A Deep Generative Distance-Based Classifier for Out-of-Domain Detection with Mahalanobis Space

Hong Xu, Keqing He, Yuanmeng Yan, Sihong Liu, Zijun Liu, Weiran Xu*
Beijing University of Posts and Telecommunications, Beijing, China
{xuhong,kqin,yanyuanmeng,liusihong,liuzijun,xuweiran}@bupt.edu.cn

Abstract

Detecting out-of-domain (OOD) input intents is critical in the task-oriented dialog system. Different from most existing methods that rely heavily on manually labeled OOD samples, we focus on the unsupervised OOD detection scenario where there are no labeled OOD samples except for labeled in-domain data. In this paper, we propose a simple but strong generative distance-based classifier to detect OOD samples. We estimate the class-conditional distribution on feature spaces of DNNs via Gaussian discriminant analysis (GDA) to avoid over-confidence problems. And we use two distance functions, Euclidean and Mahalanobis distances, to measure the confidence score of whether a test sample belongs to OOD. Experiments on four benchmark datasets show that our method can consistently outperform the baselines.

1 Introduction

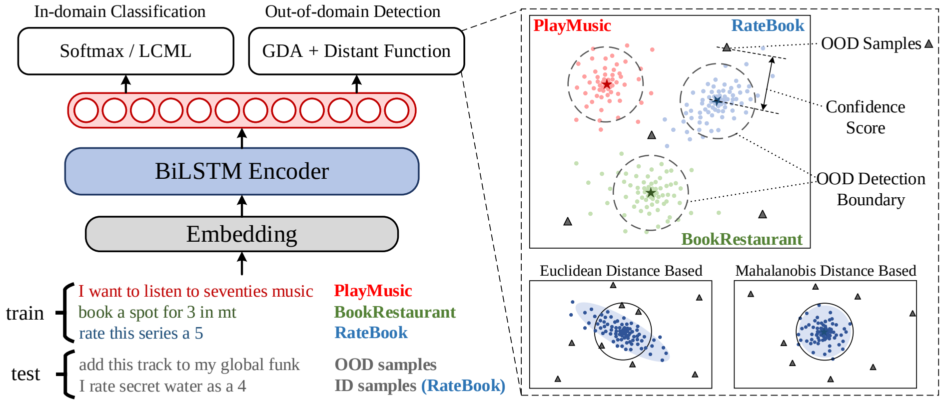
Task-oriented dialog systems such as Google's DialogFlow or Amazon's Lex have become omnipresent to let people interact with machines using natural language. Detecting unknown or OOD (Out-of-Domain) intents from user queries is an essential component that aims to know when a query falls outside their range of predefined supported intents. Correctly identifying out-of-scope cases is thus crucial in deployed systems—both to avoid performing the wrong action and also to identify potential future directions for development. Different from traditional text classification tasks, the exact number of unknown intents in practical scenarios is hard to estimate by domain experts and lack of real OOD examples always leads to poor prior knowledge about these unknown intents. These characteristics make it challenging to identify OOD samples in the task-oriented dialog system.

We classify the existing methods of detecting OOD intents into two main categories, supervised and unsupervised OOD detection. Supervised OOD detection (Scheirer et al., 2013; Fei and Liu, 2016; Kim and Kim, 2018; Larson et al., 2019; He et al., 2020c; Zheng et al., 2020) represents that there are extensive labeled OOD samples in the training data while unsupervised OOD detection (Breunig et al., 2000; Bendale and Boult, 2016; Hendrycks and Gimpel, 2017; Shu et al., 2017; Lee et al., 2018; Ren et al., 2019; Lin and Xu, 2019) means few or no labeled OOD samples except for labeled in-domain data. Unsupervised OOD detection makes it more complicated to identify unknown intents due to unseen and diverse semantic expressions. Besides, in the practical scenario, collecting large-scale OOD data is usually difficult and expensive compared to in-domain predefined intent data, especially when dealing with the rapidly evolving open-world environment. In this paper, we focus on the latter OOD setting, unsupervised OOD detection.

