JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 12, pp. 2019-2026, Dec. 2021

Check for updates

텍스트 트랜스포머 모델에서 어텐션 맵을 이용한 경사도 기반 화이트 박스 적대적 예제 생성 방안

신 초 별[1] · 문 종 섭[2*]
[1]고려대학교 정보보호대학원 석사과정
[2*]고려대학교 전자및정보공학과 교수

A Gradient Based Adversarial Example Method Using Attention Map Against Text Transformer Model

Cho-Byeol Shin[1] ·Jong-Sub Moon[2*]

[1]Master's Course, Department of Cybersecurity, Korea University School of Cybersecurity
[2*]Professor, Department of Electronics and Information Engineering, Korea University

[요 약]

트랜스포머 모델의 텍스트 데이터에 대한 적대적 예제 생성 방법은 텍스트 데이터의 이산적인 특징 때문에 블랙박스 공격 방법이 대부분이었다. 최근 트랜스포머 모델의 텍스트 데이터를 대상으로 한 경사도 기반 화이트박스 공격 방법이 발표되었는데 이는 하나의 예제 생성 마다 하나의 분포를 학습시키기 때문에 시간이 오래 걸려 효율적이지 못하다는 단점이 있다. 본 논문은 트랜스포머 모델의 어텐션 구조를 이용한 어텐션 제약조건을 제안하여 기존 화이트 박스 공격방법의 효율성을 높인다. 실험을 통해 기존의 연구결과보다 생성 시간을 6.5% 가량 단축시킬 수 있으며 생성되는 적대적 예제의 생성률을 2.4% 높일 수 있음을 입증하였다.

[Abstract]

Abstract should be placed here. These instructions give you guidelines for preparing papers for JDCS.The method of generating Adversarial examples for text data of the transformer model was mostly a black box attack method because of the discrete characteristics of text data. Recently, a gradient-based white box attack method targeting text data of a transformer model has been announced, which has the disadvantage that it takes a long time and is not efficient because it learns one distribution for each generation of an example. This paper improves the efficiency of the existing white box attack method by proposing an attention constraint using the attention structure of the transformer model. Through experiments, it has been proven that the generation time can be shortened by about 6.5% and the diversity of generated adversarial examples can be increased by 2.4% compared to the previous research results.

색인어 : 트랜스포머, 적대적 예제, 딥러닝, 텍스트 데이터, 화이트박스

Keyword : Transformer, Adversarial example, Deep learning, Text data, Whitebox

http://dx.doi.org/10.9728/dcs.2021.22.12.2019

Received 08 November 2021; **Revised** 25 November 2021
Accepted 25 November 2021

***Corresponding Author; Jong-Sub Moon**

Tel: 044-860-1423
E-mail: jsmoon@korea.ac.kr

Ⅰ. 서 론

심층신경망 (deep neural network, DNN)은 입력값에 예민하여 입력의 작은 변화가 결과값에 결정적인 영향을 끼치기도 한다. 적대적 예제(Adversarial examples)는 기계 학습 모델이 잘못된 예측을 하게 하는 작고 의도적인 섭동을 주어 생성된 입력값이다. 텍스트 데이터에 대한 적대적 예제 생성은 이미지 데이터에 대한 적대적 예제 생성시 보다 더 까다롭다. 적대적 예제는 원본 입력에 작은 변화를 주어 만든다. 즉 기존 예제와의 차이가 인지할 수 없을 정도로 작아야 한다.

이미지의 경우 이러한 차이를 측정하는 방법은 비교적 간단하지만 텍스트의 경우 그렇지 않다. 텍스트의 특성 상 기존 입력에서 변경된 토큰의 수가 적어야 할 뿐 아니라 변경된 토큰들의 의미도 유사해야 하기 때문이다. 또한 이미지에서 적대적 예제에 대한 연구는 대부분 예측 오류를 장려하는 적대적 손실을 정의한 후 손실을 최적화하여 적대적 예제를 생성하는 화이트 박스 공격 방법을 이용하였다. 하지만 자연어의 이산적 특성은 미분이 불가능하게 하여 이러한 최적화 기반 방법을 적용하기 어렵게 한다. 따라서 기존 공격 방법 대부분은 휴리스틱한 단어 대체 및 블랙박스 쿼리를 이용하여 적대적 예제를 생성하였다[2][3][4].

최근 경사도기반 화이트박스 방법을 통한 텍스트 트랜스포머 모델에 대한 적대적 예제 생성 방법 연구가 발표되었다.[1] 하지만 이는 하나의 예제 생성마다 하나의 분포를 학습시켜야 하기 때문에 충분한 공격 성능과 유사도를 가진 예제를 학습시키는 데에 많은 오버헤드가 발생하여 효율이 낮다. 본 논문은 트랜스포머 모델의 특성인 어텐션 맵(attention map)을 활용하여 기존 화이트박스 적대적 예제 생성 방법의 성능을 높일 수 있는 새로운 방안을 제시한다. 트랜스포머 모델이 학습하는 입력 토큰들 사이의 관계를 적대적 예제 생성시 반영하여 모델에 영향을 많이 미치는 토큰들 위주로 변환되게 하여 최적화 속도를 향상시켰다. 또한 기존의 방법으로는 생성하지 못했던 입력에 대한 적대적 예제를 생성함으로써 공격 성능을 높였다. 위와 같은 성능의 향상을 AG News 데이터셋[12]과 IMDB 데이터셋[13]에 대한 실험을 통해 입증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 트랜스포머 모델을 대상으로 텍스트 데이터에 대한 대표적인 적대적 예제 생성 방법을 소개한다. 3장에서는 본 논문이 기존 화이트박스 방법에 대해 어텐션 제약조건을 추가함으로써 공격 성능을 높일 수 있음을 보인다. 4장에서는 3장에서 제안된 방법을 통한 실험 결과를 보이며 마지막으로 5장에서는 결론과 앞으로의 연구 방향을 제시한다.

