



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0093739
(43) 공개일자 2022년07월05일

(51) 국제특허분류(Int. Cl.)
G06N 5/04 (2006.01) G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 5/04 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2020-0184769
(22) 출원일자 2020년12월28일
심사청구일자 2020년12월28일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
홍충선
경기도 용인시 수지구 상현로 30-10, 233동 101호
김유준
경기도 수원시 영통구 매영로 110, 101동 303호
(74) 대리인
김홍석

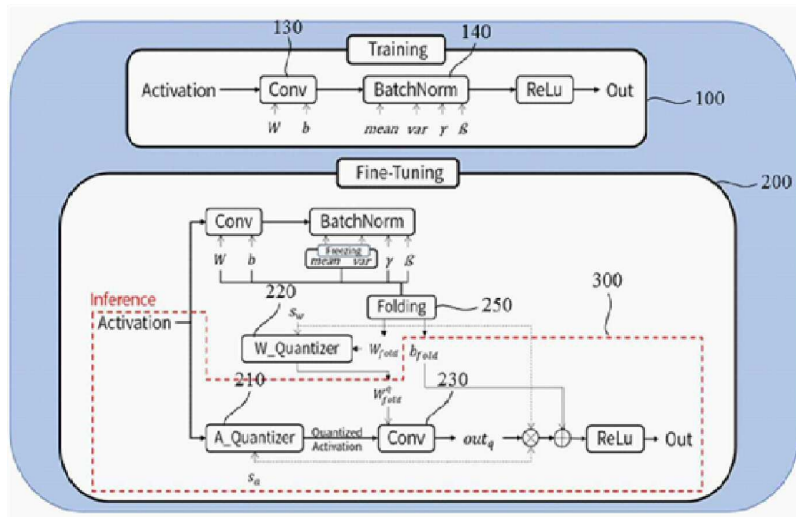
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치

(57) 요약

본 발명에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 장치가 제공된다. AI 추론 장치는 컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)가 호출되어 컨볼루션 레이어로 활성화 입력이 인가되고 배치 정규화 레이어를 통해 출력이 생성되도록 구성되는 트레이닝 모듈; 및 추론을 위한 활성화 입력 및 폴딩 모듈(folded module)의 가중치가 양자화기(quantizer)에 의해 양자화되도록 제어하고, 양자화된 활성화 입력과 가중치가 컨볼루션 연산을 통해 출력되도록 구성되는 튜닝 모듈을 포함하고, 상기 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 상기 튜닝 모듈의 컨볼루션 레이어가 추론을 위한 학습을 수행할 수 있다.

대표도 - 도1



(52) CPC특허분류
G06N 3/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711103234
과제번호	2015-0-00742-006
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	2020 Grand ICT 라이프 /컴패니온쉽 경험을 위한 지능형 인터랙션 융합 연구
기 여 율	1/2
과제수행기관명	경희대학교 산학협력단
연구기간	2020.01.01 ~ 2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116381
과제번호	2020-0-00364-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	인공지능반도체응용기술개발(R&D)
연구과제명	AI 기반 차량용 통신 기술 향상을 위한 NPU와 응용 시스템 개발
기 여 율	1/2
과제수행기관명	(주)딥엑스
연구기간	2020.04.01 ~ 2020.12.31

명세서

청구범위

청구항 1

양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 장치에 있어서,

컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)가 호출되어 컨볼루션 레이어로 활성화 입력이 인가되고 배치 정규화 레이어를 통해 출력이 생성되도록 구성되는 트레이닝 모듈; 및

추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화기(quantizer)에 의해 양자화되도록 제어하고, 양자화된 활성화 입력과 가중치가 컨볼루션 연산을 통해 출력되도록 구성되는 튜닝 모듈을 포함하고,

상기 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 상기 튜닝 모듈의 컨볼루션 레이어가 추론을 위한 학습을 수행하는 것을 특징으로 하는, AI 추론 장치.

청구항 2

제1 항에 있어서,

상기 폴디드 모듈은 상기 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴디드되어 하나의 폴디드 모듈로 형성되고, 상기 폴디드 모듈을 통한 출력이 상기 튜닝 모듈 내의 추론 모듈로 입력되는 것을 특징으로 하는, AI 추론 장치.

청구항 3

제1 항에 있어서,

상기 튜닝 모듈은,

상기 추론을 위한 활성화 입력이 인가되는 제1 양자화기; 및

상기 폴디드 모듈(folded module)의 가중치가 인가되는 제2 양자화기를 포함하는, AI 추론 장치.

청구항 4

제1 항에 있어서,

상기 튜닝 모듈은,

상기 학습가능 양자화 파라미터를 이용한 학습과 상기 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하는 것을 특징으로 하는, AI 추론 장치.

청구항 5

제1 항에 있어서,

상기 튜닝 모듈은,

상기 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도를 향상시키는, AI 추론 장치.

청구항 6

제5 항에 있어서,

상기 튜닝 모듈은,

동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하고,

상기 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하는, AI 추론 장치.

청구항 7

제1 항에 있어서,

상기 튜닝 모듈은,

학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하되,

상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행되고,

상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산 이후 배치 레이어 관련 연산 과정이 생략되는 것을 특징으로 하는, AI 추론 장치.

청구항 8

양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 로컬 디바이스에 있어서,

사용자 입력을 수신하도록 구성된 인터페이스부;

로컬 디바이스 또는 주변의 상태 정보를 센싱하도록 구성된 센서부; 및

인터페이스부와 동작 가능하게 결합되고, 사용자 입력, 상변 정보 및 컨텍스트 정보 중 적어도 하나에 기반하여 디바이스 제어를 수행함에 있어 학습을 위한 트레이닝 및 추론을 위한 학습을 수행하도록 구성된 제어부를 포함하고,

상기 제어부는,

컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)가 호출되어 컨볼루션 레이어로 활성화 입력이 인가되고 배치 정규화 레이어를 통해 출력이 생성되도록 구성되는 트레이닝 모듈; 및

추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화기(quantizer)에 의해 양자화되도록 제어하고, 양자화된 활성화 입력과 가중치가 컨볼루션 연산을 통해 출력되도록 구성되는 튜닝 모듈을 포함하고,

상기 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 상기 튜닝 모듈의 컨볼루션 레이어가 추론을 위한 학습을 수행하는 것을 특징으로 하는, 로컬 디바이스.

청구항 9

제8 항에 있어서,

상기 폴디드 모듈은 상기 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴디드되어 하나의 폴디드 모듈로 형성되고, 상기 폴디드 모듈을 통한 출력이 상기 튜닝 모듈 내의 추론 모듈로 입력되는 것을 특징으로 하는, 로컬 디바이스.

