Incremental Intent Detection for Medical Domain with Contrastive Replay Networks

Guirong Bai^{1,2}, **Shizhu He**^{1,2}, **Kang Liu**^{1,2,3}, **Jun Zhao**^{1,2}
^{1} National Laboratory of Pattern Recognition, Institute of Automation, CAS
^{2} School of Artificial Intelligence, University of Chinese Academy of Sciences
^{3} Beijing Academy of Artificial Intelligence
{guirong.bai, shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Conventional approaches to medical intent detection require fixed pre-defined intent categories. However, due to the incessant emergence of new medical intents in the real world, such requirement is not practical. Considering that it is computationally expensive to store and re-train the whole data every time new data and intents come in, we propose to incrementally learn emerged intents while avoiding catastrophically forgetting old intents. We first formulate incremental learning for medical intent detection. Then, we employ a memory-based method to handle incremental learning. We further propose to enhance the method with contrastive replay networks, which use multilevel distillation and contrastive objective to address training data imbalance and medical rare words respectively. Experiments show that the proposed method outperforms the state-of-the-art model by 5.7% and 9.1% of accuracy on two benchmarks respectively.

1 Introduction

Medical intent detection aims to identify intents of medical queries and classify them into specific categories (Chen et al., 2012; Howard and Cambria, 2013; Guo et al., 2014; Cai et al., 2017). In medical scenarios, understanding query intent is very important for medical question answer systems (Wu et al., 2020; Mrini et al., 2021). Typically, to perform medical intent detection, existing methods pre-define a class set of fixed medical intent categories in advance, and train the whole collected dataset with the fixed class set. However, novel medical intents incessantly emerge with new data in the real world. When given a query with new intent category that is out of the pre-defined class set, these models can do nothing about it.

A straightforward solution is to store and re-train the whole data every time new data and intents come in. However, it is almost infeasible with limited storage budget and computation cost in practice (Wang et al., 2019). Consider the above problems, we propose to address this issue in a *incremental learning* way (Ring et al., 1994; Thrun, 1998), where the number of intent categories is allowed to incrementally increase and the system can incessantly learn emerged novel intents from continually arriving data of new intents, which is illustrated as Figure 1.

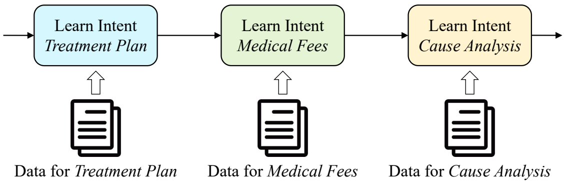

Figure 1: Incremental medical intent detection.

Naturally, one simple incremental method is to directly finetune the model by new intents data. However, this method suffers from the serious *catastrophic forgetting* problem (McCloskey and Cohen, 1989; French, 1999; Wang et al., 2019). After fitting emerged data of new intent, the performance of the model on old classes will inevitably drop a lot. There are some studies that have made effects to overcome the catastrophic forgetting problem. Typically, they are divided into parameter-based methods that preserve parameters important to the previous classes when updating (Kirkpatrick et al., 2017; Aljundi et al., 2018), and memory-based methods that store a few examples for each old class and replay them with arriving data of new classes (Rebuffi et al., 2017; Hou et al., 2018, 2019). Due to the simplicity and effectiveness, memory-based methods dominated this field.

However, when applying memory-based methods to incremental intent detection for medical domain, these methods face two new challenges – training data imbalance and medical rare words. **Training Data Imbalance:** While the amount of new classes data is often large, only a few examples for old classes are stored in the limited memory.

Findings of the Association for Computational Linguistics: ACL 2022, pages 3549 - 3556
May 22-27, 2022 ©2022 Association for Computational Linguistics

Therefore, the model can be drastically altered by the richer new classes data and ignore old classes (He and Garcia, 2009; Wu et al., 2019; Zhang et al., 2017). **Medical Rare Words:** Compared with common domains, the medical domain typically contains many domain-specific rare words[1]. These rare words are usually not well learned by the model and bring disturbance to representations of examples, which is adverse to selecting representative examples as memory for old classes replay.

Considering the above problems, we propose contrastive replay networks to enhance incremental intent detection for medical domain. Specifically, to address training data imbalance, we devise multilevel distillation to make the current model mimic the behaviors of the original model. To address medical rare words, we devise contrastive objective to push examples from the same intent close and examples from different intents further apart. Experimental results demonstrate that our method outperforms previous state-of-the-art models.