Ⅱ. 관련 연구

2-1 텍스트 모델에 대한 블랙박스 및 화이트박스 공격

텍스트 모델에 대한 적대적 예제 생성 방법은 크게 블랙박스 방법과 화이트박스 방법으로 나뉜다. 블랙박스 방법은 모델의 내부 파라미터를 알지 못한 채 입력과 그에 대해 출력되는 결과를 관찰하여 공격을 수행한다. 따라서 이러한 블랙박스 공격을 수행하기 위해 공격자는 공격 대상 모델과 유사한 결정 경계를 갖는 대체 모델을 학습시킨 후 대체 모델의 내부 정보를 이용해 적대적 데이터를 생성하고, 이를 대상 모델을 공격하는데 사용한다. 화이트박스 방법은 공격 타겟 모델에 대한 모든 정보를 알고 공격하는 방법이다.

1) 블랙박스 공격

텍스트 모델에 대한 블랙박스 적대적 예제 생성 공격으로는 TextFooler[2], BAE[3] , BERT Attack[4] 이 대표적이다.

TextFooler[2]는 모델에 쿼리를 날려 결과값에 영향을 크게 미치는 토큰을 찾는데, 원래의 입력에서 해당 토큰을 제외한 입력 결과값, 그리고 원래 입력 결과값의 차이로 각 토큰의 중요도 점수를 측정한다. 결과값의 차이가 클수록 중요도 점수를 높게 반영하여 순위를 측정한다. 중요도 순위 순으로 토큰을 비슷한 의미의 다른 토큰으로 교체하여 적대적 예제를 생성한다. 오분류가 일어날 때 까지 이러한 토큰 교체 과정을 반복한다.

BAE[3]는 기존 TextFooler[2] 에 기반을 두어 토큰의 중요도를 측정한 후 해당 토큰을 교체한다. 다만 각 토큰의 중요도를 측정할 때 토큰을 삭제한 입력과 원래 입력을 비교하는 것이 아닌 해당 토큰을 mask 토큰으로 교체하여 측정한다. TextFlooler는 토큰 교체만 가능하지만 BAE는 토큰 추가까지 가능하다. 따라서 입력과 다른 길이를 가지는 적대적 예제를 생성할 수 있어 생성되는 예제의 다양성을 높인다.

BERT-Attack[4]은 대체할 토큰을 BERT를 이용하여 생성하는 방법이다. 선행 연구들과 같이 토큰의 중요도를 평가한 후 대체할 토큰을 선택할 때 미리 학습시켜둔 BERT를 이용하여 선택한다. 미리 학습시켜 둔 BERT는 교체될 토큰의 자리에 들어갈 토큰을 예측하도록 학습되는데 이 학습시켜둔 BERT가 예측한 토큰들 중 가장 확률이 낮은 토큰을 선택하여 교체한다. 즉, 대체할 토큰 자리를 비워두고 학습시킨 BERT가 비워 둔 토큰 자리에 들어갈 토큰 후보를 예측하도록 한 후 이 예측한 토큰 후보들 중 가장 순위가 낮은 토큰을 선택하여 교체한다. 이는 기존 휴리스틱한 단어 대체 방법보다 더 높은 효율성을 가진다.

2) 화이트박스 공격

대표적인 화이트박스 공격으로는 Guo[1]가 제안한 Gradient-Based Distributional Attack (GDBA)가 있다. GDBA는 자연어의 이산적 특성으로 인해 미분이 불가능하였던 문제를 Gumbel-Softmax[9]로 해결하였다. 이를 통해 미분을 통한 최적화가 가능하게 하여 경사도 기반 적대적 예제 생성 방법을 제안하였다. 또한 기존 블랙박스 공격 방법은

토큰들의 교체가 하나씩 이루어지며 원래 입력에서 교체될 토큰들의 선택 과정과 새롭게 대체될 토큰의 선택 과정이 분리되어 이루어지는 반면, GDBA는 여러 토큰들의 교체가 및 교체될 토큰들의 선택이 복합적으로 이루어지는 장점을 가진다.

2-2 Preliminaries

이 절에서는 경사도기반 화이트박스 적대적 예제 생성 방법인 GDBA를 자세히 소개한다.

토큰 시퀀스 $z = z_1 z_2 z_3 \cdots z_{n-1} z_n$ 에 대해 $z_i \in V$ 이고, $V = 1, \cdots, v$ 는 고정된 단어 집합이다. 이 때 행렬 $\Theta \in R^{n \times v}$ 로 파라미터화된 분포 P_Θ 에 대해

$$z_i \sim Categorical(\pi_i) \quad where \quad \pi_i = Softmax(\Theta_i) \tag{1}$$

이라고 하자. GDBA는 파라미터 행렬 Θ 를 학습시킨 후 $z \sim P_\Theta$ 분포에서 샘플링하여 적대적 예제 z를 얻는 것을 목표로 한다.

z 는 카테고리컬한 분포를 따르기 때문에 경사도 기반으로 Θ를 학습시키기 위해 Gumbel-softmax를 사용한다. Gumble-softmax[9] 분포 $\widetilde{P_\Theta}$ 를 따르는 샘플 $\tilde{\pi} = \widetilde{\pi_1} \cdots \widetilde{\pi_n}$ 는 아래 (2) 식을 통해 샘플링된다.

