



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0018865
(43) 공개일자 2025년02월07일

(51) 국제특허분류(Int. Cl.)
G06N 3/048 (2023.01) G06N 3/063 (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/048 (2023.01)
G06N 3/063 (2023.01)
(21) 출원번호 10-2023-0100039
(22) 출원일자 2023년07월31일
심사청구일자 2023년07월31일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
이슬기
울산광역시 울주군 언양읍 유니스트길 50 (반연리, 울산과학기술원)
이창현
울산광역시 울주군 언양읍 유니스트길 50 (반연리, 울산과학기술원)
(74) 대리인
특허법인 수

전체 청구항 수 : 총 36 항

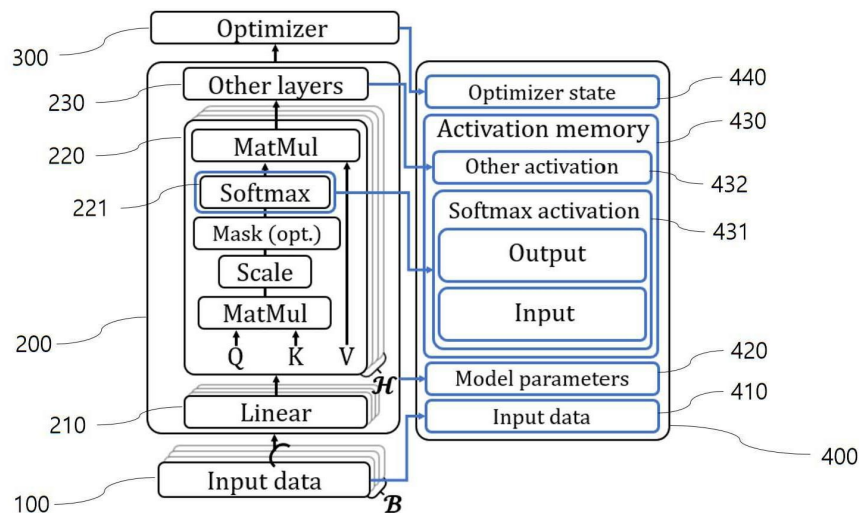
(54) 발명의 명칭 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워크 방법

(57) 요약

본 발명은 소프트맥스를 활용한 인공지능 학습에서 사용되는 메모리를 효율적으로 운용하고, 소프트맥스의 결과를 축소하여 일부 저장함으로써, 활성화에 사용되는 메모리 저장공간을 줄일 수 있는 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워크 방법에 관한 것이다.

본 발명의 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크는 입력 데이터를 수신받아 어텐션 기반의 인공지능 솔루션을 출력하는 어텐션기반 모델, 어텐션기반 모델의 출력을 토대로 인공지능 솔루션을 최적화하는 최적화부, 및 입력 데이터, 어텐션기반 모델, 및 최적화부의 정보를 저장하는 메모리로 이루어진다.

대표도



(52) CPC특허분류

G06N 3/084 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711171702
과제번호	2022R1F1A1063362
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	임베디드 멀티태스크 자연어 처리를 위한 온-디바이스 추론 및 재학습 연구
기 여 율	1/1
과제수행기관명	울산과학기술원
연구기간	2022.06.01 ~ 2023.05.31

명세서

청구범위

청구항 1

입력 데이터를 수신받아 어텐션(attention) 기반의 인공지능 솔루션을 출력하는 어텐션기반 모델;

상기 어텐션기반 모델의 출력을 토대로 상기 인공지능 솔루션을 최적화하는 최적화부; 및

상기 입력 데이터, 상기 어텐션기반 모델, 및 상기 최적화부의 정보를 저장하는 메모리;를 포함하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워크.

청구항 2

제 1항에 있어서,

상기 인공지능 솔루션은 문장 분류(Text Classification), 감성 분석(Sentiment Analysis), 및 기계 번역(Machine Translation) 중 어느 하나를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워크.

청구항 3

제 1항에 있어서,

상기 어텐션기반 모델은 선형 변환을 수행하는 선형계산층;을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워크.

청구항 4

제 3항에 있어서,

상기 어텐션기반 모델은 상기 어텐션을 추출하는 어텐션계산층;을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워크.

청구항 5

제 4항에 있어서,

상기 어텐션계산층은 소프트웨어를 계산하는 소프트웨어 계산모듈;을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워크.

청구항 6

제 5항에 있어서,

상기 어텐션기반 모델은 상기 선형계산층 및 상기 어텐션계산층을 제외한 처리를 수행하는 층인 기타층;을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워크.

청구항 7

제 6항에 있어서,

상기 메모리는 상기 입력 데이터의 정보를 저장하는 입력 데이터 저장구역을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워크.

청구항 8

제 6항에 있어서,

상기 메모리는 상기 어텐션기반 모델의 정보 중 각종 파라미터를 저장하는 모델 파라미터 저장구역을 포함하는

것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 9

제 6항에 있어서,

상기 메모리는 상기 어텐션기반 모델의 정보 중 각종 활성화 정보를 저장하는 활성화 메모리구역을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 10

제 6항에 있어서,

상기 메모리는 상기 최적화부의 정보 중 상태 정보를 저장하는 최적화 상태 저장구역을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 11

제 9항에 있어서,

상기 활성화 메모리구역은 상기 소프트맥스 계산모델의 활성화 정보를 저장하는 소프트맥스 활성화 메모리섹션을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 12

제 9항에 있어서,

상기 활성화 메모리구역은 상기 기타층의 활성화 정보를 저장하는 기타 활성화 메모리섹션을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 13

제 11항에 있어서,

상기 소프트맥스 계산모델은,

순전파과정을 처리하는 순전파과정 서브모듈; 및

역전파과정을 처리하는 역전파과정 서브모듈;을 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 14

제 13항에 있어서,

상기 순전파과정 서브모듈은 소프트맥스 계산을 수행하는 소프트맥스 계산함수를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 15

제 14항에 있어서,

상기 순전파과정 서브모듈은 상기 소프트맥스 계산함수의 출력을 압축하여 압축된 정보를 상기 소프트맥스 활성화 메모리섹션으로 저장하는 압축 파라미터 산출함수를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 16

제 15항에 있어서,

상기 역전파과정 서브모듈은 상기 소프트맥스 활성화 메모리섹션에 저장된 상기 압축한 정보를 토대로 상기 소프트맥스 계산함수의 출력의 추정치를 계산하는 추정치 계산함수를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 17

제 16항에 있어서,

상기 역전과과정 서버모듈은 상기 추정치 계산함수의 출력에 대해 경사행렬(gradient matrix)을 사용하여 상기 소프트맥스 계산함수의 입력을 예측하는 경사행렬 계산함수를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 18

제 17항에 있어서,

상기 역전과과정 서버모듈은 상기 경사행렬 계산함수의 출력을 토대로 상기 어텐션기반 모델 내 변수들의 가중치를 업데이트하는 가중치 업데이트 계산함수를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크.

청구항 19

어텐션기반 모델에서 입력 데이터를 수신받아 어텐션(attention) 기반의 인공지능 솔루션을 출력하는 어텐션기반 모델링단계; 및

최적화부에서 상기 어텐션기반 모델의 출력을 토대로 상기 인공지능 솔루션을 최적화하는 최적화단계;를 포함하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 20

제 19항에 있어서,

상기 인공지능 솔루션은 문장 분류(Text Classification), 감성 분석(Sentiment Analysis), 및 기계 번역(Machine Translation) 중 어느 하나를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 21

제 19항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 어텐션기반 모델 내의 선형계산층에서 선형 변환을 수행하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 22

제 21항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 어텐션기반 모델 내의 어텐션계산층에서 상기 어텐션을 추출하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 23

제 22항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 어텐션계산층 내의 소프트맥스 계산모듈에서 소프트맥스 계산을 수행하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 24

제 23항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 어텐션기반 모델 내의 기타층에서 상기 선형계산층 및 상기 어텐션계산층을 제외한 처리를 수행하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포

합하는 어텐션 기반 네트워킹 방법.

