Semantic Role Labeling Guided Out-of-distribution Detection

Jinan Zou[1,†] , Maihao Guo[1,†], Yu Tian[2,†], Yuhao Lin[1], Haiyao Cao[1], Lingqiao Liu[1], Ehsan Abbasnejad[1], Javen Qinfeng Shi[1,*]

[1] Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia
[2]Harvard University, Cambridge, USA
{jinan.zou, javen.shi}@adelaide.edu.au, ytian11@meei.harvard.edu

Abstract

Identifying unexpected domain-shifted instances in natural language processing is crucial in real-world applications. Previous works identify the out-of-distribution (OOD) instance by leveraging a single global feature embedding to represent the sentence, which cannot characterize subtle OOD patterns well. Another major challenge current OOD methods face is learning effective low-dimensional sentence representations to identify the hard OOD instances that are semantically similar to the in-distribution (ID) data. In this paper, we propose a new unsupervised OOD detection method, namely Semantic Role Labeling Guided Out-of-distribution Detection (SRLOOD), that separates, extracts, and learns the semantic role labeling (SRL) guided fine-grained local feature representations from different arguments of a sentence and the global feature representations of the full sentence using a margin-based contrastive loss. A novel self-supervised approach is also introduced to enhance such global-local feature learning by predicting the SRL extracted role. The resulting model achieves SOTA performance on four OOD benchmarks, indicating the effectiveness of our approach. The code is publicly accessible via `https://github.com/cytai/SRLOOD`.

Keywords: Out-of-distribution Detection, Semantic Role Labeling, Domain Shift

1. Introduction

Recent advances in natural language processing have shown tremendous improvements in various natural language classification tasks. Natural language classification is usually formulated as a close-set problem, where training and testing samples are from the same domain/distribution. Despite the accurate predictions on the inlier close-set classes, the classifier often fails to properly identify out-of-distribution (OOD) instances from other unknown/unexpected domains that deviate from the close-set training distribution, bringing risks to real-world scenarios. Tackling such failure cases is crucial to real-world safety-critical NLP applications. For instance, OOD instances can be represented by unknown sentences from different domains or distributions, such as semantically shifted sentences that can be incorrectly predicted as a part of the inlier classes, leading to potential impairment to user trust (Arora et al., 2021). Although large language models (LLMs) are revolutionizing the field of NLP, they are prone to OOD and even adversarial inputs (Wang et al., 2023). OOD detection can be applied to directly handle OOD inputs and avoid potentially harmful responses (Bai et al., 2022).

Despite the importance, little literature has addressed the problem of OOD detection in NLP. One proposed method is to train a model to increase the inter-class discrepancy of in-distribution (ID) classes and tends to depend on classification uncertainty or latent embedding distance to detect OOD instances (Zhou et al., 2021). The high classification uncertainty association with OOD instances (i.e., max softmax or energy) is intuitive, but it obtains a few caveats. One of the major issues is that classification uncertainty happens when samples are close to classification decision boundaries. However, there is no guarantee that all OOD instances will be close to classification boundaries (i.e., subtle OOD samples may share similar semantic features to ID data), leading to subpar performance in detecting OOD samples. Moreover, complicated inlier sentences containing more outlier components, such as punctuation and discourse fillers, tend to fall close to the decision boundary, which can incorrectly lead to high classification uncertainty. Latent embedding-based approaches rely on the assumption that the OOD instance resides outside a bounded or unbounded latent hyperspace constructed by the ID feature distributions (Zhou et al., 2021; Hendrycks et al., 2020; Cao and Zhang, 2022; Rawat et al., 2021; Tan et al., 2019). However, it is challenging to define such a latent hyperspace to encode all possible ID features, significantly affected by many outlier components from a sentence, and the aforementioned subtle OOD issue still exists.

In this paper, we propose a new OOD detection method designed for NLP tasks, namely Semantic Role Labeling Guided Out-of-distribution Detection (SRLOOD), simultaneously extracting, separating, and learning both global and SRL-guided local fine-grained feature representations through a margin-based contrastive loss and self-supervision. We identify a critical shortcoming in current NLP

