(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0085280
(43) 공개일자 2022년06월22일

(51) 국제특허분류(Int. Cl.)
G06T 3/40 (2006.01) *G06N 3/04* (2006.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06T 3/4053 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2020-0175216
(22) 출원일자 2020년12월15일
심사청구일자 2020년12월15일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
배성호
경기도 용인시 기흥구 예현로 15 서그내마을에스케이아파트 102동 501호
황지원
서울특별시 강남구 남부순환로363길 30 도곡쌍용예가아파트 101동 907호
(74) 대리인
특허법인지원

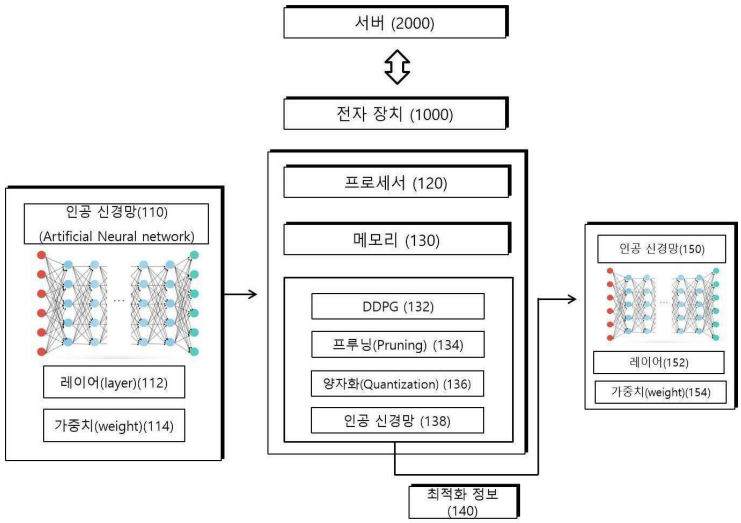
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 **초해상화를 수행하는 인공 신경망의 가중치를 처리하는 방법 및 장치**

(57) 요 약

본 개시는 초해상화를 수행하는 인공 신경망의 가중치를 처리하는 방법 및 이를 수행하는 전자 장치에 관한 것이다. 일 실시 예에 의하면, 초해상화를 수행하는 인공 신경망의 가중치를 처리하는 방법은 상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하는 단계; 상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하는 단계; 상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하는 단계; 및 상기 정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는 단계; 를 포함할 수 있다.

대 표 도



(52) CPC특허분류
G06N 3/082 (2013.01)
G06T 2207/20081 (2013.01)
G06T 2207/20084 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711103234
과제번호 IITP-2020-2015-0-00742-006
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 Grand ICT연구센터지원사업
연구과제명 라이프 컴패니온쉽 경험을 위한 지능형 인터렉션 융합 연구
기 여 율 1/2
과제수행기관명 경희대학교 산학협력단
연구기간 2020.01.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711097853
과제번호 2019-0-01768
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원 IITP
연구사업명 인공지능산업원천기술개발(R&D) 사업
연구과제명 드론에서의 실시간 고정밀 음성 방향 탐지를 위한 딥러닝 기술 연구
기 여 율 1/2
과제수행기관명 성균관대학교 산학협력단
연구기간 2019.07.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

초해상화를 수행하는 인공 신경망(Artificial Neural Network)의 가중치를 처리하는 방법에 있어서,

상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하는 단계;

상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하는 단계;

상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하는 단계; 및

상기 정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는 단계; 를 포함하는, 방법.

청구항 2

제1항에 있어서, 상기 방법은

상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들을 프루닝(Pruning)함으로써, 상기 프루닝된 레이어들의 가중치를 0으로 결정하는 단계; 및

상기 인공 신경망 내 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치를 포함하는 일부 레이어들을 포함하는 인공 신경망의 정확도가 상기 기준 상태의 인공 신경망의 정확도와 유사해지도록, 상기 프루닝된 일부 레이어들 및 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는 단계; 를 포함하는, 방법.

청구항 3

제1항에 있어서, 상기 보존 비율을 결정하는 단계는

상기 인공 신경망의 학습 주기 단위에 관한 에피소드가 진행됨에 따라 상기 보존 비율을 점차적으로 줄이는 단계; 를 포함하는, 방법.

청구항 4

제2항에 있어서, 상기 방법은

소정의 에피소드가 진행되는 동안 무작위로 상기 신경망 내 적어도 일부 레이어들의 양자화 비트를 결정하는 단계;

상기 소정의 에피소드가 진행되는 동안 무작위로 상기 신경망 내 적어도 일부 레이어들을 프루닝하는 단계;

상기 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치가 할당되는 일부 레이어들을 포함하는 인공 신경망의 정확도가 입력되면, 상기 인공 신경망 내 레이어들의 중요도를 출력하는 강화 학습 모델을 이용하여 상기 인공 신경망 내 레이어들의 중요도를 식별하는 단계; 및

상기 식별된 중요도에 기초하여 상기 프루닝된 일부 레이어들 및 상기 일부 레이어들의 양자화 비트를 갱신하는 단계; 를 포함하는, 방법.

청구항 5

제4항에 있어서, 상기 중요도를 식별하는 단계는

상기 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치가 할당되는 일부 레이어들을 포함하는 인공 신경망의 정확도에 더하여, 상기 기준 상태의 인공 신경망의 정확도 및 상기 인공 신경망의 상태 정보를 상기 강화 학습 모델에 더 입력함으로써, 상기 인공 신경망 내 레이어들의 중요도를 식별하는 단계; 를 포

함하는, 방법.

청구항 6

제2항에 있어서, 상기 프루닝된 일부 레이어들 및 상기 양자화 비트가 결정된 상기 일부 레이어들은, 상기 인공 신경망 내 non-linear 매핑 동작을 수행하는 레이어들인 것을 특징으로 하는, 방법.

청구항 7

제2항에 있어서, 상기 인공 신경망은

초해상화(Super-resolution)를 수행하는 딥러닝 기반 모델이고, 이미지가 입력되면 상기 입력된 이미지의 해상도를 변환하는데 사용되는 것을 특징으로 하는, 방법.

청구항 8

제2항에 있어서, 상기 방법은

상기 인공 신경망의 정보들 중 일부의 정보만을 포함하는 프록시 신경망 모델을 생성하는 단계;

상기 생성된 프록시 신경망 모델 내 일부 레이어들을 프루닝하고, 상기 프록시 신경망 모델 내 일부 레이어들에 대한 양자화 비트를 결정하며, 상기 프루닝된 일부 레이어들 및 상기 일부 레이어들에 대한 양자화 비트를 최적화 하는 단계; 및

상기 최적화된 양자화 비트를 가지는 일부 레이어들 및 상기 최적화된 프루닝된 일부 레이어들을 포함하는 프록시 신경망 모델을 상기 인공 신경망에 대응되도록 선형 확장하는 단계; 를 포함하는, 방법.

청구항 9

초해상화를 수행하는 인공 신경망의 가중치를 처리하는 전자 장치에 있어서,

하나 이상의 인스트럭션을 저장하는 메모리; 및

상기 하나 이상의 인스트럭션들을 실행하는 적어도 하나의 프로세서; 를 포함하고,

상기 적어도 하나의 프로세서는 상기 하나 이상의 인스트럭션을 실행함으로써,

상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하고,

상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하고,

상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하고,

상기 정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는, 전자 장치.

청구항 10

제9항에 있어서, 상기 적어도 하나의 프로세서는

상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들을 프루닝(Pruning)함으로써, 상기 프루닝된 레이어들의 가중치를 0으로 결정하고,

상기 인공 신경망 내 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치를 포함하는 일부 레이어들을 포함하는 인공 신경망의 정확도가 상기 기준 상태의 인공 신경망의 정확도와 유사해지도록, 상기 프루닝된 일부 레이어들 및 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는, 전자 장치.

청구항 11

제9항에 있어서, 상기 적어도 하나의 프로세서는

상기 인공 신경망의 학습 주기 단위에 관한 에피소드가 진행됨에 따라 상기 보존 비율을 점차적으로 줄이는 것을 특징으로 하는, 전자 장치.

청구항 12

제10항에 있어서, 상기 적어도 하나의 프로세서는

소정의 에피소드가 진행되는 동안 무작위로 상기 신경망 내 적어도 일부 레이어들의 양자화 비트를 결정하고,

상기 소정의 에피소드가 진행되는 동안 무작위로 상기 신경망 내 적어도 일부 레이어들을 프루닝하고,

상기 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치가 할당되는 일부 레이어들을 포함하는 인공 신경망의 정확도가 입력되면, 상기 인공 신경망 내 레이어들의 중요도를 출력하는 강화 학습 모델을 이용하여 상기 인공 신경망 내 레이어들의 중요도를 식별하고,

상기 식별된 중요도에 기초하여 상기 프루닝된 일부 레이어들 및 상기 일부 레이어들의 양자화 비트를 갱신하는, 전자 장치.

