텍스트 분야 적대적 공격 기법 연구 동향

김보금[1,†], 강효은[1], 김용수[1,2], 김호원[1,2,‡]
[1]부산대학교 정보컴퓨터공학부
[2]SmartM2M
{bogold[1,†], hyoeun0915[1], howonkim[1,2,‡]}@pusan.ac.kr, yongsu[1,2]@smartm2m.co.kr

Research Trends of Adversarial Attack Techniques in Text

Bo-Geum Kim[1,†], Hyo-Eun Kang[1], Yongsu-Kim[1,2], Ho-Won Kim[1,2,‡]
[1]School of Computer Science and Engineering, Pusan National University
[2]SmartM2M

요 약

인공지능 기술이 문서 분류, 얼굴 인식, 자율 주행 등 실생활 전반에 걸쳐 다양한 분야에 적용됨에 따라, 인공지능 모델에 대한 취약점을 미리 파악하고 대비하는 기술의 중요성이 높아지고 있다. 이미지 영역에서는 입력 데이터에 작은 섭동을 추가해 신경망을 속이는 방법인 적대적 공격 연구가 활발하게 이루어졌지만, 텍스트 영역에서는 텍스트 데이터의 이산적인 특징으로 인해 연구에 어려움이 존재한다. 본 논문은 텍스트 분야 인공지능 기술에 대한 적대적 공격 기법을 분석하고 연구의 필요성을 살펴보고자 한다.

1. 서론

인공지능은 실생활 속 다양한 분야에 적용되면서 빠른 기술 발전에 이바지하고 있다. 하지만 Goodfellow 등[1]이 연구를 통해 인공지능 모델이 가지는 취약성을 보였다. 이에 따라 인공지능 모델에 대한 공격 기법 파악 및 대비의 중요성이 높아지고 있다.

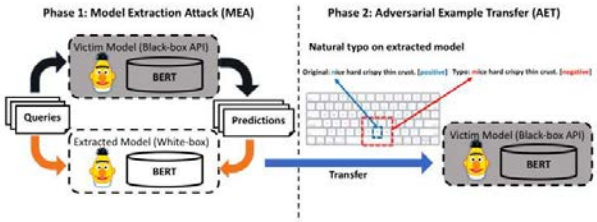
작은 섭동으로 신경망을 속이는 방법에 관한 연구는 이미지 분야에서 활발하게 진행되었다. Di 등[2]은 데이터의 이산적인 특성으로 인해 텍스트 영역에서의 적대적 공격이 어렵다고 언급하였다.

본 논문에서는 텍스트 분야에 대한 적대적 공격 기법의 연구 동향을 텍스트 단위별로 나누어 알아본다.

2. 본론

2.1 Character-level Attack

Character-level attack은 단어 내 문자열을 수정하여 적대적 예제를 생성하고 신경망을 속이는 기법[3]이다. Xuanli 등[4]은 BERT[5] 기반 서비스를 제공하는 API에 대해 적대적 공격을 수행하는 방법을 제시하였다.



(그림 1) BERT 기반 API를 이용한 공격 과정

모델 추출 공격(Model extraction attack) 단계에서 공격자가 m개의 쿼리로 이루어진 쿼리 집합을 BERT API에 전송하면 공격 대상 모델(Victim model)은 예측 결과를 출력한다. 공격자는 각 쿼리 x_i에 대해 공격 대상 모델이 출력한 사후 확률(Posterior Probability) 벡터 y_i로 이루어진 $\{x_i, y_i\}_{i=1}^{m}$을 이용해 공격 대상 모델을 재구성한다.

적대적 예제 전이(Adversarial example transfer) 단계는 모델 추출 공격 단계에서 얻은 모방 모델로부터 적대적 예제를 생성하고, 적대적 예제가 공격 대상 모델을 속일 수 있는지 확인하는 단계이다. 적대적 예제에 삽입, 삭제, 교체, 단순 오타 입력, 유사한 발음으로 인한 오타 입력 등의 방법[6]을 적용해

적대적 오타(Typo) 예제를 생성하였다. BERT-base 모델과 AG News 데이터셋을 사용하여 모델의 뉴스 토픽 예측 태스크를 수행했을 때, 적대적 예제 전이성 실험에서 47.5%의 공격 성공률을 보였다. 동일 데이터로 BERT-large 모델에서 태스크 수행 시 공격 성공률 59.3%를 달성하였다. 이를 통해 단순 오타처럼 보이는 문자열만으로도 감정 분석, 리뷰 분석 등에 사용되는 인공지능 모델을 공격할 수 있음을 보여준다.

2.2 Word-level Attack

Lee 등[7]은 베이지안 최적화에 기반을 둔 Blockwise Bayesian Attack(BBA) 프레임워크를 제안하였다. 이산적인 시퀀스 데이터에 대한 적대적 공격은 Black-box 함수를 최대화하기 위한 베이지안 최적화 단계와 원본 데이터와 적대적 예제 사이 해밍 거리(Hamming distance)를 최소화하는 단계로 나눌 수 있다. 베이지안 최적화 단계에 대한 수식 표현은 (1)과 같다.

$$\underset{s' \in S}{maximize} L(f_\theta(s'), y). \quad (1)$$

s'는 Attack space $S \triangleq \prod_{i=0}^{l-1} C(w_i)$에 속하는 텍스트를 의미한다. l은 입력 시퀀스 s를 이루고 있는 단어의 개수이며 $C(w_i)$는 원본 시퀀스의 i번째 단어인 w_i와 의미상으로 유사한 표현의 집합이다.