청구항 10

제8 항에 있어서,

상기 튜닝 모듈은,

상기 추론을 위한 활성화 입력이 인가되는 제1 양자화기; 및

상기 폴디드 모듈(folded module)의 가중치가 인가되는 제2 양자화기를 포함하는, 로컬 디바이스.

청구항 11

제8 항에 있어서,

상기 제어부는,

상기 학습가능 양자화 파라미터를 이용한 학습과 상기 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동

시에 진행하도록 상기 튜닝 모듈을 제어하는 것을 특징으로 하는, 로컬 디바이스.

청구항 12

제8 항에 있어서,

상기 제어부는,

상기 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도를 향상되도록 상기 튜닝 모듈을 제어하는, 로컬 디바이스.

청구항 13

제12 항에 있어서,

상기 제어부는,

동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하고,

상기 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하도록 상기 튜닝 모듈을 제어하는, 로컬 디바이스.

청구항 14

제13 항에 있어서,

상기 제어부는,

학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하되,

상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행되고,

상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산 이후 배치 레이어 관련 연산 과정이 생략되는 것을 특징으로 하는, 로컬 디바이스.

청구항 15

양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 방법에 있어서, 상기 방법은 로컬 디바이스의 프로세서에 의해 수행되고, 상기 방법은,

컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)를 호출하여 컨볼루션 레이어로 활성화 입력이 인가되어 배치 정규화 레이어를 통해 출력을 생성하여, 디바이스 제어를 위한 트레이닝을 수행하는 트레이닝 과정;

추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화기(quantizer)에 의해 양자화되도록 제어하는 튜닝 과정; 및

양자화된 활성화 입력과 가중치에 따른 컨볼루션 연산을 통해, 디바이스 제어와 연관된 추론을 수행하는 추론 과정을 포함하고,

상기 튜닝 과정에서, 상기 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 튜닝 모듈의 컨볼루션 레이어가 추론을 위한 학습을 수행하는 것을 특징으로 하는, AI 추론 방법.

청구항 16

제15 항에 있어서,

상기 트레이닝 과정에서, 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴디드되어 하나의 폴디드 모듈로 형성되고, 상기 폴디드 모듈을 통한 출력이 튜닝 모듈 내의 추론 모듈로 입력되고,

상기 튜닝 모듈은,

상기 추론을 위한 활성화 입력이 인가되는 제1 양자화기; 및

상기 폴디드 모듈(folded module)의 가중치가 인가되는 제2 양자화기를 포함하는, AI 추론 방법.

청구항 17

제15 항에 있어서,

상기 튜닝 과정에서,

상기 학습가능 양자화 파라미터를 이용한 학습과 상기 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하는 것을 특징으로 하는, AI 추론 방법.

청구항 18

제15 항에 있어서,

상기 튜닝 과정에서,

상기 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도를 향상시키는, AI 추론 방법.

청구항 19

제18 항에 있어서,

상기 튜닝 과정에서,

동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하고,

상기 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하는, AI 추론 방법.

청구항 20

제15 항에 있어서,

상기 튜닝 과정에서,

학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하되,

상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행되고,

상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산 이후 배치 레이어 관련 연산 과정이 생략되는 것을 특징으로 하는, AI 추론 방법.

발명의 설명

기술 분야

[0001] 본 발명은 AI 추론을 수행하는 장치에 관한 것이다. 보다 상세하게는, 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치에 대한 것이다.

배경 기술

[0002] 최근에는 인공지능(Artificial Intelligence: AI)과 딥러닝(Deep Learning) 등 머신 러닝(machine learning)에 관한 관심이 크게 증가하고 있다. 머신 러닝(machine learning; 기계학습)이란 인공지능(Artificial Intelligence: AI)의 한 종류에 해당할 수 있다. 머신 러닝은 데이터를 기반으로 컴퓨터가 스스로 학습한 내용을 바탕으로 회귀, 분류, 군집화 등의 예측 작업을 수행하는 것을 지칭한다.

- [0003] 딥 러닝(deep learning)은 사람의 사고방식을 컴퓨터에게 가르치는 기계 학습의 한 분야이다. 딥 러닝은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화(abstractions, 다량의 데이터나 복잡한 자료를 속에서 핵심적인 내용 또는 기능을 요약하는 작업)를 시도하는 기계학습(machine learning) 알고리즘의 집합으로 정의될 수 있다.
- [0004] 딥 러닝 구조는 인공신경망(ANN, artificial neural networks)에 기반하여 설계된 개념이다. 인공신경망은 가상의 뉴런을 수학적으로 모델링한 후 시뮬레이션하여 인간의 뇌와 같은 학습 능력을 갖게 하고자 하는 알고리즘이다. 딥 러닝 구조는 주로 패턴인식에 많이 사용된다. 딥 러닝에서 이용하는 인공신경망 모델은 선형 맞춤(linear fitting)과 비선형 변환(nonlinear transformation or activation)을 반복해 쌓아 올린 구조를 가진다. 딥 러닝에서 사용하는 신경망 모델은 심층 신경망(Deep Neural Network, DNN), 합성곱 신경망(Convolutional Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN), 제한 볼츠만 머신(Restricted Boltzmann Machine, RBM), 심층 신뢰 신경망(Deep Belief Network, DBN), 심층 Q-네트워크(Deep Q-Networks) 등을 들 수 있다.
- [0005] 딥 러닝의 트레이닝 과정에서는 수많은 학습 데이터를 가지고 인공신경망 모델의 파라미터를 최적화하게 된다. 딥 러닝의 트레이닝 과정에서는 (Error-backpropagation 알고리즘), 경사하강법(Gradient Decent 기법) 등이 이용될 수 있다. 학습 데이터를 이용한 트레이닝 과정이 수행되고 나면 최적화된 파라미터를 가지는 모델이 생성되며, 생성된 모델에 라벨링되지 않은 데이터를 입력하면 입력된 데이터에 상응하는 결과값(레이블)을 예측할 수 있다.
- [0006] 하지만, 학습 데이터의 수가 많을수록 보다 정확한 예측이 가능한 모델이 형성되고, 이에 따라 인공 지능(artificial intelligence, AI) 기반 추론 시, 학습 시간이 증가하는 문제점이 있다.

