Novel Class Discovery in Chest X-rays via Paired Images and Text

Jiaying Zhou$^{1, 2}$, Yang Liu$^3$, Qingchao Chen$^{1, 2, 4}$ *

$^1$National Institute of Health Data Science, Peking University, Beijing, China
$^2$Institute of Medical Technology, Peking University Health Science Center, Beijing, China
$^3$Wangxuan Institute of Computer Technology, Peking University, Beijing, China
$^4$National Key Laboratory of General Artificial Intelligence, Beijing, China

{zhoujiaying, qingchao.chen, yangliu}@pku.edu.cn

Abstract

Novel class discovery (NCD) aims to identify new classes undefined during model training phase with the help of knowledge of known classes. Many methods have been proposed and notably boosted performance of NCD in natural images. However, there has been no work done in discovering new classes based on medical images and disease categories, which is crucial for understanding and diagnosing specific diseases. Moreover, most of the existing methods only utilize information from image modality and use labels as the only supervisory information. In this paper, we propose a multi-modal novel class discovery method based on paired images and text, inspired by the low classification accuracy of chest X-ray images and the relatively higher accuracy of the paired text. Specifically, we first pretrain the image encoder and text encoder with multi-modal contrastive learning on the entire dataset and then we generate pseudo-labels separately on the image branch and text branch. We utilize intra-modal consistency to assess the quality of pseudo-labels and adjust the weights of the pseudo-labels from both branches to generate the ultimate pseudo-labels for training. Experiments on eight subset splits of MIMIC-CXR-JPG dataset show that our method improves the clustering performance of unlabeled classes by about 10% on average compared to state-of-the-art methods. Code is available at: https://github.com/zzzzzzzzjy/MMNCD-main.

Figure 1: (a) Pseudo-label accuracy of different methods and different datasets. (b) Validation set accuracy of unlabeled classes in SET1 under supervision.

Introduction

The success of deep learning based classification methods greatly depends on labeled data. However, it is difficult to gather high-quality labeled data, especially for medical data. And in real-world scenarios, it is almost impossible to collect labeled data for all classes because of missing definitions, vague categories, infinite categories, etc. To address this problem, a new paradigm called novel class discovery (NCD) has been proposed and gained significant attention due to its potential applications in various domains, such as surveillance, medical image analysis, and anomaly detection.

Given a labeled set, the goal of NCD is to discover undefined categories in the unlabeled set, which distinguishes it from semi-supervised learning. Most of the existing methods start with supervised pretraining on the labeled set, while others adopt self-supervised pretraining on the whole dataset. Two-stage methods then employ learned similarity prediction networks or feature extraction networks to classify the unlabeled set through clustering or (pairwise) pseudo-labeling (Hsu, Lv, and Kira 2017; Hsu et al. 2019; Han, Vedaldi, and Zisserman 2019). Some one-stage methods tune the feature representation while classifying/clustering by using different objective functions for labeled set and unlabeled set (Han et al. 2020; Zhong et al. 2021a; Han et al. 2021). Others unify the objective function as cross-entropy loss by assigning pseudo-labels to unlabeled samples (Fini et al. 2021; Li et al. 2022a; Yang et al. 2022). In this paper, we mainly focus on one-stage methods that unify the objective function through pseudo-labels.

Despite abundant research on the topic of NCD, most of these works are conducted on benchmark datasets of natural images. The usability and effectiveness of the NCD methods on large-scale medical image datasets are not yet known. NCD for medical images is of great importance for disease diagnosis and precision medicine because of its ability to uncover new disease types or unknown disease subtypes. In this paper, we focus on NCD in Chest X-Ray (CXR) images that are widely used in medical diagnostics for detecting lung diseases, tumors, and other abnormalities.

