국방 분야에서 텍스트 기반 딥러닝 모델에 대한 적대적 샘플에 관한 고찰

김동욱, 권 현*
육군사관학교 사이버전 전공, *육군사관학교 전자공학과
dukim0112@gmail.com, *hkwon.cs@gmail.com

A Study on Adversarial Examples for Text-Based Deep Learning Model in the Military

Dong Uk Kim, Hyun Kwon*
Korea Military Academy, *Dept. of Electrical Engineering, Korea Military Academy.

요 약

최근 딥러닝 기술을 이용하여 이미지 인식, 음성 인식, 텍스트 인식, 패턴 인식 등에 대해서 좋은 성능을 제공하고 있다. 그 중 텍스트 기반 모델 BERT 모델을 이용한 문건을 분류 또는 생성하는 분야에 적용 될 수가 있다. 이러한 텍스트 기반 딥러닝 기술이 발달함에 따라 국방 분야에 적용하기 위한 연구도 활발하게 진행되고 있다. 국방 분야에 여러 가지 문건을 분류하거나 생성하는 분야에 적용하여 전투력 증진에 기여할 수가 있다. 하지만 이런 텍스트 기반 딥러닝 모델은 적대적 샘플 공격에 대하여 취약점이 있으며 이를 통해 발생할 수 있는 위험성이 있다. 따라서 본 논문에서는 텍스트 기반 딥러닝 모델에 적대적 샘플 공격에 대한 실험적인 분석과 국방 분야에 적용 시 고려사항에 대하여 기술하였다.

Ⅰ. 서 론

4차 산업혁명이 도래되면서 국방부에서도 최첨단 과학기술을 적용한 과학기술군을 육성하기 위하여 노력하고 있다. 왜냐하면 매년 감축되는 군 병력과 사회에서의 최신 과학기술에 관심이 높아졌기 때문이다. 그러한 분위기 속 핵심기술은 인공지능 기술이라고 볼 수 있다. 국방부에서 인공지능 기술을 이용하여 AI 협업센터 등 다양한 조직을 창설하여 인공지능 기술을 접목한 과학기술을 연구하고 개발하고 있다. 이러한 인공지능 기술은 딥러닝 기술[1]을 구현되어지고 있고, 이미지 인식, 음성 인식, 텍스트 인식, 패턴 인식 분야에 좋은 성능을 제공하고 있다. 하지만 이러한 딥러닝 기술은 적대적 샘플 공격[2]에 대한 취약점이 존재하며 이는 국방 분야에서의 AI 기술에 대한 위험성이 클 수가 있다는 것을 의미한다. 하지만 이러한 적대적 샘플 공격에 관한 연구는 이미지 분야에 주로 많이 연구되었고 텍스트 기반에 연구는 많이 이뤄지지 않고 있는 실정이다.

텍스트 기반 딥러닝 기술에 대한 적대적 샘플[3]은 사람이 보기에 정상적인 문장이지만 모델에 의해서 잘못 오인식 되는 샘플을 의미한다. 이미지 분야에서 적대적 샘플과 달리 문장 중에 중요한 단어를 대체 단어로 바꿈으로써 의미상, 문법상 문제는 없으나 모델에 의해서는 다른 의미로 분류 되는 샘플이다. 이미지 분야에 비해 이산적인 접근 방식으로 적대적 샘플을 생성해야하며 입력값에 대한 모델의 신뢰값(confidence value)을 알아야 생성이 가능한 점이 있다.

본 논문에서는 텍스트 기반 딥러닝 모델에 대해서 적대적 샘플 공격에 대한 생성 방법과 실험결과 및 이에 대한 국방 분야에 대한 고찰을 다루고자 한다. 타겟 모델로는 최신 텍스트 기반 딥러닝 모델인 BERT 모델[4]을 대상으로 하였고 데이터셋은 movie review (MR)[5]를 이용하여 실험적으로 검증을 하였다.

Ⅱ. 텍스트 기반 딥러닝 모델에서의 적대적 샘플 공격

텍스트 기반 적대적 샘플은 BERT 모델과 같은 NLP(Natural Language Processing)[6]을 공격하여 사람에 보기에 문법상 의미상 변화가 없지만 모델에 의해서 기존 문장과 다르게 오인식을 일으키는 방법이다.

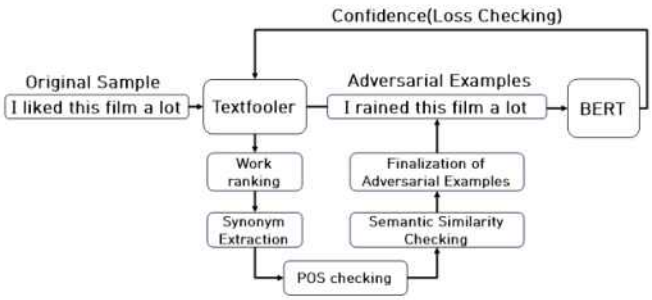

그림 1. 텍스트 도메인에서의 적대적 샘플 생성 과정 (예시)

