Depth Aware Hierarchical Replay Continual Learning for Knowledge Based Question Answering

Zhixiong Cao[1], **Hai-Tao Zheng**[1,2,‡], **Yangning Li**[1,2,‡],
Jin Xu[1], **Rongsheng Li**[1], **Hong-Gee Kim**[3]
[1] Shenzhen International Graduate School, Tsinghua Universiy
[2] Pengcheng Laboratory, Shenzhen, China
[3] Seoul National University

Abstract

Continual learning is an emerging area of machine learning that deals with the issue where models adapt well to the latest data but lose the ability to remember past data due to changes in the data source. A widely adopted solution is by keeping a small memory of previously learned data that uses replay. Most of the previous studies on continual learning focused on classification tasks, such as image classification and text classification, where the model needs only to categorize the input data. Inspired by the human ability to incrementally learn knowledge and solve different problems using learned knowledge, we considered a more practical scenario, knowledge based quesiton answering about continual learning. In this scenario, each single question is different from others(which means different fact triples to answer them) while classification tasks only need to find feature boundaries of different categories, which are the curves or surfaces that separate different categories in the feature space. To address this issue, we proposed a Depth Aware Hierarchical Replay (DAHR) framework which includes a tree structure classifier to have a sense of knowledge distribution and fill the gap between text classification tasks and question-answering tasks for continual learning, a local sampler to grasp these critical samples and a depth aware learning network to reconstruct the feature space of a single learning round. In our experiments, we have demonstrated that our proposed model outperforms previous continual learning methods in mitigating the issue of catastrophic forgetting.

Keywords: continual learning, question answering, catastrophic forgetting

1. Introduction

In conventional machine learning or deep learning tasks (Dong et al., 2023; Liu et al., 2022), the common practice involves feeding all training data to the model simultaneously to enhance the learning of input sample representations (Li et al., 2022c; Ma et al., 2022). As the volume of training data continues to grow, a proportional expansion of the model becomes necessary (De Lange et al., 2021). Typically, the idea of processing training data in discrete batches can be considered, akin to the human process of sequentially mastering different subjects in a predefined order. This concept forms the central focus of research in the field of Continual Learning, also known as Lifelong Learning and Incremental Learning.

The fundamental challenge addressed by continual learning is mitigating catastrophic forgetting(Schwarz et al., 2018), which is primarily attributed to the incongruity of feature distributions in input data. This phenomenon resembles the way humans tend to forget old information upon learning new material. Typically, these methods can broadly be categorized into three groups to address catastrophic forgetting: 1) regularization-based methods, 2) expansion-based methods, and 3) memory-based methods. The memory-based methods are among the most effective and widely used ones (Zhao et al., 2022). As illustrated in figure 1, the main idea of memory-based methods is to retrain samples or representations from already seen tasks when learning new tasks(Mundt et al., 2023). In this article, our primary focus is on the memory replay-based methods.

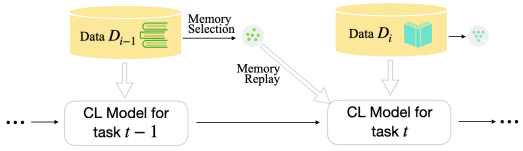

Figure 1: A commonly employed framework for continual learning with memory-replay methods.

Due to the necessity to consider each replay within memory constraints significantly smaller than the total memory capacity, the pivotal challenge faced by replay-based continual learning algorithms is how to optimally select or generate samples that represent all the acquired knowledge from this round of learning.

In past scenarios of continual learning research, Many works (Kumari et al., 2022) focused on continuous learning for recognizing image categories (e.g., sequentially learning cats, dogs, or sequentially learning from 1 to 10). In the field of Natu-

‡Corresponding authors: zheng.haitao@sz.tsinghua.edu.cn, and yn-li23@mails.tsinghua.edu.cn

LREC-COLING 2024, pages 4654–4664

ral Language Processing (NLP), (Ke et al., 2021) investigated the sequential learning of different categories of text, (Zhao et al., 2022) addressed sequential learning from diverse classes of dialogues. These studies were grounded in a strong assumption that the content learned in different batches was homogeneous. Consequently, the knowledge acquired in different rounds naturally exhibited better discrimination in high-dimensional feature spaces, with feature representations of samples from the same class being closer. Therefore, the approach to mitigating catastrophic forgetting was to identify samples near the boundaries of various categories for use as replay data (Kumari et al., 2022).In other words, for tasks focused solely on classification, it is only necessary to delineate the boundaries of distinct category information in the model's representation. Our work breaks free from this constraint by considering a more uncertain mode of continual learning. Hence, we chose knowledge-based question-answering as the research task for continual learning. In this scenario, different question-answer samples do not possess a direct category relationship (although implicit associations still exist, such as different questions having different types of relevance). This allows us to simulate a more naturally generalizable continual learning scenario.