For supervised OOD detection, classical methods such as (Fei and Liu, 2016; Larson et al., 2019), form a $(m+1)$-class classification problem where the $(m+1)$-th class represents the unseen intents. Further, Zheng et al. (2020) uses labeled OOD data to generate an entropy regularization term to enforce the predicted distribution of OOD inputs closer to the uniform distribution. The critical drawback is that collecting OOD data is usually labor-intensive and they do not guarantee similar semantics. Typically,

*The corresponding author.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1452–1460
Barcelona, Spain (Online), December 8-13, 2020



a) The overall architecture b) Generative distance-based OOD detector

Figure 1: The architecture of our proposed method. We first extract intent representation by a intent classifier pre-trained on in-domain data. Then, we use distance functions in latent space to detect unknown intents for test.

these OOD samples should barely be classified to the same intent class for their unconstrained expressions. For unsupervised OOD detection, (Hendrycks and Gimpel, 2017; Shu et al., 2017) simply use a threshold on the in-domain classifier's probability estimate. (Ren et al., 2019; Gangal et al., 2020) proposes a likelihood ratio method to effectively correct for confounding background statistics. Lin and Xu (2019) employs an unsupervised density-based novelty detection algorithm, local outlier factor (LOF) to detect unseen intents. However, such deep neural networks with the softmax classifier are known to produce highly overconfident posterior distributions even for such abnormal OOD samples (Guo et al., 2017; Liang et al., 2017; Liang et al., 2018).

In this paper, we propose a simple generative distance-based classifier to detect OOD samples. Our method can be applied to any pre-trained softmax neural classifier without re-training and avoid the issue of overconfident predictions. Specifically, we first estimate the class-conditional distribution on feature spaces of DNNs via Gaussian discriminant analysis (GDA). Then we use two distance functions, Euclidean and Mahalanobis distances, to measure the confidence score of whether a test sample belongs to OOD. Our contributions are three-fold: (1) We propose a generative distance-based classifier for OOD detection. (2) Apart from traditional Euclidean distance, we introduce Mahalanobis distance under GDA to takes into account the correlations between features and get better results. (3) Experiments conducted on four benchmark OOD datasets show the effectiveness of the proposed method.

2 Approach

Fig 1(a) shows the overall architecture of our proposed method. We first train an intent classifier on in-domain data and then use a generative distance-based OOD detector to identify unknown intents for test. Fig 1(b) shows the difference between Euclidean and Mahalanobis distances in latent space.

2.1 Neural Intent Classifier

For a fair comparison, we adopt the same network architecture BiLSTM (Mesnil et al., 2015; Liu and Lane, 2016; Weiran and Chunyun, 2016; Goo et al., 2018; Xu et al., 2019; Haihong et al., 2019; Xu et al., 2020; He et al., 2020b; He et al., 2020a) as (Lin and Xu, 2019). We train the BiLSTM on the in-domain data and employ the pre-trained classifier as a feature extractor. Specifically, we also replace the softmax loss of BiLSTM with LMCL (Nalisnick et al., 2019) which transforms softmax loss into cosine loss by applying L2 normalization on both features and weights:

$$\mathcal{L}_{LMC} = \frac{1}{N}\sum_i -\log \frac{e^{s\cdot\left(\cos\left(\theta_{y_i,i}\right)-m\right)}}{e^{s\cdot\left(\cos\left(\theta_{y_i,i}\right)-m\right)}+\sum_{j\neq y_i} e^{s\cdot\cos\theta_{j,i}}} \tag{1}$$

where N denotes the number of training samples, y_i is the ground-truth class of the i-th sample, s is the scaling factor, m is the cosine margin. For test, we simply use the pre-trained model to extract intent representations without re-training like supervised OOD detection.

2.2 Generative Distance-Based OOD Detector

In this section, we first describe the basic concept of GDA as a representative discriminative classifier. Then we dive into the details of how to estimate the class-conditional distribution on feature spaces of the pre-trained neural intent classifier via GDA. Finally, we introduce two distance measurements, Euclidean and Mahalanobis distances, to measure the confidence score of whether a test sample belongs to OOD.

In contrast to the discriminative softmax classifier, the generative classifier (Murphy, 2012) defines the class-conditional distribution $P(x|y)$ and class prior $P(y)$ in order to indirectly define the posterior distribution by specifying the joint distribution $P(x, y) = P(y)P(x|y)$. GDA is a popular generative classifier by assuming that the class conditional distribution follows the multivariate Gaussian distribution and the class prior follows Bernoulli distribution: $P(\mathbf{x} \mid y = c) = \mathcal{N}(\mathbf{x} \mid \mu_c, \mathbf{\Sigma}_c), P(y = c) = \frac{\beta_c}{\sum_{c'} \beta_{c'}}$, where μ_c and Σ_c are the mean and covariance of multivariate Gaussian distribution, and β_c is the unnormalized prior for class c.

