Knowledge-Based Systems 251 (2022) 109296

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Incorporating geometry knowledge into an incremental learning structure for few-shot intent recognition

Xin Zhang, Miao Jiang, Honghui Chen, Jianming Zheng *, Zhiqiang Pan

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, Hunan, China

ARTICLE INFO

Article history:
Received 22 March 2022
Received in revised form 15 June 2022
Accepted 16 June 2022
Available online 21 June 2022

Keywords:
Few-shot learning
Incremental learning
Intent recognition

ABSTRACT

Few-shot incremental intent recognition aims at continually identifying users' intents from utterances with limited labeled novel data. To mitigate the effect of catastrophic forgetting inherent in incremental learning, existing methods generally resort to storing informative base exemplars as the model memory that are further replied to when learning novel classes. However, merely preserving these base exemplars and ignoring the relationship between them can easily trigger some concerns in the incremental training process. First, in each iteration, the participation of novel samples further updates the embedding space to make the novel labels well separated, which instead poisons the learned labels, reducing their separability. Second, the overfitting risk is further amplified in incremental learning, causing models to be overconfident on seen base classes but hardly generalized to unseen novel classes. In view of these problems, a geometry-aware learning (GAL) model is proposed to address few-shot incremental intent recognition and bridge the two research gaps mentioned above. Specifically, in our proposal, for the embedding shifting issue, GAL constructs a geometric structure of selected exemplars based on their spatial distribution in the embedding space, which is treated as a strong constraint in the following training. For the overfitting issue, GAL introduces an episodic-based pretraining strategy as well as a multisource contrastive-based loss to enhance the lack of supervised signals in the classification of data-scarce novel labels. Experimental results on a public dataset OOS (CLINC-150) verify the effectiveness of our proposal by beating the state-of-the-art baselines. Specifically, our model outperforms the best baseline by 1.34% and 0.39% on the 5-way 1-shot meta task for the base and novel classes, respectively. It also outperforms the best baseline by 1.89% and 0.80% on the 5-way 5-shot meta task for the base and novel classes, respectively. Furthermore, cross-domain experiments between two datasets (CLINC-150 and ATIS) reflect that GAL has better generalization ability across different domains. The method is superior to traditional intent recognition models in the application scenario where the model is required to accurately distinguish both new and trained categories under the low-data dilemma. Analogously, this common application scenario can also provide a broad developing direction for the dialog recommender system, which helps it recommend continuously and accurately with only a few labeled samples.

1. Introduction

Intent recognition concentrates on recognizing users' potential purposes from utterances, e.g., detecting the intent *"text"* from the given sentence *"Text Christy and ask her what she wants for dinner"*. This task could further assist many downstream tasks, such as dialog systems [1,2] and recommender systems [3,4].

In real-world applications, however, limited human annotation cannot afford the constant emergence of new intent labels, which inevitably leads to a low-data regime in the intent recognition task. To address such a low-resource dilemma, few-shot learning (FSL) [5–9] is proposed to learn the generalization ability within limited labeled data, which is expected to have humanlike cognitive abilities [8–11]. Despite much progress, the episode training format drives existing few-shot intent recognition models to pay too much attention to the novel labels, thereby sacrificing the performance of learned labels. Such a phenomenon in which few-shot models gradually lose the classification ability of learned labels is called "catastrophic forgetting" [12]. To mitigate this forgetting trend, few-shot incremental learning (FSIL) has been proposed, which learns to classify novel categories while maintaining the recognition ability of old categories.

Existing FSIL methods can be grouped into two categories. The first one is parameter-based FSIL, which attempts to solve the FSIL

* Corresponding author.
E-mail addresses: zhangxin16@nudt.edu.cn (X. Zhang), jiangmiao20@nudt.edu.cn (M. Jiang), chenhonghui@nudt.edu.cn (H. Chen), zhengjianming12@nudt.edu.cn (J. Zheng), panzhiqiang@nudt.edu.cn (Z. Pan).

https://doi.org/10.1016/j.knosys.2022.109296
0950-7051/

task by directly adjusting the model structure or updating their parameters, such as the work proposed by Gidaris and Komodakis [13]. The second one is memory-based FSIL, which stores and replays the most informative part of the seen instances as memories to consolidate the learned knowledge, such as the proposals of Han et al. [14] and Cao et al. [15]. Compared with the parameter-based FSIL, memory-based FSIL methods are more effective in avoiding "catastrophic forgetting"; hence, they have been extensively used. However, solely recording representative instances would trigger some concerns. **First,** the update of memorized instances always lags behind that of the overall embedding space. Taking the first step increment as an example, the memorized instances only make base labels well separated in the current embedding space; however, the newly added novel labels further update this embedding space, causing asynchronous updates between the base and novel labels. Such asynchronous updates cannot guarantee the overall separability of all labels, making the embedding space shift in uncontrollable directions. **Second,** the imbalanced data distribution in few-shot incremental learning, i.e., data-abundant base classes against data-scarce novel classes, further amplifies the overfitting risk, causing models to be overconfident on base classes but hardly generalized to unseen novel classes [16].

It can be concluded that the few-shot incremental intent recognition task has the following research gaps. Existing studies only store base samples as independent individuals and ignore their relationships with each other, which leads to the embedding shift and weakening of the separability of base categories in subsequent incremental learning. In addition, due to the large amount of base category data and the small amount of novel category data, such a gap exacerbates the risk of overfitting the model on the novel category.

To address the above research gaps, we propose a geometry-aware learning (GAL) model that consists of an episodic-based feature extractor, geometry structures over learned classes, and a multisource contrastive-based loss. In particular, for the embedding shifting issue, GAL creatively introduces the geometry structure in each step increment, which not only remembers the selected instances themselves but also records their relative positions in the embedding space. With this geometry structure, GAL becomes more directed to synchronize the update of the embedding space and the memorized instances. For the overfitting risk in few-shot incremental learning, we attempt to mitigate this adverse effect from two aspects. One is employing episodic-based pretraining on data-abundant base labels, helping the feature extractor to obtain the encoding ability in the low-resource scenario. The other is proposing a multisource contrastive-based loss that pulls apart the embedding distance between the novel-label instances and the memorized exemplars to enhance the deficient supervised signals.

Few-shot incremental experiments on OOS(CLINC-150) [17] for intent recognition illustrate that our proposals achieve obvious improvements over competitive baselines. In particular, not limited to the same domain labels, GAL presents a stronger generalization ability than the other discussed models in the cross-domain experiments. In addition, the ablation study further verifies the effectiveness of the geometry structure in retaining the learned knowledge and demonstrates that the multisource contrastive-based loss can mitigate the intent-label confusion when learning novel classes. In summary, the main contributions of our paper can be briefly listed as follows:

(1) To the best of our knowledge, we are the first to introduce a geometry structure in few-shot incremental intent recognition to record the selected instances as well as their geometry relationship in the embedding space, which mitigates the embedding shifting issue stemming from asynchronous updating.
(2) We introduce episodic-based pretraining to improve the encoding ability of the feature extractor in limited resources. Although its form is the same as episodic learning, their objectives are different. The goal of episodic learning in vanilla few-shot learning only focuses on obtaining a suitable metric to deal with the novel classes. However, episodic-based pretraining in our work aims to acquire the adaptability that generalizes to the few-shot scenario in the incremental learning stage.
(3) We introduce multisource contrastive-based training to further enhance the less-supervised signals when classifying data-scarce novel labels. Traditional contrastive learning selects only positive and negative samples to self-supervise the current task. On top of this, our multisource contrastive-based learning strategy can further select from the stored base classes rather than being limited to the classes in the current task, thereby enhancing the distinction between base and novel categories.
(4) Experimental results demonstrate the effectiveness of our model against the competitive baselines in terms of accuracy in recognizing both base and novel classes. The incremental experiments in either the common setting or the cross-domain setting show that our model presents obvious advantages over baselines, which proves the effectiveness of our proposal.

2. Related work

In this section, we first briefly introduce related works on the intent recognition task in Section 2.1. Then, we summarize the incremental learning methods in Section 2.2, where we discuss traditional incremental learning and few-shot incremental learning.

2.1. Intent recognition

Intent recognition, which aims to form a semantic-frame structure [18] to capture users' intentions from utterances or queries [19], is an important task in the field of natural language understanding (NLU) [20] and is well studied in multiple applications, such as task-oriented dialog systems [21,22]. Various neural architectures have been applied to intent recognition. For instance, Sarikaya et al. [23] and Mesnil et al. [24] attempted to solve this issue with recurrent neural networks (RNNs), while Liu and Lane [25] adopted an attention-based approach and Chen et al. [26] leveraged transformer models to settle the problem [27]. As a pretask of slot filling [28], intent recognition can help the spoken language understanding system determine users' intentions in each turn of dialog so that the slot filling subsystem can accurately determine the type of slots [27]. Thus, the task-oriented dialog systems can respond to users' utterances correctly according to the above results.

When facing the low-resource dilemma, traditional methods are easily trapped in the overfitting morass. Several existing few-shot learning solutions have attempted to address the problem mainly from two aspects, i.e., data augmentation and task-adaptive training with pretrained models [37]. For the former method, Zhang et al. [38] leveraged a nearest neighbor classification schema [39] to fully utilize the limited available samples in both training and testing phases, while Peng et al. [40] attempted to generate utterances for emerging intents with GPT-2 [41]. On the other hand, Casanueva et al. [42] utilized related conversational pretraining models based on a few hundred million conversations.

However, neither the traditional methods nor the recent few-shot learning approaches take into account the demand of

Table 1
Comparison of few-shot incremental learning approaches.

