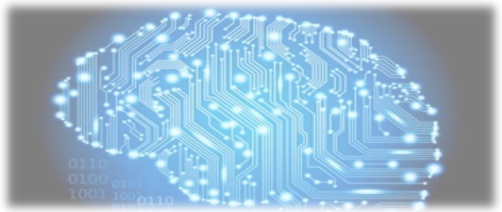
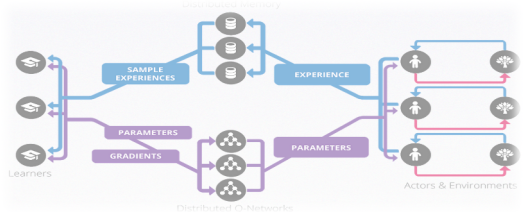



BERT를 사용한 Homoglyph Attack 복원

AutoML 연구실

이한용

2022. 1. 30.

BERT를 사용한 Homoglyph Attack 복원

To Do List

- Last Week
  1. 문제 정의
  2. 데이터 수집 계획
- This Week
  1. 문제 정의
  2. 데이터 수집 계획
  3. 선행연구 조사
  4. 모델 구축

BERT를 사용한 Homoglyph Attack 복원

0. Papers

- Boucher, Nicholas, et al. "Bad Characters: Imperceptible NLP Attacks." arXiv preprint arXiv:2106.09898 (2021).
- Deng, Perry et al. “Weaponizing Unicodes with Deep Learning -Identifying Homoglyphs with Weakly Labeled Data.” 2020 IEEE International Conference on Intelligence and Security Informatics (ISI) (2020): 1-6.

BERT를 사용한 Homoglyph Attack 복원

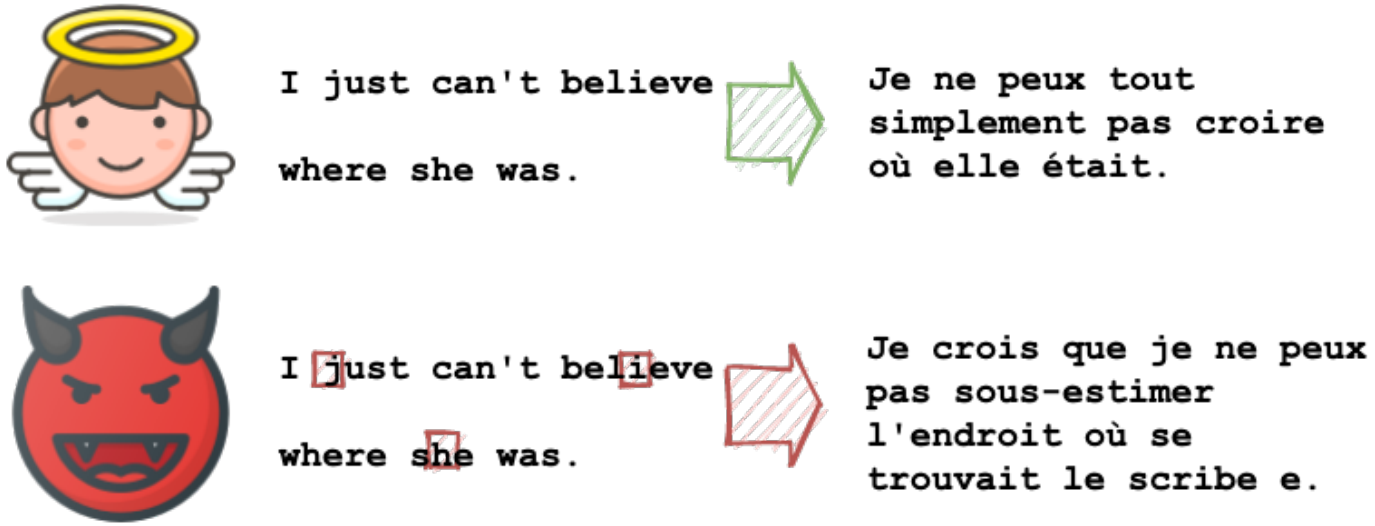
1. 문제 정의

- 스팸메일 등 에서 자동 스팸 필터링 시스템을 우회하기 위해 homoglyph등을 사용한 'Adversarial Example'이 존재함
- NLP 에서의 Adversarial Example은 다음과 같은 예시가 있음
  1. 모양이 비슷한 homoglyph를 사용 (Homoglyph Attack; Visual Spoofing)
     예시: "j" (U+006A; LATIN SMALL LETTER J) 대신
     " j " (U+FF4A; FULLWIDTH LATIN SMALL LETTER J)를 사용
  2. 렌더링 되었을 때 보이지 않는 문자(NPC, Non-Printing Character)를 단어 사이에 삽입
     예시: 단어 사이에 U+200B (ZERO WIDTH SPACE)를 삽입
  3. 문자의 순서를 바꾸거나 문자를 지우는 등의 제어문자(Control Character)를 사용