We have considered a depth-aware hierarchical replay framework for continual learning. The objective is to find better replay samples in the aforementioned uncertain feature space. Specifically, we begin by employing unsupervised clustering to capture the overall feature distribution of the samples. Subsequently, we focus on feature selection within each local cluster to represent the samples within these clusters. Moreover, we hypothesize that, within a classification tree, samples from deeper clusters are often more representative of the overall feature distribution than shallower ones. Consequently, we retrain the selected samples from different clusters with appropriate weights. Extensive experimentation demonstrates that our proposed replay approach significantly outperforms baseline methods, offering a more effective solution to the problem of catastrophic forgetting in continual learning.

Moreover, from an efficiency perspective, our memory replay framework is both concise and efficient. In a broader sense, we can consider that for memory replay tasks, we select m sufficiently representative samples from a dataset of size M for retraining. For conventional strategies, the computational complexity is limited by the square of the number of samples, which can become very costly when M is sufficiently large. However, we have effectively ensured efficiency through multi-level sampling.

Overall, the main contributions of our work in this paper are:

- We have considered conducting continual learning beyond classification tasks, exploring how to assist in mitigating catastrophic forgetting among unclassified learning samples. Additionally, we support the random permutation of datasets to simulate the process of continual learning in a more human-like and realistic manner.

- We propose a depth-aware hierarchical memory replay method **DAHR** (**D**epth **A**ware **H**ierarchical **R**eplay) for continual learning, which initially self-classifies samples, then considers locally challenging learning samples, ultimately resulting in an efficient sample replay method with weighted sampling.

- Our experiments demonstrate that our approach effectively addresses the ability to better retain previously learned content during learning without compromising new learning efficiency.

2. Related Work

2.1. Continual Learing

Continual learning aims at incrementally acquiring new knowledge, and in the meantime, mitigating the catastrophic forgetting issue(Qin et al., 2022). These methods of continual learning can be categorized into three kinds : **Expansion-based methods**(eg, (Li and Hoiem, 2017; Yoon et al., 2017; Rosenfeld and Tsotsos, 2018; Hung et al., 2019; Veniat et al., 2020; Li et al., 2022b,a, 2023e)) dynamically expand the network capacity to reduce the interference between the new tasks and the old ones. **Regularization-based methods**(eg,(Kirkpatrick et al., 2017; Lee et al., 2017; Chaudhry et al., 2018a; Dhar et al., 2019; Ritter et al., 2018; Schwarz et al., 2018; Zenke et al., 2017; Ye et al., 2023)) protect the old tasks by adding regularization terms in the loss function to penalize the model change on their important weights. **Memory-based methods** mitigate forgetting mainly by either storing a subset of examples from the past tasks in the memory from rehearsal(Rebuffi et al., 2017; Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2018b, 2019; Riemer et al., 2018; Huang et al., 2023), or synthesizing old data from generative models to perform pseudo-rehearsal(Shin et al., 2017; Zhao et al., 2022).