발명의 내용

해결하려는 과제

- [0007] 따라서, 본 발명의 목적은 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치를 제공함에 있다.
- [0008] 또한, 본 발명의 목적은 정확도는 유지하면서 4-bit 및 2-bit로 모델 압축이 가능한 Learnable Quantization Parameter를 활용하여 더욱 빠른 추론이 가능한 AI 추론 방법을 제공함에 있다.
- [0009] 또한, 본 발명의 목적은 AI 추론과 관련하여, 양자화와 Batch Folding을 동시에 진행하는 방법론을 제시한다.

과제의 해결 수단

- [0010] 상기와 같은 과제를 해결하기 위한 본 발명에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 장치가 제공된다. AI 추론 장치는 컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)가 호출되어 컨볼루션 레이어로 활성화 입력이 인가되고 배치 정규화 레이어를 통해 출력이 생성되도록 구성되는 트레이닝 모듈; 및 추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화(quantizer)에 의해 양자화되도록 제어하고, 양자화된 활성화 입력과 가중치가 컨볼루션 연산을 통해 출력되도록 구성되는 튜닝 모듈을 포함하고, 상기 양자화(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 상기 튜닝 모듈의 컨볼루션 레이어가 추론을 위한 학습을 수행할 수 있다.
- [0011] 일 실시 예에 따르면, 상기 폴디드 모듈은 상기 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴디드되어 하나의 폴디드 모듈로 형성되고, 상기 폴디드 모듈을 통한 출력이 상기 튜닝 모듈 내의 추론 모듈로 입력될 수 있다.
- [0012] 일 실시 예에 따르면, 상기 튜닝 모듈은 상기 추론을 위한 활성화 입력이 인가되는 제1 양자화기; 및 상기 폴디드 모듈(folded module)의 가중치가 인가되는 제2 양자화기를 포함할 수 있다.
- [0013] 일 실시 예에 따르면, 상기 튜닝 모듈은 상기 학습가능 양자화 파라미터를 이용한 학습과 상기 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행할 수 있다.
- [0014] 일 실시 예에 따르면, 상기 튜닝 모듈은 상기 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도를 향상시킬

수 있다.

- [0015] 일 실시 예에 따르면, 상기 튜닝 모듈은 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨벌루션 레이어를 제어하고, 상기 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨벌루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행할 수 있다.
- [0016] 일 실시 예에 따르면, 상기 튜닝 모듈은 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨벌루션 레이어를 제어하되, 상기 제2 컨벌루션 레이어를 통한 컨벌루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행되고, 상기 제2 컨벌루션 레이어를 통한 컨벌루션 연산 이후 배치 레이어 관련 연산 과정이 생략될 수 있다.
- [0017] 본 발명의 다른 양상에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 로컬 디바이스가 제공된다. 상기 로컬 디바이스는 사용자 입력을 수신하도록 구성된 인터페이스부; 로컬 디바이스 또는 주변의 상태 정보를 센싱하도록 구성된 센서부; 및 인터페이스부와 동작 가능하게 결합되고, 사용자 입력, 상변 정보 및 컨텍스트 정보 중 적어도 하나에 기반하여 디바이스 제어를 수행함에 있어 학습을 위한 트레이닝 및 추론을 위한 학습을 수행하도록 구성된 제어부를 포함한다. 상기 제어부는 컨벌루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)가 호출되어 컨벌루션 레이어로 활성화 입력이 인가되고 배치 정규화 레이어를 통해 출력이 생성되도록 구성되는 트레이닝 모듈; 및 추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화기(quantizer)에 의해 양자화되도록 제어하고, 양자화된 활성화 입력과 가중치가 컨벌루션 연산을 통해 출력되도록 구성되는 튜닝 모듈을 포함하고, 상기 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 상기 튜닝 모듈의 컨벌루션 레이어가 추론을 위한 학습을 수행할 수 있다.
- [0018] 일 실시 예에 따르면, 상기 폴디드 모듈은 상기 트레이닝 모듈 내의 컨벌루션 레이어와 배치 정규화 레이어가 폴디드되어 하나의 폴디드 모듈로 형성되고, 상기 폴디드 모듈을 통한 출력이 상기 튜닝 모듈 내의 추론 모듈로 입력될 수 있다.
- [0019] 일 실시 예에 따르면, 상기 튜닝 모듈은 상기 추론을 위한 활성화 입력이 인가되는 제1 양자화기; 및 상기 폴디드 모듈(folded module)의 가중치가 인가되는 제2 양자화기를 포함할 수 있다.
- [0020] 일 실시 예에 따르면, 상기 제어부는 상기 학습가능 양자화 파라미터를 이용한 학습과 상기 컨벌루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하도록 상기 튜닝 모듈을 제어할 수 있다.
- [0021] 일 실시 예에 따르면, 상기 제어부는 상기 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도가 향상되도록 상기 튜닝 모듈을 제어할 수 있다.
- [0022] 일 실시 예에 따르면, 상기 제어부는 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨벌루션 레이어를 제어하고, 상기 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨벌루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하도록 상기 튜닝 모듈을 제어할 수 있다.
- [0023] 일 실시 예에 따르면, 상기 제어부는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨벌루션 레이어를 제어하되, 상기 제2 컨벌루션 레이어를 통한 컨벌루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행되고, 상기 제2 컨벌루션 레이어를 통한 컨벌루션 연산 이후 배치 레이어 관련 연산 과정이 생략될 수 있다.
- [0024] 본 발명의 또 다른 양상에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 방법이 제공된다. 상기 방법은 로컬 디바이스의 프로세서에 의해 수행되고, 상기 방법은 컨벌루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)를 호출하여 컨벌루션 레이어로 활성화 입력이 인가되어 배치 정규화 레이어를 통해 출력을 생성하여, 디바이스 제어를 위한 트레이닝을 수행하는 트레이닝 과정; 추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화기(quantizer)에 의해 양자화되도록 제어하는 튜닝 과정; 및 양자화된 활성화 입력과 가중치에 따른 컨벌루션 연산을 통해, 디바이스 제어와 연관된 추론을 수행하는 추론 과정을 포함하고, 상기 튜닝 과정에서, 상기 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 튜닝 모듈의 컨벌루션 레이어가 추론을 위한 학습을 수행할

수 있다.

