ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해

ALBERT for Korean Natural Language Processing: Named Entity Recognition, Sentiment Analysis, Machine Reading Comprehension

저자 (Authors)

이영훈, 나승훈, 최윤수, 이혜우, 장두성

Young-Hoon Lee, Seung-Hoon Na, Yun-Su Choi, Hye-Woo Lee, Du-Seong Chang

출처 (Source)

한국정보과학회 학술발표논문집 , 2020.7, 332-334 (3 pages)

발행처 (Publisher)

한국정보과학회

The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09874430

APA Style

이영훈, 나승훈, 최윤수, 이혜우, 장두성 (2020). ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해. 한국정보과학회 학술발표논문집, 332-334.

ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해

이영훈[O1], 나승훈[1], 최윤수[2], 이혜우[2], 장두성[2]
[1]전북대학교, [2]KT
{ dldudgns73, nash }@jbnu.ac.kr, { yunsu.choi, lee.hyewoo, dschang }@kt.com

ALBERT for Korean Natural Language Processing: Named Entity Recognition, Sentiment Analysis, Machine Reading Comprehension

Young-Hoon Lee[O1], Seung-Hoon Na[1], Yun-Su Choi[2], Hye-Woo Lee[2], Du-Seong Chang[2]
[1]Jeonbuk National University, [2]KT

요 약

최근 자연어처리 분야에서는 Transformer 기반의 언어모델을 이용하여 사전학습을 진행하고, fine-tuning을 적용함으로 다양한 태스크에서 SOTA(State Of The Art)를 이루고 있다. 이러한 언어모델은 대용량의 말뭉치 데이터와 많은 수의 파라미터를 사용하여 학습을 진행하는데, 학습을 진행함에 있어 자원의 한계가 존재하기에 큰 모델의 학습에는 어려움이 있다. 본 논문에서는 이러한 문제를 해결하기 위해 모델의 파라미터의 수를 줄인 AL-RoBERTa를 제안하고, 여러 한국어 태스크에 적용함으로 모델의 성능을 측정하여 효과성을 보인다.

1. 서 론

최근 자연어처리 분야에서는 BERT[1], RoBERTa[2], XLNet[3] 등 Transformer[4] 기반의 언어모델(Language Model) 연구가 활발히 이뤄지고 있다. 이러한 언어모델은 대용량의 말뭉치 데이터와 많은 파라미터를 이용하여 사전학습을 진행하고 fine-tuning을 적용하여 다양한 자연어처리 태스크에서 높은 성능을 보이고 있다. 하지만 학습 데이터의 양과 모델의 파라미터 수가 많아질수록 GPU 메모리와 시간 등의 더 많은 자원이 필요하여 현실적으로 큰 모델의 학습에는 어려움이 있다.

본 논문에서는 이러한 문제점을 개선하기 위해 RoBERTa모델에 ALBERT[5]에서 제안한 Embedding parameterization과 Parameter sharing을 적용하여 모델의 파라미터 개수를 줄인 *AL-RoBERTa*를 제안하고, 점차적으로 파라미터가 증가하는 모델에 맞게 *Wider-Net*[6]을 이용하여 초기 파라미터를 설정하여 학습 효율을 높인다. 또한 사전학습된 언어모델을 여러 자연어처리 태스크에 적용하고 성능을 측정하여 효과성을 보인다.

2. 관련연구

최근 대용량 말뭉치를 이용하여 사전 학습한 Transformer [4] 기반의 언어모델이 다양한 자연어처리 태스크에서 기존 방법들을 뛰어넘는 성능을 보여주고 있다.

구글에서 발표된 BERT[1]는 학습 말뭉치 문장의 토큰들을 랜덤으로 마스킹하고 마스킹 된 위치의 토큰을 예측하는 Masked LM과 연속하는 두 문장의 순서가 적절한지를 예측하는 NSP(Next Sentence Prediction)을 이용하여 사전학습을 진행한다. 이를 개선한 RoBERTa[2]는 NSP의 효용성에 의문을 제기하며 NSP Loss를 제거하였고, 하나 이상의 문서를 이용하여 최대 토큰 길이에 가깝게 구성하는 FULL-SENTENCES와 같은 개선된 학습 방법을 제시하였다. 또한 고정된 마스크 위치를 학습하는 BERT와 달리 마스크의 위치를 동적으로 결정하는 Dynamic Masking을 통하여 성능을 향상시켰다.

