IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 March 2023, accepted 11 April 2023, date of publication 18 April 2023, date of current version 28 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3268107

RESEARCH ARTICLE

Bridge Pre-Training and Clustering: A Unified Contrastive Learning Framework for OOD Intent Discovery

YUTAO MOU[1] AND HEYANG XU[2]
[1]School of Artificial intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]International School, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Yutao Mou (myt@bupt.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFF0303300 and Grant 2019YFF0303302; and in part by DOCOMO Beijing Communications Laboratories Company Ltd., MoE-CMCC "Artificial Intelligence'' under Project MCM20190701.

ABSTRACT Discovering Out-of-Domain (OOD) intents is essential for developing new skills in a task-oriented dialogue system. Previous methods suffer from poor knowledge transferability from in-domain (IND) intents to OOD intents, and inefficient iterative clustering. In this paper, we propose an efficient unified contrastive learning framework to discover OOD intents, bridging the gap between IND pre-training stage and OOD clustering stage. Specifically, we employ a supervised contrastive learning (SCL) objective to learn discriminative pre-trained intent features for clustering. And we introduce an efficient end-to-end contrastive clustering method to jointly learn representations and cluster assignments. Besides, we propose an adaptive contrastive learning (ACL) method to automatically adjust the weights of different negative sample pairs for a given anchor according to their semantic similarities. Extensive experiments on two benchmark datasets show that our method is more robust and achieves substantial improvements over the state-of-the-art methods.

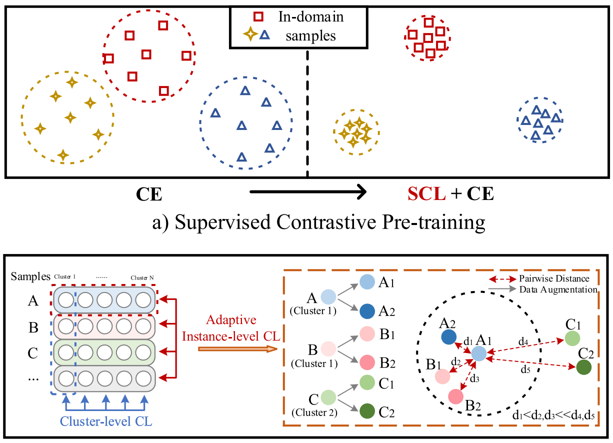
INDEX TERMS Clustering, contrastive learning, knowledge transfer, OOD intent discovery, dialogue system.

I. INTRODUCTION

Discovering Out-of-Domain (OOD) or unknown intents from user queries is an essential component in a task-oriented dialog system [1], [2], [3], [4], [5]. By grouping new unknown intents into different clusters, we may identify future development directions to improve the dialogue system. Different from normal text clustering tasks, OOD discovery needs to consider how to leverage the prior knowledge of known in-domain (IND) intents to enhance clustering unknown OOD intents, which makes it difficult to directly apply existing clustering algorithms [6], [7], [8], [9] to the OOD discovery task.

We classify the existing methods of OOD discovery into two main categories, unsupervised and semi-supervised OOD discovery. Unsupervised methods [10], [11], [12] only model OOD data but ignore prior knowledge of in-domain data thus impair final clustering performance. Therefore, recent work focus on the semi-supervised setting where there exist a few labeled IND intents [4], [5]. [5] firstly pre-trains a BERT-based [14] in-domain intent classifier using cross-entropy classification loss then uses intent representations to calculate the similarity of OOD sample pairs as weak supervised signals. The gap between pre-trained IND features and unseen OOD data makes it hard to generate high-quality pairwise pseudo labels. Then, [4] proposes an iterative clustering method, DeepAligned, to obtain pseudo classification labels. For each training epoch, they firstly perform k-means [7] on the extracted pre-trained intent features, and then use the produced aligned cluster assignments to finetune the intent classifier. However, DeepAligned learns intent representations and cluster assignments in a pipeline manner,

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.



a) Supervised Contrastive Pre-training

b) Adaptive Contrastive Clustering

FIGURE 1. The overall architecture of our proposed framework. Fig (a) denotes the SCL pre-training to learn better discriminative features. Fig (b) represents adaptive contrastive clustering to adjust sample weights given an anchor sample so that a negative sample within the same cluster gets a smaller penalty than distant negatives from other clusters.

which is notably inefficient and may cause error propagation. Generally, these semi-supervised methods don't match the IND pre-training objective in the first stage with the OOD clustering objective in the second stage and suffer from poor knowledge transferability. Therefore, in this paper, we aim to align the two-stage learning objectives and improve the efficiency and accuracy of OOD discovery via a unified contrastive learning framework.

The main challenges of OOD intent discovery are summarized as follows: (1) Knowledge Transferability. It's hard to effectively transfer prior IND knowledge to OOD data. Because classification objectives in the IND pre-training stage don't align with the clustering objectives in the OOD clustering stage, which makes the knowledge transfer from IND to OOD has a natural gap. (2) Jointly Learning Representations and Cluster Assignments. Previous OOD clustering methods [4], [6] iteratively learn intent features and cluster assignments. Limited by the inherent inefficiency of clustering algorithms like k-means, these methods suffer from lazy two-stage back-propagation signals thus result in poor performance. Consequently, it's vital to jointly learn representations and cluster assignments.

To solve these challenges, we propose an efficient unified contrastive learning framework (COD) to discover OOD intents as shown in Fig 1. For knowledge transferability, we employ a supervised contrastive learning (SCL) pre-training objective [15] to better learn discriminative pre-trained intent features for clustering. Previous cross-entropy (CE) loss only focuses on whether a sample is correctly classified and does not explicitly distinguish the margins between categories [3], [13]. In contrast, SCL aims to minimize intra-class variance by pulling together IND samples belonging to the same class and maximize inter-class variance by pushing apart samples from different classes. Moreover, pre-training with SCL aligns with the clustering objective we will discuss later, which can bridge the gap between pre-training and clustering. For jointly learning representations and cluster assignments, we introduce an efficient end-to-end contrastive clustering method to simultaneously model instance-level and cluster-level representation space. Specifically, we regard the rows of the input feature matrix of a batch of augmented examples as instance representations and the columns as cluster representations [16], [17]. Then we can construct two levels of contrastive objectives where the instance-level one helps capture low-level linguistic knowledge and the cluster-level one facilitates learning high-level semantic concepts. Further, we theoretically find the instance-level contrastive objective indeed has a negative impact on clustering performance since it pushes apart representations from different instances even while they belong to the same cluster with strong semantic similarities. Therefore, we propose an adaptive contrastive learning (ACL) method to automatically adjust the weights of different negative samples for a given anchor according to their semantic similarities. ACL aims to perform a smaller penalty on semantically similar negative intent samples but a larger penalty on distant negatives, which can be regarded as a soft negative sampling strategy. Combining the pre-training stage and the clustering stage, we propose a simple but strong unified contrastive learning framework for the OOD discovery task, which can effectively solve both knowledge transfer and clustering efficiency issues.

