Generalized Category Discovery with Decoupled Prototypical Network

Wenbin An[1,3], **Feng Tian**[2,3], **Qinghua Zheng**[2,3], **Wei Ding**[4], **QianYing Wang**[5], **Ping Chen**[6]

[1]School of Automation Science and Engineering, Xi'an Jiaotong University
[2]School of Computer Science and Technology, Xi'an Jiaotong University
[3]National Engineering Laboratory for Big Data Analytics
[4]Department of Computer Science, University of Massachusetts Boston [5]Lenovo Research
[6]Department of Engineering, University of Massachusetts Boston
wenbinan@stu.xjtu.edu.cn, {fengtian,qhzheng}@mail.xjtu.edu.cn
{wei.ding,Ping.Chen}@umb.edu, wangqya@lenovo.com

Abstract

Generalized Category Discovery (GCD) aims to recognize both known and novel categories from a set of unlabeled data, based on another dataset labeled with only known categories. Without considering differences between known and novel categories, current methods learn about them in a coupled manner, which can hurt model's generalization and discriminative ability. Furthermore, the coupled training approach prevents these models transferring category-specific knowledge explicitly from labeled data to unlabeled data, which can lose high-level semantic information and impair model performance. To mitigate above limitations, we present a novel model called *Decoupled Prototypical Network* (DPN). By formulating a bipartite matching problem for category prototypes, DPN can not only decouple known and novel categories to achieve different training targets effectively, but also align known categories in labeled and unlabeled data to transfer category-specific knowledge explicitly and capture high-level semantics. Furthermore, DPN can learn more discriminative features for both known and novel categories through our proposed *Semantic-aware Prototypical Learning* (SPL). Besides capturing meaningful semantic information, SPL can also alleviate the noise of hard pseudo labels through semantic-weighted soft assignment. Extensive experiments show that DPN outperforms state-of-the-art models by a large margin on all evaluation metrics across multiple benchmark datasets. Code and data are available at https://github.com/Lackel/DPN.

Introduction

Although modern machine learning methods have achieved superior performance on many NLP tasks such as text classification, they usually fail to recognize novel categories from newly collected unlabelled data. To cope with this limitation, some task settings were proposed to discover novel categories from unlabeled text (Lin, Xu, and Zhang 2020) or images (Yu et al. 2022). Recently, Vaze et al. (2022) comprehensively formalized these settings and proposed a task called Generalized Category Discovery (GCD). Since models trained on a labeled dataset containing only known categories may encounter newly collected unlabelled data with both known and novel categories, GCD requires models to recognize both known and novel categories from the unlabeled data without

any additional annotation, which can help to extend existing category taxonomy and reduce significant labeling cost.

To discover novel categories from unlabeled data, current methods (Lin, Xu, and Zhang 2020; Zhang et al. 2021b; Vaze et al. 2022) usually employ two steps: pretraining on labeled data for representation learning and then pseudo-label training on unlabeled data to discover categories. Although these methods can learn some discriminative features for novel categories, they usually face following limitations. First, these methods cannot transfer category-specific knowledge explicitly from labeled data to unlabeled data, which can lose high-level semantic information and impair model performance on known categories. Even though these methods can transfer some general knowledge implicitly through the pretrained feature extractor, they usually fail to transfer category-specific knowledge since they discard the pretrained classifier after pre-training and cannot align known categories in labeled and unlabeled data. Second, these methods cannot decouple known and novel categories from unlabeled data to achieve different training goals for them, which can hurt model's generalization and discriminative ability. Without considering different training objectives for known and novel categories, these methods simply treat them the same and learn about them in a coupled unsupervised manner, which can prevent models learning about discriminative features for both known and novel categories.

In this paper, we first provide insight of different training objectives for known and novel categories in GCD, which is ignored by previous methods. Since we have both labeled and unlabeled data for known categories, we want to **exploit** knowledge acquired from labeled data to improve model's **generalization ability** on these categories in a **semi-supervised manner**. Meanwhile we have only unlabeled data for novel categories, we mainly want to **explore** some novel knowledge to improve model's **discriminative ability** on these categories in an **unsupervised manner**. So learning about known and novel categories with different training approaches in a decoupled manner is crucial for GCD.

To cope with above challenges, we propose *Decoupled Prototypical Network* (DPN), a novel model which can transfer category-specific knowledge explicitly and decouple known and novel categories to achieve different training goals for them. First, we formulate a bipartite matching problem for category prototypes to align known categories in labeled

and unlabeled data, meanwhile decouple known and novel categories from unlabeled data, without introducing any additional model parameters. Specifically, we first learn a set of category prototypes for labeled data to represent known categories and another set of category prototypes for unlabeled data to represent both known and novel categories. Then we align these two sets of prototypes by formulating it as a bipartite matching problem, which can be solved efficiently with Hungarian algorithm (Kuhn 1955). And the matched prototypes in unlabeled data correspond to known categories and the unmatched prototypes correspond to novel categories. In this way, DPN can decouple known and novel categories from unlabeled data and learn about them with different training manners to achieve different training objectives. Furthermore, DPN can utilize the prototypes learned from labeled data to guide the pseudo-label training process on unlabeled data in a category-specific way. So DPN can transfer not only general knowledge through the pre-trained feature extractor, but also category-specific knowledge through prototypes to capture high-level semantics and improve model performance on known categories.

