A study of OOD detectors with Text Classifier

Sami Sabah

February 2023

Abstract

The effectiveness of machine learning methods relies on the assumption that both the training and testing data conform to the same distribution p_{xy}. However, once a model is developed, it can be utilized with any data that meets the requisite format. In practice, models often fail when presented with data from a distribution that differs from the one on which they were trained. Even more concerning, models can produce incorrect predictions with high confidence, without notifying users of the error. This represents a significant safety issue, particularly in sensitive fields such as medicine or finance, prompting the machine learning community to recognize the need for methods to identify changes in data distribution. The active area of research that tackles this issue is known as Out-of-Distribution (OOD) detection. In this article, we address this problem following the approach of [Col+22a]. Our study is divided into three parts: we first frame the problem, then benchmark OOD detectors, and finally, we investigate the effect of different aggregation methods by applying the same OOD detector with varying numbers of layers.

1 Introduction

Language models, such as those based on deep learning techniques, have shown impressive performance in various Natural Language Processing (NLP) tasks in recent years. However, concerns about fairness [CCP21; Col+22b; Pic+22] and reliability [Pic+23b; Pic+23a] have arisen as these models are increasingly applied to real-world applications. Fairness refers to the ethical and legal obligation to avoid bias and discrimination, and it is crucial to ensure that language models are free from these issues. On the other hand, reliability relates to the ability of language models to make accurate predictions, even when presented with input data that differs from the training data.

In particular, Out-of-Distribution (OOD) detection has emerged as a key issue in ensuring the reliability of language models [Gom+; DPC23; Dar+23]. The OOD problem arises when models are tested on data that comes from a different distribution than the one they were trained on. This can lead to unexpected and often incorrect predictions, which may have severe consequences, especially in sensitive domains such as healthcare or finance. Therefore, detecting OOD data and handling it appropriately is essential for safe and responsible deployment of language models.

In this context, there has been a growing interest in developing reliable language models that can OOD detection issues. In this paper, we aim to investigate the current state-of-the-art approaches to achieve OOD detection in language models and to identify the challenges and opportunities for future research.

1.1 Problem framing

In this paper, we focus on classifiers for textual data. Consider the corresponding setup : $\mathcal{X}$ the textual input space and $\mathcal{Y}$ the target space. We have N i.i.d samples $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$ that we assume follow the joint law p_{xy} and the marginals p_x, p_y. Using this datasat, we build a classifier $f_N : \mathcal{X} \to \mathcal{Y}$.

The objective of OOD detection is to construct a similarity function $s : \mathcal{X} \to \mathcal{R}_+$ that measures the proximity of any element of $\mathcal{X}$ with respect to the training in-distribution. Once we have a similarity function, according to it magnitude and having choosen a threshold γ we classify $\mathbf{x}$ as IN-DATA if $s(\mathbf{x}) > \gamma$ and OOD otherwise.

1.2 Performance evaluation

The OOD problem is a binary classification problem and classically, two threshold invariant quantities are used to evaluate the performance of a model. The first one is the area under the ROC curve $\gamma \to (TPR(\gamma), FPR(\gamma))$. The second one and that is more relevant in the case of unbalanced classification (which is usually the case in OOD detection) is the area under the Precision-Recall curve : $\gamma \to (Precision(\gamma), Recall(\gamma))$. In this article, the AUROC and the AUPRC are the two metrics that will be used.

2 Benchmark

Existing methods for OOD detection can be grouped in two line of research according to their position relative to the network. The first line of work can be assimilated to robust training techniques which consist in incorporating regularization terms. The second line of work, and our focus, is post-processing methods that can be applied on any pretrained model. Within this category, we can distinguish methods that require access to in-distribution data and those that doesn't.

