Memory Replay for Continual Medical Image Segmentation Through Atypical Sample Selection

Sutanu Bera(✉), Vinay Ummadi, Debashis Sen, Subhamoy Mandal, and Prabir Kumar Biswas

Indian Institute of Technology Kharagpur, Kharagpur, India
sutanu.bera@iitkgp.ac.in

Abstract. Medical image segmentation is critical for accurate diagnosis, treatment planning and disease monitoring. Existing deep learning-based segmentation models can suffer from catastrophic forgetting, especially when faced with varying patient populations and imaging protocols. Continual learning (CL) addresses this challenge by enabling the model to learn continuously from a stream of incoming data without the need to retrain from scratch. In this work, we propose a continual learning-based approach for medical image segmentation using a novel memory replay-based learning scheme. The approach uses a simple and effective algorithm for image selection to create the memory bank by ranking and selecting images based on their contribution to the learning process. We evaluate our proposed algorithm on three different problems and compare it with several baselines, showing significant improvements in performance. Our study highlights the potential of continual learning-based algorithms for medical image segmentation and underscores the importance of efficient sample selection in creating memory banks.

Keywords: Memory Replay · Continual Learning · Medical Image Segmentation

1 Introduction

Medical image segmentation is an essential task in clinical practice, enabling accurate diagnosis, treatment planning, and disease monitoring. However, existing medical segmentation methods often encounter challenges related to changes in imaging protocols and variations in patient populations. These challenges can significantly impact the performance and generalizability of segmentation models. For instance, a segmentation model trained on MRI images from a specific

S. Bera and V. Ummadi—These authors contributed equally to this work and share the first authorship.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_49.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 513–522, 2023.
https://doi.org/10.1007/978-3-031-43901-8_49

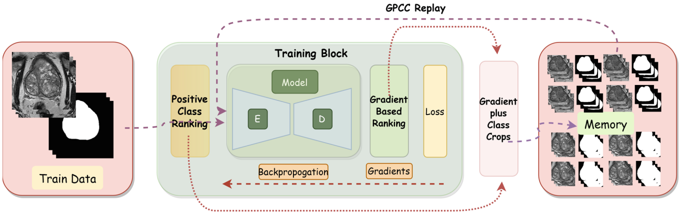

Fig. 1. Graphical summary of the concept of the proposed method. Positive Class Ranking and Gradient Based Ranking are computed online while training. Both rankings are used to create crops and stored in memory. The stored crops are to be used for replay while training on future datasets.

patient population may not perform well when applied to a different population with distinct demographic and clinical characteristics. Similarly, variations in imaging protocols, such as the use of different contrast agents or imaging parameters, can also affect the model's accuracy and reliability. To ensure accurate segmentation, it is necessary to retrain or fine-tune the model with current data before deploying it. However, this process often leads to catastrophic forgetting, where the model loses previously acquired knowledge while being trained on the current data. Catastrophic forgetting occurs due to the neural network's inability to learn from a continuous stream of data without disregarding previously learned information. Retraining the network using the complete training set, including both old and current data, is not always feasible or practical due to reasons such as the unavailability of old data or data privacy concerns. Moreover, training the network from scratch every time for every perturbation is a resource-intensive and time-sensitive process. Continual learning aims to address this limitation of catastrophic forgetting by enabling the model to learn continuously from a stream of incoming data without the need to retrain the model from scratch. Continual learning algorithms have gained significant interest lately for computer vision tasks like image denoising, super-resolution, and image classification. However, the development of efficient continual learning algorithms specifically designed for medical image segmentation has been largely overlooked in the literature. To address the above gap, our study proposes a continual learning based approach for medical image segmentation, which can be used to train any backbone network. In our approach, we leverage the recently proposed concept of the memory replay-based continual learning (MBCL) [6,15,17]. In MBCL, a memory buffer is used to store and replay previously learned data, enabling the model to retain important information while learning from new data. MBCL is, however, hampered by a few bottlenecks associated with medical images that pose a serious obstacle to its proper use in medical image segmentation. The efficiency of MBCL largely depends on the images stored in the memory bank [4,18], as the stored images must faithfully represent the previous task. It is known that medical image segmentation faces a major challenge of class imbalance. If an image with an under-representation of the positive class is stored in the memory bank, then it impedes the network from effectively remembering the previous task. In addition, not all medical images for training contribute equally to the learning process. So, images that pose greater challenges for the segmentation network should be saved in the memory bank. The importance of identifying atypical examples before creating memory banks cannot be overstated.