[†] Indicates equal contribution as first authors
[*] Corresponding author

LREC-COLING 2024, pages 14641–14651

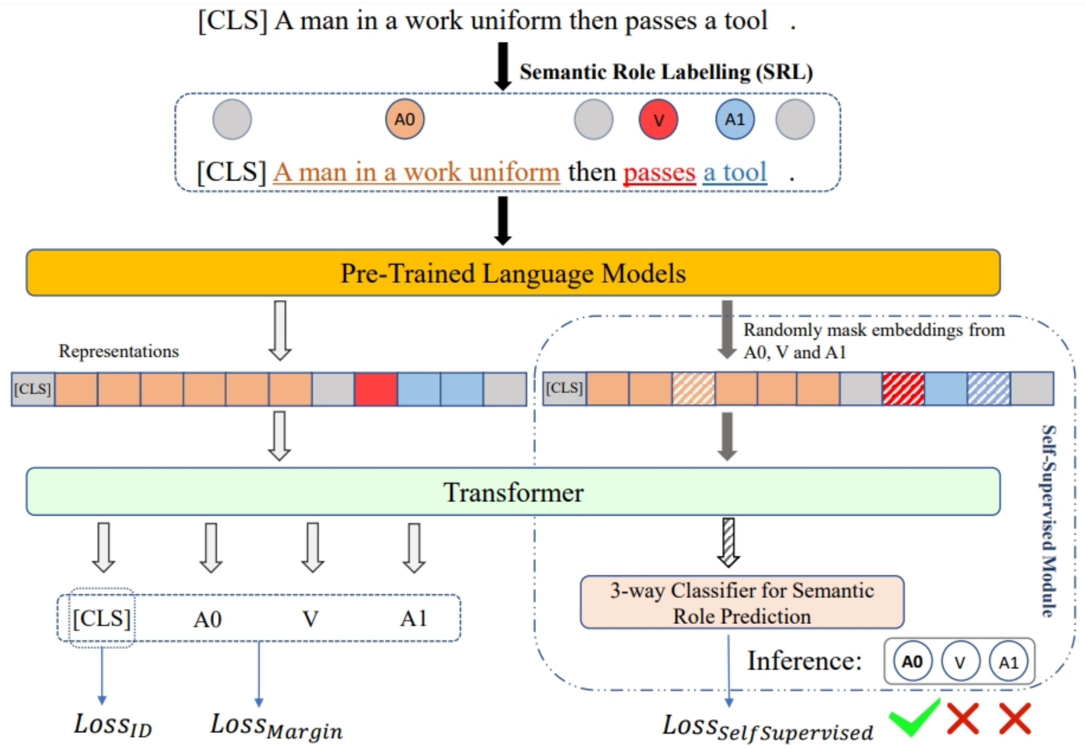

Figure 1: Model architecture of our framework. The Transformers including the pre-traiend language model and a subsequent encoder are guided by SRL, extracting global and local representations of input sequence according to the semantic roles A0, V, A1, and masking them according to the semantic roles A0, V, A1 to construct the Self-Supervised Module. An additional 3-way Classifiers take the Transformer representation of A0, V or A1 MASKs as input, and predict their semantic roles.

models' ability to detect OOD cases: they lack nuanced, low-dimensional local representations that are crucial for identifying OOD instances semantically similar to ID data. SRL emerges as a key technique, that aims at extracting vital local features while effectively omitting outlier elements, including punctuation and discourse fillers. This innovative integration of SRL into our methodology leads to a remarkable enhancement in both the performance and efficiency of OOD detection. In particular, our contributions can be summarised into three folds:

- We propose SRLOOD that learns fine-grained low-dimensional representations by increasing the inter-class discrepancies between the concatenation of the global and SRL-guided local features of different ID classes. Our proposed SRLOOD aims to effectively eliminate the outlier phrases (e.g., punctuation and discourse fillers) and extract key local semantic components (e.g., verbs and arguments) from a sentence to better characterize subtle OOD samples;

- A novel self-supervised pretext task is also proposed to strengthen the relations between different local arguments, further facilitating the optimization of SRL-guided local features; and

- A Transformer block is introduced to resemble some of the SRL-guided features from a sentence, so our model is enforced to learn discriminative representations through such strong perturbations for the better discriminability of subtle semantic features.

- Extensive experiments on four different OOD benchmarks show that our resulting model achieves the best performance on four different scoring functions.

2. Related Work

Out-of-distribution detection Machine learning aims to design models that can learn generalizable knowledge from training data. The success of machine learning models lies in the assumption that training and test data share the same distribution. However, in many real-world tasks, it is unknown whether the training and test data share the same distribution. For online LLMs such as ChatGPT that interact with users, the inputs out of the training distribution is prevalent(Wang et al., 2023). This potential distribution gap is known as OOD and can be a major issue, with the performance of classical ML models often deteriorating. To handle the OOD issue, OOD detection aims to detect whether test data is from

the training distribution. Based on the availability of OOD data, recent methods can be categorized into classification methods, density-based methods, and distance-based methods (Yang et al., 2021). Classification methods often formulate the OOD task as a one-class classification problem, then use appropriate methods to solve it (Ruff et al., 2018; Chen et al., 2021b; Tian et al., 2021; Hendrycks et al., 2019a; Lee et al., 2018; Dhamija et al., 2018; Morteza and Li, 2022; Chen et al., 2022). Hendrycks and Gimpel (2017) proposed a softmax prediction probability baseline for error and out-of-distribution detection across several architectures and numerous datasets. Density-based methods (Cao and Zhang, 2022; Abati et al., 2019; Zisselman and Tamar, 2020; Kirichenko et al., 2020) in OOD detection explicitly model the in-distribution with some probabilistic models and flag test data in low-density regions as OOD. Zong et al. (2018) utilizes a deep autoencoder to generate a low-dimensional representation and reconstruction error for each input data point, which is further fed into a Gaussian Mixture Model for anomaly detection. The main idea of distance-based methods is that the testing OOD samples should be relatively far away from the centroids of in-distribution classes (Lee et al., 2018; Chen et al., 2020; Van Amersfoort et al., 2020; Zaeemzadeh et al., 2021). Previous methods primarily studied for computer vision (Lin et al., 2021; Huang and Li, 2021; Zaeemzadeh et al., 2021; Zhou, 2022; Dong et al., 2022) and OOD detection has been overlooked in NLP. Only few works recently that adapted the solutions designed for images into the text to leverage the features representation of an entire sentence for detecting the OOD case. For example, Zhou et al. (2021) adapted a contrastive OOD detection from computer vision using a pre-trained Transformer to improve the compact news of representations and evaluate the trained classifier on the four text datasets.
In contrast, we propose a self-supervised SRL method to learn fine-grained feature representations of text data and shows that is a surprisingly effective approach for OOD detection.

Semantic Role Labeling Semantic role labelling(SRL) leads to the advancement of many NLP tasks and applications due to the clear detection of augments regarding predicates. For example, Sarzynska-Wawer et al. (2021) proposed a BERT-based model incorporating semantic role labelling, which significantly improves the text understanding ability of the model. Chen et al. (2021a) used the verb-specific semantic role, a variant of semantic role labelling, for the controllable image captioning, which is a task about image description. Conditioned on the semantic role representation. More recently, Ross et al. (2022) proposed a Tailor model for the sequence-to-sequence task, which gained a great improvement in measuring the reliance on syntactic heuristics.

