k]를 하기 수학식 2에서와 같이 가중치의 확률론적 사건 로그로 추정할 수 있다.

[0109] [수학식 2]

[0110] $$\mathrm{A}[\omega_k] = -\log Pr(\omega_k)$$

[0111] CNN의 불확실성이 크다는 것은 훈련 중 가중치의 출력 기여도에 대한 정보를 얻을 수 있다는 것을 의미한다. 다시 말해서 엔트로피가 높으면 더 많은 정보를 얻을 수 있다. 반대로 엔트로피가 낮으면, CNN은 이미 안정된 상태에 있고, 이용할 수 있는 정보가 거의 없다. 엔트로피가 높을 때 가지를 치면 가중치와 관련된 정보를 더 쉽게 얻을 수 있다. 따라서 현재 엔트로피보다 낮은 정보를 가진 가중치는 중요하지 않은 것으로 간주될 수 있다. 따라서 현재 엔트로피보다 낮은 정보를 가진 가중치가 더 잘 제거될 수 있다. 수학식 2와 같이 정보는 확률론적 분포를 사용하여 계산한다. 데이터 크기가 클 경우 확률 분포는 중심 극한 정리에 따라 가우스 분포로 가정할

수 있다. 일반적으로 CNN의 가중치는 매우 많기 때문에, 상기 연산부(210)는 다음과 같은 수학식 3과 같이 가중치가 가우스 분포를 가지고 있다는 가정 하에 가우스 분포를 추정할 수 있다.

[0112] [수학식 3]

[0113] $$G(\omega_k|\mu_\omega,\sigma_\omega) = \frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}$$

[0114] 여기서 μ_ω는 가중치 평균이고, σ_ω는 가중치 표준 편차이며, $g(\omega_k|\mu_\omega, \sigma_\omega)$는 가중치의 분포다. 일반적으로, 가중치의 값이 매우 작으면, 가중치의 산출 기여도는 중요하지 않은 것으로 간주되므로, 훈련 중에 가중치를 줄일 수 있다. 이러한 관점에서, 본 발명에서는 그 가중치의 값이 작다는 이유만으로 안정적이지 않다고 판단할 수 없기 때문에 작은 가중치 값에 대한 잠재성을 분석할 수 있다. 훈련 반복 중에는 훈련 성과가 수렴될 때까지 작은 값의 가중치의 기여도가 증가될 수도 있다. 작은 가중치의 잠재력은 엔트로피를 사용하여 추정할 수 있다. 수학식 2에서 볼 수 있듯이 가중치의 값이 작고 μ_ω에 가까울 경우 가우스 분포가 평균에 및 근처에 있을 확률이 가장 높기 때문에 가중치의 정보 양이 작다. 그것은 다음 반복에 거의 기여하지 않는 것으로 간주될 수 있다. 따라서 가지치기를 하는 것은 훈련 성능을 해치지 않는다. 상기 연산부(210)는 다음과 같은 수학식 4를 연산해 엔트로피의 값을 얻을 수 있다.

[0115] [수학식 4]

[0116] $$E[W] = -\int\left(\frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}\right)\log\left(\frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}\right)d\omega_k$$

$$= \frac{1}{2}(1 + ln(2\pi\sigma_\omega{}^2))$$

$$= ln(\sigma_\omega\sqrt{2\pi e})$$

[0117] 여기서 σ_ω는 가중치의 표준 편차다. 가중치는 가우스 분포로 가정되기 때문에 상기 연산부(210)는 다음과 같이 수학식 5에 가중치 분포를 간단한 확률 간격으로 통합함으로써 수학식 2을 사용하여 n번째 중량, ω_n의 정보를 얻을 수 있다.

[0118] [수학식 5]

[0119] $$A[\omega_n] = \left(-\int_{k=n}^{n+\varepsilon}\log\left(\frac{1}{\sigma_\omega\sqrt{2\pi}}e^{\left(-\frac{(\omega_k-\mu_\omega)^2}{2\sigma_\omega{}^2}\right)}\right)d\omega_k\right)$$

[0120] 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법은 상기 복수의 가중치들 중 가지치기할 가중치를 결정하기 위한 임계값을 연산하는 (b) 단계(S1320)를 포함할 수 있다.

[0121] S1320 단계에서, 상기 임계값은 $ln(\sigma_\omega\sqrt{2\pi e}) \times \alpha$일 수 있다.