2 Related Work

2.1 Medical Intent Detection

The goal of intent detection is to identify query intent and classify them into specific categories (Chen et al., 2012; Howard and Cambria, 2013; Guo et al., 2014; Cai et al., 2017). With artificial intelligence gradually changing the landscape of healthcare and biomedical research (Yu et al., 2018), medical intent detection (Zhang et al., 2021; Chen et al., 2020a) becomes an important task. In medical domain, query intent can be divided into many categories, such as disease description, medical fees, treatment plan, precautions, and so on, which are domain-specific with highly specialized medical knowledge (Zhang et al., 2021). Understanding medical can assist medical question answer systems and significantly improve the relevance of medical search results(Wu et al., 2020; Mrini et al., 2021).

With the development of medical systems, the categories of intent continually increase in real-world applications. It means that medical intent detection is facing new challenges of how to incrementally learning new intents while avoid forgetting old classes.

2.2 Incremental Learning

Incremental learning, which is also called continual learning or lifelong learning, is a problem that deserves effort and has been studied for a long time in machine learning (Cauwenberghs and Poggio, 2001; Kuzborskij et al., 2013). It aims to incrementally train a model on new data to learn incessantly emerging novel classes while avoiding the catastrophic forgetting problem of old classes (McCloskey and Cohen, 1989; French, 1999). Existing incremental learning methods are mainly divided into two types, parameter-based methods (Kirkpatrick et al., 2017; Aljundi et al., 2018) and memory-based methods (Rebuffi et al., 2017; Hou et al., 2019; Wang et al., 2019; Cao et al., 2020; **?**).

In parameter-based methods, these methods try to capture and preserve parameters important to the previous classes (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018). When updating the model with new data, recognized important parameters tend to be constant. For example, existing studies propose to keep the updated parameters close to the optimal parameters for the old classes when training data of new classes (Kirkpatrick et al., 2017). However, it is difficult to provide a reasonable metric to evaluate all the parameters. In memory-based methods, these methods store a few examples to replay for each old class (Castro et al., 2018; Wang et al., 2019; Han et al., 2020). When data of new classes arrives, memory-based methods learn these examples again with the new data to alleviate catastrophic forgetting. For example, existing studies propose an episodic memory replay method that randomly selects examples to store (Wang et al., 2019).

Among these methods, memory-based methods dominated this field due to their simplicity and effectiveness. However, these methods cannot handle medical term disturbance and training data imbalance in incremental intent detection for medical domain.

3 Problem Definition

Medical intent detection (MID) is a classification task. In real medical applications, new intent classes incessantly emerge. Therefore, a practical MID system should be able to incrementally learn new query intent classes. We introduce a new problem, *incremental intent detection*. Suppose that there is a class-incremental data stream, denoted as $\{\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^{(M)}\}$. Each $\mathcal{X}^{(k)}$ contains train-

[1] In our experiments, more than 60% queries contain at least one rare word (such as diseases, proteins and chemicals) in the context.

ing/validation/testing data $(\mathcal{X}^{(k)}_{train}, \mathcal{X}^{(k)}_{valid}, \mathcal{X}^{(k)}_{test})$ and its own intent class set $\mathcal{C}^{(k)}$. Note that any two intent class sets are disjoint[2], i.e., $\mathcal{C}^{(i)} \cap \mathcal{C}^{(j)} = \emptyset (i \neq j)$. At the k-th step, the MID model optimizes its parameters using the training data $\mathcal{X}^{(k)}_{train}$ and the updated model should still perform well on historical classes, i.e., classes from 1 to $k-1$. Thus, for testing at the k-th step, we evaluate the updated model on the testing data of all old classes, i.e., $\cup_{i=1}^{k} \mathcal{X}^{(i)}_{test}$. Given an input from $\mathcal{X}^{(j)} (j \leq k)$, the model needs to give a prediction from $\cup_{i=1}^{k} \mathcal{C}^{(i)}$, instead of $\mathcal{C}^{(j)}$. To alleviate catastrophic forgetting, memory-based incremental learning methods allow limited memory to store a few examples for each old class. Therefore, every time new classes data arrive, the MID model utilizes the stored old classes data and new classes data $\mathcal{X}^{(k)}_{train}$ as training data to re-train parameters. The overall training procedures are described in Appendix.

