



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0066745
(43) 공개일자 2024년05월16일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) G06F 40/40 (2020.01)
G06N 7/00 (2023.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
G06F 40/40 (2020.01)
(21) 출원번호 10-2022-0147917
(22) 출원일자 2022년11월08일
심사청구일자 없음

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
박성배
서울특별시 송파구 오금로 396, 1402호 (가락동, 에스케이허브파크)
김봉민
경기도 용인시 기흥구 서그대로53번길 3-17, 304호 (서천동, 서천빌리지)
(뒷면에 계속)
(74) 대리인
유미특허법인

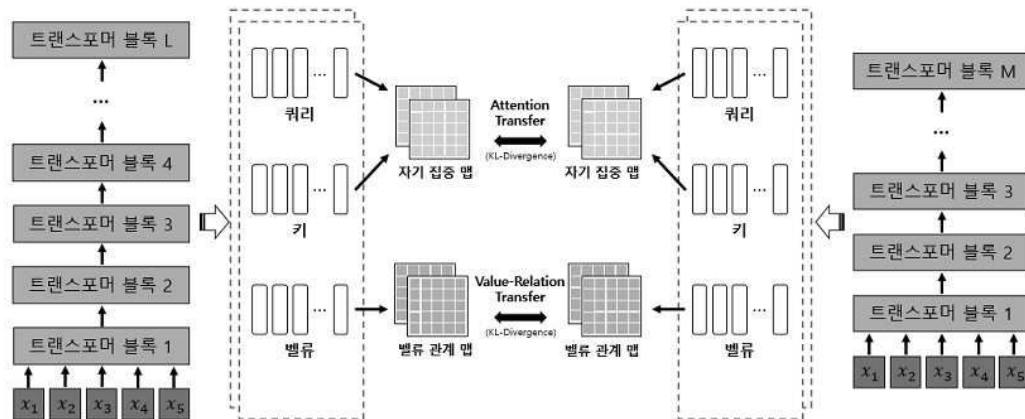
전체 청구항 수 : 총 1 항

(54) 발명의 명칭 한국어 대용량 언어 모델 경량화 기법

(57) 요약

한국어 대용량 언어 모델 경량화 기법이 개시된다. 이 기법에서, 교사 모델과 학생 모델은 왼쪽과 오른쪽으로 나뉘지며 트랜스포머 아키텍처로 구성되고, 각각의 트랜스포머 층에서 쿼리, 키 그리고 밸류 값을 얻을 수 있으며, 쿼리-키 값을 통해 정방향렬로 구성된 자기 집중 맵을 만들고, 각 층에서의 밸류 값을 이용하여 자기 내적을 통해 밸류 관계 맵을 얻으며, 교사 모델과 학생 모델의 트랜스포머 각 층에서 자기 집중 맵 및 밸류 관계 맵을 얻은 후, 학생 모델이 교사 모델의 맵 분포를 닮도록 학습하

대표도 - 도1



(52) CPC특허분류

G06N 7/01 (2023.01)

(72) 발명자

김소연

경기도 용인시 기흥구 서천동로21번길 6-1, 301호
(서천동)

홍상훈

경기도 용인시 수지구 성북1로163번길 11, 205동
101호 (성북동, 버들치마을 경남아너스빌2차)

홍충선

경기 용인시 수지구 상현로 30-10, 233동 101호 (상현동, 상현마을성원상떼빌아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711162037
과제번호	2020R1A4A1018607
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	기초연구실지원사업
연구과제명	Meta Federated Learning 기반 이동엣지 컴퓨팅시스템 핵심구조 개발
기 여 율	1/2
과제수행기관명	경희대학교(국제캠퍼스)
연구기간	2022.03.01~2023.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711174034
과제번호	RS-2022-00155911
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	인공지능융합혁신인재양성(R&D)
연구과제명	인공지능융합혁신인재양성(경희대학교)
기 여 율	1/2
과제수행기관명	경희대학교(국제캠퍼스)
연구기간	2022.07.01~2022.12.31