The novelty of our proposed method comes from three aspects: (1) We are the first to propose a unified contrastive learning framework for the OOD discovery task. In contrast, previous methods [4], [5], [6] use two indenpedent models to learn IND features and OOD clustering respectively, which make the gap between pre-trained IND features and unseen OOD data. Our method employs a unified view to solve the two problems. (2) We introduce a supervised contrastive learning pre-trained objective to learn discriminative intent features compared to the previous cross-entropy loss. (3) We propose a novel adaptive contrastive learning mechanism to perform better OOD clustering using soft negative sampling.

Our contributions are four-fold: (1) To the best of our knowledge, we are the first to propose a unified contrastive learning framework for OOD discovery, bridging the gap between pre-training and clustering. (2) We introduce a supervised contrastive learning pre-trained objective to learn discriminative intent features by maximizing inter-class variance and minimizing intra-class variance. (3) We propose a novel adaptive contrastive learning mechanism to perform soft negative sampling. (4) Experiments and analysis on two benchmark datasets demonstrate the effectiveness of our framework for OOD discovery.

II. RELATED WORK

A. INTENT MODELING

There are two main applications of intent modeling in the task-oriented dialogue system, intent classification and OOD discovery. The former aims to distinguish intent types jointly with other tasks, like slot filling [24], [25], [26]. The latter is to leverage intent representations to construct clustering signals [4], [5], [12]. In this paper, we focus on the latter application. It's important to leverage hidden semantic information to construct supervised signals for intent feature learning.

B. CLUSTERING

Most existing clustering methods are unsupervised, such as partition-based methods [7], hierarchical methods [27] and density-based methods [28], feature dimensionality reduction methods [27]. However, these methods suffer from high computational complexity and poor performance since they can't capture high-level semantics of intent features. Then deep clustering methods are proposed to leverage the strong feature modeling capability of deep neural networks (DNNs), such as JULE [29], DEC [9], DeepCluster [6]. The joint unsupervised learning (JULE) [29] combines deep feature learning with hierarchical clustering but needs huge computational and memory costs on large-scale datasets. Deep Embedded Clustering (DEC) [9] trains the autoencoder with the reconstruction loss and iteratively refines the cluster centers by optimizing KL-divergence with an auxiliary target distribution. Compared with DEC, Deep Clustering Network [6] further introduces a k-means loss as the penalty term to reconstruct the clustering loss. However, these methods follow a two-stage clustering process and only use unsupervised data.

Recent work perform semi-supervised clustering with the aid of some labeled data, such as KCL [30], CDAC+ [5], DeepAligned [4] and DKT [46]. KCL [30] uses deep neural networks to perform pairwise constraint clustering. It firstly trains an extra network for binary similarity classification with a labeled auxiliary dataset. Then, it transfers the prior knowledge of pairwise similarity to the target dataset and uses KL-divergence to evaluate the pairwise distance. CDAC+ [5] is specifically designed for discovering new intents. It uses limited labeled data as a guide to learn pairwise similarities. However, it is limited in providing specific supervised signals and fails to estimate the number of novel classes. DeepAligned is the previous mainstream baseline for OOD discovery which iteratively learns intent representations then cluster assignments. DKT is a newly proposed method, which introduces contrastive learning in OOD discovery for the first time, and uses a multi-head decoupling framework to map the shared intent representations of BERT to the instance-level and cluster-level subspaces. Our proposed COD significantly outperforms all previous methods on two benchmark datasets.

III. APPROACH

A. PROBLEM FORMULATION

In this paper, we denote OOD discovery as OOD clustering with IND pre-training unless otherwise stated. Given a set of labeled in-domain data ($\mathcal{X}_{IND}, \mathcal{Y}_{IND}$) and unlabeled OOD data ($\mathcal{X}_{OOD}, \mathcal{Y}_{OOD}$), OOD discovery aims to cluster OOD groups from unlabeled OOD data using prior knowledge from labeled IND data. Note that IND classes have no overlapping with OOD classes.

B. OVERALL ARCHITECTURE

Fig 2 displays the overall architecture of our proposed unified contrastive learning framework for OOD discovery, COD. We follow the similar two-stage framework as [5] and [4]: IND pre-training and OOD clustering. For IND pre-training, we employ a supervised contrastive learning (SCL) objective to better learn discriminative pre-trained intent features along with the traditional cross-entropy (CE) loss. For OOD clustering, we introduce an efficient end-to-end contrastive clustering method to jointly learn representations and cluster assignments. Besides, we propose an adaptive contrastive learning (ACL) loss to automatically adjust the weights of different negative sample pairs for a given anchor according to their semantic similarities. We will dive into the details in the following sections.

C. SUPERVISED CONTRASTIVE PRE-TRAINING

We firstly pre-train an intent feature extractor using labeled IND data. Specifically, we use the similar BERT [14] intent classifier following [4] for fair comparison, including the input layer, BERT encoder, and a pooling layer. Finally, we obtain the intent representation $z_i \in \mathbb{R}^H$ for the input sample x_i. Previous intent classification models [2], [4], [5], [19] always use cross-entropy objective which only focuses on whether a sample is correctly classified, and does not explicitly distinguish the margins between categories. Inspired by recent contrastive work [3], [15], [20], we employ a supervised contrastive learning (SCL) objective to learn discriminative intent features by maximizing inter-class variance and minimizing intra-class variance. We formulate SCL as follows:

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \log \frac{\exp\left(s_i \cdot s_j / \tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp\left(s_i \cdot s_k / \tau\right)} \tag{1}$$

where N_{y_i} is the total number of examples in the batch that have the same label as y_i and $\mathbf{1}$ is an indicator function. Following [21], [22], we employ simple dropout [23] as data augmentation. As Fig 1(a) shows, SCL aims to pull together IND samples belonging to the same class and push apart samples from different classes, which helps distinguish OOD cluster boundaries. In the implementation, we perform joint training both using SCL and CE. We also try other variants, such as only using SCL, firstly use SCL then CE, etc. However, simply adding SCL and CE gets the best performance. We conduct a comprehensive analysis (see Section V-A) of the effect of SCL from multiple perspectives, including IND and OOD, both achieving superior performance than CE.

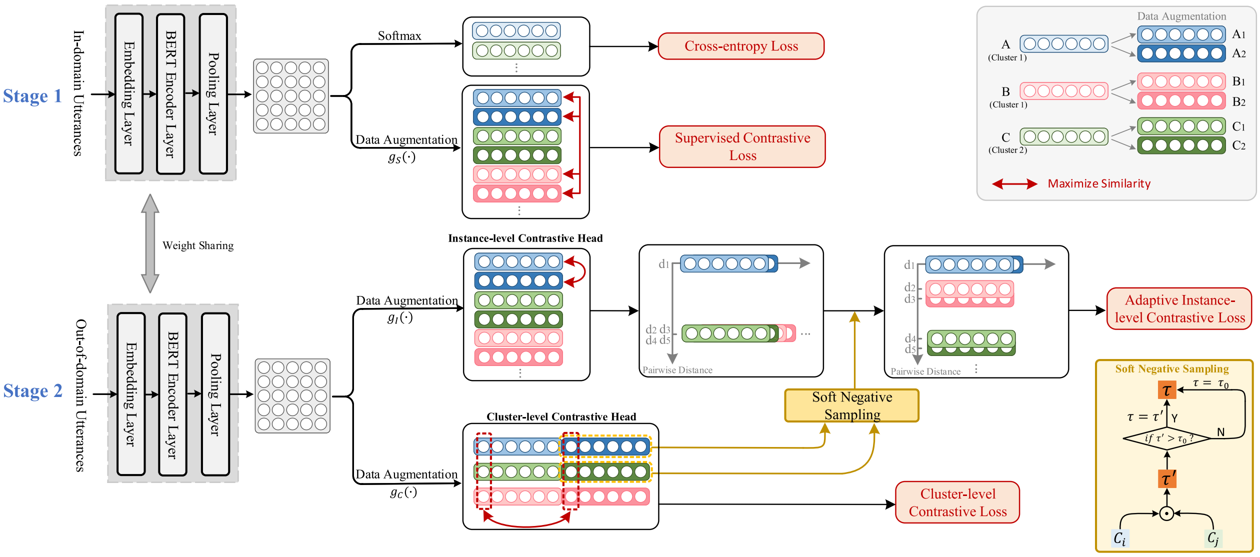


FIGURE 2. The overall architecture of our proposed unified contrastive learning framework for OOD discovery, COD. Stage 1 denotes IND pre-training and Stage 2 denotes OOD clustering.