- [0025] 일 실시 예에 따르면, 상기 트레이닝 과정에서, 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴딩되어 하나의 폴딩 모듈로 형성되고, 상기 폴딩 모듈을 통한 출력이 튜닝 모듈 내의 추론 모듈로 입력될 수 있다. 상기 튜닝 모듈은 상기 추론을 위한 활성화 입력이 인가되는 제1 양자화기; 및 상기 폴딩 모듈(folded module)의 가중치가 인가되는 제2 양자화기를 포함할 수 있다.
- [0026] 일 실시 예에 따르면, 상기 튜닝 과정에서, 상기 학습가능 양자화 파라미터를 이용한 학습과 상기 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행할 수 있다.
- [0027] 일 실시 예에 따르면, 상기 튜닝 과정에서, 상기 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도를 향상시킬 수 있다.
- [0028] 일 실시 예에 따르면, 상기 튜닝 과정에서, 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하고, 상기 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행할 수 있다.
- [0029] 일 실시 예에 따르면, 상기 튜닝 과정에서, 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어하되, 상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행되고, 상기 제2 컨볼루션 레이어를 통한 컨볼루션 연산 이후 배치 레이어 관련 연산 과정이 생략될 수 있다.

발명의 효과

- [0030] 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치는 다음과 같은 기술적 효과를 갖는다.
- [0031] 본 발명의 적어도 일 실시 예에 따르면, AI 추론 시 양자화로 인한 빠른 행렬 연산이 가능하며 Batch Normalization 관련 연산이 불필요하다.
- [0032] 본 발명의 적어도 일 실시 예에 따르면, 기존 양자화 방법보다 더욱 빠른 추론속도를 가질 수 있다. 그렇기 때문에 IoT Device, 스마트폰, 엣지 등과 같이 저 사양 컴퓨팅 능력을 가지며 실시간 서비스를 요구하는 상황에서 해당 방법은 효율적으로 사용될 수 있다.
- [0033] 본 발명의 적어도 일 실시 예에 따르면, On device AI 가 가능하게 하여 사용자의 Privacy를 보존함은 물론 네트워크, 중앙서버 등의 상황에 관계없이 최상의 AI 서비스를 로컬 디바이스에서 제공할 수 있을 것이다.
- [0034] 상술한 본 발명의 특징 및 효과는 첨부된 도면과 관련한 다음의 상세한 설명을 통하여 보다 분명해 질 것이며, 그에 따라 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다.

도면의 간단한 설명

- [0035] 도 1은 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치의 구성도를 나타낸다.
- 도 2는 본 발명에서 제안한 방법에서 activation의 양자화 범위에 따른 수렴 속도를 나타낸 것이다.
- 도 3은 본 발명에 따른 AI 추론을 수행하는 로컬 디바이스의 상세 구성을 나타낸다.
- 도 4는 본 발명에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0036] 상술한 본 발명의 특징 및 효과는 첨부된 도면과 관련한 다음의 상세한 설명을 통하여 보다 분명해 질 것이며, 그에 따라 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다.
- [0037] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시 예를 가질 수 있는바, 특정 실시 예들을 도면에 예시하고 상세한 설명에 구체적으로 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되

어야 한다.

[0038] 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용한다.

[0039] 제1, 제2등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0040] 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0041] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미가 있다.

[0042] 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않아야 한다.

[0043] 이하의 설명에서 사용되는 구성요소에 대한 접미사 모듈, 블록 및 부는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다.

[0044] 이하, 본 발명의 바람직한 실시 예를 첨부한 도면을 참조하여 당해 분야에 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 설명한다. 하기에 본 발명의 실시 예를 설명함에 있어, 관련된 공지의 기능 또는 공지의 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다.

[0045] 이하에서는, 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치에 대해 설명하기로 한다. 이와 관련하여, 본 발명에 따른 AI 추론과 관련하여, 1) IAO(Integer Arithmetic Only inference), 2) Fully Quantized Network 및 3) Learnable Quantization Parameter에 대해 설명하면 다음과 같다.

[0046] 1) IAO(Integer Arithmetic Only inference)

[0047] IAO는 현재 가장 널리 사용되는 양자화 기술로써 학습이 진행됨에 따라 AI 압축에 사용되는 매개변수들을 수학적, 통계적으로 최적화시킨다. 최적화에는 Exponential Moving Average 기법을 사용한다. IAO의 양자화 방법은 수학적 식 1 및 수학적 식 2와 같다.

수학적 식 1

$$x_{integer} = \text{round} \left(\frac{\text{clamp}(a, x, b)}{s} \right) + \text{zero_point} \quad ①$$

[0048]

수학적 식 2

$$\text{clamp}(a, b, x) = \begin{cases} a & x \leq a \\ x & a \leq x \leq b \\ b & x \geq b \end{cases} \quad ②$$

[0049]

[0050] 수학적 식 1의 x는 weight 및 activation을 나타내고 s는 양자화 scale, zero_point는 shift 연산으로써 정수이다. a는 최소 양자화 값이며, b는 최대 양자화 값이다. a와 b는 양자화 bits에 따라 다른 값을 가진다. 그리고 이 두 값은 Exponential Moving Average 기법에 의해 Update되며 s값과 zero_point는 a와 b 값에 의해 결정된다.

[0051] 2) Fully Quantized Network

[0052] Fully Quantized Network 구성 방법은 IAO의 방법과 마찬가지로 Uniform 양자화 기법을 사용할 수 있다. 하지만, 본 기법은 이미지 분류가 아닌 detection 상황에서의 양자화 기법에 적용될 수 있다. 또한 그들은

Quantization Aware Training 기반의 Batch Folding을 실험하였으며 객체 인식 상황에서의 activation의 불안정성을 해결할 방법을 제안하였다. 그들은 백분위수(Percentile)을 활용하여 activation의 최대 양자화 값과 최소 양자화 값을 수치적으로 계산하여 activation에 의해 발생하는 학습 불안정성을 해결할 수 있다.

[0053] 3) Learnable Quantization Parameter

[0054] LSQ(Learned Step Size Quantization) 방법은 layer-wise 양자화로써 quantization aware training 기반의 양자화 기법을 사용할 수 있다. 이와 관련하여, quantization scale parameter인 s 를 사용할 수 있고, s 는 학습 Loss에 의한 Back propagation으로 학습된다. 이에 따른 양자화 방법은 수학적 식 3과 같다.

수학적 식 3

$$x_q = \text{round}[\text{clamp}(\frac{x}{s}, -Q_N, Q_P)] \quad \textcircled{3}$$

[0055]

[0056] x 는 weight 및 activation을 나타내며 Q 는 최소(N), 최대(Q) 양자화 값이다. Q 는 양자화 bit 수에 따라 달라진다. s 는 scale로써 Learnable Quantization Parameter이다. 또한 그들은 학습 안정성을 위해 gradient scale이라는 기법을 사용할 수 있다.

