(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0149601
(43) 공개일자 2024년10월15일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G06N 3/0455* (2023.01)
G06N 3/096 (2023.01)
(52) CPC특허분류
G06N 3/082 (2023.01)
G06N 3/0455 (2023.01)
(21) 출원번호 10-2023-0045341
(22) 출원일자 2023년04월06일
심사청구일자 2023년04월06일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
최동완
인천광역시 미추홀구 용정공원로 33, 108동 1303호(용현동, 인천 SK Sky VIEW)
임혜민
인천광역시 미추홀구 경인남길118번길 30-1, 원하우스 205호(용현동)
(74) 대리인
양성보

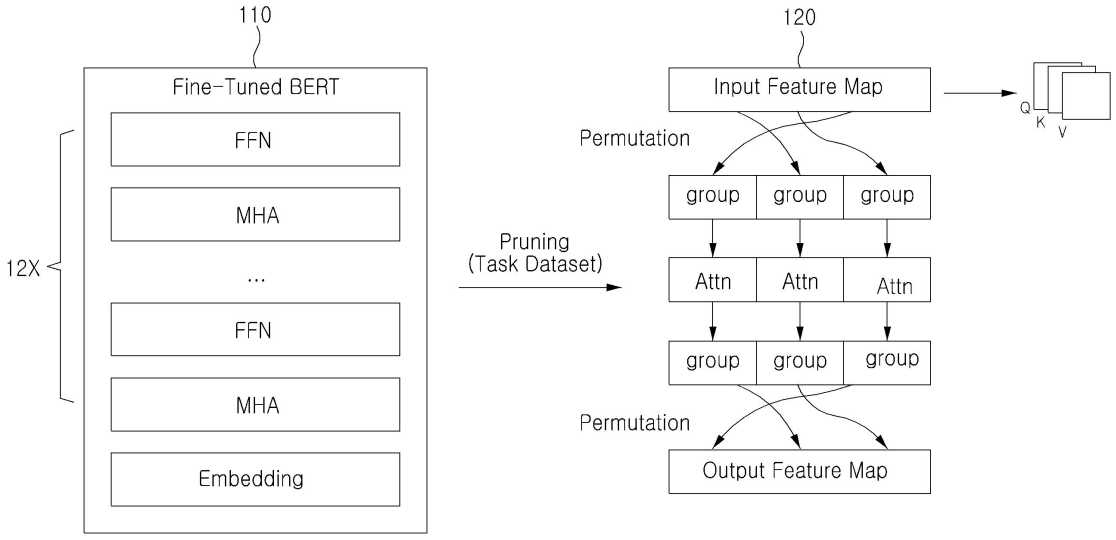
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법 및 시스템**

(57) 요 약

어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법 및 시스템이 제시된다. 본 발명에서 제안하는 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법은 PGB(Permutation Grouped BERT pruning) 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 단계 및 PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계를 포함한다.

대 표 도



(52) CPC특허분류
G06N 3/096 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193695
과제번호 2022-0-00448-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 방송통신산업기술개발사업
연구과제명 [Ezbaro] 인간처럼 회상이 가능한 인공 신경망 지속학습 플랫폼 개발(2차년도)
기 여 율 50/100
과제수행기관명 인하대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711189477
과제번호 2021R1F1A1060160
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 기본연구
연구과제명 [Ezbaro] 질의 가능한 인공지능: 상호작용형 XAI 서비스를 실현하기 위한 질의 기반 딥러닝 모델 해석 기술
기 여 율 40/100
과제수행기관명 인하대학교 산학협력단
연구기간 2023.03.01 ~ 2024.02.29
이 발명을 지원한 국가연구개발사업
과제고유번호 1711179347
과제번호 00155915
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신인재양성(교육운영수익)
연구과제명 [Ezbaro][정부+민간] 인공지능융합혁신인재양성(2차년도)
기 여 율 10/100
과제수행기관명 인하대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

PGB(Permutation Grouped BERT pruning) 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 단계; 및

PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계

를 포함하는 BERT 신경망 경량화 방법.

청구항 2

제1항에 있어서,

상기 PGB 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 단계는,

셀프 어텐션 헤드를 구할 때 사용되는 쿼리(Q), 키(K), 값(V) 중 상기 쿼리(Q) 및 키(K)에 대해서만 순서 변경 및 그룹화 수행하고, 어텐션 헤드의 최종 결괏값은 중 상기 쿼리(Q) 및 키(K)의 스케일링된 내적(scaled dot-product)에 의해 얻어진 스코어 행렬에 상기 값(V)을 곱해주어 어텐션 행렬을 산출하고,

어텐션 헤드를 그룹화할 때, 그룹화 과정을 거친 각 그룹 간에도 정보를 교환할 수 있도록 어텐션 헤드의 입력에 해당하는 어텐션 행렬의 행과 어텐션 헤드의 출력에 해당하는 행렬의 열 각각에 대하여 두 번의 순서 변경을 수행하는

BERT 신경망 경량화 방법.

청구항 3

제2항에 있어서,

상기 PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계는,

상기 산출된 어텐션 행렬에 대하여 미리 정해진 기준 이상의 중요 헤드는 블록 대각선 행렬(block diagonal matrix)로 모으고 상기 블록 대각선 행렬 외부의 헤드에 대해서는 제거하는

BERT 신경망 경량화 방법.

청구항 4

제3항에 있어서,

상기 PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계는,

미리 정해진 기준 이상의 중요 헤드를 결정하기 위해 어텐션 헤드의 입력과 출력에 대한 어텐션 헤드의 가중치 행렬에서 블록 대각선 행렬 외부의 헤드에 대한 가중치를 제거한 후 발생하는 정보 손실이 최소화되도록 하는 순서 변경 쌍을 이용하는 헤드의 순서 변경 규칙을 적용하는

BERT 신경망 경량화 방법.

청구항 5

BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 PGB(Permutation Grouped BERT pruning) 그룹화부; 및

BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 PGB 프루닝부

를 포함하는 BERT 신경망 경량화 시스템.

청구항 6

제5항에 있어서,

상기 PGB 그룹화부는,

셀프 어텐션 헤드를 구할 때 사용되는 쿼리(Q), 키(K), 값(V) 중 상기 쿼리(Q) 및 키(K)에 대해서만 순서 변경 및 그룹화 수행하고, 어텐션 헤드의 최종 결괏값은 중 상기 쿼리(Q) 및 키(K)의 스케일링된 내적(scaled dot-product)에 의해 얻어진 스코어 행렬에 상기 값(V)을 곱해주어 어텐션 행렬을 산출하고,

어텐션 헤드를 그룹화할 때, 그룹화 과정을 거친 각 그룹 간에도 정보를 교환할 수 있도록 어텐션 헤드의 입력에 해당하는 어텐션 행렬의 행과 어텐션 헤드의 출력에 해당하는 행렬의 열 각각에 대하여 두 번의 순서 변경을 수행하는

BERT 신경망 경량화 시스템.