Self-supervised Learning Self-supervised learning method and has been soaring and achieving big success in representative learning because of the powerful generalization ability. BERT (Pre-training of deep bidirectional Transformers for language understanding) proposed by Devlin et al. (2018) are fine-tuned for many downstream tasks. as a result, BERT has become a milestone of not only NLP but also the development of self-supervised learning. Baevski et al. (2022) built a platform based on a self-supervised method for either speech, text or computer vision. Previous research has shown that the self-supervised method drastically improves the OOD detection performance on the difficult near-domain outliers Hendrycks et al. (2019b); Tian et al. (2023). Self-supervised learning methods tackle the OOD in two aspects: (1) the enhancement of feature quality can improve OOD performance; (2) some well-designed surrogate tasks can help reveal the anomalies from OOD samples(Yang et al., 2021).

3. Methodology

Generally, the OOD instances can be defined as instances (x, y) sampled from an underlying distribution other than the training in-distribution $P(\mathcal{X}_{train}, \mathcal{Y}_{train})$, where $\mathcal{X}_{train}$ and $\mathcal{Y}_{train}$ are the training corpus and training label set. Specifically, an instance $(\boldsymbol{x}, y)$ is primarily deemed OOD if $y \notin \mathcal{Y}_{train}$ to be consistent with previous works (Hendrycks and Gimpel, 2017; Hendrycks et al., 2019a, 2020; Zhou et al., 2021). Following the previous work (Zhou et al., 2021), we formally define the OOD detection task. Given the main task of natural language classification, the OOD detection task is the binary classification of each instance $\boldsymbol{x}$ as either ID or OOD, judged by its OOD score computed with scoring function $f(x) \to \mathbb{R}$. A lower OOD score value indicates ID where $y \in \mathcal{Y}_{train}$ and a higher OOD score value indicates OOD where $y \notin \mathcal{Y}_{train}$ (y is the underlying label for $\boldsymbol{x}$ and is unknown at inference).

The key idea of our proposed model, SRLOOD, is extracting and learning the SRL-guided fine-grained local representation. Building on top of this representation, a novel supervised approach is introduced to enhance such local argument representation.

Our framework consists of (1) semantic role labelling, (2) an SRL-guided self-supervised module, and (3) OOD detection with OOD scoring functions

Algorithm 1 Learning Process

Input ID training set $\mathcal{D}_{\text{train}}$ and ID validation set $\mathcal{D}_{\text{val}}$.
Output A trained classifier and an OOD detector.

Load the Pre-Trained Transformer and initialize the subsequent Transformer.
for $t = 1...T$ **do**
  Sample a batch from $\mathcal{D}_{\text{train}}$.
  Calculate the ID classification loss $\mathcal{L}_{\text{ID}}$.
  Calculate the contrastive loss $\mathcal{L}_{\text{margin}}$.
  Calculate the self-supervised loss $\mathcal{L}_{\text{SSL}}$.
  $\mathcal{L}_{\text{total}} = \alpha_1\mathcal{L}_{\text{ID}} + \alpha_2\mathcal{L}_{\text{Margin}} + \alpha_3\mathcal{L}_{\text{SSL}}$.
  Update model parameters w.r.t. $\mathcal{L}_{\text{total}}$.
  if t % evaluation steps $= 0$, **then**:
    Fit the OOD detector on $\mathcal{D}_{\text{val}}$.
    Evaluate both the classifier and OOD detector on $\mathcal{D}_{\text{val}}$.
Return the best model checkpoint.

as illustrated in Figure 1.

3.1. Semantic Role Labeling

The task of SRL is to determine the underlying predictive argument structure of a sentence and to provide representations that can answer the basic questions about the meaning of the sentence, including who did what to whom (Màrquez et al., 2008). Therefore the SRL primarily extracts the essential features and passingly filters out outlier phrases (e.g., punctuation and discourse fillers). We leverage off-the-shelf SRL-BERT (Shi and Lin, 2019) to label each token sequence in a batch with Propbank (Kingsbury and Palmer, 2003) semantic roles proto-agent, verb, and proto-patient, then labeled tokens are recorded into sets A0, V, and A1 respectively, as illustrated in Figure 1. Each token sequence is fed to the pre-trained language model, whose output is fed to the Transformer block. We compute the mean of A0, V, A1 embeddings $\boldsymbol{\mu}_{A0}$, $\boldsymbol{\mu}_V$, and $\boldsymbol{\mu}_{A1}$ pooled from the Transformer's output for fine-grained feature representations.

3.2. Self Supervised Learning based on SRL

Our proposed SRLOOD framework uses SRL to extract and learn key local semantic features and use self-supervision to further strengthen such fine-grained local representations. We introduce strong perturbation by randomly masking a certain percentage of SRL-extracted local representations for better generalization on detecting hard OOD instances.

Guided by SRL, strong perturbation is independently exerted on the pre-trained language model representations of A0, V, and A1 according to a generated and recorded supervising ground truth label for each sequence. The perturbed embeddings are input to the Transformer encoder. Subsequently, we compute the mean embeddings of A0, V, or A1 from the Transformer's output and use them for an auxiliary three-way classification task. This task aims to improve the feature discriminability by predicting the semantic role of a given embedding, computing the mean embeddings, and inputting one of them to a classifier for semantic role prediction according to its self-supervising label. To this end, our framework is consisted of the pre-trained language model, the Transformer head, the SRL-guided pooling, and the 3-way self-supervised classifier. This framework is optimized by the loss functions introduced in the next section.

