GII: A Unified Approach to Representation Learning in Open Set Recognition with Novel Category Discovery

Jingyun Jia[1][0000−0003−0865−049X] and Philip K. Chan[2][0000−0002−3878−4205]

[1] Baidu Research, Sunnyvale, CA 94089 USA
jingyunjia@baidu.com
[2] Florida Institute of Technology, Melbourne FL 32901, USA
pkc@fit.edu

Abstract. In this paper, we consider the problem of Novel Class Discovery (NCD) in Open Set Recognition (OSR). Given a labeled and an unlabeled set for training, NCD aims to discover the novel categories in the unlabeled set with prior knowledge learned from the labeled set. Existing approaches tackle the NCD problems under a close-set setting, where only the existing categories from the labeled set and the novel categories from the unlabeled set will occur during the inference. This paper considers a more realistic open-set scenario. In the open-set setting, in addition to the existing and novel categories, some unknown categories absent from the training could be present during inference. To address NCD in the open-set scenario, we propose the General Inter-Intra (GII) loss, a unified approach for learning representations from both labeled and unlabeled samples. The proposed approach discovers novel categories in the training set (NCD) meanwhile recognizes the unknown categories (OSR). We evaluate GII with image and graph datasets, and the results indicate that our proposed approach is more effective than other NCD and OSR approaches.

Keywords: Novel Category Discovery · Open Set Recognition · Representation Learning.

1 Introduction

Machine learning models have achieved significant advances in various tasks in recent years. Most of these models are developed under a closed-world assumption and rely on a huge amount of data with human annotations. The real world is an open set, and humans can determine whether images belong to the same category. However, such an open-set setting brings new challenges for machine learning models. First, it is cost-inhibitive to keep manually annotating the emerging new categories. Second, it is unlikely to collect samples exhausting all the classes. In the open-set setting, an ideal machine learning model should automatically discover new categories in the training set without having access

to their labels, called novel category discovery (NCD) [Han et al.(2019)]. Meanwhile, the model should recognize the unknown classes absent from the training set, which is referred as Open Set Recognition (OSR) [Bendale and Boult(2016)].

In this paper, we focus on automatically discovering novel categories in a realistic open-set scenario. In the open-set setting, we have labeled and unlabeled samples available for training. Meanwhile, we have unknown samples that are not available in the training process. Our proposed approach has three objectives: classifying the existing categories from the labeled samples, clustering the novel categories from the unlabeled samples, and recognizing the unknown classes absent from the training set. Specifically, we introduce a one-step solution for NCD under the open-set scenario and name this solution general inter-intra (GII) loss. [Hassen and Chan(2020a)] propose inter-intra (ii) loss for OSR with labeled training samples. Ii loss maximizes the inter-class distances and minimizes the intra-class distances in the representation space to achieve inter-class separation and intra-class compactness. We generalize this idea to unlabeled samples in our work. GII consists of three components: intra-class loss for existing categories, intra-cluster loss for novel categories, and inter-category loss for all categories. We calculate their class centroids in representation space for existing categories and minimize the intra-class distance. For novel categories, we first estimate the centroids of the novel categories and cluster assignments via k-means, then we minimize the intra-cluster distance in the representation space. The assumption is that novel categories are disjointed with existing ones, so intra-category loss is designed to maximize the distance between any two categories.

Our contribution includes: first, we propose a unified approach for learning representations from both labeled and unlabeled samples for NCD under an open-set scenario. Second, to the best of our knowledge, we are the first to extend NCD to an open-set setting. Third, we experiment with the proposed approach with image and graph datasets, and the results indicate that our proposed approach is more effective than other approaches for NCD and OSR.

2 Related Work

An **Open Set Recognition (OSR)** task has two objectives: classify the known classes and recognize the unknown class absent from training. We can divide OSR techniques into three categories based on the training set compositions. The first category includes the techniques that borrow additional data in the training set. Dhamija et al. [Dhamija et al.(2018)] utilize the differences in feature magnitudes between known and borrowed unknown samples as part of the objective function. Shu et al. [Shu et al.(2018)] indicate that several manual annotations for unknown classes are required in their workflow. The second category of OSR approaches includes the research works that generate additional data in training data. Most data generation methods are based on GANs. Ge et al. [Ge et al.(2017)] introduce a conditional GAN to generate some unknown samples followed by OpenMax open set classifier. Neal et al. [Neal et al.(2018)] add another encoder network to traditional GANs to map from images to a latent

space. The third category of OSR approaches does not require additional data. Instead, it requires outlier detection for the unknown class. Hassen and Chan [Hassen and Chan(2020b)] propose ii loss for open set recognition. It first finds the representations for the known classes during training and then recognizes an instance as unknown if it does not belong to any known classes. Jia and Chan [Jia and Chan(2021)] propose MMF as a loss extension to further separate the known and unknown representations for OSR.