[0057] 한편, IAO와 같은 양자화 기법은 수학적으로 양자화 파라미터의 최적 값을 찾아내는 방법이다. 하지만 이 방법은 양자화 파라미터를 일반적인 Loss에 의한 back propagation으로 학습시키는 방법보다 정확도 측면에서 낮은 결과가 발생할 수 있다. Learnable Quantization Parameter에 의한 양자화 scale 최적화는 높은 정확도를 유지하면서 낮은 bit로의 양자화가 가능하다. 본 발명에서는 Learnable Quantization Parameter에 의한 양자화 방법에 의해 얻어진 모델을 Batch Folding으로 구현하고자 한다. 이에 따라, batch 관련 연산을 하지 않으면서 높은 정확도는 유지하면서 더욱 빠른 추론이 가능할 수 있다.

[0058] 본 발명에서 제안하는 방법은 IoT Device, 스마트폰, CCTV 등 컴퓨팅 능력이 낮은 디바이스에서 실시간 AI 서비스가 필요한 상황에서 효과가 극대화된다. 현재 대부분의 AI 추론은 AI의 계산량 때문에 높은 컴퓨팅 능력의 중앙 Cloud에서 실행된다. Cloud를 사용하면 네트워크 상황에 따라 지연시간이 발생할 수 있으며 Cloud가 멀리 위치해 있으면 실시간 AI 서비스가 불가능하다. 하지만 본 발명에서 제안하는 AI의 추론시간을 굉장히 빠르게 하기 때문에 AI가 로컬 디바이스에 위치하여 실시간 서비스를 제공할 수 있게 한다. 또한 32-bit Full-precision 모델의 정확도를 거의 유지하기 때문에 정확도 또한 굉장히 높다.

[0059] 이와 관련하여, 도 1은 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치의 구성도를 나타낸다. 즉, 도 1은 본 발명에서 제안한 방법의 Training, Fine-Tuning, Inference Flow(Quantizer: Learnable Quantization Parameter를 활용한 양자화) 과정을 나타낸다..

[0060] 도 1을 참조하면, AI 추론 장치는 트레이닝 모듈(100) 및 튜닝 모듈(200)을 포함하도록 구성될 수 있다. 트레이닝 모듈(100)은 training 과정을 수행하고, 튜닝 모듈(200)은 fine-tuning 과정을 수행할 수 있다.

[0061] 1) Training

[0062] 모델의 Training은 일반적인 모델의 Training과 같은 방법으로 진행된다. Convolution Layer - Batch Normalization Layer - Activation Function 순이다. 학습 또한 일반적인 Back propagation에 의한 Stochastic Gradient Descent로 학습된다.

[0063] 2) Fine-Tuning

[0064] Fine-Tuning 과정은 일반적인 Fine-Tuning 과정과 다른 Flow를 갖는다. Training 이후 Batch의 moving average와 variance는 training과정에서 충분히 최적화 되었다고 가정하고 freezing되어 고정된 값이 된다. Freezing을 하지 않으면 Batch 값의 불안정성에 의해 학습 진행이 어렵다. 그리고 Convolution Layer의 weight 및 bias, Batch Layer의 weight 및 bias 학습을 위해 Input은 위 방향으로 먼저 들어가게 된다. 이후 Conv Layer와 Batch Layer가 Folding 되어 하나의 Folded module이 된다. Activation 및 Folded Module의 weight는 Quantizer에 의해 양자화되고 양자화된 activation과 weight는 Integer convolution 연산으로 계산된다. Backward pass에서는 위쪽 방향의 Convolution layer의 weight 및 bias, Batch layer의 weight 및 bias 뿐만 아니라 quantizer의 learnable quantization parameter까지 Stochastic Gradient Descent에 의해 학습된다.

- [0065] 3) Inference
- [0066] 모델의 추론은 도 1과 같이 Activation - Quantizer for activation - Convolution layer(Folded layer), 양자화 파라미터(S)만으로 구성될 수 있다. 즉, Convolution 연산은 Integer 행렬 곱으로 시행되어 빠른 연산이 가능하다. 이에 더해 Batch Layer 관련 연산이 불필요하게 되어 더더욱 빠른 추론이 가능하다.
- [0067] 본 발명에 따른 AI 추론 방법을 통한 기술적 장점은 다음과 같다.
- [0068] 1) 제안한 방법은 fine-tuning 과정에서 쉽게 적용될 수 있다.
- [0069] 2) 제안한 방법은 모든 CNN, FCN 모델에 적용 가능하다.
- [0070] 3) 제안한 방법은 추론 시 Batch 관련 연산이 불필요하기 때문에 더욱 빠른 추론이 가능하다.
- [0071] 4) 제안한 방법은 Learnable Quantization Parameter에 의한 양자화 최적화에 의해 높은 정확도의 모델을 얻을 수 있다.
- [0072] 한편, 제안한 방법의 검증 및 실험을 위해 본 발명에서는 Public dataset인 Cifar10 데이터셋을 사용하였으며 사용 모델로는 VGG16, ResNet18, MobileNetv2, EfficientNetB0를 사용하였다. 모든 Training은 [120, 30]번 반복하였으며 Learning rate는 [0.01, 0.001]로 하였다. Fine-Tuning에서는 [120, 120, 120]번 반복하였으며 Learning rate는 [0.001, 0.0005, 0.0001]로 하였다.
- [0073] 표 1은 제안한 방법(Learnable Quantization Parameter를 활용한 양자화 + Batch Folding)과 IA0의 양자화 방법 + Batch Folding을 정확도 측면에서 비교한 것이다.

표 1

Method(Folding Batch norm + Quantization)	VGG16	ResNet18	MobileNetv2
32bits float	93.24	93.87	92.60
Learnable Quantization Parameter(4bits)	93.35	93.87	92.40
Asymmetric Uniform Quantization(4bits)	93.18	93.8	88.08

- [0074]
- [0075] 표 1을 참조하면, 제안한 방법이 IA0 + Batch Folding보다 VGG16은 0.17%, ResNet18은 0.07%, MobileNetv2는 3.32% 높은 것을 확인할 수 있다. IA0는 Channel wise, 제안한 방법은 Layer wise로 실험하였음에도 불구하고 제안한 방법의 정확도가 높은 것을 확인할 수 있다. 그리고 MobileNetv2는 IA0의 양자화 방법으로는 학습 진행이 불가능한 것을 확인하였다.
- [0076] EfficientNet은 swish 함수를 activation 함수로 사용하기 때문에 relu 함수와는 다르게 음수 값을 가질 수 있다. 그래서 제안한 방법의 4-bit 양자화에서 activation 양자화 범위를 다양하게 설정하여 실험한 결과를 표 2에 나타냈다.

