Software Impacts 13 (2022) 100310

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

BERT Probe: A python package for probing attention based robustness evaluation of BERT models

Shahrukh Khan *, Mahnoor Shahid, Navdeeppal Singh

Saarland University, Germany

ARTICLE INFO

Keywords:
Deep learning
BERT
Transformers
Adversarial machine learning

ABSTRACT

Transformer models based on attention-based architectures have been significantly successful in establishing state-of-the-art results in natural language processing (NLP). However, recent work about adversarial robustness of attention-based models show that their robustness is susceptible to adversarial inputs causing spurious outputs thereby raising questions about trustworthiness of such models. In this paper, we present BERT Probe which is a python-based package for evaluating robustness to attention attribution based on character-level and word-level evasion attacks and empirically quantifying potential vulnerabilities for sequence classification tasks. Additionally, BERT Probe also provides two out-of-the-box defenses against character-level attention attribution-based evasion attacks.

Code metadata

Current code version	v1.0
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2022-49
Permanent link to Reproducible Capsule	https://codeocean.com/capsule/6207048/tree/v1
Legal Code License	MIT License
Code versioning system used	Git
Software code languages, tools, and services used	Python
Compilation requirements, operating environments & dependencies	Transformers, PyTorch, and TextAttack
If available Link to developer documentation/manual	
Support email for questions	shkh00001@stud.uni-saarland.de

1. Introduction

Natural language processing has been able to achieve great progress in overcoming human-level baselines in a wide array of language tasks which was made possible mainly due to attention-based neural architectures [1]. However, recent work [2–4] indicate that attention-based neural networks can be vulnerable to evasion based adversarial attacks exploiting the attention attributions in a white-box setting. Moreover, this also poses a new challenge against white box adversaries who present a real threat to the model's robustness and performance if present. This threat model can potentially be harmful as it may enhance the efficiency of the adversary as attention attributions-based attacks require less perturbation budget and can flip the label of benign prediction by causing significant shifts in the posteriors.

In this paper, we present BERT Probe which provides researchers with a comprehensive package for evaluating attention-based vulnerabilities based on character-level and word-level evasion attacks. In both of the attacks, first attention attributions are computed, and then based on attention scores for character-level attacks character-level perturbations are performed by inserting, deleting, or replacing characters starting from the token with the highest attention. For word-level attacks, tokens with high attention scores are replaced with contextual synonyms, token candidates generated using masked language modeling (MLM). Moreover, we also provide two novel defenses against character-level defenses in BERT Probe, first implicit defense re-trains the model by extending the number of classes and adding abstain class and maps all synthetically generated adversarial examples

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: shkh00001@stud.uni-saarland.de (S. Khan), mash00001@stud.uni-saarland.de (M. Shahid), s8nlsing@stud.uni-saarland.de (N. Singh).

https://doi.org/10.1016/j.simpa.2022.100310
Received 13 April 2022; Received in revised form 28 April 2022; Accepted 4 May 2022

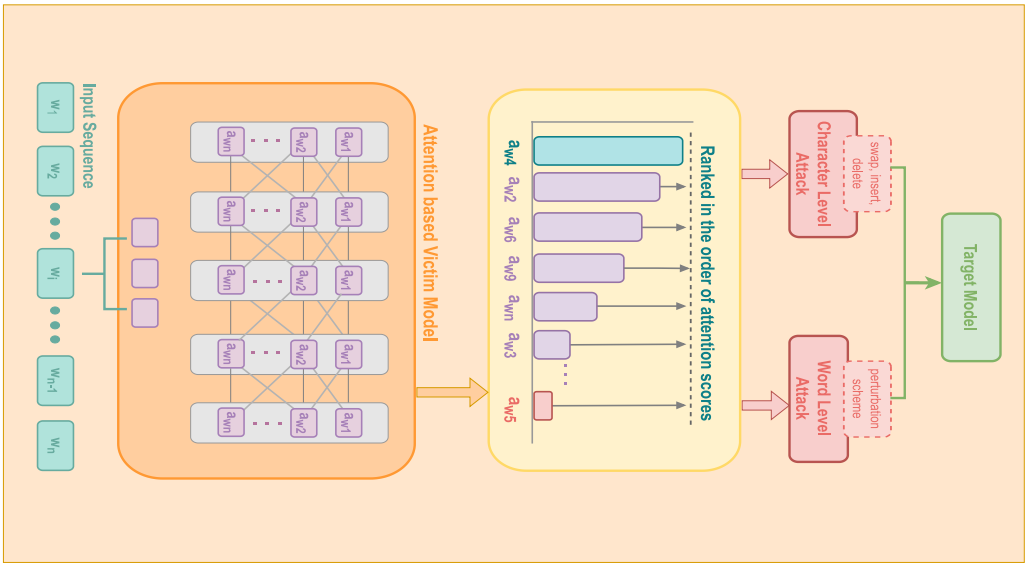

Fig. 1. BERT Probe attack schemes.

to abstain class and re-trains the model. Whereas, the other explicit defense performs adversarial pre-processing of each text sequence prior to inference to eliminate adversarial signals hence resulting in the transformation of adversarial input to benign. For attacks, we obtain the adversarial examples by attacking the model with character-level and word-level attacks using the test data, and the examples which successfully flip the prediction labels are classified as adversarial examples with the specific perturbation budget, whereas, for defenses, we obtain adversarial examples by character-level attacks on train data. Bert Probe allows users to specify the perturbation budget as a configurable option.