청구항 13

제12항에 있어서, 상기 적어도 하나의 프로세서는

상기 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치가 할당되는 일부 레이어들을 포함하는 인공 신경망의 정확도에 더하여, 상기 기준 상태의 인공 신경망의 정확도 및 상기 인공 신경망의 상태 정보를 상기 강화 학습 모델에 더 입력함으로써, 상기 인공 신경망 내 레이어들의 중요도를 식별하는 단계; 를 포함하는, 전자 장치.

청구항 14

제10항에 있어서, 상기 프루닝된 일부 레이어들 및 상기 양자화 비트가 결정된 상기 일부 레이어들은, 상기 인공 신경망 내 non-linear 매핑 동작을 수행하는 레이어들인 것을 특징으로 하는, 전자 장치.

청구항 15

제10항에 있어서, 상기 인공 신경망은

초해상화(Super-resolution)를 수행하는 딥러닝 기반 모델이고, 이미지가 입력되면 상기 입력된 이미지의 해상도를 변환하는데 사용되는 것을 특징으로 하는, 전자 장치.

청구항 16

제10항에 있어서, 상기 적어도 하나의 프로세서는

상기 인공 신경망의 정보들 중 일부의 정보만을 포함하는 프록시 신경망 모델을 생성하고,

상기 생성된 프록시 신경망 모델 내 일부 레이어들을 프루닝하고, 상기 프록시 신경망 모델 내 일부 레이어들에 대한 양자화 비트를 결정하며, 상기 프루닝된 일부 레이어들 및 상기 일부 레이어들에 대한 양자화 비트를 최적화 하고,

상기 최적화된 양자화 비트를 가지는 일부 레이어들 및 상기 최적화된 프루닝된 일부 레이어들을 포함하는 프록시 신경망 모델을 상기 인공 신경망 에 대응되도록 선형 확장하는, 전자 장치.

청구항 17

초해상화를 수행하는 인공 신경망(Artificial Neural Network)의 가중치를 처리하는 방법에 있어서,

상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하는 단계;

상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하는 단계;

상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하는 단계; 및

상기 정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는 단계; 를 포함하는, 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체.

발명의 설명

기 술 분 야

[0001] 본 개시는 인공 신경망의 가중치를 처리하는 방법 및 장치에 관한 것이다. 보다 상세하게는, 인공 신경망을 경량화 하기 위한 가중치 처리 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 인공 신경망(Artificial Neural Network)는 인공 뉴런들의 상호 연결된 집합들을 구현하기 위하여 컴퓨팅 기기 또는 컴퓨팅 기기에 의해서 수행되는 방법을 지칭할 수 있다. 인공 신경망의 일 실시 예로, 심층 신경망(Deep Neural Network) 또는 딥 러닝(Deep Learning)은 멀티 레이어 구조를 가질 수 있고, 레이어들 각각이 다수의 데이터에 따라 학습될 수 있다.

[0003] 한편 초해상화는 저해상도 이미지로부터 고해상도 이미지를 복원하는 기술을 의미한다. 초해상화에 인공 신경망 기술을 적용하기 위한 예로써 초해상화 딥러닝 모델이 연구되고 있다. 그러나 초해상화 딥러닝 모델 기법들은 층을 깊게 쌓다 보니 모델 크기가 과도하게 증가함으로써 제한된 환경에서 사용하기에는 많은 시간 및 자원이 소요되는 한계가 있다.

[0004] 이러한 과도한 모델의 크기 내지 계산량을 해결하기 위해서 다른 딥러닝 기법에서도 모델을 압축하기 위한 여러 기술들이 개발되고 있다. 일 예로, 이러한 문제를 극복하기 위해 양자화, 가지치기, 증류 기법 등이 활발하게 연구되고 있다.

[0005] 그러나, 초해상화 딥러닝 모델의 특성에 기초하여 초해상화 딥러닝 모델의 크기 문제를 해결하기 위한 기술은 개시되어 있지 않다. 따라서, 초해상화 기법을 수행하는 신경망 모델의 정확도 향상과 함께 효과적으로 모델을 압축하기 위한 기술 개발이 요구되고 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국공개특허 제2020-0115239호

발명의 내용

해결하려는 과제

[0007] 일 실시 예에 따르면, 인공 신경망의 가중치를 처리하는 방법 및 전자 장치가 제공될 수 있다.

[0008] 또한, 일 실시 예에 의하면, 초해상화를 수행하는 인공 신경망의 가중치를 처리하는 방법이 제공될 수 있다.

과제의 해결 수단

[0009] 상술한 기술적 과제를 달성하기 위한 본 개시의 일 실시 예에 따라, 초해상화를 수행하는 인공 신경망의 가중치를 처리하는 방법은 상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하는 단계; 상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하는 단계; 상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하는 단계; 및 상기 정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는 단계; 를 포함할 수 있다.

[0010] 상술한 기술적 과제를 달성하기 위한 또 다른 실시 예에 의하면, 초해상화를 수행하는 인공 신경망의 가중치를 처리하는 전자 장치는 하나 이상의 인스트럭션을 저장하는 메모리; 및 상기 하나 이상의 인스트럭션들을 실행하는 적어도 하나의 프로세서; 를 포함하고, 상기 적어도 하나의 프로세서는 상기 하나 이상의 인스트럭션을 실행함으로써, 상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하고, 상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하고, 상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하고, 상기 정

확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신할 수 있다.

[0011] 상술한 기술적 과제를 달성하기 위한 또 다른 실시 예에 의하면, 초해상화를 수행하는 인공 신경망(Artificial Neural Network)의 가중치를 처리하는 방법에 있어서, 상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하는 단계; 상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하는 단계; 상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하는 단계; 및 상기 정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신하는 단계; 를 포함하는, 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체가 제공될 수 있다.

도면의 간단한 설명

[0012] 도 1은 일 실시 예에 따른 전자 장치가 인공 신경망을 압축하기 위해 인공 신경망 내 가중치를 처리하는 과정을 설명하기 위한 도면이다.

도 2는 일 실시 예에 따른 전자 장치가 인공 신경망의 가중치를 처리하는 방법의 흐름도이다.

도 3은 또 다른 실시 예에 따른 전자 장치가 인공 신경망의 가중치를 처리하는 방법의 흐름도이다.

도 4는 일 실시 예에 따른 전자 장치가 강화 학습 모델을 이용하여 인공 신경망을 압축하는 과정을 설명하기 위한 도면이다.

도 5는 본 개시의 일 실시 예에 따른 방법에 따라 일부 레이어를 프루닝함에 따른 정확도 차이를 설명하기 위한 도면이다.

도 6은 본 개시에 따른 인공 신경망 가중치를 처리하는 방법이 적용된 신경망 모델의 성능을 보존 비율에 따라 설명하기 위한 도면이다.

도 7은 본 개시에 따른 인공 신경망 가중치를 처리하는 방법이 적용된 신경망 모델의 성능을 모델 사이즈에 따라 설명하기 위한 도면이다.

도 8은 일 실시 예에 따른 전자 장치의 블록도이다.

도 9는 일 실시 예에 따른 서버의 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 개시에 대해 구체적으로 설명하기로 한다.

[0014] 본 개시에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0015] 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0016] 아래에서는 첨부한 도면을 참고하여 본 개시의 실시예에 대하여 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 개시는 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 개시를 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0017] 도 1은 일 실시 예에 따른 전자 장치가 인공 신경망을 압축하기 위해 인공 신경망 내 가중치를 처리하는 과정을 설명하기 위한 도면이다.

[0018] 일 실시 예에 의하면, 전자 장치(1000)는 인공 신경망(Artificial Neural Network)의 가중치를 처리할 수 있다. 예를 들어, 인공 신경망은 복수의 신경망 레이어들로 구성될 수 있다. 복수의 신경망 레이어들 각각은 복수의 가중치들(weight values, weights)을 갖고 있으며, 이전(previous) 레이어의 연산 결과와 복수의 가중치들 간의 연산을 통해 신경망 연산을 수행한다. 복수의 신경망 레이어들이 갖고 있는 복수의 가중치들은 인공 신경망의 학습 결과에 의해 최적화될 수 있다. 또한, 학습 과정 동안 인공지능 모델에서 획득한 손실(loss) 값 또는 코스트(cost) 값이 감소 또는 최소화되도록 복수의 가중치들이 수정 및 갱신될 수 있다.

