효율적인 언어모델 사전학습을 위한 집중학습기법

박준형[1], 이상근[2]
[1]고려대학교 컴퓨터학과, [2]고려대학교 인공지능학과
{irish07, yalphy}@korea.ac.kr

Concentrated Learning for Efficient Pre-training of Language Models

Jun-Hyung Park[1], SangKeun Lee[2]
[1]Department of Computer Science and Engineering, Korea University
[2]Department of Artificial Intelligence, Korea University

요 약

사전학습된 언어모델은 다양한 자연어 처리 작업에서 뛰어난 성공을 거두고 있다. 그러나 성능 향상을 위해 사전학습에서 사용하는 말뭉치의 크기와 계산량이 계속해서 증가하고 있다. 우리는 효율적인 사전학습을 위해 말뭉치에서 많은 지식을 함유하는 텍스트를 추출하여 집중적으로 학습하는 집중학습 기법을 제안한다. BERT 모델을 사용한 실험을 통해 제안 기법이 일반적인 사전학습 기법보다 더 효율적으로 더 높은 자연어 이해 성능을 달성하는 것을 확인한다.

Ⅰ. 서 론

언어모델 사전학습은 자연어 추론, 감정 분류, 질의응답 등 다양한 자연어 처리 작업 성능을 향상시킬 수 있는 효과적인 수단이다 [1]. 사전학습 과정에서 언어모델은 말뭉치로부터 자연어의 구조적, 의미적 정보와 지식에 대해 학습할 수 있고, 이를 자연어 처리 작업에 활용하여 높은 성능을 달성할 수 있다. 특히 사전학습된 Transformer [2] 기반 언어모델들은 자연어 이해 벤치마크(GLUE [3] 등)에서 인간을 초월하는 점수를 달성했다. 2018년 이래로 BERT [1], RoBERTa [4] 등 사전학습된 Transformer 기반 언어모델이 계속해서 제안되어오고 있으며 다양한 자연어 처리 작업에서 최고 성능을 갱신하고 있다.

최근 연구들은 사전학습에서 사용하는 말뭉치의 크기가 커질수록 언어모델의 자연어 처리 작업의 성능이 향상된다는 것을 전제로 받아들이고 있다 [4, 5]. RoBERTa는 BERT와 비교하여 열 배 수준의 말뭉치로 사전학습되어 자연어 이해 벤치마크와 질의응답 벤치마크에서 성능을 크게 향상시키는 결과를 보여주었다. 그리고 GPT-3 [5]는 BERT와 비교하여 33배의 말뭉치로 사전학습되어 인간이 구분하지 못할 정도로 자연스러운 기사를 생성해냈고, 다양한 자연어 처리 벤치마크에서 미세조정 없이 언어모델만으로 경쟁력 있는 성능을 보여주었다.

그러나 언어모델은 전체 말뭉치를 일정 회 이상 반복적으로 학습해야 효과적인 성능을 보이기 때문에 [5], 말뭉치의 크기가 증가함과 함께 사전학습에 필요한 계산 비용 또한 증가하고 있다. 최신 언어모델들의 사전학습에 필요한 계산을 모두 수행하기 위해서 고성능의 NVIDIA V100 GPU 1대로도 1년 이상의 시간이 소요된다 [5]. 이로 인해 수백 대 이상의 GPU/TPU로 구성된 클러스터 없이는 최신 언어모델의 사전학습이 어렵고, 이는 자연어 처리 연구에 큰 걸림돌이 되고 있다.

본 연구에서 우리는 앞서 언급한 전제가 Transformer 기반 언어모델 사전학습에 있어 항상 성립하는 것이 아님을 보인다. 다양한 자연어 처리 작업에서 지식은 유용하게 사용될 수 있다. 따라서 우리는 많은 지식을 함유하는 텍스트를 추출하여 집중적으로 학습하는 집중학습기법을 제안한다. 우리는 BERT를 활용한 집중학습기법의 실증적인 평가를 통해 기존 전제와 모순되는 두 가지 현상을 관측했다. 첫째로, 우리는 사전학습 말뭉치에서 지식을 많이 함유하는 극히 일부의 텍스트만으로도 BERT의 성능을 효과적으로 향상시킬 수 있음을 발견했다. 둘째로, 우리는 사전학습 말뭉치에서 BERT 성능 향상에 도움이 되지 않는 텍스트가 있음을 발견했다. 이러한 발견들을 기반으로 우리는 큰 말뭉치와 계산 비용을 사용하여 언어모델을 사전학습하는 현재의 접근법을 재고해야 함을 주장한다. 본 연구는 단순히 대규모의 사전학습 말뭉치를 사용하기보다 소규모의 풍부한 지식을 담고 있는 텍스트를 선별하여 사전학습을 진행하는 것이 더 효과적, 효율적으로 성능을 향상시킬 수 있음을 보여준다. 본 연구의 기여점을 요약하면 다음과 같다.