$$(\widetilde{\pi_i})_j := \frac{\exp((\Theta_{i,j} + g_{i,j})/T)}{\sum_{v=1}^{V} \exp((\Theta_{i,j} + g_{i,v})/T)} \tag{2}$$

$$where\, g_{i,j} \sim Gumbel(0,1),\ T > 0$$

적대적 예제 생성을 위해 먼저 적대적 손실함수를 정의한다. 적대적 손실 함수는 마진(margin) 함수를 이용해 정의한다. 분류 모델 $h : X \rightarrow Y$ 에 대해 입력 x 에 대한 모델 h의 출력인 라벨 y 는 모델 h의 로짓 벡터 $\phi_h(x) \in R^K$ 에 대해 (3)를 만족한다.

$$y = h(x) = \arg\max_k \phi_h(x)_k \in Y \tag{3}$$

정상적인 결과값 y를 가지는 입력 x에 대해 모델 h의 오분류를 유도하기 위한 손실함수의 정의는 (4)와 같다.

$$\ell_{margin}(x, y; h) = \max(\phi_h(x)_y - \max_{k \neq y} \phi_h(x)_k + \kappa, 0) \tag{4}$$

(4)의 식에 대해 손실 함수값이 0이 될 때 모델 h 는 마진 κ 에 대해 입력 x를 오분류하게 된다. 이러한 마진 손실함수는 이미지 데이터에 대한 적대적 예제 생성 시에도 사용한다[14]. 원본 입력 x와 섭동을 준 입력 x'에 대해 두 입력의 차이를 나타내는 함수를 $\rho(x, x')$ 라고 하자. $\rho(x, x') < \epsilon$ 을 만족하면 x와 x'의 차이는 감지 할 수 없다고 할 때, 적대적 예제 생성을 위한 최적화는 식(5)를 만족하는 x'을 찾도록 진행된다.

$$\min_{x' \in X} \ell(x', y; h)\ where\ \rho(x, x') < \epsilon \tag{5}$$

이 때, 텍스트 데이터의 경우 생성되는 적대적 예제와 기존 입력과의 차이를 측정하는 것은 이미지 데이터에 비해 어렵다. 먼저 바뀐 토큰의 수가 적어야 하며 변경된 토큰과 기존 토큰의 의미가 유사해야 한다. 또한 변경된 토큰이 그 문장 속에서 자연스러운 맥락을 가질 수 있도록 해야 한다. 이를 위해 GDBA는 유사도 제약조건과 유창성 제약조건을 제안하였다.

유사도 제약조건은 BERTScore[10]를 이용한다. 입력 토큰 시퀀스 $x = x_1 \cdots x_n$ 와 $x' = x'_1 \cdots x'_n$ 에 대한임베딩 벡터를 각각 $\phi(x) = (v_1, ..., v_n)$,

$\phi(x') = (v'_1, ..., v'_n)$ 라고 할 때 두 토큰 시퀀스의 BERTScore는 (6)와 같다.

$$R_{BERT}(x, x') = \sum_{i=1}^{n} w_i \max_{j=1,\ldots,m} v_i^T v'_j \tag{6}$$
$$where\ w_i := idf(x_i) / \sum_{i=1}^{n} idf(x_i)$$

미분가능하게 하기 위해 유사도 제약조건은 $\rho(x, \pi) = 1 - R_{BERT}(x, \pi)$ 로 정의한다.

유창성 제약조건은 토큰 예측을 목표로 훈련된 Casual language models (CLMs) [17]를 이용하여 토큰 시퀀스의 가능성을 계산한다. 훈련된 CLM인 g가 주어졌을 때 시퀀스 $x = x_1 \cdots x_n$ 에 대한 유창성 제약조건은 (7)로 정의한다.

$$NLL_g(x) = -\sum_{i=1}^{n} \log p_g(x_i \mid x_1 \cdots x_{i-1}) \tag{7}$$
$$where\ \log p_g(x_i \mid x_1 \cdots x_{i-1}) = g(x_1 \cdots x_{i-1})_{x_i}$$

따라서 파라미터 행렬 Θ 는 손실함수 (8)를 이용하여 최적화한다. λ_{lm}, λ_s 는 각각 유창성 제약조건과 유사도 제약조건의 하이퍼파라미터를 의미한다.

$$L(\Theta) = E_{\tilde{\pi} \sim P(\Theta)} \ell(e(\tilde{\pi}), y; h) + \lambda_{lm} NLL_g(\tilde{\pi}) + \lambda_s \rho_g(x, \tilde{\pi}) \tag{8}$$

Θ를 최적화 한 후 적대적 예제는 적대적 분포 P_Θ 에서 샘플링하여 생성한다.

Ⅲ. 제안 방법

3-1 공격 개요

본 논문에서는 기존 GDBA가 제안한 적대적 예제 생성 방법 중 파라미터 행렬 Θ 를 최적화하는 손실함수를 수정하여 성능을 높인다.

기존 GDBA의 최적화 방법은 오분류될 수 있을만큼 충분히 적대적 손실을 감소시키는데에 시간이 오래걸리며, 주어진 반복횟수 범위 내에서 적대적 손실을 충분히 감소시키지 못해 예제 생성에 실패하는 경우가 존재한다.

본 논문은 트랜스포머모델의 구조인 어텐션 맵을 이용하여 손실함수를 개선, 적대적 손실을 빠르게 감소시킬 수 있도록 한다.

3-2 총 어텐션 점수

논문에서는 어텐션 맵을 적대적 예제 생성 공격에 이용하기 위해 총 어텐션 점수를 정의한다.

