VI-OOD: A Unified Representation Learning Framework for Textual Out-of-distribution Detection

Li-Ming Zhan[1]**,Bo Liu**[1]**, Xiao-Ming Wu**[1†]
Department of Computing, The Hong Kong Polytechnic University, Hong Kong S.A.R.[1]
{lmzhan.zhan,bokelvin.liu}@connect.polyu.edu.hk
xiao-ming.wu@polyu.edu.hk

Abstract

Out-of-distribution (OOD) detection plays a crucial role in ensuring the safety and reliability of deep neural networks in various applications. While there has been a growing focus on OOD detection in visual data, the field of textual OOD detection has received less attention. Only a few attempts have been made to directly apply general OOD detection methods to natural language processing (NLP) tasks, without adequately considering the characteristics of textual data. In this paper, we delve into textual OOD detection with Transformers. We first identify a key problem prevalent in existing OOD detection methods: the biased representation learned through the maximization of the conditional likelihood $p(y|x)$ can potentially result in subpar performance. We then propose a novel variational inference framework for OOD detection (VI-OOD), which maximizes the likelihood of the joint distribution $p(x, y)$ instead of $p(y|x)$. VI-OOD is tailored for textual OOD detection by efficiently exploiting the representations of pre-trained Transformers. Through comprehensive experiments on various text classification tasks, VI-OOD demonstrates its effectiveness and wide applicability. Our code has been released at `https://github.com/liam0949/LLM-OOD`.

Keywords: Out-of-distribution detection, large language models, representation learning

1. Introduction

Large-scale deep neural networks (DNNs) such as CNNs and Transformers, have brought about a revolutionary impact on numerous complex real-world machine learning applications. Nevertheless, a notable drawback of DNNs remains their tendency to make *overconfident* decisions, rendering them less reliable for safety-critical applications like medical diagnosis (Ulmer et al., 2020) and self-driving cars (Filos et al., 2020). It has been noted that DNNs often assign elevated confidence scores to unfamiliar inputs, leading to potential erroneous predictions when confronted with anomalous out-of-distribution (OOD) data (Nguyen et al., 2015). To address this issue, there has been active research and investigation into OOD detection in recent years (Hendrycks et al., 2022; Yang et al., 2022).

Challenge of OOD detection. OOD detection aims at solving a K-class in-distribution (ID) classification task and a binary ID *vs.* OOD discrimination task simultaneously. A commonly assumed practical setting is OOD examples are unavailable during training, which presents the major challenge for OOD detection. The mainstream methods for OOD detection commonly follow a post-hoc scheme (Hendrycks and Gimpel, 2017), which first discriminatively trains an ID K-class classifier by maximizing the conditional likelihood of $p(y|x)$ and then derives some statistics from the trained model to predictive OOD confidence scores. However, since the binary ID *vs.* OOD discrimination task is not considered in the training process, the learned representations by K-class training may be biased to the ID classes. While some attempts have been made to address this challenge by incorporating surrogate OOD datasets during the training phase, such as those described in the works by Hendrycks et al. (2019) and Lee et al. (2018a), further endeavors are required to identify appropriate OOD datasets that demonstrate significant distributional shifts compared to the ID data.

Research on textual OOD detection. The majority of recent research efforts have concentrated on detecting OOD data in visual domains, with only a limited number of studies (Hendrycks et al., 2020; Podolskiy et al., 2021a; Zhou et al., 2021a) focusing on textual OOD detection. As far as our knowledge extends, current textual OOD detection methods typically utilize general OOD detection algorithms on representations generated by Transformers (Vaswani et al., 2017). However, these methods often fail to adequately account for the unique characteristics and nuances of textual data. Moreover, although the hierarchical contextual representations of pre-trained Transformers have demonstrated remarkable effectiveness in numerous NLP tasks (Sun et al., 2019; Ma et al., 2019; Mohebbi et al., 2021; Devlin et al., 2019; Liu et al., 2019a), their potential for textual OOD detection has not been fully harnessed.

Our proposal. To tackle the aforementioned issues, we propose a variational inference framework based on Transformers for textual OOD detection. Rather than solely focusing on maximizing the conditional distribution $p(y|x)$ of ID data,

[†] Corresponding author.

LREC-COLING 2024, pages 17371–17383

our approach involves optimizing the joint distribution $p(x, y)$, which is to maximize $p(y|x)$ and $p(x)$ simultaneously. The core idea revolves around modeling the distribution of the provided ID data, which allows us to harness valuable information that might not be directly relevant for ID classification but proves significant for outlier detection. To make the joint distribution $p(x, y)$ tractable, we resort to optimizing the evidence lower bound of $p(x, y)$ derived via amortized variational inference (AVI) (Kingma and Welling, 2014). Moreover, considering the unique characteristics of textual data, we modify the approximated posterior distribution in the framework of AVI, making the posterior conditioned on a dynamic combination of intermediate layer-wise hidden states of the Transformer. The Transformer backbone functions as a shared encoder for both the ID classification head and the decoder (generator) in the AVI framework (Fig. 2).