[0019] 일 실시 예에 의하면, 전자 장치(1000)는 인공 신경망(110)을 획득하고, 획득된 인공 신경망(100)내 레이어들(112) 및 상기 레이어들의 연결 강도에 관한 가중치(114)를 처리함으로써, 인공 신경망(110)의 모델의 크기가 압축할 수 있다. 전자 장치(1000)는 인공 신경망(110)의 모델의 크기를 압축함으로써 압축된 크기의 레이어들(152) 및 상기 레이어들의 연결 강도에 관한 가중치(154)를 포함하는 인공 신경망(150)을 생성할 수 있다.

[0020] 일 실시 예에 의하면, 전자 장치(1000)가 처리하는 인공 신경망은 초해상화 딥러닝 모델일 수 있다. 또한, 일 실시 예에 의하면 인공 신경망은 심층 신경망(DNN:Deep Neural Network)를 포함할 수 있으며, 예를 들어, CNN (Convolutional Neural Network), DNN (Deep Neural Network), RNN (Recurrent Neural Network), RBM (Restricted Boltzmann Machine), DBN (Deep Belief Network), BRDNN(Bidirectional Recurrent Deep Neural Network) 또는 심층 Q-네트워크 (Deep Q-Networks) 등이 있으나, 전술한 예에 한정되지 않는다. 이하에서는 편의상 본 개시에 따른 인공 신경망은 심층 신경망인 경우를 예로 설명하기로 한다.

[0021] 일 실시 예에 의하면, 전자 장치(1000)는 프로세서(120) 및 메모리(130)를 포함할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)의 프로세서(120)는 메모리(130)에 저장된 하나 이상의 인스트럭션을 수행함으로써, 인공 신경망 내 가중치를 처리하기 위한 다양한 과정을 수행할 수 있다. 일 실시 예에 의하면, 메모리(130)에는 DDPG(Deep Deterministic Policy Gradient, 132) 알고리즘을 수행하기 위한 인스트럭션, 인공 신경망 내 소정의 레이어들을 가지치기함으로써 레이어의 가중치들을 0으로 설정하기 위한 프루닝(pruning, 134) 인스트럭션, 인공 신경망 내 레이어들을 양자화하기 위한 양자화(Quantization, 136) 인스트럭션, 압축 대상이 되는 인공 신경망 또는 압축된 인공 신경망을 포함하는 신경망(138)에 대한 정보들이 저장될 수 있다.

[0022] 일 실시 예에 의하면, 전자 장치(1000)내 프로세서(120)는 메모리(130)내 하나 이상의 인스트럭션을 수행함으로써, 인공 신경망(110)의 모델의 크기를 압축하기 위한 최적화 정보(140)를 결정하고, 결정된 최적화 정보에 기초하여 인공 신경망(110)을 압축함으로써 인공 신경망(150)을 생성할 수 있다. 일 실시 예에 의하면, 최적화 정보(140)는 인공 신경망에서 가지치기(예컨대 프루닝) 대상이 되는 레이어들의 정보 또는 인공 신경망 내 레이어들의 가중치를 양자화 하기 위해 레이어들 별로 설정되는 양자화 비트 정보 중 적어도 하나를 포함할 수 있다.

[0023] 또한, 일 실시 예에 의하면, 전자 장치(1000)는 인공 신경망 내 초반의 feature 추출 및 후반의 reconstruction을 수행하기 위한 레이어들이 아닌, 중간 부분의 non-linear 매핑을 수행하는 레이어들 및 상기 레이어들의 가중치를 상기 최적화 정보에 기초하여 처리할 수 있다. 본 개시에 따른 전자 장치(1000)는 상술한 방법에 따라 초해상화 딥러닝 모델의 특성을 적응적 양자화 기법에 반영함으로써 가지치기 여부까지 한번에 수행할 수 있으며 신경망 압축 최적화를 위해 신경망 내 레이어들에 대한 보존 비율을 조절할 수도 있다.

[0024] 도 2는 일 실시 예에 따른 전자 장치가 인공 신경망의 가중치를 처리하는 방법의 흐름도이다.

[0025] S210에서, 전자 장치(1000)는 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화 하기 위한 보존 비율을 결정할 수 있다. 일 실시 예에 의하면, 보존 비율은 신경망 내 전체 가중치 중, 양자화 이후 남아있는 가중치의 비율을 의미할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 인공 신경망의 학습 주기 단위에 관한 에피소드가 진행됨에 따라 보존 비율을 점차적으로 줄일 수 있다. 본 개시에 따른 전자 장치(1000)는 최적화 양자화 비트 검색 및 가지치기 여부를 판단하기 위해, 한번에 보존 비율을 낮추기 보다는, 에피소드가 진행됨에 따라 점차적으로 보존 비율을 줄여가면서, 최적화 비트 및 가지치기 여부를 탐색할 수 있다.

[0026] 일 실시 예에 의하면 보존 비율은 인공 신경망 내 전체 레이어에 대한 가중치들 중, 양자화된 가중치의 비율을 의미할 수 있다. 전자 장치(1000)가 인공 신경망 내 레이어들에 양자화 비트를 결정하고, 소정의 레이어들을 프루닝한 후, 인공 신경망 내 전체 레이어들의 가중치들 중 양자화 과정을 통해 남은 가중치의 비율은 보존 비율에 매칭될 수 있다.

[0027] S220에서, 전자 장치(1000)는 보존 비율에 기초하여 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정할 수 있다. 예를 들어, 전자 장치(1000)는 인공 신경망 내 레이어들의 양자화 비트를 결정한 후, 양자

화 비트에 따라 양자화된 가중치들을 이용하여 인공 신경망을 학습시킬 수 있다. 또 다른 실시 예에 의하면, 전자 장치(1000)는 인공 신경망 내 레이어들에 대한 양자화 비트를 결정함과 함께 프루닝(pruning) 동작을 수행할 수 있다. 즉, 전자 장치(1000)는 보존 비율에 기초하여 인공 신경망 내 일부 레이어에 대해서는 양자화 비트를 설정한 후, 양자화 비트에 따라 가중치를 양자화하고, 인공 신경망 내 일부 레이어들은 가중치를 0으로 설정함으로써 가지치기 과정을 수행할 수 있다.

[0028] S230에서, 전자 장치(1000)는 S230에서 결정된 양자화 비트에 기초하여 양자화된 가중치에 따라 인공 신경망을 학습 시킨 후, 학습된 인공 신경망의 정확도를 획득할 수 있다. S240에서, 전자 장치(1000)는 인공 신경망의 정확도 및 기준 상태의 인공 신경망의 정확도의 차이를 비교하고, 차이가 작아지도록 인공 신경망 내 레이어들의 양자화 비트를 갱신할 수 있다. 일 실시 예에 의하면, 기준 상태의 인공 신경망의 정확도는 양자화되지 않은 가중치에 따라 학습된 인공 신경망의 정확도를 의미할 수 있다.

[0029] 도 3은 또 다른 실시 예에 따른 전자 장치가 인공 신경망의 가중치를 처리하는 방법의 흐름도이다.

[0030] S310에서, 전자 장치(1000)는 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화 하기 위한 보존 비율을 결정할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 소정의 에피소드 동안 인공 신경망 내 전체 레이어들의 보존 비율이 설정된 보존 비율 스케줄링 정보를 생성할 수도 있다.

[0031] S320에서, 전자 장치(1000)는 보존 비율에 기초하여, 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정할 수 있다. 또한, S330에서, 전자 장치(1000)는 보존 비율에 기초하여 인공 신경망 내 레이어들 내 일부 레이어들을 프루닝함으로써, 프루닝된 레이어들의 가중치를 0으로 결정할 수 있다. 예를 들어, 전자 장치(1000)가 인공 신경망 내 레이어들에 대해 결정된 양자화 비트에 따라 레이어들을 양자화하고, 또 다른 레이어들을 가지치기한 결과, 인공 신경망 내 레이어들의 전체 가중치들 중 양자화되지 않고 남은 가중치의 비율은 보존 비율에 매칭될 수 있다.

[0032] S340에서, 전자 장치(1000)는 프루닝 과정 및 양자화 비트 설정 과정이 완료되면, 프루닝 과정 및 양자화 비트 설정 과정이 완료된 인공 신경망의 정확도를 기준 상태의 인공 신경망의 정확도와 비교함으로써 정확도 차이를 식별할 수 있다. 전자 장치(1000)는 프루닝 과정 및 양자화 비트 설정 과정이 완료된 인공 신경망의 정확도의 차이가 작아 지도록 프루닝 과정 및 양자화 비트 설정 과정을 다시 수행할 수 있다. 일 시릿 예에 의하면, 프루닝 과정 및 양자화 비트 설정 과정은 정확도의 차이가 소정의 임계치 보다 작아질 때까지 계속될 수 있다.

