The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 23, No. 4, pp.115-119, Aug. 31, 2023. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2023.23.4.115
JIIBC 2023-4-18

트랜스포머 기반 효율적인 자연어 처리 방안 연구

A Study on Efficient Natural Language Processing Method based on Transformer

임승철*, 윤성구

Seung-Cheol Lim*, Sung-Gu Youn

요 약 현재의 인공지능에서 사용되는 자연어 처리 모델은 거대하여 실시간으로 데이터를 처리하고 분석하는 것은 여러 가지 어려움들을 야기하고 있다. 이런 어려움을 해결하기 위한 방법으로 메모리를 적게 사용해 처리의 효율성을 개선하는 방법을 제안하고 제안된 모델의 성능을 확인하였다. 본 논문에서 제안한 모델의 성능평가를 위해 적용한 기법은 BERT[1] 모델의 어텐션 헤드 개수와 임베딩 크기를 작게 조절해 큰 말뭉치를 나눠서 분할 처리 후 출력값의 평균을 통해 결과를 산출하였다. 이 과정에서 입력 데이터의 다양성을 주기위해 매 에폭마다 임의의 오프셋을 문장에 부여하였다. 그리고 모델을 분류가 가능하도록 미세 조정하였다. 말뭉치를 분할 처리한 모델은 그렇지 않은 모델 대비 정확도가 12% 정도 낮았으나, 모델의 파라미터 개수는 56% 정도 절감되는 것을 확인하였다.

Abstract The natural language processing models used in current artificial intelligence are huge, causing various difficulties in processing and analyzing data in real time. In order to solve these difficulties, we proposed a method to improve the efficiency of processing by using less memory and checked the performance of the proposed model. The technique applied in this paper to evaluate the performance of the proposed model is to divide the large corpus by adjusting the number of attention heads and embedding size of the BERT[1] model to be small, and the results are calculated by averaging the output values of each forward. In this process, a random offset was assigned to the sentences at every epoch to provide diversity in the input data. The model was then fine-tuned for classification. We found that the split processing model was about 12% less accurate than the unsplit model, but the number of parameters in the model was reduced by 56%.

Key Words : machine learning, natural language processing, transformer, artificial intelligence

Ⅰ. 서 론

인공지능이 우리에게 처음 다가온 것은 2016년 3월 구글 딥마인드의 알파고가 이세돌과의 바둑 대국에서 보여준 능력이었고, 일반사람들도 스마트폰의 시리와 빅스비와 같은 인공지능 비서를 사용하면서 인공지능에 관한 이해도가 높아지고 있으며 우리 일상의 다양한 전문 분야에도 인공지능이 녹아들고 있다.[2][3] 챗 GPT가 출시되

*정회원, 우송대학교 IT융합학부 (교신저자)
접수일자 2023년 7월 14일, 수정완료 2023년 7월 31일
게재확정일자 2023년 8월 4일

Received: 14 July, 2023 / Revised: 31 July, 2023 /
Accepted: 4 August, 2023
*Corresponding Author: sclim@wsu.ac.kr
Dept. of IT Convergence, Woosong University, korea

면서 더 많은 관심을 받고 있다.

1. 트랜스포머[4] 기반 자연어 처리 모델의 크기

오픈 AI에서 개발한 GPT-2의 모델 구조를 따른 한국어 자연어 처리모델 중 하나인 koGPT의 경우에는 22GB[5]의 모델 크기를 가진다. 이런 큰 용량의 모델은 배포에 제한이 크고 이에 따라 모델의 크기를 줄여야 할 필요성이 있다.

트랜스포머 기반 모델들은 처리할 수 있는 토큰의 개수가 늘어날수록 토큰의 임베딩 크기와 레이어의 개수가 늘어나는 추세가 있다.

표 1 은 트랜스포머 기반 모델은 BERT와 GPT[6]의 파라미터 개수(P), 레이어의 개수(L), 어텐션 헤드의 개수(H), 한번에 처리가능한 토큰의 개수(T), 토큰 임베딩의 크기(E)를 보여준다.

표 1. 트랜스포머 기반 모델들의 구조요약[7][8]
Table 1. Overview of transformer-based models

	BERT Base	BERT Large	GPT-1	GPT-2	GPT-3
P	110M	340M	117M	1.5B	175B
L	12	24	12	48	96
H	12	16	12	-	96
T	512	512	512	1024	2048
E	768	1024	768	1600	12288

2. 메모리 효율적인 처리방법

본 논문에서는 BERT 기반 모델의 어텐션 헤드의 개수와 토큰의 임베딩 크기를 조정해 입력 데이터를 나누어 처리했을 때 사전훈련의 손실값의 비교와 입력 말뭉치(Corpus)의 길이가 작은 만큼 여분의 토큰이 생기므로 이 데이터를 잘 활용하기 위해 고안한 방법도 같이 소개한다. 3장의 제안한 모델 구조에서는 모델 구조와 사전 학습(Pre-training)에서 데이터의 다양성을 주기 위해 말뭉치에서 오프셋을 지정하는 방법, 미세 조정(Fine-tuning)에서의 모델 출력을 합치는 방법에 관해 설명하였다.