트랜스포머 모델에서 어텐션 맵은 토큰끼리의 관계성을 수치로 나타낸다. multi-head 어텐션 구조를 가진 모델의 경우 각 어텐션 맵은 다른 성질의 관계를 나타낸다. 각 어텐션 맵이 나타내는 토큰의 관계가 어떤 것인지 정해져 있지는 않지만 명사와 소유격의 관계, 다음 토큰끼리의 관계, 수동태 동사와 주어의 관계, 명사 의미끼리의 관계 등을 나타낸다[15].

본 논문에서는 이러한 어텐션 맵을 이용하여 각 토큰들의 관계를 적대적 입력 생성 시 반영한다. 이를 위해 토큰 사이의 관계를 점수로 계산한 어텐션 맵을 이용한다. 먼저 총 어텐션 점수를 계산하기 위해 모든 어텐션 맵을 하나의 맵으로 합친 총 어텐션 맵을 정의한다. 입력 토큰의 개수가 n이고 타겟 모델의 어텐션 맵의 개수가 l이라고 하자. i번째 어텐션 맵의 값을 [1,0] 범위의 값으로 정규화한 행렬을 $A_i \in R^{n \times n}$ 라고 하자. 이 때 총 어텐션 맵 T의 정의는 다음과 같다.

$$T \tag{9}$$

$\alpha_i \in R$ 은 i 번째 어텐션 맵에 대한 가중치를 나타낸다. 총 어텐션 점수 측정 시 각 어텐션 맵의 성질에 따라 α_i 을 조절하여 가중치를 조정할 수 있다. 이 때, 어텐션 맵은 서로 다른 토큰의 관계를 나타내기 위한 행렬이므로 같은 토큰끼리의 값을 나타내는 총 어탠션 맵의 대각원소는 0으로 설정한다. 또한 입력의 처음과 끝을 나타내는 토큰 또한 적대적 예제 생성시 고려대상이 아니므로 0으로 설정한다.

i번째 토큰과 j번째 토큰의 어텐션 점수$\tau_{i,j}\,(i<j)$ 의 값은 (10)과 같다.

$$\tau_{i,j} = T_{i,j} + T_{j,i} \quad (i,j \in (1,n)) \tag{10}$$

어텐션 점수의 값이 클수록 타겟 모델이 i번째 토큰과 j번째 토큰의 관계성이 높다고 판단함을 뜻한다.

Fig. 1.은 IMDB 데이터셋으로 훈련한 BERT에 'I enjoy this movie, and I like it.' 이라는 토큰 시퀀스에 대한 총 어탠션 맵을 나타낸 그림이다. 해당 맵에서 가장 어텐션 점수 값이 높은 'I' 와 'like' 토큰의 경우 주어와 동사 관계를 가지며 두 번째로 높은 'this'와 'enjoy' 토큰의 경우 동사와 목적어 관계를 가진다. 가장 어텐션 점수가 낮은 'movie'와 'and' 토큰 및 두 번째로 어텐션 점수가 낮은 'movie'와 '.' 의 경우 직관적으로 뚜렷한 관계를 찾을 수 없다. 따라서 총 어텐션 맵에 따른 어텐션 점수는 통사적인 의미를 잘 반영하고 있음을 알 수 있다.

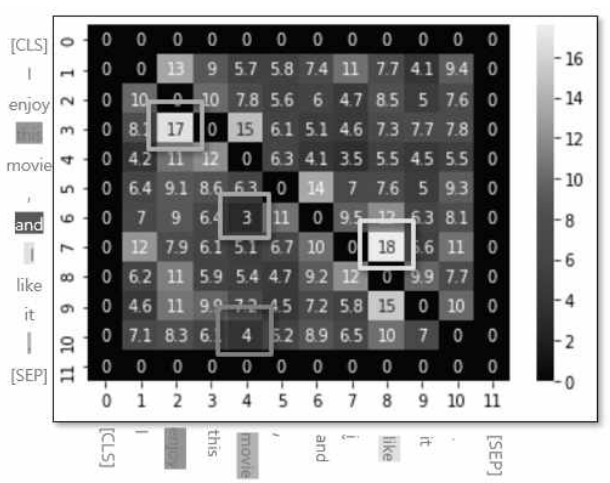

그림 1. IMDB 데이터셋으로 학습한 BERT에서 'I enjoy this movie, and I like it'. 입력의 총 어탠션 맵

Fig. 1. The total attention map for 'I enjoy this movie, and I like it' of BERT trained with IMDB dataset

3-3 어텐션 제약조건

본 논문이 GDBA와 다른 점은 어텐션 제약조건에 있다. GDBA의 방법은 하나의 예제 생성시마다 적대적 분포를 학습시켜 시간이 오래 걸린다. 또한 예제에 따라 분포 학습 후에도 적대적 손실을 충분히 감소시키지 못하여 적대적 예제를 생성하지 못하는 경우가 존재한다. 반복횟수를 증가 시켜 해결할 수 있지만 이는 하나의 예제를 생성하는데 걸리는 시간을 더 늘리기 때문에 바람직하지 않다. 따라서 적대적 예제 생성 시 결과 값에 영향을 많이 미치는 토큰들이 변경될 수 있도록 손실함수를 정의하여 더 빠르게 적대적 손실 값을 감소시킬 것으로 기대할 수 있다.

