텍스트 분류 모델 공격 기법

TextFooler

임세진

https://youtu.be/93XchRMfX4U

한성대학교 HANSUNG UNIVERSITY

CryptoCraft LAB

01. 텍스트 분야에서 적대적 공격이 어려운 이유

02. TextFooler 소개

03. 실험 결과

논문 링크 : https://arxiv.org/pdf/1510.00149.pdf

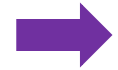
01. 텍스트 분야에서 적대적 공격이 어려운 이유

- **적대적 공격 (Adversarial Attack)**

  딥러닝 모델의 내부적인 취약점을 이용하여 만든 특정 노이즈 값을 이용해 입력 값을 생성하여

  의도적으로 딥러닝 모델의 오분류를 유발시키는 공격 기법

20	20	65	25
45	95	30	90
50	60	80	110



22	18	63	23
47	94	32	92
48	62	78	112

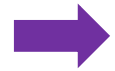
- **이미지 도메인**

  ➢ 픽셀의 값이 연속적임 ➔ 많은 픽셀에 조금씩 노이즈를 섞어도 사람 눈에 잘 띄지 않음

- **자연어 도메인**

  One-Hot Encoding

  ➢ 단어나 문자가 불연속적인 토큰임 ➔ 약간의 변화만 생겨도 사람 눈에 잘 띄는 편

"He loves you so much" 

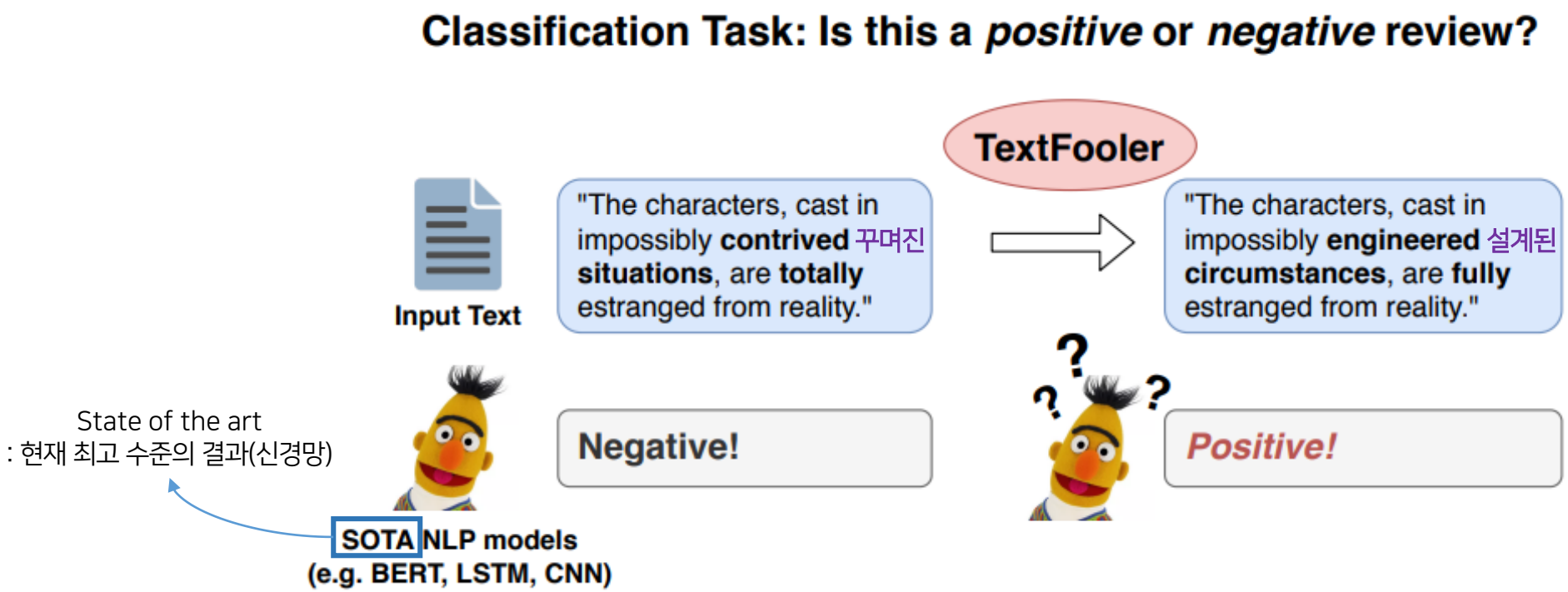
"He loves you a lot"