2.1 Without ID data

There are two methods in this category :

- In his seminal work, [HG16] noticed that even though the prediction probability in isolation can be misleading, it is lower for OOD examples than for ID examples. Based on that, he constructs the following score : $s(x) = 1 - \max_{y \in \mathcal{Y}} p_{Y|X}(y|x)$ where $p_{Y|X}(.|x)$ is the soft-probability predicted for x.
- Energy-based score are defined as the score : $s(x) = Tlog[\sum_{y \in \mathcal{Y}} exp(\frac{g_y(x)}{T})]$ where $g_y(x)$ is the logit corresponding to the class y.

2.2 With ID data

Multilayer neural networks apply successive transformations to the data to construct meaningful internal representations of it at each layer. These internal representation lie in what we call the latent space. The idea of the OOD methods of this category is that these layers carry useful information to perform OOD detection. The two major bricks of these methods are : aggregation functions and data-depths. Aggregation functions refers to how the information of the layers is processed. Data-depths are nonparametric statistics that measure the centrality of any element of R^d w.r.t a probability distribution defined on a subset of R^d (introduced initially as an extension of the notion of median to the multivariate setting). OOD detection is done in the following frame :

1. The incoming data point $\mathbf{x}$ is processed by the neural network and results in a predicted target $\hat{y}$ and the internal representations $\{\Phi_1(\mathbf{x}), ..., \Phi_L(\mathbf{x})\}$ corresponding to the L layers
2. The latent representations of $\mathbf{x}$ are aggregated via an aggregation function F :
$$F_{agg}(\mathbf{x}) := F(\Phi_1(\mathbf{x}), ..., \Phi_L(\mathbf{x}))$$
3. A similarity score is computed using a data-depth D, between $F(\mathbf{x})$ and $F(S_{\hat{y}}^{train})$ the distribution of the F-aggregation of the training distribution samples with same predicted target as $\mathbf{x}$
$$s(\mathbf{x}) := D(F_{agg}(\mathbf{x}), F_{agg}(S_{\hat{y}}^{train}))$$

To define an OOD method in this setting, we need to define the aggregation function F and the data-depth D. Existing methods differs in these previous elements. The best results are produced by the two following methods :

- In [Pod+21], the aggregation function solely relies on the last layer of the network and the data-depth function [Sta+21] is the Mahalanobis distance D_M wich is given by the formula $D_M(x, p_X) = (x - E[X])^T \Sigma^{-1}(x - E[X])$ with Σ the covariance matrix of X.
- Leveraging the observation that going deeper in the network might improve OOD detection, [Col+22a] used as an aggregation the mean function

:

$$F_{agg} = \frac{1}{L}\sum_{l=1}^{L}\Phi_l(\mathbf{x})$$

For the data-depth, the recently introduced Intregrated Rank-Weighted Depth (IRW) was used for its attractive theroetical properties (namely to not suffer from the curse of dimensionnality). It approximation is :

$$\frac{1}{n_{proj}}\sum_{k=1}^{n_{proj}}\min\{\frac{1}{n}\sum_{i=1}^{n}I\{\langle u_k, x_i - x\rangle < 0\}, \frac{1}{n}\sum_{i=1}^{n}I\{\langle u_k, x_i - x\rangle > 0\}$$

This method achieved state-of-the-art results for textual OOD detection.

3 Varying the information depth

In [Col+22a] , the mean agregation function was proposed leveraging the observation that *all* layers carry useful information for OOD detection. Using the mean function means that they implicitly assumed that all the layers carry information of the same importance regardless of their position in the network. However, the intuition and previous work suggests something else : [HG16] used the soft-probabilites and [Pod+21] used only the last layer in their detection method. It seems that end layers carry more meaningful information than those at the beginning. In order to understand better how the fine-tuning of information depth could improve OOD detection, we tried running the same OOD method with different depths. We ran the following OOD method : with the mean-aggregation method and the Mahalanobis distance. Each time, we use k layers starting from the last one :

$$F_{agg} = \frac{1}{k+1}\sum_{l=1}^{L-k}\Phi_l(\mathbf{x})$$

We use BERT model (12 layers) fine-tuned on SST2 (IN-DS) and IMDB as OOD-dataset.