To learn more discriminative features for both known and novel categories, we further propose *Semantic-aware Prototypical Learning* (SPL). Different from traditional prototypical learning (Snell, Swersky, and Zemel 2017; Li et al. 2020) which assigns each instance into a single prototype based on noisy pseudo labels, SPL assigns instances into prototypes in a soft manner based on semantic similarities between them, which can acquire meaningful semantic knowledge and alleviate the effect of pseudo-label noise. Finally, we update the prototypes learned from labeled data through *Exponential Moving Average* (EMA) algorithm. In addition to utilizing knowledge obtained from labeled data through prototypes, we also iteratively update these prototypes to capture the latest acquired knowledge so that it can generalize better to unseen instances. And the EMA algorithm can yield more consistent representations for prototypes to make the training process more stable.

Our main contributions can be summarized as follows:

- We propose a novel *Decoupled Prototypical Network* (DPN) for Generalized Category Discovery. By formulating a bipartite matching problem for category prototypes, DPN can decouple known and novel categories to achieve different training goals for them and transfer category-specific knowledge to capture high-level semantics.

- We propose *Semantic-aware Prototypical Learning* (SPL) to learn more discriminative features, which can capture semantic relationships between instances and prototypes while alleviating the effect of pseudo-label noise.

- Extensive experiments on multiple benchmark datasets show that our model outperforms state-of-the-art methods by a large margin.

Related Work

Generalized Category Discovery

Generalized Category Discovery (GCD) was first comprehensively formalized by Vaze et al. (2022). They further pro-

posed a simple baseline to use contrastive learning and semi-supervised learning to discover novel categories. Another line of work aimed to discover novel user intents from unlabeled utterances, and they usually adopted pseudo-labeling methods to train their models. For example, Lin, Xu, and Zhang (2020) adopted pseudo supervision based on pairwise similarities to guide the clustering process for all categories. Zhang et al. (2021b) performed clustering with an alignment strategy to generate pseudo labels for all unlabeled data to learn about known and novel categories. Yu et al. (2022) proposed to perform self-labeling and mutual-information maximization to learn discriminative features for both known and novel categories. However, these methods cannot explicitly transfer category-specific knowledge from labeled data to unlabeled data, which can hurt model performance on known categories. Furthermore, they learn about known and novel categories in a coupled way, which can hurt model's generalization and discriminative ability.

Prototypical Learning

Prototypical Learning (PL) assumes that each category can be represented by a single prototype in the feature space (Snell, Swersky, and Zemel 2017). Most PL methods focused on matching instances to prototypes by computing distances between them. For example, Yue et al. (2021) performed cross-domain instance-to-prototype matching for Unsupervised Domain Adaptation (UDA). Pan et al. (2019) performed PL to jointly bridge the gap between domains and construct classifiers in the target domain for UDA. There were also some works that utilized PL for interpretability. For instance, Chen et al. (2019) performed interpretable image recognition by comparing image parts to learned prototypes. Nauta, van Bree, and Seifert (2021) combined PL with decision trees to perform image recognition to address accuracy-interpretability trade-off. However, these methods simply assigned each instance into a single prototype based on one-hot pseudo labels, which can easily make models overfit on the noisy pseudo labels in the GCD setting.

Method

Problem Statement

Traditional classification models are developed based on a labeled dataset $\mathcal{D}^l = \{(x_i, y_i)|y_i \in \mathcal{Y}_k\}$, which contains only known categories $\mathcal{Y}_k$. However, in the real world, the deployed model may encounter unlabeled data $\mathcal{D}^u = \{x_i|y_i \in \{\mathcal{Y}_k, \mathcal{Y}_n\}\}$ which contains both known categories $\mathcal{Y}_k$ and novel categories $\mathcal{Y}_n$. So the aim of Generalized Category Discovery (GCD) is to recognize both known and novel categories based on $\mathcal{D}^l$ and $\mathcal{D}^u$. Finally, model performance will be tested on the testing set $\mathcal{D}^t = \{(x_i, y_i)|y_i \in \{\mathcal{Y}_k, \mathcal{Y}_n\}\}$.

