문서 요약 기반의 텍스트 데이터 증강 기법에 관한 연구

최연석
동서울대학교

yeonseok9872@gmail.com

Text summarization-based Text data augmentation technique

Choi Yeon Seok
Dongseoul Univ.

요 약

본 논문은 문서 요약을 기반으로 한 텍스트 데이터 증강 기법에 관한 것이다. 머신러닝 모델의 성능 향상을 위해 데이터 증강이 필요한데, 이미지 분야와는 달리 문서 데이터의 경우 의미가 보존되어야 하며, 이를 위한 정량적 평가가 어려운 문제가 있다. 따라서 효과적인 문서 요약을 통해 데이터를 변형하는 방법을 제안한다. 연구에서는 EDA(쉬운 데이터 증강)와 Pre-trained Transformer 모델을 사용한 데이터 증강 기법을 설명하고 실험을 통해 성능을 평가한다. 실험 결과, 추상적 요약 기법을 사용할 때 Perplexity 가 낮고 ROUGE 점수가 높을 때 모델의 성능이 향상되는 것을 확인하였다. 이를 통해 문서 요약 기반의 데이터 증강이 모델 성능 향상에 기여할 수 있음을 보여준다.

Ⅰ. 서 론

근래의 머신러닝 모델들은 많은 파라미터를 갖고 있으므로 Overfitting 되지 않도록 학습시키려면 많은 데이터가 필요하다. 외부로부터 데이터를 추가로 확보, 생성, 분류하는 데에는 많은 자원이 소비된다. 따라서 데이터 확보 비용을 줄이고 모델의 성능을 향상하기 위해 Data Augmentation 을 사용해야 한다. 이미지 분야에서는 자주 쓰이는 기법이지만 문서 데이터의 경우 문법적 요소가 포함되어 있기 때문에 작은 변화로 의미가 바뀔 수 있고, 의도가 보존되었는지 비슷함(Similarity)에 대한 정량적 평가할 수 없다는 어려움이 있다. 따라서 문서의 의미를 보존하면서 표현을 변화하여야 한다. 본 연구에서는 효율적인 문서 요약(Text Summarization) 기반 데이터 증강 기법에 대해 제안하고자 한다.

Ⅱ. 관련 연구

2.1 EDA(Easy Data Augmentation)

외부 데이터를 사용하지 않고 문서 데이터를 증강하는 간단하고 효과적인 기법을 4 가지로 설명한다. [1]

1. Synonym Replacement (SR): 문장에서 불용어가 아닌 단어 중 무작위 n 개의

단어를 임의로 선정하여 유의어로 대체한다.

2. Random Insertion (RI): 문장에서 불용어가 아닌 단어를 찾고 그 단어가 있는 문장 내 무작위 위치에 유의어로 삽입하고 n 번 반복한다.

3. Random Swap (RS): 문장에서 두 단어를 무작위로 선택하고 위치를 바꾼다. 이 작업을 n 번 반복한다.

4. Random Deletion (RD): 문장의 각 단어를 p 의 확률로 무작위로 제거한다.

원본 데이터의 50%만 사용하며 EDA 기법을 적용한 모델의 성능과 원본 데이터의 100%를 사용한 모델 성능이 같은 정확도를 달성하였다.

2.2 Data Augmentation using Pre-trained Transformer Models

Pre-trained 모델 3 가지 (AR, AE, Seq2Seq)를 이용하여 텍스트를 증강하는 방법을 제시했다. [2] 텍스트 시퀀스에 Label 을 미리 추가하면 Fine-tuning 하는 Conditional Pre-training 방법에 대한 연구와 비교하였는데 1%의 데이터만 보유한 시나리오에서 Text Classification 에 그 유효성을 입증하였다.

Ⅲ. 실험

3.1 데이터셋

AI hub 의 "문서요약 텍스트" 데이터셋을 사용하였다. 문서요약 텍스트 데이터셋은 신문기사와 기고문, 잡지기사, 법원 판결문을 비롯한 원문 데이터와 원문 데이터에서 생성한 추출요약과 생성요약으로 Human-Labeled 된 요약문으로 구성되어 있다. 데이터셋에서 뉴스 데이터 "정치, 경제, 사회, 스포츠, IT/과학" 카테고리에서 무작위로 500 개의 데이터를 추출하여 훈련 데이터 80%와 검증 데이터 20%로 분리하였다.