The success of pseudo-label-based one-stage methods relies heavily on the quality of the pseudo-labels (Lee et al. 2013). Existing methods (Fini et al. 2021; Li et al. 2022a)
formulate pseudo-label assignment as an optimal transport problem that can be solved by Sinkhorn-Knopp algorithm (Cuturi 2013). Also, they improve the reliability of pseudo-labels by constraining the consistency of the predictions of the image and its augmented view. However, the pseudo-label assignment accuracy of these methods on the CXR dataset is much lower than that on the natural image datasets with a similar number of classes, as shown in Fig. 1. We speculate that this may be due to the highly similar appearances of CXR images from different anatomical structures or classes. With these pseudo-labels, the mainstream methods do not achieve good results on the CXR dataset. Meanwhile, we find that the classification accuracy of CXR images under supervision is also not very good, at least not as good as text classification, as shown in Fig. 1. Inspired by this, we hold the hypothesis that text can help generate better pseudo-labels in some cases and aim to investigate the usage of the descriptive text paired with CXR images to improve performance. We would like to solve the following research questions: when does one modality provide better pseudo-labels than another and how to boost the image NCD performance utilizing the advantageous text information in the training process.

To answer the two questions and innovate text-augmented CXR NCD, it still entails three technical challenges: given the heterogeneous cross-modal semantic gap, (1) how to quantify the quality of pseudo-labels calculated by visual and text modalities; (2) how to transfer useful information between two modalities to improve CXR NCD performance through text; (3) the unavailable text at the test time in the real-world deployments.

Constrained by unavailable text at the test time, instead of multi-modal feature fusion, we propose to use a two-branch network, encoding the visual and text features separately. To quantify the pseudo-label quality of two modalities, we propose a novel measurement – the consistency between semantic feature structure and the pseudo-label structure in each modality. The quality of the pseudo-labels in NCD relies on two aspects: (1) how much the abstract features can distinguish the unknown classes; (2) how the semantic patterns in the local feature embeddings are able to identify and comprehend the relationship between the known and the novel classes. Pseudo-labels or the features close to them may represent abstract and categorical information but tend to lose local visual semantic features. Therefore, to bring advantages from both worlds, i.e. the categorical and local semantic features, we propose to utilize their structural consistency for pseudo-label quality identification. By comparing the consistencies in both modalities, we can identify when to transfer information from one modality to another.

To integrating effective information from both modalities and reduce the cross-modal semantic gap, we propose to synthesize pseudo-labels that guide both visual and text NCD. The synthetic pseudo-labels are the linear combination of the visual and text ones, using the quantified consistency scores respectively. The potential reason is that visual and text features are heterogeneous, although the cross-modal alignment losses are performed (Liang et al. 2022), direct transfer/distillation or feature alignment may result in degraded NCD performance. Our contributions can be summarized as follows:

- We explore NCD in CXR images and propose a method based on paired images and text.
- We introduce intra-modal consistency as a basis for pseudo-label quality measuring and weighting.
- Based on the MIMIC-CXR-JPG dataset, we set up two benchmarks which have the same known classes and different new classes. We evaluate the proposed methods on eight data splits of the two benchmarks and demonstrate significant performance improvements over the state-of-the-art methods.

**Related Work**

**Novel Class Discovery**

Novel class discovery (NCD) aims to discover new classes in an unlabeled dataset given different but related labeled classes. Existing methods can be divided into two categories: two-stage methods and one-stage methods.

The pioneering works of NCD are two-stage methods, including KCL (Hsu, Lv, and Kira 2017), MCL (Hsu et al. 2019) and DTC (Han, Vedaldi, and Zisserman 2019). The KCL and MCL utilize similarity prediction networks to generate pairwise pseudo-labels and leverage the clustering models to classify unlabeled data. The two stages adopt different objective functions. DTC first trains a model with supervised learning on the labeled set and then discovers novel visual categories using DEC (Xie, Girshick, and Farhadi 2016). MM/MP (Chi et al. 2021) trains a group of classifiers on the labeled set and fine-tunes classifiers on the unlabeled set.