명세서

청구범위

청구항 1

교사 모델과 학생 모델은 왼쪽과 오른쪽으로 나뉘지며 트랜스포머 아키텍처로 구성되고, 각각의 트랜스포머 층에서 쿼리, 키 그리고 벨류 값을 얻을 수 있으며, 쿼리-키 값을 통해 정방향렬로 구성된 자기 집중 맵을 만들고, 각 층에서의 벨류 값을 이용하여 자기 내적을 통해 벨류 관계 맵을 얻으며, 교사 모델과 학생 모델의 트랜스포머 각 층에서 자기 집중 맵 및 벨류 관계 맵을 얻은 후, 학생 모델이 교사 모델의 맵 분포를 닮도록 학습하는,

한국어 대용량 언어 모델 경량화 기법.

발명의 설명

기술 분야

[0001] 본 발명은 한국어 대용량 언어 모델 경량화 기법에 관한 것이다.

배경 기술

[0002] 언어 모델이란 단어 시퀀스에 확률을 할당하는 모델로 단어들의 조합이 얼마나 적절한지, 또는 해당 문장이 얼마나 적합하지를 알려주는 일을 한다. 최근에는 매우 많은 파라미터를 갖는 언어 모델을 구축하여 다양한 과제를 해결하는 데 사용되고 있다. 하지만 많은 수의 파라미터를 갖는 대용량 언어 모델의 경우 학습 시 필요로 하는 컴퓨터 메모리 자원이 크며, 추론 속도 또한 느려지는 단점이 존재한다. 그렇기에 대용량 언어 모델을 경량화하여 작은 파라미터를 갖지만 대용량 언어 모델과 비슷한 성능을 낼 수 있는 연구가 주목받고 있다. 공개된 한국어 경량화 모델은 단순히 작은 언어 모델(이하 학생 모델)이 대용량 언어 모델(이하 교사 모델)의 예측 출력만을 모방하도록 학습된다. 하지만 이러한 학습 방식은 어떤 과제를 풀기 위한 교사 모델의 출력만을 학생 모델이 모방하기에 근본적으로 언어 자체의 문맥화 능력을 학습한 교사 모델의 지식을 추출하기 어렵다. 또한 풀어야 하는 과제가 하나로 고정되지 않고 변경될 수 있는 환경에서는 기존의 지식 증류 방법 적용 시 예측 출력을 증류 받기 위한 아키텍처를 계속해 수정해줘야 하는 어려움이 존재한다.

발명의 내용

해결하려는 과제

[0003] 본 발명은 한국어 대용량 언어 모델을 어떻게 경량화할 것인가에 대한 것으로, 대용량 언어 모델의 성능을 유지하면서 작은 모델 파라미터 수를 갖게 하는 방법을 제안한다.

과제의 해결 수단

[0004] 본 발명은 교사 모델이 최종 예측 출력을 내기 위해 내부적으로 언어를 이해하는 과정을 학생 모델에게 증류하는 방법을 제안한다. 언어 모델은 내부적으로 트랜스포머 아키텍처로 구성되어 입력 문장을 임베딩해 쿼리, 키 그리고 벨류 3가지 구성 요소로 나누어 사용한다. 쿼리와 키를 사용해 자기 집중 맵을 만들어 입력된 문장에서 어떤 단어에 집중할지를 학습하며, 벨류를 통해 해당 단어의 중요도를 높인다. 언어 모델은 이런 내부 동작을 통해 최종 예측 응답을 출력하며, 본 발명에서는 교사 모델의 쿼리와 키를 사용해 만든 자기 집중 맵 및 벨류 관계 맵을 만들어 학생 모델이 이를 모방하도록 학습하여 교사 모델의 문맥화 능력을 그대로 학습할 수 있게 한다.

발명의 효과

[0005] 본 발명은 대용량 언어 모델의 성능은 유지하면서 낮은 파라미터 수를 갖는 한국어 경량화 언어 모델 공개함으로써, 언어 모델을 활용한 서비스를 제공하는 기업 또는 기관에서 이를 사용할 수 있다.

[0006] 대용량 언어 모델과 비슷한 성능을 유지하면서 낮은 파라미터 수를 갖기 때문에 추론 시 상대적으로 낮은 컴퓨

터 메모리를 요구하며 빠른 추론 속도를 보장할 수 있다.

