SciERC 데이터셋의 개체명 인식 Task 성능 개선

김은희[1], 황명권[1,2]
[1]한국과학기술정보연구원 인공지능기술연구단
[2]과학기술연합대학원대학교 데이터및HPC과학과
e-mail : ehkim@kisti.re.kr, mgh@kisti.re.kr

Performance Improvement of the NER Task in SciERC dataset

Eunhui Kim[1], Myunggwon Hwang[1,2]
[1]Korea Institute of Science and Technology Information
[2]University of Science and Technology

요 약

최근 버트계열의 언어 모델은 트랜스포머의 인코더모듈을 기반으로 학습한다. 최근 트랜스포머 계열의 언어 모델인 Ernie3.0, T5 및 Deberta 모델은 10종의 태스크로 구성된 SuperGLUE 벤치마크에 있어 사람의 자연어처리 인지 성능을 앞섰다. 대용량 데이터 셋으로 기 학습된 언어 모델의 파라미터를 기반으로 적은 크기의 데이터 셋에서 태스크별 이차 학습을 하는 파인튜닝 성능은 데이터 셋의 크기 차이로 과 적합 문제를 지닌다. 본 연구는 기 학습 파라미터의 높은 계층을 재초기화를 적용하여 과 적합 해소를 통해 SciERC 개체명 인식 태스크에서 현재 Rank 2위의 성능을 확인하였다.

1. 서 론

최근 트랜스포머의 인코더 모듈을 기반으로 대용량 텍스트를 학습하는 버트 계열의 언어모델은 기 학습된 파라미터들을 시작으로 태스크별 이차 학습을 하게 되고 이를 파인튜닝이라 보통 명한다. 보통 일차 학습 시 대용량의 데이터를 기반으로 학습한다. 버트의 경우 약 3만개의 사전(vocabulary)를 구성하고 33억 개의 토큰으로 학습하였다. 학습된 모델의 크기는 버트기본 모델의 경우 그 파라미터의 크기가 1.1억 개(110M)로, 512 입력 크기, 12 계층의 블록 구성, 크기 768의 은닉계층, 12개의 어텐션 모듈로 구성된다[1]. 이렇게 기 학습된 파라미터를 시작으로 파인튜닝을 태스크별로 수행할 때 그 데이터의 크기가 작은 경우(1만개 이하의 샘플인 경우), 파인튜닝 성능이 떨어진다. 이를 최적화하기 위해 최근 Zhang의 연구에서는 대량의 데이터에 과 적합 된 기 학습 파라미터를 재초기화를 적용하여 파인튜닝 성능향상의 보고가 있다[2]. 본 연구는 대용량으로 학습된 기 학습 모델로 개체명 1만개 미만인 작은 크기의 데이터 셋인 SciERC 데이터 셋의 파인튜닝 성능 향상을 위해 학습 파라미터 재 초기화 적용을 통해, SciERC 데이터 셋의 개체명 인식 태스크에서 Rank 2위의 성능을 달성한다.

2. 관련 연구

2.1 딥뉴럴 네트워크의 계층의 역할에 대한 분석

딥뉴럴 네트워크의 구성에 있어 낮은 계층과 높은 계층의 학습 파라미터 역할에 대해, 비전 분야 및 자연어처리 분야 각각에서 분석이 이뤄져 왔다. 비전분야에서는 ResNet 계열의 컨벌루셔널 뉴럴 네트워크가 인간의 인지 능력을 2015년에 넘어섰다. SuperGlue 벤치마크를 기준으로 자연어 처리 분야에서 BERT 계열의 트랜스포머 모델이 인간의 인지 능력을 2020년에 넘어섰다. 현재 3개의 모델이 SuperGlue 벤치마크 기준 인간의 자연어 인지능력을 넘어섰다. 비전 분야에 있어 딥 뉴럴 네트워크의 각 계층의 역할에 대해 2009년 분석된 내용에 따르면, 낮은 계층의 파라미터는 이미지를 구성하는 방향과 색깔의 선 등 이미지의 구성요소들을 인식하는 데 사용되고, 높은 계층의 파라미터는 특정 객체 인식이 처리됨을 뉴럴 네트워크 구성 파라미터 분석을 통해 드러낸 바 있다 [3]. 자연어 처리 분야 각 계층별 특징을 분석한 연구에 따르면, 문장 전체의 특징을 분류하는 CLS 토큰(token)은 낮은 계층에서, 그리고 입력 처리되는 두 문장을 분류하는 SEP와 같은 특정 토큰(special token)은 중간 계층에서, “. 혹은 ,” 같은 다 빈도 단어는 높은 계층에서 Attention 평균값이 높게 나옴이 보고된 바 있다[4]. 즉, 하나의 토큰들의 의미 매핑은 높은 계층에서 이뤄지고 있음을 미루어 짐작할 수 있다.