Approach Overview

An overview of our *Decoupled Prototypical Network* (DPN) is shown in Figure 1, which contains five steps (marked in the figure). First, we pretrain a feature extractor F_θ on both labeled and unlabeled data for representation learning. Second, we learn two sets of category prototypes based on F_θ for labeled and unlabeled data, respectively. Third, We
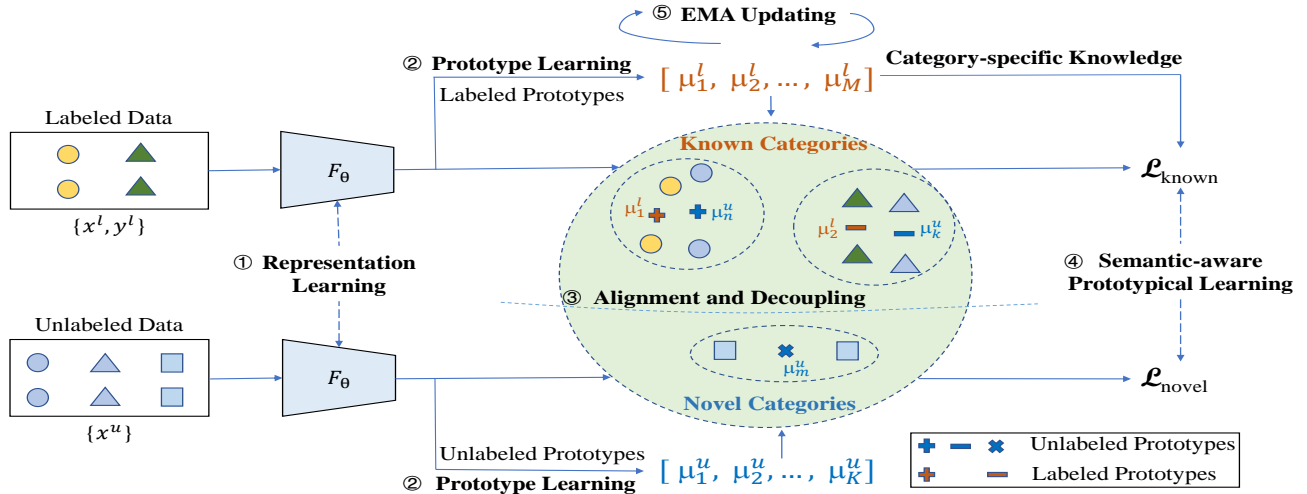
Figure 1: An overview of our model.

match these two sets of prototypes to align known categories in labeled and unlabeled data, meanwhile decouple known and novel categories from unlabeled data with our proposed "*Alignment and Decoupling*" strategy. Fourth, we propose *Semantic-aware Prototypical Learning* (SPL) to learn more discriminative features for known and novel categories in a decoupled way. We also use the prototypes learned from the labeled data to transfer category-specific knowledge explicitly. Finally, we update category prototypes of labeled data with Exponential Moving Average (EMA) algorithm to capture the latest acquired knowledge.

Representation Learning

We use the pretrained language model BERT (Devlin et al. 2018) as our feature extractor $F_\theta : \mathcal{X} \to \mathbb{R}^d$. To learn semantic representations for current tasks, we fine-tune F_θ on both labeled and unlabeled data. We employ a joint pretraining loss following Zhang et al. (2021c), which contains Cross-Entropy (CE) loss for labeled data and Masked Language Modeling (MLM) loss (Devlin et al. 2018) for all data:

$$\mathcal{L}_{pre} = \mathcal{L}_{ce}(\mathcal{D}^l) + \mathcal{L}_{mlm}(\mathcal{D}^l, \mathcal{D}^u) \qquad (1)$$

$\mathcal{D}^l$ and $\mathcal{D}^u$ are labeled and unlabeled dataset, respectively. Through pretraining, models can acquire various general knowledge for both known and novel categories and learn meaningful semantic representations for subsequent tasks.

Learning Category Prototypes

After pretraining, most existing work (Lin, Xu, and Zhang 2020; Zhang et al. 2021b, 2022) generated pseudo labels for all unlabeled data to learn about known and novel categories in a coupled way. However, they ignored the different training goals for known and novel categories, and the coupled training manner made them difficult to transfer category-specific knowledge for known categories since they cannot align known categories in labeled and unlabeled data. To mitigate above problems, we propose a novel "*Alignment and Decoupling*" strategy. Specifically, we learn two sets

of category prototypes for labeled and unlabeled data, respectively. By aligning these two sets of prototypes, we can not only align known categories in labeled and unlabeled data to transfer category-specific knowledge explicitly in a category-to-category way, but also decouple known and novel categories from unlabeled data to acquire different knowledge about them with different training manners.

For labeled data, we take average of all instance embeddings belonging to the same category as labeled prototypes $P^l = \{\mu_j^l\}_{j=1}^M$, where $\mu_j^l = \frac{1}{|C_j|} \sum_{x_i \in C_j} F_\theta(x_i)$ denotes the labeled prototype for category j, C_j denotes a set of instances from category j and $M = |\mathcal{Y}_k|$ is the number of known categories. For unlabeled data, we first perform KMeans clustering to get clusters $C^u = \{C_1^u, C_2^u, ..., C_K^u\}$, where $K = |\mathcal{Y}_k| + |\mathcal{Y}_n|$ is the number of total categories. We presume prior knowledge of K following previous works (Zhang et al. 2021b, 2022) to make a fair comparison and we tackle the problem of estimating this parameter in the experiment. Then we take average of all instance embeddings belonging to the same cluster as unlabeled prototypes $P^u = \{\mu_j^u\}_{j=1}^K$, where $\mu_j^u = \frac{1}{|C_j^u|} \sum_{x_i \in C_j^u} F_\theta(x_i)$.