Compared to two-stage methods, one-stage methods received more attention in the field recently. One-stage methods use both labeled data and unlabeled data simultaneously at some point in the optimization process. RS/AutoNovel (Han et al. 2020, 2021) may be the first work among the one-stage methods. It uses the pairwise similarity obtained by ranking statistics as supervision to discover novel classes. Follow-up work DualRS (Zhao and Han 2021) expands this method to a two-branch framework focusing on both local and global features. Afterward, NCL (Zhong et al. 2021a) further boosts the performance by leveraging the framework of contrastive learning. In addition, OpenMix (Zhong et al. 2021b) uses MixUp (Zhang et al. 2017) to generate more robust pseudo-labels for the unlabeled data.

Other one-stage methods eliminate the use of pairwise pseudo-labels and directly assign pseudo-labels to unlabeled samples. UNO (Fini et al. 2021) may be the first of these works. UNO unifies the training objective by using a multi-view self-labeling strategy to generate pseudo-labels that can be treated homogeneously with ground truth labels. Based on UNO, IIC (Li et al. 2022a) models both inter-class and intra-class constraints based on the symmetric Kullback-Leibler divergence. ComEx (Yang et al. 2022) focuses on the generalized setting of NCD (GNCD) and classifies the data with two complementary groups of classifiers with global-to-local and local-to-local regularization.
to strengthen pseudo-labels. Similar to the ComEx, some work focuses on generalized class discovery (GCD). Among them, CLIP-GCD is the first work to combine multi-modal (image and text) models in GCD. CLIP-GCD proposes a retrieval-based mechanism that leverages CLIP’s aligned visual-language representations.

Different from the previous works, we propose to solve a novel task, text-augmented NCD in medical image analysis. Our work focuses on quantifying the quality of pseudo-labels from both modalities and integrating effective information from both text and image to boost CXR NCD performance. We did not adopt the solution to retrieve text annotations for medical images from the text corpus due to unavailable cross-modal pre-trained models like CLIP for medical images. In addition, the medical image exhibits unique challenges for NCD, i.e., high semantic similarity in the anatomical structures. To our best knowledge, this is the first work that tackles the NCD in medical image analysis using both image and text.

Pseudo-Labeling in Semi-Supervised Learning

Our work is related to part of the work on semi-supervised learning involving pseudo-labeling. Among them, MixMatch (Berthelot et al. 2019b) averages and sharpens the predictions of multiple strongly augmented views as pseudo-labels. ReMixMatch (Berthelot et al. 2019a) proposes to generate the pseudo-labels with weakly augmented views and align the pseudo-label distribution with the marginal distribution of ground-truth labels. Instead of using all pseudo-labels, FixMatch (Sohn et al. 2020) retains only those with high confidence. SoftMatch (Chen et al. 2023) overcomes the trade-off between quantity and quality of pseudo-labels with truncated Gaussian weighting function and uniform alignment.

Different from these methods, we do not employ data augmentation and rely solely on the images and paired text to generate pseudo-labels. Our contribution lies in not only the quality quantification of pseudo-labels but also in proposing a new strategy to integrate two modality information for joint learning.

Method

Overall

Problem Formulation: Similar to the image-only NCD setting, our training data are split into a labeled set and an unlabeled set. The labeled set $D^l = \{ (v_1^l, t_1^l, y_1^l), ..., (v_M^l, t_M^l, y_M^l) \}$ contains paired image and text $(v, t)$ with corresponding label $y^l$ from $C^l$ classes. The unlabeled set $D^u = \{ (v_1^u, t_1^u), ..., (v_M^u, t_M^u) \}$ contains unpaired image and text $(v^u, t^u)$ from $C^u$ classes, where $C^u$ is known as a prior. The set of $C^l$ labeled classes is disjoint with the set of $C^u$ unlabeled classes. The purpose of NCD is to discover $C^u$ clusters in the unlabeled set. Following the UNO (Fini et al. 2021), we formulate this problem as learning a mapping from the sample to the complete-label set $Y = \{ 1, ..., C^l, C^l + 1, ..., C^l + C^u \}$. To generalize the method to real-world scenarios, we assume that the text is not available at the test time.

Architecture: We propose a method based on paired images and text, using the intra-modal structural consistency to generate and adjust pseudo-labels. Our network architecture is shown in Fig.2, where it consists of two branches: the image branch and the text one.