도면의 간단한 설명

도 1은 한국어 대용량 언어 모델을 경량화 하기위한 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[그림 1]은 한국어 대용량 언어 모델을 경량화 하기위한 방법을 나타낸다. 교사 모델과 학생 모델은 [그림 1]의 왼쪽과 오른쪽으로 나뉘지며 트랜스포머 아키텍처로 구성된다. 는 입력 문장의 단어를 의미하여 벡터로 임베딩 되어 트랜스포머 층의 입력으로 들어간다. 각각의 트랜스포머 층에서 쿼리, 키 그리고 벨류 값을 얻을 수 있으며, 쿼리-키 값을 통해 정방향렬로 구성된 자기 집중 맵을 만든다. 또한 각 층에서의 벨류 값을 이용하여 자기 내적을 통해 벨류 관계 맵을 얻을 수 있다. 교사 모델과 학생 모델의 트랜스포머 각 층에서 자기 집중 맵 및 벨류 관계 맵을 얻은 후, 학생 모델이 교사 모델의 맵 분포를 닮도록 학습한다(Kullback Leibler). 이와 같은 학습 방법은 기존 지식 증류 방법처럼 교사 모델의 예측 출력을 모방하도록 학습하지 않기에 아키텍처 구성의 자유도가 높으며, 언어 자체를 학습하는 과정을 모방하여 지식 증류의 근본적인 문제를 해결할 수 있다.

[표 1]은 각 과제(NSMC, Naver NER, PAWS, KorNLI, KorSTS, Question Pair, KorQuaD)에 대해 대용량 언어 모델 및 경량화 언어 모델의 성능을 나타낸다. #Param은 모델의 파라미터 개수를 의미하며 KoMiniLM이 본 발명에서 학습한 한국어 경량화 언어 모델이다. 상단의 KoBERT(KLUE) 모델을 교사 모델로 사용하여 학생 모델인 KoMiniLM을 학습하였고, KoMiniLM은 교사 모델의 약 98~99%의 성능을 회복하였다. 표의 DistilKoBERT부터 경량화 언어 모델이며, 제안 모델의 성능은 모든 과제에 대해 DistilKoBERT 보다 높고 몇 대용량 언어 모델에 비해서도 높은 성능을 보여준다.

표 1

Model	#Param	Average	NSMC (Acc)	Naver NER (F1)	PAWS (Acc)	KorNLI (Acc)	KorSTS (Spearman)	Question Pair (Acc)	KorQuaD (EM/F1)
KoBERT(KLUE)	110M	86.84	90.20±0.07	87.11±0.05	81.36±0.21	81.06±0.33	82.47±0.14	95.03±0.44	84.43±0.18 / 93.05±0.04
KcBERT	108M	78.94	89.60±0.10	84.34±0.13	67.02±0.42	74.17±0.52	76.57±0.51	93.97±0.27	60.87±0.27 / 85.01±0.14
KoBERT(SKT)	92M	79.73	89.28±0.42	87.54±0.04	80.93±0.91	78.18±0.45	75.98±2.81	94.37±0.31	51.94±0.60 / 79.69±0.66
DistilKoBERT	28M	74.73	88.39±0.08	84.22±0.01	61.74±0.45	70.22±0.14	72.11±0.27	92.65±0.16	52.52±0.48 / 76.00±0.71
KoMiniLM*	68M	85.90	89.84±0.02	85.98±0.09	80.78±0.30	79.28±0.17	81.00±0.07	94.89±0.37	83.27±0.08 / 92.08±0.06
KoMiniLM*	23M	84.79	89.67±0.03	84.79±0.09	78.67±0.45	78.10±0.07	78.90±0.11	94.81±0.12	82.11±0.42 / 91.21±0.29

교사 모델이 언어를 이해하는 동작 과정을 학생 모델에 증류하여 경량화된 한국어 언어 모델을 구축하는 기법은 본 발명에서 새로이 제안하는 기법이므로 지식재산권에 의해 보호받아야 한다.

도면

도면1

