서울대학교 대학원 융합전공 혁신의과학 전공

학부연구생인턴 보고서

연 구 기 간: 2022. 12. 26 ~ 2023. 02. 17
연 구 제 목: Dataset Is All You Need - MedNLI
성 명: 심 재 준 (인 또는 서명)
지 도 교 수 성 명: 최 진 욱 (인 또는 서명)

1. 연구요약

생물의학에서 생성되는 정보들 중 자연어처리에 접목 가능한 환자 진료기록부로부터 생성된 자연어추론 자료를 Transformer 구조의 모형인 BERT와 그 변형 모형들을 사전학습 자료의 작업 혹은 분야 특화된 상황에서 어떠한 성능을 보이는지 실험한다. 자료와 모형들의 특징을 파악하고 조합하는 과정에서 어떠한 이유에서 각 조건들을 설정했으며, 그에 대한 결과를 풀이한다. 각 모형의 구조와 자료가 어떤 상황에서 좋은 결과를 도출할 수 있는지 역시 생각해본다.

2. 서론 (연구배경 및 연구목적)

자연어처리[1] 연구는 학계와 산업계를 통틀어 활발히 연구가 진행되는 분야다. 생물의학[2] 분야에서 생성되는 정보들 중 자연어처리에 접목 가능한 부분도 존재하며, 이는 생물공학에서 진행되는 연구들 중 정보처리 연구의 일부로 존재한다.

생물의학 분야의 언어는 일반적인 언어와는 다른 특수성을 가진다. 그 중 병원 환자의 진료기록부를 대표적인 예로 들 수 있다. 진료기록부는 환자의 병원 증상과 그에 대한 의사의 소견, 치료 내용, 처방 등의 행동 기록이 담긴다. 기록하는 과정에서 정보들은 지역의 언어와 생물의학 분야의 전문용어들을 약어 [1]와 특수문자의 조합으로 표현

[1] Natural Language Processing (NLP)

[2] Biomedicine

된다. 작성자에 따라서 활용하는 언어의 조합이 다르고, 약어와 특수문자가 문맥에 따라 각각 다른 의미를 가진다. 여기서 비롯되는 상황 중 하나는 환자의 담당의가 바뀌는 상황에서 환자 진료기록부의 과거 기록에 대해 해석하며 가설을 가지는 경우 해소할 방법은 작성자에게 직접 요청하는 방법이 가장 확실하다. 여기서 요청을 받는 대상이 자연어처리 모형으로 대체 혹은 도움을 줄 수 있는 가능성에 대해 알아보려고 한다.

자연어추론 자연어처리의 많은 작업들 중 자연어추론[3]은 전재와 가설이 주어졌을 때에 이에 대한 논리적인 참과 거짓, 그리고 중립을 가리는 작업이다. 위에서 언급한 작업에 적합한 언어처리 접근 방법이다.

대표적인 자료로 SNLI[4] [2]와 MultiNLI[5] [3]가 있으며, 모두 연구를 위해 공개된 자료이다. SNLI는 Flickr[6] 출처의 약 57만개의 인간이 쓴 문장 쌍으로 이루어져 있다. 학습 모형의 성능을 평가하는 용도로 잘 정제되어 품질이 좋고, 각 분류의 분포가 고르게 된 것이 특징이다. MultiNLI는 SNLI 형태와 동일하며, 약 43만개의 문장 쌍으로 이루어져 있다. 대면 대화, 편지, 전화, 여행, 소설, 공문서 등 다양한 장르의 언어적 현상을 다루는 것이 특징이다.

생물의학 언어 자료 환자의 개인정보로 인한 접근 권한 제한과 전문성의 요구로 제3자의 주석 작업이 어려운 점과 전문가의 높은 작업 비용으로 공개 되어있는 자료가 많지 않다. 대체적으로 생물의학 분야의 자료는 단체의 자체 자료이거나 아래에서 언급할 두 가지 자료의 변형된 자료이다. 이 실험에서 사용되는 자료는 MedNLI[7] [4]이며, 이는 MIMIC-III[8] [5]의 진료기록부를 자연어추론 작업에 맞게 변형된 자료이다.

