Mahalanobis and max-softmax : but why? A comprehensive study of the benchmark scores in Adversarial Attacks Detection

Raphael Thabut
ENSAE Paris
raphael.thabut@ensae.fr

Eric Vong
ENSAE Paris
eric.vong@ensae.fr

Abstract

Transformers (Vaswani et al., 2017) and other Deep Learning architectures have gained a lot of traction lately, as we have seen with the release of Chat-GPT3 (Brown et al., 2020). Although highly performant, those black-box models are questionned on their robustness, which will condition their use on sensible tasks. With their democratization, adversarial attacks have become a growing concern.
The goal of this article is to study popular Adversarial Attack detection scores, mainly the max-softmax (Hendrycks and Gimpel, 2018), and a Mahalanobis distance score (Yoo et al., 2022), we will attempt to measure both their performances and limitations. To this end, we introduce two scores, **FtS** (first-to-second) and **Euclidian**, the first is based on the softmax output of the classifier, while the second uses its penultimate layer's output. Those scores will respectively attempt to challenge the max-softmax and the Mahalanobis scores.
The code leading to our results is available on our GitHub [1]

1 Introduction

In Natural Language Processing, an adversarial attacks is a small modification of a text which leads the model into making an incorrect classification. As an example, an attacker could modify a spam email into classifying it as a legitimate email.
Possibilites of generating attacks include synonym substitution, character-level modification, or grammatical perturbations (Pruthi et al., 2019). These attacks can be very challenging to defend against, as language is inherently complex and contains many nuances that can be difficult to capture in machine learning models.
This is especially true since the models tries to generalize from training data, as such, possibilities of out of distribution (OOD) inputs can be problematic for the robustness of the model (Hendrycks and Gimpel, 2018). In fact, the training and test data rarely characterize the entire distribution (Fang et al., 2022).
As such, research has been conducted to counteract the attacks. Two main methods have emerged: *detection* and *defense*; The former aims to discriminate the input whereas the latter aims to correctly predict the output. If recents improvement have been made in attacks defense, (Zhou et al., 2021; Keller et al., 2021; Jones et al., 2020; Jin Yong Yoo, 2018), research on detection techniques is only starting to gain momentum. (Yoo et al., 2022; Picot et al., 2023a; Colombo et al., 2022; Picot et al., 2023b).

1.1 Our Work

We aim to study the standard state-of-the-art scores used for detection, this includes the Mahalanobis-based score (Kimin Lee, 2018) and other scores relying on the softmax distribution output of our transformers classifier : max-softmax, Kullback-Leibler divergence (Darrin et al., 2023), and Wasserstein distance.
We will compare their results and attempt to measure their performance by proposing close but different metrics:

Mahanalobis-based score We aim to compare the Mahanalobis-based score, which applies Mahanalobis distance on the last layer embedding of the transformers, with euclidean distance on the same embedding. The idea being that Mahalanobis distance is a modification of the euclidian distance which takes into account correlation between variables. The objective here is to challenge the covariance matrix estimation.
Indeed, computing Mahanalobis distance involves, in this case, estimating an ill-conditionned covariance matrix and inverting it, which can lead

[1] https://github.com/rthabut/nlp_adversarial_attacks

to stability issues. The problem is tackled by using robust estimators : Minimum Covariance Determinant (Driessen, 1999), Oracle Approximating Shrinkage (Yilun Chen, 2009), Ledoit-Wolf Shrinkage Estimator (Ledoit, 2004). However, to improve stability, those estimators may include significant bias in the estimation, as we can see by observing the Oracle Approximating shrinkage coefficient, which can reach values greater than 50%. One can wonder if, with such potential issues, it is still relevant to apply Mahalanobis distance. As euclidean distance corresponds to Mahalanobis distance with the covariance matrix being the identity, comparing the two should give us insights on how successful the inverse covariance matrix estimation is. We expect Euclidean distance to be not as efficient as Mahanalobis distance, but wish to quantify the performance gained by estimating the covariance matrix.