D. ADAPTIVE CONTRASTIVE CLUSTERING

After transferring knowledge from known intents, we propose an efficient end-to-end contrastive clustering method to group similar OOD intents into the same cluster. The key challenge of OOD clustering is how to jointly learn representations and cluster assignments. Previous mainstream method DeepAlighed [4] uses the DeepCluster [6] algorithm with an aligned mechanism to iteratively learn intent representations then cluster assignments. We argue that this method suffers from poor clustering efficiency and lazy back-propagation signals. Therefore, we introduce an end-to-end contrastive clustering method [16] to mitigate the above issues. Specifically, we firstly use the pre-trained intent classifier to obtain a feature matrix given a batch of dropout-augmented OOD samples. Then we adopt two individual two-layer nonlinear MLPs $g(\cdot)$ to map the feature matrix to a new subspace where two contrastive objectives are applied. We regard the rows of the new feature matrix as instance representations and the columns as cluster representations [17]. Next, we can construct two levels of contrastive objectives where the instance-level one helps capture low-level linguistic knowledge and the cluster-level one facilitates learning high-level semantic concepts. We use different transformation MLPs for the two-level contrastive objectives, which has been proved effective by [16].

1) ADAPTIVE INSTANCE-LEVEL CONTRASTIVE LOSS

We formulate the original instance-level CL loss for a given sample z_i:

$$\ell_{i,j}^{ins} = -\log \frac{\exp\left(\operatorname{sim}\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp\left(\operatorname{sim}\left(z_i, z_k\right)/\tau\right)} \quad (2)$$

where z_i represents the transformed vector of i-th intent sample and z_j is the dropout-augmented sample. $1_{[k\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 if $k \neq i$. τ denotes a temperature parameter. Then we extend the normalized item as follows:

$$\begin{aligned} &\sum_{k=1}^{2N} 1_{[k\neq i]} \exp\left(\operatorname{sim}\left(z_i, z_k\right)/\tau\right) \\ &= \sum_{k\in\mathcal{P}} \exp\left(\operatorname{sim}\left(z_i, z_k\right)/\tau\right) + \sum_{k\in\mathcal{N}} \exp\left(\operatorname{sim}\left(z_i, z_k\right)/\tau\right) \quad (3) \end{aligned}$$

where $\mathcal{P}$ is the positive set of anchor z_i and $\mathcal{N}$ is the negative set. Original instance-level CL uses the anchor sample and its augmented sample as a positive pair, but regard the other samples in the batch as negatives, which is not suitable to OOD clustering. Because clustering tries to pull together samples within the same cluster and push apart samples from different clusters. Therefore, to decrease the weight of these false negative samples which belong to the same cluster with the anchor, we propose an adaptive contrastive loss as shown in Fig 2: Soft Negative Sampling. The main intuition is to adaptively adjust the temperature of each negative sample according to their semantic similarities:

$$\tau_{i,j} = \begin{cases} \tau' & \text{if } \tau' > \tau_0 \\ \tau_0 & \text{if } \tau' \leqslant \tau_0 \end{cases} \quad (4)$$

$$\tau' = C_i \odot C_j \quad (5)$$

where $\tau_{i,j}$ denotes the temperature between anchor z_i and other sample z_j and $\odot$ represents the dot product of z_i's cluster logits C_i and z_j's cluster logits C_j from cluster-level contrastive head. Here we use samples' cluster logits to compute their semantic similarities of belonging to the same cluster. If the similarity of (z_i, z_j) is above a fixed hyperparameter τ_0 (we set it to 0.5), then the negative sample z_j gets a relatively larger temperature and a smaller penalty so that (z_i, z_j) don't stay away from each other. By this way, we can keep pairs with the similar semantics as near as possible.

Here we give a theoretical explanation of our proposed adaptive instance-level contrastive (ACL) loss. The original instance-level contrastive loss is formulated as follows:

$$\mathcal{L}(x_i) = -\log\left[\frac{\exp\left(s_{i,i}/\tau\right)}{\sum_{k\neq i}\exp\left(s_{i,k}/\tau\right)+\exp\left(s_{i,i}/\tau\right)}\right] \tag{6}$$

where $s_{i,j} = sim(f(x_i), f(x_j))$. For convenience, we denote $s_{i,i}$ as the positive pair and $s_{i,j}, i \neq j$ as negative pairs, which is slightly different from Eq 2. Then we analyze the gradients with respect to positive samples and different negative samples following [37], [38].

$$\frac{\partial\mathcal{L}(x_i)}{\partial s_{i,i}} = -\frac{1}{\tau}\sum_{k\neq i}P_{i,k} \tag{7}$$

$$\frac{\partial\mathcal{L}(x_i)}{\partial s_{i,j}} = \frac{1}{\tau}P_{i,j} \tag{8}$$

$$P_{i,j} = \frac{\exp\left(s_{i,j}/\tau\right)}{\sum_{k\neq i}\exp\left(s_{i,k}/\tau\right)+\exp\left(s_{i,i}/\tau\right)} \tag{9}$$

We observe from Eq 8 & 9 that the gradients with respect to negative samples is proportional to the exponential term $\exp(s_{i,j}/\tau)$ since all other items are the same for all negative samples. If we increase the temperature τ of negative samples belonging to the same cluster, the gradient (penalty) gets smaller so that these false negatives within a cluster can get closer than true negatives from different clusters. Therefore, we can keep intent representations from the same cluster close and dense. For specific implementation, we use the dot product of cluster logits between two samples to measure whether they belong to the same cluster as shown in Eq 4 & 5.