MIMIC-III는 실제 현장에서 발생한 의료 정보들을 MIT 연구실[9]에서 모아둔 자료다. 2001년에서 2012년 동안 미국 보스턴에 위치한 Beth Israel Deaconess Medical Center에서 의료기록이 있는 환자 중 동의를 구한 약 4만명의 정보를 제공한다. 정보 항목은 활

[3] Natural Language Inference (NLI), Recognizing Textual Entailment (RTE)

[4] The Stanford Natural Language Inference

[5] Multi-Genre Natural Language Inference

[6] at (https://www.flickr.com/)

[7] Natural Language Inference in Clinical Texts at (https://jgc128.github.io/mednli/)

[8] Medical Information Mart for Intensive Care III at (https://mimic.mit.edu/docs/iii/)

[9] Laboratory for Computational Physiology, MIT

력 징후, 실험실 결과, 약물, 간병인 메모, 진료기록부, 사망률 등을 포함한다. 지속적인 공개 작업을 통해 현재는 MIMIC-IV까지 존재한다.

PubMed[10]는 생물의학 분야의 글을 담은 MEDLINE 약 3천5백만개 이상의 인용과 초록을 가지고 있으며, 매년 약 1백만개의 글이 추가된다. PubMed Central은 전체 글을 담고 있다. 그리고 외부의 글일 경우 인용에서 연결 주소를 활용하면 된다. 지속적인 관리가 이루어지고 있으며, 분야의 전문적인 자료에 접근하기 좋다.

Transformer 언어 모형은 문맥을 이해하는 과정에서 점차 커지며, 다양한 작업을 수행할 수 있도록 변화했다. 자연어추론 작업에서 쓰이는 가장 대표적인 모형은 Transformer [6] 기반 Encoder-Decoder 결합 구조의 사전학습[11] 모형이다. 사전학습 모형은 자연어처리에서 큰 두각을 보이고 있으며, 대체적으로 일반적인 분야의 큰 규모 말뭉치(기사, 웹글) 자료를 활용한 사전학습이 좋은 성능을 보이고 있다.

3. 실험방법

모형은 Transformer 구조의 대표적인 BERT [7] 모형과 그 구조체의 변형들을 활용한다. 자연어추론 작업에서 어떤 특징의 결과를 보여줄지, 그리고 자연어추론 자료와 생물의학 정보 자료에서의 사전학습은 어떤 특징의 결과를 보여줄지 알아본다.

학습자료 - MedNLI 자연어추론 작업이 진행되기 위한 생물의학 정보 자료의 확보는 사전 작업이 필요하다. 특히 환자의 정보를 담은 의료자료의 경우 필수적이다. MIMIC-III 자료의 경우 연구목적에서 활용으로 제한을 두고 있고, 제공받기 전 자료 사용 목적과 사용자의 정보 등을 제출하고, 허가를 받으면 활용 교육을 진행한 뒤 사용이 가능하다. MedNLI는 MIMIC-III의 약 4만명가량의 환자에 대한 진료기록부 약 2백만개 중 결정적인 추론이 가능한 부분을 선정하였고, SNLI와 동일한 형태의 전체 약 1만4천개로 구성됐다. 표 1 [4]는 자료의 일부이며, 전제[12]에 해당하는 부분이 각각 3번 연속적으로 출현하고,

[10] at (https://pubmed.ncbi.nlm.nih.gov/)

[11] Pre-Trained Model (PTM)

[12] Premise

가설[13]에 대한 참, 거짓, 중립을 결정한다. 이는 자료를 모형에 학습시에 낮은 Batch 크기를 가지는 상황에서 과적합을 초래할 수 있다. 따라서 이를 피하기 위해 순서를 무작위로 섞어서 학습을 진행한다. 자료의 구성 중 이진 구문분석자료는 문장의 의미를 파악하기에 좋은 이점을 가진 것으로 보이지만 구문분석도 Transformer 모형의 성능을 보여주는 중요한 부분이기에 사용하지 않는다.

BERT와 변형 모형 BERT는 Transformer구조를 활용한 대표적인 모형으로 큰 말뭉치를 그림 1 [7]과 같이 사전학습하고, 작업에 맞는 자료를 전이학습으로 최적화하여 사용한다. BERT-base의 경우 약 12개의 Transformer층과 약 1억개의 매개변수를 연산한다. 특징적으로 MLM[14] [7]은 자료 중 무작위 15%의 80% 위치를 [MASK] token으로 대체하여 학습하며, NSP[15] [7]는 말 그대로 다음에 올 문장을 예측하는 방법이다. 그림 2 [7]와 같이 이는 두가지 문장이 주어지는 자연어추론 작업에서 이점을 가질 것이라는 가설을 가질 수 있다.

