사전 학습 모델의 위치 임베딩 길이 제한 문제를 극복하기 위한 방법론

정민수, 허탁성, 이주환, 김지수, 이경욱, 김경선

엔에이치엔다이퀘스트

tmzkdlwm1125@gmail.com, gjxkrtjd221@gmail.com, 9521ljh@gmail.com, jisukim8873@gmail.com,

arp1710@diquest.com, kksun@diquest.com

Methodology for Overcoming the Problem of Position Embedding Length Limitation in Pre-training Models

Minsu Jeong, Tak-Sung Heo, Juhwan Lee, Jisu Kim, Kyounguk Lee, Kyungsun Kim

NHN Diquest

요 약

사전 학습 모델을 특정 데이터에 미세 조정할 때, 최대 길이는 사전 학습에 사용한 최대 길이 파라미터를 그대로 사용해야 한다. 이는 상대적으로 긴 시퀀스의 처리를 요구하는 일부 작업에서 단점으로 작용한다. 본 연구는 상대적으로 긴 시퀀스의 처리를 요구하는 질의 응답(Question Answering, QA) 작업에서 사전 학습 모델을 활용할 때 발생하는 시퀀스 길이 제한에 따른 성능 저하 문제를 극복하는 방법론을 제시한다. KorQuAD v1.0과 AIHub에서 확보한 데이터셋 4종에 대하여 BERT와 RoBERTa를 이용해 성능을 검증하였으며, 실험 결과, 평균적으로 길이가 긴 문서를 보유한 데이터에 대해 성능이 향상됨을 확인할 수 있었다.

주제어: 길이 제한 극복, 사전 학습 언어 모델, 양방향 인코더 모델, 질의 응답

1. 서론

최근 자연어 처리 분야는 사전 학습 언어 모델(Pre-trained Language Model)을 활용하여 여러 분야에서 우수한 성능을 보이고 있다 [1]. 사전 학습 언어 모델의 학습 방법은 사전 학습(Pre-training)과 미세조정(Fine-tuning) 단계로 나뉜다 [2]. 사전 학습은 대량의 텍스트 데이터를 사용하여 높은 자연어 이해 능력을 갖추는 과정을 의미하며, 미세조정은 특정 작업에 대한 성능을 더욱 높이기 위해 해당 작업의 데이터를 소량 학습하여 가중치가 조정되는 과정을 의미한다.

BERT(Bidirectional Encoder Representation from Transformer), GPT(Generative Pre-trained Transformer) 등 대다수의 사전 학습 언어 모델은 텍스트에 대해 토큰 임베딩(Token Embedding)과 위치 임베딩(Positional Embedding) 단계를 거쳐 은닉 표현(Hidden Representation)을 생성한다 [3]. 여기서 토큰 임베딩이란 문장을 Word-Piece Model이나 Byte Pair Encoding 등을 사용하여 토큰으로 분할한 후 각 토큰에 대한 임베딩을 수행하는 작업이다. 위치 임베딩은 토큰의 순서에 따라 고유한 정수 값을 할당하여 위치 정보를 부여하는 작업이다.

사전 학습 언어 모델의 입력 시퀀스(Sequence) 길이는 512, 1024 등 다양한 크기를 가질 수 있지만, 일반적으로 512를 사용하는 경우가 많다. 이는 연산 비용과 메모리 제약 등 여러 가지 제한 사항으로 인한 문제를 고려한 타협점이다. 그리고, 절대 위치 임베딩(Absolute Positional Embedding)을 사용하는 사전 학습 언어 모델에서 입력 시퀀스가 512보다 길어지는 경우, 초과된 부분은 사용되지 않는다. 만약 초과된 부분을 모두 사용한다면, 위치 정보를 나타내는 위치 임베딩 값들이 적절하게 반영되지 못하는 문제가 발생할 수 있다.