[0122] 상기 가지치기부(220)는 수학식 5을 사용하여 계산된 ω_n, $A[\omega_n]$에 대한 정보량이 수학식 4을 사용하여 계산된 임계값보다 작으면 훈련 중 중요하지 않은 가중치로 판단하여 상기 ω_n을 삭제할 수 있다. 그러나 가중치의 분포가 일반적으로 가우스 분포에 완전히 맞지 않기 때문에 수학식 5와 수학식 4를 직접 비교하면 무게의 중요성이 결정될 수 없다. 가지치기 결정은 수학식 6과 같이 조정해야 한다.

[0123] [수학식 6]

[0124] $$A[\omega_n] \leq ln(\sigma_\omega \sqrt{2\pi e}) \times \alpha$$

[0125] 여기서 α는 조정 계수이다.

[0126] 상기 가지치기부(220)는 컨벌루션 레이어의 임계치는 완전히 연결된 층보다 약간 큰 α의 값으로 설정할 수 있다. 또한, 상기 가지치기부(220)는 훈련 초기에 임계값을 결정하고 가중치를 줄이기 시작할 때 문제가 된다. 훈련을 시작할 때부터 가지치기 작업을 수행하면 가중치가 임의로 초기화되기 때문에 가중치에 대한 엔트로피 정보가 신뢰할 수 없다. 따라서, 가지치기 시작점을 찾기 위해 학습 비용과 학습 중 결과 노드 사이의 비교에서 안정적인 상태를 제공하는 학습 곡선을 사용할 수 있다.

[0127] 상기 α는 상기 시스템에 포함되는 복수의 컨벌루션 레이어 및 복수의 완전 연결 레이어가 각각의 값을 가질 수 있다.

[0128] 본 발명의 일 실시 예에 따른 엔트로피 기반 신경망(Neural Networks) 가지치기 방법은 상기 정보의 양이 임계값 이하의 값을 가지는 가중치들을 가지치기하는 (c) 단계(S1330)를 포함할 수 있다.

[0129] S1330 단계에서, 상기 가지치기부(220)는 상기 연산부(210)의 연산 결과를 이용해 가지치기할 가중치를 결정하고 제거할 수 있다. 상기 가지치기부(220)는 상기 CNN의 복수의 가중치 중 정보의 양이 상기 임계값 이하인 가중치를 제거할 수 있다.

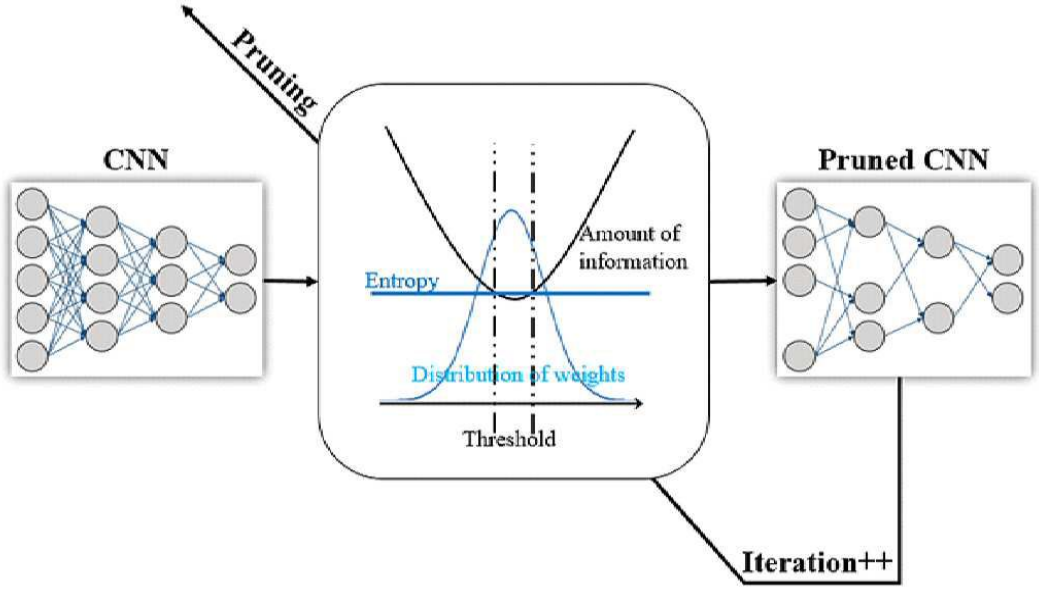
[0130] 일단 시작점이 결정되면, 본 발명의 일 실시 예에 따른 가지치기 프로세스는 정확도가 저하되지 않는 한 반복적으로 계속될 수 있다.