RoBERTa [8]는 BERT에 비해서 더 유동적이게 [MASK] token을 생성하고, NSP 과정을 제외한다. 그리고 성능을 높이기 위해 모형의 규모를 키우고, 더 긴 문장들을 가진 더 많은 말뭉치를 학습하는 방법을 제안했다. 그리고 이와 같은 방법으로 NSP를 사용하는 BERT에 비해 더 좋은 성능을 가질 수 있었다. 하지만 자연어추론에서는 이점을 가지는 요소가 줄었다는 가설을 가질 수 있다.

DeBERTa [9, 10]는 encoder에서 위치를 사용하고, 그림 3 [9]과 같이 절대적 위치를 decoder에서 활용한다. DeBERTa-V3 [10]는 ELECTRA 사전학습 방식을 채택하여, MLM를 RTD[16] [10]로 대체한다.

사전학습 - 작업 및 분야 특화자료 사전학습 모형의 경우 대체적으로 일반적인 분야의 큰 규모 말뭉치 자료를 통해 학습한다. 다수의 GPU[17]들을 통한 병렬연산이 큰 규모 모형의 회당 연산 속도를 큰 폭으로 줄여주며 사전학습이 진행된다. 사전학습 정보를 보면

[13] Hypothesis
[14] Masked Language Model
[15] Next Sentence Prediction
[16] Replaced Token Detection
[17] Graphics Processing Unit

대체적으로 공개된 사전자료들과 책의 글, 그리고 인터넷 소셜 미디어 정보 등 다양한 구성을 담고 있다. 표 2는 연구에서 활용되는 모형의 기본자료 구성이다.

반면 적용 분야에 특화된 언어로 이루어진 자료를 활용하여 사전학습된 모형은 자료의 적용 방식에 따라서 특징이 나뉜다. 연구 사례 [11]를 통해 분야에서 벗어난 언어를 포함하는 자료보다 순수하게 분야에 특화된 언어만 포함하는 자료를 학습하는 경우 관련 자연어처리 작업에서 대체적으로 우세한 성능을 보였다. 이를 지속적인 사전학습[18]과 처음부터 사전학습[19]으로 구분한다. BERT의 특징은 제한된 길이의 자체적인 어휘사전을 가지고, 자가 지도학습[20]을 통해서 지속적으로 이를 Token-embedding에 활용한다. 보통 합성어를 구분하기 위해 새로운 분절을 만드는 BPE[21] [12] 방식을 이용하기에 어휘사전은 단순히 독자적인 단어만 포함하지는 않는다. 따라서 사전학습에서 이미 어휘사전을 모두 정의했다면, 전이학습 과정에서 새로운 단어가 출현해도 이를 반영하지 못하는 특성을 가진다. 또한 사전학습에서 정의된 어휘사전을 다른 모형에서는 활용하지 못한다.

여기서 맥락 파악을 위해서는 작업 및 분야 특화자료를 처음부터 사전학습하는 상황이 이상적이라는 가설을 세울 수 있다. 그로 인하여 가지는 이점은 분야에 최적화된 어휘사전을 생성할 수 있다는 부분에서 온다. 생물의학 특화자료는 PubMed 자료가 학습된 PubMedBERT [11]가 사용될 것이며, 이는 BERT-base에 사전학습된 상태이다. MEDLINE의 초록들 중 모형의 한 번에 수용할 수 있는 token 개수 제한인 128개 이하의 단어를 가지는 자료만 고른 결과, 1천4백만개의 자료에서 32억개의 단어와 21GB 크기를 가지는 자료를 사용한다. 자연어추론 작업에 특화된 자료인 SNLI와 MultiNLI를 활용하여 사전학습된 DeBERTa-V3-base도 자료 특성에 대한 사전학습 연구로 활용된다.