청구항 7

제6항에 있어서,

상기 PGB 프루닝부는,

상기 산출된 어텐션 행렬에 대하여 미리 정해진 기준 이상의 중요 헤드는 블록 대각선 행렬(block diagonal matrix)로 모으고 상기 블록 대각선 행렬 외부의 헤드에 대해서는 제거하는

BERT 신경망 경량화 시스템.

청구항 8

제7항에 있어서,

상기 PGB 프루닝부는,

미리 정해진 기준 이상의 중요 헤드를 결정하기 위해 어텐션 헤드의 입력과 출력에 대한 어텐션 헤드의 가중치 행렬에서 블록 대각선 행렬 외부의 헤드에 대한 가중치를 제거한 후 발생하는 정보 손실이 최소화되도록 하는 순서 변경 쌍을 이용하는 헤드의 순서 변경 규칙을 적용하는

BERT 신경망 경량화 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 현재 트랜스포머(Transformer) 계열의 모델은 자연어 처리, 이미지 인식 등의 여러 인공지능 분야에서 뛰어난 성능을 보이고 있다. 해당 모델들은 다운스트림 태스크에 대해서 더 좋은 성능을 내기 위해 모델의 크기가 점점 커지고 있으며 추론(inference) 시에 상당한 연산량(computational cost)을 필요로 한다는 문제점이 있다. 이로 인해 컴퓨팅 자원이 제한된 상황에서 트랜스포머 계열의 모델을 사용하는 것은 많은 제약이 따른다. 이러한 문제를 해결하고자 트랜스포머 계열 모델의 크기 축소 및 추론 속도 향상을 위한 경량화 연구가 활발히 진행되고 있다. 경량화 기법 중 하나인 프루닝(pruning)은 신경망의 특성에 따라 크게 비구조화된 프루닝(unstructured pruning)과 구조화된 프루닝(structured pruning)으로 나뉘며 트랜스포머 모델에서도 동일하게 적용된다. 비구조화된 프루닝은 모델의 크기를 줄이기 위해 가중치의 중요도 점수를 기반으로 중요하지 않은 파라미터를 제거한 다. 이와 같은 방법으로는 Lottery Ticket 가설, L0 정규화(L0 Regularization), Movement pruning과 같은

방법들이 있으며 해당 방법들은 미세조정된(fine-tuning) 모델의 크기를 줄이는데 뛰어난 효과를 보여줬다. 하지만, 트랜스포머 모델은 표준 하드웨어에서 실행되기 위해 대부분 기존의 조밀한 구조(dense structure)를 재구성해야 되기 때문에 이전 방법론이 효과적이라고 볼 수는 없다. 반면에, 구조화된 프루닝 방법론에서는 중복되는 헤드(head) 또는 레이어(layer) 전체를 제거하거나 모델의 너비(width)와 깊이(depth)를 줄이는 것과 같은 경량화 방법들이 널리 사용되고 있다. 모델 구조의 일부를 제거하는 구조화된 프루닝은 모델의 크기와 연산량을 줄여 컴퓨팅 자원 및 추론 효율성을 높였다. 그러나 구조화된 프루닝 방법은 어텐션의 희소성 비율(sparsity ratio)을 매번 설정해 주어야 한다는 문제점이 존재한다.

선행기술문헌

비특허문헌

[0003] (비특허문헌 0001) [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

(비특허문헌 0002) [2] R. Zhao and W. Luk, "Efficient structured pruning and architecture searching for group convolution," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), pp. 1961-1970, 2019.

발명의 내용

해결하려는 과제

[0004] 본 발명이 이루고자 하는 기술적 과제는 트랜스포머 모델 기반의 사전 학습된(pre-trained) 언어 모델인 BERT에 구조화된 프루닝 기법을 적용하여, 성능 하락에 최소한의 영향을 주며 모델의 크기를 줄이고 추론 속도를 향상시키는 그룹화 방법을 통한 새로운 경량화 방법론인 PGB(Permutation Grouped BERT pruning)를 제안한다. 본 발명의 실시예에 따른 PGB를 통한 BERT 신경망 경량화 방법 및 시스템은 사전 학습된 합성곱 신경망에서 채널의 순서를 변경(permutation)하는 최적의 방법을 찾아 중요한 채널들은 그룹화(grouping)하고 중요하지 않은 채널들은 제거하는 그룹 컨볼루션(GroupConvolution) 방법을 BERT에 적용하는 프루닝 방법을 제안한다.

과제의 해결 수단

[0005] 일 측면에 있어서, 본 발명에서 제안하는 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법은 PGB(Permutation Grouped BERT pruning) 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 단계 및 PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계를 포함한다.

[0006] 상기 PGB 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 단계는 셀프 어텐션 헤드를 구할 때 사용되는 쿼리(Q), 키(K), 값(V) 중 상기 쿼리(Q) 및 키(K)에 대해서만 순서 변경 및 그룹화 수행하고, 어텐션 헤드의 최종 결괏값은 중 상기 쿼리(Q) 및 키(K)의 스케일링된 내적(scaled dot-product)에 의해 얻어진 스코어 행렬에 상기 값(V)을 곱해주어 어텐션 행렬을 산출하고, 어텐션 헤드를 그룹화할 때, 그룹화 과정을 거친 각 그룹 간에도 정보를 교환할 수 있도록 어텐션 헤드의 입력에 해당하는 어텐션 행렬의 행과 어텐션 헤드의 출력에 해당하는 행렬의 열 각각에 대하여 두 번의 순서 변경을 수행한다.

[0007] 상기 PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계는 상기 산출된 어텐션 행렬에 대하여 미리 정해진 기준 이상의 중요 헤드는 블록 대각선 행렬(block diagonal matrix)로 모으고 상기 블록 대각선 행렬 외부의 헤드에 대해서는 제거한다.

[0008] 상기 PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계는 미리 정해진 기준 이상의 중요 헤드를 결정하기 위해 어텐션 헤드의 입력과 출력에 대한 어텐션 헤드의 가중치 행렬에서 블록 대각선 행렬 외부의 헤드

에 대한 가중치를 제거한 후 발생하는 정보 손실이 최소화되도록 하는 순서 변경 쌍을 이용하는 헤드의 순서 변경 규칙을 적용한다.

[0009] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 시스템은 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 PGB(Permutation Grouped BERT pruning) 그룹화부 및 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 PGB 프루닝부를 포함한다.

발명의 효과

[0010] 본 발명의 실시예들에 따르면 트랜스포머 모델에 대한 그룹화 기반의 새로운 구조화된 프루닝 방법인 PGB(Permutation Grouped BERT pruning)를 통해 자원 제약 조건에 따라 최적의 어텐션 순서를 변경하고, 모델의 정보 손실을 최소화하기 위해 헤드의 중요도를 기반으로 불필요한 헤드에 대해 프루닝할 수 있다. 본 발명의 실시예에 따른 PGB를 통한 BERT 신경망 경량화 방법 및 시스템을 통해 단순히 모델의 가중치 또는 일부 구조를 중요도 점수 기반으로 제거하는 것이 아닌 각 헤드의 순서 변경 및 그룹화하여 프루닝하는 방법인 PGB를 통해 대부분의 언어 태스크에서의 성능 하락 및 추론 속도 문제를 해결할 수 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 일 실시예에 따른 PGB의 개념도이다.