3.3. Loss Functions

We adopt the margin-based contrastive loss that drives the model to encode tokens in the same ID class with adjacent SRL-guided comprehensive representation measured by L2 distances:

$$\mathcal{L}_{\text{margin}} = \frac{1}{md}\left[\sum_{i=1}^{m}\frac{1}{|P(i)|}\sum_{p\in P(i)}\left(\|\boldsymbol{h}_i - \boldsymbol{h}_p\|^2\right) + \sum_{i=1}^{m}\frac{1}{|N(i)|}\sum_{n\in N(i)}\left(\xi - \|\boldsymbol{h}_i - \boldsymbol{h}_n\|^2\right)_+\right] \tag{1}$$

where $P(i)$ is the subset of training data with the same class as instance i, $N(i)$ is the subset of training data with different class labels from instance i, m is the number of instances in the entire training set. The d is the dimensionality of comprehensive representation of a sequence $\boldsymbol{h} = Concat(\boldsymbol{h}_{[CLS]}; \boldsymbol{\mu}_{A0}; \boldsymbol{\mu}_V; \boldsymbol{\mu}_{A0})$, where $\boldsymbol{h}_{[CLS]}$ is the [CLS] embedding, $\boldsymbol{\mu}_{A0}$, $\boldsymbol{\mu}_V$, $\boldsymbol{\mu}_{A1}$ are the mean embeddings pooled from the Transformer's output according to A0, V, A1 respectively. The Margin Loss will give rise to clusters in the latent space of $\boldsymbol{h}$. Combined with cross-entropy losses $\mathcal{L}_{\text{ID}}$ and $\mathcal{L}_{\text{SSL}}$ from the ID sequence classification task and the self-supervised task, respectively. The total loss is their weighted sum with hyper-parameters α_1, α_2 and α_3:

$$\mathcal{L}_{\text{total}} = \alpha_1\mathcal{L}_{\text{ID}} + \alpha_2\mathcal{L}_{\text{margin}} + \alpha_3\mathcal{L}_{\text{SSL}}. \tag{2}$$

3.4. Scoring Functions

During OOD inference, we extract the local key components and features using SRL-BERT (Shi and Lin, 2019) based on a previously fine-tuned language model backbone. We compute the mean embeddings $\boldsymbol{\mu}_{A0}$, $\boldsymbol{\mu}_V$, and $\boldsymbol{\mu}_{A1}$ for local feature representations. The [CLS] embedding $\boldsymbol{h}_{[CLS]}$ is used for global feature representations. The global and local representations are then concatenated together to produce the final feature vector to represent a sentence

$$\boldsymbol{h} = Concat(\boldsymbol{h}_{[CLS]}; \boldsymbol{\mu}_{A0}; \boldsymbol{\mu}_{V}; \boldsymbol{\mu}_{A0}). \quad (3)$$

For a fair comparison, we use the same OOD scoring functions as Zhou et al. (2021). For the validation set $\mathcal{D}^{val} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}^{val}|}$, we computed the Mahalanobis distance based on the class mean embedding $\boldsymbol{\mu}_c = \mathbb{E}_{y_i=c}[\boldsymbol{h}_i]$, $c \in C$ the number of classes, and its covariance $\boldsymbol{\Sigma} = \mathbb{E}\left[\left(\boldsymbol{h}_i - \boldsymbol{\mu}_{y_i}\right)\left(\boldsymbol{h}_i - \boldsymbol{\mu}_{y_i}\right)^{\intercal}\right]$, where $i = 1, ..., C$. The OOD score S is then defined as the minimum Mahalanobis distance among the C ID classes given an instance $\mathbf{x}$ during inference:

$$S = -\min_{c=1}^{C}(\boldsymbol{h} - \boldsymbol{\mu}_c)^{\intercal}\boldsymbol{\Sigma}^{\dagger}(\boldsymbol{h} - \boldsymbol{\mu}_c), \quad (4)$$

where $\boldsymbol{\Sigma}^{\dagger}$ denotes the pseudo-inverse of the covariance matrix $\boldsymbol{\Sigma}$. Such a distance considers both the global sentence features and the SRL-guided local features, enabling better performance on OOD detection.

For cosine similarity, we compute the maximum cosine similarity of the concatenated feature representation $\boldsymbol{h}$ to instance features of the validation set $\mathcal{H}^{val} = \{(\boldsymbol{h}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{H}^{val}|}$. The OOD score is computed as

$$S = -\max_{i=1}^{|\mathcal{H}^{val}|} \cos(\boldsymbol{h}, \boldsymbol{h}_i). \quad (5)$$

Maximum Softmax Probability (MSP) (Hendrycks and Gimpel, 2017) and Energy Score (Energy) (Liu et al., 2020) represent the class of probabilistic scoring functions. Although the MSP is biased, not aligned with the density of the inputs (Liu et al., 2020), it is widely adopted as a baseline for OOD detection. For C training classes in the softmax layer, the MSP score is defined by the maximum class probability:

$$S = 1 - \max_{j=1}^{C} \boldsymbol{p}_j. \quad (6)$$

Liu et al. (2020) estimates the probability density of inputs as:

$$S = -\log \sum_{j=1}^{C} \exp(\boldsymbol{w}_j^{\intercal}\boldsymbol{h}_{softmax}), \quad (7)$$

where $\boldsymbol{w}_j \in \mathbb{R}^d$ is the weight of the j^{th} class in the softmax layer, $\boldsymbol{h}_{softmax}$ is the input to the softmax layer. A higher energy score S indicates a greater likelihood of being OOD data, thereby suggesting a lower likelihood of being ID data.

3.5. Datasets

Previous studies on OOD detection mostly focus on computer vision, while few have been made on natural language processing. Zhou et al. (2021) propose a extensive benchmarks for OOD detection on natural language processing and use different pairs of NLP datasets as ID and OOD data. Following Zhou et al. (2021), we use the same NLP datasets and same criterion on choosing ID and OOD data to evaluate our proposed method. The ID datasets correspond to three categories of natural language classification tasks as following:

- **Sentiment Analysis** Following (Zhou et al., 2021), we use SST2(Socher et al., 2013) and IMDB (Maas et al., 2011) as our ID datasets, which are both sentiment analysis datasets. Note that both datasets belong to the same task and are not condisered OOD to each other.

- **Topic Classification** Following (Zhou et al., 2021), we use 20 Newsgroup dataset (Lang, 1995) as our ID dataset, which is a dataset for topic classification containing 20 classes.