One group of existing approaches solves the **Novel Category Discovery (NCD)** problem by pairing samples and converting the NCD problem to pairwise similarity prediction problem. [Gupta et al.(2020)] utilize the Information Maximization (IM) loss in an ensemble of models to predict the similarity between two data points. [Chang et al.(2017)] propose DAC architecture, which uses the learned label features for clustering tasks. The sample pairs used for training are alternately selected and labeled by the learned features in each iteration. Another group of existing approaches solves the NCD problem using prior knowledge learned from labeled samples. For example, [Han et al.(2019)] use such prior knowledge to reduce the ambiguity of clustering by reducing its KL divergence to a sharper target distribution. [Zhao and Han(2021)] propose to apply dual ranking statistics to transfer the knowledge learned from labeled samples to unlabelled samples for pseudo-labeling. [Liu and Tuytelaars(2022)] propose ResTune to estimate a new residual feature from the pre-trained network and add it with a previous basic feature to compute the clustering objective. [Zhong et al.(2021)] introduce OpenMix to mix the unlabeled examples from an open set and the labeled examples from known classes. They follow a two-stage learning stage for the NCD problem. The model initialization stage is trained on the labeled samples in a supervised way. In the unsupervised clustering stage, they generate mixed training samples by incorporating labeled samples with unlabeled samples. The pseudo-labels of mixed samples will be more reliable than their unlabeled counterparts. In addition to pseudo-pair learning and pseudo-label learning, the loss of OpenMix is applied to the mixed samples.

3 Approach

3.1 Learning Representations of Existing and Novel Categories

Consider we have a labeled collection of instances $D^l = \{(x_i^l, y_i^l)\}_{i=1}^{N^l}$, where $y_i^l \in \{1, \ldots, C^l\}$ is the ground-truth class labels for the labeled samples, and N^l is the number of labeled samples. In addition, we have an unlabelled collection of instances $D^u = \{x_i^u\}_{i=1}^{N^u}$, where N^u is the number of unlabelled samples. Following a common assumption in other works [Han et al.(2019)], we assume that the novel categories are disjoint with the existing ones, i.e., $D^l \cap D^u = \emptyset$, also the number of novel categories C^u is known.

Our goal is to model a representation space that separates the existing categories in D^l and the novel categories in D^u. Through such representation space, we can identify if a test instance belongs to one of the existing categories, one of

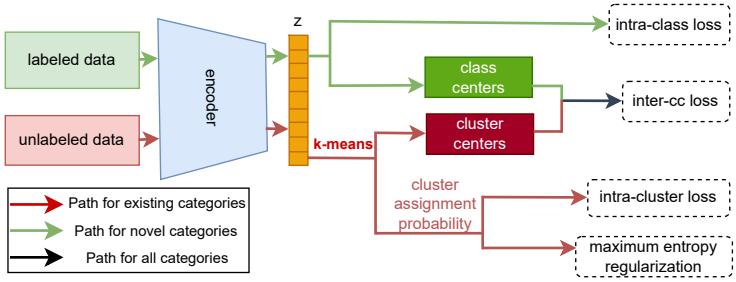


Fig. 1: Illustration of GII architecture for NCD.

the novel categories, or the unknown class. We propose an end-to-end framework to learn the representations, which provides a one-step solution for NCD under the open-set scenario. The training of the framework consists of three components: intra-class loss for the existing categories, intra-cluster loss for the novel categories, and inter-category loss for all categories. The existing categories are the classes of the labeled samples. The novel categories are the clusters of the unlabeled samples and all categories include these classes and clusters.

Intra-class loss for existing categories The intra-class component deals with the intra-spread for the labeled samples. One can expect the network to capture some informative knowledge for the existing categories through the training process, which not only helps classify labeled samples but also is beneficial to transfer the basic feature for clustering unlabeled samples. Given a labeled sample x_i^l, we use a network-based trainable encoder $f(\cdot)$ to extract its representation vector z_i^l. Thus, for existing category (or class) j, we find its centroid in the representation space as:

$$\mu_j^l = \frac{1}{N_j^l} \sum_{i=1}^{N_j^l} z_i^l, \tag{1}$$

where N_j^l denotes the number of samples in the existing category j. Then, we measure the intra-class spread as the average distance of labeled instances from their class means:

$$\text{intra-class}_j = \frac{1}{N_j^l} \sum_{i=1}^{N_j^l} \|\mu_j^l - z_i^l\|_2^2. \tag{2}$$