본 논문은 이를 어텐션 제약조건을 추가하여 해결하였다. 경사도 기반 화이트박스 적대적 예제 생성방안의 경우 블랙박스 공격 방법과 달리 적대적 예제 생성 과정에서 토큰들의

교체가 여러 개씩 동시다발적으로 일어난다. 따라서 기존의 블랙박스 공격 방법과 같이 토큰과 결과 값의 개별적인 관계를 반영하는 것은 바람직하지 않다. 따라서 본 논문은 어텐션 구조인 트랜스포머 모델의 특징을 이용하여 어텐션 점수를 통해 변경되는 토큰들과 모델의 예측값과의 관계를 복합적으로 고려할 수 있도록 어텐션 제약조건(attention constraint, ATC)을 도입하였다. 입력은 n개의 토큰을 가지며 입력을 $x = x_1 x_2 x_3 \cdots x_{n-1} x_n$이라고 하고 학습하고자 하는 적대적 분포에서 샘플링된 값을 $\tilde{\pi} = \tilde{\pi_1}\tilde{\pi_2}\tilde{\pi_3}\cdots \widetilde{\pi_{n-1}}\tilde{\pi_n}$라고 할 때, 어텐션 제약조건 ATC 의 정의는 아래 (11)와 같다.

$$(M_{diff})_{i,j} = \begin{cases} 1 & \text{if } x_i \neq \tilde{\pi}_i \text{ and } x_j \neq \tilde{\pi}_j \\ 0 & else \end{cases} \quad (11)$$

$$(i,j \in (1,n))$$

$$ATC(x,\tilde{\pi}) = T \times M_{diff}$$

어텐션 제약조건은 변경된 토큰들의 어텐션 점수의 합으로 나타낸다. 이는 원래 입력과 적대적 입력 사이에 변경된 토큰들의 어텐션 점수 및 변경된 토큰들의 개수와 비례한다. 이러한 어텐션 제약조건은 적대적 예제를 생성할 때 어텐션 점수가 낮은 토큰들끼리 교체될 수 있도록 한다. 이는 확률적으로 결과값에 영향을 크게 미치는 토큰들이 교체되도록 한다. 결과값에 영향을 많이 미치는 토큰들끼리 교체하는 것은 무작위로 토큰을 변경하는 것 보다 적대적 에제 생성 시 더 효율적이기 때문에 기존 블랙박스 공격 방법에서 많이 이용되고 있다. 또한 원래의 입력에서 변경된 토큰들의 수가 커질수록 어텐션 제약조건도 증가하기 때문에 어텐션 제약조건을 통해 기존 입력과 적대적 입력과 유사하게 할 수 있다.

3-4 적대적 예제 생성

본 논문은 2.2장에서 소개한 공격방법 GDBA의 손실함수에 어텐션 제약조건을 추가하여 성능을 개선시킨다. 따라서 파라미터 행렬 Θ 에 대한 손실 함수 정의는 (12) 와 같다. 여기서 λ_{atc} 는 어텐션 제약조건의 하이퍼파라미터를 의미한다.

$$L(\Theta) = E_{\tilde{\pi} \sim P(\Theta)} \ell(e(\tilde{\pi}), y; h) + \lambda_{lm} NLL_g(\tilde{\pi}) + \lambda_s \rho_g(x,\tilde{\pi}) + \lambda_{atc} ATC(x,\tilde{\pi}) \quad (12)$$

Ⅳ. 실험 및 평가

이 장에서는 본 논문에서 제안한 어텐션 제약조건이 적대적 예제 생성 시 미치는 영향을 분석한다. 또한 어텐션 제약조건을 추가했을 때의 공격 성능을 어텐션 제약조건을 추가하지 않은 공격 성능과 비교하여 평가한다.

4-1 실험 환경 및 실험 데이터

Table 1은 실험이 이루어진 시스템의 환경과 사용한 딥러닝 라이브러리 Pytorch[8]의 버전이다.

실험에 사용한 데이터셋은 Ag News [12] 와 IMDB[13] 데이터셋 두 종이다. Ag News는 4 개의 주제의 뉴스 기사 데이터셋으로 세계, 스포츠, 비즈니스, 과학/기술 4가지 주제로 구분된다. IMDB 데이터셋은 영화 리뷰 데이터셋으로 긍정적인 평가, 부정적인 평가 2종의 라벨로 구성되어 있다.

표 1. 실험 환경
Table 1. Experiment Environment

Experiments	Version
OS	Ubuntu 18.04.5 LTS
CPU	Intel Core i7-10700K
GPU	GeForce GTX 1080 Ti
Pytorch	1.9.1+cu102

4-2 실험 구성

본 논문에서는 공격 대상 모델을 BERT[6]로 선정하여 진행하였다. 유창성 제약조건을 측정하기 위한 CLM g는 해당 BERT와 같은 방식으로 토큰을 구분했고, WikiText-103 데이터셋으로 학습하여 진행하였다.

Θ를 최적화하는데에는 Adam optimizer[11]를 사용하였으며, 학습률(learning rate)은 0.3, 배치 사이즈는 10으로 설정한 후 실험을 진행하였다. 각 하이퍼파라미터는 실험 데이터에 따른 실험 결과를 통해 경험적으로 적절한 값을 찾았다. 유창성 제약조건에 대한 하이퍼 파라미터 $\lambda_{lm} = 1$ 로 설정하였으며 유사도 제약조건의 하이퍼 파라미터는 $\lambda_s = 50$ 로 설정한 후 진행하였다. 어텐션 제약조건에 대한 파라미터는 $\lambda_{atc} \in \{0.01, 0.005\}$ 로 설정 후 진행하였다.

어텐션 제약조건을 생성하는 어탠션 맵 가중치는 모두 1로 같게 설정한 후 실험을 진행하였다.

또한 생성된 예제와 기존 예제의 차이가 충분히 작은지 평가하기 위해 두 값의 유사도를 Universal Sentence Encoders(USE)[7]의 코사인 유사도(cosine similarity)로 측정하였다.