Softmax-based scores Softmax based scores are retrieved from the output of the last embedding layer and as such correspond to a probability distribution. One of the main score used for attack detection is the max-softmax, which selects the maximum probability on the softmax distribution.
We decided to add a small switch to it by substracting the second maximum probability of the softmax distribution to the score, the idea being that a text is more likely to be a (successful) attack if the two most probable classes have close probability since one just trespassed the other. We expect this new score to be slightly more efficient than the max-softmax.

2 Experiments & Protocol

We chose attacks on the AG-News database, a database on which we can perform topic selection between 4 topics, therefore scoring yields non trivial information. Indeed, in the case of binary classification, the max-softmax score uses all the information contained in the softmax probabilities, it only becomes interesting to try to compare it with other softmax-based scores for a number of classes $k > 2$.

2.1 The model

To characterize the result of an attack, we used a pre-trained model of BERT (Jacob Devlin, 2019) fined-tune on the AG-News dataset to perform topic selection.

2.2 The Data

2.2.1 Attacks

Loading the Attacks We retrieved 4 different attacks datasets which are available on the Github[2] used in (Kimin Lee, 2018).
Those datasets, generated with the Python library *TextAttack*, use four different attacking methods: TF-adjusted (Morris et al., 2020), Probability Weighted Word Saliency (Ren et al., 2019), Textfooler (Jin et al., 2020), and BAE (Garg and Ramakrishnan, 2020), on a dataset of 7600 samples of news article.

Building the Dataset We concatenated those datasets together, and retrieved the successfull attacks and original texts. This gives us a reasonnably balanced database containing 11402 attacks and 7600 normal inputs.

2.2.2 To compute Mahalanobis score

To use Mahalanobis score and Euclidian score, we make the assumption that the data in the penultimate layer follows a multivariate gaussian distribution: we model the class conditional probability $p_{\mu,\Sigma}(z|y=k) \sim \mathcal{N}(\mu_k, \Sigma_k)$ where y is the indicator function of whether the text is an attack.
In order to estimate the mean and the covariance matrix of that distribution, we used as a training set, a corpus of 120.000 texts proposed by AG-News as in (Kimin Lee, 2018).

2.3 Scores Used

Please note that the score taken will be the opposite of every score proposed presented below. For the sake of conciseness we will define the adapted metrics here without the minus sign.

2.3.1 Softmax-based scores

To classify, the BERT model uses for its last layer a softmax activation function, which outputs a probability distribution.
We will use this distribution output to detect attacks, which we call $s = (s_1, \ldots, s_k)$.
We define the following scores:

- **max-softmax** : The maximum probability of the softmax distribution as proposed in (Hendrycks and Gimpel, 2018):
  $max - softmax = \max(softmax)$

[2] https://github.com/bangawayoo/adversarial-examples-in-text-classification

- **KL** : The Kullback-Leibler divergence between the softmax s and the uniform distribution $U = \{1, \dots, k\}$:
$$D_{KL}\left(softmax, U\right) = \sum_{i=1}^{k} s_i \log k s_i$$

- **Was**: The Wasserstein distance between the softmax and the uniform distribution : $W\left(softmax, U\right) = \inf_{\pi \in \Gamma(softmax, U)} \int_{R^2} |u - v| \, \mathrm{d}\pi\left(u, v\right)$, where $\Gamma(softmax, U)$ is the set of distributions whose marginals are $softmax$ (resp. U) on the first (resp. second) factor.

- **Our contribution** : First-to-Second max (FtS): $FtS = softmax_{(k)} - softmax_{(k-1)}$, with $softmax_{(i)}$ statistic of order i

2.3.2 Scores using hidden layers

For the following section, x is the data observed at the penultimate layer of the network.

Mahalanobis Score : We define the score as $D_M(x) = (x - \mu)^T \Sigma^{-1} (x - \mu)$, with μ and Σ respectively the mean vector and covariance matrix estimators computed on the training dataset. To avoid ill-conditionning, we use the Oracle Approximating Shrinkage estimator to estimate the covariance matrix.
Preprocessing - We perform standardization on the data, both on the training set before computing the covariance matrix and on the test set. Then, we apply kernel-PCA with the Radial Basis Function Kernel (rbf) to reduce the dimension (which is $d = 768$) to $d' = 100$.

Euclidian Score We define the score as $D_E(x) = (x - \mu)^T (x - \mu)$, where x and μ are scaled using a min-max scaler before computation.