The contributions of this work include:

- Our proposed variational inference framework for OOD detection (VI-OOD) offers a novel and principled approach, providing a fresh perspective that is orthogonal to previous OOD detection methods (Hendrycks et al., 2020; Podolskiy et al., 2021a; Zhou et al., 2021a).
- Our instantiation of VI-OOD harnesses the rich contextual representations of pre-trained Transformers to learn more effective latent representations for text inputs. The improved representations can be readily used by various existing post-hoc OOD detection algorithms, consistently enhancing their performance in textual OOD detection.
- Our proposed method is evaluated using mainstream encoder-based and decoder-based Transformer architectures and comprehensive OOD text classification scenarios. It can offer advantages to widely utilized OOD detection algorithms, particularly for distance-based OOD detectors, such as the Mahalanobis Distance method (Lee et al., 2018b)

2. Pilot Study

2.1. Problem Statement and Motivation

Out-of-distribution (OOD) detection aims to accurately separate all class-dependent in-distribution (ID) examples as well as out-of-distribution (or anomalous) examples. Given the input space $\mathcal{X} \times \mathcal{Y}$ and an ID class label set $\mathcal{Y}_{\text{ID}} = \{y_j\}_{j=1}^{K} \subset \mathcal{Y}$, an ID training set $\mathcal{D}_{\text{ID}} = \{(x_i, y_i)\}_{i=1}^{N}$ is sampled from the distribution $p(x, y)$ of ID data where $y_i \in \mathcal{Y}_{\text{ID}}$. With $\mathcal{D}_{\text{ID}}$, an ID classifier $f_{\text{ID}} : \mathcal{X} \to \mathcal{Y}_{\text{ID}}$ is trained. During test time, since there may be a distribution shift between the training and test data in practical application scenarios (Szegedy et al., 2014; Morningstar et al., 2021), the ID classifier f_{ID} may encounter OOD samples ($y_i \notin \mathcal{Y}_{\text{ID}}$). Hence, an OOD confidence scoring function $f_{\text{OOD}} : \mathcal{X} \to \mathbb{R}$ is needed to perform ID *vs.* OOD binary classification. In this regard, OOD detection aims to solve both the K-class ID classification task and the binary outlier detection task. The ID classifier f_{ID} is commonly trained with a discriminative loss by maximizing the conditional log-likelihood of the training set:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{ID}}} \log p(y_i \mid x_i; f_{\text{ID}}, \theta), \quad (1)$$

where θ stands for all trainable parameters of f_{ID}.

The fundamental challenge of OOD detection is that at the training stage, real OOD examples are unavailable and thus cannot be effectively represented to provide necessary learning signals for the binary ID *vs.* OOD task. To address this issue, a few attempts have been made to introduce surrogate OOD datasets during training by using some datasets irrelevant to the ID data (Hendrycks et al., 2019; Lee et al., 2018a). However, it is difficult to select suitable "OOD" datasets to represent the huge space of real OOD data.

Post-hoc methods. The majority of existing OOD detection methods (Hendrycks and Gimpel, 2017; Hendrycks et al., 2019; Lee et al., 2018b; Liu et al., 2020; Hendrycks et al., 2022; Sun et al., 2021, 2022) follow a post-hoc paradigm and address the binary ID *vs.* OOD task in the inference stage. These methods propose different OOD confidence scoring functions with the trained ID classifier f_{ID}. Specifically, the parameters of the trained f_{ID} are frozen, and some statistics of specific layers of f_{ID} (usually the penultimate layer or the softmax layer) are often used as OOD confidence scores.

Motivation of this work. While post-hoc methods have shown promise, it is pointed out that the performance of f_{ID} on ID data is not a good indicator of its performance on OOD data (Hendrycks et al., 2020; Lee et al., 2018a). Specifically, the discriminative training of f_{ID} is often conducted with $p(y|z)$, where z is the latent representation obtained by passing an input x to a DNN encoder. Maximizing the conditional log-likelihood $\log p(y|z)$ is essentially maximizing the mutual information between the latent variable Z and the label variable Y, i.e., $\mathcal{I}(Z, Y)$ (Boudiaf et al., 2020). Naturally, the learned representation Z will be biased towards the ID classification task. Indeed, Kamoi and Kobayashi (2020) have demonstrated that in the Mahalanobis-distance-based OOD detection method, the principal components of ID data that are deemed least important for the ID classification task actually contain valuable information for the

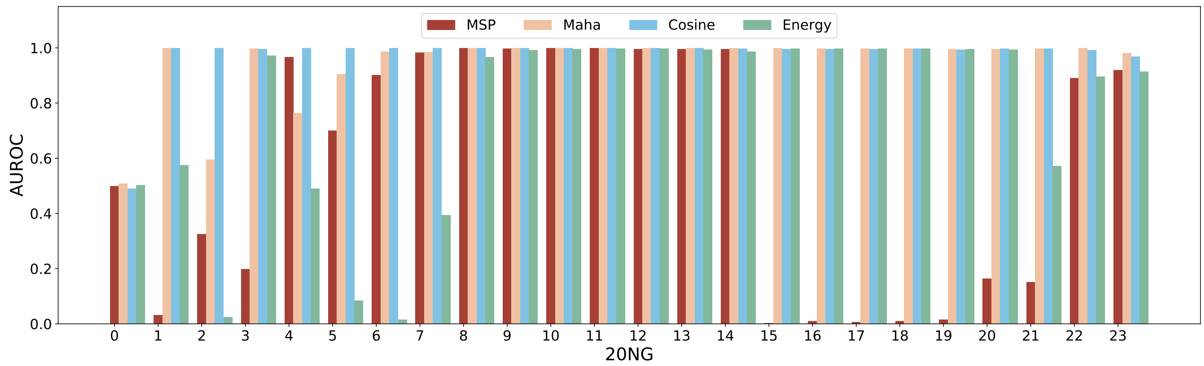

Figure 1: Investigation of OOD performance of Transformer's intermediate Hidden States: AUROC Results for 24 Layers of RoBERTa$_{\text{LARGE}}$. The figure illustrates the OOD performance evaluation across multiple layers of RoBERTa$_{\text{LARGE}}$. Higher values indicate better performance. The model undergoes fine-tuning on SST-2 and is assessed for OOD performance using the 20NG dataset. The four commonly used OOD scoring functions, namely MSP (red), Maha (light yellow), Cosine (blue), and Energy (green), are represented in the figure.

binary ID *vs.* OOD task. This information may be overlooked or discarded when training the ID classification function f_{ID} using the conditional likelihood $p(y|x)$. Furthermore, a recent study by Uppaal et al. (2023) highlights that relying solely on supervised training with ID data can lead to a degradation in the performance of OOD detection as the training progresses.

