Few-Shot Out of Domain Intent Detection with Covariance Corrected Mahalanobis Distance

Jayasimha Talur, Oleg Smirnov, Paul Missault

Amazon
{talurj, osmirnov, pmissaul}@amazon.com

Abstract

Conversational agents like chat bots and voice assistants are trained to understand and respond to user intents. On encountering an utterance with an intent different from the ones they have been trained on, these agents are expected to classify the intent as "unknown" or "out of domain". This problem is known as out of domain (OOD) intent detection. Podolskiy et al. (2021), showed that Mahalanobis distance can be used effectively for identifying OOD intents, outperforming competing approaches. However, their method fails to outperform the baselines in the practically important few-shot setting. In this paper we analyze the reason for low performance and propose a covariance corrected Mahalanobis distance for detecting out-of-domain intents.

1 Introduction

Intent classification is a key component of natural language understanding systems, such as voice assistants and chat bots. Recent advances in those systems are overwhelmingly contributed by deep learning techniques, that can learn meaningful feature representations with a minimum amount of hand-crafting (Chen et al. 2017). Chat bots typically follow the intent-response pattern, where there is a fixed or context-aware mapping between predicted intents and the responses. In addition to providing the confidence score for intent, intent detection models are also expected to produce an OOD score, that measures the likelihood of an utterance being out-of-domain. Typically, when an utterance is deemed to OOD, a fallback mechanism is triggered to either ask a clarifying question or respond with "I don't know". OOD detection can be modeled as a binary classification task, where we are interested in classifying the utterance in out-of-domain and in-domain (IND) categories. For good user experience and user trust it is important to achieve a strong trade-off between precision and recall of the OOD classifier.

Recently many methods have been proposed to detect OOD intents (Podolskiy et al. 2021; Chen and Yu 2021; Rawat, Hebbalaguppe, and Vig 2021). However, these methods require access to a large training dataset, either with out-of-domain examples, or a large unlabeled corpus. In industrial settings it is unreasonable to expect any of those assumptions to be met. For example, when bootstrapping a domain-specific chat bot, there is no access to a large training dataset of out-of-domain utterances, since the chat bot has not been in production yet. It is also difficult to obtain a large corpus of unlabeled *domain-specific* examples.

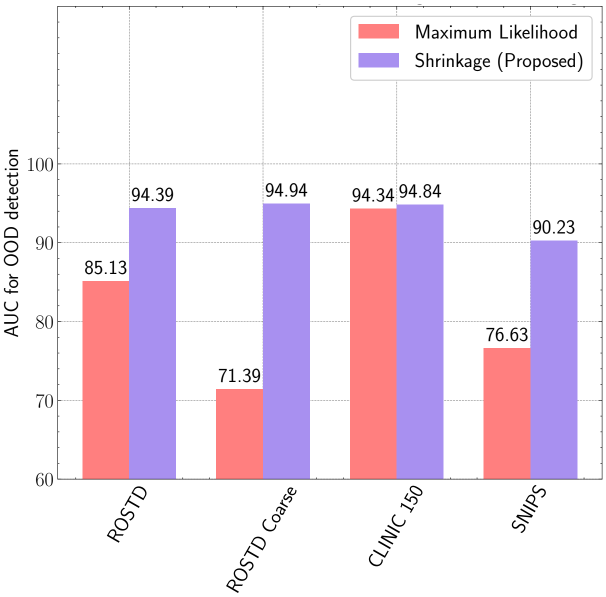

Figure 1: Performance of OOD detection for various intent classification datasets when Mahalanobis distance is calculated using Shrinkage (our proposal) and Maximum Likelihood estimators in 5-shot setting.

Specifically, for conversational agents provided as a service, such as Amazon Lex and Google Dialogflow, customers can extend the agent's capabilities by uploading custom utterances and intents datasets. In those settings both the training and validation data are scarce.