2) CLUSTER-LEVEL CONTRASTIVE LOSS

When projecting a data sample into a space whose dimensionality equals the number of clusters, the i-th element of its feature can be interpreted as its probability (logit) of belonging to the i-th cluster. Meanwhile, all the i-th elements from a batch of feature vectors (i-th column of the feature matrix) denote the i-th cluster representation accordingly. Intuitively, OOD clustering aims to pull together cluster representation pairs(positive) from the same cluster and push apart negative pairs from different clusters. We simply use dropout augmentation to get its augmented version corresponding to the cluster representation of original samples. Therefore, we formulate the cluster-level CL as follows:

$$\ell_{i,j}^{clu} = -\log\frac{\exp\left(\text{sim}\left(y_i, y_j\right)/\tau\right)}{\sum_{k=1}^{2M}1_{[k\neq i]}\exp\left(\text{sim}\left(y_i, y_k\right)/\tau\right)} \tag{10}$$

where y_i denotes i-th cluster representation (also i-th column of feature matrix) and y_j is the dropout-augmented cluster representation. M is the cluster number. To avoid the trivial solution that most instances are assigned to the single cluster, we also add an regularization item $H(y_i)$:

$$H(y_i) = -p(y_i)logp(y_i), p(y_i) = \sum_{j=1}^{N}y_{ji}/\|Y\|_1 \tag{11}$$

where y_{ji} is the (j, i) coordinate of cluster-level feature matrix Y. We simply add the above three objectives and optimize together in the experiments which still gets significant improvements. For inference, we only use the cluster-level contrastive head and compute the argmax to get the cluster results without additional k-means.

TABLE 1. Statistics of CLINC and Banking datasets.

Dataset	Classes	Training	Validation	Test	Vocabulary	Length (max/mean)
CLINC	150	18,000	2,250	2,250	7,283	28 / 8.31
BANKING	77	9,003	1,000	3,080	5,028	79 / 11.91

IV. EXPERIMENTS

A. DATASETS

We conduct experiments on two benchmark datasets, CLINC [33] and Banking [34]. CLINC contains 22,500 queries covering 150 intents and Banking contains 13,083 customer service queries with 77 intents. We show the detailed statistics in Table 1. Following previous work, to construct IND/OOD data, we divided the two datasets according to the specified OOD ratio(10%, 20%, 30% for CLINC, 10% for Banking), and the rest is IND data. For the semi-supervised setting (we mainly focus on in this paper), we use the labeled IND data for pre-training and use unlabeled OOD data for clustering. For the unsupervised setting, we only use unlabeled OOD data for clustering. We rerun all the baselines for three times using our settings and report the averaged results on the same divided IND/OOD datasets for reliable and fair evaluation. For each run, all the models use the same divided dataset. Due to limited resources, we only perform a 10% split on Banking, but pay more attention to extensive ablation studies to understand the effectiveness of our proposed method.

B. BASELINES

We mainly compare our method with semi-supervised baselines: PTK-means (k-means with IND pre-training), DeepCluster [6] and three OOD discovery methods CDAC+ [5], DeepAligned [4] and DKT [46]. We also report the results of unsupervised methods for a comprehensive comparison. For fairness, we use the same BERT backbone as the baselines. To avoid the randomness of splitting IND/OOD, we average results over three random runs following [4]. For each run, all the models use the same divided dataset. We adopt three widely used metrics to evaluate the clustering results: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). To calculate ACC, we use the Hungarian algorithm [35] to obtain the mapping between the predicted classes and ground-truth classes.

C. IMPLEMENTATION DETAILS

For a fair comparison with previous work, We use the same pre-trained BERT model (bert-base-uncased[1]) as our network backbone. During the pre-training phase, the training batch size is 128, and during the clustering phase,

[1] https://github.com/google-research/bert

TABLE 2. Performance comparison on two datasets. For CLINC, We randomly sample 10%, 20% and 30% of all classes as OOD types. For Banking, we randomly sample 10% of all classes as OOD types. We evaluate both unsupervised and semi-supervised methods. Results are averaged over three random runs. ($p < 0.05$ under t-test).

Method		CLINC-10%			CLINC-20%			CLINC-30%			Banking		
		ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
Unsup.	k-means	58.67	43.81	67.77	48.89	30.90	64.68	42.22	23.65	60.55	32.81	8.30	17.30
	DeepCluster	53.15	37.80	62.31	47.73	34.55	65.91	33.96	18.89	56.21	29.81	7.79	17.34
	DeepAligned	62.66	47.60	71.50	48.24	34.49	66.24	39.02	24.50	61.16	36.56	12.57	21.84
Semi-sup.	PTK-means	70.22	50.39	73.92	57.56	37.02	72.71	61.63	40.96	75.90	55.00	36.18	53.75
	DeepCluster	78.13	68.31	82.87	83.42	76.18	89.33	78.09	71.05	88.70	60.59	41.88	55.22
	CDAC+	88.00	75.18	88.33	84.89	75.98	89.96	73.04	64.44	87.90	77.50	60.53	71.14
	DeepAligned	95.11	89.81	94.13	93.80	90.22	95.83	91.56	86.58	94.91	77.78	66.95	76.91
	DKT	97.78	95.16	96.97	96.89	93.69	96.94	94.96	90.25	95.94	84.69	71.11	76.92
	COD(ours)	96.44	92.50	95.62	95.11	91.32	96.12	92.44	87.90	95.15	84.69	71.07	78.72
	COD w. ACL(ours)	98.22	96.15	97.79	97.11	94.43	97.63	94.81	90.30	96.10	86.56	73.31	79.03

the training batch size is 512 for CLINC-10, CLINC-30, Banking-10, and 400 for CLINC-20. The learning rate is 5e-5 in the pre-training phase and 0.0003 in the clustering phase. Notably, We use dropout [21] to construct augmented examples for contrastive learning. For the instance-level contrastive head, the dimensionality of the row space is set to 128, and the temperatures of SCL and ACL are 0.5. As for the cluster-level contrastive head, the dimensionality of the column space is naturally set to the number of clusters, and the cluster-level temperature parameter $\tau = 1.0$ is used for all datasets. The dropout value is fixed at 0.5. We use Adam optimizer [36] to train our model. We use the SC metric (see details in Appendix A) of valid OOD data (still unlabeled data) to choose the best checkpoint. The pre-training stage of our model lasts about 30 minutes and clustering runs for 10 minutes on CLINC-10%, both using a single Tesla T4 GPU(16 GB of memory). For comparison, DeepAligned almost consumes 30 minutes for clustering and similar 30 minutes for pre-training. The average value of the trainable model parameters is 17.34M.

For DKT, during the pre-training phase, the training batch size is 128, and during the clustering phase, the training batch size is 512 for CLINC-10%, CLINC-30%, Banking-10%, and 400 for CLINC-20%. The learning rate is 5e-5 in the pre-training phase and 0.0003 in the clustering phase. For the instance-level contrastive head, the dimensionality of the row space is set to 128, and the temperatures of SCL and instance-level CL are 0.5, and the cluster-level temperature parameter $\tau = 1.0$ is used for all datasets. For DeepAligned, the training batch size is 128, the learning rate is 5e-5, and the dimension of intent features is 768. For CDAC+, the training batch size is 256, and the learning rate is 5e-5. We use the same dynamic thresholds as [5]. we freeze all but the last transformer layer parameters to speed up the training procedure and improve the training efficiency with the backbone of BERT.

TABLE 3. Representation distribution of different pre-training objectives.

	Intra-class ↓	Inter-class ↑	SC ↑
No-pretraining	0.42	0.30	0.10
SCL	0.19	0.56	0.43
CE	0.20	0.57	0.31
CE+SCL	0.12	0.74	0.43

TABLE 4. Clustering performance comparison of different pre-training targets using the same clustering method.