[0131] 본 발명의 일 실시 예에 따르면, 상기 (a) ~ (c) 단계는 상기 CNN이 학습을 수행하는 중에 수행될 수 있다.

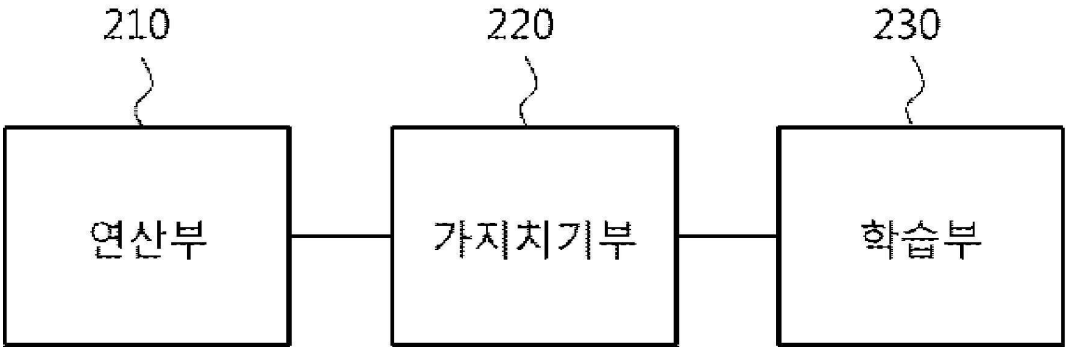
[0133] 이제까지 본 발명에 대하여 그 바람직한 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면

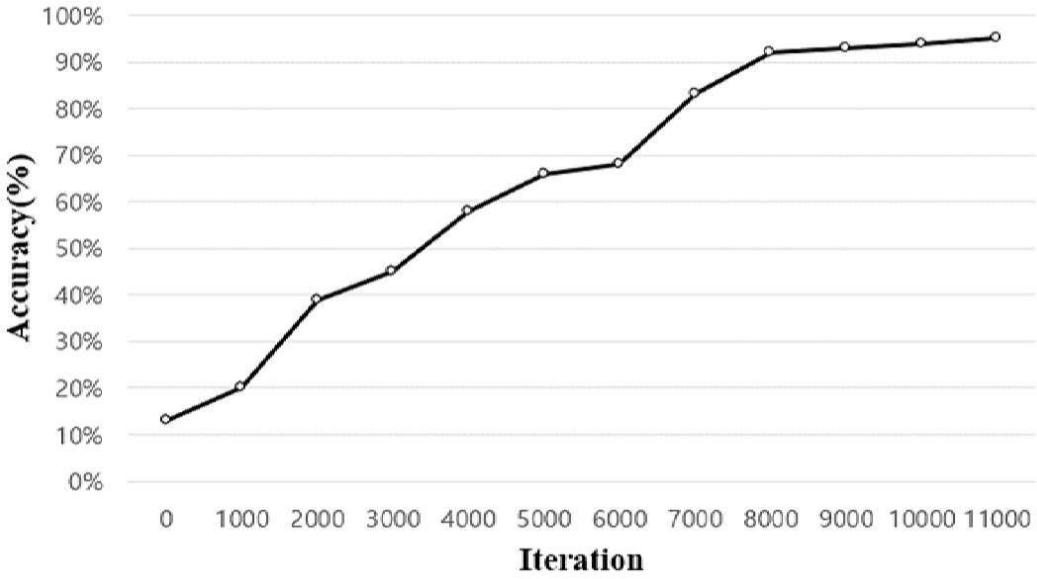
도면1


Pruning
CNN
Amount of information
Entropy
Distribution of weights
Threshold
Pruned CNN
Iteration++

