소수 클래스 데이터 증강을 통한 BERT 기반의 유형 분류 모델 성능 개선

김정우*, 장광호**, 이용태***, 박원주****
충남대학교 컴퓨터공학과*
치안정책연구소**
한국전자통신연구원***
bnm961126@gmail.com, pathfinder@police.go.kr, ytlee@etri.re.kr, wjpark@etri.re.kr

Bert-based Classification Model Improvement through Minority Class Data Augmentation

Jeong-Woo Kim*, Kwangho Jang**, Yong Tae Lee***, Won-joo Park***,
*Dept. of Computer Science Engineering, Chungnam National University
**Police Science Institute
***ETRI

요 약

자연어처리 분야에서 딥러닝 기반의 분류 모델은 획기적인 성능을 보여주고 있다. 특히 2018 년 발표된 구글의 BERT 는 다양한 태스크에서 높은 성능을 보여준다. 본 논문에서는 이러한 BERT 가 클래스 불균형이 심한 데이터에 대해 어느 정도 성능을 보여주는지 확인하고 이를 해결하는 방법으로 EDA 를 선택해 성능을 개선하고자 한다. BERT 에 알맞게 적용하기 위해 다양한 방법으로 EDA 를 구현했고 이에 대한 성능을 평가하였다.

1. 서론

최근 자연어처리 분야에서 2018 년 발표된 구글의 사전 학습 모델 BERT[1]를 이용한 연구가 지속하고 있다[2]. BERT 는 다양한 NLP 태스크에서 "state-of-the-art"를 기록하며 지속해서 인기를 끌고 있다. BERT 는 기존 언어 표현 모델과 다른 Transformer 모델을 기반으로 입력데이터를 양방향으로 학습하며 자연어에 대한 범용적인 수치 표현을 제공한다[3]. 이러한 BERT 를 기반으로 Fine-tuning 을 하면 쉽게 다양한 분야에 적용할 수 있으며 문장 분류, 질의·응답과 같은 태스크를 수행할 수 있다. 하지만 한국어 처리를 위해서는 다국어 지원 BERT 모델을 사용해야 하고 해당 모델은 최적의 성능을 보여주지는 않는다. 이러한 문제를 해결하기 위해 SKT Brain, ETRI 와 같은 기관에서 한국어 특화 BERT 를 공개했다. 이들은 구글의 BERT 모델과 같은 구조를 가지며, 한국어 텍스트를 토큰화하기 위해 토크나이저를 따로 학습해서 제공해준다.

그러나 모든 뉴럴 네트워크가 그렇듯 BERT 역시 학습시키는 데이터의 클래스가 비슷한 비율로 구성되어 있지 않으면 모델의 성능이 저하되게 된다. 이와 같은 문제를 데이터 불균형 문제라고 하며 소수 클래스에 속하는 데이터들이 다수 클래스에 속하는 데이터보다 잘못 분류되는 경우가 생길 수 있다[4]. 이를 해결하기 위해 더 많은 데이터를 수집하거나 성능지표를 변경하는 등의 다양한 기법들이 존재한다.

본 논문에서는 다양한 데이터 불균형 처리 기법을 BERT 모델 맞게 적용해 실험하고 결과를 비교하는 것을 목표로 한다. 2 장에서는 관련 연구로 언어 모델 BERT 그리고 데이터 불균형 처리의 보편적인 접근인 UnderSamplimg, OverSampling 의 적용 알고리즘에 대해 살펴보고 3, 4 장에서는 실험 환경과 결과를 마지막으로 5 장에서는 향후 연구에 대해 논의하고자 한다.