도 2는 본 발명의 일 실시예에 따른 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명의 일 실시예에 따른 멀티 헤드 어텐션에서의 순서 변경 규칙 및 그룹화 BERT 방법을 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 시스템의 구성을 나타내는 도면이다.

도 5는 본 발명의 일 실시예에 따른 3 GLUE 데이터셋에 대한 구조화된 프루닝 방법의 모델 크기에 따른 정확도 비교 결과를 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0012] 최근 사전 학습된 트랜스포머 계열의 모델은 자연어 처리, 이미지 인식 등 다양한 인공지능 분야에서 활발히 사용되고 있다. 그러나 해당 모델들은 수십억 개의 파라미터를 가지고 있어 추론 시에 상당한 연산량을 필요로 하며 자원이 제한된 환경에서 사용하기에는 많은 제약이 따른다. 이러한 문제들을 해결 하기 위해 본 발명은 트랜스포머 모델에 대한 그룹화 기반의 새로운 구조화된 프루닝 방법인 PGB(Permutation Grouped BERT pruning)를 제안한다. 제안된 방법은 자원 제약 조건에 따라 최적의 어텐션 순서를 변경하는 방법을 찾고, 모델의 정보 손실을 최소화하기 위해 헤드의 중요도를 기반으로 불필요한 헤드에 대해 프루닝한다. 다양한 비교 실험을 통해 사전 학습된 BERT 모델에 대한 기존의 구조화된 프루닝 방법보다 본 발명에서 제안한 방법이 더 우수한 성능을 보임을 확인한다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0014] 도 1은 본 발명의 일 실시예에 따른 PGB의 개념도이다.

[0015] 도 1은 특정 태스크에 맞게(task-specific) 미세조정 된(Fine Tuned) BERT 모델(110)을 프루닝하기 위한 PGB의 개념도이다. 도 1에서 볼 수 있듯이 PGB는 기존 모델($BERT_{BASE}$)을 단순히 중요도 기반으로 프루닝하는 기존 방법론과는 달리, 최대한 중요한 헤드들이 제거되지 않도록 헤드의 순서를 변경 및 그룹화(permutation grouping)하여 프루닝하는 PGB 기법(120)을 적용한다. 이를 통해 중요한 헤드가 제거되는 문제를 해결하며 중요하지 않은 헤드들은 제거하여 성능의 큰 변화 없이 효과적으로 프루닝할 수 있다.

[0016] 다중 레이어로 구성된 트랜스포머는 각 레이어마다 여러 개의 헤드와 피드포워드 네트워크(FeedForward Network)로 구성되어 있다. 하나의 헤드는 셀프 헤드 어텐션(Self-Head Attention)이라 불리며, 이는 식(1)과

같이 $Q(query)$와 $K(key)-V(value)$ 쌍을 매핑(mapping)함으로써 형성될 수 있다. 식 (1)의 d_k는 트랜스포머 모델을 구성하는 $K(key)$의 차원을 의미한다.

[0017] $$Attn(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

[0018] 트랜스포머 모델은 인코더와 디코더로 구성되어 있으며 핵심 구성요소인 멀티 헤드 어텐션(Multi-Head Attention)은 모델이 다른 위치의 표현 공간 정보에도 집중할 수 있도록 하며, 식(3)과 같이 여러 개의 $head$로 병렬 구성된다. 각 레이어의 H는 헤드의 개수를 나타내며 Q, K, V와 결괏값(output)의 가중치 행렬 크기는 식(2)와 같이 나타낼 수 있다.

[0019] $$[W_i^Q, W_i^K] \in R^{d \times d_k}$$
$$W_i^V \in R^{d \times d_v}$$
$$W^O \in R^{Hd_v \times d} \quad (2)$$

[0020] $$MHA(Q,K,V)=concat(h_1,...,h_H)W^O$$
$$where\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

[0021] 피드포워드 네트워크(FeedForward Network; FFN)는 식(4)와 같이 나타낼 수 있다. 피드포워드 네트워크의 파라미터는 W_1과 W_2 두 개의 가중치 행렬로 구성되며 크기는 $[W_1, W_2] \in R^{d \times d_{ff}}$와 같다. d와 $d_{ff}(=4 \times d)$는 각각 은닉 유닛 크기(hidden state size)와 피드포워드 네크워크(FFN) 안에 있는 중간 레이어(intermediate layer)의 뉴런의 크기를 의미한다.

[0022] $$FFN(x)=\max(0, xW_1+b_1)W_2+b_2 \quad (4)$$

[0023] 신경망 경량화 연구는 많은 파라미터를 가진 신경망 내에서 불필요한 정보를 제거하여 컴퓨팅 자원이 제한된 환경에서 딥러닝 모델이 사용될 수 있도록 하였다. 대표적인 경량화 연구로는 모델의 파라미터를 비트 단위로 표현하여 메모리를 줄이고 계산 속도를 높이는 양자화(quantization) 방법, 기존에 사전 학습된 대형 신경망(teacher)에서 상대적으로 더 작은 신경망(student)으로 지식을 전이하는 지식 증류 방법(knowledge distillation), 사전 학습된 모델에서 불필요한 가중치나 레이어 등을 제거하는 프루닝 방법(pruning)이 있다. 이와 같이 사전 학습된 대형 신경망을 압축하는데 위의 경량화 방법들을 적용한 연구들이 활발히 진행되고 있다. 또 다른 경량화 방법으로는 신경망의 입력 채널을 그룹으로 나누거나 필터 집합을 그룹으로 나눠서 모델의 파라미터와 연산량을 줄이는 GConv(Group Convolution) 방법이 있다. GConv로 신경망을 구성하는 방법으로는 from scratch로 학습된 모델 또는 사전 학습된 모델을 프루닝하는 방법이 있다. 또한, 사전 학습된 신경망 모델의 채널에 대해 순서를 다르게 하는 최적의 방법(permutation)을 찾아 채널을 그룹화하여 중 요하지 않은 채널을 프루닝하는 방법이 있다.

[0024] 프루닝(Pruning)은 신경망에서 중요하지 않은 요소를 제거하여 신경망의 크기를 줄이는 방법이다. 신경망 프루닝과 마찬가지로 트랜스포머 기반의 자연어 처리 모델 프루닝 기법도 크게 비구조화된 프루닝과 구조화된 프루닝으로 나눌 수 있다. 비구조화된 프루닝은 기존 심층 신경망(deep neural network)에서 사용된 방법처럼 모델의 크기에 따라 프루닝할 가중치를 선택(sparsity) 또는 중요도 점수를 기반으로 모델의 불필요한 가중치를 제거한다. 이 방법은 모델의 크기를 상당히 줄일 수 있지만, 추론 속도를 향상시키기 위해 비구조화된 희소 행렬(sparse matrix)을 위한 전문화된 하드웨어 장비(예를 들어, GPU)가 요구된다. 구조화된 프루닝은 모델의 특정 구조의 가중치 그룹 전체를 제거하는 방법으로, 트랜스포머 기반의 모델에서는 주로 어텐션 헤드(attention head), 레이어(layer), 은닉 유닛(hidden state) 등의 구조에 대해 프루닝한다. 이와 관련된 종래기술에서는 일부 헤드들이 성능 저하 없이 제거될 수 있으며, 중복된(redundancy) 특성을 가지는 헤드들이 존재할 수 있다.