We propose a simple yet effective algorithm for image selection while creating the memory bank. Two different ranking mechanisms, which address the bottlenecks related to medical images discussed above, are proposed to rank all the images present in the training set. Then, images to be stored in the memory bank are selected using a combined ranking. Further, we suggest the cropping of the images around the organ of interest in order to minimize the size of the memory bank. An extensive evaluation is performed on three different problems, i.e., continual prostrate segmentation, continual hippocampus segmentation, and task incremental segmentation of the prostate, hippocampus and spleen. We consider several baselines including EWC [12], L2 regularization-based [10], and representation learning-based [16]. Our method is found to outperform the conventional MBCL, and all the baseline mentioned above by a significant margin, creating a new benchmark for continual learning-based medical image segmentation.

2 Proposed Methodology

We are given a sequential stream of images from K sites, which are sequentially used to train a segmentation model. In a round $k \in [1, K]$ of this continual learning procedure, we can only obtain images and ground truths $\{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k}$ from a new incoming site (dataset) D_k without access to old data from previous sites D_{k-1}. Due to catastrophic forgetting, this type of sequential learning results in a drop in performance for all the previous sites ($\leq D_{k-1}$) after training with images from D_k site as the parameters of the previous site or task are overwritten while learning a new task. In naive memory replay-based continual learning, a memory buffer, $\mathcal{M}$, is used to store a small number of examples of past sites ($\leq D_{k-1}$), which can be used to train the model along with the new data. Unlike other tasks like image classification or image restoration, for downstream tasks like medical image segmentation, the selection of images for storing in the $\mathcal{M}$ is very crucial. A medical image segmentation (like hippocampus segmentation) approach typically has a very small target organ. It is very likely that randomly selected images for storage in the $\mathcal{M}$ will have an under-representation of the positive (hippocampus) class. Further, the contribution of each training sample is not equal towards the learning, as a network usually learns more from examples that are challenging to segment. Based on the above observations, we propose two image ranking schemes to sort the images for storing in $\mathcal{M}$ (see Fig. 1).

2.1 Positive Class Based Ranking (PCR)

In this ranking scheme, we rank an input image volume according to the percentage of voxels corresponding to the positive class available in the volume.

Let $cr_{k,i}$ be the positive class based ranking score for the sample input-ground truth pair $(x_{k,i}, y_{k,i})$ from the dataset D_k. We use the ground truth label $y_{k,i}$ to calculate the score of each volume. Let H, W and D respectively be the height, width and number of slices in the 3D volume of $y_{k,i}$. The voxel value at location (h, w, d) in the ground truth label $y_{k,i}$ is represented by $y_{k,i_{h,w,d}} \in 0, 1$. If the voxel value at a location (h, w, d) is equal to 1, then the voxel belongs to the positive class. Let us use $|y_{k,i}|$ to represent the total number of voxels in the 3D volume. For a sample pair $(x_{k,i}, y_{k,i})$, $cr_{k,i}$ is computed as follows:

$$cr_{k,i} = \frac{\sum_{h=0}^{H-1}\sum_{w=o}^{W-1}\sum_{d=0}^{D-1}(y_{k,i_{h,w,d}})}{|y_{k,i}|} \tag{1}$$

The rationale behind this ranking scheme is that by selecting volumes with a higher positive class occupancy, we can minimize the risk of underrepresentation of the positive class leading to a continuously trained network that remembers previous tasks more faithfully.