3 Results

3.1 Comparing all classifiers

Running the BERT model on the dataset, we retrieve the penultimate layer encoding the latent space and the softmax distribution, we then compute the various scores presented above.

- The Mahalanobis score slightly outperforms the euclidian distance, as such, taking into account the correlation yields for a better detection.

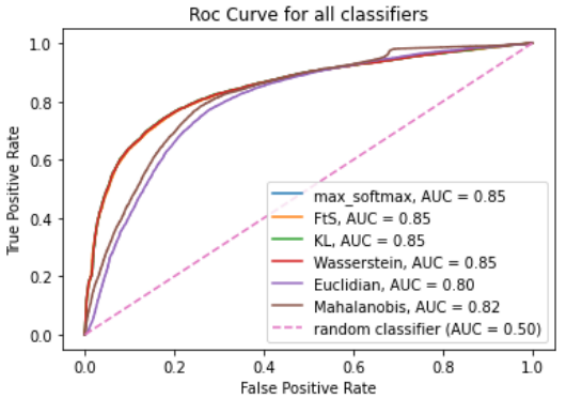

Figure 1: ROC Curve per score

- The Mahalanobis score outperforms the other scores for $FPR > 0.6$

- The softmax scores outperform the hidden layer based scores for $FPR < 0.2$

- Wasserstein, Kullback-Leibler, FtS and max-softmax scores perform very similarly and are almost indistinguishable

3.2 Softmax-based scores

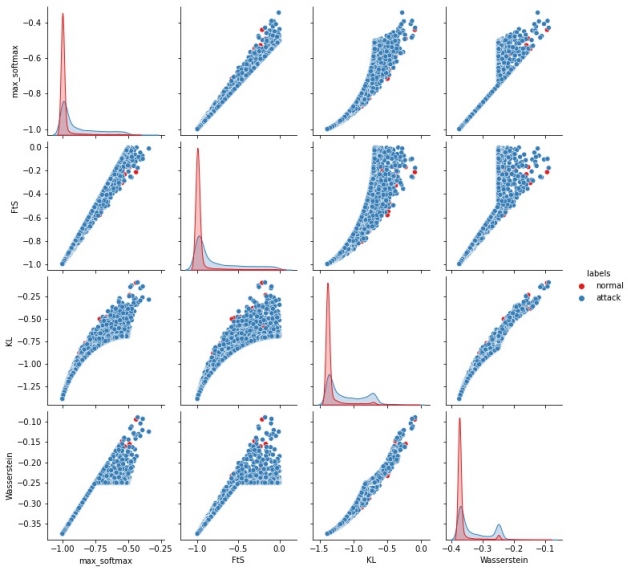

Figure 2: Pairplot of softmax-based scores

The last point is highlighted by the following pairplot, as one can observe the high amount of correlation between scores. Moreover, the confidence in the output is very high even for the successful attacked points, which questions the robustness of transformers architecture. Indeed, one would expect a lower confidence in the output for

attacked points.
The reason for such similar performances is due to the attacks' softmax distributions, which is known to produce highly overconfident predictions (Kimin Lee, 2018).

3.3 Discussing Gaussian Assumption

3.3.1 Mahalanobis distance vs Euclidian distance

As expected, the Mahalanobis distance outperforms the Euclidian distance by a thin margin. This shows that the statistical methods applied (OAS, kPCA) tackle the ill-conditionning issue enough to make the covariance estimation worth it. This solidifies the efficiency of the Mahalanobis score even in this context.

3.3.2 About Gaussian assumption

A strong hypothesis of the Mahalanobis score is the Gaussian distribution assumption. Unfortunately, Gaussian test are hard to compute in higher dimension due to the curse of dimensionality.
We use here a PCA decomposition to discuss the validity of the gaussian assumption: as a linear transformation of a gaussian vector remains a gaussian vector, if the data is gaussian, then its Principal Components should retain a gaussian distribution as well. Performing such a 2d-representation on the contour plot with covariance matrix Σ, we observe thick tails. We can therefore conclude that the data is most likely not gaussian.
Although this fact harms the validity of using distances relying on this assumption, e.g Euclidian or Mahalanobis, we observe that those methods remain efficient in the context of attack detection.