Alignment and Decoupling

Instead of introducing extra data and parameters to train a binary classifier like previous two-stage methods (Vedula et al. 2020), we propose to decouple known and novel categories in unlabeled data by aligning the learned prototypes P^u and P^l. Since unlabeled data contain all known categories (following assumptions of previous work), we can find prototypes in P^u which represent known categories. Intuitively, we think the closest prototypes in labeled and unlabeled data represent the same category. So we can formulate a bipartite matching problem to align known categories in the prototype sets P^u and P^l. Specifically, to find a bipartite matching between P^l and P^u, we search for all possible permutation $\mathcal{P}_{all}$ for the set P^u and get the optimal permutation $\hat{\mathcal{P}}$ by minimizing the

total matching cost:

$$\hat{\mathcal{P}} = \arg\min_{\mathcal{P} \in \mathcal{P}_{all}} \sum_{i=1}^{M} \mathcal{L}_{match}(\mu_i^l, \mu_{\mathcal{P}(i)}^u) \qquad (2)$$

where $\mathcal{L}_{match}(\mu_i^l, \mu_{\mathcal{P}(i)}^u)$ is point-to-point matching cost between the labeled prototype μ_i^l and the unlabeled prototype $\mu_{\mathcal{P}(i)}^u$. Here, we use the Euclidean distance as the matching cost to find the minimum distance matching:

$$\mathcal{L}_{match}(\mu_i^l, \mu_{\mathcal{P}(i)}^u) = \left\| \mu_i^l - \mu_{\mathcal{P}(i)}^u \right\|_2 \qquad (3)$$

The bipartite matching problem can be solved efficiently with the Hungarian algorithm (Kuhn 1955). After that, we can get an optimal matching between P^l and parts of P^u: $\{\mu_j^l, \mu_{\hat{\mathcal{P}}(i)}^u\}_{i=1}^{M}$. So these matched prototypes in P^u can be considered to represent known categories, denoting as $P^{uk} = \{\mu_{\hat{\mathcal{P}}(i)}^u\}_{i=1}^{M}$. And the remaining unmatched prototypes in P^u are seen to represent novel categories, denoted as $P^{un} = \{\mu_{\hat{\mathcal{P}}(i)}^u\}_{i=M+1}^{K}$. Furthermore, the unlabeled data $\mathcal{D}^u$ can also be decoupled into two parts: $\mathcal{D}^{uk}$ for data with known categories and $\mathcal{D}^{un}$ for data with novel categories, which are decided by the clusters data belong to.

Semantic-aware Prototypical Learning

By decoupling known and novel categories, we can design specific training loss for them to achieve different training goals. Specifically, we propose *Semantic-aware Prototypical Learning* (SPL) to learn about novel categories in an unsupervised manner. Then we transfer category-specific knowledge from labeled data to learn about known categories in a semi-supervised manner.

Unsupervised Learning for Novel Categories. Since data with novel categories are unlabeled, we aim to explore some novel knowledge to improve model's discriminative ability on these categories with unsupervised learning. Previous work focused on instance-level discrimination with pseudo-label training (Zhang et al. 2021b) or contrastive learning (Zhang et al. 2022). However, they ignored the high-level semantics between instances and categories. To capture high-level semantics, Prototypical Learning (PL) (Snell, Swersky, and Zemel 2017) was proposed to pull instances closer to prototypes they belong to and separate instances far away from other irrelevant prototypes:

$$\mathcal{L}_{pl} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{e^{-d(F_\theta(x_i), \mu_g^u)}}{\sum_{j=1}^{K} e^{-d(F_\theta(x_i), \mu_j^u)}} \qquad (4)$$

where x_i belongs to cluster C_g^u and μ_g^u is the corresponding prototype. $d(\cdot, \cdot)$ is a distance function (e.g., Euclidean distance (Snell, Swersky, and Zemel 2017) or cosine distance (Li et al. 2020)).

However, PL simply divides each instance into a single prototype in a hard way and does not consider semantic relationships between instances and prototypes. Furthermore, since the mapping between instances and prototypes comes from pseudo-labels, PL is easily affected by pseudo-label

Table 1: Statistics of datasets. $|\mathcal{Y}_k|$, $|\mathcal{Y}_n|$, $|\mathcal{D}^l|$, $|\mathcal{D}^u|$ and $|\mathcal{D}^t|$ represent the number of known categories, novel categories, labeled data, unlabeled data and testing data, respectively.

Dataset	$|\mathcal{Y}_k|$	$|\mathcal{Y}_n|$	$|\mathcal{D}^l|$	$|\mathcal{D}^u|$	$|\mathcal{D}^t|$
BANKING	58	19	673	8,330	3,080
StackOverflow	15	5	1,350	16,650	1,000
CLINC	113	37	1,344	16,656	2,250

noise. To mitigate above issues, we propose *Semantic-aware Prototypical Learning* (SPL). Instead of assigning an instance into a single prototype based on the noisy pseudo labels, SPL assigns each instance into all prototypes using semantic similarity as weights:

$$\mathcal{L}_{spl} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} d(F_\theta(x_i), \mu_k^u) \frac{e^{s(F_\theta(x_i), \mu_k^u)}}{\sum_{j=1}^{K} e^{s(F_\theta(x_i), \mu_k^u)}} \qquad (5)$$