Angle	Parameter-based	Memory-based
Strength	No need for extra space to store the learned knowledge	Strong memory for base knowledge
Weakness	Catastrophic forgetting is prone to occur during incremental learning	Requires additional storage space for base knowledge
Performance[a]	⋆	⋆⋆
Representative works	Gidaris and Komodakis [13], Ren et al. [29], Mazumder et al. [30] and Qi et al. [31]	Tao et al. [32,33], Dong et al. [34], Zhu et al. [35] and Zhang et al. [36]

[a]More stars indicate relatively higher performance.

maintaining the performance on known intent categories while learning novel intents. Our proposal in this paper employs a geometry-aware learning method to simultaneously deal with the continually increasing intents and low-resource problems.

2.2. Incremental learning

Incremental learning means learning from a sequence of data that appears over time [43]. van de Ven and Tolias [44] categorized incremental learning techniques into three groups: task-incremental learning [45], domain-incremental learning [46], and class-incremental learning [47–50]. In this paper, we only concentrate on the third group, which aims to learn a unified classifier for both the base and novel classes [34].

For instance, Rebuffi et al. [51] proposed a memory-based approach named iCaRL, which memorizes class exemplars to mitigate forgetting and further utilizes the nearest neighbor classification loss and the distillation loss on novel classes and the exemplars of base classes, respectively. The incremental learning method proposed in iCaRL can update the parameters corresponding to the old classes; however, it may overfit the retained old data. To address the problem in iCaRL, Lopez-Paz and Ranzato [48] proposed a model named gradient episodic memory (GEM) to alleviate forgetting. It modifies the gradient update direction of the new task with the help of inequality constraints, which only changes the parameters corresponding to the novel tasks and does not interfere with the other parameters. Moreover, Castro et al. [47] introduced end-to-end incremental learning (EEIL), which introduces cross-entropy loss and distillation loss for incremental learning in an end-to-end framework. Similar to EEIL [47], Wu et al. [52] also utilized the distillation-based method to mitigate the catastrophic forgetting problem from different perspectives. In addition, Han et al. [14] and Cao et al. [15] retained the learned knowledge by storing representative samples, which are then relayed to consolidate the memory and mitigate the forgetting problem in event detection and relation classification tasks, respectively.

With the rapid emergence of new categories and limited available samples, traditional incremental learning methods that rely on representation learning [53] cannot quickly obtain effective features from such few samples. Therefore, special methods are needed to deal with incremental learning with only a small number of samples.

Few-shot incremental learning [29,51,54] was recently proposed with the goal of dealing with incremental learning tasks with limited well-labeled training data [29] by Tao et al. [32].

It can also be regarded as a generalized few-shot learning task that can recognize novel as well as base classes at the same time. Existing few-shot incremental learning methods can mainly be divided into two categories [15], i.e., parameter-based methods [13,29–31] and memory-based methods [32–34].

Parameter-based methods attempt to solve the few-shot incremental learning problem by adjusting the model structure or only optimizing the model parameters. In detail, Gidaris and Komodakis [13] proposed leveraging an attention-based few-shot classification weight generator and redesigning the classifier of a ConvNet model as the cosine similarity function between feature representations and classification weight vectors. Inspired by attractor networks [13], Ren et al. [29] optimized a regularizer to reduce the catastrophic forgetting risk. Moreover, Mazumder et al. [30] retained the ability to classify base classes by identifying and preserving the important model parameters of old data when learning novel data. Similar to the work of Gidaris and Komodakis [13] and Qi et al. [31] described how to make *ConvNet* deal with incremental learning tasks in a low-resource scenario by directly setting the final layer weights from novel training examples during few-shot learning. However, it is difficult to measure the importance of all the parameters of a model, especially for large-scale ones.

Memory-based approaches concentrate on preventing catastrophic forgetting by storing and organizing a few seen instances, which are then utilized with novel samples to jointly train the model. For instance, Tao et al. [32] employed a neural gas network [55] to preserve the topology of the features in both base and novel classes for computer vision tasks. Furthermore, Tao et al. [33] utilized a competitive Hebbian learning (CHL) [56] method to construct and update the elastic Hebbian graph (EHG). Compared with Tao et al. [32], the elastic Hebbian graph has a more flexible strategy for setting the network weights. Inspired by Tao et al. [32] and Dong et al. [34] proposed a novel few-shot incremental learning framework named ERDIL to explore the relation between exemplars in base tasks, which transfers the learned knowledge to a new model for preserving the capability of recognizing seen classes when learning new tasks. In addition, Zhu et al. [35] added Gaussian noise [57] to the prototypes as a means of data enhancement, thereby reducing the risk of overfitting under a few samples. Zhang et al. [36] proposed a continuously evolved classifier employing a graph model [58] to propagate context information between classifiers for adaptation. However, the size of such a graph needs to be preset based on the number of incremental sessions. In conclusion, memory-based approaches have been proven to be more effective than parameter-based approaches for incremental learning tasks since the old knowledge is preserved by storing informative samples of the base classes and reused directly when learning novel data. Table 1 presents some properties of the two methods and their representative works.

Existing memory-based methods usually store the exemplars of the base classes as separate entities regardless of their relations. In addition, they do not take full advantage of the limited samples within the meta tasks in the incremental learning phase. The lack of relation constraints leads to changes in the separability between these entities during the incremental learning phase, which becomes an important cause of catastrophic forgetting. The underutilization of samples in incremental meta tasks causes the model to be overconfident in the base classes, resulting in poor performance in the novel categories. Thus, we propose constructing a geometry structure to organize exemplars of the learned classes into a whole to strengthen the memory for the base knowledge while avoiding catastrophic forgetting. Furthermore, we introduce a multisource contrastive-based loss to exploit the limited samples in each incremental meta task in hopes of improving the generalization ability on novel categories.

3. Approach

In this section, we first describe the task formulation in Section 3.1. Then, we introduce the overall geometry-aware learning framework for few-shot incremental intent recognition in Section 3.2. In particular, we introduce the episodic-based feature extractor and detail the construction procedure of the geometry structure in Section 3.3 and Section 3.4, respectively. Finally, we discuss the learning process of the novel labels in Section 3.5.

3.1. Task formulation

This subsection illustrates the task formulation of intent recognition, from the traditional intent recognition to the fewshot one and finally to the few-shot incremental one.

3.1.1. Intent recognition

For any utterance, intent recognition aims at identifying an intent label from the predefined intent label set. Formally, given a predefined intent label set C and an intent recognition training dataset $\mathcal{D}_{train}$ with N_I utterance-label pairs, $\mathcal{D}_{train} = \{(x_i, y_i), y_i \in C\}_{i=1}^{N_I}$, where x_i and y_i denote the *ith* utterance and corresponding intent label, and the traditional supervised training of an intent classification model $\mathcal{F}$ iteratively processes $\mathcal{F}(x_i) \xrightarrow{\mathcal{D}_{train}} y_i$ to reach training convergence. After such traditional training, the success of $\mathcal{F}$ generalized to the unseen testing dataset $\mathcal{D}_{test}$ is built on the data volume of $\mathcal{D}_{train}$, i.e., the larger $|\mathcal{D}_{train}|$ is, the better the generalization ability of $\mathcal{F}$.

3.1.2. Few-shot intent recognition

Since human annotation cannot catch the ever-emergence of novel intent labels, this data-volume requirement cannot always be satisfied. As a result, the intent label set can be split into two parts: data-rich base labels C_{base} and data-scarce novel labels C_{novel}. Hence, few-shot intent recognition is proposed to free this data-volume constraint and concentrates on predicting C_{novel} *with limited labeled data.*

Traditionally, few-shot intent recognition models extensively employ episodic learning as in [5] on data-rich labels C_{base} to obtain the potential generalization ability that can readily adapt to C_{novel}. In particular, episodic learning extracts a series of meta tasks from C_{base} in the training phase to imitate the low-resource scenario in C_{novel}. Formally, each meta task $\mathcal{T}$ consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The support set $\mathcal{S}$ strictly abides by the "N_s-way K_s-shot" format, which can be formulated as:

$$\mathcal{S} = \bigcup_{n=1}^{N_S} \{(x_{n,i}^s, y_n^s)\}_{i=1}^{K_S}, y_n^s \in C_{base} \tag{1}$$

where K_S is the number of instances of each intent category and N_S denotes the number of categories. The query set does not necessarily follow this format in other few-shot papers. In this paper, we still adopt the same format in the query set to be united with the support set. In addition, the label set in $\mathcal{Q}$ must be the same as that in $\mathcal{S}$ and each instance in $\mathcal{Q}$ does not appear in $\mathcal{S}$, which are as follows:

$$\mathcal{Q} = \bigcup_{n=1}^{N_S} \{(x_{n,i}^q, y_n^q)\}_{i=1}^{K_Q}, y_n^q \in C_{base} \tag{2}$$

where K_Q is the number of queries for each category. In this way, each meta task can be treated as an imitation of few-shot generalization from training to testing. After training a series of meta tasks in C_{base}, the intent recognition model $\mathcal{F}$ can obtain such generalization ability, readily testing on C_{novel}.

3.1.3. Few-shot incremental intent recognition

Despite the vast progress, vanilla few-shot intent recognition has to sacrifice some recognition accuracy on the initial categories C_{base}. In view of this drawback, few-shot incremental recognition is proposed, which endeavors to learn the base categories effectively and recognize novel categories with only limited labeled instances.

Following the meta-task format, the data-scarce label set C_{novel} is further split into C_{novel}^{train} and C_{novel}^{test}, i.e., $C_{novel} = C_{novel}^{train} \cup C_{novel}^{test}$, where the training on C_{novel}^{train} simulates the few-shot incremental scenario, while C_{novel}^{test} is used for testing the few-shot incremental ability. Hence, the training set of few-shot incremental learning D_{train} can be treated as a combination of C_{base} and C_{novel}^{train}. For the testing set $\mathcal{D}_{test}$, its support set $\mathcal{S}$ is extracted from C_{novel}^{test}, while the query set $\mathcal{Q}$ not only contains novel categories in the current meta task but also includes some base categories. Note that there is no instance overlap between $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.