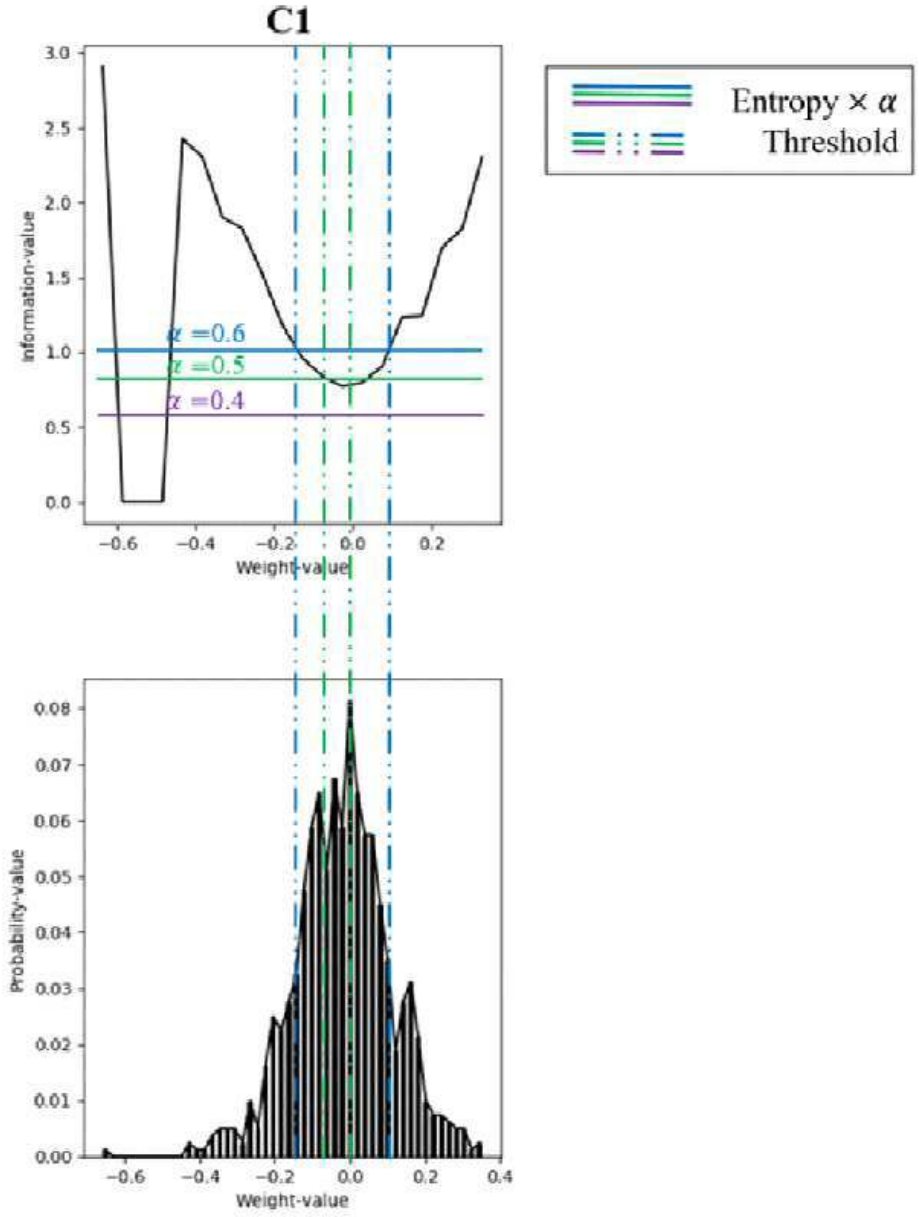
도면2


210
220
230
연산부
가지치기부
학습부

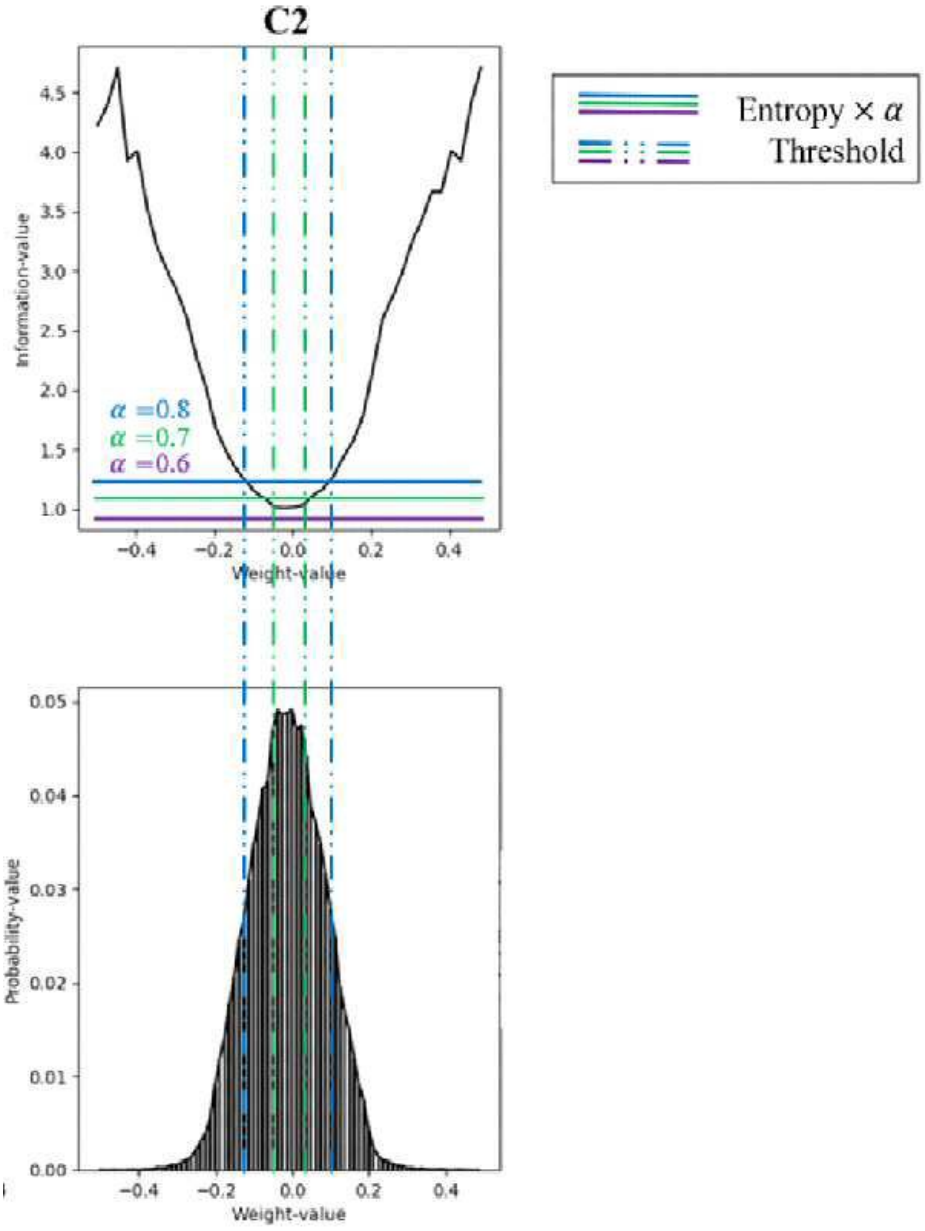
도면3


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Accuracy(%)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
Iteration