전이학습 - 미세조정 사전학습 가중치는 동결하지 않고 미세조정 방식으로 전이학습을 진행한다. 사전학습 모형의 논문은 대체적으로 전이학습시에 미세조정을 위해 어떤 설정값의 상황에서 최적의 성능을 보여주는지 명시하고 있다. 하지만 작업과 학습 자료의 상태에 따라 원하는 성능이 나오지 않을 수 있다. 또한 연구환경의 자원 한계성에 따라서 원하는 만큼의 학습을 진행하기 어려울 수 있다. 성능평가는 F1-macro를 활용했으며, 3

[18] Continual pre-training

[19] Pre-trained from scratch

[20] Self-supervised learning

[21] Byte Pair Encoding

번의 학습단계를 지나는 동안 성능의 향상이 없으면 학습을 중단하는 방식으로 Early-stop callback 함수를 활용했다. 그리고 마지막에 중간 성능평가의 결과가 가장 좋게 나온 가중치를 활용하여 최종 성능평가에 사용했다.

4. 실험결과

학습 연산은 Google Colab[22]의 NVIDIA Tesla P100 16GB GPU로 진행됐다.

결과적으로 표 3의 내용대로 주어진 환경에서 자료와 작업에 맞는 최적의 조건을 찾기 위해 추가적으로 가중치 변형 시점을 Batch의 크기대비 조절가능한 Accumulation과 학습 속도를 조절하는 Learning-rate schedule(Warm-up), Weight-decay 등의 hyperparameter를 설정하게 되었다. 학습 평가 결과와 분류결과로부터 혼동의 정도를 나타내는 정보는 Figure 1와 Figure 2참고.

5. 고찰

Figure 1의 정보를 보면 자연어추론 작업에 맞는 자료로 사전학습한 NLI-DeBERTa-V3가 성능이 가장 좋게 나왔다. DeBERTa-V3의 경우 BERT에 비해 수용 가능한 어휘사전의 크기가 약 2만개 커졌다. 따라서 사전학습 자료의 다양성도 탑재가 가능한 상황에서 작업의 특수성 또한 같이 확보하기 용이하다. NLI-DeBERTa-V3는 원하는 작업과 사전학습 자료의 형태 유사성 확보로 느린 학습을 진행할 경우 성능 향상이 가능한 모습을 보였다. 느리게 학습을 하는 경우 중간검증에서는 손실이 천천히 낮아져서 다른 학습 경우에 비해 평가 성능을 기대하지 못할 수 있지만, 결과적으로는 가장 좋았다.

처음부터 생물의학 분야의 특화 자료로 사전학습한 PubMedBERT와 기본 자료로 사전학습한 DeBERTa-V3의 성능이 비슷하게 나왔다. 이는 분야에 특화된 자료의 경우 전문용어를 어휘사전에 포함하기에 문맥을 이해하는 과정에서 더 수월했을 것이고, 이를 분류하는 부분에서 추가적인 학습이 필요했다. 구조적인 차이로는 NSP가 자연어추론 작업에서 더 이점을 가졌을 것으로 판단한다. DeBERTa-V3는 [MASK] token의 절대적 위치와 RTD로

[22] at (https://colab.research.google.com/)

성능을 확보하였을 것이다. 반면 NSP가 없는 RoBERTa의 경우 상대적으로 낮은 성능을 보였다.

Figure 2에서 Entailment와 Neutral의 분류 결과를 보면 오분류된 분포가 Contradiction보다 높다. 이는 모든 조건에서 비슷한 비율 분포를 보였으며, 자료의 특성으로 파악된다. 긍정적인 요소로는 생물의학 분야의 특성상 기술의 안정적인 상황을 기대하며, 거짓은 높은 확률로 판별 가능한 것이 긍정적인 요소로 작용할 가능성이 있다.

6. 결론

환자 진료기록부는 과거의 자료로부터 현재 판단의 도움을 얻을 수 있다. 따라서 개인의 특성을 가리기 위해서는 누적된 자료의 활용이 중요하다. 하지만 MedNLI 자료는 문장 단위로 구성되어 있기에 개인의 특성을 담기에는 어려움이 존재한다. Transformer 모형은 반면에 이어지는 문장들을 이해하기 적합하다. 따라서 문맥 이해와 정확성, 신뢰성 확보를 위해 분야에 특화된 자료와 이어지는 사전학습으로 자연어추론 자료를 사용하여 사전학습이 이루어진 DeBERTa-V3에 환자의 모든 진료기록을 포함하는 자료를 요약하여 넣어주는 과정을 포함한다면 원하는 작업에 더 가까운 결과를 도출할 수 있을 것으로 보인다.