	ACC	ARI	NMI	SC
No-pretraining	86.67	75.28	85.50	0.51
SCL	96.46	92.93	95.96	0.79
CE	96.44	92.51	95.70	0.83
CE+SCL	98.22	96.15	97.79	0.92

D. MAIN RESULTS

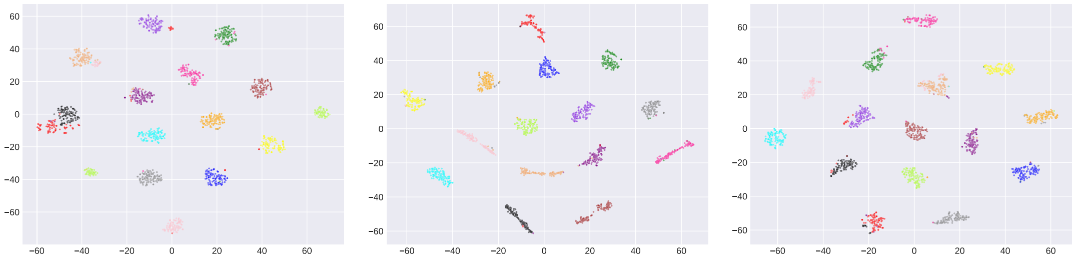
Table 2 shows the main results of our proposed method compared to the baselines. Our method consistently outperforms all the previous baselines with a large margin. For the semi-supervised setting on CLINC-10%, COD w. ACL outperforms the DeepAligned by 3.11%(ACC), 6.34%(ARI), 3.66%(NMI). On Banking, COD w. ACL also gets significant improvements of 8.78%(ACC), 6.36%(ARI), 2.12%(NMI). The results prove the effectiveness of our proposed contrastive framework for OOD discovery. Specifically, comparing COD with DeepAligned, COD gets an improvement of 1.33%(ACC), 2.69%(ARI), 1.49%(NMI) on CLINC-10%. Comparing COD with COD w. ACL, we find ACL also gets an improvement of 1.78%(ACC), 3.65%(ARI), 2.17%(NMI), which confirms adaptive contrastive learning helps learn better cluster assignments. Besides, comparing unsup COD with semi-sup COD, the latter significantly outperforms the former by 9.77%(ACC), 15.34%(ARI), 7.87%(NMI), which demonstrates the effectiveness of SCL pre-training. Overall, both COD and ACL achieve superior performance and the combination of the two is the best.

V. QUALITATIVE ANALYSIS

A. EFFECT OF SUPERVISED CONTRASTIVE LEARNING

Supervised contrastive learning (SCL) contributes to model discriminative representation. We analyze the effect of SCL from multiple perspectives.

We first analyze the spatial distribution of representations when our proposed clustering objective is not used. For in-domain data, we use the intra-class distance, which is the mean value of the Euclidean distance between each sample and its class center, and the inter-class distance, which is the mean value of the Euclidean distance between the center of each class and the center of the 3 classes closest to it.



(a) DeepAligned (b) COD(ours) (c) COD w. ACL(ours)

FIGURE 3. OOD intent visualization of different models. We use the same OOD test set of CLINC-10%.

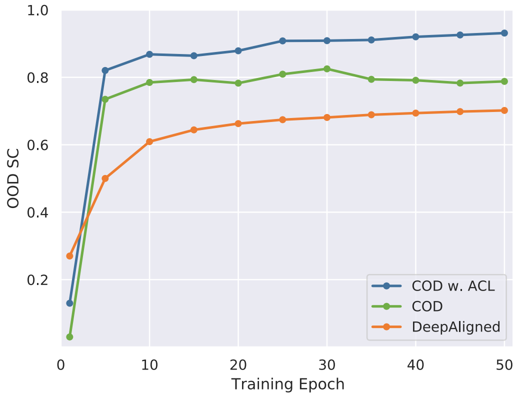


FIGURE 4. OOD SC curves in the training.

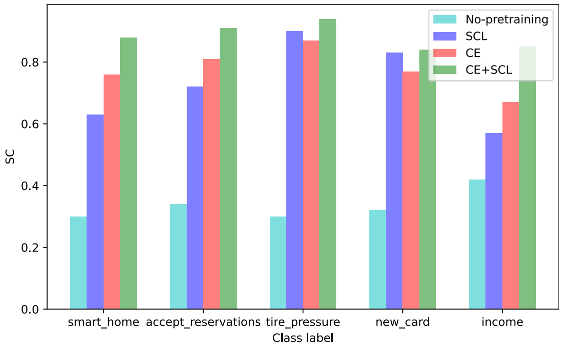


FIGURE 5. Clustering performance comparison of 5 hard OOD clusters.

For OOD data, we use the SC metric [32] for evaluating the quality of OOD clusters (see details in Appendix A). It can comprehensively consider the relationship between the intra-cluster distance and the inter-cluster distance and is used to characterize the tightness of clusters. It should be noted that the cluster label of OOD data is calculated by k-means since we aim to analyze the effect of SCL. As shown in Table 3, we use two basic settings, No-pretraining and CE (using cross-entropy as the pretraining loss). On this basis, we find that after adding SCL, each statistical indicator significantly improves. It shows that SCL is effective for improving data distribution and modeling discriminative representations.

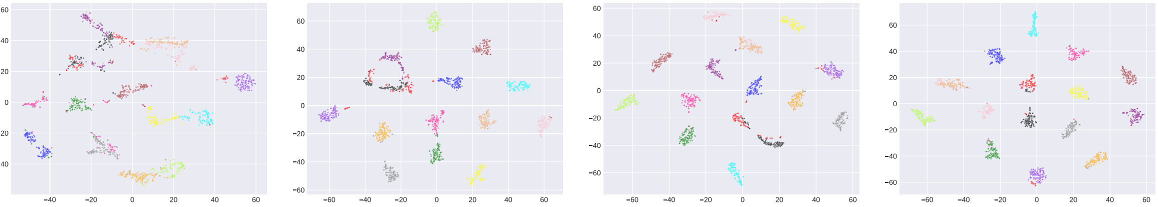
Then, we further conduct experiments on the model that added our proposed clustering method. As shown in Table 4, compared to the corresponding base setting, the addition of SCL brings consistent improvement on all metrics. It indicates that pre-training with SCL does align with the clustering objective, and can effectively bridge the gap between pre-training and clustering. Furthermore, we also independently analyze the 5 OOD clusters with the worst clustering metric (get the lowest SC in the no-pretraining setting). As shown in Fig 5, after adding the SCL training objective, we observe significant improvements in SC, which shows that our method brings obvious improvements to the OOD clusters that are difficult to cluster accurately. This is of great significance in practical applications.

B. EFFECT OF COD AND ACL

To understand the effectiveness of COD and ACL, we perform OOD intent visualization of DeepAligned, COD and COD w. ACL in Fig 3. COD is the overall contrastive learning framework for OOD discovery and ACL is our proposed adaptive instance-level contrastive (ACL) loss. Comparing COD to DeepAligned, we can observe DeepAligned gets some mixed OOD clusters (see red and black dots in Fig a) while COD successfully separates them, which indicates COD learns discriminative OOD cluster assignments. But we also find some OOD clusters have narrow distributions (see black, brown dots in Fig b). We argue it's because COD uses the original instance-level contrastive loss which pushes apart the samples within the same cluster. After using ACC, we can get a more uniform and tight distribution. The visualization proves both COD and ACL helps OOD discovery and have a mutual complementary effect on each other. We also display OOD SC curves in the training in Fig 4. Results show COD converges faster and better than DeepAligned. Note that the initial SC of COD (w. ACL) is worse than DeepAligned because we add a new cluster-level MLP head(randomly initialized) while DeepAligned directly uses k-means, but our methods still converge faster via contrastive objectives. It demonstrates the efficiency of our proposed COD.