Podolskiy et al. (2021) showed that Mahalanobis distance computed on RoBERTa (Liu et al. 2019) embeddings outperforms baseline methods for OOD detection without using any additional data. Unfortunately, Mahalanobis distance performs poorly in low resource settings (Tajwar et al. 2021). In this paper, we identify the reason for the poor performance and propose a new method, that performs OOD intent detection in low resource settings.

2 Preliminaries and Related Work

The OOD detection task is to classify a test data point into OOD and IND categories. We can broadly classify the approaches into two buckets:

- Data-centric: these methods use additional OOD data to learn representation that can better separate OOD from IND examples. Additional OOD data is obtained by either sampling from a large corpus (Hendrycks, Mazeika, and Dietterich 2019), by using a language model to generate sentences (Rawat, Hebbalaguppe, and Vig 2021), or by mining or filtering examples using sentence similarity models (Chen and Yu 2021).
- Score-based: these methods compute a *score* to decide between the IND and OOD classes. The score can be computed from the features (Lee et al. 2018), model logits (Liu et al. 2020; Liang, Li, and Srikant 2017), or the norm in the gradients space (Huang, Geng, and Li 2021).

In this work we focus on score-based methods, which don't require additional OOD data that makes them attractive for industrial applications. In the score-based methodology, given a test sample x and a decision threshold T, we are interested in constructing a score function $G : x \to \mathbb{R}$, such that $G(x) >= T$ implies that x is OOD and $G(x) < T$ implies that x is IND.

Mahalanobis distance: Let $F \in \mathbb{R}^{n \times d}$ denote n points, each represented by d dimensional features, and $y \in [1, C]$ the corresponding labels in the set of C classes. For a test feature $x \in R^{d \times 1}$, the Mahalanobis distance for OOD detection is defined by

$$d(x) = \min_{c \in [1,C]} (x - \mu_c)^T \Sigma^{-1} (x - \mu_c) \quad (1)$$

where $\mu_c \in R^{d \times 1}$ is the empirical mean of features of the corresponding class, and $\Sigma \in R^{d \times d}$ is the feature covariance matrix. Mahalanobis-based OOD detection method uses a score function $G(x) = d(x)$.

Besides OOD detection, Mahalanobis distance has been used to perform pattern recognition (De Maesschalck, Jouan-Rimbaud, and Massart 2000), anomaly detection (Zhang et al. 2015) and detecting adversarial examples (Lee et al. 2018). Mahalanobis distance is known to performs well for sufficiently large dataset sizes. However, its performance degrades rapidly in low resource settings (Tajwar et al. 2021).

To the best of our knowledge, there are no score-based methods specifically designed for few-shot out-of-domain intent detection.

3 Methodology

Before describing our method, we analyze why Mahalanobis distance performs worse when the training dataset size is small. Tajwar et al. (2021) conjectured that poor covariance estimate in low sample settings leads to bad estimate of Mahalanobis distance. The reasoning is that the rank of a $d \times d$ covariance matrix computed on n points in d dimensions is bounded from above by $\min(n - 1, d)$, since $n \ll d$ in few-shot settings, the covariance matrix becomes singular. Mahalanobis distance uses the inverse of covariance, hence estimating a “best fit” solution with pseudoinverse calculation may negatively affect the performance.

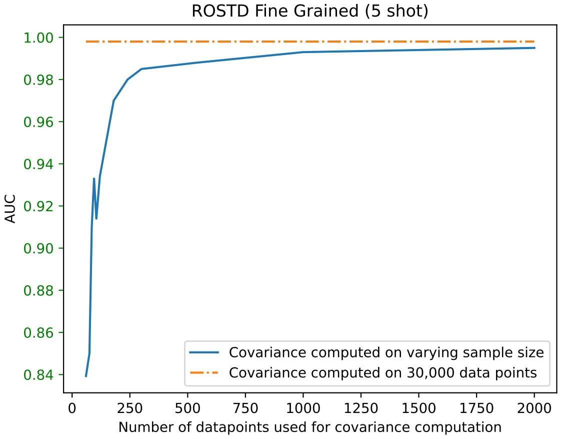

Figure 2: Performance of Mahalanobis-based OOD detection on the ROSTD dataset in 5-shot settings, as a function of the number of data points used for covariance computation. There is a sharp improvement in AUC with up to 400 data points, and only a modest improvement afterwards.