- **Question Classification** Following (Zhou et al., 2021), we use TREC-10 dataset(Li and Roth, 2002) as our ID dataset, which classifies questions based on the types of their sought-after answers.

Moreover, for the above three tasks, any pair of datasets for different tasks can be regarded as OOD to each other. Besides, following Zhou et al. (2021), we employ for additional datasets solely as the OOD data: concatenations of the premises and respective hypotheses from two **NLI** datasets RTE (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009) and MNLI(Williams et al., 2018), the English source side of **Machine Translation**(**MT**) datasets English-German WMT16 (Bojar et al., 2016) and Multi30K (Elliott et al., 2016).

3.6. Evaluation Metrics

We adopted the same two metrics (Zhou et al., 2021) commonly used for measuring OOD detection performance in machine learning researches (Hendrycks and Gimpel, 2017; Lee et al., 2018): **AUROC** and **FAR95**. **AUROC** is the area under the receiver operating characteristic (ROC) curve. It compares the true positive rate (TPR) to the false positive rate (FPR). **FAR95** is the probability of mistakenly classifying OOD as ID at a 95% TPR.

3.7. Experiments Details

We conducted all experiments based on the same codebase and used the same $\text{RoBERTa}_{\text{LARGE}}$ from previous work (Zhou et al., 2021). The Transformer encoder has 3 layers and 16 attention heads. The weights $\alpha_1 = 1$, $\alpha_2 = 3$, $\alpha_3 = 1$. The warm-up ratio for learning rate is 0.06. The batch size is 12. We use AdamW (Kingma and Ba, 2014) to optimized our model, and a learning rate of $1e-5$ and weighted decay 0.01. We pick the masking probability that optimize the average OOD detection performance, 30% for SST2 and IMDB, and 50% for TREC-10 and 20NG, guided by Figure 3. The model is trained for 10 epochs with runtime ranging from 5 hours to 10 hours on one Tesla V100 GPU. We further discuss the performance of taking different masking probabilities in Figure 3. Please note that we manually select all hyper-parameters based on the AUC and FAR performance on testing sets. The total number of model parameters is 392M. All the hyper-parameters are tuned on the development sets.

Compared Methods. We compare our method with three baselines: OOD detection using probabilities from softmax distributions w/o $\mathcal{L}_{\text{Cont}}$-MSP (Hendrycks and Gimpel, 2017), fine-tuning the Transformers with supervised contrastive loss w/ $\mathcal{L}_{\text{SCL}}$, and with margin-based loss w/ $\mathcal{L}_{\text{margin}}$ (Zhou et al., 2021).

3.8. Main Results

As demonstrated in Table 1, our SRLOOD model excels in performance compared to three diverse State-of-the-Art (SOTA) techniques across a variety of Out-of-Distribution (OOD) benchmarks. Regardless of the scoring function applied, our model delivers superior Mean Area Under the Curve (AUC) and False Acceptance Rate (FAR) results. This high performance pertains to MSP, energy, Cosine, and Mahalanobis distance scoring.

Our model, even without fine-tuning of classification logits, offers significant advancements over preceding SOTA methodologies. All reported results represent an average of five independent runs, each initiated with different random seeds. Among the various OOD detection functions, Mahalanobis and cosine distances prove to be the most effective, outstripping the MSP and energy baseline by a significant margin, as documented in (Hendrycks and Gimpel, 2017). The reason behind the superior performance of Mahalanobis and cosine distances in detecting distributional differences can be attributed to their enhanced capacity to encapsulate such differences.

When contrasted with baseline methodologies that do not employ contrastive loss, our approach notably improves the FAR by an average of 10%, 14%, 7%, and 6% for MSP, energy, Mahalanobis, and cosine distances respectively. When pitted against earlier SOTA techniques (Zhou et al., 2021) that use $\mathcal{L}_{\text{SCl}}$, $\mathcal{L}_{\text{margin}}$ losses, our strategy significantly surpasses them in terms of mean AUC and FAR across all four OOD distance measures. To be precise, our model furnishes a minimum of 9%, 11%, 1%, and 0.3% enhancements in mean FAR when utilizing MSP, energy, Mahalanobis, and cosine distances, respectively. This consistent improvement showcases the efficiency of our model in OOD detection across various distance measurements and datasets.

In terms of models trained on different In-Distribution (ID) datasets, our model, when applying Mahalanobis and cosine distances, achieves near-perfect OOD detection on SST2, IMDB, and TREC-10 datasets, aligning with the previous SOTA's performance with $\mathcal{L}_{\text{margin}}$. While the 20 Newsgroup dataset, comprising articles from multiple genres, offered room for improvement when tackled with the previous SOTA method, our approach managed to deliver near-perfect OOD detection on this dataset, surpassing the prior method's performance by a significant margin. It is noteworthy that our model achieves unrivaled performance in both AUROC and FAR95 on all four datasets evaluated.

Furthermore, we assess the performance of our model across all OOD scoring functions. This underscores the exceptional ability of our framework to perform optimally across a very diverse range of scoring functions to enhance the OOD performance. Comparisons between our approach and the prior SOTA (Zhou et al., 2021), as presented in Table 2 and 3, illustrate our model's performance on four different ID datasets, measured in terms of AUROC and FAR95, respectively. In comparison with the previous state-of-the-art performance, our method successfully lowers the FAR95 on 3 out of 4 distinct ID datasets, among which, falsely accepting OOD as ID of IMDB is drastically alleviated from 24.7% to 10.9%. These results also highlight that our method surpasses previous state-of-the-art AUROC on all 4 distinct ID datasets.