To address this issue, we propose to learn better latent representation Z for post-hoc methods by considering the distribution of ID data, i.e., maximizing the likelihoods $p(y|x)$ and $p(x)$ simultaneously[1], which is equivalent to modeling $p(x, y)$ – the joint distribution of ID data. To this end, we design a novel principled variational framework that will be elaborated in the next section.

2.2. A Closer Look at Textual OOD Detection with Transformers

In Figure 1, we study the impact of the intermediate hidden states of RoBERTa$_{\text{LARGE}}$ on textual OOD detection. Following (Hendrycks et al., 2020), we take the model trained on SST-2 as a case study. The model is trained solely with the discriminative loss. We conduct OOD detection by utilizing each hidden state of the trained model's 24 layers as representations of the input text data. Subsequently, we summarize the AUROC results obtained from four commonly used OOD detection algorithms. As the layer number increases from 0 to 23, the hidden layer is closer to the head of the model, i.e., layer 23 outputs the last hidden state.

Intermediate hidden states could help OOD detection. The results presented in Figure 1 clearly indicate that intermediate hidden states consistently outperform the final hidden states in terms of OOD performance, as observed across all four OOD detection methods. The best performance consistently occurs in the middle layers, particularly in the range of layers 9 to 13. This consistent performance is observed for all four OOD detection methods. On the other hand, as pointed out by Sun et al. (2019), intermediate hidden states of Transformers exhibit inferior performance compared to the final hidden state in ID classification tasks. Based on these observations, we make a key assumption: **intermediate hidden states contain redundant information for ID classification but crucial information for OOD detection.**

Furthermore, it is possible to address the disparities among various OOD detection methods. As depicted in Figure 1, the performance of intermediate layers (layers 9 to 14) is consistently comparable across the four OOD detection methods. For instance, the Maximum Softmax Probability (MSP) method demonstrates excellent results around layer 13, but its performance significantly deteriorates at the last layer, layer 23. These findings suggest that effectively harnessing the potential of hidden states in Transformers can alleviate the challenges associated with OOD detection.

3. Proposed Method

3.1. VI-OOD: A Variational Inference Framework for Out-of-distribution Detection

Our goal is to directly maximize the likelihood of the joint distribution $p(x, y)$ rather than $p(y|x)$. We assume that a latent variable Z is a stochastic encoding of the input sequence X. The log likelihood

[1] Note that $p(y|x) = \int_z p(y|z, x)p(z|x)\, dz$ and $p(x) = \int_z p(x|z)p(z)\, dz$.

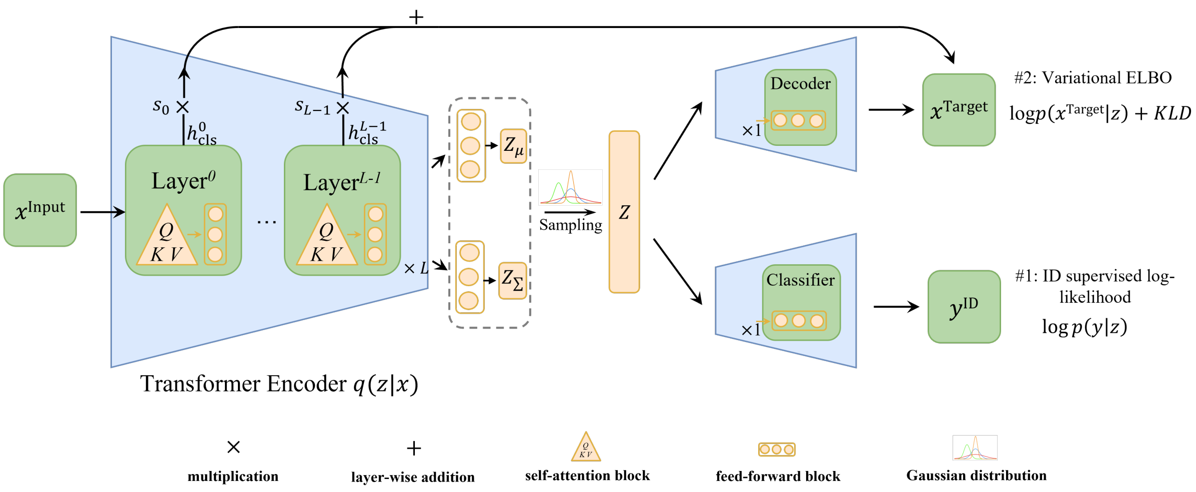

Figure 2: The architecture of our proposed framework. Our method employs an encoder-based transformer model as the backbone textual encoder. Hidden states of the [CLS] token are chosen to be textual representations. z is a latent variable conditioned on the textual representations. The in-distribution (ID) classification head $p(y|z)$ and decoder $p(x^{\text{target}}|z)$ both take z as the input. s is the hidden states combination factor and the merge representation x^{target} works as the target of the decoder.