To test this hypotheses, we perform a simple experiment:

1. Fine-tune the RoBERTa (Liu et al. 2019) base model on the ROSTD intent classification dataset in a 5-shot setting. The ROSTD dataset is discussed in Section 4.2, and the fine-tuning procedure is described in Section 4.1.
2. Extract features from the last hidden layer for the full training dataset (∼30K data points). Note that this full dataset is unavailable in practical cases, since we only observe 5 data points.
3. Compute the covariance matrices using various sample sizes of training features extracted in the previous step.
4. Use the Mahalanobis method for detecting OOD samples.

Table 1 summarizes the performance of OOD detection with respect to the metrics discussed in Section 4.2, where the covariance matrix is computed in small and large data regimes. We observe that the Mahalanobis distance-based OOD method trained with only 5 samples per class performs poorly when the covariance is computed using 60 data points. However, it achieves competitive performance provided that the covariance matrix was computed using 30K data points. As expected, the best performance is obtained when the covariance computation and model training was performed on the full dataset.

Figure 2 depicts the OOD detection performance, when the number of data points used for covariance computation is varied. We observe a sharp improvement in AUC for up to 400 data points, and only a modest improvement afterwards. This experiment empirically confirms two assumptions. Firstly, the features extracted from a model fine-tuned

Training mode	Covariance data points	AUC	PR ROC$_{ood\_neg}$	PR ROC $_{ood\_pos}$	FPR$_{ood\_neg}$	FPR$_{ood\_pos}$
5-shot	60	84.98±4.15	94.75±1.53	58.55±8.68	75.42±10.41	37.95±7.57
5-shot	30K	98.92±1.43	99.63±0.49	96.62±4.68	4.92±9.87	4.07±5.06
Full	30K	99.8±0.1	-	99.5±0.3	1.0±0.5	0.5±0.4

Table 1: Mahalanobis OOD detection performance on the ROSTD dataset with respect to the number of examples for covariance estimation and training regimes. Performance figures for the full dataset are taken from Podolskiy et al. (2021).

on only a handful of samples still have sufficient representational power to separate IND and OOD categories. Secondly, a non-invertible covariance matrix contributes to the poor performance in few-shot setting.

To overcome this, we propose to use robust covariance estimators for covariance computation. This is motivated by the fact that robust approaches provide a better estimate of the covariance matrix when $n \ll d$, compared to the baseline Maximum Likelihood estimator (MLE) in the standard Mahalanobis distance. Intuitively, this is achieved by incorporating various prior beliefs about the structure of the features space (e.g. shape of the clusters). However, in practice different assumptions lead to differences in the downstream performance. Below, we briefly review covariance estimators with their corresponding closed-form expressions listed in Table 2.

Maximum Likelihood estimator: MLE Σ_m is conventionally used for computing the Mahalanobis distance. When Σ_m is not invertible, Σ_m^{-1} can be estimated with a pseudoinverse.

Van Ness estimator: the diagonal elements of a covariance matrix represent the variance of individual features, that is typically non-zero for all elements. Van Ness estimator (Ness 1980) only retains the diagonal elements $\Sigma_{\text{Van Ness}}$ and sets non-diagonal elements to zero.

Shrinkage estimator: Shrinkage methods perform a convex combination of a singular matrix Σ_m and some stable *target* matrix. The Shrinkage estimator (Friedman 1989) $\Sigma_{\text{Shrinkage}}$ employs a diagonal target matrix, where elements on the diagonal equal to the mean of Σ_m eigenvalues.