Boxin 등[8]이 제시한 SemAttack은 원본 입력과 의미상으로 유사한 적대적 예제를 생성하는 프레임워크이다. Targeted Attack, Untargeted Attack을 수행할 때 SemAttack의 목적 함수 $g(\cdot)$는 각각 (2), (3)으로 표현된다.

$$g(x') = \max[\max\{f(x')_i : i \neq t\} - f(x')_t, -K] \quad (2)$$

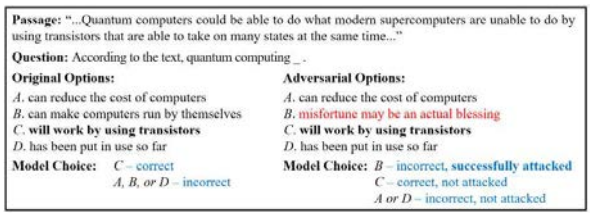
$$g(x') = \max[f(x')_t - \max\{f(x')_i : i \neq t\}, -K] \quad (3)$$

(2)는 타겟 공격 시나리오에서의 목적 함수를 의미한다. $f(x')_i$는 적대적 공격 토큰 x'에 대한 i번째 클래스의 logit을 의미하며, t는 공격자가 공격 대상 모델로부터 유도하고자 하는 잘못된 클래스를 의미한다. K는 모델이 공격자가 의도한 잘못된 클래스를 출력할 때 가지는 confidence score와 연관된 상수로, 값이 클수록 모델은 높은 confidence score로 잘못된 클래스를 출력하게 된다. (3)은 무작위 공격 시나리오에서 목적 함수를 의미한다.

Yelp 데이터셋과 오타 기반 섭동 함수 F_T, 지식 기반 섭동 함수 F_K, 문맥화된 의미론 섭동 함수 F_C를 모두 적용한 SemAttack을 BERT 모델에 적용했을 때 타겟 공격 성공률은 93.8%, 무작위 공격 성공률은 97.6%를 보였다.

2.3 Sentence-level Attack

Jieyu 등[9]은 영어 지문과 지문을 통해 해결할 수 있는 객관식 질문으로 이루어진 RACE 데이터셋[10]과 Magnet option으로 사전 학습 언어 모델을 공격하는 방법을 제시했다. 정답과 관련이 없으나 모델이 정답으로 많이 선택하는 선지인 Magnet option과 오답 선지를 바꾸는 적대적 공격을 수행하였다.


Passage: "...Quantum computers could be able to do what modern supercomputers are unable to do by using transistors that are able to take on many states at the same time..."
Question: According to the text, quantum computing _ .

Original Options:	**Adversarial Options:**
A. can reduce the cost of computers	*A*. can reduce the cost of computers
B. can make computers run by themselves	*B*. misfortune may be an actual blessing
C. **will work by using transistors**	*C*. **will work by using transistors**
D. has been put in use so far	*D*. has been put in use so far
Model Choice: *C* – correct	**Model Choice:** *B* – incorrect, **successfully attacked**
A, B, or D – incorrect	*C* – correct, not attacked
	A or D – incorrect, not attacked

(그림 2) Sentence-level Attack 예시

"A, B and C"라는 Magnet option과 RACE 테스트 데이터셋을 사용해 BERT, ALBERT[11], RoBERTa[12] 모델에 공격을 적용했을 때 ALBERT-base 모델의 정확도는 21.7%, RoBERTa-base 모델의 정확도는 16.6%, BERT-base 모델의 정확도는 9.4%로 높은 공격 성공률을 보였다.

3. 결론

본 논문은 텍스트 대상 적대적 공격 기법을 문자 단위 공격, 단어 단위 공격, 문장 단위 공격으로 세분화하여 조사하였다. 인공지능 기술이 자연어처리 분야에 많이 쓰이고 있는 만큼, 텍스트에 대한 적대적 공격 방법 연구는 필수적이다. 이미지 영역에서 주로 사용되었던 적대적 예제 생성 방법이 텍스트 영역에도 적용 가능하다는 점을 인지하고 있을 때, 더욱 강건한 모델이 만들어질 수 있을 것이다.

사사(Acknowledgement)

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(IITP-2022-0-01201)

참고문헌

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572, 2014.

[2] 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.

[3] Wang, Wenqi, et al. "Towards a robust deep neural network in texts: A survey." arXiv preprint arXiv:1902.07285, 2019.

[4] He, Xuanli, et al. "Model extraction and adversarial transferability, your bert is vulnerable!." arXiv preprint arXiv:2103.10013, 2021.

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

[6] Sun, Lichao, et al. "Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT." arXiv preprint arXiv:2003.04985, 2020.

[7] Lee, Deokjae, et al. "Query-Efficient and Scalable Black-Box Adversarial Attacks on Discrete Sequential Data via Bayesian Optimization." PMLR, Baltimore, Maryland, USA, 2022.

[8] Wang, Boxin, et al. "SemAttack: Natural Textual Attacks via Different Semantic Spaces." arXiv preprint arXiv:2205.01287, 2022.

[9] Lin, Jieyu, Jiajie Zou, and Nai Ding. "Using adversarial attacks to reveal the statistical bias in machine reading comprehension models." arXiv preprint arXiv:2105.11136, 2021.

[10] Lai, Guokun, et al. "Race: Large-scale reading comprehension dataset from examinations." arXiv preprint arXiv:1704.04683, 2017.

[11] Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942, 2019.

[12] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692, 2019.