To improve the intra-class compactness, we minimize the largest intra-class spread among the existing categories.

$$\mathcal{L}_{\text{intra-class}} = \max_{1 \le j \le C^l} \text{intra-class}_j \tag{3}$$

Intra-cluster loss for novel categories There are several differences comparing intra-cluster spread with intra-class spread. First, intra-class spread relies on labels to find class centroids. In the intra-cluster spread, we only have unlabeled samples. Thus, we use k-means to estimate the representation of cluster centroids as the centers of novel categories $\tilde{\mu}^u$. Second, we are uncertain which specific centroid is for an unlabeled sample. Thus, we calculate the soft assignment of sample x_i^u based on the distance of its representation z_i^u to the estimated centroids. Since unlabeled samples do not belong to known classes, these samples do not have a soft assignment to known classes. To calculate the soft assignment (probability), we use the softmax of the negative distance of z_i^u from all the estimated centroids. Hence, the probability of sample x_i^u belongs to novel category (or cluster) k is given by:

$$p_{ik} = P(y_i^u = k|x_i^u) = \frac{e^{-\|\tilde{\mu}_k^u - z_i^u\|_2^2}}{\sum_{t=1}^{C^u} e^{-\|\tilde{\mu}_t^u - z_i^u\|_2^2}}, \tag{4}$$

where $\tilde{\mu}_k^u$ is the estimated centroid for novel category k. Similar to the intra-class spread, we measure the intra-cluster spread as the weighted average distance of unlabeled instances from their soft assignments. Suppose we have N_u unlabeled samples, the intra-cluster spread of novel category k is calculated as:

$$\text{intra-cluster}_k = \frac{\sum_{i=1}^{N^u} p_{ik}\|\tilde{\mu}_k^u - z_i^u\|_2^2}{\sum_{i=1}^{N^u} p_{ik}}. \tag{5}$$

Then, we minimize the largest intra-cluster spread among the novel categories to achieve intra-cluster compactness. The differences between the intra-cluster spread in Equation 5 with the intra-class spread in Equation 2 are the estimated cluster centroid $\tilde{\mu}_k^u$ and the soft assignment p_{ik}.

$$\mathcal{L}_{\text{intra-cluster}} = \max_{1 \le k \le C^u} \text{intra-cluster}_k \tag{6}$$

The cluster centroids are initialized and updated by k-means. To reduce the training time, we use a scheduling function for the k-means. Intuitively, we want to update the centroids more frequently at the beginning of the training. Close to the end of the training, when the network has learned informative knowledge from the labeled samples, and the clusters of the unlabeled samples have been formed for the novel categories, we perform k-means less frequently for the centroids updates.

Finally, to avoid a trivial solution of assigning all unlabeled samples to the same class, we regularize the model with maximum entropy regularization (MER). Specifically, we use the probability p_{ik} calculated from Equation 4 as the probability of an unlabeled sample x_i^u being assigned to novel category k. MER maximizes the entropy of the output probability distribution:

$$\mathcal{R} = -H(p) = \frac{1}{N^u} \sum_{i=1}^{N^u} \sum_{k=1}^{C^u} p_{ik} \log p_{ik}. \tag{7}$$

Inter-category loss for all categories The above two components shorten the distance between representations of the same categories to ensure intra-class and intra-cluster compactness. To distribute the representations of different categories to different subspaces, we further measure the inter-category separation as the distance between the two closest category centroids. Let μ_c be the centroid of category c, where $c \in \{1, ..., C^l\} \bigcup \{1, ..., C^u\}$. The inter-category separation for category m is defined as:

$$\text{inter-category}_m = \min_{1 \leq i \leq (C^l + C^u), k \neq i} \|\mu_m - \mu_i\|_2^2. \tag{8}$$

To improve the intra-category separability, we maximize the inter-category separation in the inter-category loss:

$$\mathcal{L}_{\text{inter-category}} = - \min_{1 \leq m \leq (C^l + C^u)} \text{inter-category}_m. \tag{9}$$

GII loss function The objective function in GII combines three components, and the overall training loss of our unified framework can then be written as:

$$\mathcal{L} = \mathcal{L}_{\text{intra-class}} + \lambda_1 \mathcal{L}_{\text{intra-cluster}} + \lambda_2 \mathcal{L}_{\text{inter-category}} + \lambda_3 \mathcal{R}, \tag{10}$$

where λ_1, λ_2, and λ_3 are regularization parameters set to 1 in all our experiments.