본 연구에서는 이러한 문제를 해결하기 위해, 절대 위치 임베딩을 사용하는 사전 학습 언어 모델에서 입력 시퀀스의 길이가 초과되어도 적절한 위치 정보가 반영되도록하는 방법을 제안한다. 실험 결과로는 사전 학습된 BERT와 RoBERTa(Robustly Optimized BERT Pretraining Approach)를 사용하여 이 방법의 유용성을 검증하였으며, 입력 시퀀스 최대 길이보다 긴 길이를 갖는 텍스트 데이터에서 입력 시퀀스 길이 제한이 있는 모델보다 더욱 우수한 성능을 도출하였다. 학습 및 추론에 사용된 코드는 https://github.com/skaeads12/OvercomingPositionEmbeddingLimitation/tree/main에서 확인할 수 있다.

2. 관련 연구

2.1. Advancements in Pre-training and Transfer Learning

BERT는 양방향 트랜스포머(Transformer) 구조로 트랜스포머의 인코더(Encoder)를 활용한 사전 학습 언어 모델이다 [4]. BERT는 대량의 텍스트로부터 사전 학습을 시킨 후, 특정 작업에 대해 미세조정된다. 이는 자연어

이해, 문장 분류, 개체명 인식 등의 작업에서 뛰어난 성능을 보여준다.

GPT는 단방향 트랜스포머 구조로 트랜스포머의 디코더(Decoder)를 활용한 사전학습 언어 모델이다 [5]. GPT 또한 BERT와 같이, 대량의 텍스트로부터 사전 학습을 시킨 후, 특정 작업에 대해 미세조정된다. 이는 주어진 문맥에서 다음 단어를 예측하는 텍스트 생성 작업에서 뛰어난 성능을 보여준다.

최근 사전 학습과 전이 학습 방법에는 상당한 발전이 이루어졌다. 예를 들어, BERT의 변형인 RoBERTa는 사전 학습 단계에서 몇 가지 다른 방법을 사용하고, 더 큰 데이터셋과 초개매변수(Hyper-parameters)를 적절하게 조정하여 우수한 성능을 보인 모델이다 [6]. 또한, GPT도 시대의 변화에 따라 GPT-2, GPT-3, GPT-4 등으로 발전해 오며, 더 큰 데이터셋과 다른 사전 학습 방법을 활용하여 텍스트 생성에 우수한 성능을 보인다 [7].

뿐만 아니라, 사전 학습된 언어 모델을 다른 언어로 전이 학습하는 연구도 활발히 이루어지고 있다 [8]. 이를 통해 작은 언어나 특정 도메인의 데이터로도 효과적인 자연어 처리 모델을 구축할 수 있는 가능성이 제시되고 있다.

2.2. Attempts to Overcome the Sequence Length Limitation in the Pre-trained Model

사전 학습된 BERT 모델은 최대 길이 512의 제한을 가지고 있으나, 문서 분류나 기계 번역과 같은 작업에서는 더욱 긴 문장이나 긴 문서를 처리해야 하는 경우가 많다. 이러한 경우에는 입력 시퀀스 길이 제한으로 인해 문장이 잘리거나, 문서의 일부가 손실될 수 있어 작업의 성능이 저하되는 문제가 발생할 수 있다. 이러한 문제를 해결하기 위해 다양한 접근 방식이 제안되고 있다 [9, 10].

[9]는 문서 분류 작업을 위해 BERT를 사용한 연구이다. [9]는 문서의 어느 부분이 중요한 지 BERT를 통해 3가지 방법을 통해 검증한다. 첫 번째 방법은 문서의 앞 부분에서 나타나는 510 토큰만을 사용하는 방법이고, 두 번째 방법은 문서의 뒷부분에서 나타나는 510 토큰만을 사용하는 방법이며, 마지막 방법은 문서의 앞 부분 128 토큰과 뒷 부분 382 토큰을 종합하여 사용하는 방법이다. 510 토큰을 제외한 나머지 2개의 토큰은 BERT에서 특수 토큰으로 표현되는 [CLS], [SEP]을 사용하여 최종적으로 512 토큰을 사용한다. 이를 통해 문서를 분할하고 분할된 부분에 대해 사전 학습 언어 모델을 미세조정하여 최종 출력을 얻을 수 있다.

