Fine-grained Category Discovery under Coarse-grained supervision with Hierarchical Weighted Self-contrastive Learning

Anonymous ACL submission

Abstract

In this paper, we propose a new task named Fine-grained Category Discovery under Coarse-grained supervision (FCDC). Without asking for any fine-grained knowledge, FCDC aims at discovering fine-grained categories with only coarse-grained labeled data, which can not only reduce significant labeling costs, but also adapt to novel fine-grained categories. It is also a challenging task since performing FCDC requires models to ensure fine-grained sample separability with only coarse-grained supervision and can easily make models overfit on the training set. Considering most current methods cannot transfer knowledge from coarse-grained level to fine-grained level, we propose a novel hierarchical weighted self-contrastive network to approach the FCDC task. Inspired by the hierarchy of pre-trained models (e.g. BERT), we combine supervised learning and contrastive learning to learn fine-grained knowledge from shallow to deep. Specifically, we use coarse-grained labels to train bottom layers of our model to learn surface knowledge, then we build a novel weighted self-contrastive module to train top layers of our model to learn more fine-grained knowledge. Extensive experiments on two public datasets show both effectiveness and efficiency of our model over state-of-the-art methods.

1 Introduction

Fine-grained classification (FGC) training with fine-grained labeled data has attracted much attention in both Natural Language Processing (Munikar et al., 2019; Suresh and Ong, 2021) and Computer Vision (Wei et al., 2019; Gao et al., 2020). However, in real-world scenario, performing FGC usually faces two challenges. On the one hand, FGC methods usually rely on abundant fine-grained labeled data, which is both time and money consuming to obtain. On the other hand, performing FGC task can not discover novel fine-grained categories when data volume increases. So how to perform

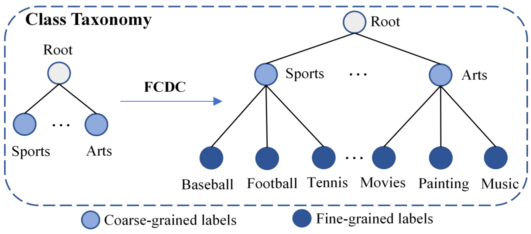

Figure 1: An example of proposed FCDC task (fine-grained clusters are discovered by the FCDC task and fine-grained label names are assigned by experts).

FGC with ability to reduce labeling costs and discover novel fine-grained categories is an important topic.

To meet above requirements, we propose a novel task named Fine-grained Category Discovery under Coarse-grained supervision (FCDC). Different from FGC, performing FCDC only needs coarse-grained labeled data, which is easier to obtain and can reduce significant labeling costs. Furthermore, performing FCDC can discover fine-grained clusters from coarse-grained labeled data and classify inputs into proper fine-grained categories. As shown in Figure 1, at training phase, only coarse-grained (e.g. sports and arts) labeled data is available. Performing FCDC firstly requires models to discover fine-grained clusters (e.g. tennis and music), then experts can assign these clusters with appropriate class names to construct the fine-grained class taxonomy. Finally, models need to predict fine-grained labels of each input in an unsupervised way at testing phase. Since performing FCDC only needs training data with coarse-grained labels, most existing text classification datasets can be directly used.

FCDC is not only more conforming to real-world scenario, but also more challenging than FGC. And the difficulties of solving FCDC task mainly lies in two aspects. Firstly, performing FCDC can easily make models overfit on the training set. Since FCDC needs models to be trained and tested on the same feature space but different label space, models can easily overfit to the coarse-grained classes in the training set (Day and Khoshgoftaar, 2017). So how to fully utilize given coarse-grained supervision meanwhile avoid overfitting is a severe challenge. Secondly, performing FCDC needs models to control both the intra-class and inter-class distance of samples with only coarse-grained supervision. Since coarse-grained classification does not care about intra-class distance (Bukchin et al., 2021), samples with the same coarse-grained labels will be close to each other and hard to be separated in the fine-grained feature space (see Figure 7). So how to control the intra-class distance to ensure fine-grained sample separability is also a serious challenge.

To cope with above challenges, we propose a novel hierarchical weighted self-contrastive network. Inspired by the hierarchy of pre-trained models such as BERT (Devlin et al., 2018) and their ability to extract features from shallow to deep (Xu et al., 2021; Jawahar et al., 2019; Leavitt and Morcos, 2020), the core motivation of our model is to learn coarse-grained knowledge by shallow layers of BERT and learn fine-grained knowledge by the rest of deep layers hierarchically. This motivation is not only consistent with the feature extraction process of pre-trained models, but also corresponding with the learning process of humans. Specifically, we use given coarse-grained labels to train shallow layers of BERT to learn some surface knowledge with supervised learning, then we propose a weighted self-contrastive module to train deep layers of BERT to learn more fine-grained knowledge with contrastive learning.

By performing supervised and contrastive learning on shallow and deep layers, our model can fully utilize given coarse-grained supervision to extract universal features on shallow layers while preserving the ability to extract fine-grained features on deep layers (Cohen et al., 2020), which can mitigate the overfitting problem. To solve the low intra-class differentiation problem, we propose a novel weighted self-contrastive module by introducing a novel strategy to generate positive samples and giving different weights to negative samples, which can better control the inter-class and intra-class distance between samples as well as improve training efficiency of our model (see Section 6.3).

The main contributions of our work can be summarized as threefold:

- To mitigate limitations of the fine-grained classification (FGC) task, we propose a novel task named Fine-grained Category Discovery under Coarse-grained supervision (FCDC), which can reduce labeling costs and adapt to novel fine-grained categories

- We propose a novel model named hierarchical weighted self-contrastive network for the FCDC task. By cooperating supervised learning and weighted self-contrastive learning, our model can ensure both inter-class and intra-class separability to facilitate the FCDC task with higher training efficiency.