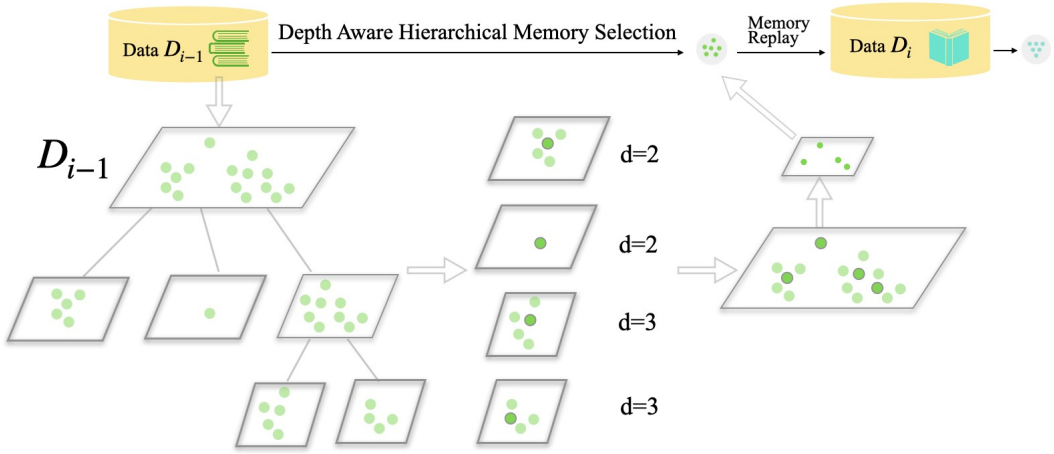

Figure 2: The DAHR replay method involves creating a tree-like structure using a recursive unsupervised classification process. Each learning round's qa data is represented at the root node. For each leaf cluster, the most representative samples are selected and aggregated. If the aggregated size exceeds the replay threshold $|M|$, some samples are randomly discarded. If the size is below the threshold, samples are randomly selected from the root node to supplement. These samples are then used in the next round of learning.

2.2. Knowledge Based Question Answering

The task of knowledge based question answering (namely KBQA) aims to answer the questions presented in natural language using the relevant facts available in knowledge base(Li et al., 2021, 2023a,c; Yu et al., 2023; Li et al., 2023d,b). Traditionally, the key tasks in addressing KBQA include entity and relationship recognition in the identification problem (Named Entity Recognition and Relationship Extraction). Past work has collected a series of datasets (Yih et al., 2016; Zhang et al., 2018; Su et al., 2016; Gu et al., 2021) as well as proposed a diversity of approaches for this task. These approaches can be roughly divided into two categories, semantic parsing based (Yih et al., 2014; Xu et al.; Xing et al., 2024) and information retrieval based(Bordes et al., 2015; Zhang et al., 2018, 2016; Tan et al., 2023).

Inspired by(Li et al., 2021), we explore continual learning in knowledge based question answering (KBQA). In contrast to this prior work, our approach is better suited for handling the replay of class-less question-answer samples, addressing the issue of catastrophic forgetting in continual learning more effectively.

3. Methodology

To alleviate the catastrophic forgetting issue of unclassified samples in the context of continual learning, we propose a continual learning framework **DAHR** (Depth Aware Hierarchical Replay) for KBQA tasks. In Section 3.1, we define the task objectives for continual learning in KBQA. In Section 3.2, we optimize the question representation through continual learning methods, obtaining vector encodings for each question-answer pair and associated knowledge from pretrained models. Section 3.3 and 3.4 introduce the DWHR replay method, which, theoretically, exhibits reasonable computational complexity while effectively addressing both local and global knowledge representations.

3.1. Problem Formulation

The knowledge based question answering (KBQA) is typically performed by analyzing mentioned entities s and entity relation r and then indexing the target entity o in the knowledge graph based on the fact triple $(s, r, o) \in (S, R, O)$ as the answer of the question given question q and a knowledge base. For instance, the mentioned entity and relation in *'who recorded the song baby'*are *'baby'* and *'song recorded by artist'*. Continuous learning in KBQA can be defined as follows: As fact triples and the relevant questions are assumed to be incrementally available in the sequence as the datasets $D_1, D_2, ..D_i, ..D_n$, where $D_i = \{Q_i, F_i\}$ and Q_i, F_i represent the set of questions, and the knowledge base represented by a set of fact tripples(s, r, o) in the $i-th$ learning task, and n denotes the total number of learning rounds in continual learning tasks. we incrementally train the model with new questions and related knowledge tripples.

Each round of training involves learning entirely new knowledge compared to the previous rounds.