청구항 25

제 24항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 어텐션기반 모델에서 상기 입력 데이터의 정보를 메모리 내의 입력 데이터 저장구역에 저장하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워킹 방법.

청구항 26

제 25항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 어텐션기반 모델에서 상기 어텐션기반 모델의 정보 중 각종 파라미터를 상기 메모리 내의 모델 파라미터 저장구역에 저장하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워킹 방법.

청구항 27

제 24항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 어텐션기반 모델에서 상기 어텐션기반 모델의 정보 중 각종 활성화 정보를 메모리 내의 활성화 메모리구역에 저장하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워킹 방법.

청구항 28

제 25항에 있어서,

상기 최적화단계에서는 상기 최적화부에서 상기 최적화부의 정보 중 상태 정보를 상기 메모리 내의 최적화 상태 저장구역에 저장하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워킹 방법.

청구항 29

제 27항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 소프트웨어 계산모듈에서 상기 소프트웨어 계산모듈의 활성화 정보를 상기 활성화 메모리구역 내의 소프트웨어 활성화 메모리섹션에 저장하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워킹 방법.

청구항 30

제 27항에 있어서,

상기 어텐션기반 모델링단계에서는 상기 기타층에서 상기 기타층의 활성화 정보를 상기 활성화 메모리구역 내의 기타 활성화 메모리섹션에 저장하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워킹 방법.

청구항 31

제 29항에 있어서,

상기 어텐션기반 모델링단계는,

상기 소프트웨어 계산모듈 내의 순전파과정 서브모듈에서 순전파과정을 처리하는 순전파과정 계산단계; 및

상기 소프트웨어 계산모듈 내의 역전파과정 서브모듈에서 역전파과정을 처리하는 역전파과정 계산단계;를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트웨어를 포함하는 어텐션 기반 네트워킹 방법.

청구항 32

제 31항에 있어서,

상기 순전파과정 계산단계에서는 상기 순전파과정 서브모듈 내의 소프트맥스 계산함수에서 소프트맥스 계산을 수행하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 33

제 32항에 있어서,

상기 순전파과정 계산단계에서는 상기 순전파과정 서브모듈 내의 압축 파라미터 산출함수에서 상기 소프트맥스 계산함수의 출력을 압축하여 압축된 정보를 상기 소프트맥스 활성화 메모리섹션으로 저장하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 34

제 33항에 있어서,

상기 역전파과정 계산단계에서는 상기 역전파과정 서브모듈 내의 추정치 계산함수에서 상기 소프트맥스 활성화 메모리섹션에 저장된 상기 압축한 정보를 토대로 상기 소프트맥스 계산함수의 출력의 추정치를 계산하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 35

제 34항에 있어서,

상기 역전파과정 계산단계에서는 상기 역전파과정 서브모듈 내의 경사행렬 계산함수에서 상기 추정치 계산함수의 출력에 대해 경사행렬(*gradient matrix*)을 사용하여 상기 소프트맥스 계산함수의 입력을 예측하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

청구항 36

제 35항에 있어서,

상기 역전파과정 계산단계에서는 상기 역전파과정 서브모듈 내의 가중치 업데이트 계산함수에서 상기 경사행렬 계산함수의 출력을 토대로 상기 어텐션기반 모델 내 변수들의 가중치를 업데이트하는 단계를 포함하는 것을 특징으로 하는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법.

발명의 설명

기술 분야

[0001] 본 발명은 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법에 관한 것으로, 상세하게는, 인공지능 학습에서 사용되는 메모리를 효율적으로 운용하는 것이다.

[0002] 또한, 본 발명은 소프트맥스의 결과를 축소하여 저장함으로써, 활성화에 사용되는 메모리 저장공간을 줄일 수 있는 메모리 효율적(*memory-efficient*) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법에 관한 것이다.

배경 기술

[0003] 딥러닝은 심층신경 네트워크(*DNN*; *Deep Neural Network*) 알고리즘과 이를 학습하는 방법을 의미하는 것으로, 초기에는 시간이 오래 소모되고 학습데이터와의 과적합(*overfitting*)이라는 단점 때문에 활용도가 낮았다.

[0004] 그러나 병렬연산이 가능한 GPU(*Graphic Processing Unit*)의 등장과 새로운 알고리즘의 등장으로 이미지 내 객체 인식 기술이 획기적으로 개선되었고, 현재는 오차율이 굉장히 낮아져 인간의 인지 오차율에 매우 근접한 수준이 되었다.

[0005] 한편, 딥러닝은 소프트맥스 함수(*softmax function*)를 이용하여 기존 연산 결과 값을 0 과 1 사이의 확률 값으로 표현할 수 있는 장점이 있다. 그러나 소프트맥스의 연산은 매우 복잡하고 메모리를 많이 차지하므로 소프트

맥스 계층의 연산을 효율적으로 수행하는 연구가 지속되어 왔다.

[0006] 그 일례로, 대한민국 공개특허공보 제10-2020-0184869호에서는 소프트맥스 계층의 출력값을 도출하기 위한 N개의 입력값을 입력받고, 입력값에 대해 수학식을 이용하여 지수 연산하되, 연산 회로 내 지수(EXP) 블록에 포함된 LUT(Loop Up Table) 데이터의 연속 근사 방식으로 순차적인 지수 연산을 수행하도록 구성된 합성곱 신경망(CNN; Convolutional Neural Network) 알고리즘의 소프트맥스 계층을 위한 연속근사 방식의 연산 방법에 관해 개시하고 있다.

[0007] 그러나, 이 경우에도 소프트맥스의 결과를 축소 저장할 수 있는 방법은 고려하지 않고 있어 메모리를 효율적으로 사용하지 못하는 단점이 있다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 공개특허공보 제10-2020-0184869호 (2020.12.28)

발명의 내용

해결하려는 과제

[0009] 본 발명의 목적은, 소프트맥스를 활용한 인공지능 학습에서 사용되는 메모리를 효율적으로 운용할 수 있는 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법을 제공하는 것이다.

[0010] 본 발명은 소프트맥스의 결과를 축소하여 일부 저장함으로써, 활성화에 사용되는 메모리 저장공간을 줄일 수 있는 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법을 제공하는데 또 다른 목적이 있다.

과제의 해결 수단

[0011] 본 발명에 따른 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크는 입력 데이터를 수신받아 어텐션(attention) 기반의 인공지능 솔루션을 출력하는 어텐션기반 모델, 어텐션기반 모델의 출력을 토대로 인공지능 솔루션을 최적화하는 최적화부, 및 입력 데이터, 어텐션기반 모델, 및 최적화부의 정보를 저장하는 메모리를 포함할 수 있다.

[0012] 여기서, 인공지능 솔루션은 문장 분류(Text Classification), 감성 분석(Sentiment Analysis), 및 기계 번역(Machine Translation) 중 어느 하나를 포함할 수 있다.

[0013] 또한, 어텐션기반 모델은 선형 변환을 수행하는 선형계산층을 포함할 수 있다.

[0014] 여기서, 어텐션기반 모델은 어텐션을 추출하는 어텐션계산층을 포함할 수 있다.

[0015] 또한, 어텐션계산층은 소프트맥스 계산을 수행하는 소프트맥스 계산모듈을 포함할 수 있다.

