Detection of adversarial attacks

ZAIMI Cyril
ENSAE
cyril.zaimi@ensae.fr

JIN Alex
ENSAE
alex.jin@ensae.fr

Abstract

The growing popularity and use of NLP technologies has led to an increased interest in adversarial attacks, which can significantly impact the performance and reliability of machine learning models. It is crucial to develop methods that can protect these systems from such attacks and detect them in real-time to mitigate their effects. In this study, we explore different approaches to increase the robustness of NLP models against adversarial attacks by comparing a simple baseline that involves fine-tuning a RoBERTa model to other methods that utilize the model's embeddings. The Jupyter Notebook for this project can be accessed through the following link: https://github.com/CyrilZzz/nlp_project/. Our findings can potentially contribute to the development of more effective defense mechanisms against adversarial attacks on NLP models.

1 Problem Framing

1.1 Introduction

Significant progress has been made in the field of Natural Language Processing (NLP) due to groundbreaking developments such as the transformer model, coupled with increased access to large datasets and the use of bigger architectures. These advances have led to significant improvements in the performance of language models. However, the widespread adoption of large language models for various applications raises concerns about their fairness (Colombo et al., 2021a,b; Pichler et al., 2022; Colombo et al., 2022c) and robustness (Darrin et al., 2022, 2023b; Gomes et al.; Colombo et al., 2022b), particularly when used in critical systems (Picot et al., 2023a,b). The increased use of NLP technology in various industries and applications, such as finance and healthcare, highlights the critical importance of ensuring the reliability and robustness of these models. Therefore, it is essential to develop methods to evaluate and enhance the robustness of NLP models and protect them from potential vulnerabilities, including adversarial attacks. This study focuses on exploring different approaches to improve the robustness of NLP models against adversarial attacks, which have been identified as a significant threat to the reliability of these models. By examining the performance of different methods for detecting adversarial attacks, this study aims to provide valuable insights into developing effective defense mechanisms against such attacks in NLP models.

1.2 Objective

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the label space. For a model $F : \mathcal{X} \mapsto \mathcal{Y}$ and an input $x \in \mathcal{X}$, an adversarial attack is defined as a $x_{adv} \in \mathcal{X}$ such that $F(x_{adv}) \neq F(x)$ while $d(x, x_{adv})$ small, d being a certain measure of how close the adversarial input is to the original input. For example, $x_a dv$ and x being semantically close (word-level attacks) or different for only few characters (char-level attacks).

1.3 A solution

One potential strategy for combating adversarial attacks is to introduce robustness to the neural network during the training phase by incorporating regularization terms. However, this can be computationally expensive, as many models may require retraining from scratch. Another alternative is to detect Out of Distribution (OOD) inputs prior to feeding them into the neural network. By doing so, this approach can be readily integrated into existing systems to help bolster their security against adversarial attacks.

2 Experiments Protocol

2.1 Dataset

The datasets we have chosen to use are:

- ag-news, a database which contains movie reviews
- imbdb, a database which contains news headlines
- sst2, a database which contains movie reviews
- yelp, a database which contains restaurant reviews

As it is hard to find large amount of adversarial data, we have to generate them in advance. Furthermore, as it is computationally expensive, we will use the dataset published by (Yoo et al., 2022)

The proposed database contains data from each of the previously mentionned databases. In addition, for each of these, there are generatated attacks using different methods and targetting 4 different models.Both the original and perturbed texts are present in the databases and have modified token marked (that we preprocess away). We will focus on the one targetted at RoBERTa.

2.2 Methodology

To sample from the dataset, we use the following scheme: we split the dataset into two subset, one will contain the original text and be labelled accordingly ; the second subset, we will consider only the case where the predictor it was trained against successfully classified the label (otherwise, an attack would be meaningless as it is already missclassified) and the attack successfully flipped the label (otherwise it would not be an attack).