Let $x \in X$ be an input and $y \in Y = 1, ..., C$ be its label. Section 2.1 gives a pre-trained softmax neural classifier: $P(y = c \mid \mathbf{x}) = \frac{\exp\left(\mathbf{w}_c^\top f(\mathbf{x})+b_c\right)}{\sum_{c'} \exp\left(\mathbf{w}_{c'}^\top f(\mathbf{x})+b_{c'}\right)}$, where $\mathbf{w}_c$ and b_c are the weight and the bias of the softmax classifier for class c, and $f(\Delta)$ denotes the output of the penultimate layer of DNNs. Assuming that a class-conditional distribution follows the multivariate Gaussian distribution, we use the pre-trained features of the softmax neural classifier $f(x)$ to estimate the parameters of the generative classifier. Following GDA, we compute the empirical class mean and covariance of training samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$:

$$\widehat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f(\mathbf{x}_i), \widehat{\mathbf{\Sigma}} = \frac{1}{N} \sum_c \sum_{i:y_i=c} \left(f(\mathbf{x}_i) - \widehat{\mu}_c\right)\left(f(\mathbf{x}_i) - \widehat{\mu}_c\right)^\top \quad (2)$$

where N_c is the number of training samples with label c.

Distance Functions: Euclidean vs Mahalanobis Using the above class-conditional Gaussian distributions, we estimate the confidence score M(x) using distance functions between test sample x and the closest class.

- Euclidean Distance: $M(\mathbf{x}) = \max_c - (f(\mathbf{x}) - \widehat{\mu}_c)^\top (f(\mathbf{x}) - \widehat{\mu}_c)$
- Mahalanobis Distance: $M(\mathbf{x}) = \max_c - (f(\mathbf{x}) - \widehat{\mu}_c)^\top \widehat{\mathbf{\Sigma}}^{-1} (f(\mathbf{x}) - \widehat{\mu}_c)$

Compared to Euclidean, Mahalanobis distance can unitless and scale-invariant, and takes into account the correlations between features. The distance metric corresponds to measuring the log of the probability densities of the test sample. Experiments confirm that OOD samples can be characterized better in the representation space of DNNs. In contrast, posterior output distribution of traditional softmax-based discriminative classifier is always susceptible to label-overfitted issues (Hendrycks and Gimpel, 2017; Liang et al., 2018) where OOD samples may get high confidence from the output probability distribution.

3 Experiments

3.1 Setup

	SNIPS	ATIS	CLINC-Full	CLINC-Imbal
Vocabulary size	11241	938	6240	5725
Avg utterance length	9	11	9	9
Intents	7	18	150	150
Training set size	13084	4978	15000	10525
Development set size	700	500	3000	3000
Testing Set Size	700	893	4500	4500

Table 1: Full statistics of the OOD datasets.

% of known intents	Snips 25%	Snips 50%	Snips 75%	ATIS 25%	ATIS 50%	ATIS 75%	CLINC-Full 25%	CLINC-Full 50%	CLINC-Full 75%	CLINC-Imbal 25%	CLINC-Imbal 50%	CLINC-Imbal 75%
MSP	0.0	6.2	8.3	8.1	15.3	17.2	0.0	21.3	40.4	0.0	27.8	40.4
DOC	72.5	67.9	63.9	61.6	62.8	37.7	-	-	-	-	-	-
DOC(softmax)	72.8	65.7	61.8	63.6	63.3	38.7	-	-	-	-	-	-
LOF(softmax)	76.0	69.4	65.8	67.3	61.8	38.9	91.1	83.1	63.5	88.4	77.6	57.5
LOF(LMCL)	79.2	84.1	78.8	68.6	63.4	39.6	91.3	83.3	62.8	88.7	78.9	56.7
GDA+Euclidean distance	85.6	85.6	82.9	77.9	**75.4***	**43.7***	91.1	84.2	64.5	91.1	81.2	60.8
GDA+Mahalanobis distance	**89.2***	**87.4***	**83.2**	**78.5***	72.8	42.1	**91.4**	**84.4**	**65.1***	**91.5**	**81.5**	**61.3***

Table 2: Macro f1-score of unknown intents with different proportions (25%, 50% and 75%) of classes are treated as known intents on SNIPS and ATIS datasets. * indicates the significant improvement over all baselines ($p < 0.05$).