of $p(x, y)$ can then be calculated by:

$$\begin{aligned}\log p(x,y) &= \log \int_z p(x,y,z)\,dz \\ &= \log \int_z p(y|z,x)p(x|z)p(z)\,dz \\ &= \log \int_z p(y|z)p(x|z)p(z)\,dz, \end{aligned} \tag{2}$$

where in the last equality we assume the Markov chain $X \leftrightarrow Z \leftrightarrow Y$, i.e., $p(y|z, x) = p(y|z)$. Since it is intractable to compute the integral in Eq. (2), we employ amortized variational inference (Kingma and Welling, 2014) to derive the lower bound of $\log p(x, y)$ as follows.

$$\log p(x,y) = \log \int_z p(y|z)p(x|z)p(z)\,dz$$

$$= \log \int_z p(y|z)p(x|z)p(z)\frac{q(z|x)}{q(z|x)}\,dz \tag{3}$$

$$= \log \mathbb{E}_{z\sim q(z|x)}\left[\frac{p(y|z)p(x|z)p(z)}{q(z|x)}\right] \tag{4}$$

$$\geq \mathbb{E}_{z\sim q(z|x)}\left[\log \frac{p(y|z)p(x|z)p(z)}{q(z|x)}\right], \tag{5}$$

where $q(z|x)$ in Eq. (3) is the amortized variational approximator of the true posterior $p(z|x)$, and Jensen's inequality is applied in Eq. (5). The last quantity in Eq. (5) is the evidence lower bound of $\log p(x, y)$, which can be rewritten as:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\mathbb{E}_z\left[\log p(y|z)\right]}_{\text{Target \#1: ID supervised training}} + \underbrace{\mathbb{E}_z\left[\log p(x|z)\right] - D_{\text{KL}}(q(z|x)||p(z))}_{\text{Target \#2: Unsupervised variational training}}, \tag{6}$$

where the first term is the ID supervised training objective, and the second and third terms correspond to the unsupervised learning objective for an amortized variational Bayesian autoencoder.

3.2. Transformer-based Textual OOD Detection with VI-OOD

Our proposed VI-OOD framework is a general probabilistic approach for learning data representations, which can be applied to various types of data, including image, textual, audio, and video. However, in this work, we focus on textual data. In the following, we outline the instantiation of VI-OOD for textual OOD detection, which involves designing the encoder (posterior approximator) $q(z|x)$, the decoder (reconstructor) $p(x|z)$, and the discriminator $p(y|z)$, as depicted in Figure 2.

Encoder for learning textual representations. Encoder-based Transformers have become a prevailing standard in learning contextual representations of text due to their excellent performance in numerous NLP tasks. Hence, the transformer architecture is a natural choice for the encoder $q(z|x)$. In this paper, we utilize models from the BERT family (Devlin et al., 2019). Given an input x, which is a sequence of tokens with a length of N, denoted as $[x_0, \cdots, x_{N-1}]$, BERT adds a special token [CLS] at the start of the input sequence, i.e., [CLS, $x_0, \cdots, x_{n-1}$]. The inclusion of the [CLS] token is intended for classification tasks. Unless otherwise specified, we use the hidden states of the [CLS] token as the textual representations. The input sequence x is passed through each layer of BERT, resulting in a series of intermediate hidden states at the [CLS] position, denoted as $h_{\text{CLS}} = [h^0_{\text{CLS}}, \cdots, h^{L-1}_{\text{CLS}}]$, where L is the total number of layers. As shown in Figure 2, we instantiate the encoder $q(z|x)$ and the prior $p(z)$ as diagonal Gaussian distributions, i.e., $\mathcal{N}(z|\mu, \Sigma)$ and $\mathcal{N}(0, I)$ respectively, where μ and Σ are obtained by mapping the last hidden state h^{L-1}_{CLS} with a single-layer MLP respectively.

Decoder for reconstructing the textual representations. In the case of image data, selecting the original input image as the decoder target for reconstruction is straightforward since it contains the most informative content. However, when working with textual data, the input token sequence only represents embeddings from a predefined dictionary, while the intermediate hidden states of the Transformer capture valuable contextual semantics. As a result, determining the appropriate reconstruction target for $p(x|z)$ in textual data poses a challenging task. To leverage the potential of the intermediate hidden states, our approach aims to condition the reconstruction target on the hidden states. Based on our preliminary experiments, we observed that different hidden layers have varying effects on different ID datasets. Consequently, it is difficult to predefine a fixed combination pattern for integrating the intermediate hidden states. Therefore, we introduce a learnable weight vector $\mathbf{s} = [s_0, \cdots, s_{L-1}] \in \mathbb{R}^L$ to dynamically integrate the intermediate hidden states of the Transformer. Then, we derive the reconstruction target:

$$x^{\text{target}} = (h^0_{\text{CLS}} \cdot s_0) + (h^1_{\text{CLS}} \cdot s_1) + \cdots (h^{L-1}_{\text{CLS}} \cdot s_{L-1}),$$

where $\cdot$ denotes multiplication. In this way, x^{target} contains rich contextualized semantic information. Referring to Figure 2, we realize the reconstructor (decoder) $p(x|z)$ as a single feed-forward block, taking a sample z from $\mathcal{N}(z|\mu, \Sigma)$ as input and outputting a reconstructed version of x^{target} to maximize $p(x^{\text{target}}|z)$. The ID classifier f_{ID} is a single-layer MLP that takes the latent representation z as input.