Algorithm 1: DAHR - Depth Aware Hierarchical Replay for Continual Learning

Input: Datasets $(D_0, D_1, \ldots, D_{n-1})$, Replay Memory M, Cluster Sample Size c
Output: Critical Samples C

for *Dataset* $i \leftarrow 0$ **to** $n-1$ **do**
  for $(x_i, s_i, q_i, o_i) \sim D_i$ **do**
    prompted representation $e_i \leftarrow$ `Prompted LM Encoder`(x_i, s_i, q_i, o_i)
  while *size of* $C_i \geq c$*:* **do**
    $C_i \leftarrow$ `Hierarchical K-Means`(e_i)
  cluster $C_i \leftarrow$ `Most Similiarity Sampling`(C_i)
  $M \leftarrow C_i$ {// updating Replay Memory }

The challenge lies in effectively learning the knowledge from the current round while mitigating catastrophic forgetting in continual learning, ensuring good performance both on the current QA data and the QA data learned from previous rounds.

$$\min_{\theta} \sum_{i=1}^{T} E_{(x,y)\sim\tilde{\mathcal{D}}_i} \ell(x, y; \theta) \tag{1}$$

$$\theta = f(\theta_0; \mathcal{D}_1, M_{\mathcal{D}_1}, ..., \mathcal{D}_{T-1}, M_{\mathcal{D}_{T-1}}, \mathcal{D}_T) \tag{2}$$

$$M_{\mathcal{D}} = \mathcal{M}(\mathcal{D}, g) \tag{3}$$

The formulas 1, 2, and 3 provide the mathematical definitions for our task. Consider a T-round continual learning scenario where $\mathcal{D}_i$ and $\tilde{\mathcal{D}}_i$ represent the training and testing data for the i-th task. Here, θ represents the model parameters after continuous learning on data from all rounds, x and y denote the question and answer samples, and ℓ represents the loss function for questions and answers. Our task objective is to find the optimal parameters that minimize the mathematical expectation of answer loss. Formulas 2, and 3 describe the acquisition of θ, where f represents the model's training process, θ_0 represents the initial parameters, and $\mathcal{D}_1, M_{\mathcal{D}_1}, ..., \mathcal{D}_{T-1}, M_{\mathcal{D}_{T-1}}, \mathcal{D}_T$ denotes the all learning data. It's important to note that learning is performed sequentially in the listed order, and simultaneous training of multiple contents is not allowed. The function g represents our data replay algorithm, which corresponds to our DAHR framework. Typically $|M_{\mathcal{D}_i}| \ll |\mathcal{D}_T|, i \in \{1, 2, .., T-1\}$, which is the key to addressing the challenge of catastrophic forgetting in continual learning.

3.2. Prompted LM Encoder for KBQA

We employ a prompt-based learning approach to embed entities and relationships related to the question into its representation. Specifically, in contrast to directly concatenating the entities and relationships with the question, we utilize a set of prompt words to assist the model in acquiring encoding suitable for the QA task. We adopt the following prompt learning strategy for initial tokens: for question x, and its subject denoted as s, its relation denoted as r, its answer denote as o. We reconstruct this input as `a question is '[x]', the relation in this question is '[q]', the subject entity in this question is '[s]'.`

3.3. Generation of Depth Aware Learning Data

In contrast to continual learning in classification scenarios, in the case of question-answering tasks, the samples in different rounds of learning are not directly related. In other words, within all the question-answer knowledge learned in this round, different samples exhibit varying degrees of relevance. For instance, some questions share similar subject domains, while others share similar topics. Therefore, we can employ a self-supervised classification approach to make it more likely for similar questions to belong to the same cluster.

As illustrated in algorithm 1, we utilize the hierarchical K-means algorithm to encode all questions for each learning task. As depicted in figure 2, when provided with a collection of questions for indexing, we initially categorize all questions into k clusters using their representations encoded by Electra(Clark et al., 2020). If a cluster contains more than c questions, we apply the K-means algorithm recursively. Each cluster containing c questions or fewer (referred to as a 'leaf cluster') serves as the input for the next step.

3.4. Most Similiarity Sampling in Cluster

For each tree generated in the previous section, it can be to some extent regarded as a representation of a cluster of similar knowledge. Often, the cost of forming such a small cluster is relatively low. Hence, we can utilize similarity-based algorithms to select the most representative samples for this cluster of knowledge. It's worth noting that such methods are particularly convenient for sample selection in the case of small clusters (especially when only one sample needs to be selected). For instance, as-

suming a sampling rate of 0.1, when compared to selecting one sample from 10 nodes, selecting two samples from a cluster of 20 nodes would need to consider the risk of high similarity between the selected samples leading to memory wastage during replay. Therefore, our method offers a straightforward and efficient approach to selecting the most critical samples for each small cluster.