[0025] 최근 구조화된 프루닝 기법은 효율성을 최대화하기 위해 모델 구조의 세분(granularity)에 따라 셀프 어텐션(self-attention) 또는 피드포워드 네트워크의 레이어, 헤드 등에 대해 통합적으로 프루닝하거나 모델의 너비와

깊이를 선택하여 동적으로 크기를 조정한 후 성능 복구 방식으로 지식 증류 방법을 사용하는 프루닝 방법 등이 있다.

[0026] 본 발명에서는 단순히 모델의 가중치(weight) 또는 일부 구조를 중요도 점수(importance score) 기반으로 제거하는 것이 아닌 각 헤드의 순서 변경 및 그룹화하여 프루닝하는 방법인 PGB를 통해 대부분의 언어 태스크(language task)에서의 성능 하락 및 추론 속도 문제를 해결한다. 또한 PGB는 기존의 GroupCNN pruning 기법을 복잡한 구조를 가진 트랜스포머에 맞게 변형한 것으로 이전에 시도된 적이 없는 새로운 방법이다.

[0028] 도 2는 본 발명의 일 실시예에 따른 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법을 설명하기 위한 흐름도이다.

[0029] 본 발명의 실시예에 따른 PGB(Permutation Grouped BERT pruning)는 합성곱 신경망을 그룹화하여 압축하는 GConv 방법을 BERT 모델에 적용하여 새로운 구조화된 프루닝 방법론을 제안한다. 일반 합성곱 레이어(convolution layer)로 쌓여있는 신경망과 비교했을 때, BERT 모델은 트랜스포머의 인코더 레이어로 구성되어 있어 합성곱 신경망 보다 훨씬 더 복잡한 계산방식을 가진다. 본 발명에서는 GConv 프루닝 방법론을 트랜스포머의 복잡한 계산 방식에 맞게 변형 하여 적용하였다.

[0030] 제안하는 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 방법은 PGB 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 단계(210) 및 PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계(220)를 포함한다.

[0031] 단계(210)에서, PGB 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행한다.

[0032] 본 발명의 실시예에 따른 구조화된 프루닝 방법의 목표는 기존 모델($BERT_{BASE}$)이 가지고 있는 정보의 손실을 최소화하며 모델의 크기 감소 및 추론 속도 향상의 효과를 내도록 하는 것이다. 도 1은 PGB 방법론의 핵심 아이디어인 멀티 헤드 어텐션의 순서 변경(permutation) 및 그룹화(grouping)를 적용하여 프루닝하는 방법에 대해 간략하게 나타낸다. 상기 식(1)과 같이 셀프 어텐션 헤드를 구할 때 사용되는 쿼리(Q), 키(K), 값(V) 중 Q와 K에 대해서만 순서 변경 및 그룹화 방법을 적용하였다. 어텐션 헤드의 최종 결괏값은 Q, K의 스케일링된 내적(scaled dot-product)에 의해 얻어진 스코어 행렬에 V를 곱해주어 어텐션 행렬을 산출하기 때문에 V는 기존과 동일하게 계산해 주었다. PGB가 어텐션 헤드에 대해 순서 변경 및 그룹화를 적용하는 방식은 다음과 같다.

[0033] $$\widetilde{X}_c = X_{\tau_{in}(c)} \quad \widetilde{MHA}_g = W_{gh} * \widetilde{X}_g \quad MHA_r = \widetilde{MHA}_{\tau_{out}(r)} \tag{5}$$

[0034] 식(5)에서 X와 o는 각각 어텐션 헤드의 입력 피쳐맵과 어텐션 헤드에 대한 출력 피쳐맵을 의미하고, 어텐션 행렬을 구하기 위한 가중치의 크기는 식(2)와 같다. τ_{in}과 τ_{out}은 각각 어텐션 헤드의 순서 변경을 위한 어텐션 행렬의 인덱스 축을 의미하고, c(열)와 r(행)은 각각 d(은닉 유닛 크기)와 헤드의 수($N_h = 12; H$)로 나타낸다. 어텐션 헤드를 그룹화할 때, 그룹화 과정을 거친 각 그룹 간에도 정보를 교환할 수 있도록 어텐션 헤드의 입력과 출력에 대해 두 번의 순서 변경 규칙 방식을 적용한다. 먼저, 행 축에 해당하는 τ_{in}에 의해 순서가 변경된 $\widetilde{X}$를 헤드의 수에 따라 G개의 그룹($\widetilde{X} \in R^{d_g}$)으로 분할한다. 멀티 헤드 어텐션은 이미 N_h개의 단일 헤드 그룹으로 나뉘어 있다고 볼 수 있기 때문에 본 발명에서는 이를 고려하여 각 그룹(g)에 해당하는 헤드 (h)에 대해서 식(6)과 같이 Q, K를 구한다.

[0035] $$Q_{gh}, K_{gh} = \sum_g X_g W_{gh},$$
$$\widetilde{MHA}_{W_{gh}^Q, W_{gh}^K, W^V, W_g^O}(\widetilde{X}) = \sum_g Att^{gh}_{W_{gh}^Q, W_{gh}^K, W_h^V, W_{gh}^O}(\widetilde{X}) \tag{6}$$

[0036] 이때 각 그룹의 가중치 행렬의 크기는 $W_{gh} \in R^{d_g \times d_h}$과 같으며 d_h는 $d_h = d_g \times \frac{N_h}{g}$와 같이 표현할 수

있다. 최종적으로 아웃풋에 해당하는 멀티 헤드 어텐션을 구할 때 τ_{out}에 의해서 각 그룹의 어텐션의 순서를 한 번 더 변경하여 특 별한 아키텍처 설계 없이 그룹 간의 정보를 교환할 수 있게 된다(식(6)의 두 번째 줄).

[0037] 단계(220)에서, PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거한다.

[0038] 본 발명의 실시예에 따른 어텐션 헤드 프루닝(attention head pruning)은 모델 내 불필요한 헤드를 프루닝의 기준에 따라 제거함으로써 모델의 크기 및 연산량을 감소시킨다. 본 발명에서는 식(7)과 같이 중요한 헤드들을 블록 대 각선 행렬(block diagonal matrix)로 모으고 대각선 외부에 대해서는 바로 제거하는 방식으로 접근하였다.

$$W_{gh} = \begin{vmatrix} W_{11} & 0 & \cdots & 0 \\ 0 & W_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_{GH} \end{vmatrix} \begin{vmatrix} X_1 \\ X_2 \\ \vdots \\ X_G \end{vmatrix} \Rightarrow g \times \frac{N_h}{g}$$

[0039] (7)