where $n = |\mathcal{D}^{un}|$ is the number of unlabeled instances belonging to novel categories, $d(\cdot, \cdot)$ is a distance function to pull instances and prototypes closer, $s(\cdot, \cdot)$ is a similarity function to evaluate semantic similarities between instances and prototypes. We use Euclidean distance as $d(\cdot, \cdot)$ and cosine similarity as $s(\cdot, \cdot)$ for novel categories in our paper. So Eq. (5) can be written as:

$$\mathcal{L}_{spl} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \|F_\theta(x_i) - \mu_k^u\|_2 \frac{e^{cos(F_\theta(x_i), \mu_k^u)/\tau}}{\sum_{j=1}^{K} e^{cos(F_\theta(x_i), \mu_j^u)/\tau}}$$
$$(6)$$

where τ is a temperature hyperparameter. By weighting with semantic similarities and assigning instances into prototypes in a soft way, SPL can capture meaningful semantic information between instances and prototypes and alleviate the effect of pseudo-label noise, especially for instances near decision boundaries. For simplicity, We denote the loss for novel categories as:

$$\mathcal{L}_{novel} = \mathcal{L}_{spl}(\mathcal{D}^{un}, P^{un}) \qquad (7)$$

Semi-supervised Learning for Known Categories. For known categories, we aim to exploit knowledge acquired from labeled data to improve model's generalization ability on these categories with semi-supervised learning. Previous approaches only used labeled data to transfer some general knowledge by pretraining models. However, they do not consider transferring high-level category knowledge explicitly for known categories to guide the pseudo-label training process, which can lose high-level semantic information and impair their model performance. To mitigate this issue, in addition to using SPL to learn some discriminative features, we further utilize the labeled prototypes P^l to guide the pseudo-label training process and transfer category-specific

Table 2: Model comparison results (%) on testing sets. Average results over 3 runs are reported.

Method	BANKING			StackOverflow			CLINC		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
DeepCluster	13.95	13.94	13.99	17.37	18.22	14.80	26.92	27.34	25.67
DCN	17.85	18.94	14.35	29.10	28.94	29.51	29.64	30.00	28.45
DEC	19.30	20.36	15.84	19.30	20.36	15.84	19.99	20.18	19.40
BERT	21.29	21.48	20.70	16.80	16.67	17.20	34.52	34.98	33.16
KM-GloVe	29.18	29.11	29.39	28.40	28.60	28.05	51.64	51.74	51.50
AG-GloVe	30.09	29.69	31.29	29.23	28.49	31.56	44.70	45.17	43.20
SAE	38.05	38.29	37.27	60.33	57.36	69.02	46.59	47.35	44.24
Semi-DC	50.73	53.37	42.63	64.90	66.13	61.20	74.52	75.60	71.34
CDAC+	53.09	55.42	46.01	76.67	77.51	74.13	69.75	70.08	68.77
Self-Labeling	56.19	61.64	39.56	71.03	78.53	48.53	72.69	80.06	49.65
DTC	56.56	59.98	46.10	70.50	80.93	51.87	76.42	82.34	58.95
DAC	63.63	69.60	45.44	70.77	76.13	54.67	84.42	89.10	70.59
Semi-KM	66.23	73.62	43.68	73.13	81.02	49.47	81.42	89.03	59.01
LASKM	67.55	75.16	44.34	74.83	82.00	53.33	79.26	89.64	48.66
DPN (Ours)	**73.61**	**80.04**	**53.99**	**83.20**	**84.18**	**80.27**	**87.62**	**91.29**	**76.79**
Improvement	+6.06	+4.88	+7.89	+6.53	+2.18	+6.14	+3.20	+1.65	+5.45

knowledge explicitly as a regularization term:

$$\mathcal{L}_{reg} = \frac{1}{r} \sum_{i=1}^{r} \sum_{k=1}^{M} (1 - cos(F_\theta(x_i), \mu_k^l)) \frac{e^{cos(F_\theta(x_i), \mu_k^l)/\tau}}{\sum_{j=1}^{M} e^{cos(F_\theta(x_i), \mu_j^l)/\tau}}$$
(8)

where $r = |\mathcal{D}^{uk}|$ is the number of unlabeled instances belonging to known categories. We use cosine distance as the distance function here to acquire different knowledge from P^u. By pulling instances closer to labeled prototypes using semantic similarity as weights, our model can transfer category-specific knowledge from labeled data to unlabeled data and further mitigate the risk of model overfitting on the noisy pseudo labels. To avoid catastrophic forgetting for knowledge acquired from labeled data, we also add cross-entropy loss during training. So the loss for known categories can be denoted as:

$$\mathcal{L}_{known} = \mathcal{L}_{spl}(\mathcal{D}^{uk}, P^{uk}) + \mathcal{L}_{ce}(\mathcal{D}^l) + \lambda \cdot \mathcal{L}_{reg}(\mathcal{D}^{uk}, P^l)$$
(9)

where λ is a weighting factor for the regularization term. In this way, we can fully utilize knowledge acquired from labeled data, labeled prototypes and unlabeled prototypes to improve model's generalization ability on known categories in a semi-supervised manner.