4 Method

In this paper, we propose **C**ontrastive **R**eply **N**etworks for incremental medical intent detection (CRN). CRN consists of three components: 1) Intent classifier with memory, 2) Multilevel distillation and 3) Contrastive objective.

4.1 Intent Classifier with Memory

4.1.1 Memory-based Framework

We use BERT (Devlin et al., 2018) as the medical intent classifier. When the new medical intent classes $\mathcal{C}^{(k)}$ arrives, the corresponding new training data is denoted as $\mathcal{X}^{(k)}_{train} = \{(X_i, Y_i), 1 \leq i \leq K\}$, where K is the number of training examples, X_i is the query and Y_i denotes the intent label of the query X_i. The memory stores the representative examples for old m classes, i.e., $m = | \cup_{i=1}^{k-1} \mathcal{C}^{(i)}|$, we denoted it as $\mathcal{P} = \{\mathcal{P}^{(1)}, \cdots, \mathcal{P}^{(m)}\}$, where $\mathcal{P}^{(i)}$ is the set of stored examples for the i-th class. We combine the stored old data and new classes data, which is denoted as $\mathcal{N} = \mathcal{P} \cup \mathcal{X}^{(k)}_{train}$, to train the current model. The current label set $\mathcal{C}^{o}$ contains all observed intent categories, i.e., $\mathcal{C}^{o} = \cup_{i=1}^{k} \mathcal{C}^{i}$. Then a softmax classifier is to predict intent categories with representations of "[CLS]" in BERT. Finally, we use the cross entropy to train the intent detection model, denoted as loss $\mathcal{L}_{ce}$. Note that the current label set size is the number of all observed classes, i.e., $|\mathcal{C}^{o}|$. The overall training procedures are illustrated in the appendix.

4.1.2 Memory Updating

To overcome the catastrophic forgetting problem of old classes, memory-based methods store examples for each old class in the memory. All classes are treated equally in our model. Therefore, if m classes have been learned so far and B is the total number of examples stored in the memory, our model will store $n = B/m$ examples for each old class. Inspired by prototype learning (Snell et al., 2017; Yang et al., 2018), we select the top n example closest to the centroid of examples (based on "[CLS]" embeddings) of each class and store them into the memory as representative ones for old classes. When new data comes in, these examples in the memory are trained with the new data. Before the next new class arrives, we remove $B/m - B/(m+t)$ stored examples of each old class in the memory and allocate space to store $B/(m+t)$ current new classes examples based on the centroid, where $t = |\mathcal{C}^{k}|$ is the number of current new classes.

4.2 Multilevel Distillation

Although storing a few examples for each old class as memory is useful to avoid catastrophic forgetting, there is a serious data imbalance problem between memory and new classes data, which makes the model have an obvious bias towards the new classes, resulting in severely forgetting classification ability of previous classes. To address it, we devise multilevel distillation to make the current model mimic the behaviors of the original model.

Inspired by (Hinton et al., 2015), we first perform it at prediction level. We encourage the current predictions on old classes to match the soft labels by the original model. Formally,

$$\mathcal{L}_{pl} = -\frac{1}{|\mathcal{N}|} \sum_{X \in \mathcal{N}} \sum_{x \in X} \sum_{i=1}^{m} \alpha_i^* \log(\alpha_i) \quad (1)$$

where α_i^* and α_i is the output probability for the i-th label of the original and current model, respectively. This loss function is performed for all arriving new classes data and stored old classes data in the memory.

Then, we also perform it at feature level. We encourage the feature representations ("[CLS]" of BERT) of the current model don't greatly deviate

[2] $\mathcal{X}^{(k)}$ contains one or more new classes represented by $\mathcal{C}^{(k)}$.

from the ones of the original model. Formally,

$$\mathcal{L}_{fl} = -\frac{1}{|\mathcal{N}|} \sum_{X \in \mathcal{N}} \sum_{x \in X} cos(f^*(x), f(x)) \quad (2)$$

where $cos(f^*(x), f(x))$ measures the cosine distance between feature vectors of the original and current model. This loss is computed for all samples from the new classes and stored examples in the memory.

4.3 Contrastive Objective

Medical rare words are usually not well learned by the model and bring disturbance to representations of examples, which is detrimental to memory selection for old classes. Inspired by (Chen et al., 2020b), we devise contrastive objective to push examples from the same intent close and examples from different intents further apart. As a result, we can obtain better representations for examples and select more representative examples for replay.