[0040] 하지만 이런 단순한 방법은 대각선 행렬 외부의 헤드들이 중요하지 않다는 것을 보장하기 어렵다는 문제점이 발생한 다. 이에 PGB는 입력과 출력에 대해서 헤드의 순서 변경 (permutation) 규칙을 적용하여 중요한 헤드들이 최대한 블록 대각선 행렬로 이동할 수 있도록 하는 최적의 헤드 순서 변경 규칙을 찾는다. 순서 변경 규칙은 가중치에 대해 한 쌍의 순서 변경에 대한 인덱스$(\tau_{in},\ \tau_{out})$를 적용한 식(8)과 같은 방법을 통해 얻을 수 있다.

$$\begin{vmatrix} \omega_{11}\ \omega_{12}\ \omega_{13} \\ \omega_{21}\ \omega_{22}\ \omega_{23} \\ \omega_{31}\ \omega_{32}\ \omega_{33} \end{vmatrix} \xrightarrow[\tau_{in}=\{1,3,2\}]{\tau_{out}=\{2,1,3\}} \begin{vmatrix} \omega_{21}\ \omega_{23}\ \omega_{22} \\ \omega_{11}\ \omega_{13}\ \omega_{12} \\ \omega_{31}\ \omega_{33}\ \omega_{32} \end{vmatrix}$$

[0041] (8)

[0042] 식(8)의 ω_{ij}는 어텐션 헤드의 가중치 행렬을 의미하고, 행 과 열은 각각 d와 헤드의 수로 나타낸다. 이때 순서 변경 규칙 적용 시 발생할 수 있는 정보 손실을 최소화하기 위해 식(9)와 같이 공식화한다. 식(9)를 통해 기존 어텐션 가중치 행렬에서 블록 대각선 행렬 외부의 가중치를 제거한 후 발생하는 정보 손실이 최소화될 수 있도록 하는 순서 변경 쌍을 수 있다.

$$\arg\min_{\tau_{in}^{*}\ \tau_{out}^{*}} \sum_{L}\sum_{h}(W) - \sum_{L}\sum_{g}(diag(\widetilde{W_{gh}}))$$

[0043] (9)

[0045] 도 3은 본 발명의 일 실시예에 따른 멀티 헤드 어텐션에서의 순서 변경 규칙 및 그룹화 BERT 방법을 설명하기 위한 도면이다.

[0046] PGB의 그룹화 프루닝 방식은 다음과 같다.

$$Groupwise\ Att(Q,K,V,M) = \begin{vmatrix} X_1 \bullet softmax(\frac{Q_{11}K_{11}^{T}}{\sqrt{d_g}})V_1 \\ \vdots \\ X_g \bullet softmax(\frac{Q_{gh}K_{gh}^{T}}{\sqrt{d_g}})V_H \end{vmatrix}$$

[0047]

[0048] 도 3과 같이 중요한 헤드들을 그룹화하여 블록 대각선 행렬로 모은다. 이때, 어텐션 헤드에 대해 순서 변경 규칙을 적용했을 때 W_{gh}의 행렬이 0이 아닌 값들로 구성된다면 해당하는 어텐션 헤드들을 하나의 그룹으로 묶어준다. 그룹에 속하지 못한 중요하지 않은 헤드들에 대해서는 마스크를 씌워 제거한다. 마스크 변수는 정규화 방식을 적용하여 0 또는 1의 값을 갖도록 학습 하였다. 추가적으로 멀티 헤드 어텐션 레이어는 추론 시에 상당한

연산량이 요구된다는 문제점을 가지고 있다. 본 발명에서는 학습 시에 특정 어텐션 레이어에 중요한 헤드가 없다면 모델의 추론 속도 증가를 위해 어텐션 레이어가 제거될 수 있도록 하였다.

[0050] 도 4는 본 발명의 일 실시예에 따른 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 시스템의 구성을 나타내는 도면이다.

[0051] 본 실시예에 따른 BERT 신경망 경량화 시스템(400)은 프로세서(410), 버스(420), 네트워크 인터페이스(430), 메모리(440) 및 데이터베이스(450)를 포함할 수 있다. 메모리(440)는 운영체제(441) 및 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 루틴(442)을 포함할 수 있다. 프로세서(410)는 PGB 그룹화부(411) 및 PGB 프루닝부(412)를 포함할 수 있다. 다른 실시예들에서 BERT 신경망 경량화 시스템(400)은 도 4의 구성요소들보다 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 예를 들어, BERT 신경망 경량화 시스템(400)은 디스플레이나 트랜시버(transceiver)와 같은 다른 구성요소들을 포함할 수도 있다.

[0052] 메모리(440)는 컴퓨터에서 판독 가능한 기록 매체로서, RAM(random access memory), ROM(read only memory) 및 디스크 드라이브와 같은 비소멸성 대용량 기록장치(permanent mass storage device)를 포함할 수 있다. 또한, 메모리(440)에는 운영체제(441)와 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화 루틴(442)을 위한 프로그램 코드가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 드라이브 메커니즘(drive mechanism, 미도시)을 이용하여 메모리(440)와는 별도의 컴퓨터에서 판독 가능한 기록 매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록 매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록 매체(미도시)를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록 매체가 아닌 네트워크 인터페이스(430)를 통해 메모리(440)에 로딩될 수도 있다.

[0053] 버스(420)는 BERT 신경망 경량화 시스템(400)의 구성요소들간의 통신 및 데이터 전송을 가능하게 할 수 있다. 버스(420)는 고속 시리얼 버스(high-speed serial bus), 병렬 버스(parallel bus), SAN(Storage Area Network) 및/또는 다른 적절한 통신 기술을 이용하여 구성될 수 있다.

[0054] 네트워크 인터페이스(430)는 BERT 신경망 경량화 시스템(400)을 컴퓨터 네트워크에 연결하기 위한 컴퓨터 하드웨어 구성요소일 수 있다. 네트워크 인터페이스(430)는 BERT 신경망 경량화 시스템(400)을 무선 또는 유선 커넥션을 통해 컴퓨터 네트워크에 연결시킬 수 있다.

[0055] 데이터베이스(450)는 어텐션 헤드 그룹화를 통한 BERT 신경망 경량화를 위해 필요한 모든 정보를 저장 및 유지하는 역할을 할 수 있다. 도 4에서는 BERT 신경망 경량화 시스템(400)의 내부에 데이터베이스(450)를 구축하여 포함하는 것으로 도시하고 있으나, 이에 한정되는 것은 아니며 시스템 구현 방식이나 환경 등에 따라 생략될 수 있고 혹은 전체 또는 일부의 데이터베이스가 별개의 다른 시스템 상에 구축된 외부 데이터베이스로서 존재하는 것 또한 가능하다.

[0056] 프로세서(410)는 기본적인 산술, 로직 및 BERT 신경망 경량화 시스템(400)의 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(440) 또는 네트워크 인터페이스(430)에 의해, 그리고 버스(420)를 통해 프로세서(410)로 제공될 수 있다. 프로세서(410)는 PGB 그룹화부(411) 및 PGB 프루닝부(412)를 위한 프로그램 코드를 실행하도록 구성될 수 있다. 이러한 프로그램 코드는 메모리(440)와 같은 기록 장치에 저장될 수 있다.

[0057] PGB 그룹화부(411) 및 PGB 프루닝부(412)는 도 2의 단계들(210~220)을 수행하기 위해 구성될 수 있다.

