**KoTextFooler: Is KoBERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification**

**Yunho Mo**

Natural Language Processing Laboratory

University of Gachon

[ahdbsgh2@naver.com](mailto:ahdbsgh2@naver.com)

**Abstract**

최근 다양한 분야에서 활용중인 딥러닝은 적대적 공격 가능성의 발견으로 위험성이 발견되고 있다. 딥러닝 알고리즘들은 보통 상대방이 감지할 수 없는 예제들을 가지고 있지만. 최첨단 모델들을 속일 수 있는 적대적 예제에 취약하다. 또한 악의적으로 조작된 예제들을 노출해서 이러한 모델들의 견고성을 향상시키며 평가를 할 수 있는 예제들은 한국어에서는 발견되지 않았다. 그래서 한국어에도 적용을 할 수 있는 KoTextFooler를 제안한다. 한국어 자연어 처리 중 분류 문제를 실험 해 보았으며, 이를 해결할 때 많이 사용하는 사전 훈련된 한국어 BERT로 실험을 진행하였다. 한국어 BERT는 hugging face에 등록된 Klue의 BERT-base 모델과 SKT의 kobert-base-v1를 사용하였으며, classification 실험은 NSMC data로 실험하였다. 실험 결과 한국어 BERT의 성능을 8퍼센트 하락시키는 결과가 도출 되었으며, 우리는 이 실험을 통해 KoTextFooler 의 장점을 보여준다. 1) 효과적 – 효과적으로 KoBERT를 공격하여 8%의 성능 하락을 보였다. 2) 효율적 – 간단한 실험으로 좋은 성능을 끌어낼 수 있다. 3) 보존성 – 기존 데이터의 형태를 보존하였다.

**Introduction**

분류 문제: 이 영화의 리뷰가 긍정인지 아닌지.



그림 1: KoTextFooler는 입력 텍스트를 약간 변경하지만 예측 결과를 완전히 바꿨습니다.

최근 다양한 분야에서 활용중인 딥러닝은 적대적 공격 가능성의 발견으로 위험성이 발견되고 있다. 딥러닝 알고리즘들은 보통 상대방이 감지할 수 없는 예제들을 가지고 있지만. 최첨단 모델들을 속일 수 있는 적대적 예제에 취약하다. 또한 악의적으로 조작된 예제들을 노출해서 이러한 모델들의 견고성을 향상시키며 평가를 할 수 있는 예제들은 한국어에서는 발견되지 않았다. 영어의 경우, TextFooler(1)와 같은 예제들이 존재한다. 그래서 우리는 한국어에도 적용을 할 수 있는 KoTextFooler를 제안한다. 한국어 자연어 처리 중 분류 문제를 실험 해 보았으며, 이를 해결할 때 많이 사용하는 사전훈련된 KoBERT로 실험을 진행하였다. 사전 훈련된 KoBERT는 저자의 깃허브에도 공개가 되어 있으며, hugging face(2)에 공개되어 있다. Hugging face는 자연어 처리를 위한 사전 학습된 모델들을 전세계 사람들과 공유를 하는 사이트이다. KoBERT는 hugging face에 등록된 Klue의 BERT-base, SKT의 kobert-base-v1 를 사용하였으며, Klue의 BERT-base는 NSMC로 pre-train된 모델이며 SKT의 kobert-base-v1는 다른 데이터로 pre-train이 되어 있다. 따라서 pre-train된 모델에도 공격이 가능한지, 또한 fine-tuning한 모델에는 어떠한 공격 결과가 있는지 실험을 하였다. classification 실험은 NSMC(네이버 영화 리뷰 말뭉치)데이터로 실험하였다. 추후에는 다른 한국어 분류 문제 데이터로 실험을 늘려볼 예정이다.