2. 관련 연구

2.1 BERT

BERT(Bidirectional Encoder Representations from Transformers)는 2018 년 구글에서 발표한 언어 모델로 자연어처리 분야에서 획기적인 성능을 보여주고 있다. BERT 는 트랜스포머(transformer)[5] 모델의 인코더 부분을 사용하여 학습되며, 학습된 언어 모델 위에 출력 레이어를 추가하고 이를 여러 태스크 에 적용한다. 한국어에 특화된 BERT 모델로는 대표적으로 ETRI 의 KorBert 가 있다. 약 3 만개의

VOCAB 으로 학습시켰으며, 한국어를 지원하는 구글의 다국어 언어모델 보다 좋은 성능을 보여준다.

2.2 UnderSampling

다수 클래스의 데이터를 줄이는 방법으로 손실로 인한 분류기 성능저하가 있을 수 있다. 군집의 중심을 기준으로 K 개의 표본을 제거하는 Cluster 기법, 무작위로 데이터를 제거하는 Random Under Sampling 등의 기법이 존재한다.

2.3 OverSampling

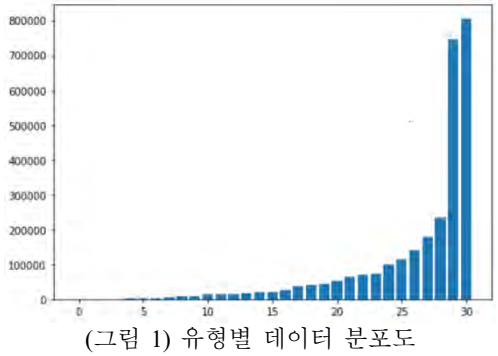
소수 클래스의 데이터를 늘리는 방법으로 일부 소수 클래스의 여러 복사본으로 훈련 데이터를 보완하는 작업의 Random OverSampling, 소수 데이터 샘플로 부터 가까운 이웃을 찾아 보간 하는 SMOTE (Synthetic Minority Over-sampling Technique)[6]와 같은 기법이 존재한다.

2.3.1 EDA

EDA(Easy Data Augmentation)[7]는 학습 데이터가 부족한 상황이나, 클래스 불균형 문제가 발생하였을 때, 현재 보유하고 있는 데이터를 변형시켜 데이터양을 늘리는 기법이다. 문장 데이터 에 적용할 수 있으며 특정 단어를 유의어로 교체하는 SR(Synonym Replacement), 임의의 단어를 삭제하는 RD(Random Deletion), 문장 내에서 단어를 선택하여 해당 단어의 유의어를 문장 내 임의의 자리에 삽입하는 RI(Random Insertion) 그리고 문장 내 임의의 두 단어의 위치를 바꾸는 RS(Random Swap) 까지 4 가지 방법이 존재한다. EDA 기법을 적용해서 학습 데이터를 다양한 형태로 늘리고 이를 통해 인공지능 모델에 대한 성능을 증가 시킬 수 있다.

3. 실험 환경 및 구성

본 연구는 특정 지역 지방경찰청의 과거 112 신고 내용을 학습하여 신규 신고 접수 시 유형을 예측하기 위한 연구이다. 실험에서 활용한 신고유형의 종류는 총 31 가지이다. ETRI 의 KorBert 를 사용해서 구현되었으며 전체 데이터의 80%인 약 200 백만여 개를 학습에 활용하였고, 20%인 50 만여개를 모델의 추론 정확도를 평가하기 위한 시험에 활용하였다.



(그림 1) 유형별 데이터 분포도

신고데이터의 유형별 분포도는 (그림 1)과 같이 불균형이 매우 심각한 분포를 갖고 있다. 위와 같이 클래스 불균형이 심한 데이터에 대해 여러 샘플링 기법을 적용하여 모델을 학습시킬 수 있다. 본 실험에서는 EDA 기법을 소수 클래스의 문장 데이터에 적용해 BERT 분류 모델에 학습시켜 성능 향상 결과를 확인한다.