[0016] 여기서, 어텐션기반 모델은 선형계산층 및 어텐션계산층을 제외한 처리를 수행하는 층인 기타층을 포함할 수 있다.

[0017] 또한, 메모리는 입력 데이터의 정보를 저장하는 입력 데이터 저장구역을 포함할 수 있다.

[0018] 여기서, 메모리는 어텐션기반 모델의 정보 중 각종 파라미터를 저장하는 모델 파라미터 저장구역을 포함할 수 있다.

[0019] 또한, 메모리는 어텐션기반 모델의 정보 중 각종 활성화 정보를 저장하는 활성화 메모리구역을 포함할 수 있다.

[0020] 여기서, 메모리는 최적화부의 정보 중 상태 정보를 저장하는 최적화 상태 저장구역을 포함할 수 있다.

[0021] 또한, 활성화 메모리구역은 소프트맥스 계산모듈의 활성화 정보를 저장하는 소프트맥스 활성화 메모리섹션을 포함할 수 있다.

- [0022] 여기서, 활성화 메모리구역은 기타층의 활성화 정보를 저장하는 기타 활성화 메모리섹션을 포함할 수 있다.
- [0023] 또한, 소프트맥스 계산모듈은 순전파과정을 처리하는 순전파과정 서브모듈 및 역전파과정을 처리하는 역전파과정 서브모듈을 포함할 수 있다.
- [0024] 여기서, 순전파과정 서브모듈은 소프트맥스 계산을 수행하는 소프트맥스 계산함수를 포함할 수 있다.
- [0025] 또한, 순전파과정 서브모듈은 소프트맥스 계산함수의 출력을 압축하여 압축된 정보를 소프트맥스 활성화 메모리 섹션으로 저장하는 압축 파라미터 산출함수를 포함할 수 있다.
- [0026] 여기서, 역전파과정 서브모듈은 소프트맥스 활성화 메모리섹션에 저장된 압축한 정보를 토대로 소프트맥스 계산함수의 출력의 추정치를 계산하는 추정치 계산함수를 포함할 수 있다.
- [0027] 또한, 역전파과정 서브모듈은 추정치 계산함수의 출력에 대해 경사행렬(gradient matrix)을 사용하여 소프트맥스 계산함수의 입력을 예측하는 경사행렬 계산함수를 포함할 수 있다.
- [0028] 한편, 역전파과정 서브모듈은 경사행렬 계산함수의 출력을 토대로 어텐션기반 모델 내 변수들의 가중치를 업데이트하는 가중치 업데이트 계산함수를 포함할 수 있다.
- [0029] 본 발명의 다른 실시예에 따른 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션(attention) 기반 네트워킹 방법은 어텐션기반 모델에서 입력 데이터를 수신받아 어텐션(attention) 기반의 인공지능 솔루션을 출력하는 어텐션기반 모델링단계 및 최적화부에서 어텐션기반 모델의 출력을 토대로 인공지능 솔루션을 최적화하는 최적화단계를 포함할 수 있다.
- [0030] 여기서, 인공지능 솔루션은 문장 분류(Text Classification), 감성 분석(Sentiment Analysis), 및 기계 번역(Machine Translation) 중 어느 하나를 포함할 수 있다.
- [0031] 또한, 어텐션기반 모델링단계에서는 어텐션기반 모델 내의 선형계산층에서 선형 변환을 수행하는 단계를 포함할 수 있다.
- [0032] 여기서, 어텐션기반 모델링단계에서는 어텐션기반 모델 내의 어텐션계산층에서 어텐션을 추출하는 단계를 포함할 수 있다.
- [0033] 또한, 어텐션기반 모델링단계에서는 어텐션계산층 내의 소프트맥스 계산모듈에서 소프트맥스 계산을 수행하는 단계를 포함할 수 있다.
- [0034] 여기서, 어텐션기반 모델링단계에서는 어텐션기반 모델 내의 기타층에서 선형계산층 및 어텐션계산층을 제외한 처리를 수행하는 단계를 포함할 수 있다.
- [0035] 또한, 어텐션기반 모델링단계에서는 어텐션기반 모델에서 입력 데이터의 정보를 메모리 내의 입력 데이터 저장구역에 저장하는 단계를 포함할 수 있다.
- [0036] 여기서, 어텐션기반 모델링단계에서는 어텐션기반 모델에서 어텐션기반 모델의 정보 중 각종 파라미터를 메모리 내의 모델 파라미터 저장구역에 저장하는 단계를 포함할 수 있다.
- [0037] 또한, 어텐션기반 모델링단계에서는 어텐션기반 모델에서 어텐션기반 모델의 정보 중 각종 활성화 정보를 메모리 내의 활성화 메모리구역에 저장하는 단계를 포함할 수 있다.
- [0038] 여기서, 최적화단계에서는 최적화부에서 최적화부의 정보 중 상태 정보를 메모리 내의 최적화 상태 저장구역에 저장하는 단계를 포함할 수 있다.
- [0039] 또한, 어텐션기반 모델링단계에서는 소프트맥스 계산모듈에서 소프트맥스 계산모듈의 활성화 정보를 활성화 메모리구역 내의 소프트맥스 활성화 메모리섹션에 저장하는 단계를 포함할 수 있다.
- [0040] 여기서, 어텐션기반 모델링단계에서는 기타층에서 기타층의 활성화 정보를 활성화 메모리구역 내의 기타 활성화 메모리섹션에 저장하는 단계를 포함할 수 있다.
- [0041] 또한, 어텐션기반 모델링단계는 소프트맥스 계산모듈 내의 순전파과정 서브모듈에서 순전파과정을 처리하는 순전파과정 계산단계 및 소프트맥스 계산모듈 내의 역전파과정 서브모듈에서 역전파과정을 처리하는 역전파과정 계산단계를 포함할 수 있다.
- [0042] 여기서, 순전파과정 계산단계에서는 순전파과정 서브모듈 내의 소프트맥스 계산함수에서 소프트맥스 계산을 수

행하는 단계를 포함할 수 있다.

- [0043] 또한, 순전파과정 계산단계에서는 순전파과정 서브모듈 내의 압축 파라미터 산출함수에서 소프트맥스 계산함수의 출력을 압축하여 압축된 정보를 소프트맥스 활성화 메모리섹션으로 저장하는 단계를 포함할 수 있다.
- [0044] 여기서, 역전파과정 계산단계에서는 역전파과정 서브모듈 내의 추정치 계산함수에서 소프트맥스 활성화 메모리 섹션에 저장된 압축한 정보를 토대로 소프트맥스 계산함수의 출력의 추정치를 계산하는 단계를 포함할 수 있다.
- [0045] 또한, 역전파과정 계산단계에서는 역전파과정 서브모듈 내의 경사행렬 계산함수에서 추정치 계산함수의 출력에 대해 경사행렬(*gradient matrix*)을 사용하여 소프트맥스 계산함수의 입력을 예측하는 단계를 포함할 수 있다.
- [0046] 여기서, 역전파과정 계산단계에서는 역전파과정 서브모듈 내의 가중치 업데이트 계산함수에서 경사행렬 계산함수의 출력을 토대로 어텐션기반 모델 내 변수들의 가중치를 업데이트 하는 단계를 포함할 수 있다.

발명의 효과

- [0047] 본 발명에 의한 메모리 효율적(*memory-efficient*) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법은 소프트맥스를 활용한 인공지능 학습에서 사용되는 메모리를 효율적으로 운용하는 장점이 있다.
- [0048] 또한, 본 발명에 의한 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법은 소프트맥스의 결과를 축소하여 일부 저장함으로써, 활성화에 사용되는 메모리 저장공간을 줄일 수 있는 장점이 있다.