- 우리는 언어모델 사전학습에서 많은 지식을 담고 있는 텍스트를 추출하여 집중적으로 학습하는 집중학습기법을 제안한다.
- 우리는 더 큰 말뭉치가 언어모델 성능을 향상시킨다는 기존 전제와는 달리 소규모의 말뭉치로도 사전학습된 언어모델의 성능을 향상시킬 수 있음을 입증한다.
- 우리는 집중학습기법을 통해 사전학습된 BERT 모델이 자연어 이해 벤치마크에서 뛰어난 성능을 달성함을 보인다.

Ⅱ. 집중학습기법

이 장에서 우리는 집중학습기법의 설계와 세부 진행 과정에 관해 설명한다. 우선 지식을 많이 함유하는 텍스트를 추출하는 기법에 관해 설명하고, 추출한 텍스트를 활용한 언어모델 사전학습 기법에 관해 설명한다.

Ⅱ-1. 지식 함유 텍스트 추출

언어모델 사전학습에 사용되는 말뭉치는 과학, 역사, 정치 등 다양한 분야와 백과사전, 소설, 기사 등 다양한 갈래의 글을 포함하고 있다. 마찬가지로, 텍스트 속에 들어있는 지식의 분야와 표현 형태 또한 다양하게 나타날 것이다. 텍스트 속에 들어있는 다양한 분야와 형태의 지식을 감지하기 위해서 우리는 3200만 개의 트리플렛으로 구성된 대규모 지식 그래프

ConceptNet [6]을 활용했다. ConceptNet은 $G \subset C \times R \times C$ 구조의 지식 그래프로 C는 자연어로 표현된 개념을 나타내고 R은 개념 사이의 관계성을 나타낸다. 하나의 트리플렛은 (c_1, r, c_2)로 구성되어 c_1이 c_2와 r의 관계성을 갖는다는 지식을 나타낼 수 있다. 예를 들어 트리플렛(river, AtLocation, waterfall)은 "강은 폭포가 있는 위치에 있다."는 지식을 표현할 수 있다.

우리는 distant supervision [7]을 바탕으로 ConceptNet을 활용하여 텍스트에 들어있는 지식을 계측한다. 다음 일련의 수식을 통해 텍스트에 들어있는 지식의 양을 측정하고 많은 지식을 포함하는 텍스트를 선별한다.

$$f(x, G) = \left|\{r \mid c_1 \subseteq x, c_2 \subseteq x, (c_1, r, c_2) \in G\}\right|$$

$$T_{extracted} = \{t \mid f(t) > \alpha, t \in T\}$$

위 수식에서 T는 사전학습 말뭉치에 들어있는 문장들의 집합을 나타내며 α는 판별 기준을 설정하는 초파라미터다. 사용하는 모든 개념과 문장들은 언어모델 학습에 필요한 방식으로 토큰화된다.

Ⅱ-2. 언어모델 사전학습

언어모델 사전학습은 기존 BERT 모델에서 사용하는 마스크 언어 모델링 방식을 사용한다. 토큰화된 문장에서 15%의 토큰을 마스킹하며 이 중 80%는 특별 토큰 중 하나인 [MASK] 토큰으로, 10%는 임의의 토큰 중 하나로 대체하며, 나머지 10%는 원본 토큰을 변경 없이 그대로 사용한다. 문맥 표현을 학습하기 위해 여러 개의 연속된 문장을 이어붙여 한 번에 최대 128 또는 512개의 토큰을 학습에 사용하는 기존 언어모델 방법론과는 달리, 우리의 방법론은 문장 단위로 학습하기 때문에 최대 64개의 토큰을 사용하여 학습을 진행한다.