표 2

Method(Folding Batch norm + Quantization)	Activation Range	EfficientNet
32bits float		91.60
	$0 \sim 2^4 - 1$	Unlearnable
Learnable Quantization Parameter(4bits)	$-1 \sim 2^4 - 2$	90.83
	$-2 \sim 2^4 - 3$	90.64
	$-2^3 \sim 2^3 - 1$	90.27
Asymmetric Uniform Quantization(4bits)	Asymmetric	51.19

- [0077]
- [0078] IA0는 Asymmetric Quantization임에도 불구하고 학습진행이 불가능한 것을 확인하였다. 제안한 방법은 4-bit 양자화에서 양자화 범위의 음수가 -1을 포함할 때 가장 높은 정확도를 보였다. 이와 관련하여, 도 2는 본 발명에

서 제안한 방법에서 activation의 양자화 범위에 따른 수렴 속도를 나타낸 것이다. 도 2를 참조하면, 파란 그래프가 -1을 포함한 범위로써 수렴속도가 가장 느리지만 최종 정확도는 가장 높은 것을 확인하였다. -2, -1을 포함한 빨간색 선은 수렴 속도도 적당하고 최종 정확도도 높은 편인 것을 볼 수 있다. 마지막으로 검은색 선은 symmetric으로 수렴속도가 가장 빠르지만 최종 정확도는 가장 낮았다.

- [0079] 진술한 기술적 특징을 참조하여, 도 1의 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치에 대해 상세하게 설명하면 다음과 같다. 이와 관련하여, 본 발명에 따른 양자화 및 배치 폴딩을 통한 AI 추론 방법에서 청구하고자 하는 주요 특징은 다음과 같다. 한편, 본 발명에서 청구하고자 하는 주요 특징은 이에 한정되는 것은 아니다.
- [0080] 1) Learnable Quantization Parameter(Learned Step size Quantization)과 Batch Folding을 동시에 진행하는 방법.
- [0081] 2) EfficientNet과 같이 swish 함수를 쓰는 모델의 양자화 범위를 변경하여 1)의 방법을 적용하는 방법.
- [0082] 도 1을 참조하면, AI 추론 장치는 트레이닝 모듈(100) 및 튜닝 모듈(200)을 포함하도록 구성될 수 있다. AI 추론 장치는 추론 모듈(300)을 더 포함하도록 구성될 수도 있다. 한편, 추론 모듈(300)은 튜닝 모듈(200) 내에 포함되도록 구현될 수도 있다. 트레이닝 모듈(100)은 training 과정을 수행하고, 튜닝 모듈(200)은 fine-tuning 과정을 수행할 수 있다.
- [0083] 트레이닝 모듈(100)은 컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)가 호출되어 컨볼루션 레이어로 활성화 입력이 인가되고 배치 정규화 레이어를 통해 출력이 생성되도록 구성될 수 있다.
- [0084] 튜닝 모듈(200)은 추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화기(quantizer)에 의해 양자화되도록 제어하고, 양자화된 활성화 입력과 가중치가 컨볼루션 연산을 통해 출력되도록 구성될 수 있다. 이 경우, 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 상기 튜닝 모듈의 컨볼루션 레이어가 추론을 위한 학습을 수행할 수 있다.
- [0085] 한편, 폴디드 모듈(250)은 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴디드되어 하나의 폴디드 모듈로 형성되고, 폴디드 모듈(250)을 통한 출력이 튜닝 모듈(200) 내의 추론 모듈(300)로 입력될 수 있다.
- [0086] 튜닝 모듈(200)은 제1 양자화기(210) 및 제2 양자화기(220)를 포함하도록 구성될 수 있다. 제1 양자화기(210)는 추론을 위한 활성화 입력이 인가되도록 구성될 수 있다. 제2 양자화기(220)는 폴디드 모듈(folded module, 250)의 가중치가 인가되도록 구성될 수 있다.
- [0087] 튜닝 모듈(200)은 학습가능 양자화 파라미터를 이용한 학습과 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행할 수 있다. 한편, 튜닝 모듈(200)은 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경되되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도를 향상시킬 수 있다.
- [0088] 튜닝 모듈(200)은 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어(230)를 제어할 수 있다. 튜닝 모듈(200)은 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨볼루션 레이어(130)와 배치 정규화 레이어(140)의 배치 폴딩을 동시에 진행할 수 있다.
- [0089] 튜닝 모듈(200)은 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어(230)를 제어할 수 있다. 이 경우, 제2 컨볼루션 레이어(230)를 통한 컨볼루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행될 수 있다. 이에 따라, 제2 컨볼루션 레이어를 통한 컨볼루션 연산 이후 배치 레이어 관련 연산 과정이 생략될 수 있다.
- [0090] 이상에서는 본 발명의 일 양상에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치에 대해 설명하였다. 이하에서는 본 발명의 다른 양상에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 로컬 디바이스에 대해 설명하였다. 도 3은 본 발명에 따른 AI 추론을 수행하는 로컬 디바이스의 상세 구성을 나타낸다.
- [0091] 도 3을 참조하면, 로컬 디바이스는 인터페이스부(1100), 센서부(1200) 및 제어부 (또는 프로세서)(1300)를 포함

하도록 구성될 수 있다. 제어부 (또는 프로세서)(1300)는 정보를 저장하도록 구성된 메모리를 포함하도록 구성될 수 있다. 또는, 메모리는 제어부 (또는 프로세서)(1300)와 별도로 배치될 수도 있다.