Datasets We perform experiments on four public benchmark OOD datasets, including SNIPS (Coucke et al., 2018), ATIS(Tür et al., 2010), CLINC-Full, and CLINC-Imbal (Larson et al., 2019). We show the detailed statistics of these datasets in Table 1. SNIPS is a personal voice assistant dataset which contains 7 types of user intents across different domains. ATIS dataset contains recordings of people making reservations with 18 types of user intent in the flight domain. CLINC-Full and CLINC-Imbal both contain 150 intents across 10 domains. CLINC-Full is balance, containing 150 samples for each intent and CLINC-Imbal has fewer training samples for each intent and is imbalance.

Baselines We compare our methods with the following state-of-the-art baselines using macro f1-scores of OOD intents. MSP (Maximum Softmax Probability) (Hendrycks and Gimpel, 2017) applies a threshold on the maximum softmax probability where the threshold is set as 0.5. DOC and DOC(softmax) are proposed by (Shu et al., 2017) to solve open-world classification. LOF(softmax) and LOF(LMCL) use local outlier factor to detect unknown intents. Similar to (Lin and Xu, 2019), we vary the number of known classes in the training set in the range of 25%, 50%, and 75% classes and use the other classes as OOD. We provide a more comprehensive comparison and implementation details of these models in the Appendix.

Implementation Details To conduct a fair comparison, we follow a similar evaluation setting as (Lin and Xu, 2019). In each experiment, we randomly sample a set of classes among all classes in the dataset, regarding them as in-domain classes and the rest as out-of-domain classes. We use train samples from only in-domain classes to train the feature extractor and use test samples both from in-domain and out-of-domain classes for OOD sample detection. We vary the proportion of in-domain classes at 25%, 50%, and 75% of all classes. For each proportion, we re-run the experiment 10 times (each with a different set of in-domain classes) and report the average F1-score on OOD sample detection.

We use the pre-trained GloVe embeddings (Pennington et al., 2014) as the word embedding matrix. For the BiLSTM encoder, we set the dimension of hidden states to 128 and use a dropout rate of 0.5. For LCML, we set the scaling factor to 30 and the cosine margin to 0.35. We use Adam optimizer (Kingma and Ba, 2014) to train our model and use a learning rate of 0.003. We train our model up to 200 epochs with an early stop of patience 20.

3.2 Results and Discussion

Main results Table 2 displays the experiment results. Our method consistently outperforms all baselines in all settings. Compared to LOF, our method improves the macro f1-score on SNIPS by 10.0%, 3.3%, and 4.4% in 25%, 50%, and 75% setting respectively. We also observe similar improvements on ATIS and CLINC datasets. The results confirm the effectiveness of our generative approach. Besides, the f1 scores of OOD intents rapidly drop as the number of known intents increases. The reason is that extensive classes of known intents will lead to overlapping with OOD classes in semantics, which causes poor performance. Still, results show that our method is more robust to class overlapping. We can also see our method better handles data imbalance comparing the average improvements on CLINC-Full (1.2%) with CLINC-Imbal (3.3%) since CLINC-Imbal is a more imbalanced dataset than CLINC-Full. Although we only focus on unsupervised OOD detection methods, we still choose supervised methods for comprehensive comparison in the Appendix, both on OOD intent classes and in-domain classes.

Visualization Fig 2 shows the visualization of learned features on Euclidean and Mahalanobis spaces

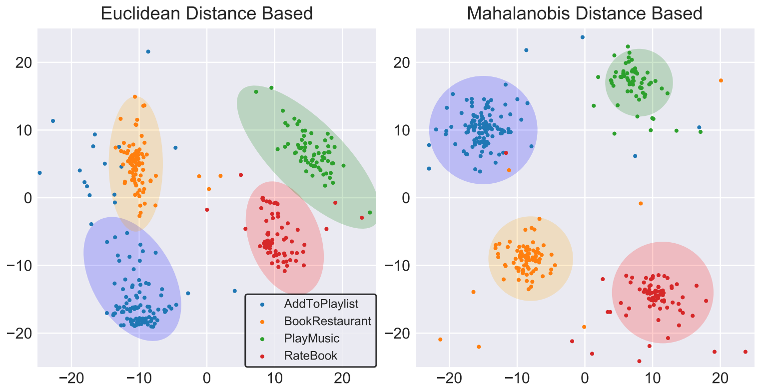


Figure 2: Visualization of learned features on Euclidean and Mahalanobis spaces using SNIPS dataset.