We employ an unsupervised classification strategy. Firstly, we obtain question representations using the model, and then we implement self-supervised classification using the K-means algorithm. Each classification process is binary. When the number of samples in a cluster exceeds N, we continue this process. Consequently, we obtain a tree-like structured labeled dataset. We refer to each final classification result $|C_i| < N$ as a node of this data tree, with the depth being the number of steps from the initial data classification to C_i.

For each node in this classification, we have two consistent characteristics: 1) For the local data, regarding a sample D_i within that node(leaf cluster), the remaining samples in that node are the most similar samples to this sample. 2) For the global data, as the data depth of the node increases, it implies that the node has the smallest total distance within the entire dataset in one task round.

4. Experiments

4.1. Dataset and Settings

We provide an extensive evaluation on Simple Question Dataset(Bordes et al., 2015), which is composed of a freebase knowledge base.

In the context of continuous learning for knowledge graph-based question-answering, we employed the Simple Question dataset, which is derived from the Freebase knowledge base. It is important to note that, unlike previous works where continuous learning was applied to classification tasks with distinct category labels, this particular task cannot be straightforwardly transferred to KBQA tasks due to the absence of clear category labels among the questions. In order to adapt this dataset to continuous learning while ensuring homogeneity between training and testing sets for each task, we sorted the dataset samples based on entities and relationships, subsequently dividing them into T subsets, each corresponding to a task in the learning process. Simultaneously, we restructured the KBQA task to infer (or extract) a unique correct question entity and relationship pair for each question within the limited (s, r, o) triple set. For the correct entities, we ensured that interfering knowledge triples were sufficiently close to the uniquely correct triples, providing a realistic measure of the difficulty in resolving KBQA questions.

Furthermore, given the necessity of employing certain stochastic algorithms in our work, including sample sampling in the stratified replay and random replay, there exist minor fluctuations in experimental results for the same model and algorithms. To ensure the accuracy and reliability of the experimental outcomes, we conducted a minimum of three experiments for each set of parameters in the results. We reported both the mean and variance of the metric values under the same model parameters and algorithms.

4.2. Baseline Methods And Training Details

We adopt the following methods as baselines in this work :

- ***Multitask*:** In this approach, it can be considered as a learning scenario with a single task, where the model learns all the question-answer data simultaneously.
- ***Fine-tuning*:** Fine-tune the model on new task data continually.
- ***Prompt-tuning*:**Differing from the previous method, this approach utilizes a simple but effective prompt encoding strategy.
- ***Random Memory Replay*:**Save $|M|$ samples randomly sampled from the training set of each task $\mathcal{T}_i$ to memory M_i and jointly train the model on new task data D_k and memory $M_{<k}$.
- ***EWC*:** Maintain the memory in the same way as Replay but use it to compute the Fisher information matrix for regularization (Kirkpatrick et al., 2017).

We employ the Electra pre-trained model as an encoder for our question-answer data. Specifically, in order to better demonstrate the benefits brought by the continuous learning approach, we have chosen the google/electra-small-discriminator model with a smaller parameter count to obtain the question encoding vectors. The encoding vector is passed through pooling and activation layers and then fed into a classifier to extract the correct (entity, relation) as an encoder-decoder structure for solving QA questions. In this structure, our training epochs are set to 5, the training and inference batch sizes are set to 50, and the learning rate is set to $5e-5$.

For the details of the memory replay algorithm, we use an unsupervised clustering algorithm, K-means, to classify the trees. Each classification is set to have 2 categories. Finally, after aggregating the samples selected from each cluster, we randomly sample and select m samples for replay.

Dataset	D_1	D_2	D_3	D_4	D_5
# q in train	18,001	22,055	10,768	11,672	15,961
# q in validation	2,250	2,757	1,346	1,459	1,995
# q in test	2,251	2,757	1,346	1,460	1,996
# relations in $\mathcal{F}_i$	1696	3,392	5,088	6,784	8,480
# relations of q	330	305	339	329	336