3.9. Detailed Comparisons

In line with (Zhou et al., 2021), we also use additional datasets exclusively as OOD data, including RTE (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), MNLI (Williams et al., 2018), WMT16 (Bojar et al., 2016), and Multi30K (Elliott et al., 2016). The comprehensive OOD detection performance on various OOD datasets is presented in Table 4 (AUC) and Table 5 (FAR). Notably, we make significant strides in improving the FAR performance by approximately 50% when employing IMDB as ID and TREC-10 as

AUROC↑ /FAR95↓		Avg	SST2	IMDB	TREC-10	20NG
w/o $\mathcal{L}_{\text{Cont}}$ (Zhou et al., 2021)	MSP	94.1/35.0	88.9/61.3	94.7/40.6	98.1/7.6	94.6/30.5
	Energy	94.0/34.7	87.7/63.2	93.9/49.5	98.0/10.4	96.5/15.8
	Maha	98.5/7.3	96.9/18.3	99.8/0.7	99.0/2.7	98.3/7.3
	Cosine	98.2/9.7	96.2/23.6	99.4/2.1	99.2/2.3	97.8/10.7
w/ $\mathcal{L}_{\text{SCl}}$ (Zhou et al., 2021)	$\mathcal{L}_{\text{scl}}$+MSP	90.4/46.3	89.7/59.9	93.5/48.6	90.2/36.4	88.1/39.2
	$\mathcal{L}_{\text{scl}}$+Energy	90.5/43.5	88.5/64.7	92.8/50.4	90.3/32.2	90.2/26.8
	$\mathcal{L}_{\text{scl}}$+Maha	98.3/10.5	96.4/26.6	99.6/2.0	99.2/1.9	97.9/11.6
	$\mathcal{L}_{\text{scl}}$+Cosine	97.7/13.0	95.9/28.2	99.2/4.2	99.0/2.4	96.8/17.0
w/ $\mathcal{L}_{\text{margin}}$ (Zhou et al., 2021)	$\mathcal{L}_{\text{margin}}$+MSP	93.0/33.7	89.7/49.2	93.9/46.3	97.6/6.5	90.9/32.6
	$\mathcal{L}_{\text{margin}}$+Energy	93.9/31.0	89.6/48.8	93.4/52.1	98.4/4.6	94.1/18.6
	$\mathcal{L}_{\text{margin}}$+Maha	99.5/1.7	99.9/0.6	100/0	99.3/0.4	98.9/6.0
	$\mathcal{L}_{\text{margin}}$+Cosine	99.0/3.8	99.6/1.7	99.9/0.2	99.0/1.5	97.4/11.8
Ours	MSP	94.8/24.7	90.8/46.4	97.0/18.3	98.6/2.5	92.9/31.4
	Energy	95.7/20.7	90.4/45.5	97.0/19.9	98.5/3.2	96.9/14.0
	Maha	99.6/0.8	99.4/2.2	99.5/0.7	99.9/0	99.1/0.8
	Cosine	99.0/3.5	98.7/6.5	98.7/4.8	99.5/0.4	98.9/2.3

Table 1: OOD Detection performance (in %) of $\text{RoBERTa}_{\text{LARGE}}$ trained on the four different ID datasets. Following the highlight standard (Zhou et al., 2021), the results of our SRL-guided Self Supervision method achieving SOTA on both evaluation metrics are highlighted in blue.

Methods	Overall	SST2	IMDB	TREC-10	20NG
(Zhou et al., 2021)	96.4	94.7	96.8	98.6	95.3
Ours	**97.3**	**94.8**	**98.1**	**99.1**	**97.0**

Table 2: Comparisons of overall AUROC↑.The results of our SRL-guided Self Supervision method achieving SOTA are highlighted in **bold**.

Methods	Overall	SST2	IMDB	TREC-10	20NG
(Zhou et al., 2021)	17.6	25.1	24.7	3.3	17.3
Ours	**12.4**	25.2	**10.9**	**1.5**	**12.1**

Table 3: Comparisons of overall FAR95↓.The results of our SRL-guided Self Supervision method achieving SOTA are highlighted in **bold**.

OOD. For other OOD datasets, our approach also yields impressive FAR enhancements ranging from 10% to 50% when using IMDB as the ID. When employing 20NG as the ID dataset, our method significantly outperforms the previous SOTA. Collectively, our technique considerably surpasses (Zhou et al., 2021), exhibiting a substantial improvement across most benchmark measures.

3.10. Ablation Studies and Other Analysis

Ablation Studies: In Table 6, we substantiate the efficacy of the various proposed components on IMDB and TREC benchmarks. Please note, the baseline method refers to a model trained using $\mathcal{L}_{\text{margin}}$ without the incorporation of SRLOOD, Transformer, and Self-supervised learning modules. The results underscore that each module imparts substantial enhancements in terms of both AUROC and FAR95 on both benchmarks. This strongly affirms the effectiveness of every component proposed.

Qualitative Analysis: In Figure 2, we present qualitative examples with IMDB functioning as the In-Distribution (ID) dataset and TREC-10 and WMT16 employed as the Out-of-Distribution (OOD) datasets. Our model exhibits superior effectiveness in identifying subtle OOD/anomalous samples, even when these contain movie content similar to the IMDB ID dataset. Conversely, the previous State-of-the-Art (SOTA) approach (Zhou et al., 2021) fails to detect the majority of these sentences.

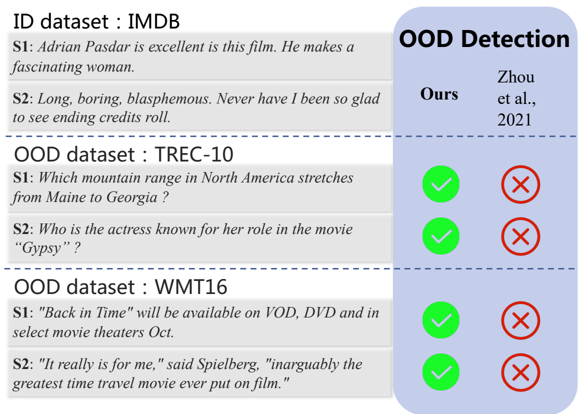

Figure 2: Qualitative examples using IMDB as the In-Distribution (ID) dataset, and TREC-10 and WMT16 as the Out-of-Distribution (OOD) datasets.