C. ESTIMATE THE NUMBER OF CLUSTER K

Since we may not know the exact number of OOD clusters, we use the following K estimation method [4] to determine the number of clusters K before clustering. The method estimates K with the aid of the well-initialized intent features. We assign a big K' as the number of clusters at first. As a good feature initialization is helpful for partition-based methods (e.g., k-means), we use the well pre-trained model to extract intent features. Then, we perform k-means with the extracted



(a) 0 epoch (NMI=30.78) (b) 6 epoch (NMI=87.38) (c) 11 epoch (NMI=88.75) (d) 18 epoch (NMI=90.08)

FIGURE 6. OOD intent visualization of different training epochs for our proposed COD w. ACL.

TABLE 5. Estimate the number of OOD clusters.

	ACC	ARI	NMI	K
PTK-means(fixed K)	70.22	50.39	73.92	15
PTK-means(auto K)	56.89	34.71	68.96	12
Relative↓	18.98%	31.12%	6.71%	-
DeepAligned(fixed K)	95.11	89.81	94.13	15
DeepAligned(auto K)	81.78	77.47	88.04	19
Relative↓	14.02%	13.74%	6.47%	-
COD w. ACL(fixed K)	98.22	96.15	97.79	15
COD w. ACL(auto K)	91.56	87.60	94.39	14
Relative↓	6.78%	8.89%	3.48%	-

features. We suppose that real clusters tend to be dense even with K', and the size of more confident clusters is larger than some threshold t. Therefore, we drop the low confidence cluster whose size is smaller than t, and calculate K with:

$$K = \sum_{i=1}^{K'} \delta\left(|S_i| >= t\right) \quad (12)$$

where $|S_i|$ is the size of the i^{th} produced cluster, and $\delta(\cdot)$ is an indicator function. It outputs 1 if condition is satisfied, and outputs 0 if not. Notably, we assign the threshold t as the expected cluster mean size $\frac{N}{K'}$ in this formula.

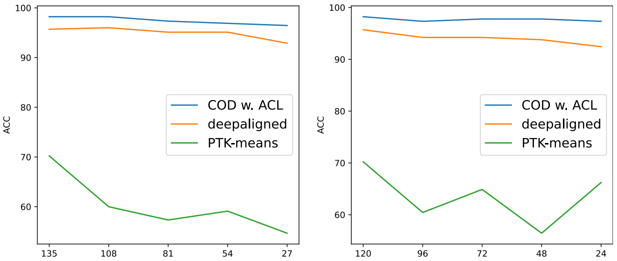
Table 5 shows the OOD clustering results using the automatic K-value estimation strategy. We find that our method both achieves the best performance under the fixed or auto K settings. Besides, under the auto K, all methods have observed a decline in the metrics, which shows that the unknown K value is a great challenge for OOD discovery. However, the reduction of our proposed method is significantly lower than other methods, indicating that our method has strong robustness to the challenge of unknown K which reflects the good practicability of our method.

D. VISUALIZATION AT DIFFERENT TRAINING EPOCHS

To see the evolution of our method in the training, we show a visualization at four different timestamps throughout the training process in Fig 6. Results show features are mixed in the beginning and cluster assignments become increasingly visible and distinct as the training process goes.

E. EFFECT OF IND DATA

We analyze the impact of in-domain data on the effect of clustering from two perspectives, number of IND classes and number of samples per class. Figure 7 (a) shows the trend of the number of different IND classes, and Figure 7 (b) shows the trend of the number of different samples in each class. Overall, the performance of our method is much better than the baselines. Moreover, with the decrease of the amount of in-domain data, all methods show varying degrees of performance fluctuation, and the fluctuation amplitude of our proposed method is the smallest, which shows that our method has little dependence on the in-domain data, and has stronger performance and robustness in the few-shot scenario.



(a) Number of classes (b) Number of samples per class

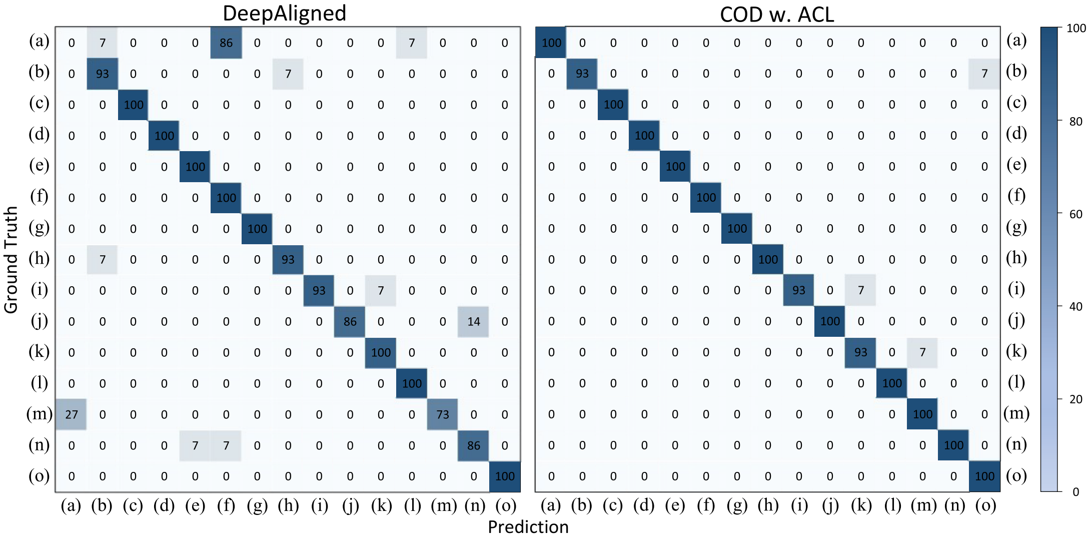
FIGURE 7. Effect of IND data for clustering. The left subfig shows the effect of number of IND classes and the right subfig shows the effect of number of IND samples per class.

F. ERROR ANALYSIS

We further analyze the error cases of DeepAligned and COD w. ACL in Fig 8. We find for semantically similar OOD intents, DeepAligned is probably confused but our COD w. ACL can effectively distinguish them. For example, DeepAligned incorrectly clusters *accept_reservation* intent into *cancel_reservation* (14% error rate) while COD w. ACL gets 100% accuracy. The result shows COD w. ACL helps separate semantically similar OOD intents. We hypothesize it's because the adaptive instance-level contrastive learning helps the model learn discriminative linguistic knowledge.

G. EFFECT OF TEMPERATURE

Fig 9 shows the effect of different temperature τ_0 of ACL. Results show for all the OOD clustering metrics ACC, ARI and NMI, $\tau_0 = 0.5$ gets the best performance. Too larger or smaller temperatures both result in a significant performance drop. Our method with τ_0 in (0.4, 0.9) outperforms the sota baselines, and τ_0 in (0.5, 0.7) brings larger improvements(above 2%), which proves τ_0 is robust. To avoid the randomness, we average results over three random runs. The standard deviation (std) of DeepAligned is 1.16, and the std of COD(t=0.5) is 0.67.