4-3 어텐션 제약조건 분석

Fig.2.는 총 어텐션 점수와 모델의 출력값의 관계를 알아보기 위한 실험 결과를 나타낸다. 토큰 시퀀스 $x = x_1 \cdots x_n$ 에 대해 I와 j 번째 토큰 x_i, x_j 에 대한 총 어텐션 점수를 $\tau_{i,j}$ 라고 할 때, 총 어텐션 점수에 따른 중요도 점수를 계산하였다. 토큰 시퀀스 x에서 토큰 x_i와 x_j를 제외한 토큰 시퀀스를 $\bar{x}$ 라고 하자.

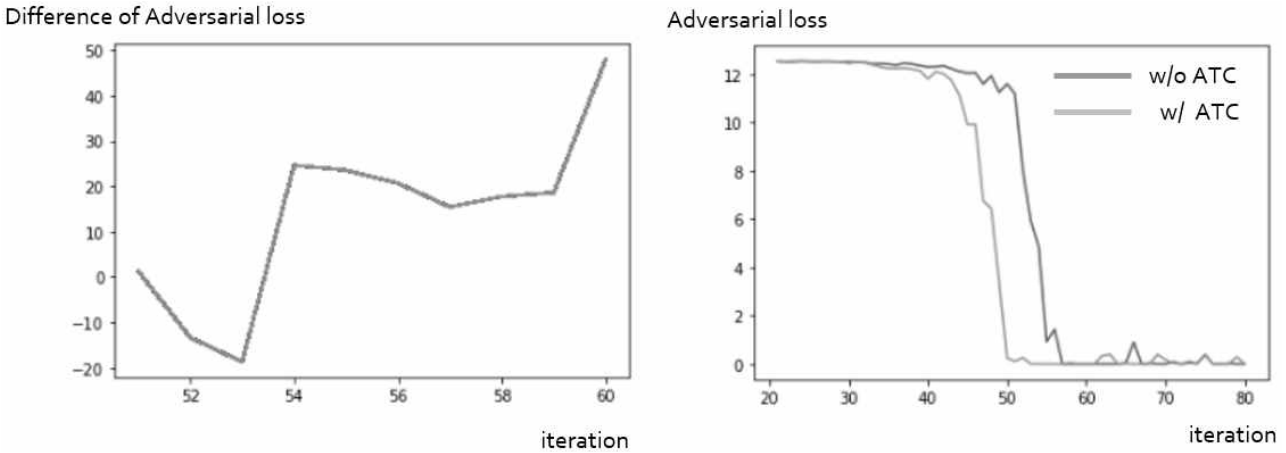

그림 2. GDBA와 본 연구에서 제안한 방법 의 적대적 손실함수 값 차이(좌), 적대적 손실함수의 감소 속도(우)
Fig. 2. The difference of adversarial loss between origin GDBA and our method(Left), Adversarial loss declining trend origin GDBA and our method (Right)

모델 h에 대한 입력의 분류 결과를 각각 $y = h(x), \overline{y} = h(\overline{x})$ 라고 하고 로짓 벡터는 $y = \phi(x), \overline{y} = \phi(\overline{x})$ 라고 하자. 이 때 $\phi_y(x)$는 라벨 y에 대한 로짓 벡터값을 나타낸다. 이 때, 총 어텐션 점수 $\tau_{i,j}$ 에 대한 중요도 점수 $I_{i,j}$ 값은 (13)과 같다.

$$I_{i,j} = \begin{cases} \phi_y(x) - \phi_{\overline{y}}(\overline{x}) & \text{if } y = \overline{y} \\ \phi_y(x) - \phi_y(\overline{x}) & \\ \quad + \phi_{\overline{y}}(\overline{x}) - \phi_{\overline{y}}(x) & \text{if } y \neq \overline{y} \end{cases} \tag{13}$$

중요도 점수가 클수록 해당 총 어텐션 점수가 모델의 결과값에 큰 영향을 미치는 것을 의미한다.

Fig. 3. 은 총 어텐션 점수에 따른 중요도 점수 분포를 나타낸 그래프이다. 이 그래프를 통해 총 어텐션 점수가 높을수록 중요도 점수가 낮게 분포되어 있으며 중요도 점수가 높을수록 총 어텐션 점수가 낮은 부분에 분포되어 있는 것을 확인할 수 있다. 이를 통해 총 어텐션 점수가 낮은 토큰들을 중심적으로 교체할 경우 입력에 대한 모델의 출력값에 더 많은 영향을 미치는 경향이 있는 것을 확인할 수 있다. 즉 각 토큰들의 관계와 모델의 출력이 유의미한 관계가 있음을 의미한다.

따라서 총 어텐션 점수가 낮은 토큰들을 먼저 교체하는 것이 적대적 예제 생성 시 더 효율적임을 확인할 수 있다. 이를 확인하기 위해 각 반복횟수에서의 어텐션 제약조건을 추가한 GDBA의 적대적 손실값과 어텐션 제약조건을 추가하지 않은 GDBA의 적대적 손실 값을 비교하였다.

Fig.2.(Right) 는 같은 AG News의 데이터에서 어텐션 제약조건을 추가한 방법의 적대적 손실과 추가하지 않은 방법의 적대적 손실 값을 그래프로 나타낸 것이다. 어텐션 제약조건을 추가한 경우 적대적 손실이 0에 가까이 가는 시점이 기존의 방법보다 약 10 반복횟수 빠른 것을 확인할 수 있다.