[0033] 보다 상세하게는, 전자 장치(1000)는 인공 신경망 내 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치를 포함하는 일부 레이어들을 포함하는 인공 신경망의 정확도가 기준 상태의 인공 신경망의 정확도와 유사해지도록, 프루닝된 일부 레이어들 및 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신할 수 있다.

[0034] 일 실시 예에 의하면, 전자 장치(1000)는 강화 학습 기반의 액터 및 크리틱 모델을 포함하는 DDPG(Deep Deterministic Policy Gradient) 모델을 이용하여, 인공 신경망 내 레이어들의 중요도를 결정하고, 결정된 중요도에 기초하여 소정의 레이어들을 프루닝하거나, 소정의 레이어들에 양자화 비트를 설정할 수 있다. 전자 장치(1000)는 DDPG 모델을 이용하여, 가지치기 내지 양자화 처리된 인공 신경망의 정확도와 기준 상태의 인공 신경망의 정확도 차이가 소정의 임계치보다 작아 질 때까지, S330 내지 S340 과정을 반복하여 수행함으로써, 인공 신경망을 압축하기 위한 최적화 정보(프루닝 레이어 및 양자화 비트 정보)를 결정할 수 있다. S350에서, 전자 장치(100) 최적화 정보를 출력할 수 있다. 예를 들어, 전자 장치(1000)는 최적화된 양자화 비트 및 프루닝 레이어에 대한 정보를 출력할 수 있다.

[0035] 도 4는 일 실시 예에 따른 전자 장치가 강화 학습 모델을 이용하여 인공 신경망을 압축하는 과정을 설명하기 위한 도면이다.

[0036] 도 4를 참조하여 전자 장치(1000)가 DDPG 모델(412)을 이용하여 인공 신경망의 모델의 크기를 효과적으로 압축하는 과정을 설명하기로 한다. 일 실시 예에 의하면 전자 장치(1000)는 인공 신경망의 학습 주기 단위에 관한 소정의 에피 소드(404) 별로 보존 비율(402)을 결정할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 인공 신경망 내 전체 레이어들의 가중치들 중 양자화 과정에 따라 남아있는 가중치의 비율인 보존 비율을 조절함으로써, 인공 신경망 내 레이어들의 가중치를 2비트(406) 또는 1비트(408)로 양자화할 수 있다.

[0037] 일 실시 예에 의하면, 전자 장치(1000)는 초기 에피소드 동안(예컨대 약 30 에피소드) 레이어들의 중요도가 결정되지 않은 상태에서는 무작위로 인공 신경망 내 적어도 일부 레이어들의 양자화 비트를 결정할 수 있다. 예를 들어, 전자 장치(1000)는 초기 에피소드가 진행되는 동안 인공 신경망 내 레이어들의 일부를 프루닝(예컨대 일

부 레이어에 대한 양자화 비트를 0으로 설정하는 동작)하거나, 일부 레이어들에 대한 양자화 비트를 1 내지 8로 설정할 수 있다.

[0038] 도 4에 도시된 바와 같이, 전자 장치(1000)는 에피 소드 초기 과정에서 너무 낮은 보존 비율(예컨대 너무 높은 압축률)로 인하여 최적의 양자화 비트를 찾지 못하는 문제를 방지하기 위해, 에피소드가 진행됨에 따라 점차적으로 보존 비율을 낮출 수 있다. 본 개시에 따른 전자 장치(1000)는 에피소드가 진행됨에 따라 점진적으로 보존 비율을 줄여가면서, 인공 신경망 모델내 레이어들의 중요도에 따라 최적의 양자화 비트를 설정함으로써, 높은 정확도를 달성함과 함께 인공 신경망을 효과적으로 압축시킬 수 있다.

[0039] 일 실시 예에 의하면, 전자 장치(1000)는 에피소드 별로 결정된 보존 비율(410)에 대한 정보를 DDPG 모델(412)로 전달할 수 있다. DDPG 모델은 강화 학습 모델로써, 액터 모델(414) 및 크리틱 모델(416)을 포함할 수 있다. 일 실시 예에 의하면, DDPG 모델(412)은 인공 신경망 내 레이어들의 중요도 정보를 결정할 수 있다. 예를 들어, DDPG 모델(412)은 중요도 정보에 기초하여 인공 신경망 내 레이어들에 대한 양자화 비트를 설정하고, 동시에 양자화 비트를 0으로 설정함으로써 가지치기 대상인 레이어들을 결정할 수도 있다.

[0040] 일 실시 예에 의하면, 강화 학습 모델인 DDPG 모델(412)은 프루닝된 일부 레이어들 및 양자화 비트에 따라 양자화된 가중치에 기초하여 학습된 인공 신경망의 정확도가 입력되면, 인공 신경망 내 레이어들의 중요도 정보를 출력할 수 있다. 또 다른 실시 예에 의하면, 강화 학습 모델인 DDPG 모델(412)은 소정의 양자화 처리 과정 및 프루닝 과정에 따라 학습된 인공 신경망의 정확도, 기준 상태의 인공 신경망의 정확도, 인공 신경망의 상태 정보가 입력되면, 해당 인공 신경망 내 레이어들의 중요도 정보를 출력하는 모델일 수 있다.

[0041] 일 실시 예에 의하면, 본 개시에 따른 DDPG 모델(412)은 레이어들의 중요도 정보, 레이어들에 대한 양자화 비트 정보, 양자화 비트가 0인 레이어들에 대한 정보를 액션 정보(418)로 하여 출력할 수 있다.

[0042] 전자 장치(1000)는 DDPG 모델에서 출력된 액션 정보(418)를 이용하여 인공 신경망(422)내 레이어들에 대해 양자화 비트를 설정할 수 있다. 예를 들어, 전자 장치(1000)는 액션 정보(418)내 레이어들 별 양자화 비트 정보 또는 양자화 비트가 0으로 설정되는 레이어 정보를 이용하여, 인공 신경망(422) 내 레이어들 중 일부 레이어에 대하여 양자화 비트를 설정하거나, 일부 레이어에 대한 양자화 비트를 0으로 설정함으로써 가지치기 대상 레이어를 설정할 수 있다.

[0043] 상술한 바와 같이, 전자 장치(1000)가 가중치를 처리하는 대상인 인공 신경망은 초해상화 딥러닝 모델로써, 저해상도 이미지(424)가 획득되면 고해상도의 이미지(426)를 출력하는 신경망 모델일 수 있다. 전자 장치(1000)는 DDPG 모델(412)로부터 출력된 액션 정보에 기초하여 인공 신경망(422)내 레이어들에 대해 양자화 비트 정보를 결정하고, 결정된 양자화 비트 정보에 따라 양자화된 가중치들에 기초하여 인공 신경망(422)을 학습시킬 수 있다.

[0044] 일 실시 예에 의하면, 전자 장치(1000)가 인공 신경망 내 레이어들에 대한 양자화 비트를 설정하는 동작은, 인공 신경망 내 레이어들에 대한 양자화 비트(432)를 0 내지 8로 설정하는 동작을 포함할 수 있다. 즉, 일 실시 예에 의하면, 전자 장치(1000)가 인공 신경망 내 레이어들에 대한 양자화 비트를 설정하는 동작은 인공 신경망 내 레이어들을 가지치기 하는 동작을 포함할 수 있다.

[0045] 전자 장치(1000)는 학습된 인공 신경망의 정확도 및 기준 상태의 인공 신경망의 정확도의 차이를 리워드 정보(420)로써 DDPG 모델(412)에 전달할 수 있다. 일 실시 예에 의하면, 리워드 정보(420)는 전자 장치(1000)가 현재 학습 시킨 인공 신경망의 상태에 관한 상태 정보를 더 포함할 수도 있다. 일 실시 예에 의하면, 상태 정보는 현재 인공 신경망 내 레이어들 정보, 레이어들 각각에 대해 설정된 양자화 비트 정보, 프루닝된 레이어들 정보 중 적어도 하나를 포함할 수 있다.