Specifically, we first perform data augmentation to obtain more examples. We use a medical dictionary that contains medical rare words[3] to randomly add, delete or replace rare words over the original examples. After that, we employ an objective to push examples from the same intent close and examples from different intents further apart:

$$\mathcal{L}_{co} = -\sum_{i} \frac{1}{N_{y_i} - 1} \sum_{j \neq i} (1_{y_i = y_j} - 1_{y_i \neq y_j}) \log s_{ij} \quad (3)$$

where $s_{ij} = \frac{\exp(f(x_i) \cdot f(x_j))}{\sum_{k \neq i} \exp(f(x_i) \cdot f(x_k))}$, $f(x)$ denotes the embeddings of "[CLS]" token of example x. Finally, our model is optimized by the total loss $\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{pl} + \mathcal{L}_{fl} + \mathcal{L}_{co}$.

5 Experiments

5.1 Benchmarks

We use two public medical datasets KUAKE-QIC in CBLUE (Zhang et al., 2021) and CMID (Chen et al., 2020a) to construct benchmarks for incremental learning setting. For a medical intent detection dataset, its intent classes are arranged in a fixed order. Then, methods are trained in a class-incremental way on the available training data[4]. Due to its long-tail frequency distribution, we use the data of the top 10/20 most frequent classes for KUAKE-QIC/CMID.

Models	KUAKE-QIC		CMID	
	Avg	Whole	Avg	Whole
Upperbound	94.6	91.6	78.0	87.2
Finetune	20.9	43.0	9.2	24.9
EWC	63.7	52.5	27.5	26.9
EMR	68.6	62.9	31.3	30.3
EMAR	71.0	65.9	33.7	32.1
CRN(Ours)	**75.8**	**71.6**	**43.5**	**41.2**

Table 1: The average accuracy (%, "Avg") on all observed classes and whole accuracy (%, "Whole") on the whole testing data at the last step.

5.2 Evaluation and Implementation

We use accuracy as the evaluation metric. Every time the model finishes training on the new classes data, we report the accuracy on the whole testing data of all observed classes. Besides, we also report the performance at the last step, containing **Average accuracy** (macro-averaging) and **Whole accuracy** (micro-averaging).

We use base BERT as the classifier. The learning rate is set to 2e-5. The batch size is 8. For the two benchmarks, both the capacity of memory is $B = 200$.

5.3 Baselines

We compare CRN with 4 baselines: 1) EWC (Kirkpatrick et al., 2017): The representative parameter-based method that keeps the network parameters close to the optimal parameters for the previous classes when training new classes data. 2) EMR (Wang et al., 2019): The representative memory-based method that avoids catastrophic forgetting via randomly storing a few examples of old classes. 3) EMAR (Han et al., 2020): The latest memory-based method that utilizes prototypes for memory reconsolidation exercise to keep a stable understanding of old classes. 4) Finetune: The lower bound that simply finetunes the model on arriving data of new classes. 5) Upperbound: The upper bound that stores and trains all observed samples.

5.4 Compared with State-of-the-art Methods

We conduct experiments on KUAKE-QIC and CMID. The accuracies over all observed classes during the whole incremental learning process are plotted in Figure 2. We also show the results at the last step in Table 1. We can find that our method outperforms all other baselines by a large margin. Specifically, compared with the state-of-the-art model EMAR, our method achieves 5.7% and 9.1% improvements of whole accuracy score on the

[3] We build it by collecting medical entities in OpenKG at http://www.openkg.cn/dataset.

[4] Only one new class is available for the model at each time, i.e., $t = |\mathcal{C}^{(k)}| = 1$.