Discriminator for ID classification and binary OOD detection. At the inference stage, we only need the trained posterior approximator (encoder) $q(z|x)$ and the ID classifier f_{ID}. Note that both the ID classification task and the binary outlier detection task are performed w.r.t. the latent variable z. For each x, we only sample one z during training and inference respectively.

4. Experiments

In this section, we present a comprehensive evaluation of textual out-of-distribution (OOD) detection with pervasive OOD detection methods. Besides, we demonstrate the effectiveness of our proposed OOD detection method on challenging natural language understanding benchmarks. To achieve a comprehensive evaluation, we employ both encoder-based and decoder-based pretrained language models as backbone models of our method. We start this section by describing our evaluation methodology and then present our experimental results.

4.1. Evaluation Methodology

4.1.1. Datasets

OOD detection in the natural language processing (NLP) domain is generally under-explored and only discussed in limited scenarios such as out-of-scope intent detection in dialogue machines (Zhan et al., 2021a; Zhang et al., 2021a; Yan et al., 2020). As such, evaluating OOD performance in the NLP domain does not have a consensus. To scale the evaluation process as general as possible, we follow the evaluation in (Hendrycks et al., 2020) and (Zhou et al., 2021b) to present our main analysis. Hendrycks et al. (2020) firstly proposes to use the sentiment analysis benchmark SST-2 as the in-distribution dataset and select five other datasets as out-distribution evaluation sets, which includes 20 Newsgroups, WMT16 and Multi30K, RTE, and SNLI. Zhou et al. (2021b) further extend this benchmark by adding more natural language understanding tasks including topic classification, and question classification.

In-distribution Tasks We use four benchmark datasets as in-distribution (ID) tasks: 20 Newsgroups (20NG) (Lang, 1995), IMDB (Maas et al., 2011), SST-2 (Socher et al., 2013) and TREC-10 (Li and Roth, 2002). When setting each of them as *in-distribution*, other ones are recognized as *out-distribution*.

Besides the above ID four tasks, we also use another four unrelated datasets as OOD test sets (not for training) for all of the four ID tasks. We refer them as the out-distribution datasets: the English source side of English-German WMT16 (Bojar et al., 2016) and English-German Multi30K (Elliott et al., 2016), and concatenations of the premise and hypothesis of RTE (Dagan et al., 2006) and MNLI (Williams et al., 2018). WMT16 and Multi30K are for machine translation while RTE and MNLI are for natural language inference. We use the respective test sets of each out-distribution dataset to measure OOD performance.

4.1.2. Baselines

To demonstrate the effectiveness of our proposed framework, we compare our method comprehensively with four commonly used OOD detection algorithms:

- **Maximum Softmax Probability (MSP)** (Hendrycks and Gimpel, 2017): The MSP confidence score leverages the maximum softmax probability outputted by the softmax function for out-of-domain detection. As correct samples tend to have higher probability scores, samples below a threshold are more likely to be outliers. Specifically, the confidence score is $\mathcal{C}(x) = \max_y p(y|x)$.

- **Mahalanobis Distance (Maha)** (Lee et al., 2018b): The Mahalanobis Distance (MD) method fits K-class conditional Gaussian distributions $\{\mathcal{N}(\mu_i, \Sigma)\}_{i=1}^K$ for the K in-distribution classes upon the output of the penultimate layer in the model. The Mahalanobis Distance and the MD confidence score are computed by:

$$\begin{aligned} \text{MD}_k(z) &= (z-\mu_k)^T \Sigma^{-1} (z-\mu_k), \\ \mathcal{C}(x) &= -\min_k \{\text{MD}_k(z)\}. \end{aligned} \tag{7}$$

- **Energy score (Energy)** (Liu et al., 2020): The energy score confidence score is inspired by the energy-based models (LeCun et al., 2006). It defines an energy of an input (x, y) as $E(x, y) = w_y^T \cdot z$, where w_y is the weight of the softmax layer for the y^{th} in-distribution class. The energy score confidence score is defined as:

$$\mathcal{C}(x) = \log \sum_i^K e^{w_i^T \cdot z}. \tag{8}$$

- **Cosine distance (Cosine)** (Zhou et al., 2021b): The cosine distance OOD confidence sore defines as the maximum cosine similarity of a test input representation with representations in the validation set, i.e., $\mathcal{C}(x) = -\max_{i=1}^V \cos(z, z_i^{val})$.

4.1.3. Metrics

We employ three commonly used metrics for OOD detection and introduce them as follows:

- **AUROC**: Area Under the Receiver Operating Characteristic curve(AUROC) reveals the relationship between True Positive Rate (TPR) (i.e., Recall) and False Positive Rate (FPR). It represents the probability of assigning a higher score to a positive example than a negative example. The pioneering work (Hendrycks and Gimpel, 2017) firstly proposed to use this metric for OOD detection. A higher AUROC score indicates a better classifier, and An AUROC score of 50% means random guessing.

- **FAR@95**: False Alarm Rate at 95% Recall(FAR@95) is the probability that a negative example is misclassified as positive when Recall or TPR is 95%. In this paper, we take the OOD class as negative.

- **AUPR**: Area Under the Precision-Recall curve (AUPR) is another commonly used metric based on the Precision-Recall Curve. It is a better indicator in the case of imbalanced in- and out-rate (Manning and Schutze, 1999). A perfect classifier has an AUPR of 100%.