In conclusion, the ultimate goal is to learn an intent-recognition model $\mathcal{F}$ parameterized by the optimized parameters h from a hypothesis space $\mathcal{H}$ [59], which is trained with a sequence of sessions to minimize the cross entropy loss function $\mathcal{L}_c(\cdot,\cdot)$ over all seen categories:

$$h^* = \underset{h\in\mathcal{H}}{argmin} \sum_{(x,y)\in\mathcal{D}_{train}} \mathcal{L}_c(\mathcal{F}_h(x), y) \tag{3}$$

where $\mathcal{F}_h(x)$ is the predicted result by the model $\mathcal{F}$ parameterized by h.

3.2. Overall framework

Traditional memory-based few-shot incremental learning models mainly depend on the stored exemplars of learned classes to mitigate "catastrophic forgetting". However, solely recording instances can trigger the asynchronous update issue, weakening the separability of learned labels. Hence, we argue that these informative exemplars should be stored as a whole. In this way, selected instances with their relationships constitute a geometry structure in the embedding space, which can better reflect the evolution of memories and the embedding space. Furthermore, since preserving base knowledge will weaken the ability to learn novel classes, we also introduce an episodic-based feature extractor as well as a multisource contrastive-based loss to mitigate the overconfidence in seen base classes.

Based on these motivations, we propose a geometry-aware learning model (GAL) for few-shot incremental intent recognition. The overall framework of our proposal is shown in Fig. 1, which has the following key modules.

Episodic-based feature extractor. The pretrained language model BERT [60,61] is employed as the backbone of the feature extractor to project given utterances into an embedding space. Then, a prototypical network utilizes the nearest neighbor theory to recognize the samples belonging to their corresponding labels. Different from the traditional epochwise training, the episodic training strategy is applied to the feature extractor, which splits all data into an "N-way K-shot" format to imitate the low-resource training scenario. The form of our strategy is similar to that of Snell et al. [5], but our target is different from theirs. Their work focuses on obtaining a suitable metric, while we hope to acquire the adaptability for the feature encoder to generalize to true few-shot tasks during the incremental learning phase.

Geometry structure. We employ the pretrained episodic-based feature extractor to select the most informative sample in each base class as its exemplar. Such exemplars and their relative position relationships construct an undirected fully connected geometry structure in the embedding space. With this geometry structure, each step of incremental training can be more

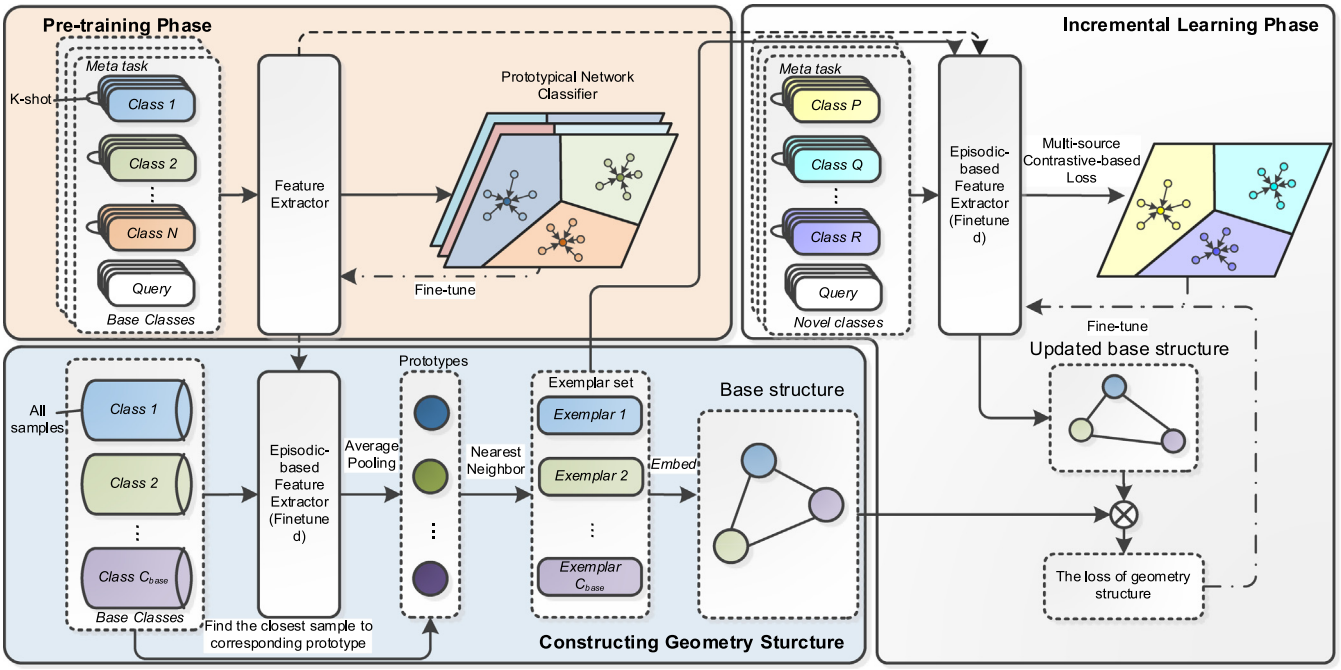

Fig. 1. The geometry-aware learning framework. The orange, blue and gray squares denote processes of the episodic-based feature extractor, the geometry and the multisource contrastive-based training, respectively.

directed without losing the base-class knowledge. To the best of our knowledge, we are the first to propose a base knowledge preservation strategy by building a geometry structure with base exemplars.

Multisource contrastive-based training. Over the geometry structure, we also introduce multisource contrastive-based training to reduce the overconfidence over base labels. Each step increment compares learned exemplars in the geometry structure with newly added novel instances, thereby reducing the confusion between learned and novel classes. It should be pointed out that although the training format is consistent with that of Schroff et al. [62], different data sources are adopted in tackling the problem of the model overconfidence in base classes in the incremental learning process.

Briefly, the whole process can be summarized as follows. First, quantities of meta tasks are constructed from the data of the base classes, each of which is fed into the feature extractor to obtain the low-dimensional continuous vectors of the original text inputs. In addition, vectors in the support set are leveraged for constructing the features of each category in prototype networks and predicting the labels of samples in the query set, which aims to obtain the episodic-based feature extractor. Then, the whole base class dataset is fed into the episodic-based feature extractor to obtain the exemplars of each base class, which are employed for constructing the geometry structure. In the incremental learning phase, the support set is constructed from the novel classes, and the query set is selected from both kinds of classes. The multisource contrastive-based loss is computed for learning novel categories, while the geometry loss is computed for retaining the base knowledge.

3.3. Episodic-based feature extractor

The traditional epochwise training mode treats all labeled data as a training epoch. Such a full-data training mode cannot help the feature extractor obtain the encoder's ability in the low-resource scenario. Different from full data training, episodic-based pretraining is adopted in this paper, which splits the full data into a series of low-data meta tasks to imitate the low-resource training in the pretraining stage, which are described as follows.

Following the format in Eqs. (1) and (2), we can first extract a set of meta tasks from the set C_{base} with put-back sampling, i.e., $S_{meta} = \{\mathcal{T}_i\}_{i=1}^{N_{meta}}$, where N_{meta} is the number of meta tasks. For any meta task $\mathcal{T} \in S_{meta}$ consisting of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, their corresponding instances, i.e., $\cup_{n=1}^{N_S}\{x_{n,i}^s\}_{i=1}^{K_S}$ and $\{x_{n,i}^q\}_{i=1}^{K_Q}$, can then be encoded into low-dimensional embeddings $f\left(x_{n,i}^s\right), f\left(x_{n,i}^q\right)$ by a feature extractor $f(\cdot)$.

Then, for each category c_n, where $n = 1, \ldots, N_S$, in $\mathcal{S}$, we perform an average pooling operation over the embedding vectors of instances belonging to c_n to acquire its "prototype" v_{p_n}:

$$v_{p_n} = \frac{1}{K_S}\sum_{i=1}^{K_S} f\left(x_{n,i}^s\right), \tag{4}$$

where $f\left(x_{n,i}^s\right)$ is the embedding of kth support instance $x_{n,i}^s$ in category c_n. Then, we utilize the nearest neighbor method to classify the instances in the query set. In particular, for each instance in the query set, we adopt the cosine similarity to compute the relation between the query instance $x_{n,i}^q$ with a specific prototype v_{p_n}, i.e.,

$$s\left(x_{n,i}^q, v_{p_n}\right) = \frac{f\left(x_{n,i}^q\right)^\top v_{p_n}}{\left\|f\left(x_{n,i}^q\right)\right\|_2 \left\|v_{p_n}\right\|_2}. \tag{5}$$

Repeating Eq. (5) over all prototype representations, we can obtain the similarity sequence $\{s\left(x_{n,i}^q, v_{p_1}\right), s\left(x_{n,i}^q, v_{p_2}\right), \ldots, s\left(x_{n,i}^q, v_{p_{N_S}}\right)\}$, from which the prototype with the greatest similarity is the prediction label for the instance $x_{n,i}^q$, i.e.,

$$\overline{y_{n,i}^q} = \underset{j=1,\ldots,N_S}{\operatorname{argmax}}\, s\left(x_{n,i}^q, v_{p_j}\right) \tag{6}$$

where $\overline{y_{n,i}^q}$ is the predicted label of input instance $x_{n,i}^q$.