DeepAligned

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)
(a)	0	7	0	0	0	86	0	0	0	0	0	7	0	0	0
(b)	0	93	0	0	0	0	0	7	0	0	0	0	0	0	0
(c)	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
(d)	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
(e)	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
(f)	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
(g)	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
(h)	0	7	0	0	0	0	0	93	0	0	0	0	0	0	0
(i)	0	0	0	0	0	0	0	0	93	0	7	0	0	0	0
(j)	0	0	0	0	0	0	0	0	0	86	0	0	0	14	0
(k)	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
(l)	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
(m)	27	0	0	0	0	0	0	0	0	0	0	0	73	0	0
(n)	0	0	0	0	7	7	0	0	0	0	0	0	0	86	0
(o)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

COD w. ACL

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)
(a)	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(b)	0	93	0	0	0	0	0	0	0	0	0	0	0	0	7
(c)	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
(d)	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
(e)	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
(f)	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
(g)	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
(h)	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
(i)	0	0	0	0	0	0	0	0	93	0	7	0	0	0	0
(j)	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
(k)	0	0	0	0	0	0	0	0	0	0	93	0	7	0	0
(l)	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
(m)	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
(n)	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
(o)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

(a)smart_home (b)spending_history (c)tire_pressure (d)lost_luggage (e)Cancel (f)reset_settings (g)book_flight (h)income
(i)bill_due (j)accept_reservations (k)expiration_date (l)timezone (m)new_card (n)cancel_reservation (o)where_are_you_from

FIGURE 8. Confusion matrix for the clustering results of DeepAligned and COD w. ACL on CLINC-10%. The percentage values along the diagonal represent how many samples are correctly clustered into the corresponding class. The larger the number, the deeper the color.

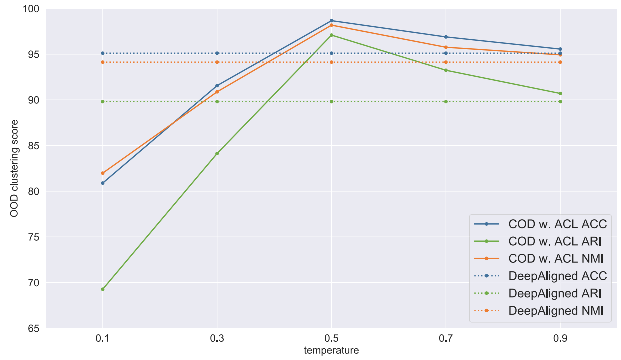


FIGURE 9. Effect of temperature τ_0 of ACL on CLINC-10%. methods. Results are averaged over three random runs. ($p < 0.05$ under t-test).

TABLE 6. Effect of different batch size. We report the results of our proposed COD w. ACL on CLINC-10%.

Batch Size	ACC	ARI	NMI
128	95.11	90.23	94.82
256	96.44	92.58	95.71
512	98.22	96.15	97.79

H. EFFECT OF DIFFERENT BATCH SIZE

Table 6 show the effect of different batch size of our proposed COD w. ACL on CLINC-10%. Results show that a larger batch size of input samples obtains a better performance on OOD discovery.

TABLE 7. Effect of different learning objectives.

Models	ACC	ARI	NMI
COD w. ACL	98.22	96.15	97.79
COD	96.44	92.50	95.62
-w/o instance-level	90.93	85.43	92.07
-w/o cluster-level (SimCLR)	90.36	82.91	90.55

I. ABLATION STUDY

In the OOD clustering stage, the intent representation of BERT output is mapped to instance-level and cluster-level subspaces respectively, and optimized with different contrastive losses. In Table 7, we remove two subspaces respectively, where w/o cluster-level means only instance-level contrastive learning used for learning representations,[2] and w/o instance-level means only cluster-level contrastive learning used for learning representations. Results show both instance-level and cluster-level contrastive losses contribute to the performance. When the cluster-level contrastive loss is removed, it is difficult for the model to learn the cluster structure from the unlabeled data, so the performance degradation is the most significant.

VI. CONCLUSION

In this paper, we propose a unified contrastive learning framework for OOD discovery, bridging the gap between pre-training and clustering. For IND pre-training, we employ a supervised contrastive learning (SCL) loss to learn discriminative intent features. For OOD clustering, we introduce an efficient end-to-end contrastive clustering method to jointly learn representations and cluster assignments. Besides, we propose an adaptive contrastive learning (ACL) method to automatically adjust the weights of different negative samples. Experiments on two benchmark datasets prove the effectiveness of our method. And extensive analyses demonstrate our method converges faster and better than the previous SOTA, helps separate semantically similar OOD intents and is robust to different IND data and K. Besides, we find even if the number of OOD clusters is not given, our method still gets relatively accurate estimation and is more

[2] This setting is equivalent to the original SimCLR.

robust to K. We also perform visualization and error analysis to understand the reason for the performance improvements. We hope to explore more self-supervised learning methods for future work.

APPENDIX A
SILHOUETTE COEFFICIENT (SC)

Following [4], we use the cluster validity index (CVI) to evaluate the quality of clusters obtained during each training epoch after clustering. Specifically, we adopt an unsupervised metric Silhouette Coefficient [32] for evaluation:

$$SC = \frac{1}{N}\sum_{i=1}^{N}\frac{b\left(\boldsymbol{I}_i\right) - a\left(\boldsymbol{I}_i\right)}{\max\left\{a\left(\boldsymbol{I}_i\right), b\left(\boldsymbol{I}_i\right)\right\}} \tag{13}$$

where $a\left(\boldsymbol{I}_i\right)$ is the average distance between $\boldsymbol{I}_i$ and all other samples in the i-th cluster, which indicates the intra-class compactness. $b\left(\boldsymbol{I}_i\right)$ is the smallest distance between $\boldsymbol{I}_i$ and all samples not in the i-th cluster, which indicates the inter-class separation. The range of SC is between -1 and 1, and the higher score means the better clustering results.

APPENDIX B
COMPARISON WITH DKT FRAMEWORK

Our proposed COD w. ACL and DKT framework are two different training strategies. They have two differences: (1) In terms of implementation, since the motivation of DKT is to decouple the shared intent representations obtained through BERT into instance-level and cluster-level representations through a multi-head framework, thus DKT maps BERT's output into two subspaces on the model structure. And in the IND pre-training stage and the OOD clustering stage, the contrastive learning objectives is designed respectively to optimize the two subspaces. However, our COD w. ACL does not adopt the multi-head framework in the IND pre-training stage, but directly uses the CE+SCL objective to constrain the representation of BERT output. (2) In terms of method, the clustering algorithm adopted by DKT is contrastive clustering [16], that is, using an instance-level CL and a cluster-level CL to optimize the instance-level and cluster-level subspaces respectively. In this paper, we propose an adaptive contrastive clustering (ACC) method, which improves the problem that traditional instance-level CL will make similar samples be pushed away as negative samples. Adaptive contrastive clustering method, which automatically adjusts the weight of different negative samples according to the semantic similarity of a given anchor, is beneficial to form a more compact cluster distribution, which is one of the innovations of this paper. We also made a theoretical analysis of this in section III-D.