4.1.4. Experimental Setup

For the encoder-based pre-trained language model, we employ the RoBERTa$_{\text{LARGE}}$ (Liu et al., 2019b) model from HuggingFace (Wolf et al., 2019) as the backbone of our framework. We use the optimizer AdamW (Loshchilov and Hutter, 2019) with a linear-scheduled learning rate 10^{-5} to fine-tune the model for 20 epochs. For the variational terms in Eq. 6, we apply a linear annealing strategy which is a common practice in variational methods (Fu et al., 2019). All reported results are obtained in 5 runs with different random seeds.

For the decoder-based large language model, we validate the effectiveness of our method on LLaMA-2-7B. To reduce training costs, we perform parameter-effective fine-tuning through the LoRA module provided by the HuggingFace PEFT package. Our hyperparameters for LoRA are set as follows: $\alpha = 16$, $r = 16$, and $lora_dropout = 0.05$. We fine-tune the model for 20 epochs with a learning rate of $1e-4$. For additional training details, please refer to the code repository we have released.

4.2. Main Results

To showcase the adaptability of our VI-OOD detection framework, we evaluate it on comprehensive datasets and compare it with competitive baselines. The summarized averaged results can be found in Table 1.

VI-OOD benefits a diverse collection of tasks and OOD score functions. According to Table 1, one notable observation is that our proposed approach consistently outperforms all compared baselines in terms of the overall average performance. This can be observed across various metrics. For instance, when comparing our method to the best-performing baseline, the Maha method, we achieve a significant reduction in the average FAR@95 from

Methods	SST-2			IMDB		
	AUROC ↑	FAR@95 ↓	AUPR ↑	AUROC ↑	FAR@95 ↓	AUPR ↑
MSP	89.85	66.20	86.40	94.30	41.90	98.80
$\text{MSP}_{\text{Contrast}}$	85.04	63.42	69.34	94.51	44.69	98.89
MSP_{VI}	**92.85**	**51.58**	**89.72**	**95.95**	**28.03**	**99.12**
Maha	97.98	11.50	97.30	99.67	0.70	99.95
$\text{Maha}_{\text{Contrast}}$	**99.42**	**2.98**	**98.73**	99.89	**0.05**	**99.97**
Maha_{VI}	99.33	3.62	98.52	**99.90**	0.21	**99.97**
Cosine	95.65	22.65	94.68	99.50	1.53	99.88
$\text{Cosine}_{\text{Contrast}}$	98.38	8.64	96.36	**99.87**	1.93	**99.96**
$\text{Cosine}_{\text{VI}}$	**98.87**	**6.62**	**98.06**	99.57	**1.43**	99.88
Energy	89.80	67.00	86.53	93.30	56.70	98.63
$\text{Energy}_{\text{Contrast}}$	84.93	63.16	69.29	94.44	44.46	98.86
$\text{Energy}_{\text{VI}}$	**92.79**	**51.25**	**89.26**	**96.05**	**27.97**	**99.12**

Methods	TREC-10			20NG		
	AUROC ↑	FAR@95 ↓	AUPR ↑	AUROC ↑	FAR@95 ↓	AUPR ↑
MSP	97.94	8.43	89.26	**93.89**	30.49	**87.39**
$\text{MSP}_{\text{Contrast}}$	98.43	4.06	**91.19**	93.19	28.00	83.17
MSP_{VI}	**98.91**	**2.77**	90.39	93.29	**25.61**	80.09
Maha	98.99	4.87	95.11	98.39	7.77	95.91
$\text{Maha}_{\text{Contrast}}$	**99.57**	0.97	**98.59**	98.78	5.89	97.29
Maha_{VI}	99.46	**0.79**	97.67	**99.80**	**0.61**	**98.93**
Cosine	98.89	3.96	94.54	97.73	10.84	88.71
$\text{Cosine}_{\text{Contrast}}$	99.14	1.42	93.34	98.03	8.86	95.27
$\text{Cosine}_{\text{VI}}$	**99.36**	**1.19**	**96.09**	**99.39**	**2.92**	**97.19**
Energy	97.19	10.07	82.16	95.76	17.93	**88.71**
$\text{Energy}_{\text{Contrast}}$	98.45	4.73	**91.18**	**96.04**	**15.70**	88.62
$\text{Energy}_{\text{VI}}$	**99.21**	**2.84**	90.84	94.34	17.04	79.67

Average	AUROC ↑	FAR@95 ↓	AUPR ↑
avg. (MSP / Maha / Cosine / Energy)	94.00 / 98.78 / 97.94 / 94.01	36.76 / 6.21 / 9.75 / 37.93	**90.46** / 97.07 / 94.45 / 89.01
$\text{avg.}_{\text{Contrast}}$ (MSP / Maha / Cosine / Energy)	92.79 / 99.17 / 98.86 / 93.47	35.04 / 3.93 / 5.21 / 32.01	85.65 / 97.43 / 96.23 / 86.99
avg._{VI} (MSP / Maha / Cosine / Energy)	**95.25** / **99.62** / **99.30** / **95.60**	**27.00** / **1.31** / **3.04** / **24.78**	89.83 / **98.77** / **97.81** / **89.72**

Table 1: Main results of our proposed framework. MSP, Maha, Energy, and Cosine are baseline methods trained with the discriminative loss, while each corresponding method with the *VI* subscript denotes the model trained with our VI framework. The *Contrast* subscript denotes the method proposed by Zhou et al. (2021b). The best result is marked in bold. At the bottom row, averaged results across four ID datasets are included. All the reported results are presented in percentage values.