02. TextFooler 소개

- TextFooler

- 텍스트 분류 모델을 속일 수 있는 강력한 공격 기법

- 원본 텍스트를 약간 변경하여 영화 리뷰 모델을 속였음(공격)



02. TextFooler 소개

- TextFooler 특징

- Effective(공격효과성) : 이전 공격 기법들보다 성공율 ↑ && 변화를 준 부분 ↓

  (즉, 단어를 조금만 바꿔도 공격 잘 됨)

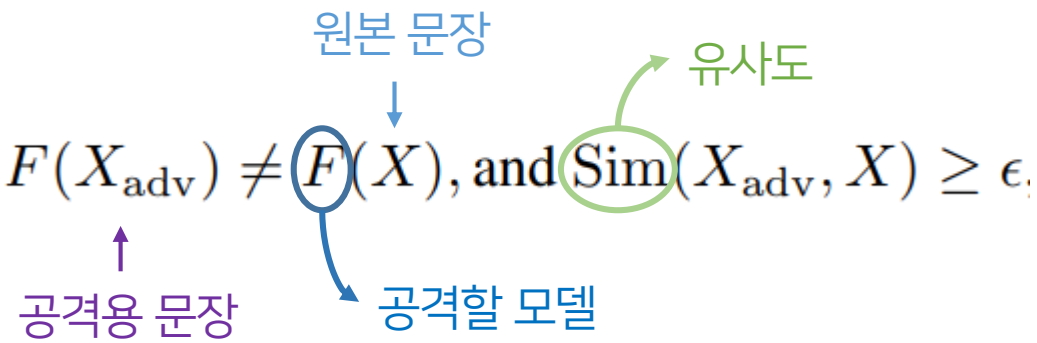
- Utility-preserving(사람이 보기에 같은 의미로 판단) : 의미 보존, 올바른 문법

- Efficient(효율성) : 다른 공격 기법들과 비교했을 때 적은 연산 복잡도

02. TextFooler 소개

- **텍스트 분야에서 공격 데이터에 대한 요구사항**

- 공격 데이터의 의미가 원본 데이터와 크게 달라지지 않으면서, 딥러닝 모델의 오분류를 유발하도록



02. TextFooler 소개

- **TextFooler를 적용할 수 있는 공격 기법의 종류**

➢ Black-box attack

: 공격하고자 하는 모델에 대해 상세 파라미터나 구조를 알지 못하는 제한적 상황에서 공격 수행

데이터 입력한 후에 분류 결과와 confidence score(그 데이터일 확률)을 알 수 있다고 가정

➢ Targeted & Non-targeted attack

: 특정 클래스로 분류되도록 공격하는 기법, 기존 클래스 외의 클래스로 분류되도록 공격하는 기법

➢ Word-wise perturbing attack

: 문장의 단어 단위로 변경을 수행하며 공격하는 기법

02. TextFooler 소개

- TextFooler 공격 알고리즘

1. 문장에 포함된 단어를 중요도에 따라 순위를 매김

2. 단어 변경

    1) 동의어 추출 : 단어와 유사한 단어 찾기 (love, like)

    2) 품사 체크 : 바꾸고자 하는 단어가 원본과 같은 품사를 갖는지 확인 for 문법 오류 방지

    3) 전체 문장이 원본과 유사한 의미를 갖는지 확인

    4) 단어 후보군 중에 모델의 결과를 바꿀 수 있는 단어를 찾음

    ➔ 즉, 클래스 분류에 많은 영향을 주는 단어 찾아내고 바꿈

02. TextFooler 소개

- **단어 중요도 순위 매기기**

- 일반적으로 문장이 n개의 단어로 구성될 때, 특정 몇 개의 단어만 실제 분류 결과에 큰 영향을 미침

- 중요한 단어 위주로 변경하면 공격이 더욱 잘 수행될 것임

- **단어 중요도 순위 구하는 방법**

- 전체 문장이 있을 때 문장에 포함된 단어를 하나씩 제외해서 타겟 모델에 넣어봄 (관사 제외)

