사전학습 **BERT** 기반 미세 조정 기법별 감정 분류 성능 평가

홍성규, 김상철*
국민대학교, *국민대학교

hsung951027@kookmin[1].ac.kr, *sckim7@kookmin.ac.kr

Evaluation of sentimental classification performace by fine-tuning methods based on Pre-trained BERT

Hong Sung Kyu, Kim Sang Chul*
Kookmin Univ., *Kookmin Univ.

요 약

BERT(Bidirectional Encoder Representations from Transformers)는 2017 년 구글에서 발표한 Transformer 기반의 양방향 문맥을 학습할 수 있도록 고안된 언어모델이다. 오늘날 언어모델의 규모가 커짐에 따라, 현실적으로 리소스 및 시간 등 학습하는데에 어려움이 수반하고 있다. 이러한 문제를 해결하기 위해 사전학습된 언어를 Fine-tuning 하는 과정에서 파라미터를 경량화 시키기 위해 대표적으로 AdapterFusion 과 LoRA 와 같은 Fully fine-tuning 을 대체하는 방법이 연구되고 있다. 본 논문에서는 새로운 Fine-tuning 기법들을 BERT 에 적용하며, 학습 시간 및 정확도에 어느정도 영향을 끼치는지 연구한다.

Ⅰ. 서 론

2017 년 구글에서 Transformer 구조[1]를 제안하면서 NLP 영역에는 기존의 RNN 혹은 LSTM 과 같은 셀들을 대체하는 새로운 패러다임이 시작되었다. 이러한 Transformer 기반의 모델들은 오늘날까지 발전해오면서 OpenAI 에서 발표한 GPT(Generative Pre–Training)[3]과 구글에서 발표한 BERT(Bidirectional Encoder Representations from Transformers)가 있다. 하지만, 그 밖에도 더욱 모델 파라미터의 규모가 커질수록 학습은 점차 어려워진다[5]. 그러므로, 모델의 전체 파라미터가 아닌 일부 파라미터만 재학습 시킬 수 있도록 새로운 Fine-tuning 기법들이 연구되고 있다. 대표적으로, Multi-task transfer learning 에 최적화하기 위한 AdapterFusion 과 Rank factorization 을 활용한 LoRA 가 있다. 본 논문에서는 두 가지 Fine-tuning 기법들을 BERT 에 적용한다.

Ⅱ. BERT (Bidirectional Encoder Representations from Transformers)

BERT 와 같은 Transformer 기반의 사전학습 언어모델(Pre-trained Language Model)은 주로 준지도 학습 방식(Semi-Supervised learning)을 따르게 된다. 이는 모델에 대량의 말뭉치(Corpus)를 비지도 학습(Unsupervised learning)을 먼저 한 후에, 해결하고자 하는 태스크에 적합한 데이터를 통해 미세 조정(Fine-tuning) 과정에서 지도 학습(Supervised learning)을 한다. BERT 는 GPT 와는 다르게 양방향으로 텍스트의 representation 을 학습하도록 설계되었다[2].

(1) BERT 의 입력

BERT 의 입력은 세 가지 임베딩을 합한 결과를 입력받는다. 첫 번째는 각 토큰을 벡터 Representation 인 Token embedding 이다. 두 번째는 Positional embedding 이다. Transformer 기반 모델은 RNN(Recurrence Neural Network) 처럼 입력을 순차적으로 받지 않기 때문에 각 토큰의 위치 정보를 부여해주어야 한다. 세 번째는 Segment embedding 이다. 문장이 여러개 입력될 경우, 각 문장을 구분하기 위해서 Segment embedding 이 사용된다.

(2) BERT 사전학습 과정

BERT 의 사전학습 과정은 Unsupervised 혹은 Self-supervised learning 으로, 대표적으로 두 가지이다. 첫 번째는 MLM(Masked Language Model) 이다. BERT 라는 이름에서 알 수 있듯이 Denosing AutoEncoder 의 역할을 한다. 그 이유는 입력 토큰에 대해 랜덤하게 마스킹 처리된 토큰들을 예측하기 때문이다. 이 과정에서 마스킹 처리를 노이즈로 해석하면 DAE 로써 해석이 가능하다. 두 번째는 NSP(Next Sentence Prediction) 이다. 한 쌍의 문장이 들어왔을 때, 두 문장이 다음 문장인지 예측하는 과정이다. 이러한 태스크는 문장 전체에 대한 이해를 학습한다.