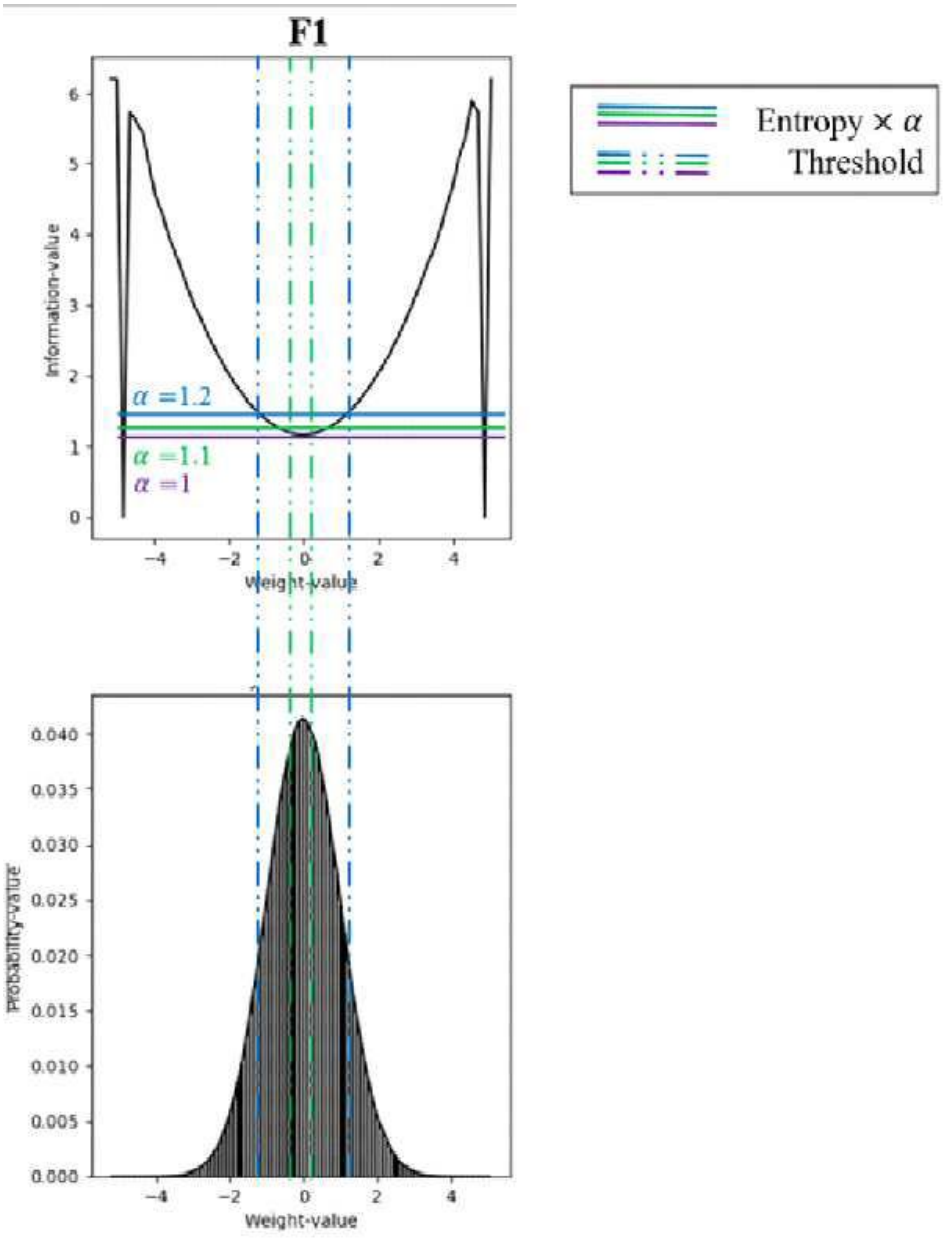
도면4a


C1
Entropy × α
Threshold
Information-value
Weight-value
α =0.6
α =0.5
α =0.4
Probability-value
Weight-value