ALBERT[5]는 RoBERTa와 마찬가지로 NSP의 문제점을 인식하고 이를 개선한 SOP(Sentence Order Prediction)를 제시하였는데, SOP는 연속되는 두 문장(Positive)과 문장 순서를 앞뒤로 바꾼 문장(Negative)을 이용하여 문장의 순서가 옳은지를 예측하는 방식이다. 또한 ALBERT에서는 히든 차원과 동일한 차원을 가지던 임베딩 차원을 따로 분리하여 더 작은 차원으로 적용하였고, 각 Layer의 파라미터를 공유하는 방식을 통하여 모델의 파라미터를 줄이고 학습시간을 단축시켜 더 큰 모델의 학습이 가능하게 했다.

Transformer 기반 언어모델의 사전학습에 사용되는 학습 데이터가 많고 모델의 크기가 클수록 성능이 증가한다고 알려져 있다. 하지만 자원의 한계가 존재하기 때문에 학습 효율성(Training efficiency)이 중요한 문제로 부각되고 있다. [7]에서는 학습 효율성을 높이기 위해 BERT의 Encoder layer를 조절하여 얕은 모델을 쌓아 깊은 모델로 만드는 Progressively Stacking을 적용하여 학습시간을 감소시켰고, [6]에서는 기존 학습된 교사(Teacher) 모델의 파라미터를

확장된 학생(Student) 모델의 초기 파라미터로 재사용하여 동일한 학습시간에 더 빠르게 모델이 학습될 수 있도록 하였다.

3. 한국어 AL-RoBERTa 모델

본 논문에서는 기존의 RoBERTa 모델에 ALBERT의 다음의 두 가지 요소를 확장 적용하여 학습을 진행하였고, 파라미터의 개수는 large 모델에서 약 20배 감소하였다.

Factorized Embedding Parameterization 기존의 BERT 기반 모델은 임베딩 차원 E의 크기가 히든 차원 H의 크기와 동일하게 고정되어 적용된다. 대체로 큰 사이즈로 구성되는 사전의 크기(Vocabulary, V)에 따라 적용되는 임베딩 파라미터 수를 줄이기 위해 ALBERT에서는 임베딩 차원을 히든 차원의 크기보다 작게 설정하여 파라미터의 수를 $O(V \times H)$에서 $O(V \times E + E \times H)$로 감소시켰다.

Cross-layer Parameter sharing BERT 모델은 여러 개의 Transformer 블록을 쌓아올려 Encoding Layer를 구성하게 되고 이들은 크게 Attention 단계와 FFN(Feed-forward Network) 단계를 거치게 된다. ALBERT에서는 이렇게 구성된 Layer의 Attention과 FFN의 파라미터를 공유하도록 학습을 진행하여 파라미터의 개수를 줄였다.

3.1 자소 단위 BPE 토크나이저

한국어 모델에서의 입력은 대부분 형태소 단위나 음절 단위의 토큰을 사용하게 되는데 형태소 단위의 토큰은 형태소 분석기의 오류의 전파와 OOV(Out Of Vocabulary)의 문제점이 존재하고, 음절 단위 토큰은 입력 토큰의 의미를 구분하는데 어려움이 있다. 따라서 본 논문에서는 [8]에서 사용된 자소 단위 BPE 토크나이저를 사용하여 학습을 진행하였다. 다음은 토크나이저 적용 예시이다.

입력 문장: 난 프랑스 영화가 이래서 좋다..
결과 토큰: _난 _프랑스 _영화 가 _이래 서 _좋다 . .

3.2 Wider-Net 기반 AL-RoBERTa 학습 안정화

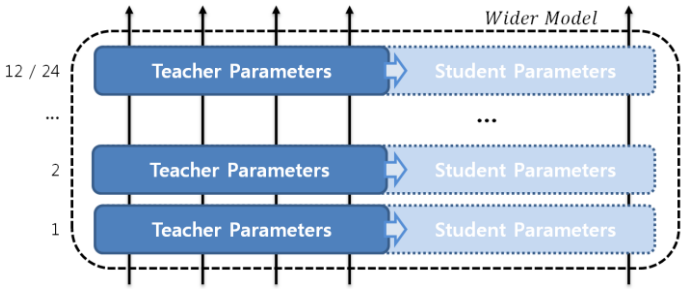

그림 1. Wider-Net 적용 Transfer Learning

본 논문에서는 약 500MB의 위키피디아 말뭉치 데이터를 사용하여 한국어 AL-RoBERTa 모델의 학습을 진행하였다. 또한 기존 학습된 모델의 재사용성과 빠른 학습을 위해 [6]에서 제안된 Wider-Net을 적용하였다. Wider-Net은 그림 1과 같이 확장 모델을 학습할 때 처음부터 다시 학습을 진행하는 것이 아닌 기존 학습된 모델의 파라미터를 이용하여 초기 파라미터를 설정함으로써 동일한 학습시간에 더 빠르게 학습이 가능하게 하였고 학습 모델의 재사용성을 확보하였다.