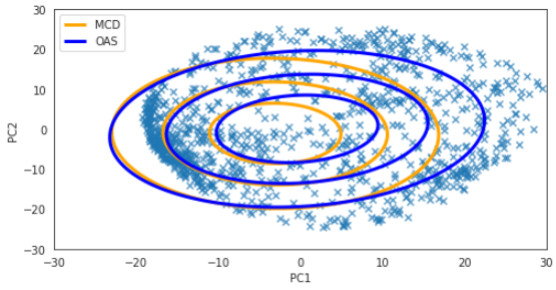

Figure 3: Contour of gaussian probability for attacked text

4 Discussion/Conclusion

Mahalanobis score With this study, we observed that class probability of attacks exhibits

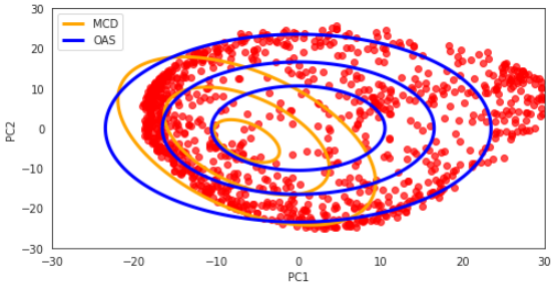

Figure 4: Contour of gaussian probability for original text

thick tails, contradicting the Gaussian assumption. But, even coupled with a context of ill-conditionned covariance matrix, we observed that the use of mahalanobis distance can constitute an efficient attack detection method as it slightly overperforms for high FPR ($FPR > 0.63$).
One natural extension would be to consider other hidden layers, as we restricted ourselves to the embedding generated by BERT's penultimate layer.
A notable shortcoming of such a score is that it relies on the embeddings. Having access to a softmax score seems more feasible than the entire hidden layer outputing the softmax distribution.

Softmax-based scores We also introduced a new softmax-based score FtS which perform very similarly to all softmax-based scores in our study. Unfortunately, the very confident predictions proposed by BERT did not allow us to compare thoroughly our score to the other benchmark softmax-based scores. It still allowed us to observe this shortcoming of the softmax function, which can prove to be problematic with regards to the robustness of models using the softmax function.

Concluding words As the development of multimodal (Garcia* et al., 2019; Colombo et al., 2021) generative models like GPT4 continues to rise, it becomes crucial to consider the potential risks associated with these technologies.
Improving and assessing robustness of transformers-based model is an open and evolving domain, and with the strong activity of this field, there is no doubt we will see, in the near future, new methods emerge for adversarial attacks which will tackle the limitations we observe today.

References

Peter J. Rousseeuw Katrien Van Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. In *TECHNOMETRICS*.

Wolf Ledoit. 2004. A well-conditioned estimator for large-dimensional covariance matrices. pages 365–411.

Yonina C. Eldar Alfred O. Hero III Yilun Chen, Ami Wiesel. 2009. Shrinkage algorithms for mmse covariance estimation. In *arXiv:0907.4698*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 6000–6010. Curran Associates.

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

Honglak Lee Jinwoo Shin Kimin Lee, Kibok Lee. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *arXiv preprint arXiv:1807.03888*.

Yanjun Qi Jin Yong Yoo. 2018. Towards improving adversarial training of nlp models. In *arXiv:2109.00544*.

Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805, 2018*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Alexandre Garcia*, Pierre Colombo*, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. *EMNLP 2019*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2752–2765, Online. Association for Computational Linguistics.

Yannik Keller, Jan Mackensen, and Steffen Eger. 2021. BERT-defense: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1616–1629, Online. Association for Computational Linguistics.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. Defense against synonym substitution-based adversarial attacks via Dirichlet neighborhood ensemble. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5482–5492, Online. Association for Computational Linguistics.

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

Pierre Colombo, Eduardo Dadalto Câmara Gomes, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Beyond mahalanobis distance for textual OOD detection. In *Advances in Neural Information Processing Systems*.

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

for Computational Linguistics: ACL 2022, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. 2022. Is out-of-distribution detection learnable? *arXiv preprint arXiv:2210.14707*.

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

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2023. 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.

(**?**)