도면의 간단한 설명

- [0049] 도 1은 본 발명의 일 실시예에 따른 메모리 효율적(*memory-efficient*) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크를 나타낸 개략구성도이다.
- 도 2는 도 1의 소프트맥스 계산모듈을 상세히 나타낸 도면이다.
- 도 3은 도 1의 소프트맥스 계산모듈의 기능을 설명하기 위한 도면이다.
- 도 4는 본 발명의 일 실시예에 따른 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법을 나타낸 순서도이다.
- 도 5는 도 4의 어텐션기반 모델링단계를 상세히 나타낸 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0050] 이하, 본 발명의 실시를 위한 구체적인 실시예를 첨부된 도면들을 참조하여 설명한다.
- [0051] 본 발명을 설명함에 있어서 제 1, 제 2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 구성요소들은 용어들에 의해 한정되지 않을 수 있다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제 1 구성요소는 제 2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소도 제 1 구성요소로 명명될 수 있다.
- [0052] 어떤 구성요소가 다른 구성요소에 연결되어 있다거나 접속되어 있다고 언급되는 경우는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해될 수 있다.
- [0053] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.
- [0054] 본 명세서에서, 포함하다 또는 구비하다 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것으로서, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해될 수 있다.
- [0055] 또한, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

- [0057] 이하, 첨부된 도면을 참조하여 본 발명에 따른 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법에 대해 상세히 설명한다.
- [0058] 심층 신경망(DNN; Deep Neural Network)의 학습 시에 메모리를 차지하는 요소는 일반적으로 학습하고자 하는 데이터, 인공지능 모델이 가지고 있는 파라미터(parameter), 최적화 상태의 가중치 업데이트(Optimizer state Weight Update)를 위해 저장되는 상태 정보, 및 활성화 메모리(Activation memory)에 저장되는 요소가 있으며, 활성화 메모리에 저장된 요소는 순전파과정에서 저장되고 역전파과정에서 사용되고 소거된다.
- [0059] 본 발명에 의한 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크는 소프트맥스 출력을 일부의 값만 저장한 후(Forward pass), 출력이 다시 사용되어야 할 때(Backward pass) 저장된 값으로부터 근사된 값들로 저장되지 않은 값들을 대체한다.
- [0060] 여기서, 소프트맥스(softmax)는 어텐션(attention) 메커니즘에서 어텐션 스코어를 구하기 위해서 반드시 포함되는 부분이며, 마찬가지로 어텐션 메커니즘을 사용하는 딥러닝 모델들에서 반드시 포함되는 활성화 함수이다.
- [0061] 또한, 딥러닝 모델이 학습할 때 하드웨어 메모리에서는 여러가지 값들이 저장되게 되는데, 대표적으로 '입력 데이터', '모델 파라미터', '활성화 메모리', '옵티마이저 상태'가 있다.
- [0062] '입력 데이터'는 학습하고자 하는 목표, '모델 파라미터'는 모델을 구성하고 학습의 매체가 되는 요소이며, '활성화 메모리'와 '옵티마이저 상태'는 '모델 파라미터'를 업데이트 하기 전에 사용했던 요소들을 저장하는 값을 의미한다.
- [0063] 이 중 가장 많은 비율을 차지하는 것은 '활성화 메모리'이며, 많은 양의 정보를 학습하고자 하는 모델일수록 더욱 커지게 되며, 이와 같이 반드시 저장 되어야 하는 요소들로 인해 딥러닝 모델들을 학습하기 위해서는 많은 양의 메모리 리소스가 필요하게 된다.
- [0064] 본 발명의 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크에 의하면, 가장 많은 비율을 차지하는 활성화 메모리를 효율화함으로써, 이러한 문제점을 해소할 수 있다.
- [0065] 또한, 본 발명에 의한 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크는 고용량의 '활성화 메모리'를 사용하는 대규모 언어 처리 모델들의 근간이 되는 어텐션 메커니즘에 적합하다.
- [0066] 대규모 언어 처리 모델(Large Language Model)은 많은 양의 정보를 학습하고자 모델의 규모는 더욱 커지고 있으며, 대표적인 예로 GPT-3, MT-NLG, BERT, ALBERT, XLNet 등이 있다.
- [0067] 해당 모델들은 영세한 사용자들이 학습할 수 없을 정도로 거대해지고 있으며, 어텐션(attention) 메커니즘을 근간으로 하기 때문에 소프트맥스 출력의 값으로 인한 활성화 메모리도 많아져 효율화가 필요하다.
- [0068] 즉, 인공지능 솔루션에 따라 활성화 메모리의 사용이 파라미터와 상태 저장에 사용되는 메모리 대비 1.5배 내지 3.5배가 요구되고 있으며, 따라서 활성화 메모리의 사용 용량을 최소화 하는 것이 효율적인 메모리 사용을 위해 중요하다.
- [0069] 본 발명의 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크에 의하면, 이와 같은 활성화에 사용되는 메모리 저장공간을 줄일 수 있다.
- [0070] 이하, 도 1 내지 도 3을 참조하여 본 발명의 일 실시예에 따른 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크를 설명한다.
- [0071] 도 1은 본 발명의 일 실시예에 따른 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크를 나타낸 개략구성도이며, 도 2 및 도 3은 도 1을 상세히 설명하기 위한 세부 도면이다.
- [0072] 먼저, 도 1을 참조하면, 본 발명의 일 실시예에 따른 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크는 입력 데이터(100)를 수신받아 어텐션(attention) 기반의 인공지능 솔루션을 출력하는 어텐션 기반 모델(200), 어텐션 기반 모델(200)의 출력을 토대로 문장 분류(Text Classification), 감성 분석(Sentiment Analysis), 및 기계 번역(Machine Translation) 등과 같은 인공지능 솔루션을 최적화하는 최적화부(300), 및 입력 데이터(100), 어텐션 기반 모델(200), 및 최적화부(300)의 정보를 저장하는 메모리(400)로 이루어진다.
- [0073] 여기서, 어텐션 기반 모델(200)은 선형 변환을 수행하는 선형계산층(210), 어텐션을 추출하는 어텐션계산층(220), 및 선형계산층(210) 및 어텐션계산층(220)을 제외한 처리를 수행하는 층인 기타층(230)으로 이루어진다.