Given a CXR image $v$, the semantic embedding $z_v \in \mathbb{R}^k$ is firstly obtained via the visual encoder $E_v$, and the projection head $Proj_v$, i.e., $z_v = Proj_v(E_v(v))$. Then, the two visual classification heads, labeled head $h_v$ and unlabeled head $g_v$, predict their categorical contents (logits), $I_{h_v}$ and $I_{g_v}$, leveraging the semantic embeddings. Finally, we concatenate both the logits from $h_v$ and $g_v$ as follows: $I_v = [h_v, g_v]$ and get the probability distribution $p_v = \sigma(I_v/\tau)$ via a softmax layer $\sigma$ with $\tau$ as the temperature parameter. Considering the unavailable text at test time, we utilize a parallel text branch adopting the same architecture. To be specific, the text semantic embedding $z_t \in \mathbb{R}^k$ is obtained from text encoder $E_t$ and projection head $Proj_t$. The text classification logits are predicted as $I_t = [h_t, g_t]$ via the text classification heads $h_t$(labeled head), $g_t$(unlabeled head), so as the probability predictions $p_t = \sigma(I_t/\tau)$.

Given our setup of CXR NCD using text data, it is essential to quantify the quality of pseudo-labels from both the visual and text modalities. As it is challenging to quantify the heterogeneous feature structures, we propose a solution to calculate the consistency index between the semantic feature structure and the pseudo-label structure. A higher consistency score may indicate better pseudo-label quality, as the score compromises the capabilities of distinguishing the unknown classes and capturing the relationship between the known and the unknown classes. The cross-modal comparison in the space of our proposed consistency alleviates the cross-modal embedding gap in the processing. More technical details are in the following section.

Once the quality of pseudo-labels from both image and text branches is quantified, we propose to generate synthetic pseudo-labels as supervision of both branches. The synthetic pseudo-labels are integrated by weighting the pseudo-labels from images and text using the calculated consistency scores. Our design is able to transfer effective information between image and text branches by scheduling based on the quantified quality of each modality. Note this procedure effectively reduce the inter-modal gap as well.

Pseudo-Labeling with Estimated Prior Distribution

Pseudo-labeling is a key step of NCD. Following prior works (Fini et al. 2021; Li et al. 2022a; Yang et al. 2022), we re-formulate the clustering problem into an optimal transport (OT) problem that finds the optimal transportation between the sample distribution and the class distribution. Formally, given logits $L_v$ of a batch of data with batch size $B$, we select the logits of all the unlabeled samples: $L_v^u = [I_v^u, ..., I_v^{B^u}]$, our goal is to assign pseudo-labels $\hat{Y}_v^u = [\hat{y}_v^u, ..., \hat{y}_v^{B^u}]$, where the rows of $L_v$ represent logits while the rows of $\hat{Y}_v$ represents the pseudo-labels for unlabeled samples, and $B^u$ is the number of unlabeled samples in the batch. The problem can be solved by the Sinkhorn-Knopp algorithm as follows:
Due to the low pseudo-label accuracy generated by image branch discussed in the introduction, we aim to improve the pseudo-label quality by leveraging the text information and interaction between the two branches. A straightforward idea is to take the average of the pseudo-labels of both branches so that all the high-quality pseudo-labels are retained to some extent. However, the unreliable pseudo-labels are equally retained and transferred. It tends to propagate and accumulate erroneous information and prevent taking advantage of high-quality pseudo-labels. Therefore, we propose a novel and robust measurement to quantify the pseudo-label quality from both modalities. Then, leveraging this quantification score, advantageous and complementary information from both modalities can be scheduled and utilized to boost the NCD performance.