Total Loss. Overall, the training objective of our model can be formulated as:

$$\mathcal{L}_{dpn} = \mathcal{L}_{novel} + \mathcal{L}_{known}$$
(10)

Updating Category Prototypes

Despite of utilizing labeled prototypes to guide the decoupling and pseudo-label training process, we also update them periodically to capture the latest acquired knowledge so that

they can generalize better to unseen instances. However, directly overwriting the original values can lead to less consistent representations for prototypes and make them easily affected by outliers, which will make training unstable. So we use Exponential Moving Average (EMA) algorithm to learn more robust prototypes for labeled data:

$$\mathcal{P}_{t+1}^l \leftarrow \alpha \cdot \mathcal{P}_t^l + (1 - \alpha) \cdot \mathcal{P}_{t+1}^l$$
(11)

where α is a momentum factor. By adjusting α, we can learn more consistent and robust prototype representations. Here, we fix unlabeled prototypes P^u to avoid the time-consuming clustering process and the influence of randomly permuted cluster indices, without influencing our model performance. After training, we apply a clustering algorithm (e.g., KMeans) to obtain cluster assignments for testing data.

Experiments

Experimental Setup

Datasets. We evaluate our model on three benchmark datasets. **BANKING** is an intent classification dataset in the banking domain released by Casanueva et al. (2020). **StackOverflow** is a technical question classification dataset processed by Xu et al. (2015). **CLINC** is a text classification dataset with diverse domains released by Larson et al. (2019). Statistics of these datasets can be found in Table 1.

Comparison Methods. We compare our model with various baselines and state-of-the-art (SOTA) methods.
Unsupervised Methods. (1) KM-GloVe: KMeans (MacQueen et al. 1967) with GloVe embeddings (Pennington, Socher, and Manning 2014). (2) AG-GloVe: Agglomerative Clustering (Gowda and Krishna 1978) with GloVe embeddings. (3) SAE: Stacked Auto Encoder. (4) DEC: Deep Embedding Clustering (Xie, Girshick, and Farhadi 2016). (5)

DCN: Deep Clustering Network (Yang et al. 2017). (6) KM-BERT: KMeans with BERT embeddings (Devlin et al. 2018). (7) DeepCluster: Deep Clustering (Caron et al. 2018). **Semi-supervised Methods.** (1) KM-Semi: KMeans with BERT pretrained on labeled data. (2) DeepCluster-Semi: Deep Clustering pretrained on labeled data. (3) DTC: Deep Transfer Clustering (Han, Vedaldi, and Zisserman 2019). (4) CDAC+: Constrained Adaptive Clustering (Lin, Xu, and Zhang 2020). (5) DAC: Deep Aligned Clustering (Zhang et al. 2021b). (6) Self-Labeling: Self-Labeling Framework (Yu et al. 2022). (7) LASKM: Label Assignment with Semi-supervised KMeans (Vaze et al. 2022).

Evaluation Metrics. We measure accuracy between ground-truth labels and model's cluster assignments on the testing set with Hungarian algorithm (Kuhn 1955). (1) All: accuracy for all instances. (2) Known: accuracy for instances with known categories. (3) Novel: accuracy for instances with novel categories.

Implementation Details. We use the pretrained BERT model (bert-base-uncased) implemented by Pytorch (Wolf et al. 2019) and adopt its suggested hyper-parameters. We pretrain model on all Transformer layers and only fine-tune on the last three Transformer layers to speed up calculation. We use the *[CLS]* token of the last Transformer layer as instance feature. For model optimization, we use AdamW optimizer. Early stopping is used during pretraining, which is decided by model performance on the validation set which only contains known categories. For comparison methods, we use the implementations and hyper-parameters in their original papers. And some implementations are based on Zhang et al. (2021a). For hyper-parameters, λ is set to 1, α is set to 0.999 and τ is set to 0.1. Training epochs for StackOverflow, BANKING and CLINC dataset are set to 10, 60 and 80. The wait patience for early stopping is set to 20. The learning rate for pretraining is set to $5e^{-5}$ and the learning rate for training is set to $1e^{-5}$. For masked language modeling, the mask probability is set to 0.15 following previous works.

Experimental Results

Main Results. The results are shown in Table 2. Overall, our model outperforms all comparison methods on all datasets and evaluation metrics with a large margin. First, our model achieves the best results on accuracy for all instances (denoted as 'All'). Specifically, our model outperforms the SOTA model by **5.26%** of average accuracy on three benchmark datasets, which can reflect the effectiveness of our model on both known and novel categories. Second, our model achieves the best performance on accuracy for known categories (denoted as 'Known'). In detail, our model improves average accuracy of known categories by **2.90%** over the SOTA method since we decouple known categories from unlabeled data and fully utilize acquired knowledge to learn about them in a semi-supervised manner. By aligning prototypes learned from labeled and unlabeled data, our model can decouple known and novel categories from unlabeled data effectively, so that we can obtain different knowledge with different training approaches. Furthermore, by aligning known categories in labeled and unlabeled data, our

Table 3: Results (%) of different model variants.