Ledoit-Wolf estimator: Ledoit and Wolf (2004) proposed a method to compute the shrinkage coefficient $\hat{\alpha}$, that minimizes the expected mean square error between $\Sigma_{\text{Shrinkage}}$ and the unobserved true covariance matrix Σ^*. We refer the interested reader to Ledoit and Wolf (2004, 2003) for the exact expression for $\hat{\alpha}$ and its derivation.

Estimator	Formula
Maximum Likelihood	$\Sigma_m = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$
Van Ness	$\Sigma_{\text{Van Ness}} = \beta \operatorname{diag}(\Sigma_m)$
Shrinkage	$\Sigma_{\text{Shrinkage}} = (1-\alpha)\Sigma_m + \alpha \frac{Tr(\Sigma_m)}{d} \mathbb{I}$
Ledoit-Wolf	$\Sigma_{\text{Ledoit-Wolf}} = (1-\hat{\alpha})\Sigma_m + \hat{\alpha} \frac{Tr(\Sigma_m)}{d} \mathbb{I}$

Table 2: Robust covariance estimators. $\beta \in R$ and $\alpha \in (0, 1)$ are the hyper-parameters of these estimators.

4 Experiments

Low sample covariance matrix estimators $\Sigma_{\text{Van Ness}}$, $\Sigma_{\text{Shrinkage}}$, and $\Sigma_{\text{Ledoit-Wolf}}$ can be used as drop-in replacements for MLE Σ_{m} in the Mahalanobis distance. We compare Mahalanobis distance-based OOD detection with those 4 alternative covariance estimation methods. Additionally, we benchmark the candidate methods against the energy-based OOD detection (Liu et al. 2020), and the gradient norm approach (Huang, Geng, and Li 2021). Finally, we compare with the Maximum Softmax Probability (MSP) approach (Hendrycks and Gimpel 2016), which was shown to be a strong baseline for OOD detection.

4.1 Training Procedure

In all our experiments, we fine-tune the RoBERTa (Liu et al. 2019) base model for intent classification using the cross-entropy loss. We start from model weights which were pretrained on five English-language corpora of varying sizes and domains, totaling over 160GB of uncompressed text.

Fine-tuning was performed using the AdamW (Loshchilov and Hutter 2018) optimizer with learning rate of $2e^{-5}$ and linear decay for 45 epochs. We repeat all experiments 20 times, sampling a new training set at each iteration. For hyperparameters, we use a fixed shrinkage coefficient $\alpha = 0.1$, and a fixed Van Ness hyperparameter $\beta = 1.0$ in all the experiments. As suggested in Liu et al. (2020); Huang, Geng, and Li (2021), we set the temperature $\tau = 1$ for the energy and gradient norm baselines.

During X-shot training, covariance matrix of size $d \times d$ is computed from a data matrix of size $(X \cdot N_c) \times d$, where $d = 768$ for RoBERTa embeddings and N_c is the number of classes. On each iteration of the experiment, 1) we randomly sample X data points per class from the training set, and 2) compute covariance matrix using only $X \cdot N_c$ data points using corresponding estimators.

4.2 Datasets and Metrics

We evaluate our approach on the following datasets:

1. **CLINC150** (Larson et al. 2019): was proposed to evaluate the performance of task-oriented dialogue systems on out-of-domain queries. The dataset contains 150 intents spanning over 10 domains.
2. **ROSTD** (Schuster et al. 2018): was developed to test cross-lingual transfer learning for multilingual task oriented dialog. The dataset was later extended by Gangal et al. (2019) by adding OOD intents to English language utterances.