2.2 Gradient Based Ranking (GBR)

Here, we intend to identify the examples which the segmentation network finds hard to segment. In this ranking, we leverage the relationship between example difficulty and gradient variance, which has been previously observed in other studies [2]. Specifically, the neural network tends to encounter high gradients on examples that are hard to learn. With this motivation, we devise a simple method for calculating the score of every sample based on gradients, outlined in Algorithm 1. This algorithm enables the calculation of the gradient-based score during the network's training in real-time. In the algorithm, $f(\theta)$ refers to the image segmentation network with parameter θ and $L_k(\theta) = \frac{1}{n_k}\sum_{i=1}^{n_k} l(f(\theta; x_{k,i}), y_{k,i})$ denotes the loss function employed to train $f(\theta)$. $gr_{k,i}$ is the *gradient based score* assigned to a sample $D_{k,i}$, where $i \in [1, n_k]$ and $k \in [1, K]$. The corresponding

Algorithm 1. Gradient-based Sample Score

```
for $k = 1 \ldots K$ do                                        ▷ Incoming dataset $D_k$
    $n_k = |D_k|$
    $\theta \leftarrow \theta_0$                                ▷ Parameters randomly initialized
    $pg \leftarrow 0$                                           ▷ Prevoius gradients initially 0
    for $epoch = 1 \ldots E$ do                                 ▷ Traning for $E$ epochs
        for $i = 1 \ldots n_k$ do
            $cg = \frac{\partial L_k(f(\theta; x_{k,i}), y_{k,i})}{\partial \theta}$    ▷ Loss function gradients
            $gr_{k,i} \mathrel{+}= |pg_{k,i} - cg|$             ▷ Absoulte gradient difference for each sample
            $pg_{k,i} \leftarrow cg$                            ▷ Set current gradients $cg$ to previous gradients
            $\theta \leftarrow \theta - \eta * cg$              ▷ $\eta$ is learning rate
        end for
    end for
end for
```

gradients for the current sample are represented by cg, and $gr_{k,i}$ accumulates the absolute gradient difference between cg and previous gradients $pg_{k,i}$ for all training epochs. Essentially, a high value of $gr_{k,i}$ signifies a large gradient variance and implies that the example is difficult.

2.3 Gradient Plus Class Score (GPCS) Sampling for Memory

Once the score $gr_{k,i}$ and $cr_{k,i}$ are available for a dataset D_k, they are normalized to $[0, 1]$ range and stored for future use. If p samples are to be selected for the replay memory $\mathcal{M}$, then $\frac{p}{2}$ will be selected using $gr_{k,i}$ and other $\frac{p}{2}$ using $cr_{k,i}$. However, we do not store entire image volumes in the memory. We propose a straightforward yet effective strategy to optimally utilize the memory bank size for continual medical segmentation tasks. Typically, the organ of interest in these tasks occupies only a small portion of the entire image, resulting in significant memory wastage when storing the complete volume. For example, the hippocampus, which is a common region of interest in medical image segmentation, occupies less than 0.5% of the total area. We propose to store only a volumetric crop of the sequence where the region of interest is present rather than the complete volume. This enables us to make more efficient use of the memory, allowing significant amounts of memory for storing additional crops within a given memory capacity. Thus, we can reduce memory wastage and optimize memory usage for medical image segmentation.

Consider that an image-label pair $(x_{k,i}, y_{k,i})$ from a dataset D_k has dimensions $H \times W \times D$ (height×width×sequence length). Instead of a complete image-label pair, a crop-sequence of dimension $h_c \times w_c \times d_c$ with $h_c \leq H, w_c \leq W, d_c \leq D$ is stored. We also use an additional hyperparameter, *foreground background ratio* $fbr \in [0, 1]$ to store some background areas, as described below

$$fbr = \frac{Number\ of\ crops\ having\ RoI}{Number\ of\ crops\ not\ having\ RoI} \tag{2}$$

Using only foreground regions in problems such as task incremental learning may result in a high degree of false positive segmentation. In such cases, having a few crops of background regions helps in reducing the forgetting in background regions as discussed in Sect. 3.3.