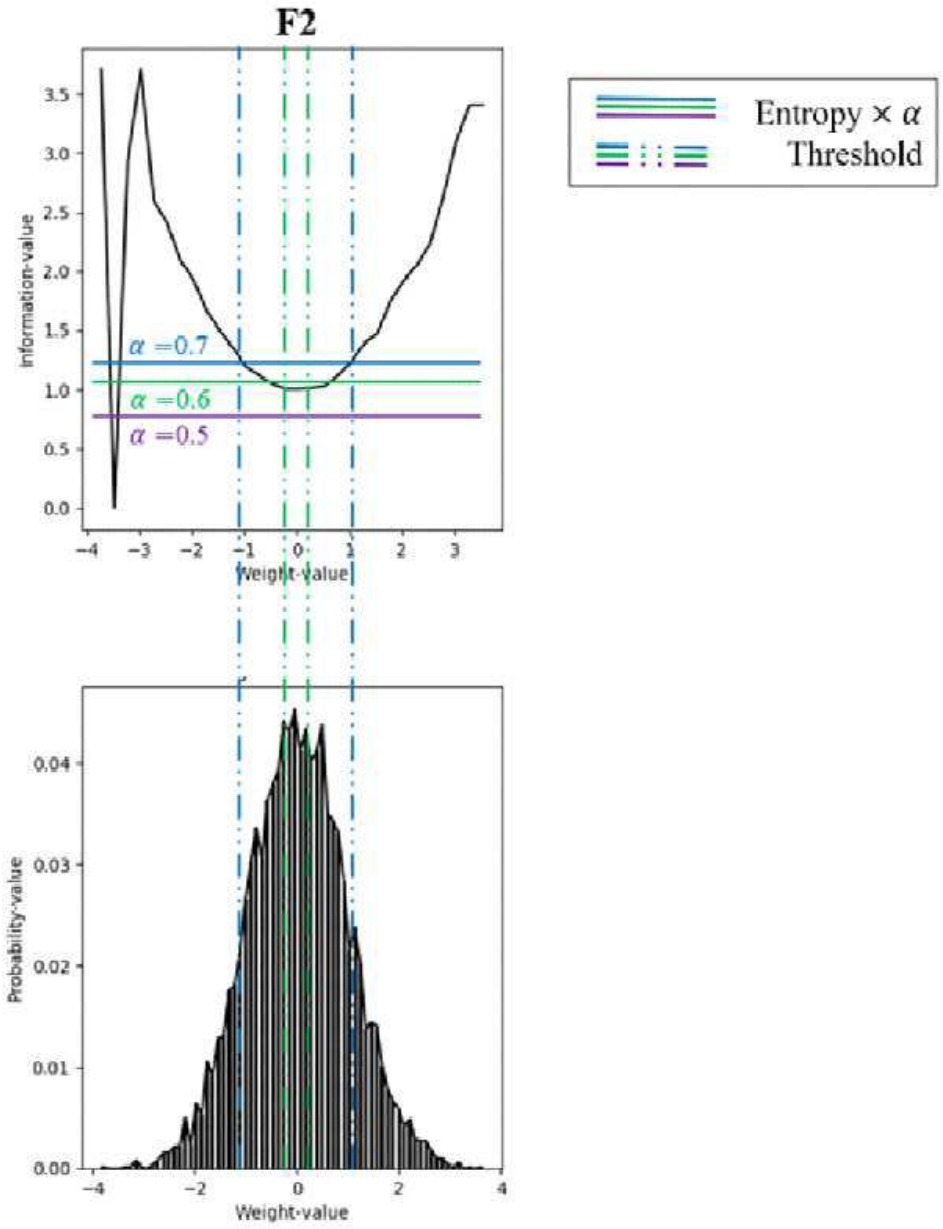
도면4b


C2
Entropy × α
Threshold
Information-value
α =0.8
α =0.7
α =0.6
Weight-value
Probability-value
Weight-value

도면4c


F1
Information-value
Weight-value
α =1.2
α =1.1
α =1
Entropy × α
Threshold
Probability-value
Weight-value

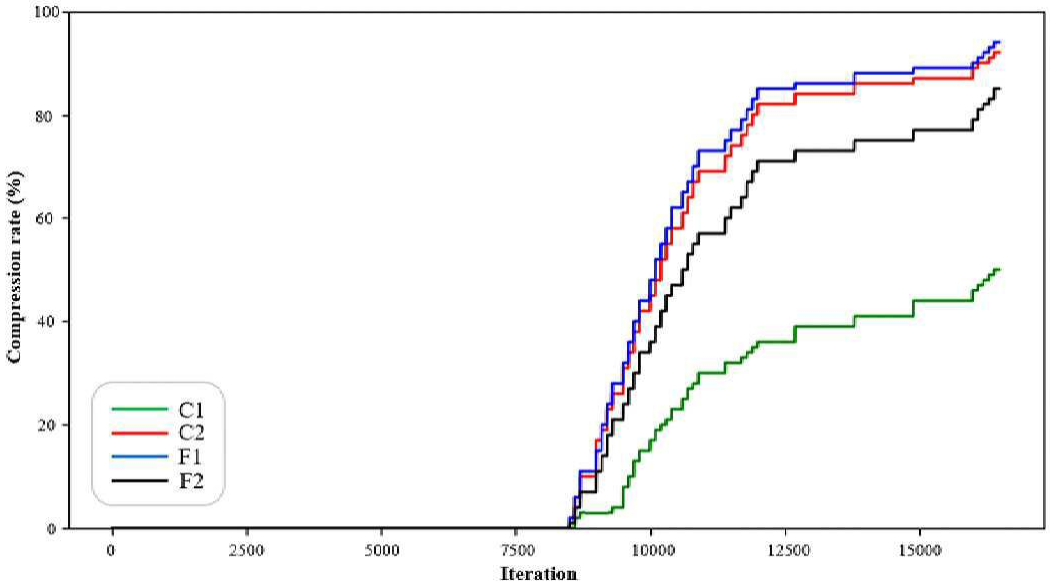
도면4d



도면5

$\alpha - value$	C1	C2	F1	F2	$\alpha - value$	C1	C2	F1	F2
0.1	98.89%	98.77%	99.30%	99.95%	0.7	65.94%	**98.12%**	99.18%	95.55%
0.2	98.79%	98.72%	99.28%	99.95%	0.8	35.78%	94.21%	99.22%	95.05%
0.3	98.82%	98.69%	99.26%	99.92%	0.9	18.16%	71.42%	99.14%	91.01%
0.4	98.81%	98.7%	99.21%	98.94%	1	0%	52.91%	99.09%	68.51%
0.5	**97.75%**	98.54%	99.25%	98.82%	1.1	0%	26.25%	**97.53%**	26.12%
0.6	82.26%	98.66%	99.26%	**98.05%**	1.2	0%	12.35%	82.73%	15.73%