6.21% to 1.31%, resulting in a relative increase of 78.9%. Similarly, for the second best baseline, the Cosine score function, our method demonstrates substantial improvement by reducing the average FAR@95 from 9.75% to 3.04%. Moreover, there are significant performance gains in terms of AUROC as well. For example, the average AUROC score for the Cosine method increases from 97.94% to 99.3% with the use of our method. It is worth noting that our method achieves these improvements without the need for real OOD examples, which makes these results even more encouraging. Upon closer examination of each of the four in-distribution (ID) datasets, it becomes apparent that detecting out-of-distribution (OOD) test examples using the model trained on TREC-10 is comparatively easier than with the other datasets. In fact, all OOD score functions achieve AUROC scores above 97%. Improving upon these already competitive results poses a significant challenge. However, our method still manages to outperform all four score functions on TREC-10. Notably, for the Energy score function, our method enhances the AUROC score from 97.19% to 99.21%, while simultaneously reducing the FAR@95 score from 10.07% to 2.84%. Furthermore, for the Maha method on TREC-10, our method achieves a near-perfect FAR@95 score of 0.79%.

Superior Performance of Our Method with Large Decoder-Only Models. To further validate our method's efficacy, we conduct experiments on SST-2 and TREC-10 using LLaMA-2-7B. For training the in-distribution classifier, we utilize *LlamaForSequenceClassification* from the Hugging Face transformers package (Wolf et al., 2020).

Results are presented in Table 2. As shown, when using SST-2 as the in-distribution (ID) dataset, our method significantly outperforms the baselines across all three metrics. In the case of TREC-10, our method elevates the Mahalanobis (Maha) and Cosine OOD scores to nearly perfect levels. However, for logit-based OOD scores, specifically MSP and Energy, our method demonstrates slightly inferior performance relative to the respective baselines. This discrepancy may stem from the fact that Maha and Cosine benefit more from the enriched information in sentence representations, whereas this richer information introduces more ambiguity

	SST-2			TREC-10		
Methods	AUROC ↑	FAR@95 ↓	AUPR ↑	AUROC ↑	FAR@95 ↓	AUPR ↑
MSP	78.22	70.40	74.34	**99.29**	**0.64**	**98.79**
MSP$_{\text{VI}}$	**83.05**	**65.32**	**69.54**	98.69	5.69	94.48
Maha	46.08	84.41	47.34	84.41	71.94	64.13
Maha$_{\text{VI}}$	**95.41**	**12.47**	**85.38**	**99.96**	**0.12**	**99.39**
Cosine	89.24	26.52	79.47	98.24	11.06	93.05
Cosine$_{\text{VI}}$	**95.95**	**7.21**	**85.70**	**100**	**0**	**99.92**
Energy	70.13	67.78	64.87	**99.86**	**0.14**	**99.27**
Energy$_{\text{VI}}$	**80.16**	**66.90**	**65.76**	98.64	7.40	92.44

Table 2: Results of our proposed framework on LLaMA-2-7B, fine-tuned with a classification head. The best result is marked in bold. All the reported results are presented in percentage values.

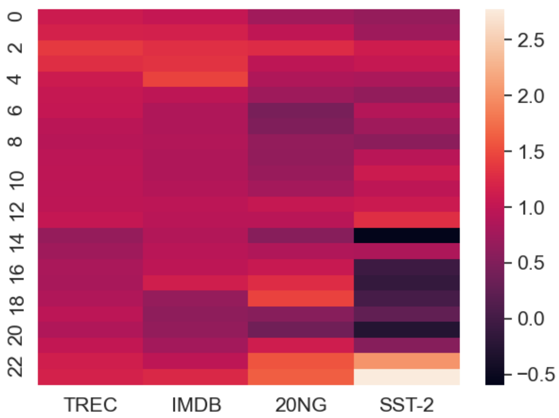

Figure 3: Heatmap of the hidden state combination factor s. The horizontal axis stands for four ID tasks and the vertical axis represents the layer number.

in the logits.

4.3. ID Classification Performance

In this subsection, we investigate the ID classification performance. Besides the binary ID *vs.* OOD task, OOD detection also concerns the ID classification task. We summarize the test accuracy of the corresponding ID test sets for the four ID datasets in Table 3. It can be seen that for 20NG, SST-2, TREC-10, and IMDB, ID test performances are very similar and all the gaps are lower than 1%. Therefore, models trained with our proposed $p(x, y)$ target do not bring significant detrimental impacts to ID classification. However, although we consider these gaps can be ignored in practical applications, it also indicates that our method can be further improved in further works.

4.4. The Combination Factor s

Finally, we analyze the learned combination vector $\mathbf{s}$ in our framework. We visualize the heatmap of the learned $\mathbf{s}$ for the in-distribution (ID) tasks in Figure 3. It is evident that the hidden state combination patterns vary significantly across different ID tasks. This observation confirms that our proposed combination vector can automatically adapt and learn appropriate combination policies for distinct ID tasks. This analysis provides further evidence of the flexibility of our framework in effectively leveraging the potent hidden states of pre-trained models.

Test Accuracy	SST-2	IMDB	TREC-10	20NG
$p(y\|x)$	96.21	95.33	97.8	93.99
$p(x, y)$	96.38	94.54	97.0	93.35

Table 3: Performance comparison of the ID K-class classifier for different training objectives. $p(y|x)$ is the commonly used discriminative objective and $p(x, y)$ is our proposed objective.