Finally, we employ the cross entropy to compute the loss, i.e.,

$$\mathcal{L}_{pre} = \sum_{n=1}^{N_S}\sum_{i=1}^{K_Q} -\log\frac{\exp\left(s\left(x_{n,i}^q, v_{p_n}\right)\right)}{\sum_{j=1,\ldots,N_k}\exp\left(s\left(x_{n,i}^q, v_{p_j}\right)\right)}. \tag{7}$$

In detail, if Eq. (7) is minimized, the value of the numerator is maximized, which means that the similarity between queries and prototypes of the same class will be maximized.

After being trained with quantities of meta tasks in S_{meta}, the initial feature extractor is optimized as an episodic-based feature extractor $f_0(\cdot)$ that is more compatible with the few-shot scenario. In the subsequent process, $f_0(\cdot)$ will be utilized to build the geometry structure of learned classes.

3.4. Geometry structure

To mitigate the risk of catastrophic forgetting, we select the informative instances of learned classes and construct a geometry structure with them. Specifically, we first feed all the instances belonging to $C_{base} = \{L_i\}_{i=1}^{N_{base}}$ into the episodic-based feature extractor $f_0(\cdot)$. Following Eq. (4), we can obtain the episodic-based prototypes for all categories in C_{base}, i.e., p_{L_i}, where $i = 1, \ldots, N_{base}$.

Next, we still adopt the nearest neighbor principle to select the most informative instance for each category, called the exemplar. We argue that the most informative sample for the specific category is the one closest to its prototype. The process can be formulated as follows:

$$x_{L_i} = \underset{x_i \in X_i}{argmax}\, s\left(x_i, p_{L_i}\right) \tag{8}$$

where x_{L_i} is the selected exemplar and X_i is the instance set whose category is L_i.

With this extraction, we employ memories $\mathcal{E}_0$ to store the exemplars of all categories at the 0*th* step increment, i.e., $\mathcal{E}_0 = \left[x_{L_1}, x_{L_2}, \ldots, x_{L_{N_{base}}}\right]^\top$. For these stored exemplars, their relationships can better reflect the learned knowledge over the base classes. In this light, we employ a matrix $\mathcal{R}_0$ to store their relations that adopt a normalized inner product over any two exemplars, i.e.,

$$\mathcal{R}_0 = g\left(\{r_{i,j}\}\right) = g\left(\{\frac{f_0\left(x_{L_i}\right)^\top f_0\left(x_{L_j}\right)}{\left\|f_0\left(x_{L_i}\right)\right\|_2 \left\|f_0\left(x_{L_j}\right)\right\|_2}\}\right), i, j = 1, \ldots, N_{base} \tag{9}$$

where $g(\cdot)$ denotes the operation normalizing a matrix by row and $\mathcal{R}_0 \in R^{N_{base} \times N_{base}}$.

With the selected instances $\mathcal{E}_0$ and their relationships $\mathcal{R}_0$, we can obtain the 0*th* step geometry structure, i.e., $G_0 = \{\mathcal{E}_0, \mathcal{R}_0\}$. To guarantee that the learned categories can be retained to the greatest possible extent, we store the $(t-1)$ *th* geometry matrix G_{t-1} as a constraint to the t_{th} step incremental training.

$$\mathcal{L}_{geo}^t\left(G_t, G_{t-1}\right) = \left\|\left[\mathcal{R}_t\right]_{a_{t-1} \times a_{t-1}} - \mathcal{R}_{t-1}\right\|_F, \tag{10}$$

where a_{t-1} is the row/column number of the matrix $\mathcal{R}_{t-1}$, $\|\cdot\|_F$ is the Frobenius norm and $[\cdot]_{a \times a}$ is a submatrix operation that returns an $a \times a$ submatrix.

3.5. Multisource contrastive-based training

Traditionally, given an incremental meta task $\mathcal{T}_{inc} = \{\mathcal{S}_{inc}, \mathcal{Q}_{inc}\}$, the support set $\mathcal{S}_{inc}$ still follows the "N_S-way K_S-shot" format, i.e.,

$$\mathcal{S}_{inc} = \bigcup_{n=1}^{N_S} \{\left(x_{n,i}^{s,inc}, y_n^{s,inc}\right)\}_{i=1}^{K_S}, y_n^{s,inc} \in C_{novel}^{train} \tag{11}$$

while the query set $\mathcal{Q}_{inc}$ is different from the setting in Eq. (2), including both the incremental novel classes for training and learned base classes, i.e.,

$$\mathcal{Q}_{inc} = \bigcup_{n=1}^{N_I} \{\left(x_{n,i}^{q,inc}, y_n^{q,inc}\right)\}_{i=1}^{K_Q}, y_n^{q,inc} \in C_{base} \cup C_{novel}^{train}, \tag{12}$$

where K_Q is the number of queries for each category, and hence $N_s < N_I \leq N_{base} + N_s$. For this incremental meta task, traditional training still adopts the cross entropy in Eqs. (5) and (7), denoted as $\mathcal{L}_{inc}$.

However, as described in Section 3.4, the geometry structure imposes a strong training constraint on learned categories instead of reducing the distinguishability of the novel categories. Hence, we introduce a multisource contrastive-based training that enhances the supervision signal in the classification of limited novel categories to balance the classification performance between base classes and novel classes. Specifically, for each incremental meta task, we can still compute the prototypes for all novel classes based on $\mathcal{S}_{inc}$. With the introduced base classes in $\mathcal{Q}_{inc}$, we can finally obtain the testing prototypes in the meta task $\mathcal{T}_{inc}$, i.e., $v_{p_n}^{inc}$, where $n = 1, \ldots, N_I$. Then, for each query instance $x_{n,i}^{q,inc}$, the prototype of its corresponding ground truth $v_{p_n}^{inc}$ can be treated as the positive instance. The negative instances come from two resources, including the learned base classes and the remaining novel classes in the current meta task, i.e., $v_{p_j}^{inc}$, where $j = 1, \ldots, N_I$ and $j \neq n$. Therefore, this multisource contrastive-based loss can be defined as:

$$\mathcal{L}_{tri} = \sum_{n=1}^{N_I} \sum_{i=1}^{K_Q} \sum_{j=1,\ldots,N_I, j \neq n} \max\left(0, \delta - s\left(x_{n,i}^{q,inc}, v_{p_n}^{inc}\right) + s\left(x_{n,i}^{q,inc}, v_{p_j}^{inc}\right)\right), \tag{13}$$

where δ is a hyperparameter meaning the margin between positive and negative samples. With this multisource contrastive-based loss, the few-shot incremental model can better distinguish the learned base classes and newly added novel classes.

Finally, the final objective function for each increment is formulated as:

$$\mathcal{L}\left(\mathcal{S}_{inc}, \mathcal{Q}_{inc}\right) = \mathcal{L}_{inc} + \lambda \cdot \mathcal{L}_{geo} + (1 - \lambda) \cdot \mathcal{L}_{tri}, \tag{14}$$

where λ is a trade-off coefficient to balance the contributions of geometry loss $\mathcal{L}_{geo}$ and multisource contrastive-based loss $\mathcal{L}_{tri}$, which helps the model regulate contributions between the base and novel knowledge. In detail, $\mathcal{L}_{geo}$ is used for retaining the base knowledge obtained in the pretraining phase, and $\mathcal{L}_{tri}$ is utilized for optimizing the model during the incremental learning phase.

For clarification, we summarize the training process of the GAL model in Algorithm 1. Specifically, we leverage the base categories to pretrain an episodic-based feature extractor in Lines 2 to 12. Then, we select the base exemplars from the base data in Line 13 and obtain the geometry structure in Line 14. Finally, the incremental learning process is performed in Lines 15 to 25.

4. Experiments

In this section, we introduce the used dataset and the evaluation metrics of the experiments in Section 4.1 and the discussed baseline models in Section 4.2. In addition, the research questions and model configuration are elaborated in Sections 4.3 and 4.4, respectively.

4.1. Datasets and evaluation metrics

We utilize the OOS (CLINC-150) dataset [17] to evaluate the capability of our GAL model and baselines. CLINC-150 contains 22,500 labeled utterances covering 150 intent categories, which are divided into 10 general domains. In addition, there are also some utterances labeled "out of scope" in the dataset, which is a noisy label with multiple ambiguous categories. To facilitate evaluating the performance of the models, the samples labeled "out of scope" are not used for training or testing.

In detail, we first divide the dataset into a training set and a test set by category, which is separated by a proportion of 120:30. The training set is further divided into two parts that contain 90 classes and 30 classes, where the former part is the base-class set C_{base} and the latter one is the pseudonovel-class set C_{novel}^{train} to simulate the incremental learning scenario. Furthermore, inside each base class, the 150 instances are divided into three parts, containing 100, 20, and 30 utterances, which are used for training, validation, and testing, respectively. For the evaluation metrics, following Ren et al. [16], we measure the model performance by comparing the classification accuracy over both the base and novel classes.

Algorithm 1 Geometry-aware learning

Input: the class sets C_{base} and C_{novel}^{train}; the max iteration step of pretraining $iter_{pre}$; the max iteration step of incremental learning $iter_{inc}$; the initial feature extractor $f(\cdot)$ and intent recognition model $\mathcal{F}$

Output: Optimal intent recognition model $\mathcal{F}^*$.