도면6



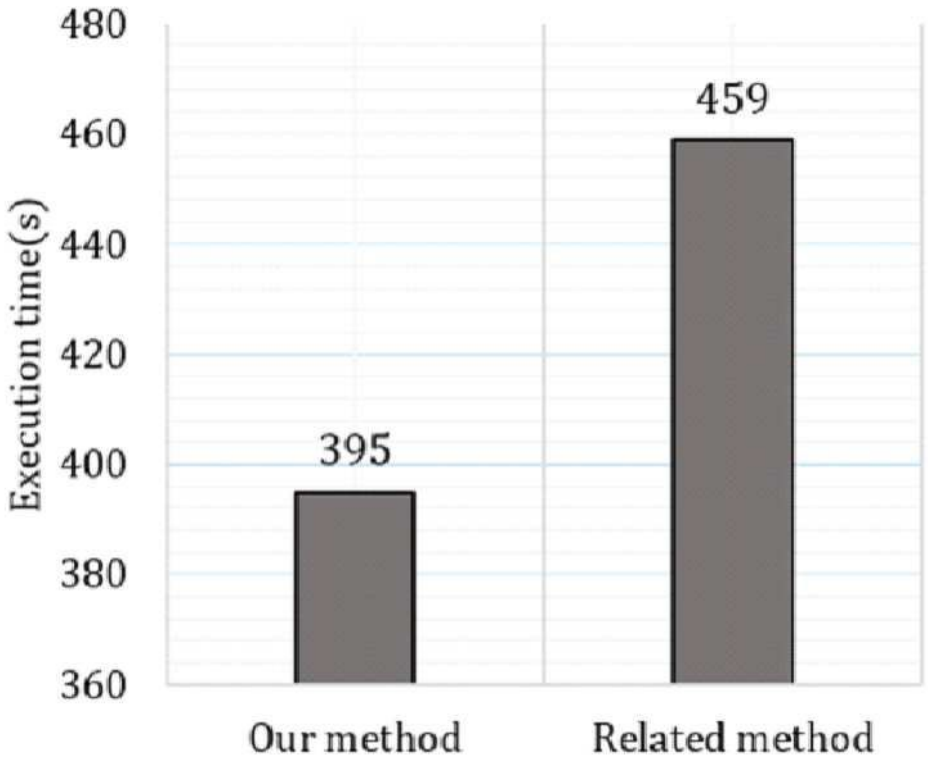
도면7

Iteration	8000	9000	10000	11000	12000	13000	14000	15000	16000	17000	18000
Accuracy	82.34%	86.5%	88.34%	91.34%	94.77%	96.49%	98.25%	98.76%	99.15%	**99.23%**	98.10%
Compression rate	15%	19%	26%	39%	56%	71%	86%	89%	92%	**94%**	95%

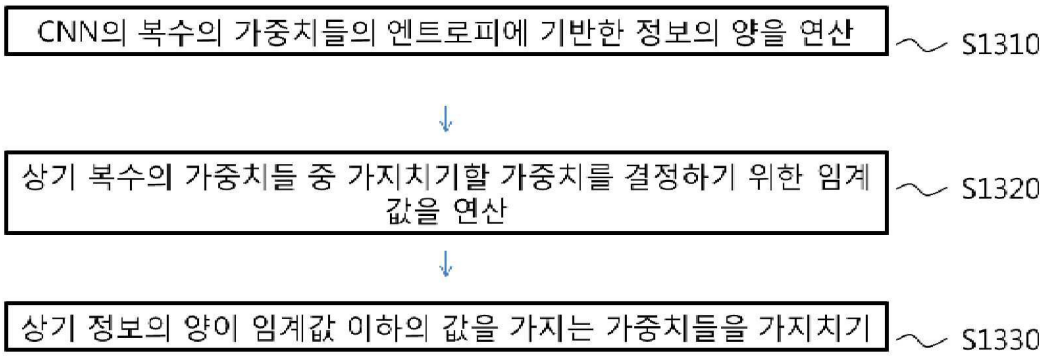
도면8

Layer	The number of weights	The number of pruned weights of related method	The number of pruned weights of our method	Compression improvement
C1	500	170	230	35%
C2	25000	22000	23000	5%
F1	400000	368000	376000	2%
F2	5000	4050	4250	5%

도면9


480
460
440
420
400
380
360
Execution time(s)
395
459
Our method
Related method

도면10


CNN의 복수의 가중치들의 엔트로피에 기반한 정보의 양을 연산
S1310
상기 복수의 가중치들 중 가지치기할 가중치를 결정하기 위한 임계값을 연산
S1320
상기 정보의 양이 임계값 이하의 값을 가지는 가중치들을 가지치기
S1330