Before introducing the details of our proposal, it seems essential to define the spectrum of pseudo-label quality in the NCD problem, especially without the ground-truth labels in novel classes. Different from the semi-supervised problem, in NCD, the capability to extend and explore new classes leverages not only the abstract information (e.g., logits) but also the local visual semantics that are shared among unknown and known categories (Sun et al. 2023). The mainstream theory (Li et al. 2022b) demonstrated that the NCD performance depends on how similar/shareable local semantic attributes are across known and novel classes. Therefore, there are at least two aspects to quantify the quality of the pseudo-labels: (1) how much the abstract features/logits/pseudo-labels) can distinguish the unknown classes; (2) how the semantic patterns in the local feature embeddings are able to identify and comprehend the relationship between the known and the novel classes.

To bring advantages from both worlds, i.e. the abstract and local feature embeddings, we propose to utilize their structural consistency for pseudo-label quality identification. To put it simply, we propose a new hypothesis that if samples with more similar local semantic attributes/features, their structural similarity of pseudo-labels should be maintained. Therefore, the higher the consistency between the local semantic structure and pseudo-label structure is, the better the pseudo-label quality is. And vice versa.
Specifically, let $Z_u = [z_1^u, ..., z_B^u] \in \mathbb{R}^{B \times k}$ be the embeddings of unlabeled images in the batch, the local semantic similarity can be calculated as:

$$Sim_{embu} = Z_u^T Z_u$$ (3)

where $Sim_{embu} = z_i^u \cdot z_j^u$ represents the similarity between $i$th image embedding and $j$th image embedding.

Similarly, given pseudo-labels $\hat{Y}_v = [\hat{y}_1^v, ..., \hat{y}_B^v] \in \mathbb{R}^{B \times C_v}$ of unlabeled images, we can get similarity matrix of pseudo-labels as follow:

$$Sim_{plu} = \hat{Y}_v^T \hat{Y}_v$$ (4)

where $Sim_{plu} = \hat{y}_i^v \cdot \hat{y}_j^v$ represents the similarity between $i$th pseudo-label and $j$th pseudo-label.

For the $i$th image, we propose to use JS-divergence to measure the consistency $Con_i$ of embedding similarity and pseudo-label similarity, where

$$Con_i = \max (m, 1 - \lambda D_{JS}(Sim_{plu} || Sim_{embu}))$$ (5)

and $m$ is a threshold that prevents consistency from being 0.

Following the above steps, we can calculate the embedding similarity $Sim_{embu}$ and pseudo-label similarity $Sim_{plu}$ of the text modality as well. And for the $i$th text data, we use JS-divergence to measure the consistency $Con_i$ of embedding similarity and pseudo-label similarity, where

$$Con_i = \max (m, 1 - \lambda D_{JS}(Sim_{plu} || Sim_{embu}))$$ (6)

After calculating the intra-modal consistency, we propose to generate synthetic pseudo-labels that guide the learning process for both modalities. Specifically, using the quality indexes as pseudo-label weights, the synthetic pseudo-labels of the $i$th sample pair can be expressed as:

$$pl^i = \frac{Con_v}{Con_v + Con_t} \hat{y}_v^i + \frac{Con_t}{Con_v + Con_t} \hat{y}_t^i$$ (7)

We train both branches using synthetic pseudo-labels. Following UNO(Fini et al. 2021), for samples from the labeled set, we zero-pad $\hat{y}_i$, i.e., $y = [0_{C_v}, \hat{y}_i]$; for samples from unlabeled set, we zero-pad $pl^i$, i.e., $y = [0_{C_v}, pl^i]$. Then we can train the whole network using standard cross-entropy:

$$L_{img} = -\frac{1}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} y_b^b(c) \log(p_t^b(c))$$ (8)

$$L_{text} = -\frac{1}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} y_b^t(c) \log(p_t^t(c))$$ (9)

$$L_{cls} = L_{img} + L_{text}$$ (10)

where $C = C_v + C_t$, $y_b^b(c)$ is the c-th element of the label $y_b$ of the b-th sample in a batch, $p_t^b(c)$ is the c-th element of b-th image’s prediction $p_t^b$, $p_t^t(c)$ is the c-th element of b-th text’s prediction $p_t^t$.