제안 모델에 Wider-Net을 적용하기 위해서 학습이 완료된 large 모델(12 layers, 1024 hidden size, 12 attention heads, 4096 intermediate size)을 교사 모델로 사용하여 학생 모델인 xlarge의 초기 파라미터로 사용하였다. 또한 확장 모델의 동일한 출력을 위해서 확장된 가중치를 1/n로 나누었고 동일한 attention heads를 가지도록 설정하여 학습을 진행하였으며, 확장된 파라미터가 동일한 기울기로 학습되는 것을 막기 위해 각 가중치의 스케일을 고려해 확장 파라미터에 Gaussian Distribution을 노이즈로 추가하여 적용하였다. 실험에 사용된 모델의 파라미터 개수와 임베딩, 히든 차원의 크기는 다음의 표 1과 같다.

표 1. 실험 모델별 파라미터 개수와 설정

Model		#Param	#Layer	H	E	Param sharing
RoBERTa	base	110M	12	768	768	False
	large	340M	24	1024	1024	
	xlarge	1278M	24	2048	2048	
AL-RoBERTa	base	11M	12	768	128	True
	large	17M	24	1024	128	
	xlarge	55M	24	2048	128	
	xxlarge	207M	12	4096	128	

4. 한국어 자연어처리 실험

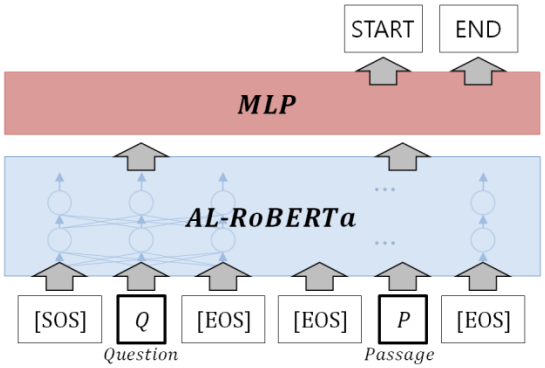

그림 2. 기계독해 모델 구조

본 논문에서는 감성분석과 개체명 인식, 기계독해에 fine-tuning 적용하여 사전학습 모델의 성능을 확인하였고 각 태스크 별 모델의 구조는 그림 2와 유사하게 구성된다. 형태소-태그[9,10]는 사전(Vocabulary)을 형태소와 BPE로 구성한 하이브리드 토크나이저를 사용한 모델이며, 본 논문에서는 자소단위 BPE를 사용하였다.

4.1 한국어 영화리뷰 감성분석

한국어 감성분석에 사용한 데이터 셋은 네이버 영화리뷰 감성분석 데이터[11]를 사용하였으며, 사전학습을 적용하기 위해 "[SOS] Context [EOS]"를 입력으로 사용하여 AL-RoBERTa의 마지막 레이어의 출력 값을 얻고 양방향 LSTM을 적용하여 최종 히든 상태로 fine-tuning을 진행하였다.

표 2. 한국어 감성분석 실험 결과

Model	#Param	학습데이터 크기	정확도
BERT(형태소태그)[9]	110M	500M	86.57

RoBERTa(형태소태그)[10]	110M	15G	89.88
RoBERTa	110M	500M	85.17
AL-RoBERTa	**55M**	**500M**	**86.05**

한국어 감성분석 실험 결과 BERT 모델과 비교하여 적은 수의 파라미터를 가짐에도 비슷한 성능을 보였다. 또한 동일 데이터와 토크나이저를 이용하여 학습한 RoBERTa의 결과와 비교하였을 때 더 높은 성능을 보였다.

4.2 한국어 개체명 인식

개체명 인식에 사용한 데이터 셋은 ETRI 엑소브레인 언어분석 말뭉치 데이터를 사용하였으며 감성분석과 마찬가지로 "[SOS] Context [EOS]"가 입력으로 들어가게 된다. AL-RoBERTa의 마지막 레이어의 출력 값에 양방향 LSTM을 이용해 인코딩하고, CRF를 이용하여 출력층을 구성하였다.