이 방법[7]은 총 5가지 단계를 통해서 공격을 실시한다. 먼저 word ranking을 이용하여 문장에서 단어들의 중요도를 선정하여 변경할 단어를 결정한다. 그 이후에 synonym extraction을 이용하여 중요도가 높은 단어와 유사한 단어 후보군을 여러 개를 선정한다. 그 이후에 Part-of-Speech checking으로 품사를 확인하여 문맥적 의미, 문법적 오류 여부를 확인한다. Semantic similarity checking을 하여 변경한 단어를 대입했을 때, 문맥상 의미가 기존 문장과 유사한지 여부를 판단한다. 가장 오인식을 잘 일으키는 샘플을 적대적 샘플로 생성하는 방법이다. 이와 같은 방법에 대하여 movie review 데이터셋을 적용하여 적대적 샘플을 생성하였다. movie review 파일은 총 9000개의 훈련데이터와 1000개의 테스트 데이터로 긍정문과 부정문으로 모델에 의해서 분류되는 데이터셋이다. 타겟 모델은 BERT 모델으로 설정하였고 활성함수는 gelu 함수[8]를 사용하였고 파라미터는 similarity score threshold는 0.7과 synonyms은 50으로 설정하였다. 적대적 샘플의 공격을 실시해보면 아래 그림 2와 같은 결과 예시가 나온다.

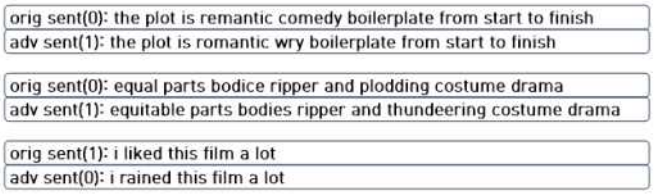

그림 2. 적대적 샘플을 공격한 예시. orig sent는 원본 문장을 의미하고 adv sent는 적대적 문장을 의미한다. (0: 부정문을 의미하고 1: 긍정문을 의미한다.)

그림에서 보면, 사람이 보기에는 동일한 두 개 문장이지만 모델에 의해서 각각 잘못 오인식 되는 것을 볼 수가 있다. 테스트 데이터 904개 중 107개의 분류 결과를 변경할 수 있었고, 정확도는 90.4%에서 79.7%로 감소시키는 공격 효과를 볼 수가 있었다. 이와 같이 텍스트 기반 딥러닝 모델에서 특정 중요단어를 바꿈으로써 모델에 의해서 오인식 되는 샘플을 생성할 수가 있다. 3장에서는 이러한 텍스트 기반 딥러닝 모델에서의 적대적 공격에 대한 국방분야에 고려사항을 설명되어 있다.

Ⅲ. 국방 분야에서의 적대적 샘플을 적용한 고찰

국방 분야에서 적대적 샘플에 대한 공격을 일으킬 경우 문제를 야기할 수가 있다. 예를 들어, 특정 인재를 선발하는 AI 면접에서 프로그램 상 특정 단어를 조작하여 평가하는 결과를 바꿀 수가 있다. 유능한 사람을 떨어지게 하고 무능한 사람을 합격하도록 할 수가 있다. 또한 육군 진급 체계에서 각 장교평가가 정리된 문서들을 분석하는 AI 기술에 악의적인 공격자가 잘못 오인식하게 하여 분류작업을 조작할 수가 있다. 따라서 국방 분야에 AI 면접과 문서 분류에 대한 기술을 접목하고 있는 상황에서 텍스트 기반 딥러닝 모델에서의 적대적 샘플 방어 방법에 대한 연구가 필요하다. AI를 이용한 방어체계와 공격체계에도 치명적인 문제를 야기할 수 있다. AI를 방어체계와 공격체계에 활용하기 위해서는 아군과 적군을 구별하는 방식을 학습시키는 과정이 필요하다. 이 과정에 적대적 샘플로 인해 아군과 적군을 식별하는 방식을 조작하여 방어체계를 무력화 시키고 오히려 적군의 또 하나의 공격수단으로 활용될 수 있다. 또한 지능미사일의 표적 식별 및 추적 시스템에서 표적에 대한 정보를 조작하여 잘못된 표적에 공격했음에도 공격을 성공했다는 잘못된 정보를 전달하도록 만들 수 있다. 즉 딥러닝 모델에 대한 연구뿐만이 아니라 기존에 입력된 정보에 대한 적대적 샘플 방어 방법에 대한 연구도 필요하다.

Ⅳ. 결론

본 논문에서 국방 분야에서 텍스트 기반 딥러닝 모델에 대한 적대적 샘플 관련 내용을 다루었다. 최신 BERT 모델에 대하여 적대적 샘플 공격을 실험적으로 분석을 하였고 국방 분야에서 고려해야하는 사항을 기술하였다. 향후 연구에는 BERT 모델이 영어기반으로 되어 있지만 한국어 기반 BERT 모델을 대상으로 보다 실제적인 모델을 대상에 공격을 할 예정이다.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2021R1I1A1A01040308)

참 고 문 헌

[1] Liu, Weibo, et al. "A survey of deep neural network architectures and their applications." Neurocomputing 234 (2017): 11-26.

[2] Kwon, Hyun, Hyunsoo Yoon, and Ki-Woong Park. "POSTER: Detecting audio adversarial example through audio modification." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.

[3] Kwon, Hyun. "Friend-Guard Textfooler Attack on Text Classification System." IEEE Access (2021).

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[5] Batanović, Vuk, Boško Nikolić, and Milan Milosavljević. "Reliable baselines for sentiment analysis in resource-limited languages: The serbian movie review dataset." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016.

[6] Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." Journal of the American Medical Informatics Association 18.5 (2011): 544-551.

[7] Jin, Di, et al. "Is bert really robust? a strong baseline for natural language attack on text classification and entailment." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 05. 2020.

[8] Hendrycks, Dan, and Kevin Gimpel. "Gaussian error linear units (gelus)." arXiv preprint arXiv:1606.08415 (2016).