- Extensive experiments on two public datasets show that our model significantly advances best compared methods with more than 20% improvement on accuracy and gets double training efficiency than state-of-the-art contrastive learning methods.

2 Related work

2.1 Contrastive learning

Contrastive Learning (CL) aims at grouping similar samples closer and separating dissimilar samples far from each other in a self-supervised way(Le-Khac et al., 2020; Jaiswal et al., 2021; Liu et al., 2021), which has gained popularity in both Natural Language Processing (NLP) (Mikolov et al., 2013; Wu et al., 2020; Meng et al., 2021) and Computer Vision (CV) (Chen et al., 2020a; Chen and He, 2021; Chen et al., 2017). One critical point for CL is to build high-quality positive and negative samples. One simple way to construct negative samples is to use other in-batch data as negatives (Chen et al., 2017). To keep consistency of representations of negatives, He et al. (2020) built a dynamic queue with momentum-updated encoder to make representations of negatives change slowly. However, these methods considered all negatives equally important, which may lose discriminative information of negatives. As for positive samples, in CV, one common way is taking two different transformations of the same image as the query and positive sample (Dosovitskiy et al., 2014). And in NLP, augmentation techniques such as word deletion (Wu et al., 2020), back translation (Sennrich et al., 2015), adversarial attack (Yan et al., 2021) and dropout (Gao et al., 2021) had been proposed to generate positives. Although there are some recent works (Bae et al., 2021; Kim et al., 2021) using outputs from the different levels of a network as positives, which are similar to our self-contrastive strategy, we have different motivations: they aim at providing more high-quality positives for representation learning but we aim at better adjusting intra-class distance for the FCDC task.

2.2 Novel Category Discovery

With data volume increases, novel categories especially novel fine-grained categories may be introduced into datasets (Mekala et al., 2021). To discover novel categories without human annotation, most previous work adopted clustering methods and transfer learning (Pan and Yang, 2009) to generate pseudo labels for unlabeled data to train their models (Zhan et al., 2020). For example, Zhang et al. (2021) proposed an alignment strategy to perform DeepCluster (Caron et al., 2018) to discover novel categories. Ge et al. (2020) proposed a mutual mean teaching network to refine noisy pseudo labels to perform unsupervised person re-identification. Recently, Two similar tasks as ours are proposed. Bukchin et al. (2021) proposed to perform fine-grained image classification under coarse-grained supervision with angular contrastive learning, and they perform this task as a few-shot learning task (Wang et al., 2019) which needs extra fine-grained labels for each categories. Mekala et al. (2021) proposed to perform fine-grained text classification with coarse-grained annotations, and they need extra fine-grained label hierarchy and corresponding surface names to assist in the task. These two tasks both rely on extra fine-grained knowledge from human annotations, which is usually unavailable when novel categories appear in real-world applications. Comparatively, our FCDC task does not require any fine-grained knowledge, which is more adapted to the novel fine-grained category discovery scenarios.

3 Problem Formulation

The proposed FCDC task has two objectives: discovering fine-grained classes from scratch and classifying inputs into proper fine-grained categories. Denote by $\mathcal{Y}_{coarse} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_M\}$ a set of coarse-grained classes. The training set of our problem is a set of texts $\mathcal{D}_{train} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$ with their coarse-grained labels $\{c_1, c_2, ..., c_N\}$, where $c_i \in \mathcal{Y}_{coarse}$. Different from previous tasks (Bukchin et al., 2021; Mekala et al., 2021) where the fine-grained label set $\mathcal{Y}_{fine} = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_K\}$ is already known, FCDC task assumes we do not have any knowledge about fine-grained labels. So performing FCDC requires models to perform clustering methods (e.g. K-Means) to discover fine-grained clusters as well as assign inputs into different fine-grained clusters with only coarse-grained labels. The number of fine-grained clusters k can be estimated by elbow method (Kodinariya and Makwana, 2013) or gap statistic (Tibshirani et al., 2001) and we assume it is known in FCDC following previous works (Lin et al., 2020; Zhang et al., 2021). After discovering fine-grained clusters, experts can assign these clusters with appropriate class names and map these fine-grained classes $\mathcal{Y}_{fine}$ into sub-classes of coarse-grained classes $\mathcal{Y}_{coarse}$. In this way, our task can construct fine-grained class taxonomy (e.g. Figure 1) automatically, in the meanwhile, classify inputs into proper fine-grained categories.

Novel fine-grained categories can be introduced when data volume increases, our task can discover these novel categories by re-estimating the number of clusters k_{novel} and re-clustering based on k_{novel}. Specifically, we first use the algorithm introduced in (Zhang et al., 2021) to estimate the approximate value k_{app}, then we perform clustering with a set of values near k_{app} and select k_{novel} by the unsupervised metric Silhouette Coefficient (Wold et al., 1987). Different from traditional classification tasks which focus on a fixed label set, our task can adapt to novel fine-grained categories and expand the fine-grained label set automatically.