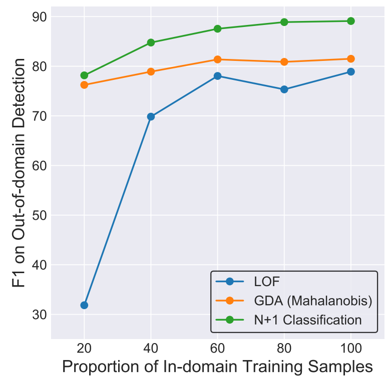


Figure 3: Effect of labeled in-domain data size.

using t-SNE (Maaten and Hinton, 2008). We can see that Mahalanobis re-scales pre-trained features to unit variance and takes into account the correlations of different feature dimensions and has a larger intra-class margin than Euclidean distance.

Effect of labeled in-domain data size Fig 3 shows the effect of different sizes of in-domain data. We choose LOF, N+1 classification (treating OOD as $(N+1)$-th class) and our method for comparison. All the methods achieve better performance along with the increase of in-domain data. We hypothesize that more labeled in-domain data can train a better feature extractor. Besides, our method outperforms LOF with a large margin in the few-shot scenario (only a few labeled in-domain samples), even close to supervised N+1 classification. It proves our method is more robust to the feature extractor than LOF.

% of known intents	50			75		
Macro f1-score	overall	seen	unseen	overall	seen	unseen
GDA+Mahalanobis distance	80.2	80.1	84	79.4	79.6	65.7
N+1 classification(2000)	64.6	64.6	67.7	65.7	65.7	66.6
N+1 classification(4000)	45.3	44.9	77.7	46.3	46.1	78.9

Table 3: Comparison between our unsupervised OOD detection method and supervised N+1 classification.

Unsupervised OOD vs Supervised OOD Table 3 shows the comparison between our method and N+1 classification, where 2000 and 4000 represent the number of labeled OOD samples of training data. We can see that for the overall f1-score, our method outperforms N+1 classification. Besides, although more labeled OOD data can facilitate unseen intent detection, it undermines the performance of seen intent classes. It is dangerous in a practical scenario. Another drawback of supervised OOD detection is re-training when new OOD samples are added to the dataset. Collecting OOD data is labor-intensive and expensive.

4 Conclusion

In this paper, we focus on the unsupervised OOD detection and propose a simple but strong generative distance-based classifier to detect OOD samples. We combine the advantages of DNNs and Gaussian discriminant analysis (GDA) to avoid previous over-confidence problems and apply Mahalanobis distance to measure the confidence score. Experiments show that our method achieves better performance and is more robust to data imbalance and poor feature extractor as well as the few-shot scenario.

Acknowledgments

We thank all anonymous reviewers for their constructive feedback. This work was partially supported by National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701, DOCOMO Beijing Communications Laboratories Co., Ltd.

References

Abhijit Bendale and Terrance E. Boult. 2016. Towards open set deep networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. Lof: identifying density-based local outliers. In *SIGMOD '00*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *HLT-NAACL*.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *AAAI*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

E Haihong, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471.

K. He, Y. Yan, H. Xu, S. Liu, Z. Liu, and W. Xu. 2020a. Learning label-relational output structure for adaptive sequence labeling. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Keqing He, Weiran Xu, and Yuanmeng Yan. 2020b. Multi-level cross-lingual transfer learning with language shared and specific knowledge for spoken language understanding. *IEEE Access*, 8:29407–29416.

Keqing He, Yuanmeng Yan, and Weiran Xu. 2020c. Learning to tag OOV tokens by integrating contextual representation and background knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 619–624, Online, July. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136.

Joo-Kyung Kim and Young-Bum Kim. 2018. Joint learning of domain classification and out-of-domain detection with dynamic class weighting for satisficing false acceptance rates. *ArXiv*, abs/1807.00072.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Stefan Larson, Anish Mahendran, Joseph Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *EMNLP/IJCNLP*.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *ArXiv*, abs/1807.03888.