5. Related Work

OOD detection based on density estimation. Besides the problem setting discussed in Section (2), another line of works tries to address the OOD detection problem by solving a more general problem – density estimation. Unlike the setting of our work, the focus of these works is solely on the binary classification task of distinguishing between in-distribution (ID) and OOD samples, disregarding the ID classification task. Their learning target is the density function of the training set – $p_{ID}(x)$ – such that OOD examples are assumed to yield lower probabilities than the ID ones. However, in high dimensional spaces, this assumption is not held in practice and many previous works (Choi et al., 2018) have found that OOD examples may be assigned higher likelihoods than ID examples. Recent works (Ren et al., 2019; Nalisnick et al., 2019; Morningstar et al., 2021) are still trying to correct this pathology.

In particular, numerous prior studies have leveraged the density estimation capabilities of variational autoencoders (VAEs) for OOD detection. For instance, Floto et al. (2023) enhance VAEs for OOD detection by substituting the standard Gaussian prior with a more versatile tilted Gaussian distribution. Likelihood Regret (Xiao et al., 2020) and Likelihood ratios (Ren et al., 2019) adopt a similar perspective of training two distinct models–one capturing the semantic content of the data, and the other capturing background information. Their major difference is the training data of the background model and semantic model.

OOD detection in NLP. OOD detection in the NLP domain has recently attracted increased attention (Liu et al., 2023). OOD intent detection (Zhang et al., 2021b; Zhan et al., 2021b) investigates the OOD detection problem for anomalous utterances in dialogue systems. Podolskiy et al. (2021b) empirically find out that Mahalanobis Distance is the best performing OOD scoring function for OOD intent detection. A few attempts has been made to study the general textual OOD detection problem. Hendrycks et al. (2020) point out that pre-trained Transformers are more robust for OOD detection than previous model architectures (Hochreiter and Schmidhuber, 1997). Zhou et al. (2021b) and Cho et al. (2022) employ a contrastive regularizer to learn better representations for textual OOD detection. Uppaal et al. (2023) conduct an evaluation on RoBERTa and point out that ID fine-tuning may pose a detrimental effect on textual OOD detection.

6. Conclusion

This paper concentrates on exploring Out-of-Distribution (OOD) detection within Natural Language Processing (NLP) classification tasks using Transformer-based large language models (LLMs). Building on our detailed analysis of hidden states in Transformers, we introduce a Variational Bayesian framework named VI-OOD. This framework optimizes the joint distribution $p(x, y)$ during the training phase. Our methodology is reinforced by both experimental evidence and theoretical analysis, underscoring its validity. We have rigorously tested our approach with mainstream Transformer architectures, encompassing both encoder-based and decoder-based models. Comprehensive experiments on diverse textual classification tasks affirm the efficacy and superiority of our OOD detection framework.

This research is dedicated to enhancing AI safety and the robustness of models. As such, our findings are poised to benefit various AI applications without presenting a direct risk of misuse. Furthermore, our proposed methodology relies exclusively on open-source benchmarks for training data, avoiding the introduction of additional datasets for training the OOD detector. As such, our approach sidesteps potential ethical concerns associated with data collection. Additionally, by building upon open-source LLMs, our method avoids substantial increases in resource consumption, aligning with principles of sustainable and responsible AI development.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. This research was partially supported by the grant of HK ITF ITS/359/21FP.

7. References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. 2020. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, pages 548–564. Springer.

Hyunsoo Cho, Choonghyun Park, Jaewook Kang, Kang Min Yoo, Taeuk Kim, and Sang-goo Lee. 2022. Enhancing out-of-distribution detection in natural language understanding via implicit layer ensemble. *arXiv preprint arXiv:2210.11034*.

Hyunsun Choi, Eric Jang, and Alexander A Alemi. 2018. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton,* *UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Angelos Filos, Panagiotis Tigas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. 2020. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Griffin Floto, Stefan Kremer, and Mihai Nica. 2023. The tilted variational autoencoder: Improving out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 240–250. Association for Computational Linguistics.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2022. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ryo Kamoi and Kei Kobayashi. 2020. Why is the mahalanobis distance effective for anomaly detection? *CoRR*, abs/2003.00402.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pages 331–339. Morgan Kaufmann.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018a. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018b. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7167–7177.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2023. How good are large language models at out-of-distribution detection? *arXiv preprint arXiv:2308.10261*.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from BERT: an empirical study. *CoRR*, abs/1910.07973.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the role of BERT token representations to explain sentence probing results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 792–806. Association for Computational Linguistics.

Warren R. Morningstar, Cusuh Ham, Andrew G. Gallagher, Balaji Lakshminarayanan, Alexander A. Alemi, and Joshua V. Dillon. 2021. Density of states estimation for out of distribution detection. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 3232–3240. PMLR.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. 2019. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 427–436. IEEE Computer Society.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021a. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13675–13682. AAAI Press.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021b. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics - 18th China National Conference, CCL 2019,* *Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pages 194–206. Springer.

Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 144–157.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. 2020. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In *Machine Learning for Health*, pages 341–354. PMLR.

Rheeya Uppaal, Junjie Hu, and Yixuan Li. 2023. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. In *Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhisheng Xiao, Qing Yan, and Yali Amit. 2020. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33:20685–20696.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y. S. Lam. 2020. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1050–1060. Association for Computational Linguistics.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. 2022. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y.S. Lam. 2021a. Out-of-scope intent detection with self-supervision and discriminative training. 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 3521–3532, Online. Association for Computational Linguistics.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y.S. Lam. 2021b. Out-of-scope intent detection with self-supervision and discriminative training. 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 3521–3532, Online. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021a. Deep open intent classification with adaptive decision boundary. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14374–14382.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021b. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14374–14382.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021a. Contrastive out-of-distribution detection for pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1100–1111. Association for Computational Linguistics.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021b. Contrastive out-of-distribution detection for pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111.