Table 1: Simple Question Dataset setting in our experiment

	$FA(\uparrow)$		$OA(\uparrow)$		$Forgotness(\downarrow)$	
$\|M\|$ for Replay (%)	2%	10%	2%	10%	2%	10%
multitask	86.4 ± 0.2	86.4 ± 0.2	86.2 ± 0.2	86.2 ± 0.2	-3.5 ± 0.4	-3.5 ± 0.4
fine-tuning	58.9 ± 0.8	58.9 ± 0.8	59.4 ± 0.9	59.4 ± 0.9	25.9 ± 0.9	25.9 ± 0.9
prompt-tuning	60.3 ± 2.1	60.3 ± 2.1	60.9 ± 1.3	60.9 ± 1.3	24.5 ± 1.4	24.5 ± 1.4
RMR	64.9 ± 2.4	$\mathbf{73.1 \pm 0.8}$	69.5 ± 1.0	76.8 ± 0.2	20.3 ± 3.3	12.3 ± 1.3
Ours	$\mathbf{67.8 \pm 2.8}$	72.6 ± 1.2	$\mathbf{72.1 \pm 2.1}$	$\mathbf{77.2 \pm 2.1}$	$\mathbf{19.6 \pm 3.5}$	$\mathbf{9.2 \pm 3.8}$

Table 2: The main results table includes the results of the baseline method and our method for three evaluation metrics across different memory sizes. Multitask training all data simultaneously represents the theoretical upper bound for the effectiveness of lifelong learning.'FA' represents the average accuracy in testing with question-answering data from different rounds after the final round of learning. 'OA', on the other hand, represents the overall average accuracy in all testing rounds.

In the experiments, we considered the method's performance under different memory constraints by using two replay ratios, $M_{|D|} = |D| * 0.02$ and $M_{|D|} = |D| * 0.1$.

4.3. Metrics

In the context of continual learning, after each learning round, we evaluate the knowledge acquired from previous tasks to assess the extent of forgetting. Consequently, each set of experiments includes C_{n+1}^2 results, where each result represents the accuracy (denoted as $a_{i,j}$) of evaluating in the $j-th$ task after learning the $i-th$ task. Based on these results, we introduce two key evaluation metrics: the final accuracy (FA), the overall accuracy (OA), and the total forgetness (Fg). These metrics provide a comprehensive and holistic assessment of the effectiveness of continual learning algorithms.

$$FA = \frac{1}{T}\sum_{j=1}^{T} a_{T,j}, OA = \frac{2}{(N+1)*N}\sum_{j=1}^{i}\sum_{i=1}^{T} a_{i,j} \tag{4}$$

$$Fg = \frac{1}{T-1}\sum_{j=1}^{T-1} \max_{l\in\{1,\dots T-1\}} (a_{l,j} - a_{T,j}) \tag{5}$$

4.4. Main Result

As shown in the table 2, we have reported the experimental results of different methods across three metrics. It is evident that our method has achieved the best performance across all three metrics.

As shown in the table 3, we compared random sampling and our DAHR sampling under the same parameters (aside from the replay method). Our experimental results demonstrate that our method significantly mitigates catastrophic forgetting in continual learning. We also observed an intriguing phenomenon: for different tasks, such as Task 1 and Task 2, it turns out that, for Task 1, subsequent learning on Task 2 exhibits faster forgetting (Task 1 experiences approximately a 15% accuracy drop after four rounds of continual learning, while Task 2 encounters a 40% drop in accuracy during three rounds of continual learning). Our analysis suggests that this phenomenon is likely due to varying degrees of overall knowledge distribution inconsistency during five rounds of different question-answer knowledge acquisition. Nevertheless, the DAHR replay method we propose consistently outperforms random sampling, achieving better results in continual learning.

Moreover, we found that for the final-round accuracy(FA), the overall accuracy(OA) better reflects the effectiveness of our method. We analyze that this is likely due to the significant variations in feature distributions of unlabeled learning samples, as in each learning iteration, the model places more emphasis on gradient updates from the most recent round of data. Therefore, when the knowledge distribution from previous rounds significantly differs from the current round, the average accuracy evaluation metric for the current round is likely to show a noticeable decrease compared to the previous round. Furthermore, since the data for each task in

	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	$Mean$
RMR	80.81					80.81
DAHR	**82.69**					**82.69**
RMR	75.65	79.05				77.40
DAHR	**78.57**	**80.43**				**79.50**
RMR	**62.31**	63.37	81.92			69.20
DAHR	61.62	**69.54**	**84.03**			**71.73**
RMR	64.93	56.59	72.51	85.87		69.98
DAHR	**65.12**	**57.94**	**74.19**	**85.97**		**70.81**
RMR	64.33	39.60	56.10	**82.19**	80.41	64.53
DAHR	**67.21**	**39.88**	**69.23**	81.46	**81.11**	**67.78**