The representation z is learned by three components together. Specifically, $\mathcal{L}_{\text{intra-cluster}}$ is applied to unlabelled data but indirectly uses information from labeled data via z and $\mathcal{L}_{\text{intra-class}}$. The features learned from the labeled data help cluster the unlabeled data. Meanwhile, $\mathcal{L}_{\text{intra-cluster}}$ further reduces intra-cluster spread, which also influences representation z. The influence on representation z from unlabeled samples can benefit not only the representation of unlabeled samples but also the representation of labeled samples. More details are in the analysis in Section 4.4. In addition, since $L_{\text{inter-category}}$ increases separation among classes (existing categories) and clusters (novel categories), it uses information from the labeled data to help separate classes from clusters. GII is a unified approach for learning representations from both labeled and unlabeled samples.

4 Experimental Evaluation

In this section, our proposed GII is evaluated on image and graph datasets.
MNIST [Ronen et al.(2018)] contains 70,000 handwritten digits from 0 to 9. Each example in the MNIST dataset is a 28x28 grayscale image.
Fashion-MNIST [Ronen et al.(2018)] is associated with 10 classes of clothing images. It contains 60,000 training and 10,000 testing examples. In the Fashion-MNIST dataset, each example is a 28x28 grayscale image.
Microsoft Challenge (MS) [Ronen et al.(2018)] contains disassembled malware samples from 9 families. We use 10260 samples that can be correctly parsed and then extracted their FCGs for the experiment as in [Hassen and Chan(2017)].

Android Genome (AG) consists of 1,113 benign android apps and 1,200 malicious android apps. Our colleague provides the benign samples, and the malicious samples are from [Zhou and Jiang(2015)]. We select nine families with a relatively larger size for the experiment to be fairly split into the training set and the test set. The nine families contain 986 samples in total. We first use [Gascon et al.(2013)] to extract the function instructions and then generate the FCGs as in [Hassen and Chan(2017)].

4.1 Implementation details

To simulate an open-set scenario, we randomly select six classes from the datasets as existing categories. Moreover, we randomly select another two classes from the datasets as novel categories by removing their labels. These eight classes participate in the training, while the rest are considered unknowns that only exist in the test set.

As shown in Figure 1, labeled and unlabeled data share the same encoder. For the MNIST and Fashion-MNIST datasets, the padded input layer of the encoder is of size (32, 32), followed by two non-linear convolutional layers with 32 and 64 nodes. We also use the max-polling layers with kernel size (3, 3) and strides (2, 2) after each convolutional layer. We use two fully connected non-linear layers with 256 and 128 hidden units after the convolutional component. Then we have an eight-dimensional representation layer after the encoder. We use the Relu activation function for all the non-linear layers and set the Dropout rate as 0.2 for the fully connected layers. We use Adam optimizer with a learning rate of 0.001. We use a contamination ratio of 0.001 for the unknown class threshold selection. We sort the output probability of training data in ascending order and pick the 0.1 percentile of the probability as the threshold. For the FCG datasets (MS and Android), the padded input layer is in the size of (67, 67). The padded input layer is then flowed by two non-linear convolutional layers with 32 and 64 nodes. We apply the max-polling layers with kernel size (3, 3) and strides (2, 2). We also add batch normalization after each convolutional layer to complete the convolutional block. After the convolutional block, we only use one fully connected non-linear layer with 256 hidden units for the graph dataset. Next, we add an eight-dimensional representation layer after the encoder. We use the Relu activation function and set the Dropout rate as 0.2. We use Adam as the optimizer with a learning rate of 0.001. Finally, we use a contamination ratio of 0.01 for the unknown class threshold selection. Moreover, as mentioned in section 3.1, we use a scheduling function for the k-means updates in the NCD process. In the experiments, we apply k-means every ten iterations in the first 5000 iterations, then reduce the frequency to every 100 iterations in the rest of the training process.

4.2 Comparison methods

We compare the proposed with ii loss without sharpening on the unlabeled samples (No sharpening), cluster loss, and supervised OSR. For a fair comparison

with "No sharpening", we first pre-train the encoder with labeled samples using ii loss [Hassen and Chan(2020b)]. After obtaining the representations of the unlabeled samples, we find the novel cluster centroids and assignments via k-means directly in the representation space without further sharpening. Finally, we apply the same OSR process. Cluster loss is proposed to sharpen the distribution of unlabeled samples through the clustering process [Liu and Tuytelaars(2022)]. We compare our proposed intra-cluster loss with cluster loss by substituting the inter-cluster loss term with cluster loss in our overall loss function in Equation 10. Moreover, as the cluster loss measures the KL-divergence between two distributions, which is on a different scale with other terms (intra-class and inter-category), we set λ_1 differently for different datasets. That is, all three terms in our GII are based on distances in the same representation space Z. Hence, GII provides a unified approach to representation learning for both labeled and unlabeled samples.