3 Experimental Details

3.1 Datasets

We conduct experiments to evaluate the effectiveness of our methods in two different types of incremental learning tasks. To this end, we use seven openly available datasets for binary segmentation tasks: four prostate datasets (Prostate158 [1], NCI-ISBI [3], Promise12 [14], and Decathlon [5]), two hippocampus datasets (Drayd [8] and HarP [7]), and one Spleen dataset from Decathlon [5]. The first set of experiments involved domain incremental prostate segmentation, with the datasets being trained in the following order: Prostate158 → ISBI → Promise12 → Decathlon. The second experiment involved domain incremental hippocampus segmentation, with the datasets being trained in the order HarP → Drayd. Finally, we conducted task incremental segmentation for three organs - prostate, spleen, and hippocampus - following the sequence: Promise12 (prostate) → MSD (spleen) → Drayd (hippocampus).

3.2 Evaluation Metrics

To evaluate the effectiveness of our proposed methods against baselines, we use the segmentation evaluation metric Dice Similarity Coefficient (DSC) along with standard continual learning (CL) metrics. The CL metrics [9] comprise Average Accuracy (ACC), Backward Transfer (BWT) [13], and Average Forgetting (AFGT) [19]. BWT measures the model's ability to apply newly learned knowledge to previously learned tasks, While AFGT measures the model's retention of previously learned knowledge after learning a new task.

3.3 Training Details

In this work, we consider a simple segmentation backbone network UNet, which is widely used in medical image segmentation. Both the proposed and baseline methods were used to train a Residual UNet [11]. All the methods for comparison (Tables 2, 3 and 4), except for Sequential (SGD), are trained using the Adam optimizer with a learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.00001. Sequential(SGD) employs an SGD optimizer with the same hyperparameters as Adam. The training loss used is DiceCELoss, which combines Dice loss and Cross Entropy loss. During task incremental segmentation training using GPCC, the fbr parameter is set to 0.7, while the default value of 1.0 is used for other tasks. In parts of our GPCC experiments, where we examine continual prostate segmentation as well as continual prostate, spleen, and hippocampus segmentation, both the height (h_c) and width (w_c) of each volume are fixed to 160. In continual hippocampus segmentation experiments, these values are reduced to 128. Note that although we perform the experiments using UNet network, any sophisticated network like VNet, or DeepMedic can also be trained continually using our method.

4 Results and Discussion

Ablation Study of Our Atypical Sample Selection:[1] In order to evaluate the effectiveness of every module proposed in our work, an ablation study is conducted. The results of the analysis shown in Table 1 indicates that randomly storing samples leads to a significant decrease in performance during in the earlier trained domains due to insufficient representation of the domain distributions. Both PCR and GBR shows improvements in ACC over random replay

[1] All experimental values reported here are the average over four random seeds.

Table 1. Ablation study on continual prostate segmentation. The best DSC, ACC and AFGT are presented in bold. GPCC gives the best performance in general.

Learning Method	Dataset wise DSC scores (%)				CL Metrics		
	Prostate158 (↑)	NCI-ISBI (↑)	Promise12 (↑)	Decathlon (↑)	ACC (↑)	BWT (↑)	AFGT (↓)
Random Replay(3)	67.5 ± 3.3	85.7 ± 3.0	74.8 ± 10.3	85.7 ± 1.3	78.4 ± 2.8	−6.8 ± 3.2	7.7 ± 2.9
PCR Replay(3)	**82.9** ± 1.4	81.6 ± 1.6	77.5 ± 3.8	82.8 ± 1.8	81.2 ± 2.5	−3.6 ± 1.3	4.6 ± 1.4
GBR Replay(3)	82.4 ± 1.2	83.9 ± 1.8	78.6 ± 3.4	81.6 ± 1.4	81.6 ± 2.2	−2.1 ± 0.9	2.7 ± 0.7
GPCC Replay(6)	82.1 ± 1.8	**85.9** ± 2.3	**80.8** ± 5.2	**86.3** ± 1.5	**82.8** ± 1.3	−1.6 ± 0.7	**2.3** ± 0.3

by **2.8**% and **3.2**%, respectively. GPCC provides further enhancements across all domains, resulting in a final accuracy gain of **4.4**% over the random replay.