물리적인 환경 구성의 한계로 인하여 각 사전학습 모형의 미세조정 권장 지침사항을 모두 담기에는 어려움이 있었다. 의도에 맞게 성능을 파악하기엔 오차가 존재하지만, 결과를 구분하기엔 적당한 타협점을 찾을 수 있었다.

참고 문헌

[1] O. D. Kuzmina, A. D. Fominykh and N. A. Abrosimova, "Problems of the English Abbreviations in Medical Translation," *Procedia - Social and Behavioral Sciences,* vol. 199, pp. 548-554, 2015.

[2] S. R. Bowman, G. Angeli, C. Potts 그리고 C. D. Manning, "A large annotated corpus for learning natural language inference," %1 *Association for Computational Linguistics*, 2015.

[3] A. Williams, N. Nangia 그리고 S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," %1 *Association for Computational Linguistics*, New Orleans, Louisiana, 2018.

[4] A. Romanov 그리고 C. Shivade, "Lessons from Natural Language Inference in the Clinical Domain," %1 *arXiv*, 2018.

[5] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi 그리고 R. G. Mark, "MIMIC-III, a freely accessible critical care database," Scientific Data, 2016.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser 그리고 I. Polosukhin, "Attention Is All You Need," %1 *arXiv*, 2017.

[7] J. Devlin, M.-W. Chang, K. Lee 그리고 K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," %1 *arXiv*, 2019.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer 그리고 V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," %1 *arXiv*, 2019.

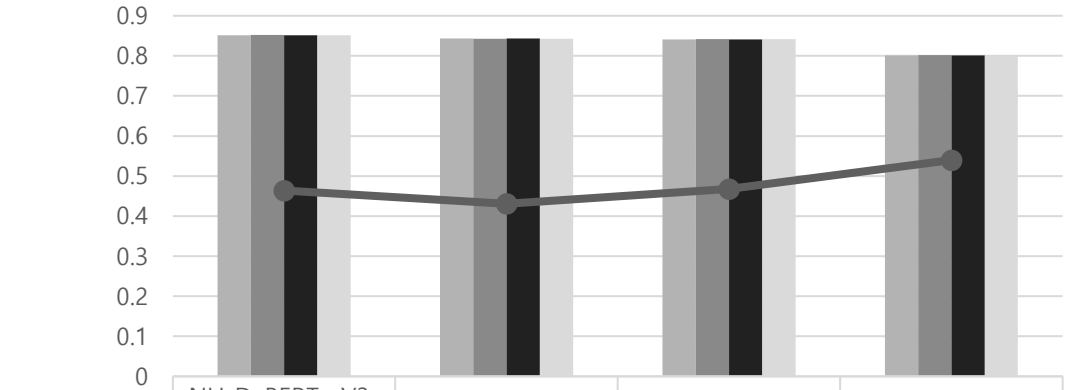
[9] P. He, X. Liu, J. Gao 그리고 W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," %1 *arXiv*, 2021.

[10] P. He, J. Gao 그리고 W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," %1 *arXiv*, 2021.

[11] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao 그리고 H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," %1 *ACM Trans. Comput. Healthcare*, 2022.

[12] R. Sennrich, B. Haddow 그리고 A. Birch, "Neural Machine Translation of Rare Words with Subword Units," %1 *arXiv*, 2016.

부록



	NLI-DeBERTa-V3-base	DeBERTa-V3-base	PubMedBERT-base	RoBERTa-base
accuracy	0.851617	0.843179	0.841069	0.800985
precision	0.852126	0.842684	0.842066	0.801805
recall	0.851617	0.843179	0.841069	0.800985
f1	0.851417	0.842889	0.84144	0.801239
loss	0.464043	0.430728	0.467523	0.539905