표 3. 한국어 개체명 인식 실험 결과

Model	#Param	학습데이터 크기	F1
BERT(형태소태그)[9]	110M	500M	91.58
RoBERTa(형태소태그)[10]	110M	15G	94.79
RoBERTa	110M	500M	91.94
AL-RoBERTa	**55M**	**500M**	**91.87**

표 3은 한국어 개체명에 대한 실험 결과로 다른 모델들에 비해 약 2배 적은 파라미터를 가짐에도 동일한 데이터로 학습하였을 때 더 높거나 비슷한 성능을 보였다.

4.3 한국어 기계독해

한국어 기계독해 태스크는 KorQuAD[12] 데이터 셋을 사용하여 평가를 진행하였다. 입력 토큰 시퀀스는 "[SOS] Question [EOS] [EOS] Passage [EOS]"로 구성되며, AL-RoBERTa의 마지막 출력 값에 MLP를 이용하여 start point, end point를 각각 얻어내어 정답을 예측한다. 표 4에서 [13]은 google-multilingual 모델을 사용한 것으로, 상위 100개 언어의 위키피디아 말뭉치의 대용량의 데이터를 가지고 학습을 진행한 모델이다.

표 4. 한국어 기계독해 실험 결과 : KorQuAD

Model	#Param	학습데이터 크기	EM	F1
BERT+MCAF (google-multilingual)[13]	110M	N/A	83.01	91.43
BERT-ETRI[14]	110M	23.5G	84.82	92.74
KT RoBERTa[8]	110M	18G	87.11	94.47
RoBERTa	110M	500M	78.63	88.25
AL-RoBERTa	**55M**	**500M**	**82.98**	**91.44**

실험 결과 대용량 데이터를 사용한 여러 모델과 비교하였을 때, 학습 데이터와 모델 파라미터의 차이가 큼에도 불구하고 비슷한 성능을 보여주었고, 특히 동일 데이터와 토크나이저를 이용하여 학습한 RoBERTa 모델과 비교하였을 때 큰 성능 향상을 보였다.

5. 결 론

본 논문에서는 사전학습 언어모델의 자원의 한계를 해결하기 위해 RoBERTa 모델에 파라미터 수를 줄인 AL-RoBERTa 모델을 제안하고 여러 한국어 태스크에 적용하여 성능을 얻었다.

모델은 적은 수의 파라미터와 데이터를 가짐에도 기존의 모델과 비슷한 성능을 보였고, 특히 KorQuAD에서 동일 데이터를 사용하여 학습한 RoBERTa 모델과 비교하였을 때 높은 성능향상을 보여주었다. 향후 계획으로는 더 많은 데이터를 이용하여 더 큰 학습 파라미터 가지는 모델의 학습을 진행하여 본 모델의 효과성을 검증할 예정이다.

참고문헌

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[3] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5754-5764).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[5] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[6] Chen, T., Goodfellow, I., & Shlens, J. (2015). Net2net: Accelerating learning via knowledge transfer. arXiv preprint arXiv:1511.05641.

[7] Gong, L., He, D., Li, Z., Qin, T., Wang, L., & Liu, T. (2019, May). Efficient training of bert by progressively stacking. In International Conference on Machine Learning (pp. 2337-2346).

[8] 최윤수, 이혜우, 김태형, 장두성, 이영훈, 나승훈. (2019). RoBERTa 를 이용한 한국어 기계독해. 한국정보과학회 학술발표논문집, 353-355.

[9] 박광현, 나승훈, 신종훈, 김영길. (2019). BERT 를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정. 한국정보과학회 학술발표논문집, 584-586.

[10] 민진우, 나승훈, 신종훈, 김영길. (2019). RoBERTa 를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱. 한국정보과학회 학술발표논문집, 407-409.

[11] https://github.com/e9t/nsmc

[12] 임승영, 김명지, 이주열. (2018). KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋. 한국정보과학회 학술발표논문집, 539-541.

[13] 박광현, 나승훈, 최윤수, 장두성. (2019). BERT 와 Multi-level Co-Attention Fusion 을 이용한한국어 기계독해. 한국정보과학회 학술발표논문집, 643-645.

[14] 이동헌, 박천음, 이창기, 박소윤, 임승영, 김명지, 이주열. (2019). BERT 를 이용한 한국어 기계 독해. 한국정보과학회 학술발표논문집, 557-559.