Different masking ratios:

Figure 3 illustrates three distinct masking probabilities: 0.3, 0.5, and 0.7. A higher value corresponds to stronger perturbation of the key feature representations, and we select the optimal perturbation for each ID task, guided by both the average AUROC and FAR95 over all four OOD scores, as

AUROC	SST2				IMDB				TREC-10				20NG			
	MSP	Energy	Maha	Cosine	MSP	Energy	Maha	Cosine	MSP	Energy	Maha	Cosine	MSP	Energy	Maha	Cosine
SST2	-	-	-	-	-	-	-	-	**97.8**/96.2	**97.3**/96.6	**99.8**/98.4	**99.1**/97.8	**96.5**/96.3	**99.0**/98.1	99.2/99.5	**99.0**/99.0
IMDB	-	-	-	-	-	-	-	-	**99.5**/99.3	92.0/99.7	**99.7**/99.6	**99.6**/99.3	93.6/94.5	96.7/96.9	**99.7**/99.0	**98.8**/98.4
TREC-10	**96.0**/95.1	**96.1**/94.9	**99.8**/99.5	**99.6**/99.0	**98.8**/93.8	**98.9**/93.3	99.9/100	99.8/100	-	-	-	-	**99.3**/88.0	**99.9**/92.4	99.3/99.6	**99.5**/96.5
20NG	**96.8**/95.2	**97.0**/95.0	**100**/100	99.9/100	**96.5**/95.4	**96.6**/95.3	99.8/100	98.0/99.9	**99.6**/99.2	99.7/99.8	**100**/99.8	**100**/99.7	-	-	-	-
MNLI	**83.0**/82.8	**82.8**/82.7	98.4/99.8	96.6/99.5	**95.7**/92.4	**95.6**/91.7	99.8/100	96.9/99.9	**98.0**/96.6	**98.0**/97.6	**99.8**/99.2	**99.0**/98.8	**92.1**/91.0	**96.8**/94.2	**98.9**/98.4	**99.1**/97.2
RTE	**89.4**/87.4	**88.2**/87.5	99.9/100	99.5/99.9	**96.1**/92.9	**96.0**/92.1	99.9/100	98.4/99.9	**98.7**/96.6	**98.6**/98.1	**99.9**/99.6	**99.6**/99.2	**85.5**/84.5	**92.1**/88.7	**98.7**/98.2	**98.5**/95.6
WMT16	**85.5**/83.9	**84.3**/84.0	98.9/99.9	97.3/99.4	**96.9**/92.9	**96.8**/92.2	99.9/100	99.8/99.9	**97.8**/97.1	97.8/98.0	**99.9**/99.4	**99.5**/99.1	**91.9**/88.3	**96.8**/92.5	**99.0**/98.5	**98.8**/96.7
Multi30K	**94.2**/93.5	**93.7**/93.6	99.5/100	99.3/99.9	**98.1**/95.9	**98.3**/95.7	99.9/100	99.8/100	**99.1**/97.9	**99.1**/98.9	100/99.5	**99.9**/99.3	91.6/93.7	**96.9**/96.0	99.0/99.1	**98.6**/98.3
Avg	**90.8**/89.7	**90.4**/89.6	99.4/99.9	98.7/99.6	**97.0**/93.9	**97.0**/93.4	99.9/100	98.7/99.9	**98.6**/97.6	**98.5**/98.4	**99.9**/99.3	**99.5**/99.0	**92.9**/90.9	**96.9**/94.1	**99.1**/98.9	**98.9**/97.4

Table 4: OOD detection AUROC (%) of ours and w/ $\mathcal{L}_{\text{margin}}$ (Zhou et al., 2021). The results of our SRL-guided Self Supervision method achieving SOTA on AUROC are highlighted in **bold**.

FAR95	SST2				IMDB				TREC-10				20NG			
	MSP	Energy	Maha	Cosine	MSP	Energy	Maha	Cosine	MSP	Energy	Maha	Cosine	MSP	Energy	Maha	Cosine
SST2	-	-	-	-	-	-	-	-	**5.3**/11.9	**6.8**/10.4	**0**/1.6	**0.4**/6.9	20.5/13.7	**4.3**/5.3	**0**/1.2	**0.1**/12.6
IMDB	-	-	-	-	-	-	-	-	**0**/0.5	**0**/0.2	**0**/0	**0**/0	31.6/23.6	14.5/11.4	**0.5**/4.7	**3.5**/7.4
TREC-10	**23.2**/35.3	**23.0**/35.0	**0**/2.4	**0**/4.3	**0.6**/50.0	**0.8**/54.0	**0**/0	**0**/0	-	-	-	-	**3.5**/37.2	**0**/13.8	**0**/1.4	**0**/4.4
20NG	**15.7**/36.4	**13.7**/36.3	**0**/0	**0**/0	**22.7**/37.8	**24.6**/33.1	**0**/0	6.3/0	**0**/0.6	**0**/0.2	**0**/0	**0**/0	-	-	-	-
MNLI	68.7/64.6	67.8/64.3	7.8/0.4	22.0/2.6	**32.6**/52.2	**34.8**/83.8	**0**/0.1	14.9/0.9	**3.9**/9.6	**4.5**/6.7	**0**/0.7	**1.5**/1.9	38.1/37.4	**16.9**/24.7	**0.1**/9.6	**0.3**/16.7
RTE	58.5/58.3	**57.5**/57.7	**0**/0	0.9/0.3	**29.3**/52.9	**32.6**/54.3	**0**/0	5.9/0.3	**2.8**/9.8	**3.8**/6.2	**0**/0.1	**0.2**/0.5	**51.3**/52.9	**30.6**/35.4	**2.6**/11.1	**4.5**/24.2
WMT16	68.3/64.3	67.0/64.1	4.8/0.5	15.3/3.0	**18.9**/53.7	**21.2**/55.7	**0**/0	2/0.4	**5.3**/7.9	7.2/5.7	**0**/0.5	**0.6**/1.3	**37.3**/45.3	**15.8**/27.8	**2.0**/7.5	**4.3**/18.7
Multi30K	44.0/36.3	43.8/35.4	0.2/0	1/0.3	**5.6**/30.9	**5.7**/31.9	**0**/0	**0**/0	**0.3**/5.3	**0.3**/2.6	**0**/0	**0**/0.2	37.9/27.8	16.2/12.0	**0.7**/6.9	**3.9**/8.7
Avg	**46.4**/49.2	**45.5**/48.8	2.2/0.6	6.5/1.7	**18.3**/46.3	**19.9**/52.1	**0**/0	4.8/0.2	**2.5**/6.5	**3.2**/4.6	**0**/0.4	**0.4**/1.5	**31.4**/32.6	**14.0**/18.6	**0.8**/6.0	**2.3**/11.8