Figure 1

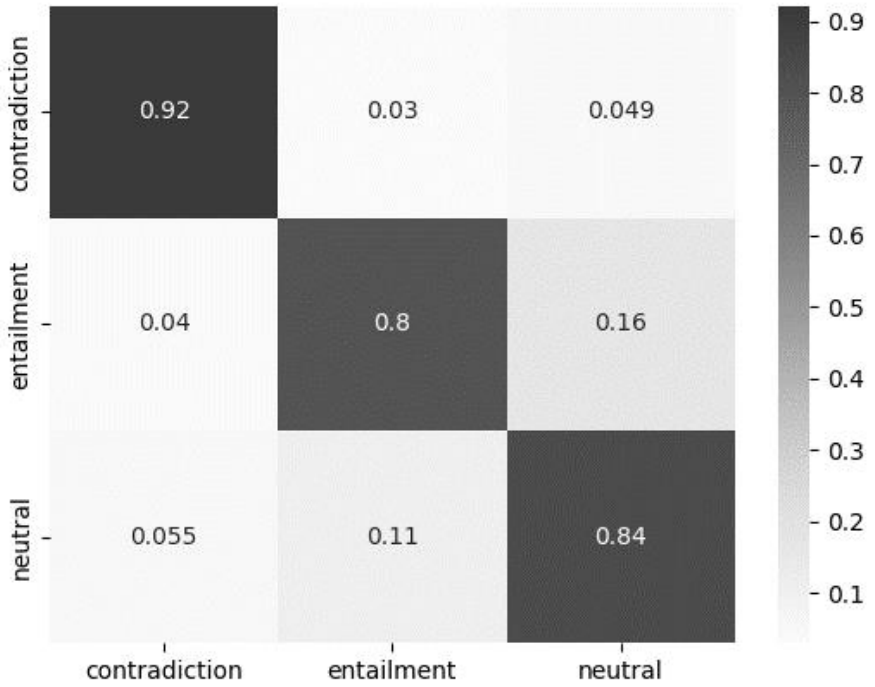


Figure 2

#	Premise	Hypothesis	Label
1	ALT , AST , and lactate were elevated as noted above	patient has abnormal lfts	entailment
2	Chest x-ray showed mild congestive heart failure	The patient complains of cough	neutral
3	During hospitalization , patient became progressively more dyspnic requiring BiPAP and then a NRB	The patient is on room air	contradiction
4	She was not able to speak , but appeared to comprehend well	Patient had aphasia	entailment
5	T1DM : x 7yrs , h/o DKA x 6 attributed to poor medication compliance , last A1c [** 3-23 **] : 13.3 % 2	The patient maintains strict glucose control	contradiction
6	Had an ultimately negative esophagogastroduodenoscopy and colonoscopy	Patient has no pain	neutral
7	Aorta is mildly tortuous and calcified .	the aorta is normal	contradiction

표 1

Model	Wiki+Book 16GB	OpenWebText 38GB	Stories 31GB	CC-News 76GB
BERT	✓			
RoBERTa	✓	✓	✓	✓
DeBERTa	✓	✓	✓	
DeBERTaV3	✓	✓	✓	✓

표 2

Model	epochs	train_batch	eval_batch	accumulation	learning_rate	weight_decay	warmup_ratio	warmup_steps	metric
NLI-DeBERTa-V3-base	3	8	32	1	5.00E-06	0.1	0.01	84	f1
DeBERTa-V3-base	3	8	32	4	3.00E-05	0.5	0.001	140	f1
PubMedBERT-base	3	16	64	4	3.00E-05	0.3	0.01	105	f1
RoBERTa-base	3	16	64	4	3.00E-05	0.9	0.001	105	f1