[0046] DDPG 모델(412)은 리워드 정보(420)에 기초하여, 인공 신경망 내 레이어들의 중요도를 다시 결정할 수 있다. 예를 들어, DDPG 모델(412)은 리워드 정보에(420)에 기초하여 인공 신경망 내 레이어들의 중요도를 다시 결정하고, 결정된 중요도에 기초하여 인공 신경망 내 레이어들에 대한 양자화 비트를 다시 설정할 수 있다. DDPG 모델(412)은 리워드 정보(420)에 따라 인공 신경망 내 레이어들의 중요도를 갱신하고, 갱신된 중요도에 따라 인공 신경망 내 레이어들의 양자화 비트를 갱신하는 동작을 반복함으로써, 인공 신경망을 효과적으로 압축하기 위한 중요도 정보를 결정할 수 있다.

[0047] 도 4에서 도시된, 전자 장치(1000)가 프루닝 하거나 양자화 비트를 설정하는 레이어들은 인공 신경망 내 non-linear 매핑 동작을 수행하는 레이어들일 수 있다. 또한, 일 실시 예에 의하면, 전자 장치(1000)는 복잡도가 높은 신경망 모델을 그대로 사용함에 있어 최적의 양자화 비트를 찾거나 미세 조정 과정에 많은 연산이 소요되는

경우, 경량화 대상 신경망 모델에 대한 프록시 신경망 모델을 생성하고, 생성된 프록시 신경망 모델에 대하여도 4에서 상술한 과정을 수행할 수 있다. 일 실시 예에 의하면 프록시 신경망 모델은 인공 신경망 모델 내 일부 레이어들에 대한 정보가 선형적으로 압축된 신경망 모델일 수 있다.

[0048] 예를 들어, 전자 장치(1000)는 프록시 신경망 모델에 대하여 보존 비율을 결정하고, 결정된 보존 비율에 기초하여 프록시 신경망 모델 내 일부 레이어들을 프루닝하거나, 프록시 신경망 모델 내 일부 레이어들에 대한 양자화 비트를 설정할 수 있다. 전자 장치(1000)는 프루닝 과정 및 양자화 비트 설정 과정이 완료된 신경망 모델을 학습시킨 후, 강화 학습 모델인 DDPG 모델을 이용하여 결정되는 리워드 정보에 기초하여, 프루닝 과정 및 양자화 비트 설정 과정을 반복할 수 있다. 전자 장치(1000)는 프루닝 과정 및 양자화 비트 설정 과정이 완료된 프록시 신경망 모델의 정확도 및 기준 상태의 인공 신경망의 정확도의 차이가 소정의 임계치 이하가 될 때까지 프루닝된 레이어들 및 일부 레이어들에 대한 양자화 비트를 최적화할 수 있다.

[0049] 전자 장치(1000)는 프루닝 과정 및 양자화 비트 설정 과정이 완료된 프록시 신경망 모델의 정확도 및 기준 상태의 인공 신경망의 정확도의 차이가 소정의 임계치 이하가 되면, 프록시 신경망 모델을 초기 인공 신경망 모델에 대응되도록 선형(Linear) 확장함으로써 최종 압축된 신경망 모델을 생성할 수 있다.

[0050] 도 5는 본 개시의 일 실시 예에 따른 방법에 따라 일부 레이어를 프루닝함에 따른 정확도 차이를 설명하기 위한 도면이다.

[0051] 일 실시 예에 의하면, 도 5를 참조하면 단일층 가지치기 시 이미지 분류 모델(510)의 성능과 초해상화 모델(520)의 성능이 도시된다. 이미지 분류 모델 및 초해상화 모델의 성능 차트를 참조하면, 가로축에 기재된 가지치기 대상이된 레이어들 번호에 따라 변하는 각 신경망 모델의 성능 차이가 세로축으로 도시된다. 가지치기가 되는 층에 따른 정확도가 불규칙하게 변하는 이미지 분류 모델(510)과 달리 본 개시에 따른 초해상화 신경망 모델의 경우, 가지치기가 되는 층에 따른 정확도가 상대적으로 균일하게 나타남을 알 수 있다.

[0052] 즉, 본 개시에 따른 초해상화 기법을 수행하는 초해상화 인경 신경망 모델은 초반의 feature-extraction, 후반의 reconstruction 부분과 달리 중간의 non-linear mapping 부분에서는 많은 잔차 연결(residual connection)로 인해 하나의 층을 가지치기하여도 성능 하락이 크지 않은 것을 볼 수 있다. 이는 초해상화 모델이 일반적으로 예측에 있어서 단계별 중요도의 차이가 확실하며, 특정 단계(non-linear mapping)의 단일 층을 통째로 가지치기했을 때 성능 하락을 최소화하며 압축률을 극대화할 수 있음을 의미한다. 따라서, 본 개시에 따른 전자 장치(1000)는 강화학습을 사용한 적응적 층간 양자화 기법 및 층별 가지치기 방법을 하나의 탐색공간에 넣음으로써 신경망 모델을 효과적으로 압축할 수 있다.

[0053] 도 6은 본 개시에 따른 인공 신경망 가중치를 처리하는 방법이 적용된 신경망 모델의 성능을 보존 비율에 따라 설명하기 위한 도면이다.

[0054] 도 6을 참조하면 본 개시에 따른 방법이 적용된 신경망 모델(602)들의 보존 비율 변화(604)에 따른 성능 변화(PESNR/SSIM)가 도시된다. 예를 들어, 각 신경망 모델 MSRN, DDBPN 및 EDSR 모델에 대한 보존 비율이 100%인 경우 SET 5에서 PESNR/SSIM은 32.26/0.8960, 32.25/0.8957, 32.46/0.8968로 나타날 수 있다. 보존 비율이 0.1로 감소함에 따라 각 신경망 모델 MSRN, DDBPN 및 EDSR 모델의 정확도는 32.17/0.8947, 32.21/0.8947, 32.22/0.8948로 미세하게 낮아지는 것을 관측할 수 있다. 즉 본 개시에 따른 방법이 적용된 신경망 모델들은 보존 비율이 낮아짐에 따라 그 정확도의 하락 폭이 크지 않은 효과가 있다.

[0055] 도 7은 본 개시에 따른 인공 신경망 가중치를 처리하는 방법이 적용된 신경망 모델의 성능을 모델 사이즈에 따라 설명하기 위한 도면이다.

[0056] 도 7을 참조하면 기존의 신경망 모델들(702, 704, 706, 708)과 비교되는 본 개시에 따른 방법이 적용된 신경망 모델(722, 724, 726)들의 모델 크기(712) 변화에 따른 성능 변화 (PESNR/SSIM)가 도시된다. 예를 들어, SR Resinet 모델이 사용된 본원 신경망 모델과 기존의 연구에서 진행된 Binart SR 신경망 모델들(702, 704)의 성능을 비교하면, 각각의 full precision과 비교해서, 높은 압축률에 비해 성능 하락이 더 적은 것을 볼 수 있다.

[0057] 도 8은 일 실시 예에 따른 전자 장치의 블록도이다.

[0058] 일 실시 예에 의하면, 전자 장치(1000)는 프로세서(1300) 및 메모리(1700)를 포함할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 인공 신경망의 가중치를 처리하기 위한, AI 프로그램이 탑재되고 음성 인식 기능을 포함하는 스마트폰, 태블릿 PC, PC, 스마트 TV, 휴대폰, 미디어 플레이어, 서버, 마이크로 서버, 기타 모바일 또는 비모바일 컴퓨팅 장치일 수 있으나, 이에 제한되지 않는다.

[0059] 프로세서(1300)는, 통상적으로 전자 장치(1000)의 전반적인 동작을 제어한다. 일 실시 예에 의하면, 본 개시에 따른 프로세서(1300)는 메모리(1700)에 저장된 프로그램들을 실행함으로써, 도 1 내지 도 7에 기재된 전자 장치(1000)의 기능을 수행할 수 있다. 또한, 프로세서(1300)는 하나 또는 복수의 프로세서로 구성될 수 있고, 하나 또는 복수의 프로세서는 CPU, AP, DSP(Digital Signal Processor) 등과 같은 범용 프로세서, GPU와 같은 그래픽 전용 프로세서일 수 있다. 일 실시 예에 의하면, 프로세서(1300)가 범용 프로세서, 인공지능 프로세서 및 그래픽 전용 프로세서를 포함하는 경우, 인공지능 프로세서는 범용 프로세서 또는 그래픽 전용 프로세서와 별도의 칩으로 구현될 수도 있다.

[0060] 일 실시 예에 의하면, 프로세서(1300)가 복수의 프로세서 또는 그래픽 전용 프로세서로 구현될 때, 복수의 프로세서 또는 그래픽 전용 프로세서 중 적어도 일부는 전자 장치(1000) 및 전자 장치(1000)와 연결된 다른 전자 장치 또는 서버에 탑재될 수도 있다.