Table 5: OOD detection FAR95 (%) of ours and w/ $\mathcal{L}_{\text{margin}}$ (Zhou et al., 2021). The results of our SRL-guided Self Supervision method achieving SOTA on FAR95 are highlighted in **bold**.

Baseline	SRL	SSL	IMDB (AUROC↑/FAR95↓)				TREC (AUROC↑/FAR95↓)			
			MSP	Energy	Maha	Cosine	MSP	Energy	Maha	Cosine
✓			93.9/46.3	93.4/52.1	**1/0**	**99.9/0.2**	97.6/6.5	98.1/4.6	99.3/0.7	99.0/1.5
✓	✓		95.6/25.4	95.6/25.5	99.7/0.4	99.6/1.3	97.6/6.9	97.6/5.1	99.2/0.6	99.3/0.7
✓	✓	✓	**97.0/18.3**	**97.0/19.9**	**99.9/0**	98.7/4.8	**98.6/2.5**	**98.2/3.2**	**99.9/0**	**99.5/0.4**

Table 6: Ablation study of our method on IMDB and TREC.

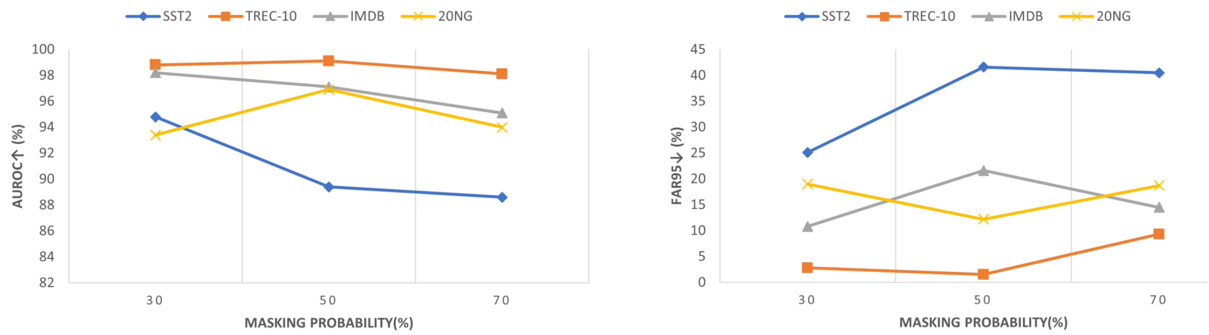

Figure 3: Performance on different masking probabilities on four benchmark datasets.

depicted in Figure 3. The results indicate that a lower masking ratio, specifically 30%, may yield superior performance due to its ability to retain most of the original information. On the other hand, more intense perturbation could potentially lead to a loss of crucial features.

4. Conclusion

In this paper, we propose a simple yet effective approach called Semantic Role Labeling Guided Out-of-distribution Detection (SRLOOD), which learns from both global and SRL-guided local fine-grained feature representation to detect OOD instances in NLP. The model jointly optimizes both global and local representations using a margin-based contrastive loss and self-supervised loss. The resulting model is able to effectively extract the key semantic roles and eliminate outliers from a sentence to detect subtle OOD samples effectively. The resulting model shows State-of-the-Art performance on four different OOD benchmarks with four different OOD scoring functions, indicating the effectiveness of our proposed SRLOOD framework.

5. Bibliographical References

Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. 2019. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 481–490.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 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.

Senqi Cao and Zhongfei Zhang. 2022. Deep hybrid models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743.

Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. 2021a. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16856.

Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. 2020. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference on Computer Vision*, pages 572–588. Springer.

Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. 2021b. Deep one-class classification via interpolated gaussian descriptor. *arXiv preprint arXiv:2101.10043*.

Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. 2022. Deep one-class classification via interpolated gaussian descriptor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 383–392.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

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

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. 2018. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31.

Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. 2022. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19217–19227.

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, Berlin, Germany. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*.

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, Online. Association for Computational Linguistics.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2019a. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019b. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32.

Rui Huang and Yixuan Li. 2021. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8710–8719.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. 2020. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33:20578–20589.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*.

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

Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. 2021. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15313–15323.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc.

Andrew L. 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, Portland, Oregon, USA. Association for Computational Linguistics.

Peyman Morteza and Yixuan Li. 2022. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34(2):145–159.

Mrinal Rawat, Ramya Hebbalaguppe, and Lovekesh Vig. 2021. Pnpood : Out-of-distribution detection for text classification via plug andplay data augmentation. *CoRR*, abs/2111.00506.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.

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.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3566–3572, Hong Kong, China. Association for Computational Linguistics.

Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. 2023. Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. *Medical image analysis*, 90:102930.

Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. 2021. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. *arXiv preprint arXiv:2111.12264*.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*.

Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. 2021. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9452–9461.

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

Yibo Zhou. 2022. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7379–7387.

Ev Zisselman and Aviv Tamar. 2020. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13994–14003.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*.