Dataset (**5-shot**)	OOD method	AUC ↑	PR ROC $_{ood\_pos}$ ↑	FPR $_{ood\_pos}$ ↓
ROSTD	MLE	85.13±5.41	58.49±10.74	36.16±9.47
	Van Ness	93.87±3.41	81.70±9.06	20.45±9.70
	Shrinkage	**94.39±2.88**	**83.32±7.79**	**19.25±9.09**
	Ledoit-wolf	94.33±2.95	83.19±7.90	19.40±9.14
	Grad Norm	93.47±3.32	81.04±8.76	22.68±10.06
	Energy	93.92±3.04	81.20±9.13	20.01±9.03
	MSP	92.13±3.41	78.23±8.34	25.69±9.07
SNIPS	MLE	76.63±9.47	49.68±13.00	56.53±11.65
	Van Ness	**90.46±3.31**	**73.98±8.21**	**29.77±8.35**
	Shrinkage	90.23±3.33	73.35±8.17	30.00±8.40
	Ledoit-wolf	90.23±3.34	73.38±8.18	29.91±8.42
	Grad Norm	88.87±4.76	72.04±10.92	37.29±15.39
	Energy	88.92±5.74	70.32±11.53	33.30±14.38
	MSP	89.27±4.05	71.71±8.46	33.14±11.13
ROSTD Coarse	MLE	71.39±8.70	40.68±9.97	65.87±12.45
	Van Ness	94.79±2.71	82.83±8.78	16.06±6.93
	Shrinkage	**94.94±2.85**	**83.85±8.76**	15.92±6.88
	Ledoit-wolf	94.93±2.85	83.83±8.78	**15.91±6.84**
	Grad Norm	92.38±3.59	77.69±9.90	23.84±10.32
	Energy	93.43±3.02	78.65±9.55	19.66±7.89
	MSP	92.96±2.78	77.05±8.92	19.54±6.46
Clinic 150	MLE	94.34±0.45	78.23±2.07	23.05±2.02
	Van Ness	94.42±0.39	79.11±1.60	23.08±2.00
	Shrinkage	94.84±0.41	81.44±1.83	21.90±1.88
	Ledoit-wolf	94.76±0.42	81.07±1.78	22.29±1.93
	Grad Norm	94.94±0.40	81.48±1.67	**21.67±2.21**
	Energy	**94.95±0.38**	**81.56±1.67**	21.85±2.23
	MSP	94.08±0.42	78.41±1.64	25.21±1.88

Dataset (**10-shot**)	OOD method	AUC ↑	PR ROC $_{ood\_pos}$ ↑	FPR $_{ood\_pos}$ ↓
ROSTD FINE	MLE	94.44±2.45	80.30±8.35	14.98±5.45
	Van Ness	96.79±1.43	89.04±5.33	10.26±3.25
	Shrinkage	**97.37±1.23**	**90.85±4.94**	**8.52±2.96**
	Ledoit-wolf	97.27±1.27	90.51±5.05	8.78±3.02
	Grad Norm	96.17±1.42	88.14±4.42	13.33±4.15
	Energy	96.98±1.23	89.83±4.49	10.10±3.47
	MSP	95.32±1.55	84.32±5.49	13.39±3.15
SNIPS	MLE	85.27±9.24	62.58±13.87	40.23±21.25
	Van Ness	90.57±4.45	74.35±10.99	28.81±9.62
	Shrinkage	**90.68±4.51**	**74.79±11.21**	**28.44±9.77**
	Ledoit-wolf	90.67±4.51	74.70±11.19	28.48±9.49
	Grad Norm	85.73±7.27	68.93±12.46	51.42±19.06
	Energy	89.75±5.48	73.57±12.16	32.89±15.11
	MSP	89.27±4.61	70.81±10.67	32.63±13.17
ROSTD COARSE	MLE	87.23±8.20	64.96±11.07	35.16±20.79
	Van Ness	95.99±1.87	85.98±6.21	11.40±4.76
	Shrinkage	**96.48±1.82**	**87.96±5.92**	**10.26±4.83**
	Ledoit-wolf	96.45±1.83	87.86±5.92	10.34±4.83
	Grad Norm	93.38±3.02	81.31±7.46	23.97±13.61
	Energy	95.35±2.07	84.01±8.36	13.99±5.55
	MSP	94.58±2.21	81.00±7.22	14.53±5.27
CLINIC 150	MLE	**96.07±0.25**	85.54±1.20	**18.03±1.10**
	Van Ness	95.73±0.26	83.96±1.10	18.94±0.96
	Shrinkage	96.05±0.24	**85.55±1.11**	18.12±1.04
	Ledoit-wolf	96.00±0.24	85.37±1.07	18.30±1.05
	Grad Norm	95.81±0.27	85.30±1.01	18.70±1.19
	Energy	95.99±0.24	85.34±1.03	18.54±1.09
	MSP	95.25±0.28	82.59±1.09	20.05±1.19