Fig.2.(Left) 는 생성되는 적대적 예제에서 반복횟수에 따른 이러한 적대적 손실의 차이를 그래프로 나타낸 것이다. 차이값이 음수일 경우 기존 방법의 적대적 손실이 작고, 양수일 경우 어텐션 제약조건을 추가한 방법의 적대적 손실이 작음을 뜻한다. 이 그래프를 통해 어텐션 제약조건이 추가되었을 때 적대적 손실이 더 빠르게 감소함을 알 수 있다.

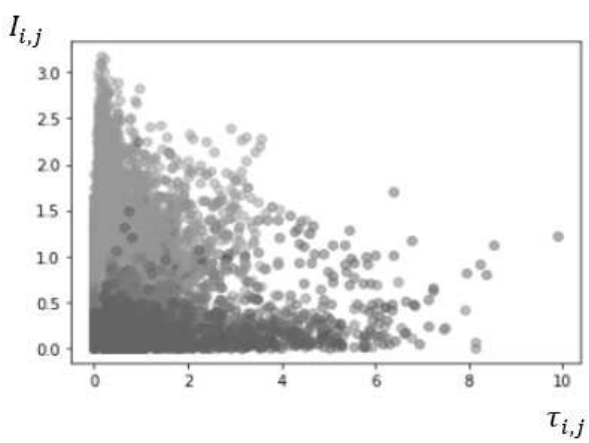

그림 3. 전체 어텐션 점수와 중요도 점수와의 관계
Fig. 3. The relationship with total attention score and important score

4-4 공격 성능 측정

기존 방법은 하나의 적대적 예제를 생성하는 데에 하나의 분포를 학습시켜야 하기 때문에 효율성이 낮다. 또한 생성시 지정한 반복횟수를 거쳐 최적화한 후에도 적대적 손실을 충분히 감소시키지 못하여 적대적 예제 생성에 실패하는 경우도 존재한다. 본 논문이 제안한 어텐션 제약조건을 추가할 경우 같은 수준의 적대적 예제를 만드는 데에 드는 시간을 줄일 수 있음을 보이고 기존의 방법으로는 적대적 손실을 충분히 감소시키지 못해 생성하지 못했던 적대적 예제를 생성할 수 있음을 확인한다.

표 2. 어텐션 제약조건을 추가한 방법과 추가하지 않은 방법으로 생성한 AG News 데이터셋에 대한 적대적 예제 예시
Table 2. Adversarial Examples on AG News generated with attention constraint and without attention constraint.

Attack	Prediction‘	Text
Original	Sports	callender wins job as starter frustration set in quickly for andre callender. he had already waited a whole year, and now he had to wait another game to play college football.
GDBA w/o ATC	Sports (Fail)	callender wins job as starter frustration set **mount** quickly for andre **callserer.** he had already waited a **a** year, and now **now** had **needed** wait another game to play college football.
GDBA w/ ATC	Business (Success)	callender wins **awards** as **ceo** frustration set in quickly for **bill** callender. he had already waited a whole year, **yet** now had to wait another game **pass** play college football.
Original	Sports	giants give up right to void bonds'deal (ap) ap – barry bonds will have two more seasons to break hank aaron's career home run record with the san francisco giants, who decided tuesday to drop their right to void the final year of his contract.
GDBA w/o ATC	Sports (Fail)	giants give up right to void bonds'deal (ap) ap – barry bonds will have two more seasons to break **cs** aaron's career home run record **leader** the san francisco giants, who decided tuesday to drop their right to void the final year of his contract.
GDBA w/ ATC	World (Success)	giants give **choose** right to void bonds'deal (ap) ap – barry bonds will play two more seasons **hr** break hank aaron's career home run record with the san francisco giants, who decided tuesday to drop their right to void the final year of his contract.

Table 2. 는 같은 토큰 시퀀스에 대해 기존의 방법과 어텐션 제약조건을 추가한 방법에서 생성한 적대적 예제이다. 변경된 토큰의 수는 같으나 기존의 방법은 적대적 손실을 충분히 감소시키지 못해 모델의 오분류를 유도하지 못한 반면 어텐션 제약조건을 추가했을 경우 모델의 오분류를 유도했음을 알 수 있다.

Table 3 은 AG News와 IMDB 데이터셋에 대해 입력 데이터 중 기존 방법으로 모델의 오분류에 성공한 적대적 예제 생성률 및 적대적 예제와 기존 입력의 유사도를 나타내었다. 여기서 오분류률은 입력 데이터가 최적화 과정을 거친 후 샘플링되었을 때 대상 모델에서 잘못된 결과값으로 출력되는 비율을 나타낸다. 즉 이러한 오분류율은 곧 적대적 예제 생성률을 의미하며 높을수록 같은 데이터셋에서 더 다양한 적대적 예제를 생성할 수 있음을 의미한다.

AG News 의 데이터셋을 대상으로 적대적 예제를 생성했을 때 어텐션 제약조건을 추가할 경우 100번의 반복횟수를 거쳐 최적화한 분포를 통해 샘플링한 예제가 모델에서 오분류 될 확률은 75.7%였으며 기존의 방법은 73.3%로 적대적 예제 생성률을 2.4% 증가시켰다. 또한 어텐션 제약조건을 추가한 후 반복횟수를 90으로 줄일 경우 오분류율은 73.6%로 기존 방법보다 0.3퍼센트 향상되었으며 기존의 방법보다 소요 시간이 6.5% 단축되었음을 확인할 수 있다. 또한 코사인 유사도도 0,82로 같았다. 즉 어텐션 제약조건을 추가할 경우 같은 수준의 유사도를 가진 적대적 예제를 더 다양하고 빠르게 생성할 수 있음을 확인할 수 있다.