Table 3: For each task with the parameter $M = 0.02 * |\mathcal{D}|$, we compared the evaluation results of our Depth Aware Hierarchical Replay (DAHR) method with Random Memory Replay (RMR) across different rounds of question-answer knowledge training end testing. Additionally, the average values for this row, which represent the final accuracy (FA) metric at different task rounds, are reported in the rightmost column.

k	$FA(\uparrow)$	$OA(\uparrow)$	$Fg(\downarrow)$
2	**66.29**	**71.90**	**21.05**
4	65.12	70.46	24.39
8	62.05	68.75	25.82

Table 4: The impact of different levels of branching factor on the experimental results. Parameter k determines the number of categories for each clustering, thus shaping the structure of the clustering tree and the depth of the sample.

c	$FA(\uparrow)$	$OA(\uparrow)$	$Fg(\downarrow)$
100	**65.77**	**71.90**	**22.38**
200	64.87	70.26	23.39
400	62.05	70.75	27.02

Table 5: The impact of different cluster size thresholds on the experimental results. Parameter c determines the size of leaf clusters and the depth of the tree, thereby shaping the structure of the classification tree.

the learning process of different methods in the experiment is the same, this issue does not affect our validation of the proposed method's effectiveness.

4.5. Ablation Study

In our proposed method, it is evident that the recursive classification of question-answer data in each round is a key step. In our ablation experiments, we examined the critical parameters for constructing the question-answer sample tree, namely k (the number of clusters in each classification round) and c (the threshold for cluster size beyond which further recursive classification is not performed). As shown in Tables 4 and 5, we considered $c = 100, c = 200, c = 400$ and $k = 2, k = 3, k = 4$, and present the experimental results for our main method with a sample replay rate of 2%.

Based on the results of our ablation experiments, we found that, for a 2% replay rate in the continuous learning framework, as discussed in Section 3.4, it is not advisable to have cluster sizes that are too large. Instead, they should be just right to ensure that each cluster independently selects approximately one central sample to reduce the information redundancy caused by overly similar replay samples, which could lead to losses for the model. In particular, when $c = 400$, there is a significant increase in the degree of forgetting(from 22.39 to 27.02), indicating that the selection of highly similar samples indirectly results in a meaningful reduction in the effective number of samples selected, rather than effectively retaining representations of previously learned question-answer knowledge.

Additionally, we observed that when considering different values of k, $k = 2$ yielded the best experimental results. We speculate that this is likely because, in the scenario with $k = 2$, the question-answer sample collections for each round can achieve the maximum depth within the classification tree. From an extreme standpoint, if we assume $k \to |\mathcal{D}|$, where n is the total number of samples and no further classification is needed, the depth of every sample becomes 1, essentially reducing the algorithm to the baseline. Therefore, we infer that our depth parameter is effective in expressing the importance of samples. When a sample attains a greater depth within this tree-like structure, it signifies that it holds a more central position within the overall sample characteristics.

5. Conclusion

In this work, we innovatively explore the scenario of continuous learning in knowledge based question answering tasks. Our proposed DAHR(Depth Aware Hierarchical Replay) method is concise and efficient, specifically tailored for selecting op-

timal replay samples from unlabeled QA data. Through comprehensive experiments, we have demonstrated that our approach significantly mitigates catastrophic forgetting. Additionally, the ablation studies conducted in our work provide consistent support for the motivation behind our proposed method, affirming its effectiveness.

6. Limitation

In this work, we have not yet discussed the impact of data augmentation. Data augmentation, as a common strategy, can often improve the performance of model methods. For our study, we have not explored whether combining our replay method with data augmentation could complement each other, potentially offering a more substantial mitigation of catastrophic forgetting. Additionally, we have used a single model or clustering algorithm in our experiments, such as Electra in the encoder and the K-means clustering algorithm. We have not extensively validated our approach using multiple parallel pre-trained model encoders and unsupervised clustering algorithms. We will explore them in future work.

7. Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer NetworkShenzhenMinistry of Educationthe Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), the Major Key Project of PCL for Experiments and Applications (PCL2021A06), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008).

8. Bibliographical References

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018a. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018b. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. 2019. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385.

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. 2019. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5138–5146.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. A survey of natural language generation. *ACM Comput. Surv.*, 55(8):173:1–173:38.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 260–269.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2023. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. *CoRR*, abs/2308.10855.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. 2019. Compacting, picking and growing

for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32.