**Experiment**

**Experiment Setup**

**Datasets**

MIMIC-CXR-JPG Dataset (Johnson et al. 2019b)

This dataset contains 377,110 chest X-ray images from 65,379 patients. Each image is provided with 14 labels derived from two natural language processing tools applied to the corresponding free-text radiology reports. In our experiment, we only focus on investigating images from the frontal view. Based on the relationship between classes and the number of samples in classes, 11 classes were selected. We divided these classes into three groups, one group as labeled classes and the remaining two groups as unlabeled classes. The labeled classes include No Finding, Atelectasis, Lung Opacity and Edema. The first group of unlabeled classes are all lung diseases, including Consolidation, Pneumonia, Pneumothorax. The second group of unlabeled classes are diseases that occur in other anatomical structures, including Cardiomegaly, Enlarged Cardiomediatinum, Fracture, Pleural Effusion. We refer to the combination of the labeled classes and the first group of unlabeled classes as SET1 and the combination of the labeled classes and the second group of unlabeled classes as SET2. We adjust the number of samples for each class and obtain eight different dataset splits. The details are shown in the Table 1.

**Chest Ima Genome Dataset** (Wu et al. 2021)

Chest Ima Genome dataset is automatically constructed from the MIMIC-CXR dataset (Johnson et al. 2019a). This dataset uses a rule-based text-analysis pipeline to correlate anatomies with various CXR attributes extracted from text reports. To reduce noise and prevent label leakage, we filter the attributes and formalize the text report into the form of “Anatomy-1: Attribute-1, Anatomy-k: Attribute-j”, where k and j mean the j-th attribute of k-th anatomy. All our experiments are conducted based on formalized text.

**Evaluation Metrics**

Adhering to the evaluation protocols employed in existing studies (Fini et al. 2021; Li et al. 2022a), our experiments are also conducted under both task-aware and task-agnostic protocols. Under the task-aware protocol, we are aware of if the paired image and text originate from the labeled set or the unlabeled set. Conversely, under the task-agnostic protocol, such information is unavailable. We use the whole clustering accuracy to evaluate the performance of our method on unlabeled sets. It is defined as:

$$Cluster_Acc = \max_{perm_p \in P} \frac{1}{N} \sum_{i=1}^{N} 1\{y_i = perm(pl_i)\}$$ (11)

where $y_i$ and $pl_i$ represent the ground-truth label and pseudo-label of sample $(v_i, t_i)$. $P$ indicates the set of all permutations. The optimal permutation can be calculated by the Hungarian algorithm (Kuhn 2005).

**Implementation Details**

We use ResNet-50 (He et al. 2016) as the image encoder and BioClinicalBERT (Aalsentzer et al. 2019) as the text encoder. We train our model in two stages. First, we finetune the encoders following GLoRia (Huang et al. 2021) with all training data. Then we conduct novel class discovery on our network with 200 epochs. All experiments are conducted with a fixed batch size of 128.
In this setting, our method is significantly better at classifying the unlabeled test sets using the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

Comparison with State-of-the-Arts

We compare our method with the current state-of-the-art methods, including AutoNovel(Han et al. 2021), NCL(Zhong et al. 2021a), UNO(Fini et al. 2021), IIC(Li et al. 2022a) and ComEx(Yang et al. 2022) besides K-Means(McQueen 1967). We also combine our method with pseudo-label assignment based on novel class distribution prior estimation in BYOP(Yang et al. 2023) for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.
<table>
<thead>
<tr>
<th>Method</th>
<th>SET1</th>
<th>SET2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>K-means (McQueen 1967)</td>
<td>38.9</td>
<td>37.1</td>
</tr>
<tr>
<td>AutoNovel (Han et al. 2021)</td>
<td>38.7</td>
<td>36.8</td>
</tr>
<tr>
<td>NCL (Zhong et al. 2021a)</td>
<td>39.6</td>
<td>37.1</td>
</tr>
<tr>
<td>UNO (Fini et al. 2021)</td>
<td>43.6</td>
<td>37.5</td>
</tr>
<tr>
<td>UNO/MMKD</td>
<td>42.4</td>
<td>37.3</td>
</tr>
<tr>
<td>IIC (Li et al. 2022a)</td>
<td>43.8</td>
<td>40.1</td>
</tr>
<tr>
<td>ComEx (Yang et al. 2022)</td>
<td>42.7</td>
<td>41.4</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>56.9</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Ours+BYOP (Yang et al. 2023)</strong></td>
<td>56.1</td>
<td><strong>58.8</strong></td>
</tr>
</tbody>
</table>