$i = 0,\ t = 0$;
while $i < iter_{pre}$ **do**
  Sample a set of meta tasks S_{meta} from C_{base}
  for all $\mathcal{T}_i \in S_{meta}$ **do**
    Encode instances in $\mathcal{T}_i$ with $f(\cdot)$;
    Obtain the prototype v_{p_n} based on Eq. (4)
    Compute $\mathcal{L}_{pre}$ based on Eq. (7)
  end for
  Compute $\nabla_f \sum_{\mathcal{T}_{pre}} l(\hat{y}, y)$ and update f with ∇_f
  $i = i + 1$;
end while
Obtain the pretrained feature extractor $f_0(\cdot)$.
Select exemplars based on Eq. (8).
Construct the geometry structure G_t based on Eq. (9).
while $t < iter_{inc}$ **do**
  Sample a set of incremental meta tasks S_{inc};
  for all $\mathcal{T}_{inc} \in S_{inc}$ **do**
    Obtain the prototype $v_{p_n}^{inc}$based on Eq. (4);
    Compute $\mathcal{L}_{geo}^t$ based on Eq. (10);
    Compute $\mathcal{L}_{tri}^t$ based on Eq. (13);
    Obtain the final loss $\mathcal{L}(\mathcal{S}_{inc}, \mathcal{Q}_{inc})$ based on Eq. (14).
  end for
  Compute $\nabla_{\mathcal{F}} \sum_{\mathcal{T}_{inc}} \mathcal{L}(\mathcal{S}_{inc}, \mathcal{Q}_{inc})$ and update $\mathcal{F}$ with $\nabla_{\mathcal{F}}$;
  $t = t + 1$;
end while
return Optimal intent recognition model$\mathcal{F}^*$

4.2. Model summary

We validate the effectiveness of our GAL model by comparing it with the following five competitive baselines:

- **ProtoAug,** i.e., prototypical augmentation [35], is a solution to the incremental learning (IL) problem. In the deep feature space, a prototype is saved for each corresponding category in the old task. In the training phase of the new task, the prototype is augmented by Gaussian noise and participates in the classification of the new task samples to reduce the speed of information forgetting in the old task. The existence of prototypes also makes it a nonexemplar method, avoiding the need to store all the old task training sets.

- **MatchNet**, i.e., matching networks [63], concentrates on calculating the point-to-point similarity. It constructs a framework for a few-shot classification task, which maps a labeled small support set and unlabeled instances to their labels and obviates the demand for fine tuning to adapt to novel categories.

- **Siamese** [64] focuses on measuring the similarity between two inputs. The twin neural network has two inputs, which are fed into two neural networks and then mapped to the same new space to form new representations. The similarity between the two inputs is evaluated through the calculation of the loss function.

- **Imprint** [31] learns base classes through supervised pretraining and represents novel classes by prototypical averaging directly. It utilizes a fully connected layer to classify the samples by concatenating the base and novel embeddings.

- **IncreProtoNet** [16] is a two-phase prototypical network with prototype attention alignment and triplet loss. The first phase pretrains the model and selects the informative samples, and the second phase merges the base and novel prototypes with an attention mechanism.

- **CEC** [36] is a two-phase few-shot learning framework that employs a graph model to propagate context information between classifiers for adaptation. It adopts a simple but effective decoupled learning strategy of representations and classifiers in which only the classifiers are updated in each incremental session, avoiding knowledge forgetting in the representations.

4.3. Research questions

We examine the effectiveness of our proposals and concentrate on the following research questions to guide our experiments:

RQ1 Does GAL improve the overall performance compared to competitive baselines for few-shot incremental intent recognition?

RQ2 How does our model perform when the training set and test set are from different domains?

RQ3 Which module plays the largest role in few-shot incremental intent recognition?

RQ4 How does GAL perform with different lengths of utterances?

4.4. Model configurations

Following the common practice of few-shot incremental learning experiments [16,31], we discuss two types of meta tasks with different numbers of "ways" and "shots", including "5-way 1-shot" and "5-way 5-shot". For all discussed models, we employ the same feature extractor (i.e., the BERT-base-uncased encoder [60,65]) and the same function (i.e., the cosine similarity) to measure the distance to guarantee a fair comparison of the performance of the baselines and our proposal.

Moreover, we apply an early-stop mechanism in the pretraining as well as the incremental training phase, i.e., the model optimization stops when no loss decay is returned. In addition, the feature extractor of our proposal is first pretrained with 2000 iterations, and all models, including baselines, are trained with 10,000 iterations, each of which includes 4 episodes. We evaluate the models per 1000 iterations with 500 meta tasks extracted from the validation set to test the performance. Furthermore, the hyperparameters δ and λ are searched in {0.3, 0.4, 0.5} and {0.5, 0.6, 0.7, 0.8}, respectively, and are finally fixed to 0.5 and 0.7. The experiments were performed on a computer with an Intel i9-11900K CPU, 64 GB RAM, and a single NVIDIA RTX 3090 GPU. The deep learning framework is PyTorch.

Table 2
Overall performance in terms of accuracy (%) as well as the 95% confidence interval on the CLINC-150 dataset for two types of meta tasks. The results produced by the best performer and baseline in each column are boldfaced and underlined, respectively.

Models	5-way 1-shot			5-way 5-shot		
	Base	Novel	Both	Base	Novel	Both
MatchNet	89.36 ± 0.13	3.44 ± 0.43	77.09 ± 0.29	90.37 ± 0.11	49.78 ± 0.32	84.57 ± 0.19
Siamese	89.24 ± 0.17	61.78 ± 0.18	85.32 ± 0.17	92.39 ± 0.14	82.52 ± 0.20	90.98 ± 0.16
Inprint	77.38 ± 0.34	58.75 ± 0.46	74.72 ± 0.37	77.23 ± 0.29	60.29 ± 0.31	74.81 ± 0.29
IncreProtoNet	90.69 ± 0.25	78.55 ± 0.27	88.96 ± 0.25	93.00 ± 0.19	86.48 ± 0.09	92.07 ± 0.16
ProtoAug	37.60 ± 0.60	41.32 ± 0.54	38.13 ± 0.59	38.36 ± 0.46	48.08 ± 0.46	39.75 ± 0.46
CEC	93.76 ± 0.35	77.61 ± 0.43	91.46 ± 0.37	94.45 ± 0.29	85.07 ± 0.21	93.11 ± 0.27
GAL (Ours)	**95.08 ± 0.17**	**78.94 ± 0.10**	**92.71 ± 0.15**	**96.34 ± 0.10**	**87.08 ± 0.14**	**95.02 ± 0.11**

5. Results and discussion

In this section, we discuss the overall performance of the models in Section 5.1 and analyze the performance of all discussed models in the cross-domain setting in Section 5.2. Furthermore, the importance of different modules and the impact of utterance length on the model are analyzed in Section 5.3 and Section 5.4, respectively.

5.1. Overall evaluation

To answer **RQ1**, we evaluate the few-shot incremental intent recognition performance of our proposal as well as five baselines for two types of meta tasks. The overall performance of the discussed models is shown in Table 2. Generally, as the shot number increases from 1 to 5, the performance of all models also improves. This phenomenon can be explained by the fact that relatively more training labeled data can reduce the overfitting risk and increase the learning capability for the few-shot models, which is especially evident for MatchNet.

Let us first consider the baselines. From Table 2, we can observe that MatchNet exhibits higher improvements than the Imprint baseline in terms of accuracy on base classes of both types of meta tasks. The advantages of MatchNet indicate that the ad hoc similarity matching computation can well preserve the learned knowledge, which can help to boost the performance on base categories. Among the baselines, the best performers in the base classes and novel classes are IncreProtoNet and CEC, respectively. The advantages of IncreProtoNet and CEC can be explained by the fact that the pretraining operation can better memorize the knowledge of base classes. Hence, we only consider IncreProtoNet and CEC for comparisons in later experimental analyses.

Next, we compare the baselines against the GAL model. Obviously, GAL is the best performer among all discussed models in all cases. Compared with the baseline IncreProtoNet, GAL achieves 0.39% and 0.60% improvements in terms of accuracy on the novel classes in the "5-way 1-shot" meta task and "5-way 5-shot" meta task. Compared with the baseline CEC, GAL achieves 1.32% and 1.89% improvements in terms of accuracy on the base classes in the two types of meta tasks. Such leading performance proves the effectiveness of our proposed model. In particular, the recognition accuracy of GAL for novel categories is obviously higher than that of the traditional few-shot learning models, i.e., MatchNet and Siamese. The advantage shows that the GAL model can effectively learn novel categories and has a stronger learning ability for them than the naive few-shot learning baselines. In addition, our proposed GAL model is also more accurate than the incremental learning models, i.e., Imprint, IncreProtoNet, ProtoAug, and CEC, in recognizing base classes, verifying that GAL has a stronger memory of old knowledge than other discussed incremental learning baselines. Furthermore, compared to other baselines, GAL has the smallest confidence interval on all types of classes and meta tasks, which indicates that GAL not only has the highest recognition accuracy but the smallest variance in all discussed settings.

5.2. Cross-domain performance

To answer **RQ2**, apart from our original CLINC-150 dataset, we adopt an extra intent recognition dataset, i.e., ATIS [66], from the airline travel domain, which consists of 18 types of intents. Furthermore, since the number of samples of some categories in ATIS is too low (less than 15) to construct the training and testing meta tasks, we delete such 6 categories and only utilize the remaining 12. To examine the general domain adaptation ability that is trained on the broad domain and tested on both broad and specific professional domains, we consider the cross-domain experiment between CLINC-150 and ATIS. In detail, the categories in CLINC-150 and ATIS are regarded as base and novel categories, respectively. Different from the cross-domain experiments in vanilla few-shot learning that only evaluate the performance on the target domain, our incremental experiment examines the model performance on both source and target domains.