- [0074] 이때, 어텐션계산층(220)은 소프트맥스 계산을 수행하는 소프트맥스 계산모듈(221)을 포함하는데, 소프트맥스 계산모듈(221)에서 순전파과정과 역전파과정이 이루어지며, 도 2 및 도 3에서 순전파과정과 역전파과정을 상세히 설명한다.
- [0075] 한편, 메모리(400)는 입력 데이터(100)의 정보를 저장하는 입력 데이터 저장구역(410), 어텐션기반 모델(200)의 정보 중 각종 파라미터를 저장하는 모델 파라미터 저장구역(420), 어텐션기반 모델(200)의 정보 중 각종 활성화 정보를 저장하는 활성화 메모리구역(430), 및 최적화부(300)의 정보 중 상태 정보를 저장하는 최적화 상태 저장구역(440)으로 이루어지고, 이때, 활성화 메모리구역(430)은 소프트맥스 계산모듈(221)의 활성화 정보를 저장하는 소프트맥스 활성화 메모리섹션(431) 및 기타층(230)의 활성화 정보를 저장하는 기타 활성화 메모리섹션(432)을 포함한다.
- [0076] 본 발명에서는, 소프트맥스 계산모듈(221)에서 순전파과정과 역전파과정에서 사용되는 소프트맥스 활성화 메모리섹션(431)의 저장 용량을 줄일 수 있으며, 이에 대해 아래의 도 2 및 도 3에서 상세히 설명한다.
- [0078] 도 2는 도 1의 소프트맥스 계산모듈(221)을 상세히 나타낸 도면이다.
- [0079] 도 2에서 볼 수 있는 바와 같이, 소프트맥스 계산모듈(221)은 순전파과정(Forward pass)을 처리하는 순전파과정 서브모듈(500) 및 역전파과정(Backward pass)을 처리하는 역전파과정 서브모듈(600)을 포함한다.
- [0080] 여기서, 순전파과정 서브모듈(500)은 소프트맥스 계산을 수행하는 소프트맥스 계산함수(510) 및 소프트맥스 계산함수(510)의 출력을 압축하여 압축된 정보를 소프트맥스 활성화 메모리섹션(431)으로 저장하는 압축 파라미터 산출함수(520)를 포함한다.
- [0081] 한편, 역전파과정 서브모듈(600)은 소프트맥스 활성화 메모리섹션(431)에 저장된 압축한 정보를 토대로 소프트맥스 계산함수(510)의 출력의 추정치를 계산하는 추정치 계산함수(610), 추정치 계산함수(610)의 출력에 경사행렬(gradient matrix)을 적용하여 소프트맥스 계산함수(510)의 입력을 예측하는 경사행렬 계산함수(620), 및 가중치를 업데이트하는 가중치 업데이트 계산함수(630)로 이루어진다.
- [0082] 여기서, 순전파과정에 사용되는 소프트맥스 계산함수(510)는 아래의 [수학식 1]로 정의되며, 보다 확률적이고 정규화스럽게 통일해주어 분류(classification) 문제에서 적용하기 적합하다.

수학식 1

$$s_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}} \text{ for all } i = 1, 2, \dots, n \text{ s.t. } 0 \leq s_i \leq 1 \text{ and } \sum_{i=1}^n s_i = 1$$

- [0084] 여기서, n 은 입력데이터 개수, z_i 는 입력데이터, s_i 는 출력데이터.

- [0086] 또한, 압축 파라미터 산출함수(520)는 아래의 [수학식 2]로 정의되는데, 이때, 압축 파라미터 산출함수(520)의 출력의 일부는 그대로 저장하되, 나머지 부분은 데이터 값을 추정할 수 있는 파라미터 α 와 β 를 추가하는 것이다.

수학식 2

$$\alpha = h_m, \quad \beta = \frac{1}{N} \frac{1}{1-h_m} \left(\ln \frac{h_m}{\ell_m} + 2(h_m - h_{m+1}) \right)$$

- [0087]

[0088] 여기서, h_m 은 저장된 상위값 중 가장 작은값, h_{m+1} 는 저장되지 않은 값들중 가장 큰 값, ℓ_m 은 저장된 하위값 중 가장 큰값, N 은 소프트맥스 출력의 개수.

[0090] 또한, h_m 및 ℓ_m 의 경우 내림차순으로 정렬된 출력들의 저장되지 않은 중간 부분을 근사하기 위해서 중간부분과 맞닿은 상위값 중 최소값, 하위값 중 최대값을 사용한 것이다.

[0091] 한편, 상기 [수학식 2]는 입력의 크기가 데이터마다 다르기 때문에, 제약 조건을 설정하고 제한을 둘 수 있는 특징(Range value), 해당 식으로부터 근사된 값들의 총합이 1로 맞춰지게 제한을 두는 특징(Normalization Factor), 및 h_m 과 h_{m+1} 의 차이를 일부 고려하여 근사되는 값들의 감소 경향을 고려하는 특징(Rate value)이 있다.

[0092] 더욱 상세하게는, 입력의 크기는 데이터 마다 다르기 때문에, 정량적으로 N 의 값을 정수로 가져가버리면 근사화 되는 범위가 입력에 따라서 서로 다르게 커지기 때문에 문제가 발생할 수 있는데, 이를 해결하기 위해서 x축 0의 기울기와 직교하는 기울기를 가진 x까지만 범위로 설정한다.

[0093] 또한, 근사된 값과 기존의 값의 오차를 줄이고 소프트맥스의 기본 성질을 충족시키기 위해서 근사함수로부터 추출되는 값들과 저장되어 있던 값들의 총 합을 1로 맞춰주도록 되어 있다.

[0094] 한편, h_m 과 ℓ_m 만으로 저장되지 않은 중간값들을 모두 근사하기는 어렵기 때문에 h_{m+1} 즉, 저장되지 않은 값 중 가장 큰 값과 저장된 값 중 상위 최소값의 차이를 고려해 근사되어야 할 값들의 경향성을 예측하기 위해 사용한다.

[0095] 한편, 역전파과정에 사용되는 추정치 계산함수(610)는 아래의 [수학식 3]으로 정의되는데, 이때, 압축 파라미터 산출함수(520)의 출력으로 저장된 데이터를 토대로 소프트맥스 계산함수(510)의 출력을 순서대로 추정하는 것으로서, 도 3에서 상세히 설명한다.

수학식 3

$$s'_k = \alpha e^{-\beta \cdot k} \text{ for } k = [0, 1, \dots, N - 1]$$

[0097] 여기서, N 은 추정치데이터 개수 = 입력데이터 개수 - $2m$

[0099] 한편, 지수 분포의 경우 소프트맥스 출력의 내림차순의 변화를 적절히 표현할 수 있고, 기울기가 매개변수로부터 빠르게 변하며, 적은 매개변수를 이용해 효율적으로 계산할 수 있는 장점을 가지고 있다.

[0100] 일반적으로, 지수 분포의 경우 하나의 매개변수만 가지고 있기 때문에 미세 조정이 어려운 특성이 있으나, 본 발명에 의한 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크는 저장된 값들로부터 결정되는 매개변수를 2개의 값, 즉 α 및 β 로 세분화하고 있어 미세 조정이 수월한 장점이 있다.

[0101] 추정치 계산함수(610)의 결과는 아래의 [수학식 4]로 표현되며, 가중치 업데이트를 위한 기초 데이터로서, 경사행렬 계산함수(620)에 제공된다.

수학식 4

$$\vec{s}' = [\underbrace{h_1 \ h_2 \ \dots \ h_m}_{m \text{ highests}} \ \underbrace{s'_1 \ s'_2 \ \dots \ s'_{n-2m}}_{n-2m \text{ approximations}} \ \underbrace{\ell_m \ \ell_{m-1} \ \dots \ \ell_1}_{m \text{ lowests}}]$$

여기서, h 는 최상위의 s_i , ℓ 는 최하위 s_i , m 은 s_i 를 그대로 사용할 개수, n 은 입력데이터 개수, s' 는 추정된 데이터.

상기 [수학식 4]는 소프트맥스 출력의 값들을 크기 순으로 정렬하고 n 개의 출력 중 상위 m 개의 값과 하위 m 개의 값을 선정하는 모습을 보여준다. 이렇게 선정된 $2m$ 개의 값들은 활성화 메모리 즉, 순전파(Forward pass) 과정에서 저장되고 $n-2m$ 개의 출력은 저장하지 않는다.

여기서, m 의 경우 사용자가 원하는 값으로 조정 가능하며, 이는 모델과 데이터의 특성을 고려하여 선정할 수 있다. 이렇게 저장된 값을 역전파(Backward pass) 과정에서 경사 업데이트(Gradient update)를 위해 다시 사용하게 될 때 원래의 값과 비슷한 근사된 값을 얻을 수 있으며, 이에 대해서는 도 3 에서 상세히 설명한다.

또한, 경사행렬 계산함수(620)는 아래의 [수학식 5]로 정의되며, 경사행렬 계산함수(620)의 결과로 어텐션기반 모델(200)에 사용된 변수들의 가중치들이 업데이트된다.