IMDB 데이터셋의 경우 텍스트 데이터의 평균 길이가 AG News 데이터보다 길어 어텐션 제약조건을 계산하는 데에 소요시간은 증가하였으나 같은 반복횟수 내에서 오분류율을 최대 0.9% 향상시킬 수 있었다.

V. 결 론

본 논문에서는 텍스트 트랜스포머 모델에 대한 경사도 기반 화이트 박스 적대적 예제 생성방법에 어텐션 맵을 이용한 어텐션 제약조건을 손실함수에 추가하여 적대적 예제 생성률을 높일 수 있음을 입증하였다. 또한 어텐션 제약조건을 이용하여 토큰의 교체가 효율적으로 일어나도록 하였고 이에 따른 적대적 손실의 감소 속도가 빨라짐을 입증하였다. 결과적으로 이는 적대적 예제 생성 시간을 단축시킬 수 있음을 보였다.

향후 연구를 통해 본문에서 제시한 총 어텐션 점수와 결과로짓 값의 연관관계의 수학적인 의미를 분석하고 이를 일반화하여 더욱 다양한 모델에 적용하는 연구를 통해 제시한 공격 방안을 발전시킬 수 있을 것이다. 또한 본 논문의 연구방안은 타겟 미지정 공격방안이지만 타겟 지정 공격을 사용하는 화이트 박스 공격 방안에 대해서도 연구를 진행할 수 있다.

표 3. 어텐션 제약조건을 추가한 GDBA와 추가하지 않은 GDBA 공격 성능 평가
Table 3. Evaluation of attack algorithm GDBA without attention constraint and with attention constraint.

Task	Attack Alg (iteration num)	Clean cc	missclassfication rate	Cosine Sim	Time Avg
AG News	GDBA (iter=100)	95.1	73.3	**0.82**	11.26
	GDBA w/ ATC (iter = 100)		**75.7**	0.81	11.67
	GDBA w/ ATC (iter = 90)		73.6	**0.82**	10.53
	GDBA w/ ATC (iter = 80)		69.3	0.81	**9.46**
IMDB	GDBA (iter=100)	92.0	98.2	0.90	**28.20**
	GDBA w/ ATC (iter = 100)		**99.1**	**0.91**	34.10
	GDBA w/ ATC (iter = 90)		97.4	0.90	30.71
	GDBA w/ ATC (iter = 80)		92.4	0.90	27.38

참고문헌

[1] Guo C., Sablayrolles, A., Jégou, H., & Kiela, D. “Gradient-based Adversarial Attacks against Text Transformers”. arXiv preprint arXiv:2104.13733. Apr, 2021

[2] D Jin, Z Jin, JT Zhou, P Szolovits, “Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment”, *AAAI Technical Track: Natural Language Processing,* Vol. 34 No. 05: AAAI-20 Technical Tracks 5,2020. https://doi.org/10.1609/aaai.v34i05.6311

[3] S Garg, G Ramakrishnan, “Bae: Bert-based adversarial examples for text classification”, arXiv preprint arXiv:2004.01970, Oct, 2020

[4] L Li, R Ma, Q Guo, X Xue, X Qiu, “Bert-attack: Adversarial attack against bert using bert”, arXiv preprint arXiv:2004.09984, Oct, 2020

[5] Ian J Goodfellow, Jonathon Shlens, Christian Szegedy, “Explaining and harnessing adversarial examples.”. arXiv preprint arXiv:1412.6572.Dec,2014.

[6] J Devlin, MW Chang, K Lee, K Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv:1810.04805, Oct, 2018

[7] D Cer, Y Yang, S Kong, N Hua, N Limtiaco, “Universal sentence encoder”, arXiv:1803.11175,Mar,2018

[8] A Paszke, S Gross, F Massa, A Lerer, “Pytorch: An imperative style, high-performance deep learning library”, *Advances in Neural Information Processing Systems* 32 ,NeurIPS, 2019

[9] E Jang, S Gu, B Poole, “Categorical reparameterization with gumbel-softmax”, arXiv preprint arXiv:1611.01144, Nov, 2016

[10] T Zhang, V Kishore, F Wu, KQ Weinberger, “Bertscore: Evaluating text generation with bert”, arXiv preprint arXiv:1904.09675 , Apr, 2019

[11] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980. Dec, 2014.

[12] Xiang Zhang, Junbo Zhao, and Yann LeCun. “Character-level convolutional networks for text classification.”, arXiv preprint arXiv:1509.01626,2015.

[13] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. “Learning word vectors for sentiment analysis.”, *In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150,2011.

[14] Nicholas Carlini, David Wagner. “Towards evaluating the robustness of neural networks.” I*n 2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE ,2017.https://ieeexplore.ieee.org/document/7958570

[15] K Clark, U Khandelwal, O Levy, CD Manning, “ What does bert look at? an analysis of bert's attention”, arXiv preprint arXiv:1906.04341,Jun, 2019

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras,, Adrian Vladu. “Towards deep learning models resistant to adversarial attacks.” arXiv preprint arXiv:1706.06083.Jun, 2017.

[17] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9., 2019.

신초별(Cho-Byeol Shin)

2020년 : 서울시립대학교 자연과학대학

2020년～현 재 고려대학교 정보보호대학원 석사과정

※관심분야 : 정보보호, 시스템 및 네트워크 보안, 딥러닝 등

문종섭(Jong-Sub Moon)

1981년 : 서울대학교 자연과학대학

1983년 : 서울대학교 대학원

1992년 : Ilinois Institute of Technology

1993년～현 재 고려대학교 전자및정보공학과 교수

※관심분야 : 정보보호, 시스템 및 네트워크 보안, 패턴인식 등