표 3

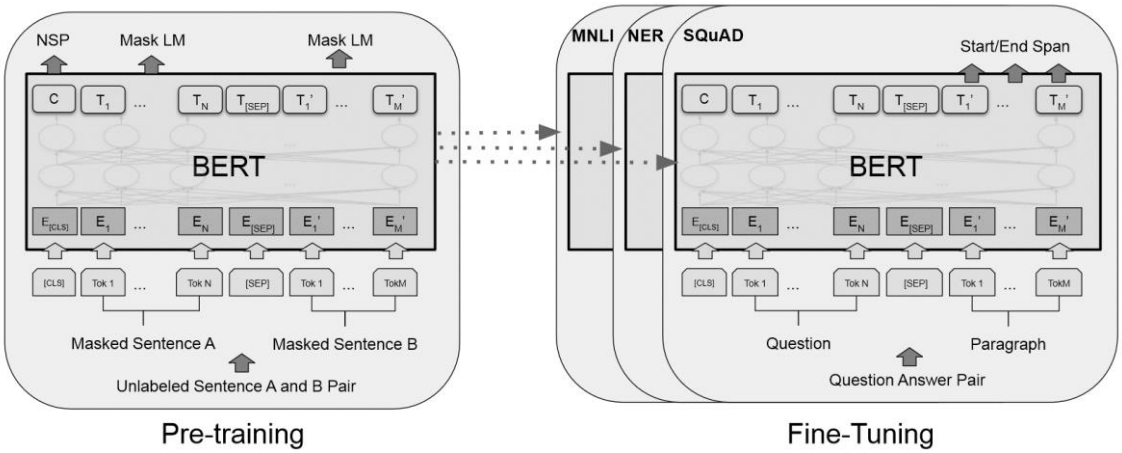


그림 1

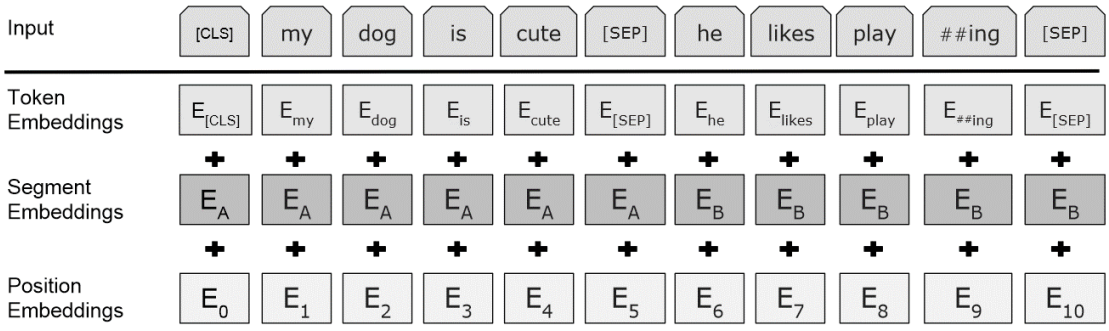
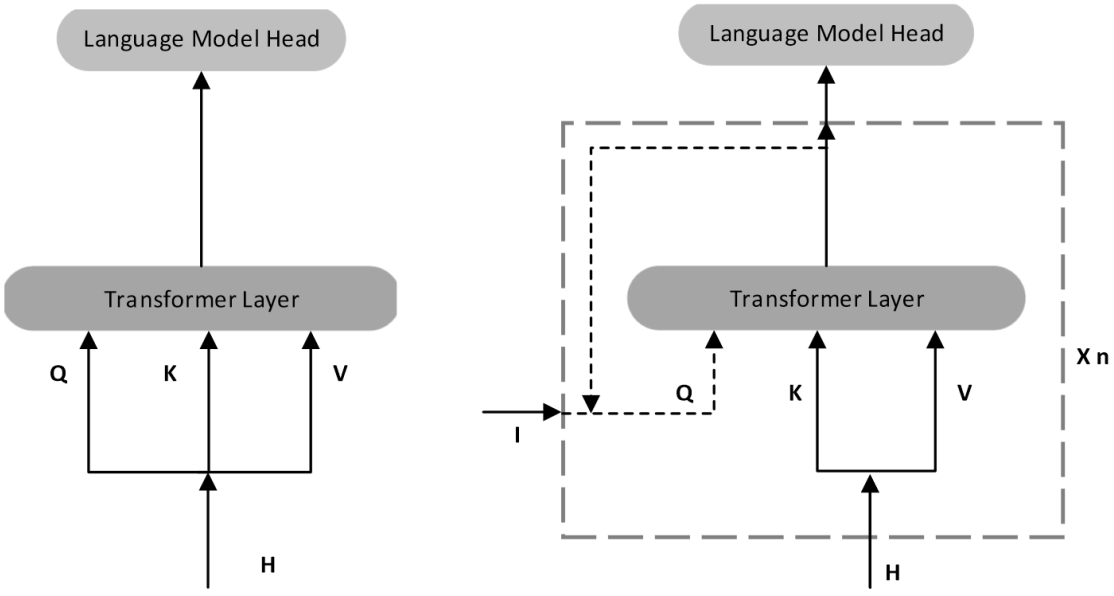


그림 2



(a) BERT decoding layer

(b) Enhanced Mask Decoder

그림 3