[10]은 긴 비정형 텍스트 구조를 갖는 임상 문서를 고정 시퀀스 단위로 분할하여 분할된 값 모두 BERT의 입력값으로 사용하는 연구로, 이를 D2SBERT(Document-to-Sequence BERT)로 정의하였다. D2SBERT로부터 추출된 분할된 값들의 [CLS] 토큰들을 하나로 결합하고, 이에 대해 시퀀스 단위의 어텐션(Attention)을 사용하여 미세조정을 하였으며, 512 토큰만을 사용하는 BERT에 비해 우수한 성능을 보였다.

이러한 방법들은 입력 시퀀스의 길이 제한으로 인한 성능 저하 문제를 극복하기 위해 다양한 방법을 제시하고 있으며, 긴 문장이나 문서를 처리하는 자연어 처리 작업에서 성능 향상을 모색하고 있다.

3. 방법론

절대 위치 임베딩을 사용하는 사전 학습 모델을 미세조정할 때, 입력 최대 길이는 사전 학습 시에 사용된 최대 길이를 따른다. 이는 위치 임베딩의 가중치 행렬(Weight Matrices)이 사전에 정의되어, 행렬의 구조를 임의로 변경할 수 없다는 제약 때문이다. 본 연구에서는 위치 임베딩의 새로운 가중치 행렬을 기존보다 큰 크기로 정의하고, 사전 학습된 절대 위치 임베딩의 가중치를 온전히 전달하는 방법을 소개한다.

위치 임베딩은 토큰의 위치 정보를 표현하기 위해 사용되는 기법으로, 각 위치에 고유한 임베딩 벡터를 할당하여 토큰의 위치를 표현한다. 이는 토큰 임베딩과 달리 의미와는 무관하게 순서에만 의존한다. 이를 통해 모델은 토큰의 순서를 이해하고 문장 내의 상호작용을 파악할 수 있다. 임베딩은 각 정수와 일대일로 대응되는 벡터 집합을 의미하며, 수식 (1)과 같이 정의된다.

$$EMB_n(p) \in \mathbb{R}^{d_{model}} \tag{1}$$

수식 (1)에서 n은 사전 학습 모델의 최대 길이를 의미하고, p는 절대 위치에 해당하는 정수 값을 의미하며, $[0,n) \in \mathbb{N}$의 범위를 가진다. 그리고 d_{model}은 임베딩 차원을 의미한다. 즉, $EMB_n(p)$는 최대 길이가 n이고 입력된 p에 해당하는 임베딩 벡터를 반환하는 함수이다. 일반적으로 사용되는 BERT-base는 $n = 512, d_{model} = 768$로 정의된다.

본 연구에서는 n을 확장하더라도 확장된 n에 대해 동일한 분포를 가지는 위치 정보를 부여하는 방법을 제안한다. 사전 학습에 사용된 초매개변수 n을 $2n$으로 확장시키고자 할 때, 최대 길이가 $2n$인 새로운 위치 임베딩은 수식 (2, 3)과 같이 정의된다.

$$EMB_{2n}(2p) = EMB_n(p) \tag{2}$$

$$EMB_{2n}(2p+1) = \frac{EMB_n(p) + EMB_n(p+1)}{2} \tag{3}$$

수식 (4)는 위치 값 p의 범위를 $p \in [0, 2n)$으로 확장하고, 수식 (2, 3)을 일반화한 것이다.

$$EMB_{2n}(p) = \frac{EMB_n\left(\left\lfloor \frac{p}{2} \right\rfloor\right) + EMB_n\left(\left\lceil \frac{p}{2} \right\rceil\right)}{2} \tag{4}$$

여기서 $\lfloor x \rfloor$는 x에 대한 버림 함수(Floor Function), $\lceil x \rceil$는 x에 대한 올림 함수(Ceil Function) 기호이다. 이를 통해 임베딩의 최대 길이를 확장하면서 동일한 분포를 유지할 수 있다.