Table 3: Comparison of covariance-corrected Mahalanobis distance methods for OOD detection with the baselines in **5-shot** and **10-shot** settings.

3. **ROSTD-COARSE**: following Podolskiy et al. (2021), we also experiment with a coarsened version of ROSTD with only 3 intent classes. We use the same set of OOD intents for testing for both fine-grained and coarsened version of the dataset.
4. **SNIPS** (Coucke et al. 2018): contains 7 intents with approximately 2000 utterances per intent. Since the dataset does not provide IND and OOD split, we follow the protocol from Podolskiy et al. (2021) by randomly taking 5 intents as IND and the remaining 2 intents as OOD. We sample different OOD and IND intents on each iteration of the experiment.

Since OOD detection is framed is a binary classification problem, we evaluate performance in terms of AUC, PR ROC, and FPR@95%TPR metrics. For FPR, the decision threshold is chosen so that the True Positive Rate is 95%. We omit 95% suffix from the metric names to avoid repetition in notation. We report measurements for the cases when the OOD class is treated as positive, and when the OOD class is treated as negative. This is indicated with ood_pos and ood_neg suffixes correspondingly.

5 Results

Table 3 compares the performance of different variants of the Mahalanobis distance on benchmarking datasets. We observe that covariance corrected methods (Shrinkage, Ledoit-Wolf and Van Ness) outperform other OOD detection techniques on 3 out of 4 datasets on all metrics of interest. The table is summarized in Figure 1 for AUC using the MLE and Shrinkage estimators.

Furthermore, in 5-shot setting the covariance correction outperforms MLE on all datasets. However, MLE performs competitively with covariance correction when the number of samples used for covariance estimation increases. This phenomenon can be seen for the CLINC150 dataset in 10-shot setting, where the 768 × 768 dimensional covariance matrix was computed by using $150 \cdot 10 = 1500$ samples (150 classes and 10 samples per class).

To assess the impact of OOD performance on the number of intent classes, we train several models by sampling subset of intents from the CLINC150 dataset. In Figure 3 we observe that Shrinkage estimator outperforms MLE by a large margin, when the number of intents is small and it converges to MLE performance as the number of classes increase. The effectiveness of OOD detection degrades as the number of in-domain (IND) classes increases, which is yet another interesting phenomenon. The reason is that, when the number of OOD samples are fixed, an increase in the size of IND samples leads to greater confusion with OOD class, because more samples end up on the IND/OOD boundary. Quantitatively, going from 3 to 20 IND classes causes the minimum distance from a query OOD example to the closest IND centroid decreases by 20%, this confuses the methods that rely on a fixed score threshold. When all 150 classes are used then, the average IND-OOD Euclidean distance decreases by 37%. Moreover, as the amount of data increases, the performance of the robust estimators degrades, but at different rates.

6 Conclusion

We have demonstrated that in few-shot settings Mahalanobis distance computed using robust covariance estimators consistently outperforms the Maximum Likelihood estimator baseline. According to our experiments, the Shrinkage estimator excels in 5-shot and 10-shot settings across various datasets. The suggested approach is computationally cheap,