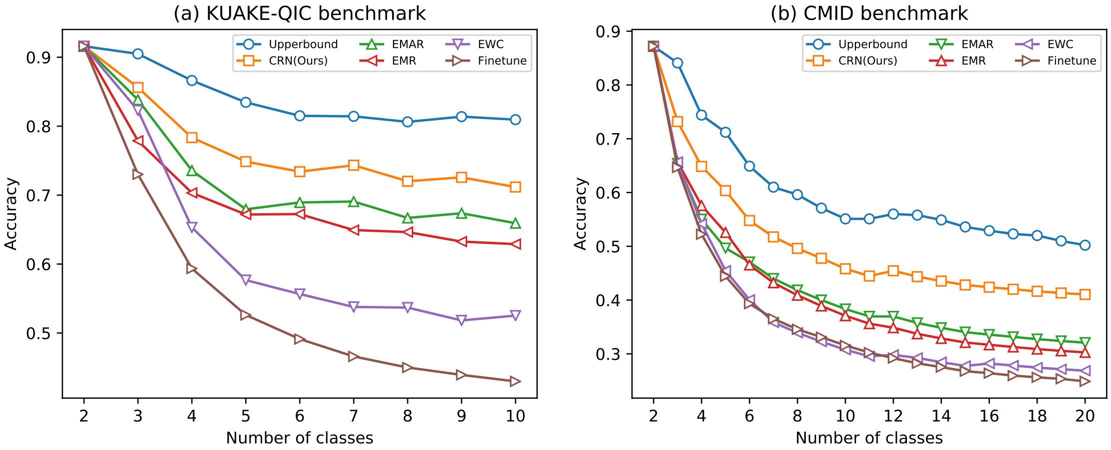

Figure 2: The performance on (a) KUAKE-QIC and (b) CMID. Our method CRN outperforms others.

Models	KUAKE-QIC		CMID	
	Avg	Whole	Avg	Whole
CRN	**75.8**	**71.6**	**43.5**	**41.2**
w/o PL	72.5	66.8	37.7	37.0
w/o FL	72.3	67.1	37.1	36.3
w/o CO	73.4	69.7	40.6	39.3

Table 2: Ablation studies for the main components.

KUAKE-QIC and CMID, respectively. It demonstrates the effectiveness of our proposed CRN. Besides, Finetuning always obtains the worst performance on both benchmarks and becomes the lower bound, which indicates that the catastrophic forgetting problem is serious. Moreover, the large gap between all methods and Upperbound indicates that this issue is still challenging.

5.5 Ablation Experiment

To investigate the effectiveness of the different parts in our method, we conduct ablation studies. The results are shown in Table 2. (1) Effectiveness of distillation at prediction level: The performance drops with removing $\mathcal{L}_{pl}$ ("w/o PL"). It demonstrates that it is useful to handle training data imbalance. (2) Effectiveness of distillation at feature level: The performance drops with removing $\mathcal{L}_{fl}$ ("w/o FL"). It demonstrates that it is effective to address training data imbalance. (3) Effectiveness of contrastive objective: The performance drops with removing $\mathcal{L}_{co}$ ("w/o CO"). It demonstrates that it is helpful to handle medical rare words. We report extra experiments in Appendix.

6 Conclusion

We explore to incrementally learning medical intent detection. We propose contrastive replay networks to handle training data imbalance and medical rare words. Experiments demonstrate that our method outperforms previous state-of-the-art models.

Acknowledgements

This work is supported by the Natural Key R&D Program of China (No.2020AAA0106400), the National Natural Science Foundation of China (No. 61922085, No.61976211) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006). This research work was supported by the independent research project of National Laboratory of Pattern Recognition, the Youth Innovation Promotion Association CAS.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154.

Ruichu Cai, Binjun Zhu, Lei Ji, Tianyong Hao, Jun Yan, and Wenyin Liu. 2017. An cnn-lstm attention approach to understanding user query intent from online health communities. In *2017 ieee international conference on data mining workshops (icdmw)*, pages 430–437. IEEE.

Pengfei Cao, Yubo Chen, Jun Zhao, and Taifeng Wang. 2020. Incremental event detection via knowledge consolidation networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 707–717, Online. Association for Computational Linguistics.

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-end incremental learning. In *Proceedings of* *the European conference on computer vision (ECCV)*, pages 233–248.

Gert Cauwenberghs and Tomaso Poggio. 2001. Incremental and decremental support vector machine learning. *Advances in neural information processing systems*, pages 409–415.

Long Chen, Dell Zhang, and Levene Mark. 2012. Understanding user intent in community question answering. In *Proceedings of the 21st international conference on world wide web*, pages 823–828.

Nan Chen, Xiangdong Su, Tongyang Liu, Qizhi Hao, and Ming Wei. 2020a. A benchmark dataset and case study for chinese medical question intent classification. *BMC Medical Informatics and Decision Making*, 20(3):1–7.