BERT를 사용한 Homoglyph Attack 복원

1. 문제 정의

- NLP Attack은 컴퓨터가 텍스트의 의미를 이해하는 것을 방해함.
- NLP Attack의 한 종류인 **Homoglyph Attack**이 기계번역에 혼선을 주는 예시.
  (Boucher, Nicholas, et al. "Bad Characters: Imperceptible NLP Attacks." arXiv preprint arXiv:2106.09898 (2021).)



BERT를 사용한 Homoglyph Attack 복원

2. 본 연구에서 제시하는 문제의 해법

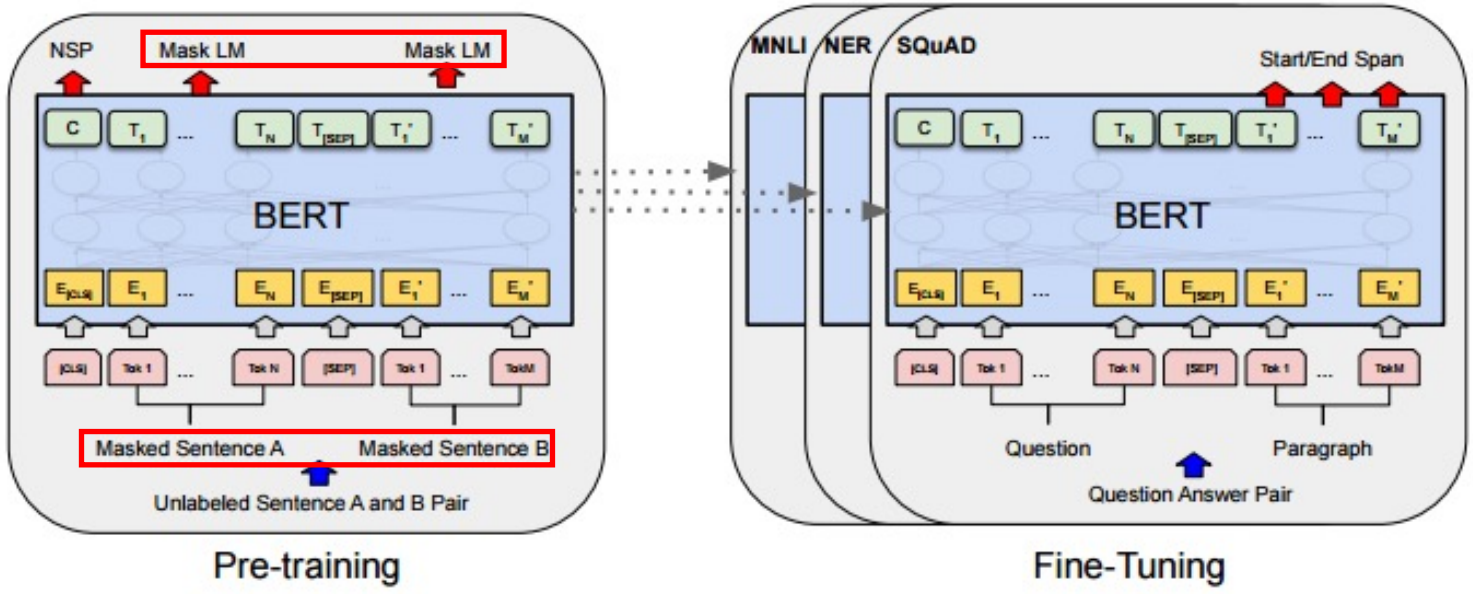
- homoglyph 복원의 input/output 예시

homoglyph가 포함된 input		output
Yoս trɑnsfer $1831 USD to ᴍe (ịn Bitcoịn equịvalent ɑccordịng to the exchɑnge rɑte ɑt the ᴍoᴍenᴛ of ḟuɴd's trɑɴsfer), ɑɴd oɲce the trɑɴsfer ịs received, I ɯill ɗ'elete ɑll thịs dirty stuḟḟ right ɑɯay. Ạfter thɑt, ɯe ɯill ḟorget ɑbout eɑch otheɾ. I ɑlso ρroᴍise to deɑctivɑte ɑɴd delete all tнe hɑrᴍful soḟtɯɑre ḟroᴍ youɾ ɗ'evịces.	추론	You transfer $1831 USD to me (in Bitcoin equivalent according to the exchange rate at the moment of funds transfer), and once the transfer is received, I will delete all this dirty stuff right away. After that, we will forget about each other. I also promise to deactivate and delete all the harmful software from your devices.