In addition, we experiment on fully supervised OSR and use the results as the upper bounds of NCD and OSR performances. In the supervised OSR experiments, we apply ii loss on eight labeled categories in the training process. The remaining categories are considered as the unknown class.

4.3 Evaluation Criteria

As mentioned above, we simulate an open-set scenario for all the datasets. Moreover, we randomly select two classes in the training set as novel categories and remove their class labels. We simulate three open-set groups for each dataset and then repeat each group 10 runs, so each dataset has results for 30 runs. We calculate the average results of the 30 runs for performance evaluation.

We calculate the accuracy (ACC) scores under different types of categories: existing categories ($\mathrm{ACC_E}$), novel categories ($\mathrm{ACC_N}$) and the unknown category ($\mathrm{ACC_U}$). Specifically, we evaluate the classification accuracy of existing categories and the recognition accuracy of the unknown category. Moreover, we evaluate the model performance on novel categories with clustering accuracy. Clustering accuracy is widely used in NCD problems. To find the optimal match between the class labels and the cluster labels, the ACC of novel categories is defined as $\mathrm{ACC_N} = \max_{\mathrm{perm}\in P} \frac{1}{N} \sum_{i=1}^{N} \delta(\mathrm{perm}(\hat{y}_i) = y_i)$, where N is the total number of unlabeled samples; δ is the Kronecker delta response; $\hat{y}_i$ denotes the predicted cluster label; $\mathrm{perm}(\cdot)$ is the permutation operation and P is the set of all permutations of the class assignments in the test set. The score ranges between 0 and 1, and a higher value means better performance. The Hungarian algorithm is used to optimize the permutations for faster computation.

To further evaluate our approach on OSR, we measure the AUC scores under 100% and 10% False Positive Rate (FPR). While the AUC score under 100% FPR is commonly used in model performance measurements, the AUC score under 10% FPR is more meaningful for malware detection applications.

Table 1: The average ACC scores of 30 runs. The upper bounds results are trained with fully supervised learning, and the values in boldface are the highest in each column.

Image Dataset	MNIST					Fashion-MNIST				
	ACC_E	ACC_N	ACC_{E+N}	ACC_U	ACC_{E+N+U}	ACC_E	ACC_N	ACC_{E+N}	ACC_U	ACC_{E+N+U}
No sharpening	$0.733_{\pm 0.078}$	$0.800_{\pm 0.091}$	$0.697_{\pm 0.078}$	$0.767_{\pm 0.015}$	$0.615_{\pm 0.060}$	$0.598_{\pm 0.068}$	$0.668_{\pm 0.089}$	$0.539_{\pm 0.098}$	$0.786_{\pm 0.008}$	$0.468_{\pm 0.079}$
Cluster loss	$0.752_{\pm 0.161}$	$0.625_{\pm 0.125}$	$0.687_{\pm 0.166}$	$0.751_{\pm 0.031}$	$0.624_{\pm 0.127}$	$0.820_{\pm 0.062}$	$0.608_{\pm 0.104}$	$0.757_{\pm 0.052}$	$0.698_{\pm 0.049}$	$0.628_{\pm 0.042}$
GII (ours)	$\mathbf{0.936}_{\pm 0.08}$	$\mathbf{0.854}_{\pm 0.088}$	$\mathbf{0.909}_{\pm 0.089}$	$\mathbf{0.817}_{\pm 0.070}$	$\mathbf{0.810}_{\pm 0.069}$	$\mathbf{0.875}_{\pm 0.047}$	$\mathbf{0.808}_{\pm 0.084}$	$\mathbf{0.847}_{\pm 0.051}$	$\mathbf{0.797}_{\pm 0.003}$	$\mathbf{0.687}_{\pm 0.034}$
Upper bound (supervised)	$0.983_{\pm 0.001}$	$0.977_{\pm 0.004}$	$0.981_{\pm 0.001}$	$0.937_{\pm 0.012}$	$0.935_{\pm 0.012}$	$0.896_{\pm 0.018}$	$0.967_{\pm 0.005}$	$0.914_{\pm 0.014}$	$0.822_{\pm 0.011}$	$0.770_{\pm 0.016}$
Malware Dataset	MS					AG				
	ACC_E	ACC_N	ACC_{E+N}	ACC_U	ACC_{E+N+U}	ACC_E	ACC_N	ACC_{E+N}	ACC_U	ACC_{E+N+U}
No sharpening	$0.732_{\pm 0.131}$	$0.625_{\pm 0.180}$	$0.717_{\pm 0.132}$	$0.763_{\pm 0.112}$	$0.653_{\pm 0.166}$	$0.680_{\pm 0.167}$	$0.708_{\pm 0.140}$	$0.602_{\pm 0.176}$	$0.798_{\pm 0.027}$	$0.564_{\pm 0.193}$
Cluster loss	$0.880_{\pm 0.117}$	$0.602_{\pm 0.183}$	$0.818_{\pm 0.106}$	$0.758_{\pm 0.096}$	$0.742_{\pm 0.094}$	$0.779_{\pm 0.146}$	$0.601_{\pm 0.177}$	$0.734_{\pm 0.120}$	$0.773_{\pm 0.063}$	$0.684_{\pm 0.118}$
GII (ours)	$\mathbf{0.942}_{\pm 0.026}$	$\mathbf{0.630}_{\pm 0.143}$	$\mathbf{0.895}_{\pm 0.054}$	$\mathbf{0.834}_{\pm 0.071}$	$\mathbf{0.811}_{\pm 0.078}$	$\mathbf{0.944}_{\pm 0.013}$	$\mathbf{0.714}_{\pm 0.080}$	$\mathbf{0.906}_{\pm 0.020}$	$\mathbf{0.831}_{\pm 0.048}$	$\mathbf{0.820}_{\pm 0.034}$
Upper bound (supervised)	$0.960_{\pm 0.016}$	$0.916_{\pm 0.035}$	$0.950_{\pm 0.020}$	$0.903_{\pm 0.035}$	$0.899_{\pm 0.035}$	$0.922_{\pm 0.012}$	$0.712_{\pm 0.080}$	$0.898_{\pm 0.021}$	$0.908_{\pm 0.013}$	$0.904_{\pm 0.012}$