[0058] BERT 신경망 경량화 시스템(400)은 PGB 그룹화부(411) 및 PGB 프루닝부(412)를 포함할 수 있다.

[0059] PGB 그룹화부(411)는 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행한다.

[0060] PGB 그룹화부(411)는 셀프 어텐션 헤드를 구할 때 사용되는 쿼리(Q), 키(K), 값(V) 중 상기 쿼리(Q) 및 키(K)에 대해서만 순서 변경 및 그룹화 수행하고, 어텐션 헤드의 최종 결괏값은 중 상기 쿼리(Q) 및 키(K)의 스케일링된 내적(scaled dot-product)에 의해 얻어진 스코어 행렬에 상기 값(V)을 곱해주어 어텐션 행렬을 산출한다.

[0061] PGB 그룹화부(411)는 어텐션 헤드를 그룹화할 때, 그룹화 과정을 거친 각 그룹 간에도 정보를 교환할 수 있도록 어텐션 헤드의 입력에 해당하는 어텐션 행렬의 행과 어텐션 헤드의 출력에 해당하는 행렬의 열 각각에 대하여

두 번의 순서 변경을 수행한다.

[0062] PGB 프루닝부(412)는 BERT 모델의 크기 및 연산량을 감소시키기 위해 상기 순서 변경 및 그룹화된 BERT 모델 내의 어텐션 헤드를 프루닝의 기준에 따라 제거한다.

[0063] PGB 프루닝부(412)는 상기 산출된 어텐션 행렬에 대하여 미리 정해진 기준 이상의 중요 헤드는 블록 대각선 행렬(block diagonal matrix)로 모으고 상기 블록 대각선 행렬 외부의 헤드에 대해서는 제거한다.

[0064] PGB 프루닝부(412)는 미리 정해진 기준 이상의 중요 헤드를 결정하기 위해 어텐션 헤드의 입력과 출력에 대한 어텐션 헤드의 가중치 행렬에서 블록 대각선 행렬 외부의 헤드에 대한 가중치를 제거한 후 발생하는 정보 손실이 최소화되도록 하는 순서 변경 쌍을 이용하는 헤드의 순서 변경 규칙을 적용한다.

[0065] 본 발명에서는 순서 변경 규칙(permutation)에 따라 어텐션 헤드를 그룹화하여 BERT 모델을 경량화하는 PGB의 성능을 확인하기 위해 사전 학습된 언어 모델에서 주로 사용되는 GLUE(General Language Understanding Evlauation) 벤치마크를 사용하여 실험하였다. GLUE 벤치마크 데이터셋에 포함된 9개 중 5개의 데이터셋(SST-2, QQP, RTE, CoLA, MRPC)을 사용하였다. 사용한 데이터들은 다음과 같이 세 가지의 카테고리로 분류할 수 있다:

[0066] 1. 단일 문장 태스크(Single Sentence Task): Stanford Sentiment Treebank(SST-2)와 Corpus of Linguistic Acceptability(CoLA).

[0067] 2. 자연어 추론(Natural Language Inference): 텍스트 관련 인식(RTE)

[0068] 3. 문장 쌍의 유사도(Sentence Pair Similarity)와 환언(Paraphrase): 퀴라(Quora) 질문 쌍(QQP)과 Microsoft Research Paraphrase Corpus(MRPC).

[0069] 실험에서 사용된 데이터셋의 크기 등의 태스크에 대한 자세한 사항은 표 1과 같다.

[0070] <표 1>

[0071]

Dataset	QQP	RTE	SST-2	CoLA	MRPC
#Train	364k	2.5k	67k	8.5k	3.7k
#Dev	40k	276	872	1k	408
#Test	391k	3k	1.8k	1k	1.7k
Epochs	3	3	3	3	3
Batch Size	32	32	32	32	32
Learning Rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Optimizer	Adam with $\epsilon=1\times10^{-8}$				
Eval Metric	Acc.	Acc.	Acc.	Mattew's corr.	Acc.

[0072] PGB를 구현하기 위해 Python 3.7.15, PyTorch 1.13.1, CUDA 11.6 버전을 사용하였으며, 구현 코드는 최신 자연어 처리 기술에 해당하는 Huggingface의 트랜스포머 라이브러리를 기반으로 하였다. 모든 실험에는 NVIDIA GeForce RTX 3090 24G(GPU)와 NVIDIA Quadro RTX 6000 24G(GPU)를 사용했다. 실험에 사용한 모델은 트랜스포머의 Encoder만으로 구성된 사전 학습된 언어 모델인 $BERT_{BASE}$를 사용하였으며 GLUE 벤치마크 데이터셋으로 모델을 학습시켰다. $BERT_{BASE}$는 12개의 인코더 레이어로 구성되어 있으며 은닉 유닛의 크기(hidden state size;d)는 768이다. $BERT_{BASE}$의 각 인코더 레이어는 12개의 어텐션 헤드로 구성되며, 인코더의 피드포워드 네트워크는 3072개의 뉴런$(d_{ff}=4\times d)$으로 구성되어 있다. 본 발명의 실험은 모델의 임베딩 레이어(Embedding layer)의 가중치를 고정시킨 상태에서 진행 되었다.

[0073] 모델 압축 방법의 성능 평가 지표로 압축률뿐만 아니라 추론 속도 또한 주된 평가 지표라고 할 수 있다. 본 발명에서는 프루닝하지 않은 $BERT_{BASE}$ 모델을 기본으로 하여 기존의 구조화된 프루닝 방법들과 PGB의 성능을 비교하였다. 비교군으로는 사전 학습된 언어 모델의 너비(width)와 깊이(depth)를 선택하여 프루닝하는 DynaBERT, coarse-grained(e.g. layer)와 fine-grained(e.g. 어텐션 헤드, 은닉 유닛) 모듈에 대해 동시에 프루닝하는 CoFi방법론을 채택하였다. 정확한 결과 비교를 위해 DynaBERT와 CoFi 프루닝 방식의 코드를 수정하여 피드포워드 네트워크(FeedForward Network)의 프루닝을 제외한 어텐션 헤드 구조에 대한 프루닝만 진행하였다. 공정한 비교를 위해 각 방법론마다 추가적으로 성능을 높이기 위해 사용된 데이터 증강(data augmentation)을 제외하여 실험한 결과에 대해 비교 하였다. 또한 공정한 성능 비교를 위해 추론 속도를 측정하는데 NVIDIA GeForce RTX 3090 24G(GPU)를 사용하였다

[0074] 도 5는 본 발명의 일 실시예에 따른 3 GLUE 데이터셋에 대한 구조화된 프루닝 방법의 모델 크기에 따른 정확도 비교 결과를 나타내는 그래프이다.

[0075] 도 5(a)는 3 GLUE 데이터셋(SST-2)에 대한 구조화된 프루닝 방법의 모델 크기에 따른 정확도이고, 도 5(b)는 3 GLUE 데이터셋(RTE)에 대한 구조화된 프루닝 방법의 모델 크기에 따른 정확도이다.