EDA 기법을 적용하기 위해서는 변형시킬 단어를 선택하기 쉽도록 증강할 대상 문장을 토크나이징을 해야 한다. 여기서 토크나이징이란 문장을 토큰으로 쪼개어 벡터화하는 작업이다. 이를 위해 사용한 Tokenizer 는 치안 도메인 용어를 추가한 자체 Tokenizer 및 ETRI 엑소브레인 연구팀에서 공개한 KorBERT Tokenizer 를 사용한다. 각각의 Tokenizer 를 통해서 만들어진 문장 토큰에 EDA 기법을 적용한다. 선택하는 EDA 기법은 RD, RS 을 선택했다. RD 은 문장 내 임의의 토큰을 삭제하는 기법이고, RS 은 문장 내 두 토큰을 선택하여 위치를 바꾸는 기법이다. RD 에서 문장의 각 토큰에 대해 20%의 확률로 토큰을 삭제하도록 구현했다. 유의어 사전이 필요한 나머지 두 기법 SR, RI 는 실험에서 제외했다. 문장이 짧고, 품질이 좋지 않은 신고 데이터의 특성을 반영하여, 같은 문장을 계속 학습시켜 생기는 오히려 성능이 떨어지는 과적합 문제를 방지하기 위해 문장당 16 개 만큼 데이터를 증강하였다. 그리고, 단어만을 변형시키지 않고 불용어, 조사에 대해서도 변형하도록 구현하여 최대한 다양한 문장을 만들 수 있도록 하였다. RD, RS 를 통해서 학습 데이터를 증가시킬 수 있고, 이를 통해서 BERT 분류 모델의 성능 향상을 기대할 수 있다. 증강 대상은 (그림 1)의 유형별 분포 중 데이터 개수가 적은 최하위 4 개가 대상이며, 증강 이전의 측정된 유형별 예측 성능 결과는 (표 1)과 같다. 여기서 매크로 평균은 모델이 분류하고자 하는 전체 31 가지 신고 유형별 f1-score 에 대한 평균이다

(표 1) 하위 4 개 유형별 데이터 개수와 성능

유형	유형별 훈련 데이터 개수	F1-score
유형 0	46	0
유형 1	230	0.75
유형 2	317	0.45
유형 3	630	0.52
Macro Avg.	-	0.774
Accuracy	-	0.876

경찰청 신고 데이터 보안을 위해 데이터 개수가 적은 4 개의 유형에 대해서만 유형 이름은 삭제하고 성능 결과만 제시한다.

4. 실험 결과

(표 2)는 데이터 증강을 거치지 않은 데이터로 학습하여 생성한 Base 모델과 치안 도메인 용어를

기반으로 학습된 자체 Tokenizer 기반으로 토크나이징된 문장에 RS 를 적용한 모델 1, RD 를 적용한 모델 2 이다.

(표 3)은 Korbert Tokenizer 기반으로 토크나이징된 문장에 RS 를 적용한 모델 3, RD 를 모델 4 의 성능을 유형별 f1-score 로 비교한 결과이다. RD 을 구현할 때 토큰을 삭제할 확률은 20%로 모델 2, 모델 4 에 동일하게 설정했다.

(표 2) 자체 Tokenizer 기반의 EDA 적용 모델 f1-score

	Base	모델 1(RS)	모델 2(RD)
유형 0	0	0.167	0.267
유형 1	0.75	0.743	0.748
유형 2	0.45	0.434	0.481
유형 3	0.52	0.511	0.611

(표 2)의 성능향상 결과를 보았을 때 RS, RD 기법 모두 유형 0 에 대해서는 성능향상을 보였다. 유형 1 에서는 두 기법 모두 근소하게 하락하였고, 나머지 두 유형에 대해서는 RS 는 성능하락, RD 는 성능향상이 된 것을 확인할 수 있었다.

(표 3) KoBert Tokenizer 기반의 EDA 적용 모델 f1-score

	Base	모델 3(RS)	모델 4(RD)
유형 0	0	0.533	0.4
유형 1	0.75	0.812	0.729
유형 2	0.45	0.487	0.496
유형 3	0.52	0.487	0.522

(표 3)의 성능향상 결과를 보았을 때 RS, RD 기법 모두 유형 0, 유형 2 에 대해 성능향상을 보였다. 유형 1 의 경우 RS 를 적용했을 때만 성능향상을 보였고, 유형 3 의 경우 RD 를 적용했을 때 근소하게 성능향상을 보였다.