수학식 5

$$\frac{\partial \vec{s}}{\partial \vec{z}_{k-1}} = \begin{bmatrix} s_1(1-s_1) & -s_1s_2 & \cdots & -s_1s_n \\ -s_2s_1 & s_2(1-s_2) & \cdots & -s_2s_n \\ \vdots & \vdots & \ddots & \vdots \\ -s_ns_1 & -s_ns_2 & \cdots & s_n(1-s_n) \end{bmatrix}$$

여기서, s 는 추정된 s_i , z 는 입력 데이터, n 은 입력 데이터 개수.

여기서, 경사행렬 계산함수(620)은 모델의 파라미터를 학습하기 위해서 소프트맥스 출력(output)으로부터 구성되는 기울기 행렬이다. 이는 역전파(Backward pass) 즉, 모델의 파라미터를 업데이트 할 때 사용되며 이 과정에서 순전파(Forward pass)에서 저장해두었던 소프트맥스 출력을 상기 [수학식 5]와 같은 형태로 구성하게 되는데, 이 과정의 수행을 위해 소프트맥스 계산함수(510)의 출력을 저장하는 것이다.

도 3은 도 1의 소프트맥스 계산모듈(221)의 기능을 설명하기 위한 도면이다. 도 3에서 알 수 있는 바와 같이, 추정치 계산함수(610)의 결과인 수학식 4는 압축 파라미터 산출함수(520)의 결과의 일부와 압축 파라미터 산출함수(520)의 결과를 예측한 일부로 표현될 수 있다.

소프트맥스의 기본적인 성질인 출력의 총합은 1이 되는 것인데, 총합이 1이 된다는 것은 특정 값이 커지면 나머지 값들은 상대적으로 작아지게 되는 것을 의미한다.

또한, 어텐션 메커니즘에서는 특정한 입력값들이 커지는 방향성으로 학습하게 되는데, 즉, 여러가지 입력들 중 가장 적절한 연관성을 가진 입력이 높은 값을 가지게 되는 것이다.

- [0116] 이를 기반으로 소프트맥스의 출력을 크기순으로 정렬하게 된다면, 특정 값이 높은 경우 다음으로 큰 값은 상대적으로 작기 때문에 급격한 차이가 생기거나, 어느 하나의 값도 두드러진 크기를 가지지 않는다면 평이한 수준의 차이가 발생하게 된다.
- [0117] 이때, 가장 큰 값과 가장 작은 값을 알고 있다면 그中间的 값들은 어느 정도 일정한 기울기를 가지고 정렬될 수 있고, 또한 특정한 값으로부터 기울기가 급격하게 변할 수 있는 지수 분포(Exponential distribution)로부터 근사화 될 수 있다.
- [0118] 즉, 압축 파라미터 산출함수(520)의 결과의 일부는 최상위 요소(D100) 및 최하위 요소(D300)를 사용하고, 최상위 요소(D100)와 최하위 요소(D300)의 사이는 예측된 요소(D200)로 예측한 값을 사용한다.
- [0119] 이때, 예측된 요소(D200)의 구간 중 압축 파라미터 산출함수(520)의 결과인 소프트맥스 출력(D400)과, 예측된 결과인 예측된 소프트맥스 출력(D500)을 비교해 보면 서로 매우 유사하게 근접함을 알 수 있다.
- [0120] 즉, 본 발명에 의한 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크는 소프트맥스의 결과를 축소하여 일부 저장함으로써, 활성화에 사용되는 메모리 저장공간을 줄일 수 있는 장점이 있다.
- [0122] 도 4는 본 발명의 일 실시예에 따른 메모리 효율적(memory-efficient) 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법을 나타낸 순서도이며, 도 5는 도 4를 상세히 설명하기 위한 세부 순서도이다.
- [0123] 이하, 도 4 내지 도 5를 참조하여 본 발명의 일 실시예에 따른 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법을 설명한다.
- [0124] 먼저, 도 4를 참조하면, 본 발명에 따른 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법은 어텐션기반 모델(200)에서 입력 데이터(100)를 수신받아 어텐션(attention) 기반의 인공지능 솔루션을 출력하는 어텐션기반 모델링단계(S200) 및 최적화부(300)에서 어텐션기반 모델(200)의 출력을 토대로 인공지능 솔루션을 최적화하는 최적화단계(S300)로 이루어진다.
- [0125] 여기서, 인공지능 솔루션은 문장 분류(Text Classification), 감성 분석(Sentiment Analysis), 및 기계 번역(Machine Translation) 중 어느 하나를 포함한다.
- [0126] 또한, 어텐션기반 모델링단계(S200)에서는 어텐션기반 모델(200) 내의 선형계산층(210)에서 선형 변환을 수행하고 어텐션기반 모델(200) 내의 어텐션계산층(220)에서 어텐션을 추출한다. 이때, 어텐션계산층(220) 내의 소프트맥스 계산모듈(221)에서 소프트맥스 계산을 수행하고, 어텐션기반 모델(200) 내의 기타층(230)에서 선형계산층(210) 및 어텐션계산층(220)을 제외한 처리를 수행한다.
- [0127] 여기서, 소프트맥스 계산모듈(221)에서는 순전파과정과 역전파과정이 이루어지며, 도 5에서 순전파과정과 역전파과정을 상세히 설명한다.
- [0128] 한편, 입력 데이터(100), 어텐션기반 모델(200), 및 최적화부(300)에서 사용되는 정보는 메모리(400)에 저장되어 사용되는데, 어텐션기반 모델링단계(S200)에서는 어텐션기반 모델(200)에서 입력 데이터(100)의 정보를 메모리(400) 내의 입력 데이터 저장구역(410)에 저장하고, 어텐션기반 모델(200)에서 어텐션기반 모델(200)의 정보 중 각종 파라미터를 메모리(400) 내의 모델 파라미터 저장구역(420)에 저장하는 한편, 어텐션기반 모델(200)에서 어텐션기반 모델(200)의 정보 중 각종 활성화 정보를 메모리(400) 내의 활성화 메모리구역(430)에 저장한다.
- [0129] 한편, 최적화단계(S300)에서는 최적화부(300)에서 최적화부(300)의 정보 중 상태 정보를 메모리(400) 내의 최적화 상태 저장구역(440)에 저장한다.
- [0130] 더욱 상세하게는, 어텐션기반 모델링단계(S200)에서는 소프트맥스 계산모듈(221)에서 소프트맥스 계산모듈(221)의 활성화 정보를 활성화 메모리구역(430) 내의 소프트맥스 활성화 메모리섹션(431)에 저장하고 기타층(230)에서 기타층(230)의 활성화 정보를 활성화 메모리구역(430) 내의 기타 활성화 메모리섹션(432)에 저장한다.
- [0131] 여기서, 소프트맥스 계산모듈(221)에서 순전파과정과 역전파과정에서 사용되는 소프트맥스 활성화 메모리섹션(431)의 저장 용량을 줄일 수 있으며, 이하 도 5에서 상세히 설명한다.