3. SciERC 데이터 셋의 NER 태스크 파인 튜닝(Fine-Tuning) 성능 최적화

3.1 SciERC 데이터 셋과 NER 태스크

SciERC 데이터 셋은 12개의 인공지능 컨퍼런스의 요약문 데이터를 활용한 정보 추출 학습데이터이다. 전체 500개의 논문 요약문을 ACL RD-Tec2.0(2016)의 가이드에 따라 개체명(entity), 관계정보(relation information), 동일지시어(coreferences)에 대한 정보가 태깅(혹은 주석처리) 되어 있

다[5]. SciERC 데이터 셋은 일반적인 지식정보(Knowledge Information)가 인명, 지리 정보 등을 포함하는 것과는 달리, 과학기술 데이터에 초점을 맞추어진 태깅 가이드에 따른다. SciERC 데이터 셋의 개체명 정보는 7종류로 Task, Method, Evaluation Metric, Material, Other Scientific Terms, Generic으로 구분 지어진다. 즉, SciERC 데이터 셋에서 NER 태스크는 토큰의 Entity 타입을 8종류로 분류하는 문제이다. 전체 500개의 요약문은 2,867개의 문장으로 구성되어 있으며, 전체 개체명은 8,094개로 기 학습 데이터 셋에 비해 매우 작은 데이터 셋이다.

3.2 SciERC 데이터 셋의 NER태스크 파인튜닝 성능 최적화 과정

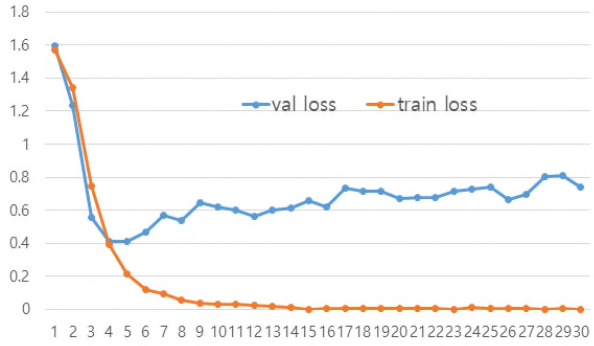
앞서 관련 연구를 통해 기 학습된 파라미터를 시작으로 파인튜닝 태스크 수행 시, 그 데이터 셋의 크기 차이가 클 경우 학습 성능의 저하 방지를 위해 출력 계층을 초기화한 구성 혹은 학습 시 학습 순서를 달리한 파라미터 재구성 방법이 있고, 최근에는 학습 파라미터의 일부를 재 초기화하는 방법을 적용하는 연구가 있다. 또한 딥러닝의 여러 계층에 대한 분석 내용을 요약해 보면 저 계층은 범용 적인 의미(즉, 전체 이미지의 기본 구성 요소 혹은 문서 전체의 카테고리)를 드러내고, 높은 계층일수록 지엽적인 의미(즉, 이미지 객체 단위 혹은 토큰의 의미)를 드러내고 있음을 확인할 수 있다.

본 연구에서는 이러한 특징들을 고려하여서 과학기술 데이터 셋으로 기 학습된 SciBert모델과 Scivocab[6]으로 학습을 시작하되 상위 계층의 파라미터를 재초기화 하는 방법을 적용하여 학습 성능을 향상시킬 수 있는 접근 방법으로 SciERC 데이터 셋에서 NER 태스크 수행 시 성능을 향상시켜 과학기술 데이터 셋에서 언어모델의 기본 태스크 성능 향상을 확인하였다.