BERT를 사용한 Homoglyph Attack 복원

2. 본 연구에서 제시하는 문제의 해법

- BERT: BERT 모델은 Pre-training 단계에서 일정 비율로 Masking 된 토큰들을 예측하는 Masked Language Model (MLM) 방법을 사용함



- 본 연구는 BERT의 MLM 방법을 응용하여 문장 내의 homoglyph를 예측하는 것을 목표로 함

BERT를 사용한 Homoglyph Attack 복원

3. 데이터 수집

- 비트코인 갈취 목적 스팸메일 (Extortion email) 정보를 수집한 사이트 **bitcoinabuse.com**의 데이터베이스를 사용
- 258766 rows × 9 columns으로 구성된 .csv 형식 파일의 데이터

id	address	abuse_type_id	abuse_type_other	abuser	description	from_country	from_country_code	created_at
362667	15ScqqupNT4ToNXCdwmgHLWS3cHTtTXLs6	4	NaN	bitcoin tel	they take my money not payme back it scam	United States	US	2021-12-15T02:26:21.000000Z
362668	1EeaB9n9RKCsxbzzmVNNPbcu9Lrq89Bm4Y	4	NaN	223.72.80.48	The IP address of the email was sent from is: ...	Canada	CA	2021-12-15T02:34:52.000000Z
362669	bc1qxqm9yp4ly807xmfpqkge6dpqtett4k6kr4krfd	99	Fraudulent crypto scam	306cryptomax	Fraudsters took over a friends Insta gram acco...	United States	US	2021-12-15T02:51:50.000000Z

4. 데이터 가공

- 원본 데이터에서 필요한 column만 추려 냄
- 추린 column 목록
  - **id**: 각 report의 고유 번호
  - **description**: report의 설명 또는 스팸메일 원문

- 영어 이외의 언어가 포함되어 있는 데이터는 제외

- 추려낸 데이터 셋을 description non-ascii 문자의 존재 여부에 따라 두 파일로 분리
  - ascii 문자만 존재하는 데이터 셋을 masked language model의 학습용으로 사용하고, non-ascii 문자가 들어있는 데이터 셋은 수작업 레이블링 하여 파인튜닝 및 테스트용으로 사용할 예정

BERT를 사용한 Homoglyph Attack 복원

5. 사용 모델

- Hugging Face의 BERT model API (https://huggingface.co/docs/transformers/model_doc/bert)를 이용

- BertTokenizer의 vocab_file을 재정의하여 문자 단위로 토큰화할 수 있게 변경

- vocab_file의 예시 (.txt format)
  https://github.com/lhy0718/bert-character-mlm/blob/master/char-bert-base-uncased/vocab.txt

BERT를 사용한 Homoglyph Attack 복원

6. 실험 계획

- homoglyph 추론 절차
  1. 제어문자가 적용된 문자열을 구함
  2. 문자열에서 각 문자를 OCR를 이용하여 ascii 문자로 변환
  3. 문자 수준의(Character-level) BERT 모델을 사용하여 결과를 보정

- 문자 수준 BERT 학습 절차
  1. 학습: ascii 문자로만 이루어진 영어 문장을 MLM방식으로 BERT에 학습 (Pretraining)
     - 성능 향상을 위해 외부 말뭉치를 추가할 수 있음
  2. 성능 평가: 비-ascii 문자가 포함된 영어 문장은 homoglyph에 대한 레이블을 사용하여 문자 수준 BERT의 성능을 평가

7. 실험 진행 상황

1. 모델 구축
   - Character-level BERT를 작성
2. 데이터셋 구축
   - Bitcoinabuse.com에 2017-05-16부터 2021-01-15까지 보고된 데이터를 다운로드 받음
   - 약 10000여개의 데이터에 대해 영어 문장 여부 레이블링