4 Proposed Approach

As shown in Figure 2, our model mainly contains three components: BERT, Dynamic Queue and Momentum BERT. BERT is used to perform supervised learning at Layer L to learn coarse-grained knowledge and perform weighted self-contrastive learning at output layer to learn more fine-grained knowledge. Dynamic Queue can store more negative samples grouping by their coarse-grained labels. Momentum BERT is used to update representations of samples in Dynamic Queue following the settings in MoCo (He et al., 2020). Inspired by the "shallow to deep" learning process of humankind and the ability of pre-trained models to extract features from shallow to deep (Jawahar et al., 2019; Xu et al., 2021), a core motivation of our model is to learn fine-grained knowledge in a progressive way. Specifically, our model can learn coarse-

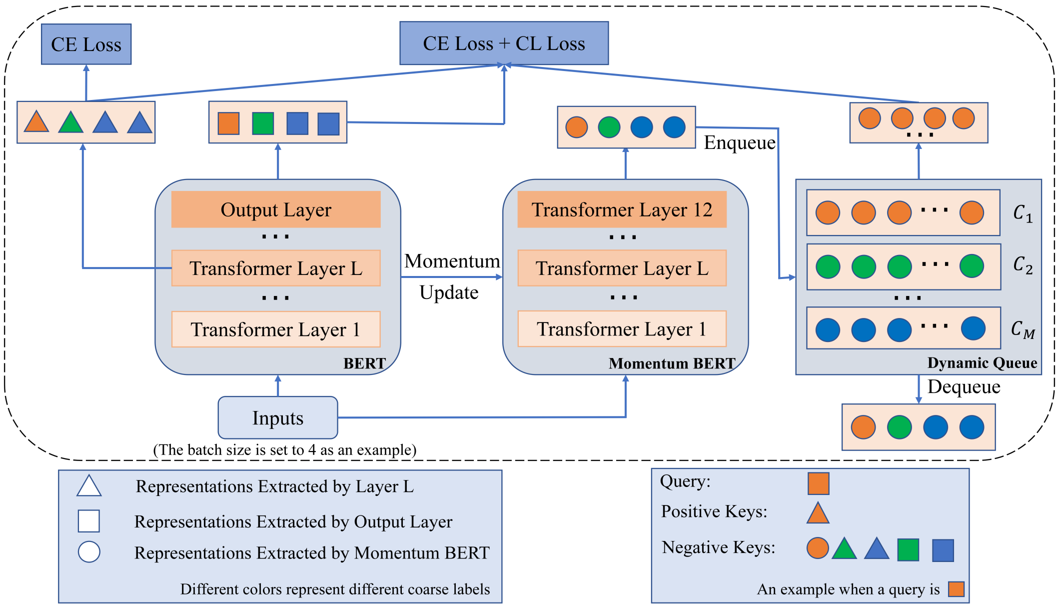

Figure 2: The overall architecture of our model. CE and CL mean Cross Entropy and Contrastive Learning, respectively.

grained knowledge with supervised learning at shallow layers and learn more fine-grained knowledge based on learned coarse-grained knowledge with weighted self-contrastive learning at deep layers.

4.1 Supervised Learning

We firstly perform supervised learning on transformer layer L of BERT to learn coarse-grained knowledge. Given the *i-th* document $\mathcal{D}_i$ with its coarse-grained label c_i, we use all token embeddings from the L layer of BERT as its shallow features. Then we apply a mean-pooling layer to get its shallow feature representation h_i^L:

$$h_i^L = mean\text{-}pooling(BERT_L(\mathcal{D}_i)) \quad (1)$$

where $h_i^L \in \mathbb{R}^h$ is the hidden state of the feature representation, h is the dimension of hidden representations. Then we can perform supervised learning using cross entropy loss on coarse-grained labels to get supervised loss $\mathcal{L}_{sup}^L$:

$$z_i^L = \sigma(W_a h_i^L + b_a) \quad (2)$$

$$\mathcal{L}_{sup}^L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp((z_i^L)^{c_i})}{\sum_{j=1}^{K} exp((z_i^L)^j)} \quad (3)$$

where $z_i^L \in \mathbb{R}^m$ is the output logits, m is the number of coarse classes. σ is the Tanh activation function, $W_a \in \mathbb{R}^{h*m}$ and $b_a \in \mathbb{R}^m$ are learnable weights and bias terms respectively, $(z_i)^j$ is the j-th element of output logits z_i.

4.2 Weighted Self-contrastive Learning

Denote the coarse-grained inter-class and intra-class distance by d_{coarse} and d_{fine}, respectively. Supervised learning on coarse-grained labels can ensure $d_{coarse} \gg 0$ but will also make $d_{fine} \approx 0$, which can bring difficulties for fine-grained categorization. So how to increase d_{fine} to ensure the separability of fine sub-classes is a severe challenge. In the meanwhile, increasing d_{fine} without restraint will result in overlapping between different coarse classes and therefore lead to misclassification. So how to constrain d_{fine} to ensure the proper classification on coarse-grained classes is another challenge. In summary, our total goal can be described as:

$$0 \ll d_{fine} < d_{boundary} \ll d_{coarse} \quad (4)$$

where $d_{boundary}$ is a threshold to ensure samples fall into proper coarse-grained classes.

To achieve above objectives, we propose a weighted self-contrastive module by introducing a novel generation strategy for positive samples and a weighting strategy for negative samples.