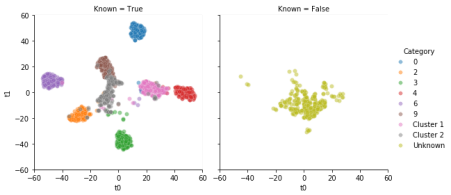
Table 2: The average ROC AUC scores of 30 runs at 100% and 10% FPR. The upper bounds results are trained with fully supervised learning, and the values in boldface are the highest in each column.

	MNIST		Fashion-MNIST		MS		AG	
FPR	100%	10%	100%	10%	100%	10%	100%	10%
No sharpening	$0.439_{\pm 0.127}$	$0.004_{\pm 0.003}$	$0.418_{\pm 0.073}$	$0.003_{\pm 0.001}$	$0.528_{\pm 0.122}$	$0.007_{\pm 0.004}$	$0.293_{\pm 0.214}$	$0.000_{\pm 0.000}$
Cluster loss	$0.413_{\pm 0.231}$	$0.007_{\pm 0.009}$	$0.620_{\pm 0.084}$	$0.008_{\pm 0.003}$	$0.651_{\pm 0.271}$	$0.018_{\pm 0.015}$	$0.507_{\pm 0.283}$	$0.007_{\pm 0.015}$
GII (ours)	$\mathbf{0.829}_{\pm 0.104}$	$\mathbf{0.047}_{\pm 0.016}$	$\mathbf{0.674}_{\pm 0.040}$	$\mathbf{0.012}_{\pm 0.004}$	$\mathbf{0.858}_{\pm 0.086}$	$\mathbf{0.028}_{\pm 0.015}$	$\mathbf{0.885}_{\pm 0.090}$	$\mathbf{0.016}_{\pm 0.020}$
Upper bound (supervised)	$0.966_{\pm 0.010}$	$0.078_{\pm 0.003}$	$0.676_{\pm 0.062}$	$0.015_{\pm 0.002}$	$0.945_{\pm 0.045}$	$0.062_{\pm 0.017}$	$0.963_{\pm 0.013}$	$0.052_{\pm 0.015}$