Similar to the results in Section 5.1, all models consistently reach higher recognition accuracy for the "5-way 5-shot" meta task than for the "5-way 1-shot" meta task, especially on the novel categories. The phenomenon illustrates that the increase in labeled data can also assist in boosting the model performance in the few-shot incremental cross-domain scenario. This indicates that the learned meta-knowledge from a broad daily life domain cannot be totally adapted to another professional domain. According to Fig. 2, GAL outperforms all the discussed baselines on both types of meta tasks. Although it also shows a performance decrease in the novel categories, it still remains the most accurate among all the discussed models in cross-domain experiments on both types of meta tasks. Specifically, it outperforms the best baseline IncreProtoNet by 3.63% on the "5-way 1-shot" meta task. Furthermore, there are some notable phenomena in Fig. 2. The performance of some models in the base categories is obviously degraded. For instance, in the "5-way 5-shot" meta task, the accuracy of MatchNet on recognizing base classes can reach nearly 80%, while that on novel classes is only less than 40%. This could be explained by the fact that after learning base knowledge from the source domain, the generalization ability of MatchNet in the target domain is limited. On the other hand, the accuracy of ProtoAug in recognizing novel classes is higher than that on base classes. This is because models focusing on learning novel categories suffer from a problem of catastrophic forgetting on the base classes in the cross-domain scenario.

In addition, following Alvi et al. [10], we further evaluate the classification performance of the models in cross-domain experiments, which is illustrated in Table 4. In this evaluation, we treat each meta task used for testing as a whole to compute the relevant metrics, including specificity, precision, and AUC (area under the curve). It is obvious that our proposal outperforms all the discussed baselines in all three metrics involved. Specifically, the advantages of our model are mainly reflected in the precision. It outperforms the best baseline by 2.97% on the "5-way 1-shot" meta task and 0.76% on the "5-way 5-shot" meta task, indicating

Table 3
Results in terms of the accuracy of GAL without different modules for the "5-way 1-shot" and "5-way 5-shot" meta tasks. Models with the worst drop in each column are appended with ▼.

Model	5-way 1-shot			5-way 5-shot		
	Base	Novel	Both	Base	Novel	Both
GAL	89.58 ± 0.36▼	80.06 ± 0.31	88.22 ± 0.15▼	90.17 ± 0.10▼	88.12 ± 0.19	89.88 ± 0.12▼
w/o geometry structure	(−5.50%)	(+0.12%)	(−3.95%)	(−6.17%)	(+1.04%)	(−5.14%)
GAL	96.01 ± 0.19	72.18 ± 0.26▼	92.60 ± 0.22	96.06 ± 0.61	83.08 ± 0.44▼	94.20 ± 0.57
w/o multisource contrastive-based loss	(+0.93%)	(−6.76%)	(0.43%)	(−0.28%)	(−4.00%)	(−0.82%)
GAL	90.63 ± 0.25	75.41 ± 0.31	88.46 ± 0.27	91.11 ± 0.57	84.88 ± 0.39	90.22 ± 0.46
w/o both components	(−4.45%)	(−3.53%)	(−3.71%)	(−5.23%)	(−2.20%)	(−4.47%)
GAL	95.08 ± 0.17	78.94 ± 0.10	92.71 ± 0.15	96.34 ± 0.10	87.08 ± 0.14	95.02 ± 0.11

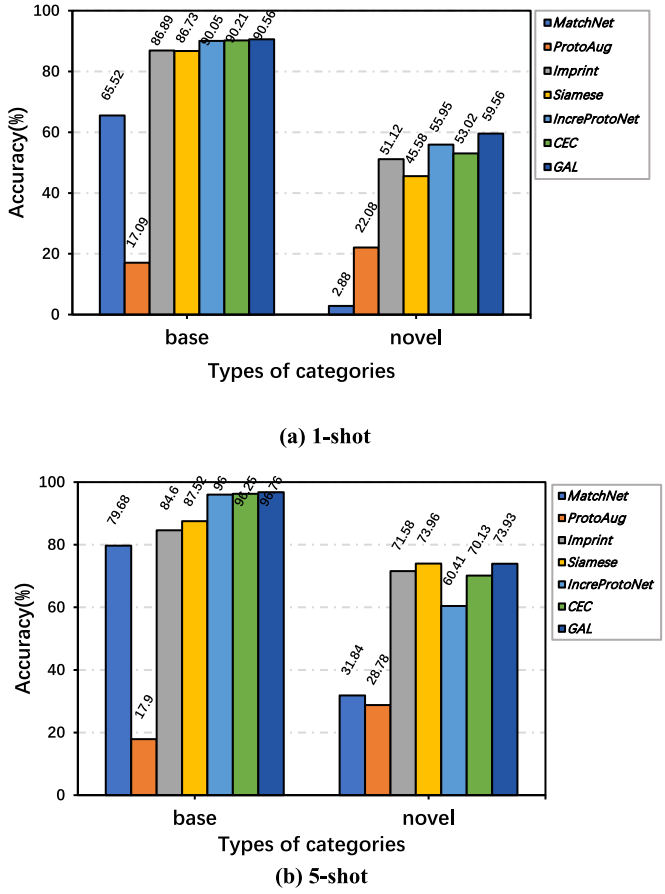

Fig. 2. The performance degradation of the discussed models under the "5-way 1-shot" and "5-way 5-shot" settings.

that our model has smaller performance fluctuations in both types of meta tasks. Furthermore, advantages in the precision indicate that our model has a stronger generalization ability in cross-domain settings than other discussed baselines.

5.3. Ablation study

To analyze the importance of each module in the GAL model, we examine the model performance on both base and novel classes for two types of meta tasks after removing two fundamental components of GAL separately, i.e., the geometry structure and the multisource contrastive-based loss for learning novel categories. Furthermore, we removed these two modules together to verify the performance of the model when they were present together.

Table 4
Classification performance in cross-domain experiments between CLINC-150 and ATIS for two types of meta tasks. The results produced by the best performer and baseline in each column are boldfaced and underlined, respectively.

Models	5-way 1-shot			5-way 5-shot		
	Specificity	Precision	AUC	Specificity	Precision	AUC
MatchNet	98.72	52.80	96.99	99.14	71.17	99.08
Siamese	99.46	83.45	<u>98.75</u>	99.53	<u>93.78</u>	<u>99.54</u>
Inprint	97.81	82.36	90.21	97.93	85.41	93.43
IncreProtoNet	<u>99.48</u>	88.29	93.96	<u>99.78</u>	90.69	94.18
ProtoAug	97.17	16.68	83.20	97.62	17.38	84.96
CEC	99.18	<u>89.51</u>	96.34	99.25	91.37	98.49
GAL (Ours)	**99.75**	**92.48**	**99.24**	**99.81**	**94.54**	**99.77**

The results in Table 3 show that removing any component can cause a severe drop in performance on one type of category and only a small improvement in performance on the other one, which demonstrates that the geometry structure and multisource contrastive-based loss are both important in improving the few-shot incremental intent recognition performance of GAL. The roles of various modules are reflected in different aspects of model performance. In detail, we find that the geometry structure module plays the main role in improving the recognition accuracy of GAL on base classes, since the model performance on base classes shows an obvious decrease after removing it regardless of the types of meta tasks. For example, the performance of the model without the geometry structure shows 5.50% and 6.17% drops in the "5-way 1-shot" and "5-way 5-shot" meta tasks, respectively. Moreover, it is worth mentioning that the performance drop of GAL in the "5-way 5-shot" meta task is more serious than that in the "5-way 1-shot" meta task. This phenomenon illustrates that the more learnable instances of novel categories are introduced, the greater the risk of catastrophic forgetting. Different from the geometry structure, the multisource contrastive-based loss has an essential effect on learning novel categories. Specifically, the performance of the variant without multisource contrastive-based loss shows obvious drops in the novel categories on both types of meta tasks. Moreover, the decrease in the "5-way 1-shot" meta task is worse than that in the "5-way 5-shot" meta task. The results can be explained by the fact that the fewer available samples are introduced, the fewer features of the current class are captured; thus, the multisource contrastive-based loss is more needed to help boost the performance of learning novel categories.

On the whole, the relationship between the geometry structure and the multisource contrastive-based loss is similar to a game, since removing one of them will make the effect of the other one more prominent. In detail, GAL without multisource contrastive-based loss obtains a 0.93% performance boost on base categories in the "5-way 1-shot" meta task, while GAL without geometry structure gains 1.12% and 1.04% performance improvements on novel categories in both types of meta tasks.