4.2.1 Negative Key Generation

Given the *i-th* document $\mathcal{D}_i$, we use all token embeddings from the output layer of BERT as its deep features. As same as the previous extraction process for shallow features, we apply a mean-pooling layer to get its deep feature representation $h_i^o \in \mathbb{R}^h$:

$$h_i^o = mean\text{-}pooling(BERT_o(\mathcal{D}_i)) \quad (5)$$

In-batch negative keys Given h_i^o with its coarse-grained label c_i as a query, we treat both shallow and deep features of other in-batch samples as its in-batch negative keys, where $k_-^{in}(i) = \{h_j^L, h_j^o\}_{j=1...N, j\neq i}$. In this way, we can increase the distance between samples so that satisfying $d_{fine} \gg 0$ and $d_{coarse} \gg 0$. To satisfy $d_{coarse} \gg d_{fine}$, we propose a weighting strategy by giving more weights to samples with the same coarse-grained labels as the query q to decrease their distance. So k_-^{in} can be divided into two groups according to the coarse-grained labels:

$$k_-^{diff}(i) = \{k \in k_-^{in}(i) : c_k \neq c_i\} \quad (6)$$

$$k_-^{same}(i) = \{k \in k_-^{in}(i) : c_k = c_i\} \quad (7)$$

Momentum negative keys To provide more negative keys, we build a momentum BERT and a set of dynamic queues $\{\mathcal{Q}_i\}_{i=1}^M$ to store previous samples grouped by their coarse-grained labels following Bukchin et al. (2021), where *M* is the number of coarse-grained classes. Specifically, given h_i^o with its coarse-grained label c_i as a query, we treat samples from the queue $\mathcal{Q}_{c_i}$ as its momentum negative keys:

$$k_-^m(i) = \{k \in \mathcal{Q}_{c_i}\} \quad (8)$$

Feature representations of samples in dynamic queues are extracted by momentum BERT, and the parameters of momentum BERT are updated in a momentum way following He et al. (2020). At the end of each iteration, the dynamic queues will be updated by adding novel samples and removing the earliest samples. Since samples in $k_-^m(i)$ have the same coarse-grained label as the query, they are much harder to be separated and beneficial to better representation learning.

The overall negative keys for the query h_i^o is :

$$k_-(i) = \{k_-^{diff}(i), k_-^{same}(i), k_-^m(i)\} \quad (9)$$

4.2.2 Positive Key Generation

By weighting different negative samples, we can satisfy the condition $0 \ll d_{fine} \ll d_{coarse}$. But increasing d_{fine} without restraint will violate the condition $d_{fine} < d_{boundary}$ and make some samples fall into incorrect coarse-grained classes. To solve this problem, we propose a self-contrastive strategy by treating shallow features of a query as its positive keys. Specifically, given the deep feature representation h_i^o for document $\mathcal{D}_i$ as a query, we treat h_i^L as its positive key:

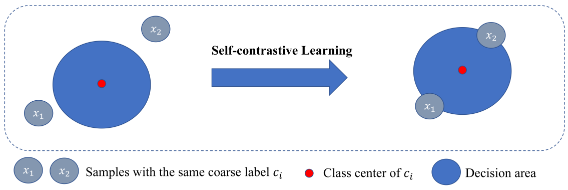

Figure 3: The effectiveness of our self-contrastive module, which can ensure both inter-class distance and proper coarse-grained classification.

$$k_+(i) = h_i^L \quad (10)$$

After supervised learning on coarse-grained labels, h_i^L can be very close to the class center of c_i, so pulling h_i^o close to h_i^L will also pull h_i^o close to the class center of c_i. In this way, we can increase d_{fine} with restraint and satisfy the condition $d_{fine} < d_{boundary}$ without computing the value of $d_{boundary}$, which is shown in Figure 3. Another advantage for our self-contrastive strategy is that we can get double training efficiency than traditional data augmentation methods (Wu et al., 2020; Gao et al., 2021) since we only need to perform feed-forward and back-forward propagation only once to get and update both queries and positive keys. (discussed in Section 6.3)

4.2.3 Contrastive Loss

Given the query h_i^o with its positive key $k_+(i)$ and negative keys $k_-(i)$, the contrastive loss of our weighted self-contrastive module is:

$$\mathcal{L}_{cont} = \sum_{i=1}^{N} -log \frac{e^{sim(h_i^o, h_i^L)/\tau}}{\sum_{l \in k_-(i)} \alpha_l \sum_{k \in l} e^{sim(h_i^o, h_k)/\tau}} \quad (11)$$

where $\{\alpha_l\}$ are weighting factors for different negative keys, $sim(h_i, h_j)$ is cosine similarity $\frac{h_i^T h_j}{\|h_i\| \cdot \|h_j\|}$ and τ is a temperature hyperparameter.

By weighting different negative keys and selecting shallow features as positive keys, our weighted self-contrastive module can satisfy the goal in Inequation 4 and provide conditions for subsequent fine-grained categorization.

4.3 Overall Loss

We further find that adding supervised learning on coarse-grained labels at the output layer can boost

Dataset	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{F}\rvert$	# Train	# Dev	# Test
CLINC	10	150	18,000	1,000	10,00
WOS	7	33	8,362	1,185	2,420

Table 1: Statistics of datasets. # indicates the number of samples in each set. $\lvert\mathcal{C}\rvert$, $\lvert\mathcal{F}\rvert$ means the number of coarse-grained and fine-grained classes, respectively.

our model performance, since it can guarantee samples to be classified into proper coarse-grained categories. So the overall loss for our hierarchical weighted self-contrastive network is:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{sup}^{L} + \gamma_2 \mathcal{L}_{sup}^{o} + \gamma_3 \mathcal{L}_{cont} \tag{12}$$

where $\mathcal{L}_{sup}^{o}$ is the cross entropy loss at the output layer and γ_1, γ_2, γ_3 are weighting factors.

By performing supervised learning on shallow layers and weighted self-contrastive learning on deep layers, our model can learn fine-grained knowledge based on learned coarse-grained knowledge and ensure both inter-class and intra-class separability to facilitate FCDC task.