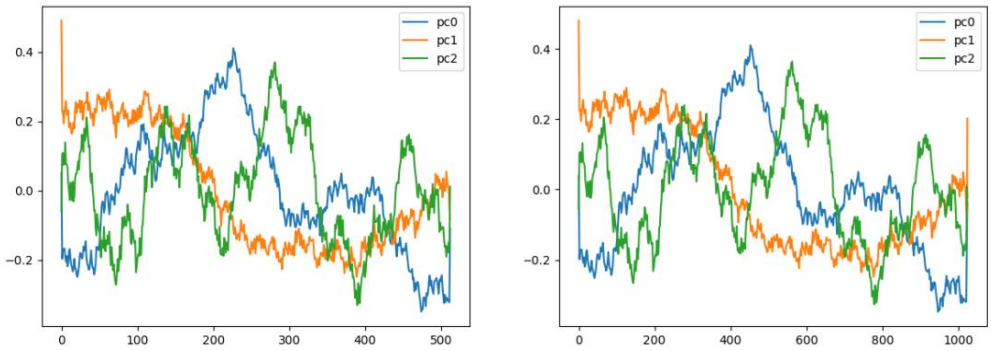

그림 1 (좌) 기존의 최대 길이 512 BERT 모델의 위치 임베딩 가중치 값 분포. **(우)** 제안 방법론으로 확장한 최대 길이 1,024 BERT 모델의 위치 임베딩 가중치 값 분포.

만약 n을 kn으로 확장하고자 할 때, $EMB_{kn}(p)$는 수식 (5, 6)과 같이 표현된다.

$$k \in [2, 4, 8, 16, \cdots] \quad (5)$$

$$EMB_{kn}(p) = \frac{EMB_{\frac{kn}{2}}\left(\left\lfloor \frac{p}{2} \right\rfloor\right) + EMB_{\frac{kn}{2}}\left(\left\lceil \frac{p}{2} \right\rceil\right)}{2} \quad (6)$$

4. 실험 결과

4.1. 사용 데이터셋

본 연구에서는 제안 방법론의 효과를 검증하기 위해 상대적으로 긴 시퀀스를 처리해야 하는 질의 응답 문제의 데이터셋을 활용하였다. 총 5가지 종류의 데이터셋을 사용하였으며, 이에는 KorQuAD v1.0 데이터셋 [11]과 AIHub에서 확보한 4종류의 데이터셋(기계독해, 뉴스 기사 기계독해 데이터, 일반상식, 행정 문서 대상 기계독해 데이터)이 포함된다. 검증 데이터가 포함된 데이터셋은 학습과 검증을 위해 9:1의 비율로 나누어 활용하였고, 검증 데이터가 없는 데이터셋의 경우 학습, 검증, 시험을 위해 각각 8:1:1의 비율로 나누어 사용하였다. 표 1은 각 데이터셋에 대한 통계 정보이다.

데이터셋	데이터 수		평균 토큰 길이	> 512 (%)[1]
KorQuAD v1.0	학습	54,366	272.70	3.18
	검증	6,041		
	시험	5,774		
기계독해	학습	272,070	368.81	15.88
	검증	34,009		
	시험	34,009		
뉴스 기사 기계독해 데이터	학습	244,728	420.05	28.12
	검증	27,192		
	시험	33,992		
일반상식	학습	80,214	234.66	0.33
	검증	10,027		
	시험	10,027		
행정 문서 대상 기계독해 데이터	학습	248,578	478.04	26.02
	검증	27,620		
	시험	34,524		

표 1 사용 데이터셋에 대한 통계 정보

[1] 문서의 길이가 512 토큰이 넘는 비율. 토큰 길이가 512 이상인 문서의 수 / 전체 문서의 수 (%).