[0061] 예를 들어, 프로세서(1300)는, 메모리(1402)에 저장된 하나 이상의 인스트럭션들을 실행함으로써, 상기 인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정하고, 상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정하고, 상기 결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득하고, 상기 정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신할 수 있다.

[0062] 또한, 일 실시 예에 의하면 프로세서(1300)는 상기 보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들을 프루닝(Pruning)함으로써, 상기 프루닝된 레이어들의 가중치를 0으로 결정하고, 상기 인공 신경망 내 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치를 포함하는 일부 레이어들을 포함하는 인공 신경망의 정확도가 상기 기준 상태의 인공 신경망의 정확도와 유사해지도록, 상기 프루닝된 일부 레이어들 및 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신할 수 있다.

[0063] 또한, 프로세서(1300)는 상기 인공 신경망의 학습 주기 단위에 관한 에피소드가 진행됨에 따라 상기 보존 비율을 점차적으로 줄일 수 있다.

[0064] 또한, 프로세서(1300)는 소정의 에피소드가 진행되는 동안 무작위로 상기 신경망 내 적어도 일부 레이어들의 양자화 비트를 결정하고, 상기 소정의 에피소드가 진행되는 동안 무작위로 상기 신경망 내 적어도 일부 레이어들을 프루닝하고, 상기 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치가 할당되는 일부 레이어들을 포함하는 인공 신경망의 정확도가 입력되면, 상기 인공 신경망 내 레이어들의 중요도를 출력하는 강화 학습 모델을 이용하여 상기 인공 신경망 내 레이어들의 중요도를 식별하고, 상기 식별된 중요도에 기초하여 상기 프루닝된 일부 레이어들 및 상기 일부 레이어들의 양자화 비트를 갱신할 수 있다.

[0065] 일 실시 예에 의하면, 프로세서(1300)는 상기 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치가 할당되는 일부 레이어들을 포함하는 인공 신경망의 정확도에 더하여, 상기 기준 상태의 인공 신경망의 정확도 및 상기 인공 신경망의 상태 정보를 상기 강화 학습 모델에 더 입력함으로써, 상기 인공 신경망 내 레이어들의 중요도를 식별할 수 있다.

[0066] 일 실시 예에 의하면, 프로세서(1300)가 처리하는 인공 신경망 모델은 초해상화(Super-resolution)를 수행하는 딥러닝 기반 모델이고, 이미지가 입력되면 상기 입력된 이미지의 해상도를 변환하는데 사용될 수 있다.

[0067] 일 실시 예에 의하면, 적어도 하나의 프로세서(1300)는 상기 인공 신경망의 정보들 중 일부의 정보만을 포함하는 프록시 신경망 모델을 생성하고, 상기 생성된 프록시 신경망 모델 내 일부 레이어들을 프루닝하고, 상기 프록시 신경망 모델 내 일부 레이어들에 대한 양자화 비트를 결정하며, 상기 프루닝된 일부 레이어들 및 상기 일부 레이어들에 대한 양자화 비트를 최적화 하고, 상기 최적화된 양자화 비트를 가지는 일부 레이어들 및 상기 최적화된 프루닝된 일부 레이어들을 포함하는 프록시 신경망 모델을 상기 인공 신경망 모델에 대응되도록 선형 확장함으로써 최종 압축된 신경망 모델을 생성할 수 있다.

[0068] 일 실시 예에 의하면 전자 장치(1000)는 네트워크 인터페이스(미도시)를 더 포함할 수도 있다. 네트워크 인터페이스는 전자 장치(1000)가 다른 장치(미도시) 및 서버(2000)와 통신을 하게 하는 하나 이상의 구성요소를 포함할 수 있다. 다른 장치(미도시)는 전자 장치(1000)와 같은 컴퓨팅 장치이거나, 센싱 장치일 수 있으나, 이에 제한되지 않는다. 예를 들어, 통신부(미도시)는, 근거리 통신부, 이동 통신부를 포함할 수 있다.

[0069] 메모리(1700)는, 프로세서(1300)의 처리 및 제어를 위한 프로그램을 저장할 수 있고, 전자 장치(1000)로 입력되

거나 전자 장치(1000)로부터 출력되는 데이터를 저장할 수도 있다. 또한, 메모리(1700)는 인공 신경망을 구성하는 레이어들, 레이어들에 포함된 노드들 및 레이어들의 연결 강도에 관한 가중치들에 대한 정보와 가중치 정규화 함수, 양자화 함수, 식별된 가중치의 타입 및 식별된 가중치의 타입에 따라 부호화된 가중치 값들을 저장할 수 있다. 즉, 메모리(1700)는 인공 신경망 내 가중치들이 수정 및 갱신될 경우, 수정 및 갱신된 가중치에 관한 정보를 더 저장할 수 있다.

[0070] 또한, 메모리(1700)는 인공 신경망 모델에 대한 정보, DDPG 모델에 대한 정보, DDPG 모델로부터 출력된 최적화 정보(레이어 별 양자화 비트 정보, 프루닝 대상 레이어 정보), 프록시 신경망 모델 정보, 최적화 압축된 신경망 모델에 대한 정보를 더 저장할 수도 있다.

[0071] 메모리(1700)는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(RAM, Random Access Memory) SRAM(Static Random Access Memory), 롬(ROM, Read-Only Memory), EEPROM(Electrically Erasable Programmable Read-Only Memory), PROM(Programmable Read-Only Memory), 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다.

[0072] 도 9는 일 실시 예에 따른 서버의 블록도이다.

[0073] 일 실시 예에 따르면, 서버(2000)는 네트워크 인터페이스(2500), 데이터 베이스(Data Base, 2700) 및 프로세서(2300)를 포함할 수 있다. 네트워크 인터페이스(2500)는 상술한 전자 장치(1000)의 네트워크 인터페이스(미도시)에 대응될 수 있다. 예를 들어, 네트워크 인터페이스(2500)는 전자 장치(1000)로부터 인공 신경망의 레이어들 및 레이어들에 포함된 노드에 관한 정보 또는 신경망 내 레이어들의 연결 강도에 관한 가중치 값들을 수신할 수 있다.

[0074] 데이터 베이스(2700)는 도 8에 도시된 전자 장치의 메모리(1700)에 대응될 수 있다. 예를 들어, 데이터 베이스(2700)는 프로세서(2300)의 처리 및 제어를 위한 프로그램을 저장할 수 있고, 전자 장치(1000)로 입력되거나 전자 장치(1000)로부터 출력되는 데이터를 저장할 수도 있다. 또한, 데이터 베이스(2700)는 인공 신경망을 구성하는 레이어들, 레이어들에 포함된 노드들 및 레이어들의 연결 강도에 관한 가중치들에 대한 정보, 레이어들 별 양자화 비트 정보, 프루닝 대상 레이어들 정보 등을 더 저장할 수도 있다. 또한, 데이터 베이스(2700)는 인공 신경망 모델을 압축하기 위한 최적화 정보들이 수정 및 갱신되는 경우 수정된 최적화 정보들을 더 저장할 수 있다. 최적화 정보들을 네트워크 인터페이스(2500)를 통하여 전자 장치(1000)로 전송될 수 있다.

[0075] 프로세서(2300)는 통상적으로 서버(2000)의 전반적인 동작을 제어한다. 예를 들어, 프로세서(2300)는, 서버(2000)의 DB(2700)에 저장된 프로그램들을 실행함으로써, DB(2700) 및 네트워크 인터페이스(2500) 등을 전반적으로 제어할 수 있다. 또한, 프로세서(2300)는 DB(2700)에 저장된 프로그램들을 실행함으로써, 도 1 내지 도 8에서의 전자 장치(1000)의 동작의 일부를 수행할 수 있다.