(3) BERT 미세조정 과정

BERT 의 Fine-tuning 단계는 매우 단순하다. 이미 사전학습이 되어 있는 BERT 모델의 최상단에 해결하고자 하는 태스크에 적절한 Head 를 최상단에 배치하면 된다. 해당 과정은 사전학습 과정보다 적은 epoch 이 소요되기 때문에,

[1] *: 교신저자

상대적으로 적은 리소스로 학습이 가능하다. huggingface(https://huggingface.co/)에 배포되어 있는 사전학습된 모델을 Python 패키지인 transformers 를 통해 불러올 수 있다.

Ⅲ. Fine-tuning Approach 1 : AdapterFusion

최근 언어모델의 파라미터 규모가 점점 커짐에 따라 Fine-tuning 이 현실적으로 불가능해지고 있다. 계산을 위한 많은 컴퓨팅 리소스 혹은 시간이 소요되기 때문이다. 그 뿐만 아니라, 순차적 미세조정(Sequential fine-tuning) 혹은 다중-태스크 학습(Multi-task learning)과 같은 전이 학습에도 어려움이 있다. 이러한 문제를 해결하기 위해 2021 년 Jonas pfeiffer 에 의해 AdapterFusion 이라는 대안이 제시되었다[4]. AdapterFusion 는 Transformer 기반 모델에 적용할 수 있으며, 매커니즘은 크게 두 단계로 구성되어 있다. 첫 번째로, Transformer 레이어에 Task-specific adapter 를 주입한다. 두 번째는 Fusion 레이어를 적용하여, 장착된 Adapter 들에 Attention 매커니즘을 적용하여 현재 모델의 태스크에 적합한 Adapter 에 가중치를 부여하여 활성화 시킨다. AdapterFusion 은 기존 Transformer 기반의 언어모델의 Fully fine-tuning 을 하지않고, 모델의 가중치를 동결(Freeze)한 다음에, Adapter 와 Fusion 레이어에 대한 가중치만 학습하여 매우 적은 가중치(Single-Task adapter 기준, Pre-trained model 의 3.6%의 가중치)만을 학습한다. Adapter 가 적용된 모델은 AdapterHub(https://adapterhub.ml/)에서 Python 패키지인 adapter-transformers 를 통해 불러올 수 있다.

Ⅳ. Fine-tuning Approach 1 : LoRA

LoRA(Low-Rank Adaptation)은 2020 년 Aghajanyan 의 Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning 논문으로부터 영감은 받아 제안된 논문이다. LoRA 의 핵심은 $W_0 + \Delta W$ 에 Rank factorization 을 적용하여 $W_0 + BA$ 로 수식을 변경하는 것이다. ($B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$) 여기서, W_0 는 사전학습된 가중치 행렬을 의미하고, ΔW 는 누적된 그레디언트의 업데이트를 의미한다. 그리고, W_0 을 동결한 상태에서, B 와 A만 업데이트하기 때문에 기존 Fully fine-tuning 보다 적은 파라미터로 학습할 수 있게 된다. LoRA 는 AdapterHub 가 아닌, 마이크로소프트에서 공개한 깃허브 레포지토리(https://github.com/microsoft/LoRA)에서 적용이 가능하다. 현재 LoRA 는 Pytorch 프레임워크에서만 적용 가능하며, 사용가능한 레이어는 nn.Linear, nn.Embedding, nn.Conv2d 가 있다.

Ⅴ. 실험과정

실험을 위해 nsmc(네이버 영화 리뷰 데이터셋 : https://github.com/e9t/nsmc)를 통한 감정분류(Sentimental Classification)테스크를 진행했다. 실험을 진행한 모델은 세 가지로 다음과 같다.