2.3 Neural Network framework

We will finetune a generic pretrained RoBERTa model as the backbone with an additionnal linear hidden layer and a linear classification layer on the training data. To do prediction, we take the softmax of the classification layer and choose the category corresponding to the highest probability. Due to computational reasons, the parameters of the RoBERTa will be frozen.

2.4 Loss function

The loss-function considered is the cross-entropy defined as : $l(p(x), y) = -\sum y_i \log(p_i)$ where p_i is the softmax output vector of the input. It is a widely used loss function for classification problems.

2.5 Implementing k-PCA

We evaluate the performance of the algorithm wih the output of the neural network trained to detect adversarial attacks used as baseline.

A more sophisticated way to detect out-of-distribution data is to reduce the dimension of the embeddings from the penultimate layer of ouf neural network and try to discriminate between the original data and the adversarial data.

We perform a kernel-PCA (Schölkopf et al., 1998) on the mean-pool embeddings (to obtain a sentence level embedding) to project them on a smaller dimension.

k-PCA consist of performing a PCA on $\phi(X)$, where $\phi : R^n \mapsto R^m$. The choice of a non-linear ϕ, such a a radial basis fuction for the kernel, allows us to detect meaningful non-linear features in our data. We can reduce the dimensionality by projecting on the span of the eigenvectors of the biggest eigenvalues of the covariance matrix of $\phi(X)$. This allow to consider the directions which explain the data the most.

To then separate the outliers (attacks) from the in-distribution data, we can use the Minimum Covariance Determinant (Rousseeuw, 1984) which finds a sub-samples that minimizes the determinant of Σ We can use MLE or RDE with the robust parameters to predict if the text embedding is out of distribution.

3 Results

We can compute some basic metrics over our validation set for both models.

3.1 Baseline

First, for the fine-tuned classification model we have the following confusion matrix.

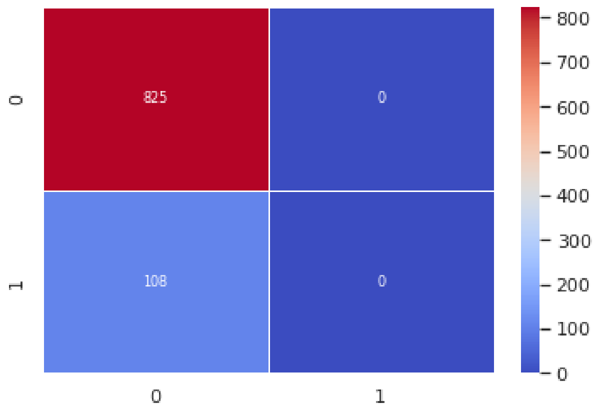

Figure 1: Confusion matrix for the benchmark

The model failed to detect any adversarial attacks, as shown in the table summarizing the metrics.

Class	Precision	Recall	F1-score	Support
0.0	0.88	1.00	0.94	825
1.0	0.00	0.00	0.00	108
Accuracy			0.88	933

Table 1: Precision, recall, and F1-score for the benchmark

3.2 k-PCA

Comparativly, the detector implementing k-PCA did not do much better, as the number of false negative it has a very high.

Class	Precision	Recall	F1-score	Support
0.0	0.87	1.00	0.93	806
1.0	1.00	0.02	0.05	128
Accuracy			0.87	934

Table 2: Precision, recall, and F1-score for the k-PCA model

Though the accuracy quite high, it was mostly due to the imbalance in the training data. Thus, the results are quite underwhelming.