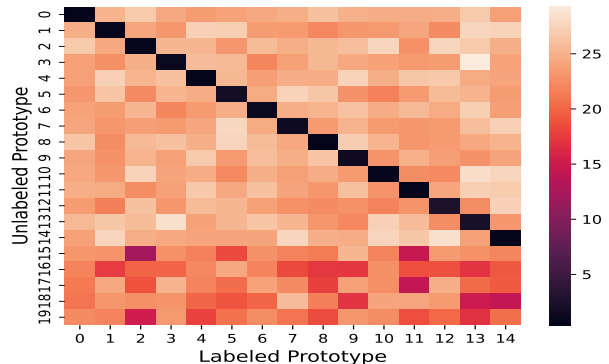| Model	All	Known	Novel
Ours	83.20	84.18	80.27
w/o Cross Entropy	82.53	83.47	79.73
w/o EMA	82.00	81.87	74.00
w/o Decoupling	81.10	82.93	75.60
w/o Soft Assignment	80.80	83.87	71.60
w/o Semantic Weights	70.50	72.53	64.40

Figure 2: Distances between 15 labeled prototypes and 20 aligned unlabeled prototypes. Darker colors represent closer distances.

model can transfer category-specific knowledge from labeled data to unlabeled data explicitly through prototypes. Third, our model also gets the best performance on accuracy for novel categories (denoted as 'Novel'). More specifically, average improvement of **6.49%** can be seen on the accuracy of novel categories compared with the SOTA method, due to our *Semantic-aware Prototypical Learning* (SPL). By measuring semantic similarities between instances and prototypes, our model can assign each instance into different prototypes in a soft manner, which can capture semantic information and mitigate the effect of pseudo-label noise.

Ablation Study. The performance of variants of our model on the StackOverflow dataset is shown in Table 3. Overall, removing different components from our model will affect model performance more or less, which can show the effectiveness of different components of our model. Specifically, (1) Removing **Cross Entropy** loss $\mathcal{L}_{ce}$ in Eq. (9) has minimal influence since our model can also obtain knowledge for known categories through labeled prototypes. (2) Removing **EMA** updating for labeled prototypes in Eq. (11) will lead to inconsistent representations for prototypes and hurt model performance. (3) Removing **Decoupling** will prevent the category-specific knowledge transfer in Eq. (8) and hurt model performance on both known and novel categories. (4) Removing **Soft Assignment** in Eq. (5) can easily make models overfit on the noisy pseudo labels generated by clustering and damage model performance, especially for novel categories. (5) Removing **Semantic Weights** $s(F_\theta(x_i), \mu_k^u)$ in Eq. (5) can lose semantic information between instances and
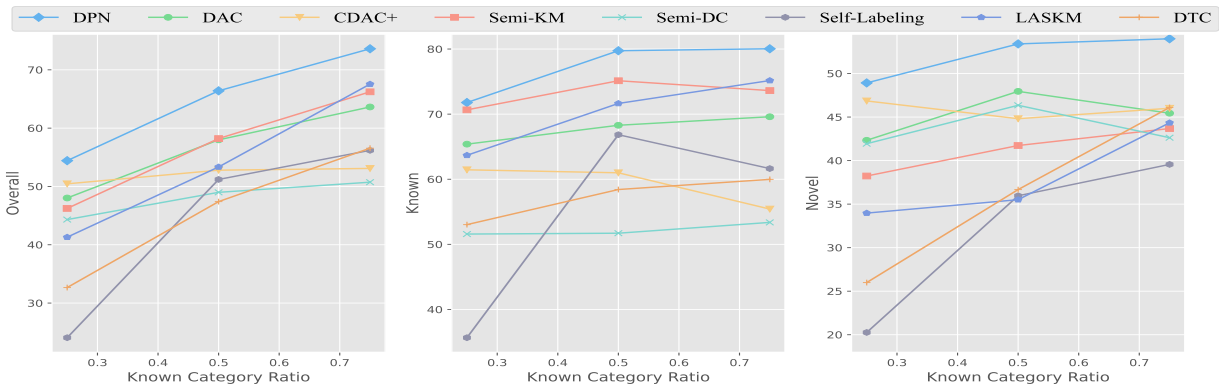
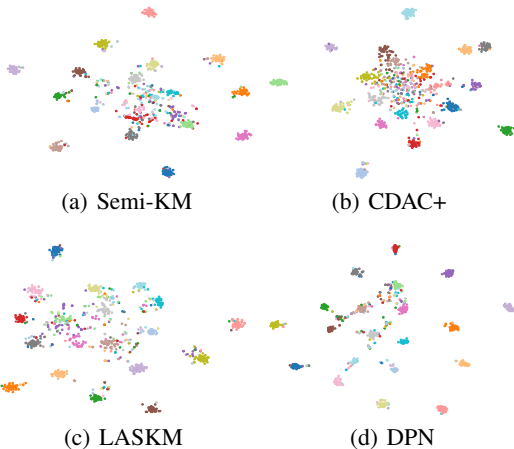Figure 3: Effect of known category ratio on the BANKING dataset.