Table 2: Comparison of state-of-the-art methods on eight splits of SET1 and SET2 using task-aware protocol. Cluster accuracy is reported on the unlabeled test set. The optimal results from the 5 runs are reported in the table. Noting that UNO/MMKD means taking the frozen text branch as **teacher** and the image branch as **student** and distilling the knowledge at the logits end.

<table>
<thead>
<tr>
<th>Method</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lab</td>
<td>Unlab</td>
<td>All</td>
<td>Lab</td>
</tr>
<tr>
<td>UNO (Fini et al. 2021)</td>
<td>38.0</td>
<td>37.5</td>
<td>37.8</td>
<td>39.1</td>
</tr>
<tr>
<td>IIC (Li et al. 2022a)</td>
<td>39.5</td>
<td>37.6</td>
<td>38.7</td>
<td>40.8</td>
</tr>
<tr>
<td>ComEx (Yang et al. 2022)</td>
<td>41.0</td>
<td><strong>38.9</strong></td>
<td>40.1</td>
<td>41.8</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>46.3</strong></td>
<td>38.7</td>
<td><strong>43.1</strong></td>
<td><strong>50.9</strong></td>
</tr>
</tbody>
</table>

Table 3: Comparison with some state-of-the-art methods on four splits of SET1 under the task-agnostic protocol. Both the classification accuracy of the labeled test sets ("Lab") and the clustering accuracy of the unlabeled test sets ("Unlab") are reported.

<table>
<thead>
<tr>
<th>Method</th>
<th>SET1_I</th>
<th>SET1_II</th>
<th>SET1_III</th>
<th>SET1_IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL avg.</td>
<td>54.8</td>
<td>53.6</td>
<td>54.8</td>
<td>51.2</td>
</tr>
<tr>
<td>Logits avg.</td>
<td>52.7</td>
<td>52.4</td>
<td>51.5</td>
<td>51.8</td>
</tr>
<tr>
<td>Weighting</td>
<td><strong>56.9</strong></td>
<td><strong>53.7</strong></td>
<td><strong>55.7</strong></td>
<td><strong>52.6</strong></td>
</tr>
</tbody>
</table>

Table 4: Ablation study performed on four splits of SET1 on synthetic pseudo-label generation. **PL avg.** means averaging pseudo-labels from two branches. **Logits avg.** means averaging unlabeled logits from two branches to generate pseudo-labels. **Weighting** means weighting by intra-modal consistency. The results are reported on the unlabeled test set using the task-aware protocol.

**Why not swapping predictions like UNO (Fini et al. 2021)?** While text can be viewed as a form of augmentation, we believe it is fundamentally different from augmented images. There is an information difference between image modality and text modality, so we want to achieve both intra-modal and inter-modal optimization when preserving the supervised information from both modalities.

**Conclusion**

In this paper, we propose a method for NCD in CXR images based on paired images and text. During pretraining, we perform multi-modal contrastive learning on the training set to mitigate the bias towards labeled classes. In the discovery phase, we generate pseudo-labels on both image branch and text branch, respectively, and weight the pseudo-labels by intra-modal consistency. In this way, pseudo-labels that combine information from both modalities are used for training on both branches. Through extensive experiments and analysis, we illustrate the effectiveness of our approach. Our method achieves the best performance on the novel class discovery on CXR images, which is simple but valid.
Acknowledgements
This work was supported by the grants from the National Natural Science Foundation of China (62201014), Beijing Advanced Discipline Construction Project (BMU2019GJXXK001) and PKU-OPPO Innovation Fund (BO202103).

References