2. Functionalities and key features

BERT Probe offers both word-level and character-level attacks to evaluate the robustness of attention-based models. Moreover, it also provides out-of-the-box two novel character-level defenses to evaluate the efficacy of existing and novel character-level attacks.

2.1. Attacks

Fig. 1 illustrates the schematic workflow of both word-level and character-level attacks possible using the BERT probe. The perturbation schemes in the word-level attack which are white-box extensions of the *BERT-based Adversarial Examples for Text Classification*[5] include a masked language modeling (MLM) based relaxed attack which replaces each word starting from the token with the highest attention score with candidate tokens generated candidate tokens using MLM. Whereas, the constrained variant of the same word-level attack enforces Parts-of-Speech (POS) constraints on generated candidate tokens from MLM. Whereas for character-level attack BERT probe offers a white-box variant of *Combating Adversarial Misspellings with Robust Word Recognition*[6].

2.2. Defenses

BERT Probe offers two variants of character-level defenses namely explicit and implicit character-level defenses. Explicit defense first projects the train dataset vocabulary to a latent space using character-level word-embeddings extracted from the Siamese BERT [7] and then at inference time each input sequence is tokenized and projected to the same latent space as the vocabulary, words within vocabulary result in cosine similarity score of 1.0 whereas the out-of-vocabulary (OOV) words in the input are replaced with their closest neighbor in the latent space. Fig. 2 demonstrates the workflow of explicit defense.

For implicit character-level defense we create a new (untrained) classifier C′ from the original classifier C by extending the number of classes it is able to predict by one. The new class is labeled 'ABSTAIN', representing that the classifier abstains from making a prediction. Using C we create the adversarial examples. We mix these with the normal examples from the dataset (of C), where the adversarial examples have the abstain label, to create a new dataset. We then simply train on this dataset.

3. Impact overview

Adversarial examples and hence the robustness of the corresponding neural network architectures in NLP are less explored compared to their counterparts in computer vision. Low robustness is undesirable in production settings, thus evaluating the robustness of NLP models is crucial for their deployment. The attacks provided in BERT Probe lets users evaluate the robustness of such NLP models, find corner cases, when they do not work as intended, and understand their behavior better. Users can choose to apply the defense solutions available in BERT Probe if any such vulnerabilities are identified using the attacks. As such BERT Probe can be used to build more robust models for a variety of NLP classification tasks. Furthermore, with the extensibility provided by the programming language used, namely, Python, BERT Probe is ideal for creating attacks and defenses in future work for evaluating and increasing the robustness of NLP models. To facilitate this further, we intend to make BERT Probe available in PyPI, making the package easily accessible and usable.

As an example, [8] used BERT Probe to evaluate the robustness of German language hate-speech attention based classifiers, showing the ease with which such models can be tricked. Furthermore, [8] used the defenses available in BERT Probe as a solution against the attacks.

4. Conclusion and future work

BERT Probe is a comprehensive toolkit for researchers and users to evaluate the robustness of attention models to white-box attention attributions based on adversarial attacks, whilst also offering novel character-level defenses to benchmark the effectiveness of white-box character-level attacks or employ them against character-level attacks. Moreover, we also aim to expand BERT Probe to cover word-level defenses. Lastly, we also intend to make the BERT Probe package accessible through PyPI in order to ensure flexible distribution and installation.

Fig. 2. Explicit character-level defense.

CRediT authorship contribution statement

Shahrukh Khan: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Software, Validation. **Mahnoor Shahid:** Visualization, Investigation, Reviewing and editing. **Navdeeppal Singh:** Methodology, Software, Data curation, Writing – original draft, Software, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, 2017.

[2] Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, Cho-Jui Hsieh, On the robustness of self-attentive models, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1520–1529.

[3] Siddhant Garg, Goutham Ramakrishnan, BAE: BERT-based adversarial examples for text classification, 2020, CoRR, arXiv:2004.01970, URL https://arxiv.org/abs/2004.01970.

[4] Danish Pruthi, Bhuwan Dhingra, Zachary C. Lipton, Combating adversarial misspellings with robust word recognition, 2019, CoRR, arXiv:1905.11268, URL http://arxiv.org/abs/1905.11268.

[5] Siddhant Garg, Goutham Ramakrishnan, BAE: BERT-based adversarial examples for text classification, 2020, CoRR, arXiv:2004.01970, URL https://arxiv.org/abs/2004.01970.

[6] Danish Pruthi, Bhuwan Dhingra, Zachary C. Lipton, Combating adversarial misspellings with robust word recognition, 2019, CoRR, arXiv:1905.11268, URL http://arxiv.org/abs/1905.11268.

[7] Nils Reimers, Iryna Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[8] Shahrukh Khan, Mahnoor Shahid, Navdeeppal Singh, White-box attacks on hate-speech BERT classifiers in german with explicit and implicit character level defense, in: BOHR International Journal of Intelligent Instrumentation and Computing, 2022, BOHR Publishers, 2022.