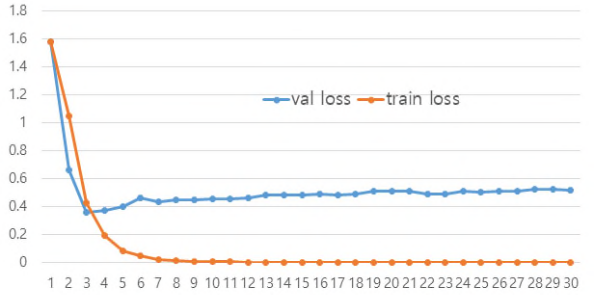
4. 실험

4.1 재초기화로 과 적합 방지를 통한 NER 태스크 파인튜닝 성능 개선

본 연구에서는 재초기화 과정이 과 적합 방지에 어떤 효과를 가져 오는지 확인하기 위하여, SciBert 기본 12 계층, Scivocab 기 학습 파라미터를 이용하여 재초기화 없이 학습한 경우와 재초기화 하여 학습한 경우 학습 손실과 검증 손실을 전체 데이터 셋의 70%와 30%로 나누어서 그 성능을 비교하였다. 그림3의 (a)는 재초기화 없이 SciBert, Scivocab 기 학습 파라미터를 학습한 경우이고, (b)는 SciBert, Scivocab 기 학습 파라미터 중 마지막 12번째 계층을 xavier normalization으로 재초기화 하여 학습한 경우를 비교한 것이다. 실험을 통해 확인할 수 있듯이 재초기화 과정은 과 적합 해소 효과가 있음을 확인할 수 있다.

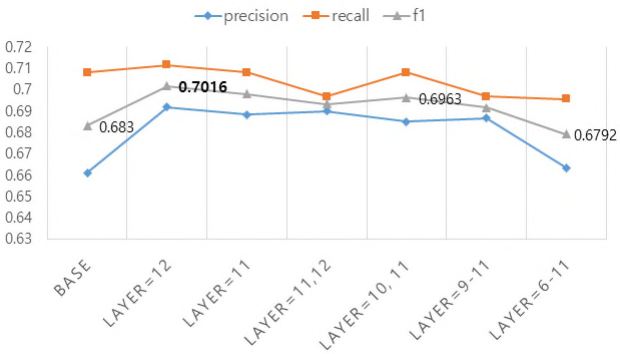


(a) 학습 손실과 검증 손실의 변화 (재초기화 하지 않은 경우)



(b) 학습 손실과 검증 손실의 변화 (재초기화 한 경우)

(그림 1) 재초기화를 통한 학습 손실로 확인되는 과 적합 해소 효과



(그림 2) 계층별 재초기화를 통한 NER 태스크 성능 변화

4.2 계층별 초기화 적용 시 파인튜닝 성능의 변화

관련 연구에서 비전 및 자연어처리 분야에서 딥러닝의 각 계층의 역할에 대한 분석을 통해 모델의 상위 계층으로 갈수록 지엽적인 데이터의 특징을 표현 학습하고 있음을 미루어 짐작할 수 있다. 이를 통해 재 초기화 과정은 특정 데이터 셋의 지엽적인 특징들에 임의의 특징을 부여 하게 되므로, 특정 태스크에 기 학습 파라미터의 영향을 줄이고 새로운 데이터 셋에 맞춘 향상된 성능을 기대할 수 있다. 예상 분석과 일치하게 그림 2의 실험 결과는 SciERC 데이터 셋에 대해 재초기화를 통해 초기화 과정이 없는 학습에 비해

1.76%향상된 성능으로, 현재 SciERC 리더 보드의 SOTA 모델인 SpERT의 70.33%의 F1 score보다 0.17%차이로 70.16%의 정확도로 2위의 성능을 보임을 확인 할 수 있다.

5. 결론

과학기술에 특화된 SciERC 데이터 셋의 NER 태스크는 전체 NER 개체명의 개수가 8,094개인 매우 작은 데이터 셋에 해당된다. 이러한 작은 데이터 셋을 기 학습된 파라미터들로 학습시킬 경우 과 적합 문제로 성능 향상의 제약이 따른다. 본 연구에서는 과 적합 해소를 위해 기 학습 파라미터 중 높은 계층을 재 초기화 하는 방법을 적용하여 SciERC 데이터 셋의 개체명 인식 태스크에서 2021년 10월 18일 기준 Rank 2위의 성능을 확인하였다.

Acknowledgement

본 연구는 2021년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.

참 고 문 헌

[1] Jacob Devlin, Ming-Wei Chang, Keton Lee, Kristina Toutanova,"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the Conf. of the Association for Computational Lingustics: Human Language Technonlogies, 2019.

[2] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weiberger, Yoav Artzi, "Revisiting Few-sample BERT fine-tuning," ICLR 2021.

[3] Honglark Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng, "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations," Proceedings of the 26th ICML, 2009.

[4] Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning, "What Does Bert Look At? An Analysis of BERT's Attention," Proceedings of the Second BloackBoxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pp. 276-286, 2019.

[5] Yi Luan, Luheng He, Mari Ostendorf, Hannaneh Hajishirzi, "Multi-task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction," Proceedings of the 2018 Conf. on EMNLP, pp. 3219-3232, 2018.

[6] Iz Beltagy, Kyle Lo, Arman Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," EMNLP 2019.