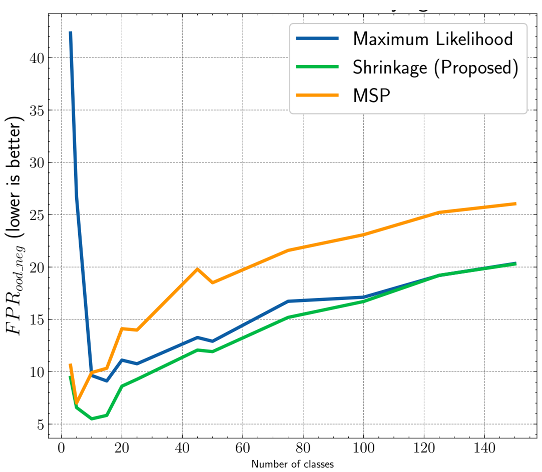

Figure 3: Performance on the CLINC150 dataset with respect to the number of training classes in **10-shot** setting.

with an additional overhead of one matrix-vector multiplication operation per class, and does not require any auxiliary data nor modifications to the training procedure.

References

Chen, D.; and Yu, Z. 2021. GOLD: Improving Out-of-Scope Detection in Dialogues using Data Augmentation. *CoRR*, abs/2109.03079.

Chen, H.; Liu, X.; Yin, D.; and Tang, J. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explor. Newsl.*, 19(2): 25–35.

Coucke, A.; Saade, A.; Ball, A.; Bluche, T.; Caulier, A.; Leroy, D.; Doumouro, C.; Gisselbrecht, T.; Caltagirone, F.; Lavril, T.; et al. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 12–16.

De Maesschalck, R.; Jouan-Rimbaud, D.; and Massart, D. L. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1): 1–18.

Friedman, J. H. 1989. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84: 165–175.

Gangal, V.; Arora, A.; Einolghozati, A.; and Gupta, S. 2019. Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection In Task Oriented Dialog. *CoRR*, abs/1912.12800.

Hendrycks, D.; and Gimpel, K. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *CoRR*, abs/1610.02136.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. G. 2019. Deep Anomaly Detection with Outlier Exposure. *ArXiv*, abs/1812.04606.

Huang, R.; Geng, A.; and Li, Y. 2021. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. *CoRR*, abs/2110.00218.

Larson, S.; Mahendran, A.; Peper, J. J.; Clarke, C.; Lee, A.; Hill, P.; Kummerfeld, J. K.; Leach, K.; Laurenzano, M. A.; Tang, L.; and Mars, J. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. 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)*, 1311–1316. Hong Kong, China: Association for Computational Linguistics.

Ledoit, O.; and Wolf, M. 2003. Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management*, 30.

Ledoit, O.; and Wolf, M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88: 365–411.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 7167–7177. Red Hook, NY, USA: Curran Associates Inc.

Liang, S.; Li, Y.; and Srikant, R. 2017. Principled Detection of Out-of-Distribution Examples in Neural Networks. *CoRR*, abs/1706.02690.

Liu, W.; Wang, X.; Owens, J. D.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. *CoRR*, abs/2010.03759.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ness, J. V. 1980. On the dominance of non-parametric Bayes rule discriminant algorithms in high dimensions. *Pattern Recognit.*, 12(6): 355–368.

Podolskiy, A. V.; Lipin, D.; Bout, A.; Artemova, E.; and Piontkovskaya, I. 2021. Revisiting Mahalanobis Distance for Transformer-Based Out-of-Domain Detection. In *AAAI*.

Rawat, M.; Hebbalaguppe, R.; and Vig, L. 2021. PnPOOD : Out-Of-Distribution Detection for Text Classification via Plug andPlay Data Augmentation. *CoRR*, abs/2111.00506.

Schuster, S.; Gupta, S.; Shah, R.; and Lewis, M. 2018. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. *CoRR*, abs/1810.13327.

Tajwar, F.; Kumar, A.; Xie, S. M.; and Liang, P. 2021. No True State-of-the-Art? OOD Detection Methods are Inconsistent across Datasets. *CoRR*, abs/2109.05554.

Zhang, Y.; Du, B.; Zhang, L.; and Wang, S. 2015. A low-rank and sparse matrix decomposition-based Mahalanobis distance method for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3): 1376–1389.