결과에서 확인할 수 있듯이 기존의 학습데이터보다 더 많은 데이터를 학습시켰음에도 불구하고 일부 유형의 성능이 하락하는 이유는 RS, RD 과 같은 기법들이 새로운 문장을 무제한으로 만들기에는 무리가 있고, 같은 문장을 계속해서 학습시켜 과적합 현상이 일어난 것으로 추측된다. (표 4)는 모델 별로 측정한 매크로 평균과 Accuracy 이다.

(표 4) 모델 별 측정 성능

	Normal	모델 1	모델 2	모델 3	모델 4
Macro Avg.	0.774	0.775	0.785	0.792	0.787
Accuracy	0.876	0.876	0.876	0.876	0.876

(표 2)와 (표 3)의 결과에서 확인되는 평균 f1-score 의 향상 수치와는 달리 (표 4)에서 확인되는 전체 31 종에 대한 Macro 평균의 성능은 큰 향상 폭을 보여주지 않는다. 이유는 소수 클래스 범주의 데이터가 증가함에 따라 해당 유형에 대한 가중치가 증가하여 일부 유형이 f1-score 가 다소 하락되는 현상이 발생하였다. 모델 1, 모델 2, 모델 4 는 약 0.01 의 성능 향상을 보여주었고, 이들 중 모델 3 이 0.792 으로 약 0.02 의 가장 크게 향상된 성능을 확인할 수 있었다. Accuracy 는 0.876 으로 모든 모델이 매우 유사한 성능을 보여주었다.

5. 결론

딥러닝 기반의 분류 모델은 클래스 불균형이 심한 데이터에 대해 소수 클래스에 속하는 데이터들이 다수 클래스에 속하는 데이터보다 잘못 분류될 가능성이 크다는 문제점이 있다.

본 논문에서는 Oversampling 중의 하나인 EDA 기술을 적용하여 클래스 불균형 문제를 해결하고자 하였다. 결과적으로 실험한 모든 모델에서 소수 클래스의 f1-score 가 소폭 향상된 것을 확인할 수 있었다. 가장 좋은 성능 향상을 보여준 모델은 KorBert Tokenizer 를 기반으로 RS 을 통해 문장을 증강한 모델이며, Macro 평균이 0.792 로 가장 높았다. 향후 연구에서 연구 도메인에 적용 가능한 유의어 사전을 구축을 할 수 있다면 새로운 단어를 만들 수 있고, 이를 통해서 RS, RI 등의 기법을 적용하여 또 다른 성능 향상 결과를 기대할 수 있을 것이다. 또한 본 연구에서 사용하지 않은 SMOTE, UnderSampling 과 같은 다른 Sampling 기법의 적용을 통해서도 성능 향상을 기대해볼 수 있다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2018-0-00440, 위험 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발)

이 논문은 한국전자통신연구원에서 공개한 한국어 언어모델(KorBERT)를 사용함(No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

참고문헌

[1] J. Devlin, M.W. Chang, K. Lee and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805, 2019.

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

[3] 황상흠, 김도현. 한국어 기술문서 분석을 위한 BERT 기반의 분류모델. 한국전자거래학회지, 25(1), 203-214, 2020.

[4] 김경민,장하영,박정완,황성택,장병탁. 불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법. 한국정보과학회 학술발표논문집, 686-688, (3 pages), 2013.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. Neural Information Processing Systems (NIPS), pp. 5998-6008, 2017.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. arXiv :1106,1813, 2011.
[7] Jason Wei, Kai Zou. EDA: Easy Data Augmentation Techniques. for Boosting. Performance on Text Classification Tasks arXiv:1901.11196, 2019.