Shiyu Liang, Yixuan Li, and R. Srikant. 2017. Principled detection of out-of-distribution examples in neural networks. *ArXiv*, abs/1706.02690.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv: Learning*.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *ACL*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech 2016*, Sep.

Laurensvander Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Z. Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:530–539.

Kevin P. Murphy. 2012. Machine learning - a probabilistic perspective. In *Adaptive computation and machine learning series*.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. 2019. Do deep generative models know what they don't know? *ArXiv*, abs/1810.09136.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. *ArXiv*, abs/1906.02845.

Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. 2013. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1757–1772.

Lei Shu, Hu Xu, and Bing Liu. 2017. Doc: Deep open classification of text documents. In *EMNLP*.

Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? *2010 IEEE Spoken Language Technology Workshop*, pages 19–24.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274.

Xu Weiran and Zhang Chunyun. 2016. Trigger word mining for relation extraction based on activation force. *International Journal of Communication Systems*, 29(14):2134–2146.

Weiran Xu, Xiusen Gu, and Guang Chen. 2019. Generating emotional controllable response based on multi-task and dual attention framework. *IEEE Access*, 7:93734–93741.

Weiran Xu, Chenliang Li, Minghao Lee, and Chi Zhang. 2020. Multi-task learning for abstractive text summarization with key information guide network. *Journal on Advances in Signal Processing*, 2020(1).

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209.

A Appendix

A.1 Settings

Baselines We compare our method with the following state-of-the-art methods on the performance of detecting OOD samples:

1. **Maximum Softmax Probability (MSP)** A method proposed in (Hendrycks and Gimpel, 2017), which simply utilizes the maximum probability from softmax distributions as the confidence score and sets a threshold for out-of-distribution detection. If a sample has a lower confidence score than the threshold, it will be regarded as an out-of-distribution sample.

2. **Deep Open Classification (DOC)** A method proposed in (Shu et al., 2017) to solve the open-world classification problem. It replaces softmax with sigmoid activation function at the final layer and computes the confidence threshold for each class to further tighten the decision boundary of the sigmoid function.

3. **DOC (softmax)** A variant of DOC, which replaces the sigmoid activation function with the softmax at the final layer.

4. **LOF (LCML)** A method proposed in (Lin and Xu, 2019). It first uses Large Cosine Margin Loss (LCML) (Wang et al., 2018) to train a feature extractor, and then uses a density-based novelty detection algorithm Local Outlier Factor (LOF) (Breunig et al., 2000) for out-of-distribution detection.

5. **LOF (softmax)** A variant of LOF (LCML), which replaces LCML with the normal softmax loss to train the feature extractor.

A.2 Case Study

Table 4 displays the case study of our method vs softmax+threshold. For case #1, ”add to my playlist all funked up this track”, softmax+threshold predicts a wrong intent *PlayMusic* with a strong confidence score 0.87. It confirms that samples from OOD may also receive a high detection score if they share similar patterns and phrases with some in-domain samples, which we call it the label-confident problem in this paper.

#1 Sentence: *add to my playlist all funked up this track*
⋆ **Predicted by softmax classifier**: PlayMusic with probability: **0.87**
⋆ **Predicted by our proposed method**: OOD sample
Ground truth class: AddToPlayList (out-of-domain class)
#2 Sentence: *i want to add michelle heaton to this is chopin*
⋆ **Predicted by softmax classifier**: PlayMusic with probability: **0.86**
⋆ **Predicted by our proposed method**: OOD sample
Ground truth class: AddToPlayList (out-of-domain class)
#2 Sentence: *i m looking for a movie called salvage mice*
⋆ **Predicted by softmax classifier**: SearchScreeningEvent with probability: **0.87**
⋆ **Predicted by our proposed method**: OOD sample
Ground truth class: SearchCreativeWork (out-of-domain class)
#2 Sentence: *show me the movie operetta for the theatre organ*
⋆ **Predicted by softmax classifier**: SearchScreeningEvent with probability: **0.87**
⋆ **Predicted by our proposed method**: OOD sample
Ground truth class: SearchCreativeWork (out-of-domain class)

Table 4: OOD detection examples from SNIPS dataset, where the in-domain classes includes *PlayMusic* and *SearchScreeningEvent*. The [GREEN] and [RED] highlight indicate correct and incorrect predictions, respectively.