



Restoration of Homoglyph Attack with BERT (4)



AutoML Lab

Hanyong Lee

2022. 2. 22.

Restoration of Homoglyph Attack with BERT (4)

✓ To Do List

- Last Week
 1. Making Translation Dictionary for baseline algorithms
 2. Preprocessing the dataset
- This Week
 1. Refining English data
 2. Splitting ASCII-only / homoglyph set

Restoration of Homoglyph Attack with BERT (4)

1. Refining English data – Introduction

1. The Problem

- Spam Dataset from BitcoinAbuse.com is a multilingual dataset.
- Refining English data is required.
- However, the dataset (~260k sentences) that need to be collected by hand is huge.

2. Solution

- Building a classifier model.
- Existing models are not presumed to classify datasets containing homoglyph sentences.

Restoration of Homoglyph Attack with BERT (4)

2. Refining English data – Process

- Training small dataset(~10k sentences) + Predicting the rest(~250k sentences)
- Model Training Process
 1. Train a custom BERT tokenizer using the whole dataset.
 2. Pretrain a BERT model as a Masked Language Model from scratch using the whole dataset.
 3. Finetune the pretrained BERT model as a text classifier using the hand-labeled dataset.

Restoration of Homoglyph Attack with BERT (4)

3. Refining English data – Training the BERT tokenizer

- Train BERT tokenizer as **WordPiece** Tokenizer
- The BitcoinAbuse.com dataset is presumed to have different vocabulary from existing BERT tokenizer.
- Training Arguments
 - vocab size: 30522
 - special tokens: ["[UNK]", "[CLS]", "[SEP]", "[PAD]", "[MASK]"]

Restoration of Homoglyph Attack with BERT (4)

4. Refining English data – Pretraining the BERT Masked Language Model

- BERT MLM is used for finetuning the text classifier.
- MLM Training Process
 1. Tokenizing the sentences.
 2. Masking the tokens randomly.
 3. Training a task to predict the masking tokens.
- Training Settings
 - no validation set (: The model will operate within the dataset.)
 - train epochs: 5
 - max sequence length: 512
 - train batch size per device: 16
 - Using two GeForce RTX 3090 GPUs

Restoration of Homoglyph Attack with BERT (4)

4. Refining English data – Pretraining the BERT Masked Language Model

- Training Result
 - Taking 3858 seconds (~ 1hr)
 - CrossEntropy loss: 4.8590

Restoration of Homoglyph Attack with BERT (4)

5. Refining English data – Finetuning the BERT Text Classifier

- Text Classifier = BERT MLM + one Linear Layer
- Binary Classification model (English, non-English)
- Training Setting
 - 13444 train + 1494 dev + 1660 test sentences
 - learning_rate=5.0e-6,
 - mini_batch_size=8,
 - max_epochs=20,
 - optimizer=torch.optim.AdamW
 - Using one GeForce RTX 3090 GPU

Restoration of Homoglyph Attack with BERT (4)

5. Refining English data – Finetuning the BERT Text Classifier

- Training Result
 - Best Model at epoch 6 (F1-score for development set: 0.996)
 - Taking 1hr 7min 11sec
 - Test Result:
 - F1-score (micro) 0.9898

Label	Precision	Recall	F1-score	Support
English	0.9887	0.9938	0.9913	972
Non-English	0.9912	0.9840	0.9876	688

Restoration of Homoglyph Attack with BERT (4)

6. Splitting ASCII-only / Homoglyph Set

- Splitting English dataset into ASCII-only / homoglyph set
 - ASCII-only set: 165906 examples
 - homoglyph set: 17791 examples
- The ASCII-only set will be used as MLM training set.
- The homoglyph characters in homoglyph set will be labeled manually and used as test set.