- [0092] 인터페이스부(1100)는 사용자 입력을 수신하도록 구성될 수 있다. 센서부(1200)는 로컬 디바이스 또는 주변 상태 정보를 센싱하도록 구성될 수 있다. 제어부 (1300)는 인터페이스부(1100) 및 센서부(1200)와 동작 가능하게 결합될 수 있다.
- [0093] 제어부 (1300)는 사용자 입력, 상변 정보 및 컨텍스트 정보 중 적어도 하나에 기반하여 디바이스 제어를 수행함에 있어 학습을 위한 트레이닝 및 추론을 위한 학습을 수행하도록 구성될 수 있다. 제어부 (1300)는 트레이닝 모듈(100) 및 튜닝 모듈(200)을 포함할 수 있다. 이와 관련하여, 트레이닝 모듈(100) 및 튜닝 모듈(200)은 도 1의 트레이닝 모듈(100) 및 튜닝 모듈(200)에 대응할 수 있다.
- [0094] 도 1 및 도 3을 참조하면, 트레이닝 모듈(100)은 컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)가 호출되어 컨볼루션 레이어로 활성화 입력이 인가되고 배치 정규화 레이어를 통해 출력이 생성되도록 구성될 수 있다. 튜닝 모듈(200)은 추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화(quantizer)에 의해 양자화되도록 제어하고, 양자화된 활성화 입력과 가중치가 컨볼루션 연산을 통해 출력되도록 구성될 수 있다. 이 경우, 양자화(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 상기 튜닝 모듈의 컨볼루션 레이어가 추론을 위한 학습을 수행할 수 있다.
- [0095] 한편, 폴디드 모듈(250)은 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴디드되어 하나의 폴디드 모듈로 형성되고, 폴디드 모듈(250)을 통한 출력이 튜닝 모듈(200) 내의 추론 모듈(300)로 입력될 수 있다.
- [0096] 튜닝 모듈(200)은 제1 양자화기(210) 및 제2 양자화기(220)를 포함하도록 구성될 수 있다. 제1 양자화기(210)는 추론을 위한 활성화 입력이 인가되도록 구성될 수 있다. 제2 양자화기(220)는 폴디드 모듈(folded module, 250)의 가중치가 인가되도록 구성될 수 있다.
- [0097] 제어부(1300)는 학습가능 양자화 파라미터를 이용한 학습과 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행하도록 튜닝 모듈(200)을 제어할 수 있다. 한편, 제어부(1300)는 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도가 향상되도록 튜닝 모듈(200)을 제어할 수 있다.
- [0098] 제어부(1300)는 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어(230)를 제어할 수 있다. 제어부(1300)는 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨볼루션 레이어(130)와 배치 정규화 레이어(140)의 배치 폴딩을 동시에 진행하도록 튜닝 모듈(200)을 제어할 수 있다.
- [0099] 제어부(1300)는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어(230)를 제어할 수 있다. 이 경우, 제2 컨볼루션 레이어(230)를 통한 컨볼루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행될 수 있다. 이에 따라, 제2 컨볼루션 레이어를 통한 컨볼루션 연산 이후 배치 레이어 관련 연산 과정이 생략될 수 있다.
- [0100] 이하에서는 본 발명의 또 다른 양상에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 방법에 대해 설명한다. 이와 관련하여, 도 4는 본 발명에 따른 양자화 및 배치 폴딩(batch folding)을 통해 AI 추론을 수행하는 방법의 흐름도이다. AI 추론 방법은 로컬 디바이스의 프로세서에 의해 수행될 수 있다. AI 추론 방법은 트레이닝 과정(S100), 튜닝 과정(S200) 및 추론 과정(S300)을 포함하도록 구성될 수 있다. 도 1, 도 3 및 도 4를 참조하여, AI 추론 방법에 대해 상세히 설명하면 다음과 같다.
- [0101] 트레이닝 과정(S100)에서, 컨볼루션 레이어와 배치 정규화 레이어를 구성하고, 활성화 함수(activation function)를 호출하여 컨볼루션 레이어로 활성화 입력이 인가되어 배치 정규화 레이어를 통해 출력을 생성하여, 디바이스 제어를 위한 트레이닝을 수행할 수 있다.
- [0102] 튜닝 과정(S200)에서, 추론을 위한 활성화 입력 및 폴디드 모듈(folded module)의 가중치가 양자화(quantizer)에 의해 양자화되도록 제어할 수 있다. 추론 과정(S300)에서, 양자화된 활성화 입력과 가중치에 따른 컨볼루션 연산을 통해, 디바이스 제어와 연관된 추론을 수행할 수 있다. 한편, 튜닝 과정(S200)에서, 양자화기(quantizer)를 통해 입력되는 학습가능 양자화 파라미터(quantization parameter)를 포함하여 튜닝 모듈의 컨

볼루션 레이어가 추론을 위한 학습을 수행할 수 있다.