5 Experiments

5.1 Datasets

To evaluate effectiveness of our model, we conduct experiments on two public datasets. Statistics of two datasets can be found in Table 1.

CLINC is an intent classification dataset released by Larson et al. (2019).

Web of Science (WOS) is a paper classification dataset released by Kowsari et al. (2017). And we use the WOS-11967 version.

5.2 Compared Methods

Since FCDC needs models to discover fine-grained categories with no fine-grained labeled data, We compare our model with a set of self-supervised methods.

Baselines We firstly perform FCDC with BERT in unsupervised way, coarse-supervised way and fine-supervised way as baselines.

Self-supervised Methods DeepCluster (Caron et al., 2018), CDAC+ (Lin et al., 2020) and DeepAligned (Zhang et al., 2021) are self-supervised methods using self-training techniques and achieve state-of-the-art results in many category discovery tasks. Ancor (Bukchin et al., 2021) is a self-supervised method designed for few-shot fine-grained classification with coarse-grained labels. SimCSE (Gao et al., 2021) and Delete One Word (Wu et al., 2020) are contrastive learning methods in NLP with different data augmentation techniques and achieve good performance in many representation learning tasks. For a fair comparison, we use the same BERT model as ours to extract features for all compared methods and adopt hyper-parameters in their original paper.

Self-supervised + Cross Entropy To investigate the influence of coarse-grained supervision on compared models, we further add the cross entropy loss on coarse-grained labels to their loss function.

5.3 Evaluation Metrics

Since no fine-grained knowledge is available for FCDC task, we need to perform clustering to discover fine-grained categories. To evaluate the performance of clustering, we use two broadly used evaluation metrics: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). To evaluate the performance of fine-grained classification, we use the metric Accuracy (ACC), which is obtained by Hungarian algorithm (Kuhn, 1955).

5.4 Main Results

The average results over 5 runs are reported in Table 2. From the results we can draw following conclusions.

Our model significantly outperforms other compared methods across all datasets. We get more than 20% improvement on metrics ACC and ARI, and more than 10% improvement on the metric NMI. We contribute the reasons of better performance of our model to the following three points. Firstly, we propose a hierarchical architecture to learn fine-grained knowledge from shallow to deep, which is consistent with both the feature extraction process of pre-trained language models and the learning process of human beings. Secondly, we perform supervised learning with coarse-grained labels at shallow layers, which can help to learn coarse-grained knowledge and lay the foundation for learning fine-grained knowledge on deeper layers. Thirdly, we propose a weighted self-contrastive module to better learn fine-grained knowledge at deep layers. Specifically, we propose a weighting strategy for negative samples to better control both inter-class and intra-class distance, and in the meanwhile, we propose a self-contrastive strategy to generate positive samples so that we can avoid the overlap between different coarse classes and meanwhile get double training efficiency than traditional contrastive methods.

Methods	CLINC			WOS		
	ACC	ARI	NMI	ACC	ARI	NMI
Unsupervised	33.38	16.42	63.46	32.32	18.21	47.12
Coarse Supervised	45.91	32.27	75.04	39.42	33.67	61.60
Fine Supervised	96.84	95.03	98.50	83.64	72.01	81.46
CDAC+	25.44	13.06	62.21	23.97	12.14	36.56
DeepCluster	26.40	12.51	61.26	29.17	18.05	43.34
DeepAligned	29.16	14.15	62.78	28.47	15.94	43.52
SimCSE	40.22	23.57	69.02	25.87	13.03	38.53
Ancor	45.60	33.11	75.23	41.20	37.00	65.42
Delete One Word	47.11	31.28	73.39	24.50	11.68	35.47
DeepCluster + CE	30.28	13.56	62.38	38.76	35.21	60.30
CDAC+ + CE	34.40	17.73	64.21	32.32	18.21	47.12
DeepAligned + CE	42.09	28.09	72.78	39.42	33.67	61.60
Ancor + CE	44.44	31.50	74.67	39.34	26.14	54.35
Delete One Word + CE	47.87	33.79	76.25	41.53	33.78	61.01
SimCSE + CE	52.53	37.03	77.39	41.28	34.47	61.62
Ours	**74.15**	**64.67**	**89.00**	**68.00**	**56.15**	**73.73**

Table 2: Model comparison results (%) on test sets. Average ACC, ARI and NMI over 5 runs are reported. '+ CE' means adding coarse-grained supervision with cross entropy loss. The statistical significance test results are shown in Appendix A.2 and all the p-values are less than 10^{-8}, which means our improvement is significant.

Fine-supervised BERT can be seen as upper bound of the FCDC task since it trains models with fine-grained labeled data. Self-training methods perform badly on all datasets and evaluation metrics since they rely on abundant labeled data to generate high-quality pseudo labels for unlabeled data. Contrastive learning methods perform better than self-training methods since they do not need fine-grained labels to initialize their models. However, their performance is still much worse than ours since they can not fully utilize given coarse-grained labels to control inter-class and intra-class distance between samples. We can also find that model performance of most compared methods increases with the addition of coarse-grained supervision, which means coarse-grained supervision can boost model performance on fine-grained tasks.

6 Discussion

6.1 Ablation Study

To investigate contributions of different components to our model, we compare the performance of our model with its variants on the the CLINC dataset. As shown in Table 3, removing different components from our model will affect model performance more or less, which can indicate the effectiveness of different components in our model. Removing Momentum Encoder has minimal impact on our model, since our model is insensitive to the number of negative samples (More details in Appendix A.4). Removing weighting strategy or cross entropy loss at shallow layers will also hurt model performance since they can help to learn coarse-grained knowledge and lay the foundation for learning fine-grained knowledge. Above all, removing self-contrastive strategy or cross entropy loss at output layer results in a significant decrease in model performance, since these two components are responsible for controlling intra-class and inter-class distance, respectively, which are two most important objectives for the FCDC task.