[0076] 사전 학습된 대형 언어 모델인 $BERT_{BASE}$에 프루닝 기법을 적용한 결과에 대한 성능 비교를 위해 이전 방법들 과의 비교 실험을 진행하였다. 표 2는 GLUE 벤치 데이터셋을 사용하여 특정 태스크에 맞게(task-specific) 미세조정된 BERT 모델에 본 발명에서 제안한 PGB 방법론과 다른 구조화된 프루닝 방법을 적용하여 정확도 및 추론 속도 변화율에 대한 실험 결과를 나타낸다. 각 방법은 특정 태스크에 대해서 미세조정된 BERT 모델을 경량화하는 구조화된 프루닝 방법이다. 표 2에서 확인할 수 있듯이 본 논문에서 제안하는 PGB는 다른 방법들과 비교했을 때 가장 낮은 성능 하락을 보인다.

[0077] <표 2>

[0078]

$Pruning_{MHA} \setminus Task$		SST-2 (67k)	QQP (364k)	RTE (2.5k)	CoLA (8.5k)	MRPC (3.7k)
	#Params	Acc.	Acc.	Acc.	Matthews' corr.	Acc.
$BERT_{BASE}$	85M	93.1	91.2	70	56.2	85
CoFi	64M	89.6	89.3	54	50.5	74.7
speedup		1.3x	1.2x	1.7x	2.0x	1.0x
DynaBERT	65M	90.8	89.8	59.5	50.0	82.3
speedup		1.0x	1.0x	1.0x	1.0x	1.0x
PGB(ours)	63M	**91.1**	**90.4**	**61.7**	**52.3**	80.3
speedup		**3.2x**	**2.5x**	**3.5x**	**3.6x**	**2.8x**

[0079] 또한, 추론 속도 변화율에서도 $BERT_{BASE}$ 및 이전 방법들보다 2~3 배 정도 증가한 것을 확인할 수 있다. 이러한 결과가 나타나는 이유는 CoFi 방법론은 중요도에 따라 레이어와 헤드에 대해서 제거하지만 매번 모델의 희소성 비율을 지정하고, DynaBERT는 제거할 멀티 헤드 어텐션의 크기(너비, 깊이)를 지정해야 하기 때문이다. 도 5는 SST-2, RTE 데이터셋에 대해 PGB와 다른 프루닝 방법들의 모델 크기 변화에 따른 성능 비교를 측정한 것이다. 이와 같이, 본 발명에서 PGB 방법론은 이전 구조화된 프루닝 방법들보다 모델 크기 변화에 따른 성능 하락이 적은 것을 확인할 수 있다.

[0081] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0082] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0083] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판

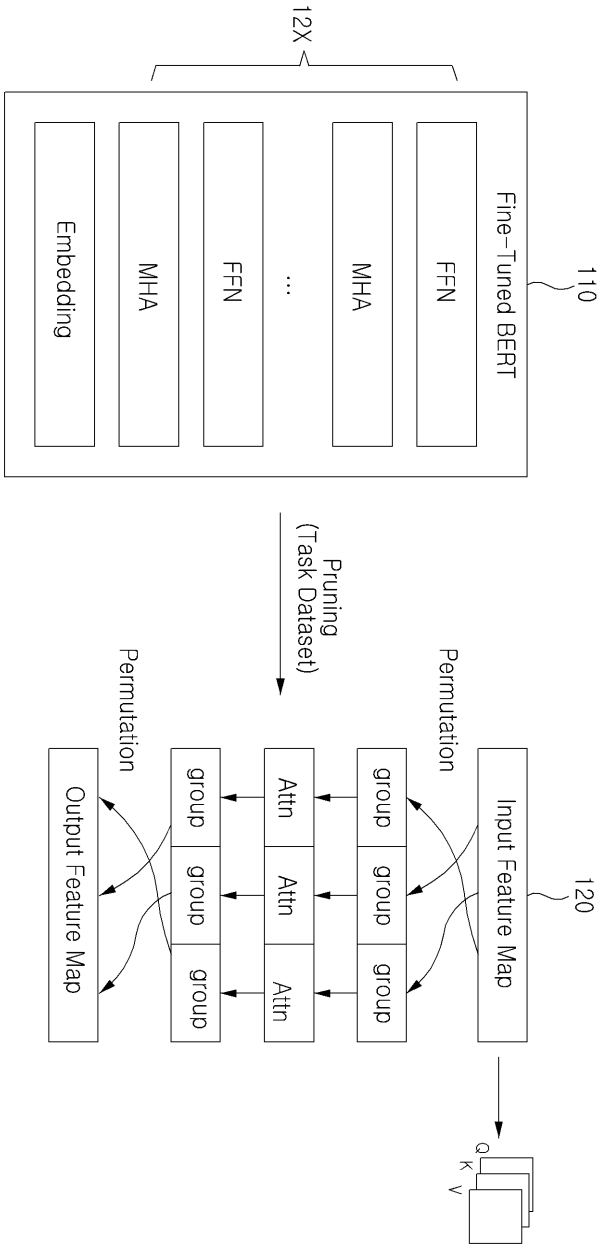
독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical 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.

[0084] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

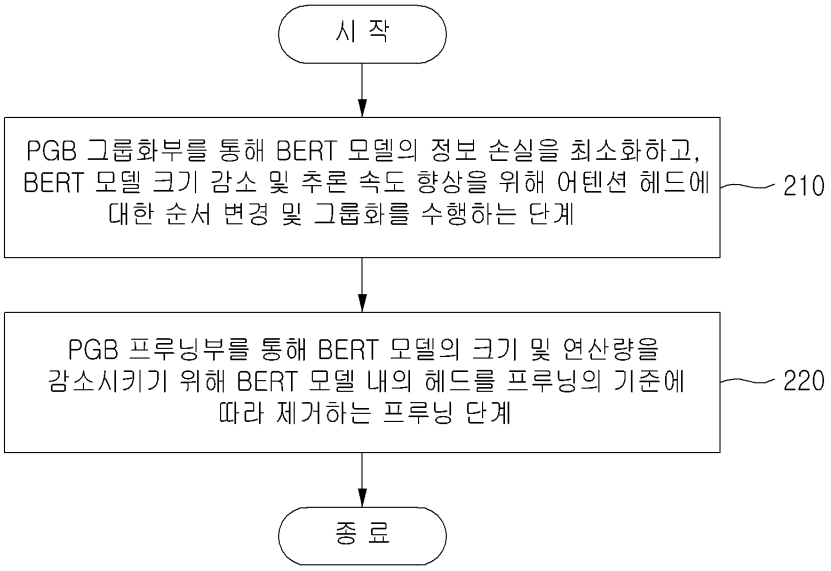
[0085] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