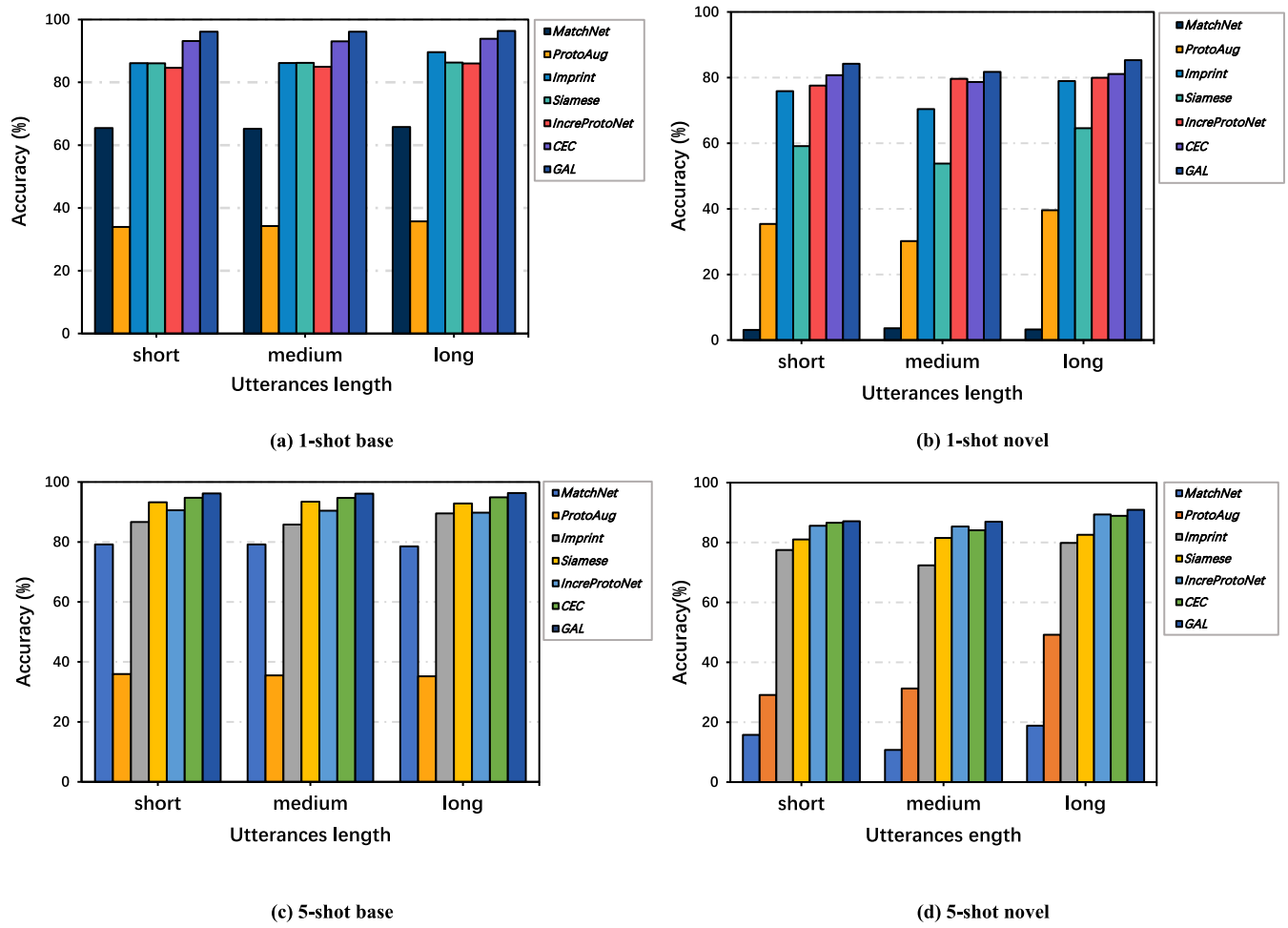

Fig. 3. Accuracy(%) of all discussed models with different length of utterances on the "5-way 1-shot" and "5-way 5-shot" meta tasks.

5.4. Impact of the utterance length

To understand the scalability of our GAL model in applying to utterances with different lengths (**RQ4**), we divide the utterances in the test set of CLINC-150 into short (no more than 5 words), medium (6 to 15 words) and long utterances (more than 15 words) and then report the separate results in Fig. 3.

As shown in Fig. 3a and c, the utterance length has no obvious impact on recognizing base classes, and the performance of the same model under different lengths of utterances fluctuates within a small range. Furthermore, it is obvious that the Siamese model has a performance drop second only to MatchNet on the 1-shot novel categories. In contrast to the other discussed models, ProtoAug is the only model that outperforms in the base categories vs. the novel categories regardless of the types of meta tasks. The phenomenon shows that it has the greatest risk of catastrophic forgetting.

From Fig. 3, we can see that GAL always achieves the highest accuracy on both types of categories regardless of the types of meta tasks. Moreover, GAL can achieve the best accuracy on long utterances, but its advantage over other models is more reflected in medium-length utterances. Following GAL, the second-best performers on the base classes and novel classes are CEC and IncreProtoNet, respectively. The advantages of these two baselines and our model demonstrate that the two-stage training strategy outperforms the single-stage training adopted by the other discussed baselines. Furthermore, among all meta tasks (base or novel classes), the largest performance gap between CEC and GAL appears in the base categories in the "5-way 1-shot" meta task, which illustrates that compared with CEC, the GAL model can acquire the features of categories more rapidly in the low-data regime. For the novel categories, a similar trend is shown in Fig. 3b and d. In particular, almost all models achieve the worst performance on medium-length novel categories, which may be due to the high diversity caused by a large number of medium-length samples in the test set, since it is difficult for the model to fully capture the characteristics of each category.

6. Conclusion and future work

In this paper, we introduce a geometry-aware learning (GAL) model for the few-shot incremental intent recognition task. Based on metric learning theory, GAL introduces a geometry structure to model the relationships between exemplars of base categories in the embedding space and applies a multisource contrastive-based loss to help it remember the base knowledge and adapt to the low-resource scenario. Therefore, GAL can mitigate the risk of catastrophic forgetting as well as overfitting. Experimental results conducted on the CLINC-150 dataset show the effectiveness of GAL against all discussed baselines. In addition, extensive cross-domain experiments on the ATIS dataset demonstrate that the generalization ability of GAL is also better than that of all baselines. However, although our approach achieves the best performance compared with the mentioned baselines, it still deals with a single incremental session with dynamically changing categories. For future work, we would like to explore multisession

few-shot incremental learning, in which the novel classes in the current session become the base classes for the next session. It is a challenging task, and current methods still cannot obtain satisfactory results because of the contradiction between learning novel categories and memorizing learned categories.

CRediT authorship contribution statement

Xin Zhang: Conceptualization, Methodology, Formal analysis, Validation, Writing – original draft. **Miao Jiang:** Methodology, Validation, Data curation. **Honghui Chen:** Investigation and Resources. **Jianming Zheng:** Conceptualization, Formal analysis. **Zhiqiang Pan:** Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Jolly, T. Falke, C. Tirkaz, D. Sorokin, Data-efficient paraphrase generation to bootstrap intent classification and slot labeling for new features in task-oriented dialog systems, in: COLING (Industry), International Committee on Computational Linguistics, 2020, pp. 10–20, http://dx.doi.org/10.18653/v1/2020.coling-industry.2.

[2] T. Shao, F. Cai, W. Chen, H. Chen, Self-supervised clarification question generation for ambiguous multi-turn conversation, Inf. Sci. 587 (2022) 626–641, http://dx.doi.org/10.1016/j.ins.2021.12.040.

[3] S. Vargas, P. Castells, D. Vallet, Intent-oriented diversity in recommender systems, in: SIGIR, 2011, pp. 1211–1212, http://dx.doi.org/10.1145/2009916.2010124.

[4] Y. Lin, F. Lin, W. Zeng, J. Xiahou, L. Li, P. Wu, Y. Liu, C. Miao, Hierarchical reinforcement learning with dynamic recurrent mechanism for course recommendation, Knowl. Based Syst. 244 (2022) 108546, http://dx.doi.org/10.1016/j.knosys.2022.108546.

[5] J. Snell, K. Swersky, R.S. Zemel, Prototypical networks for few-shot learning, in: NIPS, 2017, pp. 4077–4087.

[6] Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM Comput. Surv. (Csur) 53 (2020) 1–34, http://dx.doi.org/10.1145/3386252.

[7] Y. Xie, H. Xu, J. Li, C. Yang, K. Gao, Heterogeneous graph neural networks for noisy few-shot relation classification, Knowl. Based Syst. 194 (2020) 105548, http://dx.doi.org/10.1016/j.knosys.2020.105548.

[8] J. Zheng, F. Cai, W. Chen, et al., Taxonomy-aware learning for few-shot event detection, in: WWW, ACM / IW3C2, 2021, pp. 3546–3557.

[9] C. Song, F. Cai, J. Zheng, et al., Metric sentiment learning for label representation, in: CIKM, ACM, 2021, pp. 1703–1712.

[10] A.M. Alvi, S. Siuly, H. Wang, Developing a deep learning based approach for anomalies detection from EEG data, in: WISE (1), Springer, 2021, pp. 591–602, http://dx.doi.org/10.1007/978-3-030-90888-1_45.

[11] A.M. Alvi, S. Siuly, H. Wang, Neurological abnormality detection from electroencephalography data: a review, Artif. Intell. Rev. 55 (2022) 2275–2312, http://dx.doi.org/10.1007/s10462-021-10062-8.

[12] R.M. French, Catastrophic forgetting in connectionist networks, Trends Cogn. Sci. 3 (1999) 128–135, http://dx.doi.org/10.1162/089976660260028700.

[13] S. Gidaris, N. Komodakis, Dynamic few-shot visual learning without forgetting, in: CVPR, IEEE, 2018, pp. 4367–4375, http://dx.doi.org/10.1109/CVPR.2018.00459.

[14] X. Han, Y. Dai, T. Gao, Y. Lin, Z. Liu, P. Li, M. Sun, J. Zhou, Continual relation learning via episodic memory activation and reconsolidation, in: ACL, Association for Computational Linguistics, 2020, pp. 6429–6440, http://dx.doi.org/10.18653/v1/2020.acl-main.573.

[15] P. Cao, Y. Chen, J. Zhao, T. Wang, Incremental event detection via knowledge consolidation networks, in: EMNLP (1), Association for Computational Linguistics, 2020, pp. 707–717, http://dx.doi.org/10.18653/v1/2020.emnlp-main.52.

[16] H. Ren, Y. Cai, X. Chen, G. Wang, Q. Li, A two-phase prototypical network model for incremental few-shot relation classification, in: COLING, International Committee on Computational Linguistics, 2020, pp. 1618–1629, http://dx.doi.org/10.18653/v1/2020.coling-main.142.

[17] P.R. Cavalin, V.H.A. Ribeiro, A.P. Appel, C.S. Pinhanez, Improving out-of-scope detection in intent classification by using embeddings of the word graph space of the classes, in: EMNLP (1), 2020, pp. 3952–3961, http://dx.doi.org/10.18653/v1/2020.emnlp-main.324.

[18] G. Tur, R. De Mori, Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, John Wiley & Sons, 2011.

[19] Z. Ding, Z. Yang, H. Lin, J. Wang, Focus on interaction: A novel dynamic graph model for joint multiple intent detection and slot filling, in: IJCAI, ijcai.org, 2021, pp. 3801–3807, http://dx.doi.org/10.24963/ijcai.2021/523.