(a) Semi-KM   (b) CDAC+

(c) LASKM   (d) DPN

Figure 4: The t-SNE visualizations of embeddings.

Table 4: Estimation of the number of categories.

	CLINC	BANKING	StackOverflow
Ground Truth	150	77	20
DAC	130	66	15
Ours	137	67	18
Error	8.7%	13.0%	10.0%

Influence of Known Category Ratio. To investigate the influence of known category ratio on model performance, we vary it in the set $\{0.25, 0.50, 0.75\}$. As shown in Figure 3, our model achieves the best performance under different settings on all evaluation metrics, which can show effectiveness and robustness of our model.

Feature Visualization. In Figure 4, we use t-SNE to illustrate embeddings learned by different methods on the Stack-Overflow dataset. It clearly shows that our model can learn more separable features for different categories than comparison methods, which can further demonstrate the effectiveness of our model.

Conclusion

In this paper, we propose *Decoupled Prototypical Network* (DPN) for Generalized Category Discovery. Inspired by the different training goals for known and novel categories, we propose to decouple them from unlabeled data to acquire different knowledge. Then we formulate a bipartite matching problem for prototypes to align known categories in labeled and unlabeled data to transfer category-specific knowledge, meanwhile decouple known and novel categories to acquire different knowledge. Furthermore, we propose *Semantic-aware Prototypical Learning* (SPL) to learn more discriminative features from unlabeled data. By assigning each instance into different prototypes using semantic similarities as weights, our model can capture meaningful semantic information to learn about known and novel categories and alleviate the effect of noisy pseudo labels. Experimental results on three benchmark datasets show that our model outperforms state-of-the-art methods by a large margin.

prototypes and largely degrade model performance.

Effectiveness of Alignment and Decoupling Strategy. To investigate the effectiveness of our *Alignment and Decoupling* strategy, we visualize the heat map of distances between labeled prototypes and aligned unlabeled prototypes on the StackOverflow dataset in Figure 2. We can see that our strategy can align these two sets of prototypes effectively. Since prototypes correspond to categories, our model can align known categories in labeled and unlabeled data in a category-to-category way (the black grids) to transfer category-specific knowledge. Meanwhile, our model can decouple known and novel categories from unlabeled data, so that we can obtain different knowledge for them with different training objectives.

Estimating the Number of Categories. Here, we use the algorithm proposed in DAC (Zhang et al. 2021b) to tackle the problem of estimating the number of categories K from unlabelled dataset $\mathcal{D}^u$. The results are reported in Table 4. We can see that our model achieves lower error rates on all datasets than DAC, which means that our model can fully acquire knowledge from both labeled and unlabeled data to learn better representations to estimate the number of categories.

Acknowledgments

References

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149.

Casanueva, I.; Temčinas, T.; Gerz, D.; Henderson, M.; and Vulić, I. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gowda, K. C.; and Krishna, G. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2): 105–112.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8401–8409.

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

Larson, S.; Mahendran, A.; Peper, J. J.; Clarke, C.; Lee, A.; Hill, P.; Kummerfeld, J. K.; Leach, K.; Laurenzano, M. A.; Tang, L.; et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Lin, T.-E.; Xu, H.; and Zhang, H. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8360–8367.

MacQueen, J.; et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.

Nauta, M.; van Bree, R.; and Seifert, C. 2021. Neural prototype trees for interpretable fine-grained image recognition.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14933–14943.

Pan, Y.; Yao, T.; Li, Y.; Wang, Y.; Ngo, C.-W.; and Mei, T. 2019. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2239–2247.

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

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7492–7501.

Vedula, N.; Gupta, R.; Alok, A.; and Sridhar, M. 2020. Automatic discovery of novel intents & domains from text utterances. *arXiv preprint arXiv:2006.01208*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Xu, J.; Wang, P.; Tian, G.; Xu, B.; Zhao, J.; Wang, F.; and Hao, H. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 62–69.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, 3861–3870. PMLR.

Yu, Q.; Ikami, D.; Irie, G.; and Aizawa, K. 2022. Self-Labeling Framework for Novel Category Discovery over Domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yue, X.; Zheng, Z.; Zhang, S.; Gao, Y.; Darrell, T.; Keutzer, K.; and Vincentelli, A. S. 2021. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13834–13844.

Zhang, H.; Li, X.; Xu, H.; Zhang, P.; Zhao, K.; and Gao, K. 2021a. TEXTOIR: An integrated and visualized platform for text open intent recognition. *arXiv preprint arXiv:2110.15063*.

Zhang, H.; Xu, H.; Lin, T.-E.; and Lyu, R. 2021b. Discovering New Intents with Deep Aligned Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, J.; Bui, T.; Yoon, S.; Chen, X.; Liu, Z.; Xia, C.; Tran, Q. H.; Chang, W.; and Yu, P. 2021c. Few-shot intent detection via contrastive pre-training and fine-tuning. *arXiv preprint arXiv:2109.06349*.

Zhang, Y.; Zhang, H.; Zhan, L.-M.; Wu, X.-M.; and Lam, A. 2022. New Intent Discovery with Pre-training and Contrastive Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.