  - 전체 문장 : He loves me
  - Test1 : loves me
  - Test2 : He me
  - Test3 : He loves

- 단어의 중요도가 높다면 그 단어가 없을 때 결과값이 크게 달라질 것이라는 아이디어

$$X_{\setminus w_i} = X \setminus \{w_i\} = \{w_1, \ldots, w_{i-1}, w_{i+1}, \ldots w_n\}$$

$$I_{w_i} = \begin{cases} F_Y(X) - F_Y(X_{\setminus w_i}), & \text{if } F(X) = F(X_{\setminus w_i}) = Y \\ (F_Y(X) - F_Y(X_{\setminus w_i})) + (F_{\bar{Y}}(X_{\setminus w_i}) - F_{\bar{Y}}(X)), & \\ \quad \text{if } F(X) = Y, F(X_{\setminus w_i}) = \bar{Y}, \text{ and } Y \neq \bar{Y}. & \end{cases}$$

결과값이 달라졌을 때 높은 값을 갖게 됨

I_{w_i} : 중요도

02. TextFooler 소개

2. 단어 변경

1) 동의어 추출 : 단어와 유사한 단어 찾기 (love, like) 후보군 추출 (단어를 벡터화 할 수 있는 신경망을 이용)

2) 품사 체크 : 바꾸고자 하는 단어가 원본과 같은 품사를 갖는지 확인 for 문법 오류 방지

3) 전체 문장이 원본과 유사한 의미를 갖는지 확인

4) 단어 후보군 중에 모델의 결과를 바꿀 수 있는 단어를 찾음

➔ 즉, 클래스 분류에 많은 영향을 주는 단어 찾아내고 바꿈

Initialization: $X_{\text{adv}} \leftarrow X$

FINCANDIDATES $\leftarrow \{\}$
for c_k in CANDIDATES **do**
  $X' \leftarrow$ Replace w_j with c_k in X_{adv}
  if $\text{Sim}(X', X_{\text{adv}}) > \epsilon$ **then**
    Add c_k to the set FINCANDIDATES
    $Y_k \leftarrow F(X')$ 신경망에 넣어보고
    $P_k \leftarrow F_{Y_k}(X')$ 결과값 저장
  end if
end for

앞에서 저장한 후보군에서 하나씩 뽑아 테스트

1) 후보군이 모델의 결과값을 바꾸면 그 중에서 유사도가 높은 단어 선택 [공격 성공]

2) 결과값을 바꾸지 못할 경우, confidence score를 최대한 낮추는 단어 선택

1의 과정에서 중요도가 높은 단어들 위주로 2의 과정 수행

중간에 결과값을 바꾸면 멈추고 그 문장을 공격용 데이터로 정함

03. 실험결과

- 실험에 사용한 데이터셋

평균 문장의 길이(단어 수)

Task	Dataset	Train	Test	Avg Len
Classification	AG's News	120K	7.6K	43
	Fake News	18.8K	2K	885
	MR	9K	1K	20
	IMDB	25K	25K	215
	Yelp	560K	38K	152
Entailment	SNLI	570K	3K	8
	MultiNLI	433K	10K	11

두개의 문장이 있을 때 두 문장의 관계성(논리적으로 같음, 모순, 중립)을 찾는 문제

평균적으로 분류되는 클래스의 수는 4개 이하 ➔ 자연어 처리 분야는 이미지 분야와 다르게 적은 클래스 상에서도 공격이 쉽지 않음

03. 실험결과

- 공격 대상 모델 (state of art model)

	WordCNN	WordLSTM	BERT
AG	92.5	93.1	94.6
Fake	99.9	99.9	99.9
MR	79.9	82.2	85.8
IMDB	89.7	91.2	92.2
Yelp	95.2	96.6	96.1
	InferSent	**ESIM**	**BERT**
SNLI	84.6	88.0	90.7
MultiNLI	71.1/71.5	76.9/76.5	83.9/84.1