Table 3: Results (%) of different model variants. '-' means that we remove the component from our model.

Model	ACC	ARI	NMI
ALL	74.15	64.67	89.00
- Momentum	72.06	62.71	88.52
- Weighting	71.75	62.99	88.47
- $\mathcal{L}_{sup}^{L}$	71.02	62.22	87.50
- Self-Contrast	53.21	40.05	75.36
- $\mathcal{L}_{sup}^{o}$	50.27	32.65	74.51

6.2 Novel Fine-grained Category Discovery

As introduced in Section 3, performing the FCDC task can discover novel fine-grained categories

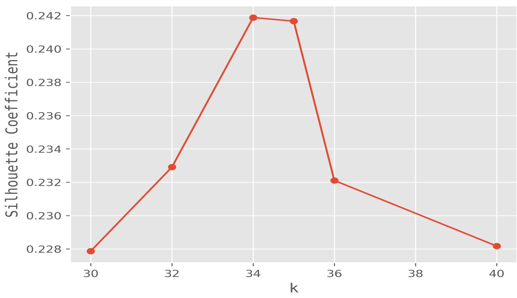

Figure 4: Impact of different batch sizes on our model.

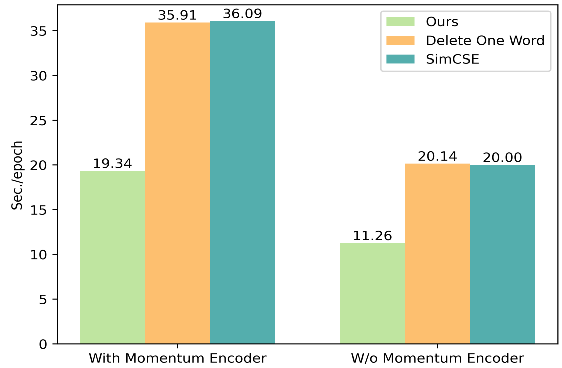

Figure 5: Training efficiency compared with other contrastive methods.

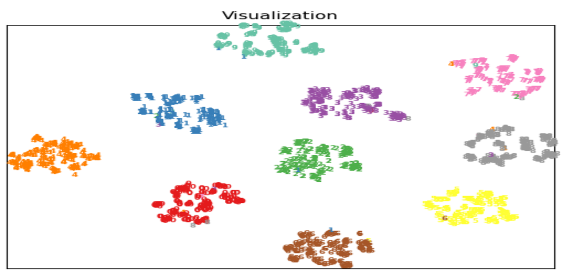

Figure 6: TSNE visualization of representations learned by our model. Each color indicates a ground-truth coarse-grained category.

from novel data. We perform experiments on the WOS dataset by randomly setting 4 fine-grained categories as novel categories and corresponding data as novel data. The approximate value k_{app} estimated by our model is 35. Then we perform clustering with a set of $k = \{30, 32, 34, 35, 36, 40\}$, and the results are shown in Figure 4. The number of fine-grained categories k_{novel} estimated by our model equals to the ground truth 34, which can show the effectiveness of our model.

6.3 Training Efficiency

In this section, we compare the training efficiency of our model with contrastive methods SimCSE and Delete One Word on the CLINC dataset. We test all methods using the BERT base model trained on the same hardware platform (an AMD EPYC CPU 7702 and a RTX 3090 GPU) with the batch size 128. Average results over 100 epochs are shown in Figure 5. Compared with SimCSE and Delete One Word, our model gets double training efficiency both when adding or removing Momentum Encoder, which benefits from our self-contrastive strategy. Traditional contrastive methods like SimCSE rely on data augmentation techniques to generate positive keys, which needs to perform feed-forward and back-forward propagation twice for queries and keys, respectively. Comparatively, our model utilizes shallow features of queries as positive keys, which only needs to perform feed-forward and back-forward propagation once to get and update both queries and positive keys.

6.4 Visualization

We visualize the learned embeddings of our model on the CLINC dataset using t-SNE (Van der Maaten and Hinton, 2008) in Figure 6. It can be seen that our model can ensure both inter-class and intra-class distance to facilitate the FCDC task. Specifically, our model can separate different coarse-grained categories with a large margin benefiting from the supervised learning on coarse-grained labels. In the meanwhile, different from traditional supervised learning methods which usually ignore the intra-class distance, our model can better increase the distance of samples within the same coarse-grained categories to ensure the intra-class separability, which benefits from the proposed weighted self-contrastive module.

7 Conclusion

In this paper, we propose a novel task named Fine-grained Category Discovery under Coarse-grained supervision (FCDC), which can reduce significant labeling costs and adapt to novel fine-grained categories. We further propose a hierarchical weighted self-contrastive network to approach the FCDC task. By performing multi-task learning on shallow and deep layers of pre-trained models, our model can learn fine-grained knowledge from shallow to deep with only coarse-grained supervision. Extensive experiments on two public datasets show that our approach is more effective and efficient than state-of-the-art methods.

References

Sangmin Bae, Sungnyun Kim, Jongwoo Ko, Gihun Lee, Seungjong Noh, and Se-Young Yun. 2021. Self-contrastive learning. *arXiv preprint arXiv:2106.15499*.

Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. 2021. Fine-grained angular contrastive learning with coarse labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8730–8740.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. 2020. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):1–13.