AUROC
0.771
0.784
0.792
0.792
0.780
0.773
0.752
0.737
0.735
0.745

AUROC
0.757
0.758

Table 1: AUROC from $k = 0$ (top) to $k = 12$ (bottom).

3.1 Results and discussion

The AUROC results of the different runs are presented in tab1. We see that there is a slight improvement of the AUROC when we incorporate supplementary layers until the fourth one supporting that end layers carry more useful information. We also observe a decrease of the performance of the OOD detector after the fourth layer. This last observation suggests that the optimal choice may be out of the couple *all layers/last layer* and that an intermediate depth could be the best one. However, we must keep in mind that these results are specific to the aggregation method and distance used. For this latter matter, these results can be understood as an evidence of the fact that the Mahalanobis distance is not well suited for incorporating additionnal information. This observation has already been done in [Col+22a] where they show that Integrated Rank-Weighted distance behave much better when all the layers are used. These consideration remind us that in OOD detection, a careful attention should be given to the specificities of each aggregation method, distance, benchmark and model as results in a case could not reproduce in others (happily or sadly). In this regard, experiments such as the one we did improves our understanding of the way information flows in network and carrying out experiments of the same style in othe settings would be useful to imporove OOD detection.

4 An interesting futur research direction

In the futur, we would like to engage new work inspired by the work of [Col+22a]. The mean agregation function was proposed leveraging the observation that *all* layers carry useful information for OOD detection. On the other hand, using the mean function assumes that all the layers carry information of the same value regardless of their position in the network. However, this is not what suggest the intuition and the previous work : [HG16] used the soft-probabilites and [Pod+21] used only the last layer in their detection method. Thus, combining the two observations that all layers carry useful information but that end layers carry more meaningful one, an interesting futur research direction would be a new aggregation function which is a weighted average of all the layers :

$$F_{agg} = \frac{1}{L}\sum_{l=1}^{L}\gamma_l \Phi_l(\mathbf{x})$$

with γ_l an increasing sequence.

Bibliographie

[HG16] Dan Hendrycks and Kevin Gimpel. “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks”. In: *CoRR* abs/1610.02136 (2016). arXiv: `1610.02136`. URL: `http://arxiv.org/abs/1610.02136`.

[CCP21] Pierre Colombo, Chloe Clavel, and Pablo Piantanida. “A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations”. In: *ACL 2021* (2021).

[Pod+21] Alexander Podolskiy et al. “Revisiting Mahalanobis Distance for Transformer-Based Out-of-Domain Detection”. In: *CoRR* abs/2101.03778 (2021). arXiv: `2101.03778`. URL: `https://arxiv.org/abs/2101.03778`.

[Sta+21] Guillaume Staerman et al. “A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions”. In: *arXiv e-prints* (2021), arXiv–2103.

[Col+22a] Pierre Colombo et al. “Beyond Mahalanobis Distance for Textual OOD Detection”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: `https://openreview.net/forum?id=ReB7CCByD6U`.

[Col+22b] Pierre Colombo et al. “Learning Disentangled Textual Representations via Statistical Measures of Similarity”. In: *ACL 2022* (2022).

[Pic+22] Georg Pichler et al. “A Differential Entropy Estimator for Training Neural Networks”. In: *ICML 2022*. 2022.

[DPC23] Maxime Darrin, Pablo Piantanida, and Pierre Colombo. “Rainproof: An Umbrella To Shield Text Generators From Out-Of-Distribution Data”. In: *arXiv preprint arXiv:2212.09171* (2023).

[Dar+23] Maxime Darrin et al. “Unsupervised Layer-wise Score Aggregation for Textual OOD Detection”. In: *arXiv preprint arXiv:2302.09852* (2023).

[Pic+23a] Marine Picot et al. “A Simple Unsupervised Data Depth-based Method to Detect Adversarial Images”. In: (2023).

[Pic+23b] Marine Picot et al. “Adversarial Attack Detection Under Realistic Constraints”. In: (2023).

[Gom+] Eduardo Dadalto Câmara Gomes et al. “A Functional Perspective on Multi-Layer Out-of-Distribution Detection”. In: ().