Ⅲ. 실험

우리는 본 논문에서 제안한 집중학습기법으로 학습된 모델과 일반사전학습으로 학습된 모델, 그리고 지식 함유 텍스트를 추출하지 않고 임의의 문장을 추출하여 학습하는 임의추출학습 기법으로 학습된 모델을 비교하여 자연어 이해 성능 향상의 효율성을 측정했다. 자연어 이해 성능을 측정하기 위해 GLUE 벤치마크 [3]를 사용하며 GLUE 점수와 사용한 말뭉치의 크기, 계산 비용을 통해 성능 향상의 효율성을 비교했다. 사전학습된 언어모델로 Hugging Face[1] 라이브러리에서 배포한 BERT 모델을 사용했으며 배치 사이즈 128, 학습률 3e-5를 사용했다. 다른 학습 관련 초파라미터는 기존 BERT 모델과 동일하게 설정했다. 집중학습을 위한 초파라미터 α는 1.0으로 설정했다. 텍스트 추출과 사전학습 시 BERT에서 사용한 말뭉치를 사용했다. 실험은 Pytorch 라이브러리를 통해 구현되었으며 NVIDIA RTX 3090 네 대를 사용하여 이루어졌다.

실험 결과는 표 1, 2에 보고되어 있다. 실험 결과에 따르면 BERT-base와 BERT-large 두 개의 모델에서 우리가 제안한 집중학습기법이 다른 기법들을 뛰어넘는 성능을 달성했다. 특히 집중학습기법은 일반사전학습 대비 0.3% 수준의 말뭉치만 사용해서 더 높은 성능을 달성했다. 이 결과로부터 집중학습기법이 말뭉치 크기 측면에서 더 효율적으로 성능을 향상시킨다는 것을 확인할 수 있다. 특히 일반사전학습으로 20배의 계산량(FLOPs)을 사용하여 모델을 학습시켰을 때에도 집중학습기법이 더 높은 성능을 달성했다. 이로부터 우리가 제안한 집중학습기법이 일반사전학습 기법보다 더 효율적인 기법임을 확인할 수 있다. 그리고 지식 함유 텍스트가 아니라 임의의 문장들을 선별해서 학습했을 때 다른 조건이 모두 같음에도 GLUE 점수가 가장 떨어지는 현상을 관측했다. 이로부터 우리가 선별한 지식 함유 텍스트가 사전학습 말뭉치에 들어있는 텍스트보다 평균적으로 성능 향상에 더 기여한다는 것을 알 수 있다.

학습 기법	말뭉치 크기	FLOPs	GLUE 점수
일반사전학습	16GB	9.6e15	82.0
일반사전학습 20x	16GB	1.9e17	82.8
임의추출학습	50MB	9.6e15	81.7
집중학습	50MB	9.6e15	**83.1**

표 1 BERT-base GLUE 점수

학습 기법	말뭉치 크기	FLOPs	GLUE 점수
일반사전학습	16GB	2.9e16	83.8
일반사전학습 20x	16GB	5.7e18	84.3
임의추출학습	50MB	2.9e16	82.2
집중학습	50MB	2.9e16	**84.5**

표 2 BERT-large GLUE 점수

Ⅲ. 결론

본 논문에서는 많은 지식을 담고 있는 텍스트를 집중적으로 사전학습하는 집중학습 기법을 제안하여 말뭉치가 클수록 언어모델 성능이 좋아진다는 기존 전제가 성립하지 않는 경우를 확인했다. 본 논문의 결과는 현재 계속해서 늘어나고 있는 사전학습 말뭉치의 크기와 사전학습 계산량을 크게 줄일 수 있다는 가능성을 보여준다. 추후 우리는 집중학습 기법을 더 다양한 모델과 지식 그래프, 사전학습 말뭉치에 적용할 수 있도록 확장할 계획이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2021-2016-0-00464)

참 고 문 헌

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[3] Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." arXiv preprint arXiv:1804.07461 (2018).

[4] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[5] Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

[6] Speer, Robyn, Joshua Chin, and Catherine Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge." Thirty-first AAAI conference on artificial intelligence. 2017.

[7] Mintz, Mike, et al. "Distant supervision for relation extraction without labeled data." Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009.

1) https://huggingface.co/