Oscar Day and Taghi M Khoshgoftaar. 2017. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4:1–42.

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

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27:766–774.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Yu Gao, Xintong Han, Xun Wang, Weilin Huang, and Matthew Scott. 2020. Channel interaction networks for fine-grained image categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10818–10825.

Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies*, 9:2.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. *arXiv preprint arXiv:2106.07345*.

Trupti M Kodinariya and Prashant R Makwana. 2013. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, , Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2:83–97.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*.

Matthew L Leavitt and Ari Morcos. 2020. Selectivity considered harmful: evaluating the causal impact of class selectivity in dnns. *arXiv preprint arXiv:2003.01262*.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8360–8367.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Dheeraj Mekala, Varun Gangal, and Jingbo Shang. 2021. Coarse2fine: Fine-grained text classification on coarsely-grained annotated data. *arXiv preprint arXiv:2109.10856*.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Sebastian Raschka, Joshua Patterson, and Corey Nolet. 2020. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Varsha Suresh and Desmond C Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. *arXiv preprint arXiv:2109.05427*.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Xiu-Shen Wei, Jianxin Wu, and Quan Cui. 2019. Deep learning for fine-grained image analysis: A survey. *arXiv preprint arXiv:1907.03069*.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

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

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. 2021. Unsupervised out-of-domain detection via pre-trained transformers. *arXiv preprint arXiv:2106.00948*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. 2020. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6688–6697.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Appendix

A.1 Implementation Details

We use the pre-trained BERT model (bert-base-uncased) implemented by Pytorch (Wolf et al., 2020) as our backbone and adopt most of its suggested hyper-parameters. We also freeze most of its model parameters and only fine-tune the last four transformer layers to speed up calculations. We use the cuml library (Raschka et al., 2020) to perform K-Means on GPU to speed up calculations. Early stopping is used in our experiment, which is decided by model performance on the validation set. We use the AdamW optimizer with 0.01 weight decay. Gradient clipping is also used with the norm 1.0. For hyperparameters, temperature τ is set to 0.1, layer L is set to 11, and the weighting

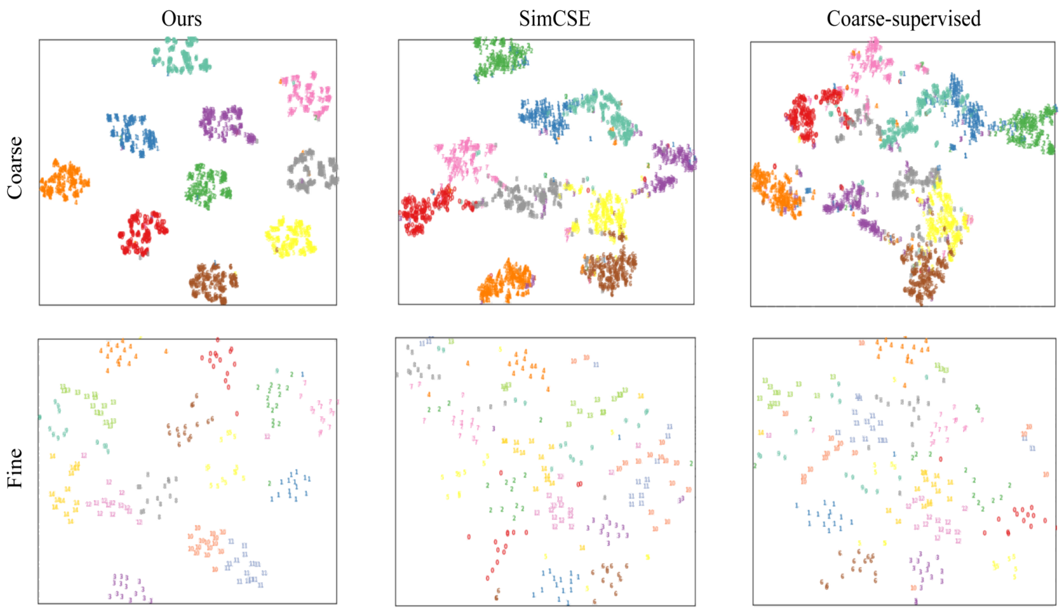

Figure 7: TSNE visualization of representations learned by our model, SimCSE and Coarse-supervised BERT. Top : each color indicates a ground-truth coarse-grained category. Bottom: each color indicates a ground-truth fine-grained category.

factors α_l for $\{k_-^{diff}(i), k_-^{same}(i), k_-^{m}(i)\}$ are set to $\{1.0,\ 1.1,\ 1.0\}$, weighting factors $\{\gamma_1, \gamma_2, \gamma_3\}$ are set to $\{0.001,\ 1,\ 0.008\}$. The training batch size is set to 128, and the testing batch size is set to 64. The momentum queue size for each coarse-grained category is set to 128, and the momentum factor for Momentum BERT is set to 0.9. The hidden dimension h is 768, the learning rate is set to $5e^{-5}$, the dropout rate is set to 0.1. The maximum training epoch is set to 100 and the wait patience for early stopping is set to 10 for all models.

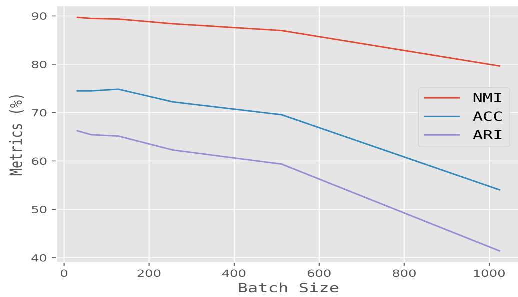

Figure 8: Impact of different batch sizes on our model.