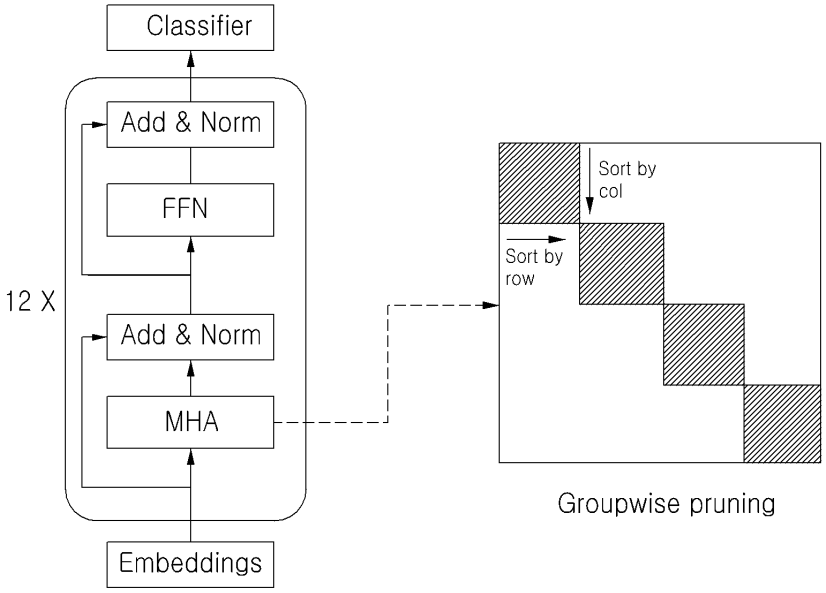
도면1


12X
Embedding
MHA
FFN
...
MHA
FFN
Fine-Tuned BERT
110
Pruning
(Task Dataset)
Permutation
Output Feature Map
group
group
group
Attn
Attn
Attn
group
group
group
Permutation
Input Feature Map
120
Q
K
V

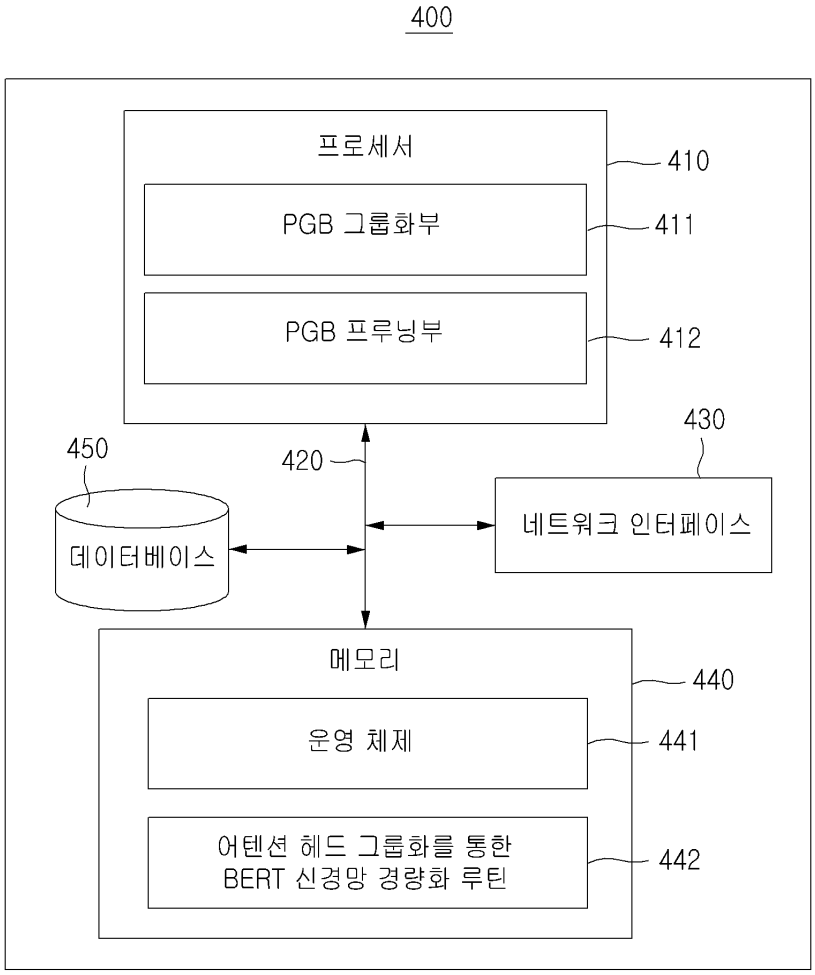
도면2


시 작
PGB 그룹화부를 통해 BERT 모델의 정보 손실을 최소화하고, BERT 모델 크기 감소 및 추론 속도 향상을 위해 어텐션 헤드에 대한 순서 변경 및 그룹화를 수행하는 단계
210
PGB 프루닝부를 통해 BERT 모델의 크기 및 연산량을 감소시키기 위해 BERT 모델 내의 헤드를 프루닝의 기준에 따라 제거하는 프루닝 단계
220
종 료

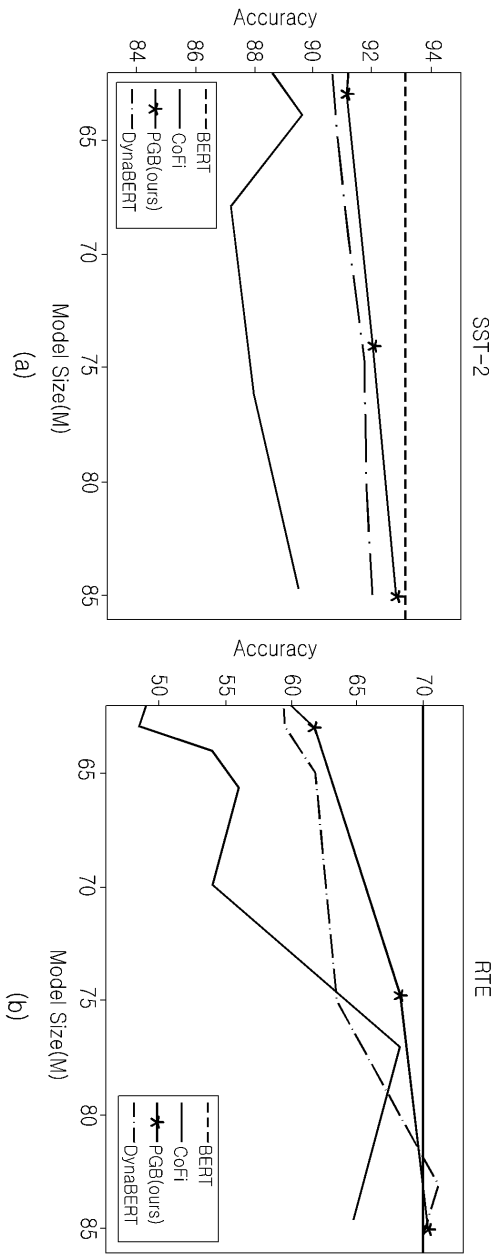
도면3


Classifier
Add & Norm
FFN
12 X
Add & Norm
MHA
Embeddings
Sort by col
Sort by row
Groupwise pruning

도면4


400
프로세서
410
PGB 그룹화부
411
PGB 프루닝부
412
430
450
420
데이터베이스
네트워크 인터페이스
메모리
440
운영 체제
441
어텐션 헤드 그룹화를 통한
BERT 신경망 경량화 루틴
442

도면5


Accuracy
SST-2
Model Size(M)
(a)
BERT
CoFi
PGB(ours)
DynaBERT
Accuracy
RTE
Model Size(M)
(b)
BERT
CoFi
PGB(ours)
DynaBERT