3.2 데이터 증강

요약문을 생성하는 방식에 따라 extractive summarization(이하 ext)와 abstractive summarization(이하 abs)로 나눌 수 있다. [3] 본 연구에서는 ext 기법 Textrank 와 abs 기법 KoBART 와 T5 을 훈련데이터에 적용하여 증강 데이터를 생성하고 증강 데이터는 훈련 데이터에 포함되었다. 그 결과 생성된 증강데이터의 Perplexity 와 ROUGE 점수는 표 1, 표 2 과 같았다.

Table 1. Perplexity by Summarization method.

Summarization Method	Perplexity
Raw data	1.5349
Human Generated	1.3940
TextRank	1.3841
T5	1.3626
KoBART	1.4039

Table 2. ROUGE Score by Summarization method.

	ROUGE-1	ROUGE-2	ROUGE-L
Human Generated	0.1366	0.0678	0.1319
TextRank	0.5408	0.5118	0.5408
T5	0.1750	0.1228	0.1738
KoBART	0.1071	0.0717	0.1054

Table 3. F1 Score by Summarization method.

Data	F1-Score
Raw data	0.7919
Raw data + Human Generated	0.8950
Raw data + TextRank	0.9349
Raw data + T5	0.8787
Raw data + KoBART	0.8431

3.3 실험 결과

증강된 데이터가 성능향상에 영향을 주었는지 확인하기 위해 한국어로 Pre-trained BERT 모델로 문서 분류 모델 학습하여 평가하였다. 각 기법으로 증강된 훈련 데이터 별로 학습한 뒤 Accuracy 와 F1-Score 를 측정하고 Perplexity, ROUGE 점수와 비교하였다.

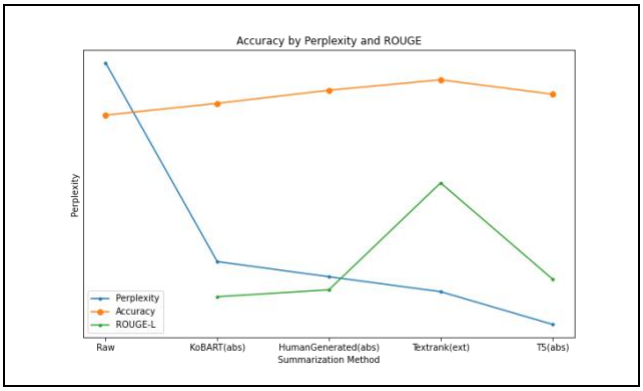

Fig 1. Accuracy by Perplexity and ROUGE.

그림 1 에서 문장들의 Perplexity 와 ROUGE 점수에 따른 상관관계를 파악한 결과, 추상적 요약 기법을 사용하였을 때는 Perplexity 가 낮고 ROUGE-L 점수가 높을 때, 즉 문장이 어색하지 않고 잘 요약된 정도가 성능에 영향을 주는 것을 확인할 수 있었다. 또한 추출적 요약 기법이 상대적으로 모델 성능에 많은 영향을 주는것을 확인할 수 있었다.

Ⅳ. 결론

본 연구에서는 머신러닝 모델 성능 향상을 위해 문서 요약 기반의 데이터 증강을 적용하였다. 적용한 요약 기법은 TextRank 와 T5, KoBART 이며 실험 결과 각 기법이 생성한 요약문의 품질을 파악할 수 있는 지표 Perplexity 와 ROUGE 점수가 모델 성능과 비례함을 확인하였다. 또한 추상적 요약기법보다 추출적 요약 기법을 적용했을 때 모델 성능이 향상되는 것을 확인하였다.

향후 원본 데이터와 증강 데이터의 비율에 따른 모델 성능에 상관관계를 밝히고자 한다. 또한 더 많은 요약 기법에 따른 모델 성능의 상관관계와 가진 데이터의 양이 얼마일 때부터 유의미한 성능 향상이 발생하거나 그렇지 않은 지점에 대한 연구로 확대하고자 한다.

참 고 문 헌

[1] Jason Wei, Kai Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, 2019, (https://doi.org/10.48550/arXiv.1901.11196).

[2] Varun Kumar, Ashutosh Choudhary, Eunah Cho, Data Augmentation using Pre-trained Transformer Models, 2020, (https://doi.org/10.48550/arXiv.2003.02245).

[3] Venkat N. Gudivada, "Handbook of Statistics", Handbook, Volume 38(2-215), 2018. (https://doi.org/10.1016/bs.host.2018.07.010)