**Method**

먼저 문장을 형태소 분석을 통해 형태소 단위로 분절한다. 형태소 분석은 KoNLPy(4)의 Mecab을 사용한다. 그 후 카이스트 대학에서 만든 한국어 기반 WordNet을 사용한다. 한국어 기반의 WordNet은 어휘 의미망으로써, 어의 중의성 해소나 정보추출 등과 같은 다양한 자연어 처리에 폭넓게 사용된다. 여기서는 동의어를 추출하여 문장에 WordNet에 나온 단어가 있으면 교체해주는 방식으로 사용한다.

**Experiments**

실험에 사용한 데이터셋은 NSMC(Naver sentiment movie corpus) v1.0(5)이다. 데이터는 ID, document, label로 구성되어 있으며 학습데이터는 120K개의 리뷰데이터를, 검증데이터는 30K의 리뷰데이터를 사용하였다.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | train | test | Avg Len |
| NSMC | 150k | 50k | 140 |

표 1: 데이터 셋 분포

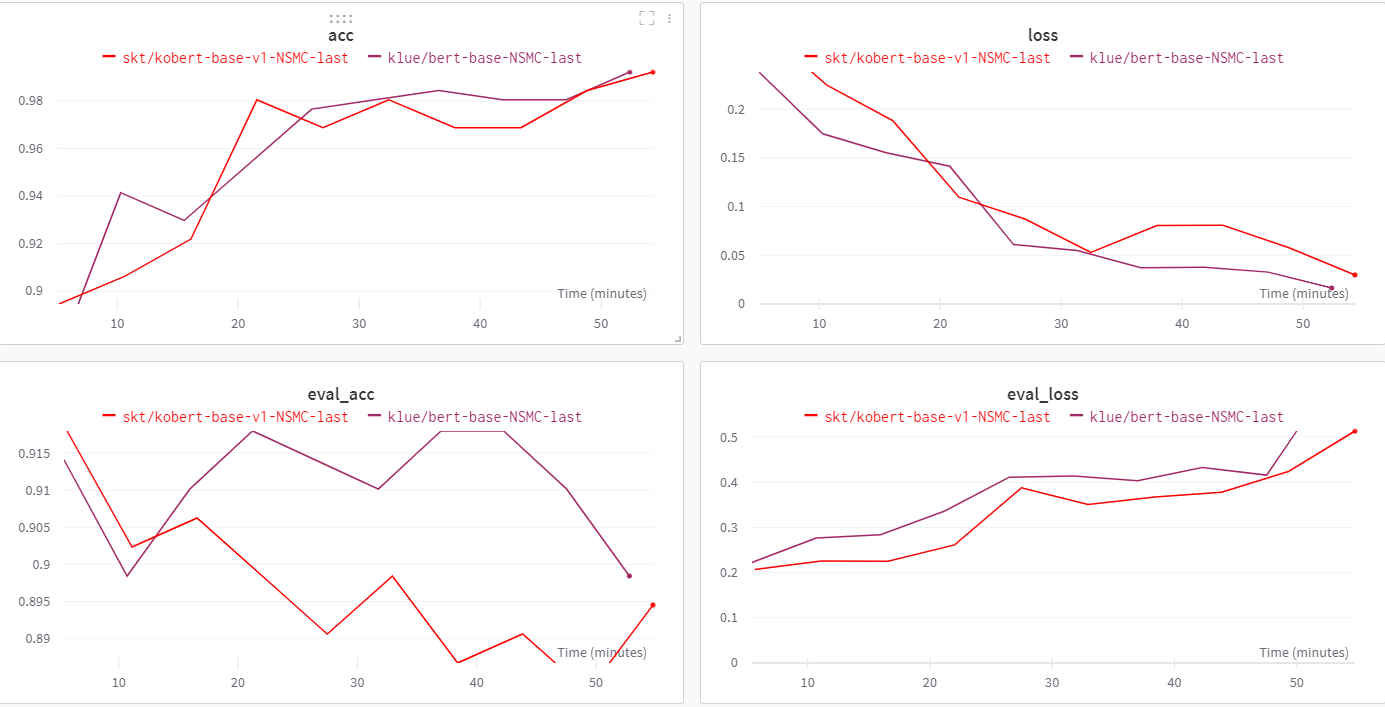


그림 1: klue/bert-base, skt/kobert-base-v1 NSMC fine-tuning 결과.