4.2. 모델 구조

BERT 모델을 기반으로 제안 방법론을 적용하였을 때, 위치 임베딩 값의 분포는 그림 1과 같다.

그림 1은 위치 임베딩의 768차원 값을 주성분 분석(Principal Components Analysis, PCA)하여, 주성분 3개 차원을 2차원 그래프로 그린 것이다. 가로 축은 입력 위치 값(p), 세로 축은 각 위치에 대응하는 임베딩 ($EMB(p)$) 값이다. 그림 1과 같이, 기존 최대 길이 512의 BERT 위치 임베딩 값과 제안 방법론을 적용하였을 때 위치 임베딩 값의 분포가 거의 일치함을 확인할 수 있다.

그림 2는 실험을 위한 질의 응답 모델이며 이는 Bidirectional Encoder의 시퀀스 출력을 Span Prediction 분류기에 입력하여 정답의 시작 위치와 종료 위치를 예측하도록 구현된다.

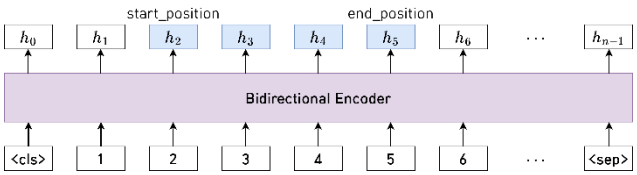

그림 2 질의 응답 모델 개요도

4.3. 결과 및 분석

사전 학습 모델은 BERT와 RoBERTa를 사용하여 실험하였으며, 두 모델을 제안 방법론에 적용하여 기존 최대 길이 512 모델(BERT-512, RoBERTa-512)과 확장된 최대 길이 1,024 모델(BERT-1024, RoBERTa-1024)을 비교하였다.

데이터셋	성능 지표	모델			
		BERT		RoBERTa	
		512	1,024	512	1,024
KorQuAD v1.0	EM (%)	78.13	77.23	80.90	78.28
	F1 (%)	88.91	88.04	90.64	88.77
기계독해	EM (%)	56.13	59.36	58.92	59.39
	F1 (%)	80.28	84.18	82.11	84.78
뉴스 기사 기계독해 데이터	EM (%)	56.96	63.02	58.92	59.39
	F1 (%)	72.11	79.91	82.11	84.78
일반상식	EM (%)	71.81	70.94	73.25	70.62
	F1 (%)	85.68	84.79	86.24	84.46
행정 문서 대상 기계독해 데이터	EM (%)	68.44	72.24	70.59	72.88
	F1 (%)	84.12	89.52	86.59	90.15

표 2 실험 결과

표 2는 기존 최대 길이인 512를 사용한 모델과 확장된 최대 길이인 1,024를 사용한 모델을 비교한 결과를 나타내며, 정량적 성능 비교를 위해 Exact Match(EM)와 F1-Score(F1)를 측정하였으며, Exact Match는 수식 (7)과 같이 표현되며, F1-Score는 수식 (8)과 같이 표현된다.

평균적으로 문서의 길이가 짧은 KorQuAD v1.0과 일반상식 데이터 실험에서는 최대 길이가 512인 모델이 더 높은 성능을 보였으며, 평균적으로 문서의 길이가 긴 기계독해, 뉴스기사 기계독해, 행정 문서 대상 기계독해 데이터 실험에서는 최대 길이가 1,024인 모델이 더 높은 성능을 보였다.

$$\mathrm{EM} = \frac{n \text{ of exactly matched samples}}{n \text{ of sample}} \times 100 \quad (7)$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precison} + \mathrm{recall}} \quad (8)$$

이는 제안 방법론을 활용하여 상대적으로 긴 시퀀스를 처리하는 것이 기존 모델을 512 길이로 제한하여 입력하는 것보다 효과적이라는 것을 의미한다. 반대로 평균적으로 짧은 시퀀스를 보유한 데이터셋에서는 성능이 하락하였으며, 이는 처리 길이가 확장되면서 입력되는 [PAD] 토큰의 수가 함께 증가하여 발생한 노이즈(Noise)가 학습을 방해한 것으로 예상된다. 따라서 [PAD] 토큰이 비정상적으로 대량 입력될 환경이 아닌, 컨텍스트의 길이가 보장된다면 제안 방법론이 유용하게 작용한다.