5장의 학습결과에서는 사전 학습 단계에서의 NSP (Next Sentence Prediction), MLM(Masked Language Model)을 합친 손실(loss)값과 미세 조정에서의 분류 정확도를 서술하였다.

II. 관련 연구

본 논문에서는 모델 연산시 임베딩 크기와 어텐션 헤드의 개수를 조절하여 분할 처리의 효율성을 높였다.

어텐션 헤드는 입력 토큰과 다른 토큰 간의 관계를 표현하는데, 트랜스포머 기반 모델에서는 이런 어텐션 헤드를 여러 개 사용하여 토큰들간의 다양한 관점에서의 관계를 표현하기 위해 멀티 헤드 어텐션 레이어를 구성한다.

자연어 처리에서 임베딩 벡터는 입력의 의미를 인식하는 최소 단위다. 일반적으로 임베딩 크기가 클수록 의미를 세밀하게 표현하는 것이 가능하다.

BERT에는 그 기반이 되는 트랜스포머와 다르게 3종류의 임베딩 레이어가 존재하는데 입력 문장을 토큰화시킨 토큰 임베딩 레이어, 토큰의 위치를 표현하기 위한 포지셔널 임베딩 레이어, 선행 학습혹은 일부 미세조정에서 두 쌍의 문장이 이어지는 여부를 알기 위한 세그먼트 임베딩 레이어가 있다.

기반이 되는 모델인 트랜스포머에는 토큰 임베딩 레이어만 존재하고 포지셔널 임베딩은 학습되지 않는 상수인 포지셔널 인코딩을 토큰 임베딩에 더해주는 것이 BERT와 다른 점 중 하나이다.

III. 제안한 모델 구조

BERT 모델에서 표 2 와 같이 구조를 변경해 학습을 진행하였다.

표 2는 BERT Base와 본 논문에서 어텐션 헤드의 개수(H)와 토큰 임베딩의 크기(E), 처리할 수 있는 토근 개수(T)에 따라서 성능 변화를 관측하기 위해 만든 모델 A, 모델 B 간의 구조적 차이를 보여준다.

표 2. 실험 모델의 구조
Table 2. Structure of an experimental model

	BERT Base	모델 A	모델 B
H	12	8	16
E	768	256	512
T	512	128	256

모델 A의 경우에는 어텐션 헤드의 개수가 8개, 토큰 임베딩 크기가 256으로 모델 B에 비해 50% 가량 작은

크기를 가진다. 각각 128개, 256개의 토큰을 처리할 수 있으므로 모델 A를 입력을 나누어 처리하기 위한 모델로 사용했다.

어텐션 헤드의 개수가 많은 경우에는 각 토큰 간의 관계를 많은 관점에서 바라보는 것으로 해석할 수 있는데 짧은 문장에서는 긴 문장의 경우보다 더 적은 관점이 존재한다고 가정하였다.

모델 A의 경우에는 입력토큰의 최대 개수는 128개이다. 또한 상술했듯이 적은 입력 토큰에는 많은 관점이 필요하지 않다고 가정했기 때문에 어텐션 헤드의 개수 또한 줄였고 이에 따라 모델의 임베딩 크기 또한 줄였다. 총 2천 3백만개 가량의 파라미터를 가졌다.

모델 B의 경우는 모델 A와 비교하기 위한 비교 대상으로 입력토큰의 최대 개수는 256개이고 어텐션 헤드의 개수는 모델 A보다 많은 16개로 설정하였다. 이에 따라 모델의 임베딩 크기 또한 512차원으로 늘렸다. 총 5천 3백만개 가량의 파라미터를 가졌다.

Ⅳ. 제안한 모델 학습 방법

1. 학습환경

사전 훈련에서 모델 A의 경우 RTX 3080 10GB GPU 를 통해 약 130억개의 토큰을 학습하였고, 모델 B의 경우 RTX 4090 24GB GPU 1개를 통해 약 130억개의 토큰을 훈련하였다.

미세 조정에서는 RTX 3080 GPU를 통해 진행하였다.