REFERENCES

[1] T.-E. Lin and H. Xu, "Deep unknown intent detection with margin loss," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 5491–5496. [Online]. Available: https://aclanthology.org/P19-1548

[2] H. Xu, K. He, Y. Yan, S. Liu, Z. Liu, and W. Xu, "A deep generative distance-based classifier for out-of-domain detection with Mahalanobis space," in *Proc. 28th Int. Conf. Comput. Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, Dec. 2020, pp. 1452–1460.

[3] Z. Zeng, K. He, Y. Yan, Z. Liu, Y. Wu, H. Xu, H. Jiang, and W. Xu, "Modeling discriminative representations for out-of-domain detection with supervised contrastive learning," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.* Association for Computational Linguistics, 2021, pp. 870–878. [Online]. Available: https://aclanthology.org/2021.acl-short.110

[4] H. Zhang, H. Xu, T.-E. Lin, and R. Lv, "Discovering new intents with deep aligned clustering," in *Proc. AAAI*, 2021, pp. 1–9.

[5] T.-E. Lin, H. Xu, and H. Zhang, "Discovering new intents via constrained deep adaptive clustering with cluster refinement," in *Proc. AAAI*, 2020, pp. 1–8.

[6] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. ECCV*, 2018, pp. 139–156.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Statist. Prob.*, 1967, pp. 281–297.

[8] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5880–5888.

[9] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," 2015, *arXiv:1511.06335*.

[10] D. Hakkani-Tür, Y.-C. Ju, G. Zweig, and G. Tur, "Clustering novel intents in a conversational interaction system with semantic parsing," in *Proc. Interspeech*, Sep. 2015, pp. 1–5.

[11] A. Padmasundari and S. Bangalore, "Intent discovery through unsupervised semantic text clustering," in *Proc. Interspeech*, Sep. 2018, pp. 1–5.

[12] C. Shi, Q. Chen, L. Sha, S. Li, X. Sun, H. Wang, and L. Zhang, "Autodialabel: Labeling dialogue data with unsupervised learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 684–689.

[13] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Minneapolis, MN, USA: Association for Computational Linguistics, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[15] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2020, *arXiv:2004.11362*.

[16] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 1–9.

[17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," 2021, *arXiv:2103.03230*.

[18] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8400–8408.

[19] Z. Zeng, K. He, Y. Yan, H. Xu, and W. Xu, "Adversarial self-supervised learning for out-of-domain detection," in *Proc. NAACL*, 2021, pp. 5631–5639.

[20] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," 2020, *arXiv:2011.01403*.

[21] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," 2021, *arXiv:2104.08821*.

[22] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "ConSERT: A contrastive framework for self-supervised sentence representation transfer," in *Proc. ACL/IJCNLP*, 2021, pp. 5065–5075.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] K. He, Y. Yan, and W. Xu, "Learning to tag OOV tokens by integrating contextual representation and background knowledge," in *Proc. ACL*, 2020, pp. 619–624.

[25] Y. Yan, K. He, H. Xu, S. Liu, F. Meng, M. Hu, and W. Xu, "Adversarial semantic decoupling for recognizing open-vocabulary slots," in *Proc. EMNLP*, 2020, pp. 6070–6075.

[26] K. He, J. Zhang, Y. Yan, W. Xu, C. Niu, and J. Zhou, "Contrastive zero-shot learning for cross-domain slot filling with adversarial attack," in *Proc. COLING*, 2020, pp. 1461–1467.

[27] K. C. Gowda, "A feature reduction and unsupervised classification algorithm for multispectral data," *Pattern Recognit.*, vol. 17, no. 6, pp. 667–676, 1984.

[28] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, pp. 1–6.

[29] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5147–5156.

[30] Y.-C. Hsu, Z. Lv, and Z. Kira, "Learning to cluster in order to transfer across domains and tasks," 2017, *arXiv:1711.10125*.

[31] D. Zhang, F. Nan, X. Wei, S.-W. Li, H. Zhu, K. McKeown, R. Nallapati, A. O. Arnold, and B. Xiang, "Supporting clustering with contrastive learning," in *Proc. NAACL*, 2021, pp. 5419–5430.

[32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[33] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1311–1316. [Online]. Available: https://www.aclweb.org/anthology/D19-1131

[34] I. Casanueva, T. Temcinas, D. Gerz, M. Henderson, and I. Vulic, "Efficient intent detection with dual sentence encoders," in *Proc. 2nd Workshop Natural Lang. Process. Conversational AI*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.nlp4convai-1.5

[35] H. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, pp. 83–97, Mar. 1955.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.

[37] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2495–2504.

[38] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. ICML*, 2020, pp. 1–11.

[39] A. V. Aho and J. D. Ullman, *The Theory of Parsing, Translation and Compiling*, vol. 1. Englewood Cliffs, NJ, USA: Prentice-Hall, 1972.

[40] A. P. Association, *Publications Manual*. Washington, DC, USA: American Psychological Association, 1983.

[41] A. K. Chandra, D. C. Kozen, and L. J. Stockmeyer, "Alternation," *J. Assoc. Comput. Machinery*, vol. 28, no. 1, pp. 114–133, 1981.

[42] G. Andrew and J. Gao, "Scalable training of L1 -regularized log-linear models," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 33–40.

[43] D. Gusfield, *Algorithms on Strings, Trees and Sequences*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[44] M. Sadegh Rasooli and J. Tetreault, "Yara parser: A fast and accurate dependency parser," 2015, *arXiv:1503.06733*.

[45] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, Nov. 2005.

[46] Y. Mou, K. He, Y. Wu, Z. Zeng, H. Xu, H. Jiang, W. Wu, and W. Xu, "Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 46–53.

YUTAO MOU received the first B.S. degree in telecommunication engineering and management from the Beijing University of Posts and Telecommunications and the second B.S. degree (Hons.) from Queen Mary University in London, in 2021. He is currently pursuing the master's degree in frontier research of task-oriented dialogue systems with the Pattern Recognition and Intelligent Systems Laboratory (PRIS), Beijing University of Posts and Telecommunications.

From 2019 to 2020, he was a Research Assistant with the Speech and Language Technology Research Center (CSLT), Tsinghua University, where he focused on the development of speech recognition and speech synthesis systems. He has published seven papers at international conferences on natural language processing, such as ACL2022, EMNLP2022, and COLING2022. His main research interests include open-world natural language understanding and cross-domain dialogue state tracking in the task-oriented dialogue systems, especially the OOD generalization problem under the open domain setting, self-supervised contrastive learning, OOD intent detection, OOD intent discovery, new slot discovery, few-shot/zero-shot slot filling, and dialogue state tracking.

Mr. Mou received several awards and honors, including the Track2 First Place of Semi-Supervised and Reinforced Task-Oriented Dialog Systems Challenge (SereTOD) held on EMNLP2022 Workshop, the First Prize of the 10th Chinese Undergraduate Mathematical Competition, and the Second Prize of the 2020 American Undergraduate Mathematical Modeling Competition.

HEYANG XU is currently pursuing the bachelor's degree in telecommunication engineering and management with the Beijing University of Posts and Telecommunications. Her main research interests include OOD intent detection and OOD intent discovery.