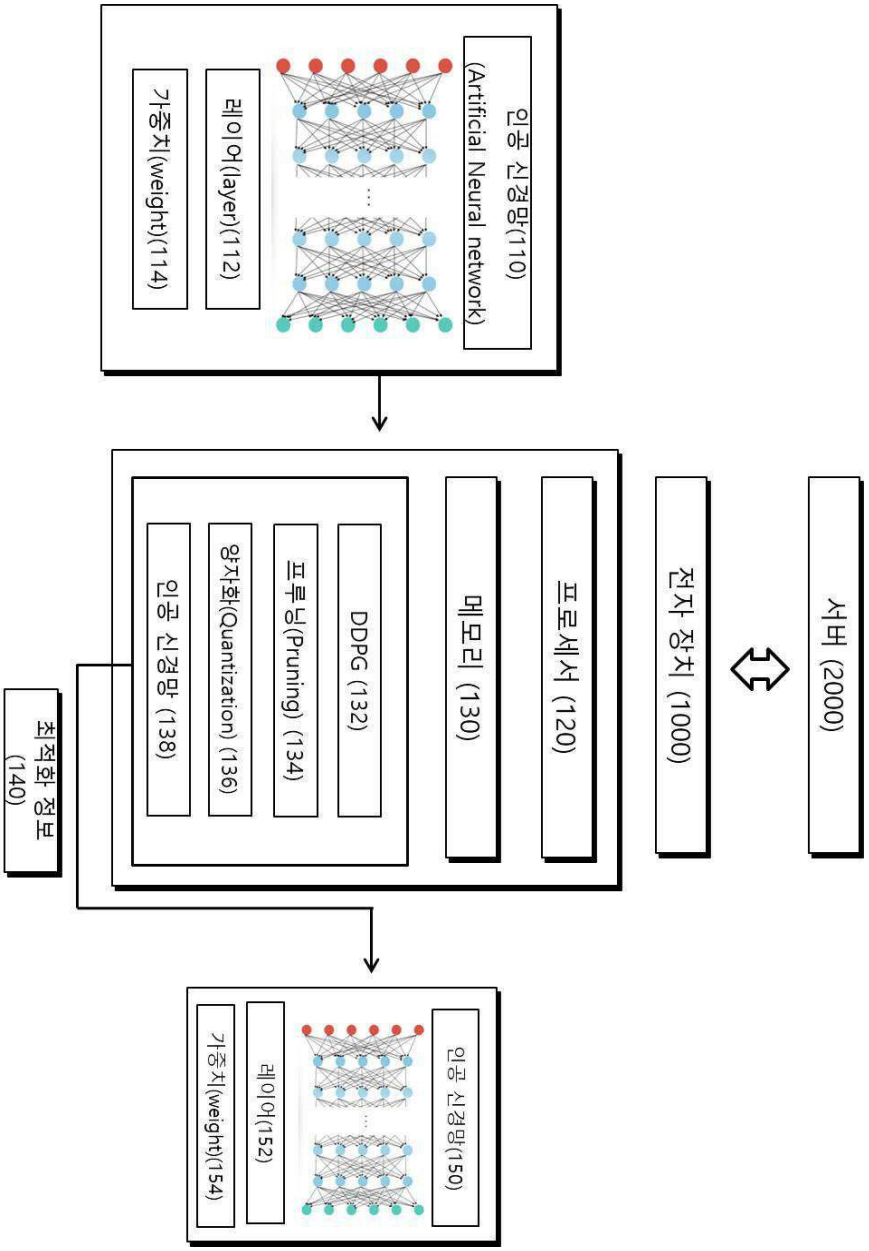
[0076] 일 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 개시를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0077] 또한, 상기 일 실시 예에 다른 방법을 수행하도록 하는 프로그램이 저장된 기록매체를 포함하는 컴퓨터 프로그램 장치가 제공될 수 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

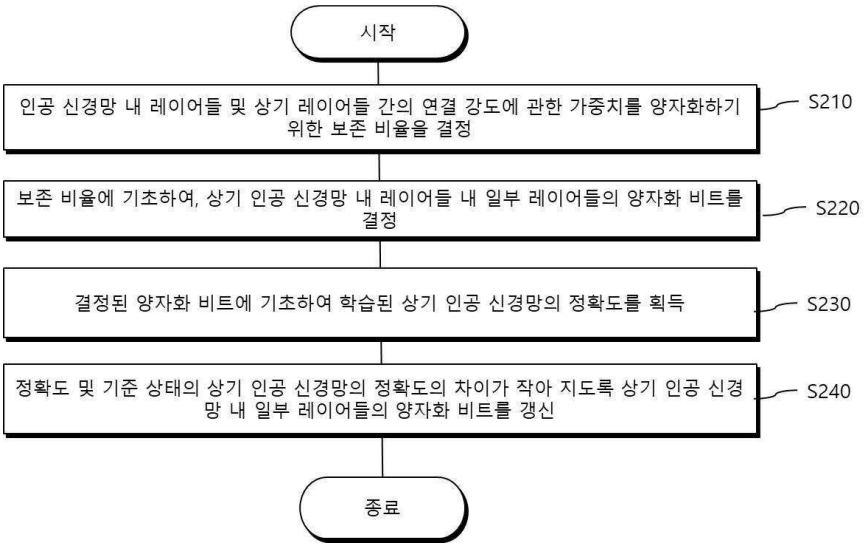
[0078] 이상에서 본 개시의 실시예에 대하여 상세하게 설명하였지만 본 개시의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 개시의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 개시의 권리범위에 속한다.

도면

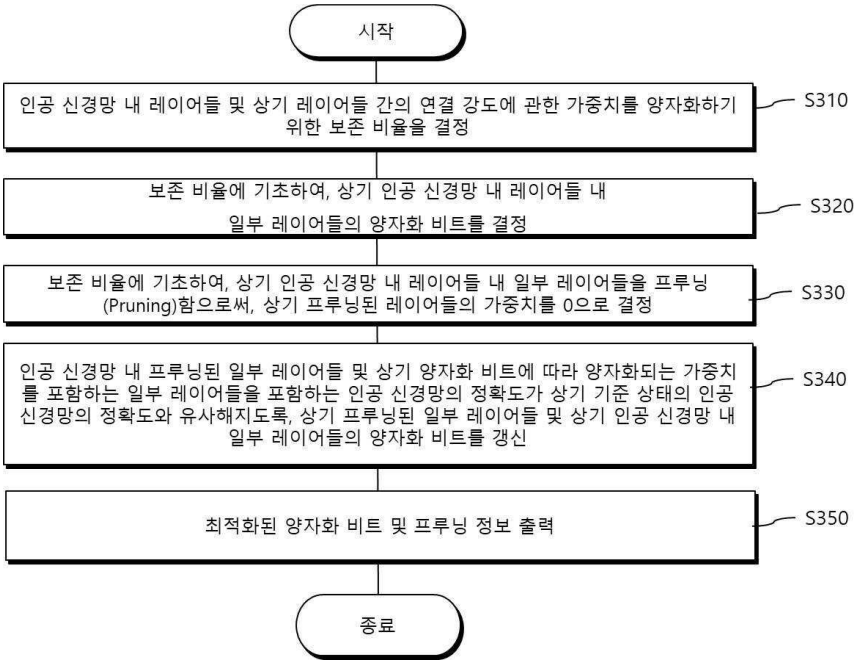
도면1


서버 (2000)
전자 장치 (1000)
프로세서 (120)
메모리 (130)
DDPG (132)
프루닝(Pruning) (134)
양자화(Quantization) (136)
인공 신경망 (138)
최적화 정보 (140)
인공 신경망(110)
(Artificial Neural network)
레이어(layer)(112)
가중치(weight)(114)
인공 신경망(150)
레이어(152)
가중치(weight)(154)