2. 사전 훈련

사전 훈련은 위키피디아의 영어 데이터셋을 바탕으로 학습을 진행했다. 입력 토큰의 개수가 짧은 모델의 특성상 학습데이터를 토큰의 개수에 맞게 자르면 버려지는 토큰이 많아지게 된다. 이런 데이터셋을 활용하기 위해 토큰의 특정 구간을 오프셋으로 지정하고 매 에폭마다 오프셋이 임의로 선택되도록 학습데이터의 전처리를 구성하였다.

이와 같은 데이터셋의 전처리는 짧은 토큰을 입력으로 받는 모델에서 데이터의 다양성을 높여주어 모델이 다양한 데이터셋을 통해 학습이 가능하도록 만들어 준다.

다음 그림 1과 같이 문장을 토큰화하고 그중 문장의 앞에서부터 절반 지점에서 임의로 오프셋을 선택한 후 필요한 개수만큼 발췌해 학습 데이터셋으로 사용하는 방식을 취하여 NSP와 MLM 방식을 통해 사전 학습을 진행하였다.

그림 1에서 n은 토큰화된 글에서 전체 토큰의 개수를 의미한다.

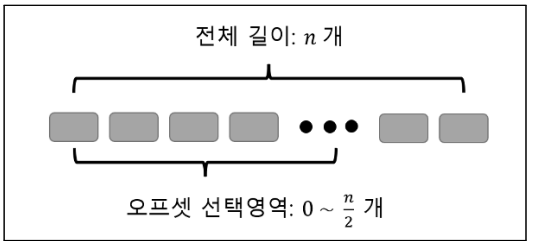

그림 1. 사전 학습 데이터의 오프셋 설정
Fig. 1. Configuration the offset of the pre-training data

3. 미세 조정

IMDb(Internet Movie Database) 감정분석 데이터셋에서 모델 B의 처리 가능 토큰 개수를 넘어가지 않는 데이터를 이용하여 감정 분류 모델을 미세조정 하였다. 모델 A에 사용되는 데이터의 경우에는, 마침표를 기준으로 우선적으로 문장을 분리하고 하나의 문장이 모델의 처리가능한 토큰의 개수를 넘어설때는 임의의 지점에서 문장을 분리하였다. 이후 분리된 문장마다 다른 의미를 가질 수 있으므로 별도의 모델을 통해 분리된 데이터의 라벨을 다시 부여했다. 모델 B의 경우에는 데이터셋에 원래대로 부여된 라벨을 그대로 사용하였다.

각 데이터에 대해 예측을 진행할 때 모델 A는 아래 같은 수식 1을 따라서 최종 출력을 산출한다. y_i는 모델의 분할 처리한 각각의 출력을 의미하고 i는 각 출력에 대한 인덱스, Y는 분할처리 모델 A의 최종 출력이다.

$$Y = \frac{\sum_{i=1}^{I} \log(y_i)}{I} \quad (1)$$

Ⅴ. 학습 결과

사전학습에서 모델 A와 모델 B는 아래 그림 2, 그림 3과 같은 손실을 기록하였다.

또한 그래프 뒤에 음영을 주어 분산을 표현하였다.

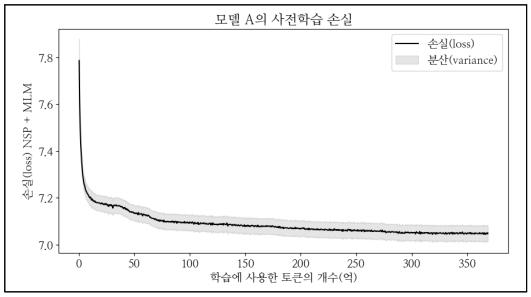

그림 2. 모델 A의 사전학습 손실
Fig. 2. Pre-training loss of Model A

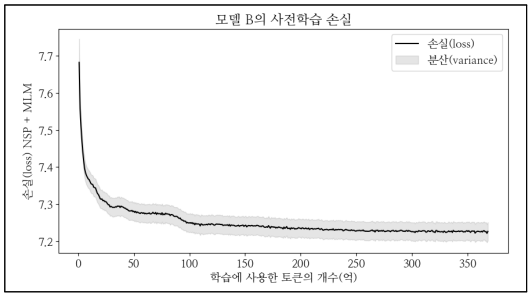

그림 3. 모델 B의 사전학습 손실
Fig. 3. Pre-training loss of Model B

미세 조정에는 모델 A의 경우 최대 13,861개의 분할된 문장으로 이루어진 테스트 데이터에서 86.5%의 정확도를 모델 B의 경우 13,659개의 문장으로 이루어진 테스트 데이터에서 최대 97.8%가량의 정확도를 나타내었다.