- [0133] 도 5는 도 4의 어텐션기반 모델링단계(S200)를 상세히 나타낸 순서도이다.
- [0134] 도 5에서 알 수 있는 바와 같이, 어텐션기반 모델링단계(S200)는 소프트맥스 계산모듈(221) 내의 순전파과정 서브모듈(500)에서 순전파과정을 처리하는 순전파과정 계산단계(S500) 및 소프트맥스 계산모듈(221) 내의 역전파과정 서브모듈(600)에서 역전파과정을 처리하는 역전파과정 계산단계(S600)를 포함한다.
- [0135] 여기서, 순전파과정 계산단계(S500)에서는 순전파과정 서브모듈(500) 내의 소프트맥스 계산함수(510)에서 소프트맥스 계산을 수행하고 순전파과정 서브모듈(500) 내의 압축 파라미터 산출함수(520)에서 소프트맥스 계산함수(510)의 출력을 압축하여 압축된 정보를 소프트맥스 활성화 메모리섹션(431)으로 저장한다.
- [0136] 즉, 순전파과정에 사용되는 소프트맥스 계산함수(510)는 상술한 [수학식 1]로 정의되어, 보다 확률적이고 정규화스럽게 통일해주어 분류(classification) 문제에서 적용하기 적합하다.
- [0137] 또한, 압축 파라미터 산출함수(520)는 상술한 [수학식 2]로 정의되는데, 이때, 압축 파라미터 산출함수(520)의 출력의 일부는 그대로 저장하되, 나머지 부분은 데이터 값을 추정할 수 있는 파라미터 α 와 β 를 추가하는 것이며, 도 2 및 도 3에서 이미 설명한 바가 있으므로 상세한 설명은 생략한다.
- [0138] 한편, 역전파과정 계산단계(S600)에서는 역전파과정 서브모듈(600) 내의 추정치 계산함수(610)에서 소프트맥스 활성화 메모리섹션(431)에 저장된 압축한 정보를 토대로 소프트맥스 계산함수(510)의 출력의 추정치를 계산하고, 역전파과정 서브모듈(600) 내의 경사행렬 계산함수(620)에서 추정치 계산함수(610)의 출력에 대해 경사행렬(gradient matrix)을 사용하여 소프트맥스 계산함수(510)의 입력을 예측하는 한편, 역전파과정 서브모듈(600) 내의 가중치 업데이트 계산함수(630)에서 경사행렬 계산함수(620)의 출력을 토대로 어텐션기반 모델(200) 내 변수들의 가중치를 업데이트한다.
- [0139] 즉, 역전파과정에 사용되는 추정치 계산함수(610)는 상술한 [수학식 3]으로 정의되는데, 이때, 압축 파라미터 산출함수(520)의 출력으로 저장된 데이터를 토대로 소프트맥스 계산함수(510)의 출력을 순서대로 추정한다.
- [0140] 여기서, 추정치 계산함수(610)의 결과는 상술한 [수학식4]로 표현되며, 가중치 업데이트를 위한 기초 데이터로서, 경사행렬 계산함수(620)에 제공된다.
- [0141] 또한, 경사행렬 계산함수(620)는 상술한 [수학식 5]로 정의되고, 경사행렬 계산함수(620)의 결과로 어텐션기반 모델(200)에 사용된 변수들의 가중치들이 업데이트되며, 도 2 및 도 3에서 이미 설명하였으므로 상세한 설명은 생략한다.
- [0142] 이와 같은 본 발명에 의한 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워킹 방법은 소프트맥스의 결과를 축소하여 일부 저장함으로써, 활성화에 사용되는 메모리 저장공간을 줄일 수 있는 장점이 있다.
- [0144] 이상과 같이 본 발명에 따른 메모리 효율적 학습을 위한 소프트맥스를 포함하는 어텐션 기반 네트워크 및 네트워킹 방법은 소프트맥스를 활용한 인공지능 학습에서 사용되는 메모리를 효율적으로 운용하는 장점이 있으며, 소프트맥스의 결과를 축소하여 일부 저장함으로써, 활성화에 사용되는 메모리 저장공간을 줄일 수 있는 장점이 있다.
- [0146] 본 발명의 기술 분야에서 통상의 지식을 가진 자는 여기에 개시된 실시예들과 관련하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 프로세서들, 수단들, 회로들 및 알고리즘 단계들이 전자 하드웨어, (편의를 위해, 여기에서 소프트웨어로 지칭되는) 다양한 형태들의 프로그램 또는 설계 코드 또는 이들 모두의 결합에 의해 구현될 수 있다는 것을 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 호환성을 명확하게 설명하기 위해, 다양한 예시적인 컴포넌트들, 블록들, 모듈들, 회로들 및 단계들이 이들의 기능과 관련하여 위에서 일반적으로 설명되었다. 이러한 기능이 하드웨어 또는 소프트웨어로서 구현되는지 여부는 특정한 애플리케이션 및 전체 시스템에 대하여 부과되는 설계 제약들에 따라 좌우된다. 본 발명의 기술 분야에서 통상의 지식을 가진 자는 각각의 특정한 애플리케이션에 대하여 다양한 방식으로 설명된 기능을 구현할 수 있으나, 이러한 구현 결정들은 본 발명의 범위를 벗어나는 것으로 해석되어서는 안 될 것이다.
- [0147] 여기서 제시된 다양한 실시예들은 방법, 장치, 또는 표준 프로그래밍 및/또는 엔지니어링 기술을 사용한 제조

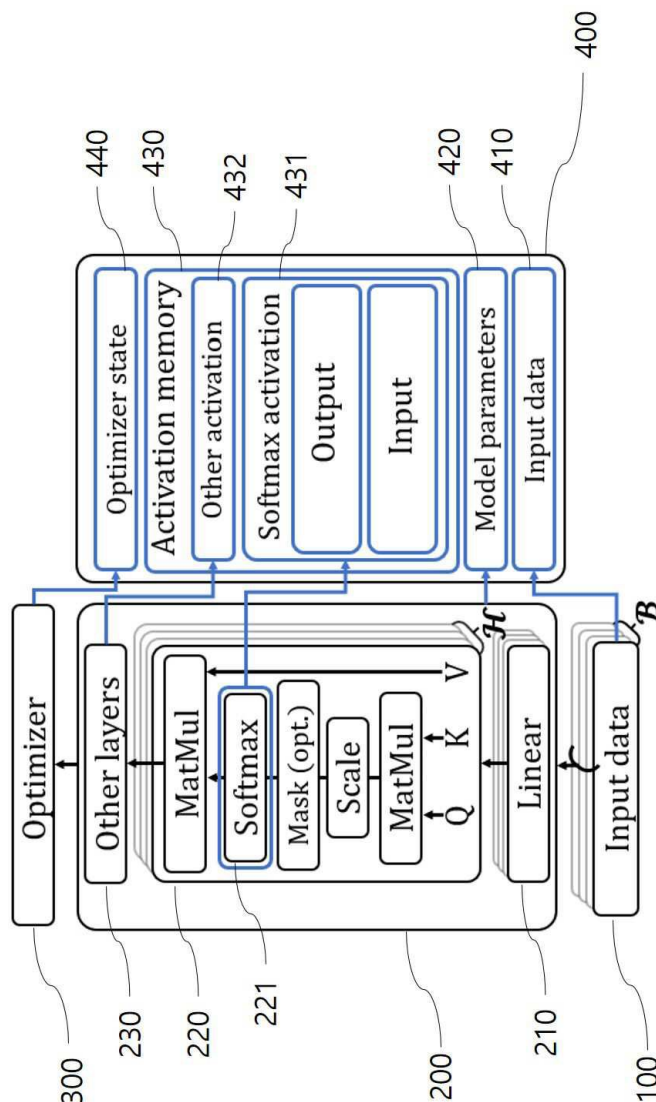
물품(article)으로 구현될 수 있다. 용어 제조 물품은 임의의 컴퓨터-판독가능 저장장치로부터 액세스 가능한 컴퓨터 프로그램, 캐리어, 또는 매체(media)를 포함한다. 예를 들어, 컴퓨터-판독가능 저장매체는 자기 저장 장치(예를 들면, 하드 디스크, 플로피 디스크, 자기 스트림, 등), 광학 디스크(예를 들면, CD, DVD, 등), 스마트 카드, 및 플래시 메모리 장치(예를 들면, EEPROM, 카드, 스틱, 키 드라이브, 등)를 포함하지만, 이들로 제한되는 것은 아니다. 또한, 여기서 제시되는 다양한 저장 매체는 정보를 저장하기 위한 하나 이상의 장치 및/또는 다른 기계-판독가능한 매체를 포함한다.