5. 결론

본 연구에서는 사전 학습 모델에서 기존 최대 입력 길이 512를 극복하고, 상대적으로 긴 시퀀스를 처리할 수 있는 새로운 절대 위치 임베딩 방법론을 제안하였다. 제안한 방법론의 유용성을 검증하기 위해, 사전 학습 모델인 BERT와 RoBERTa에 기계독해 데이터셋을 사용하여 성능을 측정하였다.

실험은 기존 최대 길이가 512인 모델과 제안한 방법론인 최대 길이가 1024로 확장시킨 모델을 사용하여 성능을 비교하였다. 실험 결과, 평균적으로 문서의 길이가 짧은 데이터셋에서는 최대 길이가 512인 모델이 더 높은 성능을 보였으며, 문서의 길이가 긴 데이터셋에서는 최대 길이가 1,024로 확장시킨 모델이 더 높은 성능을 보였다. 이러한 결과로 제안 방법론은 주로 긴 시퀀스를 처리해야 하는 환경에서 유용하게 적용될 수 있으며, 시퀀스 길이에 따라 적절하게 선택하여 사용하는 것이 필요하다.

언어 모델은 모델의 크기 뿐만 아니라, 텍스트의 길이에 따라 시간과 계산 비용에 큰 영향을 미친다. 즉, 긴 텍스트를 사용하여 언어 모델을 사전 학습한다면 이는 더 많은 시간과 계산 비용이 소모된다는 것이다.

향후 연구에서는 512의 시퀀스 길이를 가진 모델이 아닌 더 짧은 시퀀스 길이인 256에 대해 언어 모델을 직접 사전 학습하여, 본 연구에서 제안한 방법을 적용해 성능을 측정해볼 것이며, 이를 통해, 기존 사전 학습 언어 모델과 더 적은 자원을 사용하여 구축된 언어 모델의 성능 비교를 통해 유용성을 검증할 것이다. 또한, 사전 학습 언어 모델에서의 텍스트 길이를 확장시키는 다른 방법들을 모색할 것이다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00956, 소셜네트워크에서의 온라인그루밍 위험성 자가탐지 기술 개발)

참고문헌

[1] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, Xuanjing HuangI. Mani and T. Maybury, “Pre-trained Models for Natural Language Processing: A Survey”, SCIENCE CHINA Technological Sciences,2020, 63, 1872-1897

[2] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, Qing He, “A Comprehensive Survey on Transfer Learning”, arXiv preprint arXiv:1911.02685

[3] Kawin Ethayarajh, “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”, arXiv preprint arXiv:1909.00512
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv preprint arXiv:1810.04805
[5] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving language understanding by generative pre-training.", 2018
[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv preprint arXiv:1907.11692
[7] OpenAI, “GPT-4 Technical Report”, arXiv preprint arXiv:2303.08774
[8] Telmo Pires, Eva Schlinger, Dan Garrette, “How multilingual is Multilingual BERT?”, arXiv preprint arXiv:1906.01502
[9] Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang, “How to Fine-Tune BERT for Text Classification?”, arXiv preprint arXiv:1905.05583
[10] Tak-Sung Heo, Yongmin Yoo, Yeongjoon Park, Byeong-Cheol Jo, Kyungsun Kim, “Medical Code Prediction from Discharge Summary: Document to Sequence BERT using Sequence Attention”, arXiv preprint arXiv:2106.07932
[11] Seungyoung Lim, Myungji Kim, Jooyoul Lee, “KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension”, arXiv preprint arXiv:1909.07005