A.2 Statistical Significance Results

To assess the significance of our experimental results, we perform t-tests between our model and other compared methods on all datasets and evaluation metrics. The p-values are shown in Table 4 and Table 5. Specifically, the p-values are distributed between 10^{-16} to 10^{-9}, so we can conclude that the performance improvement of our model over compared methods is statistically significant.

A.3 Impact of Batch Sizes

To investigate the influence of batch sizes on our model, we plot the figure of model performance with different batch sizes. As shown in Figure 8, the performance of our model shows similar decreasing tendency on three metrics. Different from traditional insight that contrastive learning benefits from larger batch sizes (Chen et al., 2020b), larger batch sizes are harmful to our model. When batch size is small (< 128), our model gets the best performance. As batch size increases, our model performance drops quickly, especially when batch size is larger than 512. One possible reason is that when batch size increases, it will be difficult to control the distance between samples in the fine-grained feature space to ensure both inter-class and intra-class separability.

A.4 Impact of Momentum Queue Sizes

To investigate the influence of Momentum Queue size on our model, we plot the figure of model performance with different Momentum Queue sizes on CLINC dataset in Figure 9. The performance of our model does not change much with different Momentum Queue sizes on all three metrics. Since different Momentum Queue sizes mean different

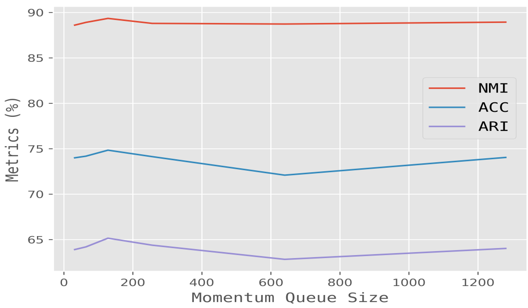

Figure 9: Impact of Momentum Queue Sizes.

number of negative samples that Momentum Queue can provide, we can draw the conclusion that our model is not sensitive to the number of negatives, which is consistent with the conclusion in Section 6.1. The insensitivity to negative samples of our model can ensure that it works well even with small data volume or limited hardware resource.

A.5 Visualization

We further visualize the learned embeddings of our model and compared methods using t-SNE in Figure 7. Firstly, our model can separate different coarse-grained categories with a larger margin than SimCSE and Coarse-supervised BERT (Top in Figure 7), which benefits from our strategy of combining supervised learning and contrastive learning in a hierarchical way. Furthermore, our model can also separate different fine-grained categories with a larger margin than SimCSE and Coarse-supervised BERT (Bottom in Figure 7). Compared with traditional supervised learning methods and contrastive learning methods, our model can better increase distance of samples from different fine-grained categories to ensure the intra-class separability, which benefits from the proposed weighted self-contrastive module. In summary, our model can better control both inter-class and intra-class distance than traditional supervised learning methods and contrastive learning methods to perform the FCDC task.

Methods	ACC	ARI	NMI
CDAC+	1.4×10^{-12}	6.7×10^{-13}	7.3×10^{-13}
DeepCluster	3.7×10^{-12}	1.7×10^{-13}	5.5×10^{-13}
DeepAligned	5.2×10^{-12}	5.6×10^{-14}	8.7×10^{-13}
SimCSE	5.8×10^{-11}	3.4×10^{-14}	7.6×10^{-12}
Ancor	2.2×10^{-10}	4.6×10^{-12}	1.5×10^{-10}
Delete One Word	2.0×10^{-10}	3.2×10^{-12}	5.4×10^{-11}
DeepCluster + CE	4.9×10^{-11}	1.7×10^{-13}	7.7×10^{-13}
CDAC+ + CE	9.5×10^{-11}	3.3×10^{-13}	2.3×10^{-10}
DeepAligned + CE	9.5×10^{-11}	2.4×10^{-12}	4.0×10^{-11}
Ancor + CE	1.6×10^{-10}	5.3×10^{-12}	1.1×10^{-10}
Delete One Word + CE	2.4×10^{-10}	9.4×10^{-12}	4.8×10^{-11}
SimCSE + CE	6.4×10^{-9}	2.3×10^{-11}	2.1×10^{-12}

Table 4: Statistical significance results on CLINC dataset.

Methods	ACC	ARI	NMI
CDAC+	1.0×10^{-12}	2.8×10^{-13}	6.7×10^{-16}
DeepCluster	2.7×10^{-12}	8.7×10^{-13}	3.3×10^{-13}
DeepAligned	9.1×10^{-13}	5.7×10^{-13}	5.7×10^{-12}
SimCSE	8.7×10^{-11}	3.2×10^{-13}	1.0×10^{-12}
Ancor	1.4×10^{-10}	2.1×10^{-11}	1.1×10^{-10}
Delete One Word	1.1×10^{-12}	2.5×10^{-13}	5.3×10^{-15}
DeepCluster + CE	6.5×10^{-12}	1.0×10^{-13}	2.3×10^{-10}
CDAC+ + CE	5.4×10^{-12}	9.0×10^{-16}	9.7×10^{-12}
DeepAligned + CE	7.7×10^{-12}	5.9×10^{-12}	5.2×10^{-10}
Ancor + CE	7.6×10^{-12}	5.9×10^{-13}	1.2×10^{-11}
Delete One Word + CE	1.4×10^{-11}	6.2×10^{-12}	3.5×10^{-10}
SimCSE + CE	1.3×10^{-11}	7.9×10^{-11}	1.1×10^{-10}

Table 5: Statistical significance results on WOS dataset.