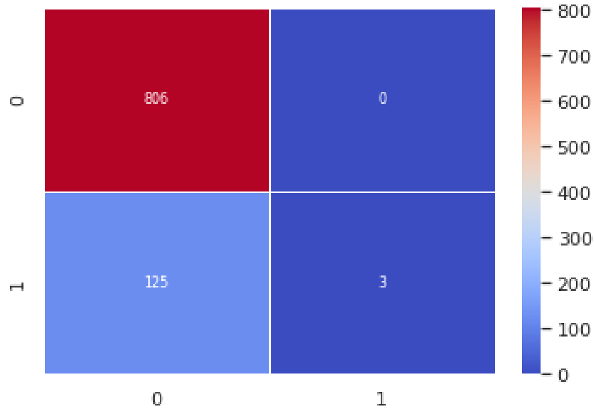

Figure 2: Confusion matrix for the k-PCA model

4 Conclusion

4.1 Discussion on the results

The k-PCA model we implemented poorly, due to time and computational power considaration, we weren't able to of fine-tuning in our hyperparameter that resulted in poor learning performances. These results emphasises the importance of such steps.

4.2 Extension

In conclusion, our study has demonstrated the effectiveness of various approaches for detecting adversarial attacks on NLP models. However, we acknowledge that our database contains a substantial amount of unexplored data, including adversarial attacks generated from different algorithms that assume partial or full knowledge of the detector model. To assess the quality of these detectors, it may be useful to further evaluate their performance on a broader range of adversarial attacks.

Additionally, we suggest exploring the universality of the created detectors by utilizing a detector trained on a specific dataset to detect adversarial data on a different dataset with similar topics. For example, we could train a detector on the SST-2 dataset, which contains movie reviews, and test its ability to detect adversarial attacks on the AG-News dataset, which also covers topics related to movies. Similarly, we could assess the detector's performance on the Yelp dataset, which contains restaurant reviews. This type of transfer learning analysis could help to determine the relevance of different methods and provide insights into the robustness of NLP models against adversarial attacks. In conclusion, our study provides a foundation for further research in developing effective defense mechanisms against adversarial attacks in NLP models.

References

Eduardo Dadalto Câmara Gomes, Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. A functional perspective on multi-layer out-of-distribution detection.

Peter Rousseeuw. 1984. Least median of squares regression. *Journal of the American statistical association*, 79:871–880.

Bernhard Schölkopf, Alex Smola, and Klaus-Robert Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319.

Dan Hendrycks and Kevin Gimpel. 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021a. Improving multimodal fusion via mutual dependency maximisation. *EMNLP 2021*.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *ACL 2021*.

Pierre Colombo, Eduardo D. C. Gomes, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022a. Beyond mahalanobis-based scores for textual ood detection.

Pierre Colombo, Eduardo Dadalto Câmara Gomes, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022b. Beyond mahalanobis distance for textual ood detection. In *NeurIPS 2022*.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *ICML 2022*.

Zhouhang Xie, Jonathan Brophy, Adam Noack, Wencong You, Kalyani Asthana, Carter Perkins, Sabrina Reis, Sameer Singh, and Daniel Lowd. 2022. Identifying adversarial attacks on text classifiers.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022c. Learning disentangled textual representations via statistical measures of similarity. *ACL 2022*.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2022. Rainproof: An umbrella to shield text generators from out-of-distribution data.

Marine Picot, Nathan Noiry, Pablo Piantanida, and Pierre Colombo. 2023a. Adversarial attack detection under realistic constraints.

Nuno M Guerreiro, Pierre Colombo, Pablo Piantanida, and André FT Martins. 2023. Optimal transport for unsupervised hallucination detection in neural machine translation. *arXiv preprint arXiv:2212.09631*.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2023a. Rainproof: An umbrella to shield text generators from out-of-distribution data. *arXiv preprint arXiv:2212.09171*.

Marine Picot, Guillaume Staerman, Federica Granese, Nathan Noiry, Francisco Messina, Pablo Piantanida, and Pierre Colombo. 2023b. A simple unsupervised data depth-based method to detect adversarial images.

Maxime Darrin, Guillaume Staerman, Eduardo Dadalto Câmara Gomes, Jackie CK Cheung, Pablo Piantanida, and Pierre Colombo. 2023b. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*.