03. 실험결과

- 공격 수행 결과

정답 클래스가 아닌 클래스로 오분류 되면 성공으로 판단

	WordCNN					WordLSTM					BERT				
	MR	IMDB	Yelp	AG	Fake	MR	IMDB	Yelp	AG	Fake	MR	IMDB	Yelp	AG	Fake
Original Accuracy	78.0	89.2	93.8	91.5	96.7	80.7	89.8	96.0	91.3	94.0	86.0	90.9	97.0	94.2	97.8
After-Attack Accuracy	2.8	0.0	1.1	1.5	15.9	3.1	0.3	2.1	3.8	16.4	11.5	13.6	6.6	12.5	19.3
% Perturbed Words	14.3	3.5	8.3	15.2	11.0	14.9	5.1	10.6	18.6	10.1	16.7	6.1	13.9	22.0	11.7
Semantic Similarity	0.68	0.89	0.82	0.76	0.82	0.67	0.87	0.79	0.63	0.80	0.65	0.86	0.74	0.57	0.76
Query Number	123	524	487	228	3367	126	666	629	273	3343	166	1134	827	357	4403
Average Text Length	20	215	152	43	885	20	215	152	43	885	20	215	152	43	885

	InferSent		ESIM		BERT	
	SNLI	MultiNLI (m/mm)	SNLI	MultiNLI (m/mm)	SNLI	MultiNLI (m/mm)
Original Accuracy	84.3	70.9/69.6	86.5	77.6/75.8	89.4	85.1/82.1
After-Attack Accuracy	3.5	6.7/6.9	5.1	7.7/7.3	4.0	9.6/8.3
% Perturbed Words	18.0	13.8/14.6	18.1	14.5/14.6	18.5	15.2/14.6
Semantic Similarity	0.50	0.61/0.59	0.47	0.59/0.59	0.45	0.57/0.58
Query Number	57	70/83	58	72/87	60	78/86
Average Text Length	8	11/12	8	11/12	8	11/12

03. 실험결과

- 공격용 문장이 생성된 모습

Movie Review (Positive (POS) ↔ Negative (NEG))	
Original (Label: NEG)	The characters, cast in impossibly ***contrived situations***, are ***totally*** estranged from reality.
Attack (Label: POS)	The characters, cast in impossibly ***engineered circumstances***, are ***fully*** estranged from reality.
Original (Label: POS)	It cuts to the ***knot*** of what it actually means to face your ***scares***, and to ride the ***overwhelming*** **metaphorical wave** that life wherever it takes you.
Attack (Label: NEG)	It cuts to the ***core*** of what it actually means to face your ***fears***, and to ride the ***big*** **metaphorical wave** that life wherever it takes you.
SNLI (Entailment (ENT), Neutral (NEU), Contradiction (CON))	
Premise	Two small boys in blue soccer uniforms use a wooden set of steps to wash their hands.
Original (Label: CON)	The boys are in band ***uniforms***.
Adversary (Label: ENT)	The boys are in band ***garment***.
Premise	A child with wet hair is holding a butterfly decorated beach ball.
Original (Label: NEU)	The ***child*** is at the ***beach***.
Adversary (Label: ENT)	The ***youngster*** is at the ***shore***.

03. 실험결과

- 기존의 다른 공격 기법들과 결과 비교

더 적은 단어만 바꾸고 더 높은 성공률

Dataset	Model	Success Rate	% Perturbed Words
IMDB	(Li et al. 2018)	86.7	6.9
	(Alzantot et al. 2018)	97.0	14.7
	Ours	**99.7**	**5.1**
SNLI	(Alzantot et al. 2018)	70.0	23.0
	Ours	**95.8**	**18.0**
Yelp	(Kuleshov et al. 2018)	74.8	-
	Ours	**97.8**	**10.6**

03. 실험결과

- 양도성 (transferability)

		WordCNN	WordLSTM	BERT
IMDB	WordCNN	—	84.9	90.2
	WordLSTM	74.9	—	87.9
	BERT	84.1	85.1	—
		InferSent	**ESIM**	**BERT**
SNLI	InferSent	—	62.7	67.7
	ESIM	49.4	—	59.3
	BERT	58.2	54.6	—

Q & A