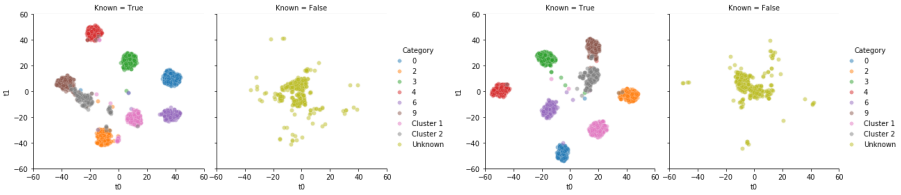
4.4 Experimental Results and Analysis

We test our proposed method on image and malware datasets for 30 runs. Table 1 shows the average accuracy scores of different methods. Notably, we measure the average clustering/classification accuracy on the existing/novel set and the combined set (ACC_{E+N}). Moreover, considering an open-set scenario, we measure the average accuracy of the unknown set, and the set contains all the existing, novel, and unknown categories (ACC_{E+N+U}). Comparing the ACC under existing categories (ACC_E) and novel categories (ACC_N), we observe that our proposed GII outperforms both ii loss without sharpening and cluster loss in NCD. Also, comparing the ACC under the unknown category (ACC_U), we observe that GII achieves the best performance in OSR. The upper bound performances are generated from supervised ii loss, where we utilize the labels of novel categories in the training set. We can see that GII has comparable performances with the supervised training in some datasets. In particular, GII obtains higher accuracy than supervised learning in the combined novel and existing categories (ACC_{E+N}) in the AG dataset.

In addition to the ACC scores, we measure the AUC ROC scores under different FPR values: 100% and 10% in Table 2. The AUC ROC measures OSR at various threshold settings. Similar to the ACC scores, our proposed GII outperforms ii loss without sharpening and cluster loss in the AUC ROC scores. Furthermore, comparing GII with supervised learning, we observe that GII can achieve comparable OSR performance in the Fashion-MNIST dataset.



(a) No sharpening

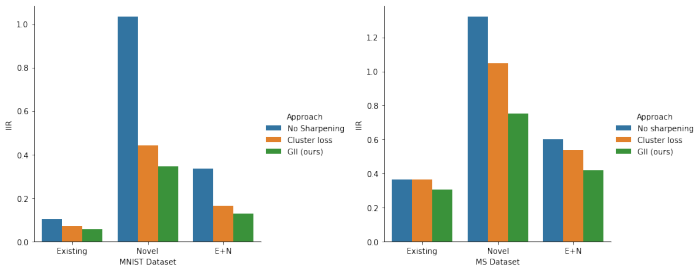


(b) Cluster loss (c) GII (ours)

Fig. 2: The t-SNE plots of the representations of MNIST test samples.

Our experiment results indicate that GII outperforms ii loss without sharpening and cluster loss in terms of performances in NCD and OSR. Specifically, ii loss without sharpening can be considered as an ablation study to investigate our approach without intra-cluster loss. We plot the t-SNE plots of the representations of samples from different categories in the MNIST test set, as shown in Figure 2. The left subplots are the representations of the samples from existing categories ("0", "2". "3", "4", "6" and "9") and novel categories ("cluster 1" and "cluster 2"). The right subplots show the representations of samples from unknown categories, which only exist in the test set. Comparing Figure 2a with Figures 2b and 2c, we can see that samples from the two clusters result in more compact intra-cluster spread with cluster loss and GII. The reason is that cluster and GII sharpen the distributions of the unlabeled samples while "No sharpening" does not change the distributions of the unlabeled samples. Furthermore, it can be seen that GII forms better clusters compared with cluster loss. GII generates a more discriminative boundary for the samples in cluster 2 (grey) and the samples in class "9" (brown). The reason is that GII forms a tighter cluster for cluster 2. Thus a more accurate cluster centroid is estimated and used in the inter-category loss. Also, comparing the representations in the right subplots, we find that the representations of unknown samples learned by ii loss without sharpening and GII are more concentrated around the origin. In contrast, those learned by GII are more widespread.

Besides visually evaluating representations via t-SNE plots, we also evaluate intra-inter ratio (IIR) [Jia and Chan(2022)] with test samples to measure the representation quality learned by different approaches. IIR measures the representation quality by calculating the ratio between intra-category spread and inter-category separation, and a lower value means better representations. Fig-



(a) MNIST (b) MS

Fig. 3: Intra-inter ratio (IIR) of the representations in different categories

ure 3 shows the IIR values of different datasets. From Figure 3, within the novel categories across four datasets, cluster loss or GII has (large) improvements in IIR over no sharpening, which indicates the benefit of representation learning with unlabeled samples via cluster loss or GII. However, GII yields a larger benefit than cluster loss. More interestingly, within the existing categories across datasets, we observe improvements in IIR with GII over no sharpening. That is, the unlabeled samples via GII help improve the representations of samples from labeled classes. Hence, not only the representations of unlabeled samples benefit from representation learning from unlabeled samples via GII, the representations of labeled samples also benefit.

5 Conclusion

We have presented a generic one-step representation learning approach to tackle the challenging problem of novel category discovery under an open-set scenario. Our proposed approach consists of three components. First, we achieve intra-class spread for labeled samples by minimizing the intra-class distance. Second, we estimate the novel category centroids and propose intra-cluster loss for the unlabeled samples to discover novel categories. Third, we separate different categories by maximizing the intra-category distance such that all the categories inhabit the same representation space. Last, we evaluated our approach on image and graph datasets, and the results indicate that the proposed approach obtained superior results in NCD and OSR compared with other approaches.