실험에 사용된 모델은 hugging face에서 불러온 Klue의 BERT-base, SKT의 kobert-base-v1를 사용하였다.

|  |  |  |
| --- | --- | --- |
| 모델명 | Fine tuning accuracy | After attack accuracy |
| klue/bert-base | 91.80 | 87.21 |
| skt/kobert-base-v1 | 90.63 | 82.73 |

표 2: 실험 결과

그림 1을 보면 klue/bert-base는 3epoch이 점수가 제일 높고 오버피팅이 발생하지 않기 때문에 으로 3epoch에서 실험하였고, skt/kobert-base-v1는 2epoch이 점수가 제일 높고 오버피팅이 발생하지 않기 때문에 2epoch에서 실험진행하였다. klue/bert-base는 NSMC 데이터로 pre-train하고 fine-tuning해서 좀 더 기존의 데이터에 강인한 모델이다. 그 결과 4.5%정도의 성능 하락이 있었다. skt/kobert-base-v1는 한국어 위키피디아 데이터로 pre-train하고 NSMC 데이터로 Fine-tuning한 모델이다. Klue/bert-base보다 좀 더 약한 모델이다. 그 결과 8%정도의 성능 하락이 있었다. 공격이 성공적이라는 것을 보여준다.

|  |  |  |
| --- | --- | --- |
| 모델명 | Fine tuning accuracy | After attack accuracy |
| klue/bert-base | 91.80 | 87.21 |
| skt/kobert-base-v1 | 90.63 | 82.73 |
| BERT(IMDB) | 90.9 | 13.6 |

표 3: 영어 실험과 한국어 실험의 비교.

영어 실험과 한국어 실험의 차이점이 큰 이유는 영어의 WordNet(6)은 150,000개의 115,000개의 동의어집합(synset)과 총 207,000개의 단어-의미 쌍이 조직되어 있지만. 한국어의 경우 8,270개 단어의 9,714개의 동의어집합과 총 20,415개의 단어-의미 쌍이 조직되어 있다. 10배 이상의 차이가 난다. 영어와 한국어의 차이점이라고 할 수 있다.

**Conclusion**

이 실험을 통해 한국어 WordNet의 한계점을 알 수 있었고 한국어에 대한 어려움을 느낄 수 있었다. 영어는 WordNet 기반으로 동의어를 교체하였을 때. 더 좋은 성능이 나왔지만 이는 영어의 WordNet의 크기가 한국어보다 10배 이상 크기 때문이라고 생각한다. 한국어도 WordNet을 늘릴 수 있다면 보다 좋은 공격을 할 수 있을 것이다. 그리고 영어 실험에서 진행한 중요한 단어 위주로 변경하는 것은 한국어 WordNet이 커진다면 진행을 할 수 있다. 영어와 한국어의 어휘 차이로 인해서 영어와 한국어의 성능차이가 크지만 점차 발전하는 한국어 자연어 처리 기술은 추후에 더 한국어 어휘에 맞는 공격 기법을 만들 것이다.

Referance

(1) Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment, Di Jin,et al, AAAI 2020, Computation and Language (cs.CL); Artificial Intelligence (cs.AI); Machine Learning (cs.LG), arXiv:1907.11932

(2) hugging face, https://huggingface.co/

(3) Korean WordNet, http://wordnet.kaist.ac.kr/

(4) KoNLPy, https://konlpy.org/ko/latest/#

(5) NSMC v1.0 (Naver sentiment movie corpus v1.0), <https://github.com/e9t/nsmc>

(6) English WordNet, <https://wordnet.princeton.edu/>