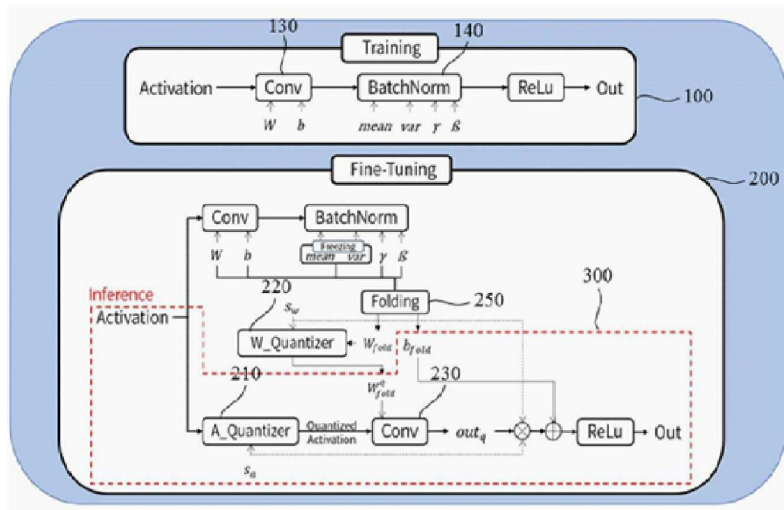
- [0103] 트레이닝 과정(S100)에서, 트레이닝 모듈 내의 컨볼루션 레이어와 배치 정규화 레이어가 폴딩되어 하나의 폴딩 모듈로 형성되고, 상기 폴딩 모듈을 통한 출력이 튜닝 모듈 내의 추론 모듈로 입력될 수 있다. 한편, 도 1 및 도 3을 참조하면, 튜닝 모듈(200)은 제1 양자화기(210) 및 제2 양자화기(220)를 포함하도록 구성될 수 있다. 제1 양자화기(210)는 추론을 위한 활성화 입력이 인가되도록 구성될 수 있다. 제2 양자화기(220)는 폴딩 모듈(folded module, 250)의 가중치가 인가되도록 구성될 수 있다.
- [0104] 튜닝 과정(S200)에서, 학습가능 양자화 파라미터를 이용한 학습과 상기 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행할 수 있다. 한편, 튜닝 과정(S200)에서, 학습가능 양자화 파라미터의 activation 양자화 범위를 동적으로 변경하되, 상기 activation 양자화 범위의 하한을 음수(negative number)로 설정하여 학습 정확도를 향상시킬 수 있다.
- [0105] 튜닝 과정(S200)에서, 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어를 제어할 수 있다. 튜닝 과정(S200)에서, 동적으로 변경된 양자화 범위를 갖는 학습가능 양자화 파라미터를 이용한 학습과 제1 컨볼루션 레이어와 배치 정규화 레이어의 배치 폴딩을 동시에 진행할 수 있다.
- [0106] 튜닝 과정(S200)에서, 학습가능 양자화 파라미터를 포함하여 추론을 위한 학습을 수행하도록 제2 컨볼루션 레이어(230)를 제어할 수 있다. 이 경우, 제2 컨볼루션 레이어(230)를 통한 컨볼루션 연산은 정수 행렬 곱(integer matrix multiplication)을 통해 수행될 수 있다. 이에 따라, 제2 컨볼루션 레이어를 통한 컨볼루션 연산 이후 배치 레이어 관련 연산 과정이 생략될 수 있다.
- [0107] 이상에서는 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치, 로컬 디바이스 및 AI 추론 방법에 대해 설명하였다.
- [0108] 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치의 활용분야는 다음과 같다. 실시간 AI 서비스가 필요한 IoT Device, CCTV, Edge, 자율주행차 등에서 제안하는 방법을 사용하면 실시간 AI 서비스가 가능하다. Cloud의 도움을 받는 로컬 디바이스는 네트워크 상황, Cloud 환경의 상황, Cloud까지의 거리 등 지연시간을 발생시키는 많은 요인이 존재한다. 하지만 AI가 로컬 디바이스에 위치해 있다면 주변 상황에 따른 지연시간이 발생하지 않음으로 실시간 AI 서비스가 가능하다. 제안한 방법은 학습과정에서 Batch의 장점을 유지하면서 Batch관련 연산을 없애고 Learnable Quantization Parameter를 활용하여 최적의 양자화를 시행함으로써 컴퓨팅이 낮은 디바이스에서도 충분히 AI 활용이 가능할 것이다.
- [0109] 본 발명에 따른 특허기술의 사업화 전망과 관련하여, On device AI 관련 반도체 기술들과 더불어 AI 압축 알고리즘의 발전이 가능하다. 이와 관련하여, 맞춤형 AI 칩은 물론 general 칩에서도 AI의 빠른 추론이 가능하게 함으로써 반도체 기술과 함께 발전해 나아가야 한다.
- [0110] 본 발명에 따른 양자화 및 배치 폴딩을 통해 AI 추론을 수행하는 장치는 다음과 같은 기술적 효과를 갖는다.
- [0111] 본 발명의 적어도 일 실시 예에 따르면, AI 추론 시 양자화로 인한 빠른 행렬 연산이 가능하며 Batch Normalization 관련 연산이 불필요하다.
- [0112] 본 발명의 적어도 일 실시 예에 따르면, 기존 양자화 방법보다 더욱 빠른 추론속도를 가질 수 있다. 그렇기 때문에 IoT Device, 스마트폰, 엣지 등과 같이 저 사양 컴퓨팅 능력을 가지며 실시간 서비스를 요구하는 상황에서 해당 방법은 효율적으로 사용될 수 있다.
- [0113] 본 발명의 적어도 일 실시 예에 따르면, On device AI 가 가능하게 하여 사용자의 Privacy를 보존함은 물론 네트워크, 중앙서버 등의 상황에 관계없이 최상의 AI 서비스를 로컬 디바이스에서 제공할 수 있을 것이다.
- [0114] 상술한 본 발명의 특징 및 효과는 첨부된 도면과 관련한 다음의 상세한 설명을 통하여 보다 분명해 질 것이며, 그에 따라 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다.
- [0115] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시 예를 가질 수 있는바, 특정 실시 예들을 도면에 예시하고 상세한 설명에 구체적으로 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0116]

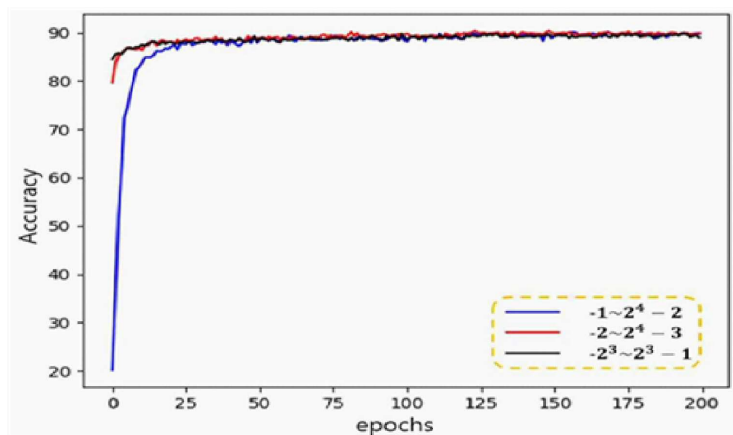
소프트웨어적인 구현에 의하면, 본 명세서에서 설명되는 절차 및 기능뿐만 아니라 각각의 구성 요소들에 대한 설계 및 파라미터 최적화는 별도의 소프트웨어 모듈로도 구현될 수 있다. 적절한 프로그램 언어로 쓰여진 소프트웨어 어플리케이션으로 소프트웨어 코드가 구현될 수 있다. 상기 소프트웨어 코드는 메모리에 저장되고, 제어부(controller) 또는 프로세서(processor)에 의해 실행될 수 있다.

도면

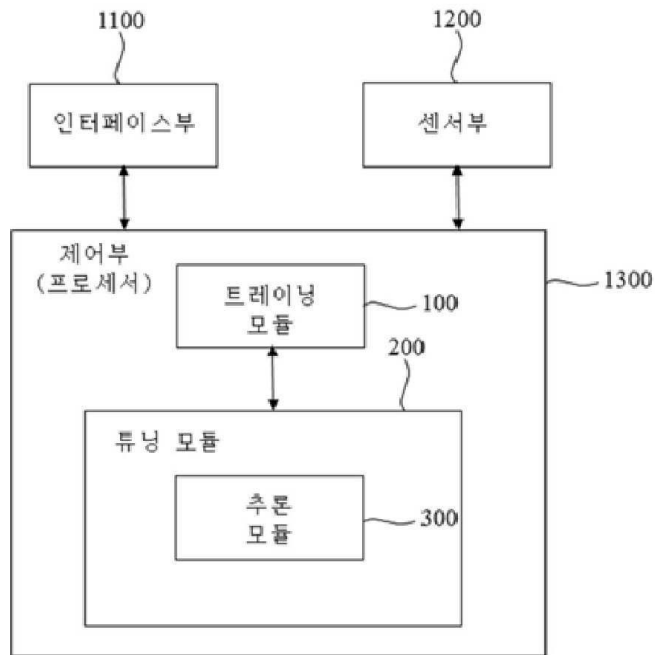
도면1



도면2



도면3



도면4