References

Bendale and Boult(2016). Abhijit Bendale and Terrance E Boult. 2016. Towards open set deep networks. In *Proc. of the IEEE conf. on computer vision and pattern recognition*. 1563–1572.

Chang et al.(2017). Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2017. Deep Adaptive Image Clustering. In *IEEE Intl. Conf. on Computer Vision, ICCV Italy*. IEEE Computer Society, 5880–5888.

Dhamija et al.(2018). Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult. 2018. Reducing Network Agnostophobia. In *Advances in Neural Information Processing Systems 31*. 9175–9186.

Gascon et al.(2013). Hugo Gascon, Fabian Yamaguchi, Daniel Arp, and Konrad Rieck. 2013. Structural detection of android malware using embedded call graphs. In *Proc. of the 2013 ACM Workshop on Artificial Intelligence and Security (AISec)*. 45–54.

Ge et al.(2017). Zongyuan Ge, Sergey Demyanov, and Rahil Garnavi. 2017. Generative OpenMax for Multi-Class Open Set Classification. In *British Machine Vision Conference*.

Gupta et al.(2020). Divam Gupta, Ramachandran Ramjee, Nipun Kwatra, and Muthian Sivathanu. 2020. Unsupervised Clustering using Pseudo-semi-supervised Learning. In *8th Intl. Conf. on Learning Representations, ICLR, Ethiopia*.

Han et al.(2019). Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to Discover Novel Visual Categories via Deep Transfer Clustering. In *IEEE/CVF Intl. Conf. on Computer Vision, ICCV, Korea (South)*. IEEE, 8400–8408.

Hassen and Chan(2017). Mehadi Hassen and Philip K. Chan. 2017. Scalable Function Call Graph-based Malware Classification. In *Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy,*. 239–248.

Hassen and Chan(2020a). Mehadi Hassen and Philip K. Chan. 2020a. Learning a Neural-network-based Representation for Open Set Recognition. In *Proc. of the SIAM Intl. Conf. on Data Mining, SDM USA*. SIAM, 154–162.

Hassen and Chan(2020b). Mehadi Hassen and Philip K. Chan. 2020b. Learning a Neural-network-based Representation for Open Set Recognition. *Proc. SIAM Intl. Conf. Data Mining* (2020), 154–162. arXiv:1802.04365

Jia and Chan(2021). Jingyun Jia and Philip K. Chan. 2021. MMF: A Loss Extension for Feature Learning in Open Set Recognition. In *Intl. Conf. on Artificial Neural Networks, Proc. Part II*. 319–331.

Jia and Chan(2022). Jingyun Jia and Philip K. Chan. 2022. Feature Decoupling in Self-supervised Representation Learning for Open Set Recognition. *CoRR* abs/2209.14385 (2022). arXiv:2209.14385

Liu and Tuytelaars(2022). Yu Liu and Tinne Tuytelaars. 2022. Residual Tuning: Toward Novel Category Discovery Without Labels. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–15.

Neal et al.(2018). Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. 2018. Open set learning with counterfactual images. In *Proc. of the European Conf. on Computer Vision (ECCV)*. 613–628.

Ronen et al.(2018). Royi Ronen, Marian Radu, Corina Feuerstein, Elad Yom-Tov, and Mansour Ahmadi. 2018. Microsoft Malware Classification Challenge. *CoRR* abs/1802.10135 (2018). arXiv:1802.10135

Shu et al.(2018). Yu Shu, Yemin Shi, Yaowei Wang, Yixiong Zou, Qingsheng Yuan, and Yonghong Tian. 2018. ODN: Opening the Deep Network for Open-Set Action Recognition. In *2018 Intl. Conf. on Multimedia and Expo (ICME)*. IEEE, 1–6.

Zhao and Han(2021). Bingchen Zhao and Kai Han. 2021. Novel Visual Category Discovery with Dual Ranking Statistics and Mutual Knowledge Distillation. In *Annual Conf. on Neural Information Processing Systems, NeurIPS, virtual*. 22982–22994.

Zhong et al.(2021). Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. 2021. OpenMix: Reviving Known Knowledge for Discovering Novel Visual Categories in an Open World. In *Conf. on Computer Vision and Pattern Recognition, CVPR, virtual*. Computer Vision Foundation / IEEE, 9462–9470.

Zhou and Jiang(2015). Yajin Zhou and Xuxian Jiang. 2015. Android Malware Genome Project. `http://www.malgenomeproject.org/`