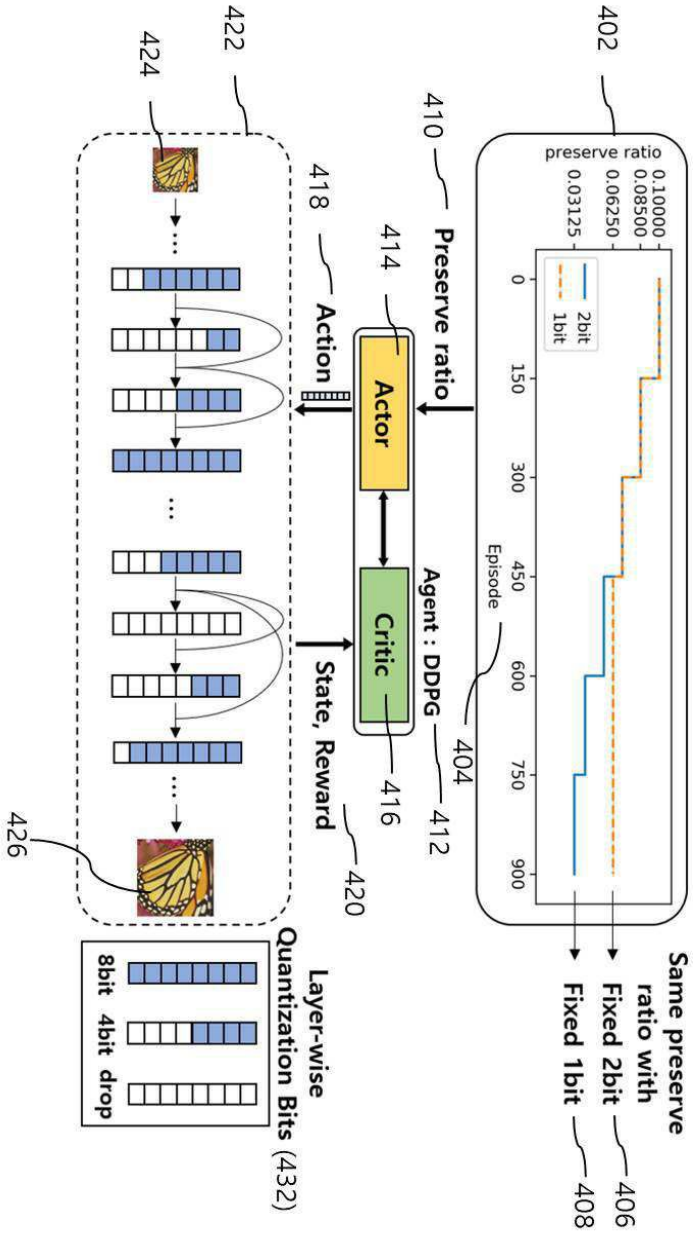
도면2


시작
인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정
S210
보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정
S220
결정된 양자화 비트에 기초하여 학습된 상기 인공 신경망의 정확도를 획득
S230
정확도 및 기준 상태의 상기 인공 신경망의 정확도의 차이가 작아 지도록 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신
S240
종료

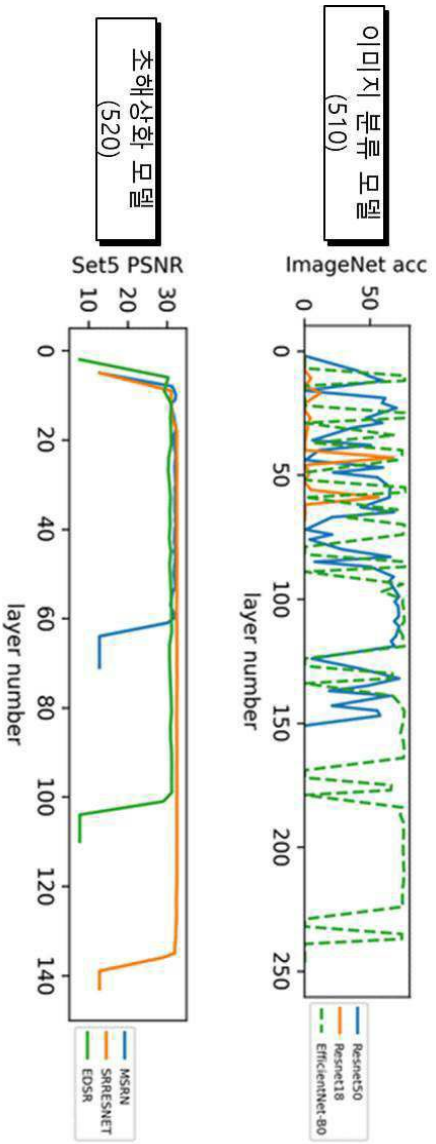
도면3


시작
인공 신경망 내 레이어들 및 상기 레이어들 간의 연결 강도에 관한 가중치를 양자화하기 위한 보존 비율을 결정
S310
보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들의 양자화 비트를 결정
S320
보존 비율에 기초하여, 상기 인공 신경망 내 레이어들 내 일부 레이어들을 프루닝(Pruning)함으로써, 상기 프루닝된 레이어들의 가중치를 0으로 결정
S330
인공 신경망 내 프루닝된 일부 레이어들 및 상기 양자화 비트에 따라 양자화되는 가중치를 포함하는 일부 레이어들을 포함하는 인공 신경망의 정확도가 상기 기준 상태의 인공 신경망의 정확도와 유사해지도록, 상기 프루닝된 일부 레이어들 및 상기 인공 신경망 내 일부 레이어들의 양자화 비트를 갱신
S340
최적화된 양자화 비트 및 프루닝 정보 출력
S350
종료

도면4


402
preserve ratio
0.10000
0.08500
0.06250
0.03125
2bit
1bit
0
150
300
450
600
750
900
Episode
404
410
Preserve ratio
414
Actor
418
Action
Agent : DDPG
412
Critic
416
State, Reward
420
Same preserve ratio with
Fixed 2bit
406
Fixed 1bit
408
422
424
426
Layer-wise Quantization Bits (432)
8bit
4bit
drop

도면5



도면6

(602) (604)

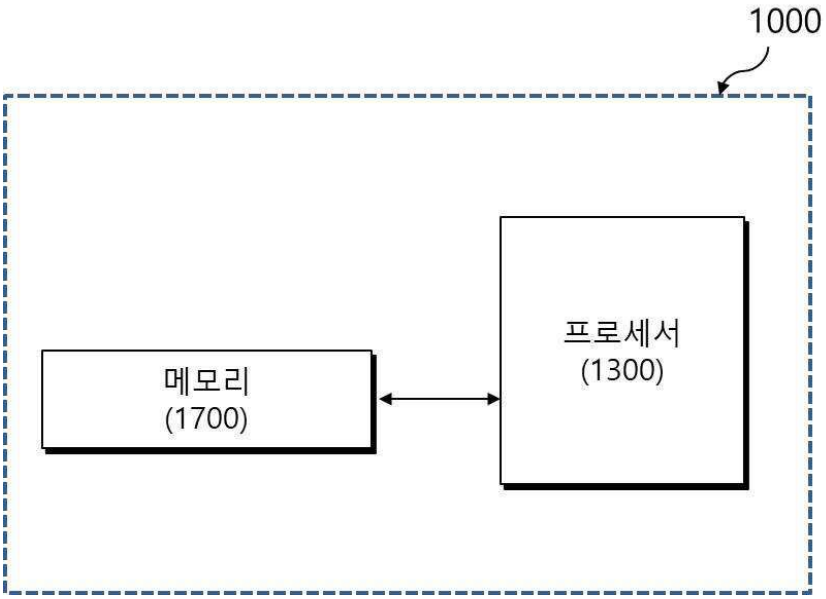
모 델	보존 비율	Set5	Set14	B100	Urban100
MSRN	1 (Full)	32.26/0.8960	28.63/0.7836	27.61/0.7380	26.22/0.7911
	0.1	32.17/0.8947	28.62/0.7823	27.59/0.7367	26.09/0.7869
	0.0625	31.97/0.8916	28.51/0.7793	27.52/0.7340	25.87/0.7792
	0.03125	31.60/0.8856	28.19/0.7714	27.30/0.7266	25.29/0.7580
DDBPN	1 (Full)	32.25/0.8957	28.68/0.7840	27.64/0.7380	26.31/0.7916
	0.1	32.21/0.8947	28.66/0.7830	27.62/0.7371	26.24/0.7896
	0.0625	32.00/0.8918	28.55/0.7800	27.55/0.7346	25.98/0.7819
	0.03125	31.42/0.8806	28.15/0.7682	27.28/0.7244	25.32/0.7567
EDSR	1 (Full)	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033
	0.1	32.22/0.8948	28.68/0.7849	27.65/0.7395	26.38/0.7962
	0.0625	30.89/0.8664	27.88/0.7614	27.17/0.7206	25.25/0.7524
	0.03125	-	-	-	-

도면7

(712)

	Model	Model size	Set5	Set14	B100	Urban100
702	BinarySR Full	5.499MB	36.36/0.952	32.16/0.904	-	29.96/0.901
704	BinarySR[]	0.928MB	35.66/0.946	31.56/0.897	-	28.76/0.882
X2	Ours Full	5.499MB	37.04/0.9561	32.66/0.9091	31.55/0.8918	29.78/0.9010
722	Ours 3bit	0.556MB	36.63/0.9533	32.37/0.9059	31.33/0.8890	29.41/0.8947
724	Ours 2bit	0.350MB	35.97/0.9494	32.02/0.9015	31.09/0.8852	28.85/0.8853
726						
706	BinarySR Full	6.089MB	31.10/0.875	27.59/0.764	-	25.09/0.750
708	BinarySR[]	1.518MB	30.34/0.864	27.16/0.756	-	24.48 / 0.728
x4	Ours Full	6.089MB	31.41/0.8828	28.07/0.7691	27.28/0.7263	25.13/0.7524
	Ours 3bit	0.6130MB	31.15/0.8771	27.88/0.7632	27.15/0.7218	24.90/0.8947
	Ours 2bit	0.384MB	30.28/0.8582	27.38/0.7503	26.84/0.7106	24.37/0.7176

도면8



도면9