아래 그림 4, 그림 5는 모델 A, 모델 B의 테스트 데이터셋에 대한 정확도를 나타낸다. 각각 29번째 에폭 44번째 에폭에서 최대 정확도를 기록한 것을 알 수 있다.

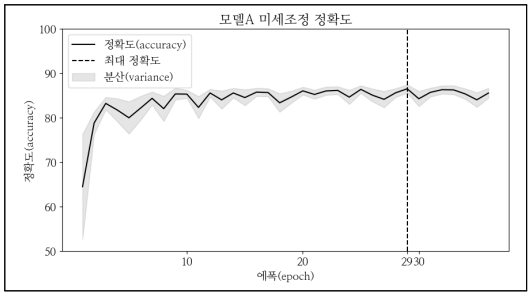

그림 4. 모델 A의 미세조정 정확도
Fig. 4. Fine-tuning Accuracy of Model A

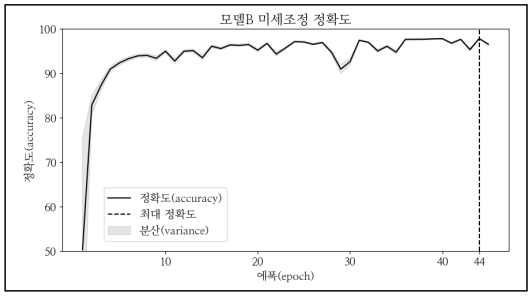

그림 5. 모델 B의 미세조정 정확도
Fig. 5. Fine-tuning Accuracy of Model B

VI. 결 론

근래의 자연어처리 모델은 거대한 크기를 가지고 이는 다양한 분야에서의 활용에 제약이 생기는 원인 중 하나이다. 이런 제약사항을 해결하기 위해 본 논문에서는 입력을 분할하여 처리하는 방법을 제안하였다.

실험에서는 전체 문장을 한 번에 처리하는 모델 B와 분할 처리를 하는 모델 A를 훈련시켰고 입력 문장의 길이가 적은 모델 A의 경우에는 매 에폭마다 임의에 오프셋을 주어 학습 데이터에 다양성을 주었다.

미세조정에서 모델 A는 모델 B에 비해서 비교적 낮은 정확도를 기록했지만, 메모리 자원이 제한적이고 처리해야하는 문장의 길이가 긴 환경에서 사용이 가능할 것이다.

본 논문에서는 이진 분류로만 미세 조정을 실시하였다. 추후 연구에서는 다중 분류, 회귀, 문장 임베딩등의 다양한 미세 조정에서 분할처리를 위한 개선된 방법을 연구하고 적용하여 좀 더 성능을 개선하는 방법에 대해 연구할 예정이다.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", n Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171-4186. DOI: https://doi.org/10.48550/arXiv.1810.04805

[2] You-bin Lee, Sung-Joon Lee, Byung-Won On, "Comparison of Q-learning and SARSA Reinforcement LearningAlgorithms Performance in Pendlum Game",

The Proceedings of the 2021 KIIT Autumn Conference, pp.431-434, Jun 2021.

[3] Bae Kyungyeol, Cho Jungkeun, Yoo Byung Joo, "A Study on Establishment of AI Development Strategy for Ground Operations innovation Applying PEST - 7S - SWOT" Journal of the Korea Academia-Industrial cooperation Society, Vol.22 No.6, pp.67-74, Jun 2021. DOI: https://doi.org/10.5762/KAIS.2021.22.6.67

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", Advances in neural information processing systems, Vol.30 pages 6000-6010. DOI: https://doi.org/10.48550/arXiv.1706.03762

[5] kakobrain (2022) "kogpt". https://github.com/kakaobrain/kogpt.git

[6] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training", 2018

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya sutskever, "Language Models are Unsupervised Multitask Learners", 2019

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al, "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, Vol.34, No.159, pp.1877-1901, Dec 2020

저 자 소 개

임 승 철(정회원)

- 1985년 2월 : 한양대학교 전자공학과 학사
- 1990년 2월 ~ 1996년 2월 : 한국전자통신연구원 선임연구원
- 1994년 8월 : 전북대학교 정보통신공학과 석사 박사
- 2006년 3월 ~ 현재 : 우송대학교 IT융합학부 교수
- 주관심분야 : 이동통신, 네트워크보안, 인공지능

윤 성 구(준회원)

- 2018년 3월 ~ 현재 : 우송대학교 IT융합학부 학사재학
- 주관심분야 : 인공지능

※ 이 논문은 2023학년도『우송대학교 교내 학술연구조성비』지원에 의해서 수행됨