[0148] 제시된 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조는 예시적인 접근들의 일례임을 이해하도록 한다. 설계 우선순위들에 기반하여, 본 발명의 범위 내에서 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조가 재배열될 수 있다는 것을 이해하도록 한다. 첨부된 방법 청구항들은 샘플 순서로 다양한 단계들의 엘리먼트들을 제공하지만 제시된 특정한 순서 또는 계층 구조에 한정되는 것을 의미하지는 않는다.

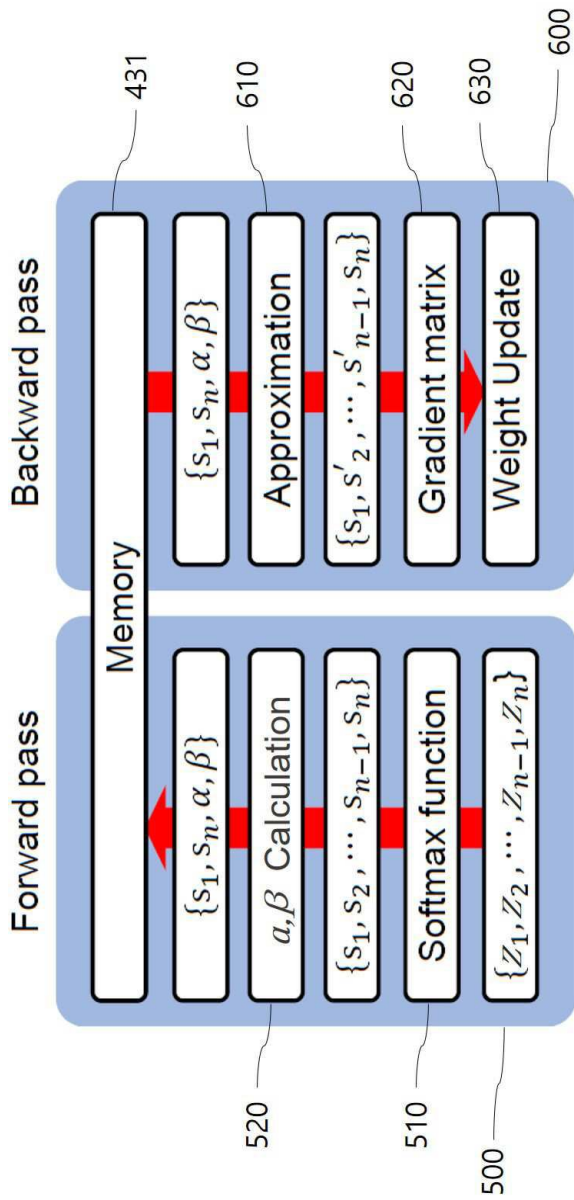
[0149] 제시된 실시예들에 대한 설명은 임의의 본 발명의 기술 분야에서 통상의 지식을 가진 자가 본 발명을 이용하거나 또는 실시할 수 있도록 제공된다. 이러한 실시예들에 대한 다양한 변형들은 본 발명의 기술 분야에서 통상의 지식을 가진 자에게 명백할 것이며, 여기에 정의된 일반적인 원리들은 본 발명의 범위를 벗어남이 없이 다른 실시예들에 적용될 수 있다. 그리하여, 본 발명은 여기에 제시된 실시예들로 한정되는 것이 아니라, 여기에 제시된 원리들 및 신규한 특징들과 일관되는 최광의의 범위에서 해석되어야 할 것이다.

도면

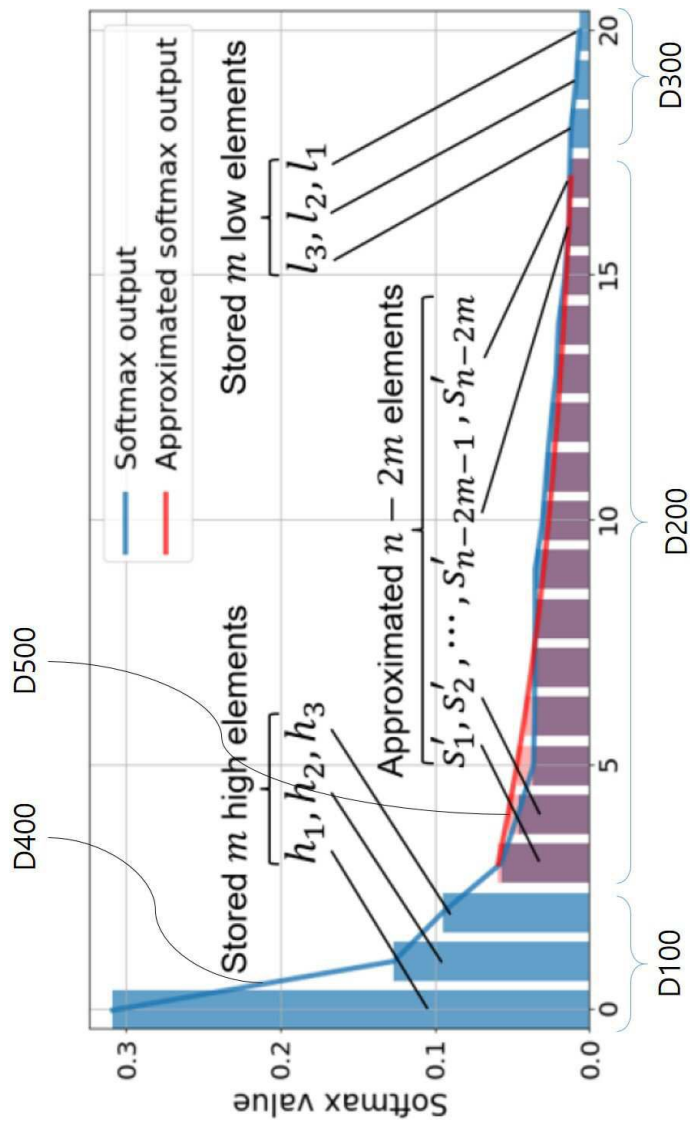
도면1



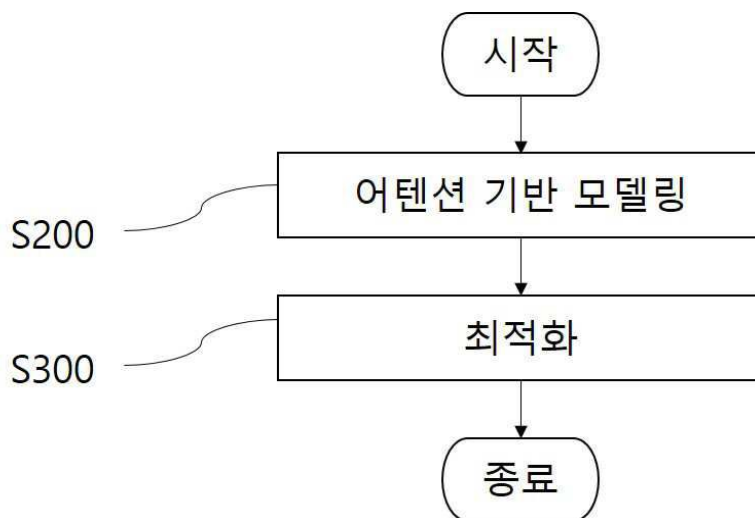
도면2



도면3



도면4



도면5