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

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

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Daniel Guo, Gokhan Tur, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 554–559. IEEE.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. Continual relation learning via episodic memory activation and reconsolidation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6429–6440.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2018. Lifelong learning via progressive distillation and retrospection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 437–452.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839.

Newton Howard and Erik Cambria. 2013. Intention awareness: improving upon situation awareness in human-centric environments. *Human-centric Computing and Information Sciences*, 3(1):1–17.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Ilja Kuzborskij, Francesco Orabona, and Barbara Caputo. 2013. From n to n+ 1: Multiclass transfer incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3358–3365.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021. A gradually soft multi-task and data-augmented approach to medical question understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1505–1515, Online. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Mark Bishop Ring et al. 1994. Continual learning in reinforcement environments.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4080–4090.

Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 796–806.

Chaochen Wu, Guan Luo, Chao Guo, Yin Ren, Anni Zheng, and Cheng Yang. 2020. An attention-based multi-task model for named entity recognition and intent analysis of chinese online medical questions. *Journal of Biomedical Informatics*, 108:103511.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382.

Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. 2018. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3482.

Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Lei Li, Xiang Chen, Shumin Deng, Luoqiu Li, Xin Xie, Hongbin Ye, Xin Shang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. 2017. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418.

Models	CRN		EMAR	
	Avg	Whole	Avg	Whole
50	66.8	61.9	60.9	60.4
100	69.2	64.7	61.8	61.2
150	72.8	68.5	65.2	63.2
200	**75.8**	**71.6**	**71.0**	**65.9**

Table 3: The effect of the number of stored examples. We compare our CRN with EMAR on KUAKE-QIC.

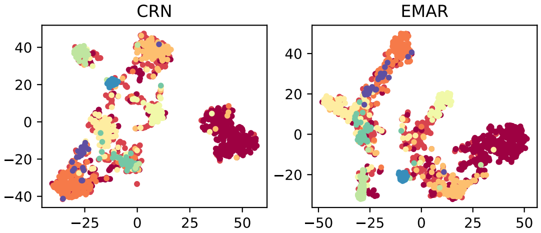

Figure 3: Feature spaces learned by CRN and EMR, respectively.

Below we provide some extra experiments for discussion and overall training procedures for understanding.

A The Effect of the Number of Stored Examples

To show the effect of different numbers of stored examples, we compare our CRN with another memory-based method EMAR on KUAKE-QIC , where the memory size to store examples is from 50 to 200. We can observe the results in Table 3.

First, the more examples stored, the better performance for both memory-based methods. We also see that our method performs better, which demonstrates the effectiveness of our method. Even with fewer examples stored, our CRN still performs better, demonstrating that our method is effective to address training data imbalance.

B Visualization

To show the effectiveness of introduced contrast objective for medical rare words, we also give the visualization in Figure 3 to show the feature spaces learned by our CRN (Ours) and EMAR (i.e., the latest representative work of memory-based method) on KUAKE-QIC.

Specifically, we extract the representations of "[CLS]" and use t-SNE to implement visualization. We can see that the feature space of CRN is more sparse and features from different intents are more distinguishable. However, the features learned by EMAR are more difficult to distinguish. Examples from the same intent in CRN are closer than ones in EMAR. It is because medical rare words are easy to bring disturbance and make the queries confused in EMAR. This result shows that our contrast objective in CRN can learn better representations for queries against medical rare words.

Algorithm 1 Training Procedures

```
Require: arriving training data $\mathcal{X}_{train}^{(k)}$ at the $k$-th step, the number of new classes $t = |\mathcal{C}^{(k)}|$, memory capacity $B$, current model $\mathcal{M}$, current reserved example sets $\mathcal{P} = (\mathcal{P}^{(1)}, \cdots, \mathcal{P}^{(m)})$, $m$ observed classes
combining training data $\mathcal{X}_{train}^{(k)} \cup \mathcal{P}$
Update the model $\mathcal{M}$ with loss $\mathcal{L}$
for $c = 1, \cdots, m$ do
    Remove stored examples for each old class $c$ until the number reaches $B/(m+t)$
end for
for $c = m+1, \cdots, m+t$ do
    Update the centroid
    Select the top $B/(m+t)$ examples close to the centroid to store in the memory
end for
```

C Training Procedures

Algorithm 1 describes the overall training procedures of incremental learning. After training a intent detection model with limited intents, new intents come in. Through our method, the model incrementally learns new intents while avoiding catastrophically forgetting old intents.