[20] W.A. Abro, G. Qi, Z. Ali, Y. Feng, M. Aamir, Multi-turn intent determination and slot filling with neural networks and regular expressions, Knowl. Based Syst. 208 (2020) 106428, http://dx.doi.org/10.1016/j.knosys.2020.106428.

[21] H. Weld, X. Huang, S. Long, J. Poon, S.C. Han, A survey of joint intent detection and slot-filling models in natural language understanding, 2021, CoRR abs/2101.08091.

[22] M. Firdaus, A. Kumar, A. Ekbal, P. Bhattacharyya, A multi-task hierarchical approach for intent detection and slot filling, Knowl. Based Syst. (2019) 183, http://dx.doi.org/10.1016/j.knosys.2019.07.017.

[23] R. Sarikaya, G.E. Hinton, B. Ramabhadran, Deep belief nets for natural language call-routing, in: ICASSP, IEEE, 2011, pp. 5680–5683, http://dx.doi.org/10.1109/ICASSP.2011.5947649.

[24] G. Mesnil, X. He, L. Deng, Y. Bengio, Investigation of recurrentneural-network architectures and learning methods for spoken language understanding, in: INTERSPEECH, ISCA, 2013, pp. 3771–3775.

[25] B. Liu, I.R. Lane, Attention-based recurrent neural network models for joint intent detection and slot filling, in: INTERSPEECH, ISCA, 2016, pp. 685–689, http://dx.doi.org/10.21437/Interspeech.2016-1352.

[26] Q. Chen, Z. Zhuo, W. Wang, BERT for joint intent classification and slot filling, 2019, CoRR abs/1902.10909.

[27] S. Louvan, B. Magnini, Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey, in: COLING, International Committee on Computational Linguistics, 2020, pp. 480–496, http://dx.doi.org/10.18653/v1/2020.coling-main.42.

[28] H. Liu, F. Zhang, X. Zhang, S. Zhao, X. Zhang, An explicit-joint and supervised-contrastive learning framework for few-shot intent classification and slot filling, in: EMNLP (Findings), Association for Computational Linguistics, 2021, pp. 1945–1955, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.167.

[29] M. Ren, R. Liao, E. Fetaya, R.S. Zemel, Incremental few-shot learning with attention attractor networks, in: NeurIPS, 2019, pp. 5276–5286.

[30] P. Mazumder, P. Singh, P. Rai, Few-shot lifelong learning, in: AAAI, AAAI Press, 2021, pp. 2337–2345.

[31] H. Qi, M. Brown, D.G. Lowe, Low-shot learning with imprinted weights, in: CVPR, IEEE, 2018, pp. 5822–5830, http://dx.doi.org/10.1109/CVPR.2018.00610.

[32] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, Y. Gong, Few-shot class-incremental learning, in: CVPR, Computer Vision Foundation/ IEEE, 2020, pp. 12180–12189, http://dx.doi.org/10.1109/CVPR42600.2020.01220.

[33] X. Tao, X. Chang, X. Hong, X. Wei, Y. Gong, Topology-preserving class-incremental learning, in: ECCV (19), Springer, 2020, pp. 254–270, http://dx.doi.org/10.1007/978-3-030-58529-7_16.

[34] S. Dong, X. Hong, X. Tao, X. Chang, X. Wei, Y. Gong, Few-shot class-incremental learning via relation knowledge distillation, in: AAAI, AAAI Press, 2021, pp. 1255–1263.

[35] F. Zhu, X. Zhang, C. Wang, F. Yin, C. Liu, Prototype augmentation and self-supervision for incremental learning, in: CVPR, IEEE, 2021, pp. 5871–5880.

[36] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, Y. Xu, Few-shot incremental learning with continually evolved classifiers, in: CVPR, IEEE, 2021, pp. 12455–12464.

[37] J. Zhang, T. Bui, S. Yoon, X. Chen, Z. Liu, C. Xia, Q.H. Tran, W. Chang, P.S. Yu, Few-shot intent detection via contrastive pre-training and fine-tuning, in: EMNLP (1), Association for Computational Linguistics, 2021, pp. 1906–1912, http://dx.doi.org/10.18653/v1/2021.emnlp-main.144.

[38] J. Zhang, K. Hashimoto, W. Liu, C. Wu, Y. Wan, P.S. Yu, R. Socher, C. Xiong, Discriminative nearest neighbor few-shot intent detection by transferring natural language inference, in: EMNLP (1), Association for Computational Linguistics, 2020, pp. 5064–5082, http://dx.doi.org/10.18653/v1/2020.emnlp-main.411.

[39] A. Ayub, A.R. Wagner, Tell me what this is: Few-shot incremental object learning by a robot, in: IROS, IEEE, 2020, pp. 8344–8350, http://dx.doi.org/10.1109/IROS45743.2020.9341140.

[40] B. Peng, C. Zhu, M. Zeng, J. Gao, Data augmentation for spoken language understanding via pretrained language models, in: Interspeech, ISCA, 2021, pp. 1219–1223, http://dx.doi.org/10.21437/Interspeech.2021-117.

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (9) (2019).

[42] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, I. Vulić, Effcient intent detection with dual sentence encoders, in: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, 2020, pp. 38–45, http://dx.doi.org/10.18653/v1/2020.nlp4convai-1.5.

[43] A. Cheraghian, S. Rahman, P. Fang, S.K. Roy, L. Petersson, M. Harandi, Semantic-aware knowledge distillation for few-shot class-incremental learning, in: CVPR, Computer Vision Foundation/ IEEE, 2021, pp. 2534–2543.

[44] G.M. van de Ven, A.S. Tolias, Three scenarios for continual learning, 2019, CoRR abs/1904.07734.
[45] A. Chaudhry, P.K. Dokania, T. Ajanthan, P.H.S. Torr, Riemannian walk for incremental learning: Understanding forgetting and intransigence, in: ECCV(11), Springer, 2018, pp. 556–572, http://dx.doi.org/10.1007/978-3-030-01252-6_33.
[46] F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, in: ICML, PMLR, 2017, pp. 3987–3995.
[47] F.M. Castro, M.J. Marín-Jiménez, N. Guil, C. Schmid, K. Alahari, End-to-end incremental learning, in: ECCV (12), Springer, 2018, pp. 241–257, http://dx.doi.org/10.1007/978-3-030-01258-8_15.
[48] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continual learning, in: NIPS, 2017, pp. 6467–6476.
[49] Z. Tan, K. Ding, R. Guo, H. Liu, Graph few-shot class-incremental learning, in: WSDM, ACM, 2022, pp. 987–996.
[50] C. Xia, W. Yin, Y. Feng, P.S. Yu, Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system, in: NAACL-HLT, Association for Computational Linguistics, 2021, pp. 1351–1360, http://dx.doi.org/10.18653/v1/2021.naacl-main.106.
[51] S. Rebuffi, A. Kolesnikov, G. Sperl, C.H. Lampert, icarl: Incremental classifier and representation learning, in: CVPR, IEEE Computer Society, 2017, pp. 5533–5542, http://dx.doi.org/10.1109/CVPR.2017.587.
[52] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, Y. Fu, Large scale incremental learning, in: CVPR, Computer Vision Foundation/ IEEE, 2019, pp. 374–382, http://dx.doi.org/10.1109/CVPR.2019.00046.
[53] Q. Hu, F. Lin, B. Wang, C. Li, Mbrep: Motif-based representation learning in heterogeneous networks, Expert Syst. Appl. 190 (2022) 116031, URL: https://doi.org/10.1016/j.eswa.2021.116031.
[54] A. Ayub, A.R. Wagner, Cognitively-inspired model for incremental learning using a few examples, in: CVPR Workshops, IEEE, 2020, pp. 897–906, http://dx.doi.org/10.1109/CVPRW50498.2020.00119.
[55] B. Fritzke, A growing neural gas network learns topologies, in: NeurIPS, MIT Press, 1994, pp. 625–632.
[56] T. Martinetz, Competitive hebbian learning rule forms perfectly topology preserving maps, in: International Conference on Artificial Neural Networks, Springer, 1993, pp. 427–434, http://dx.doi.org/10.1007/978-1-4471-2063-6_104.
[57] A.M. Alvi, S.F. Basher, A.H. Himel, T. Sikder, M. Islam, R.M. Rahman, An adaptive grayscale image de-noising technique by fuzzy inference system, in: ICNC-FSKD, IEEE, 2017, pp. 1301–1308, http://dx.doi.org/10.1109/FSKD.2017.8392954.
[58] J. Zheng, F. Cai, Y. Ling, et al., Heterogeneous graph neural networks to predict what happen next, in: COLING, International Committee on Computational Linguistics, 2020, p. 328–338.
[59] Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: A survey on few-shot learning, in: ACM Comput. Surv., vol. 53, 2020, pp. 63:1–63:34, http://dx.doi.org/10.1145/3386252.
[60] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACLHLT (1), Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/n19-1423.
[61] J. Zheng, F. Cai, H. Chen, Incorporating scenario knowledge into a unified fine-tuning architecture for event representation, in: SIGIR, ACM, 2020, pp. 249–258.
[62] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: CVPR, IEEE Computer Society, 2015, pp. 815–823, http://dx.doi.org/10.1109/CVPR.2015.7298682.
[63] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, in: NIPS, 2016, pp. 3630–3638.
[64] G. Koch, R. Zemel, R. Salakhutdinov, et al., Siamese neural networks for one-shot image recognition, in: ICML Deep Learning Workshop, Lille, 2015.
[65] T. Wolf, et al., Transformers: State-of-the-art natural language processing, in: EMNLP (Demos), Association for Computational Linguistics, 2020, pp. 38–45, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6.
[66] C.T. Hemphill, J.J. Godfrey, G.R. Doddington, The ATIS spoken language systems pilot corpus, in: HLT, 1990.