(1) bert-base-multilingual-cased
(2) bert-base-multilingual-cased + ko/wiki@ukp pfeiffer adapter
(3) Bert-base-multilingual-cased + LoRA linear layers

실험은 Tesla V100 GPU 에서 15 만건의 학습 데이터를 5 만건씩 줄여가면서 하이퍼 파라미터를 수정하면서 진행했다. 학습을 위한 하이퍼 파라미터는 AdamW 옵티마이저와 learning rate 은 $2e^{-5}$, epsilon 은 $1e^{-8}$, epochs 는 4 로 지정했다. 성능 평가 지표는 정확도(Accuracy)를 사용했으며, LoRA 의 하이퍼 파라미터는 r 를 15 만건의 데이터를 학습한 케이스에서만 튜닝했다.

Ⅵ. 실험결과

	Training time per epoch (mm:ss)	Test accuracy	Precision	Recall	F1-Score
Vanilla	17:01 ~ 17:04	0.871	0.864	0.882	0.869
Adapter Fusion	12:29 ~ 12:30	0.858	0.856	0.863	0.855
LoRA (r=64)	10:59 ~ 11:01	0.726	0.745	0.692	0.711
LoRA (r=32)	10:52 ~ 10:53	0.722	0.741	0.688	0.707
LoRA (r=16)	10:50 ~ 10:51	0.716	0.751	0.651	0.691

표 1 15 만건에 대한 학습 결과

	Training time per epoch (mm:ss)	Test accuracy	Precision	Recall	F1-Score
Vanilla	11:19 ~ 11:28	0.864	0.859	0.873	0.862
Adapter Fusion	08:20 ~ 08:21	0.850	0.849	0.853	0.847
LoRA (r=64)	07:19 ~ 07:21	0.683	0.692	0.662	0.670

표 2 10 만건에 대한 학습 결과

	Training time per epoch (mm:ss)	Test accuracy	Precision	Recall	F1-Score
Vanilla	05:40 ~ 05:42	0.848	0.840	0.863	0.847
Adapter Fusion	04:10 ~ 04:12	0.830	0.836	0.824	0.826
LoRA (r=64)	03:40	0.634	0.620	0.706	0.653

표 3 5 만건에 대한 학습 결과

실험은 학습 문장 15 만건중 5 만건씩 줄여가며 진행했다. 15 만건의 문장에 대해 실험을 진행할 때, LoRA 의 r 값이 줄어들수록 가중치 행렬의 크기가 줄어들어 정확도가 낮아지는 것을 확인했다. Vanilla 기법은 Fully fine-tuning 기법을 의미하며, 다른 기법대비 학습하는 파라미터의 수가 많아 정확도는 높지만, 학습 시간은 상대적으로 많이 소요된다. 그리고, 학습 데이터의 양이 줄어들수록 AdapterFusion 과 LoRA 와 같은 Fine-tuning 기법들의 학습 시간 감소 비율이 줄어든 것을 확인할 수 있다.

Ⅶ. 결론

AdapterFusion 과 LoRA 를 적용했을 때 학습 속도는 줄어들지만, 정확도는 Fully fine-tuning 을 상회하지 못한다. 이는 절대적으로 학습해야하는 파라미터 수가 줄어들어 나타난다고 추측한다. 본 논문에서 언급한 Fine-tuning 기법들은 학습 데이터셋의 수가 많고, 모델의 사이즈가 큰 경우에 효과적이다. 최근 화두가 되고 있는 준지도 학습 기반 딥러닝 모델들의 규모가 커짐에 따라 Fine-tuning 기법들은 필수적인 요소가 될 것이다.

ACKNOWLEDGMENT

"본 연구는 2022 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었으며 (2022-0-00964), 또한 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2022-2018-0-01396)

참 고 문 헌

[1] Ashish Vaswani, “Attention Is All You Need”, 2017

[2] Jacob Devlin, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2018

[3] Tom B. Brown, “Language Models are Few-Shot Learners”, 2020

[4] Jonas pfeiffer, “AdapterFusion: Non-Destructive Task Composition for Transfer Learning”, 2021

[5] Edward Hu, “LoRA: Low-Rank Adaptation Of Large Language Models”, 2021