Zixuan Ke, Bing Liu, Hu Xu, and Lei Shu. 2021. CLASSIC: Continual and contrastive learning of aspect sentiment classification tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6871–6883, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Lilly Kumari, Shengjie Wang, Tianyi Zhou, and Jeff A Bilmes. 2022. Retrospective adversarial replay for continual learning. *Advances in Neural Information Processing Systems*, 35:28530–28544.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems*, 30.

Yangning Li, Jiaoyan Chen, Yinghui Li, Yuejia Xiang, Xi Chen, and Hai-Tao Zheng. 2023a. Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Yangning Li, Jiaoyan Chen, Yinghui Li, Tianyu Yu, Xi Chen, and Hai-Tao Zheng. 2023b. Embracing ambiguity: Improving similarity-oriented tasks with contextual synonym knowledge. *Neurocomputing*, 555:126583.

Yangning Li, Yinghui Li, Xi Chen, Hai-Tao Zheng, and Ying Shen. 2023c. Active relation discovery: Towards general and label-aware open relation extraction. *Knowl. Based Syst.*, 282:111094.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023d. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. *CoRR*, abs/2308.06966.

Yinghui Li, Shulin Huang, Xinwei Zhang, Qingyu Zhou, Yangning Li, Ruiyang Liu, Yunbo Cao, Hai-Tao Zheng, and Ying Shen. 2023e. Automatic context pattern generation for entity set expansion. *IEEE Trans. Knowl. Data Eng.*, 35(12):12458–12469.

Yinghui Li, Yangning Li, Yuxin He, Tianyu Yu, Ying Shen, and Hai-Tao Zheng. 2022a. Contrastive learning with hard negative entities for entity set expansion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1077–1086. ACM.

Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Yangning Li, Shulin Huang, Ruiyang Liu, Chao Li, Yunbo Cao, and Haitao Zheng. 2022b. Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 238–249. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022c. The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3202–3213. Association for Computational Linguistics.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2021. Incremental knowledge based question answering. *arXiv preprint arXiv:2101.06938*.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7):100520.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Yangning Li, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11,*

2022, pages 576–589. Association for Computational Linguistics.

Martin Mundt, Yongwon Hong, Iuliia Pliushch, and Visvanathan Ramesh. 2023. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *Neural Networks*, 160:306–336.

Yujia Qin, Jiajie Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. ELLE: Efficient lifelong pre-training for emerging data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2789–2810, Dublin, Ireland. 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.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*.

Hippolyt Ritter, Aleksandar Botev, and David Barber. 2018. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31.

Amir Rosenfeld and John K Tsotsos. 2018. Incremental learning through deep adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):651–663.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 562–572.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, and Fei Huang. 2023. DAMO-NLP at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 2014–2028. Association for Computational Linguistics.

Tom Veniat, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2020. Efficient continual learning with modular networks and task-driven priors. *arXiv preprint arXiv:2012.12631*.

Peng Xing, Yinghui Li, Shirong Ma, Xinnian Liang, Haojing Huang, Yangning Li, Hai-Tao Zheng, Wenhao Jiang, and Ying Shen. 2024. Mitigating catastrophic forgetting in multi-domain chinese spelling correction by multi-stage knowledge transfer framework. *CoRR*, abs/2402.11422.

K Xu, L Wu, Z Wang, M Yu, L Chen, and V Sheinin. Exploiting rich syntactic information for semantic parsing with graph-to-sequence model. arxiv 2018. *arXiv preprint arXiv:1808.07624*.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6174–6189. Association for Computational Linguistics.

Wen-tau Yih, Xiaodong He, and Christopher Meek. 2014. Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 643–648.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. 2017. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng,

Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. *CoRR*, abs/2308.10529.

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

Yuanzhe Zhang, Kang Liu, Shizhu He, Guoliang Ji, Zhanyi Liu, Hua Wu, and Jun Zhao. 2016. Question answering over knowledge base with neural attention combining global knowledge information. *arXiv preprint arXiv:1606.00979*.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yingxiu Zhao, Yinhe Zheng, Zhiliang Tian, Chang Gao, Jian Sun, and Nevin L. Zhang. 2022. Prompt conditioned VAE: Enhancing generative replay for lifelong learning in task-oriented dialogue. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11153–11169, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

9. Language Resource References

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.