Comparison with Baselines: We consider several benchmark continual learning algorithms including L2 Regularization, EWC and Representation Replay as our baselines. First, we compare the performance of our method on continual prostate segmentation tasks. The objective comparison among different methods on this task is shown in Table 2. The proposed methods perform on par or better with joint learning, which is considered an upper bound in continual learning settings. Compared with Sequential (SGD) and Random Replay(3)[2], our method with GPCC(6) shows an ACC improvement of **9.2**% and **4.4**% respectively. With a memory footprint, GPCC with six crops of 160 × 160 × D is **65**% lighter compared to Random Replay (3) with an average image size of 384 × 384 × D. Next, the objective comparison of different methods of continual hippocampus segmentation is shown in Table 3. This task poses a significant challenge due to the small region of interest (RoI) in whole brain MRI scans. Our proposed approach outperforms all other baseline methods in terms of CL metrics. When comparing GPCC Replay with six crops of 128 × 128 × D images to Random Replay(3) using full-size images of 233 × 189 × D, we find that GPCC Replay is more memory-efficient consuming **26**% less memory while still achieving an **1.1**% ACC performance improvement. While some baselines showed higher DSC scores on the second domain, a high value of AFGT indicates their inability to

Table 2. Objective comparison of different methods on continual prostate segmentation. The best DSC (in continual), ACC and AFGT are presented in bold.

Learning Method	Dataset wise DSC scores(%)				CL Metrics		
	Prostate158 (↑)	NCI-ISBI (↑)	Promise12 (↑)	Decathlon (↑)	ACC (↑)	BWT (↑)	AFGT (↓)
Individual	82.3 ± 1.3	79.4 ± 1.6	87.5 ± 2.3	81.5 ± 2.1	–	–	–
Joint	83.5 ± 1.4	86.8 ± 1.8	82.6 ± 2.6	86.4 ± 1.5	–	–	–
Sequential(SGD)	61.8 ± 2.7	82.4 ± 2.4	66.4 ± 9.4	83.6 ± 1.5	73.6 ± 2.3	−10.6 ± 1.4	11.8 ± 1.2
L2 Regularization	36.7 ± 1.6	59.8 ± 5.2	59.3 ± 10.6	78.9 ± 4.8	58.7 ± 3.8	0.13 ± 0.0	**0.1** ± 0.0
EWC	61.6 ± 3.2	82.6 ± 2.6	64.2 ± 10.9	85.3 ± 1.3	73.4 ± 3.1	−4.6 ± 1.7	6.07 ± 1.8
Representation Replay	60.2 ± 2.3	65.6 ± 3.2	58.3 ± 6.7	72.1 ± 2.6	64.0 ± 6.2	−10.3 ± 2.5	11.3 ± 3.4
Random Replay(3)	67.5 ± 3.3	85.7 ± 3.0	74.8 ± 10.3	85.7 ± 1.3	78.4 ± 2.8	−6.8 ± 3.2	7.7 ± 2.9
GPCC Replay(6)	**82.1** ± 1.8	**85.9** ± 2.3	**80.8** ± 5.2	**86.3** ± 1.5	**82.8** ± 1.3	−1.6 ± 0.7	2.3 ± 0.3

[2] 3 (the number within brackets) is number of volumes stored in memory buffer.

Table 3. Objective comparison of different methods on continual hippocampus segmentation. The best DSC (in continual), ACC and AFGT are presented in bold.

Learning Method	Dataset wise DSC scores (%)		CL Metrics		
	HarP (↑)	Drayd (↑)	ACC (↑)	BWT (↑)	AFGT (↓)
Individual	80.5 ± 2.0	79.1 ± 1.1	-	-	-
Joint	82.1 ± 0.9	80.2 ± 0.8	-	-	-
Sequential(SGD)	55.2 ± 2.3	76.5 ± 1.2	63.4 ± 1.5	−15 ± 2.1	15.3 ± 1.9
EWC	71.3 ± 0.8	83.4 ± 0.7	77.3 ± 1.1	−7.6 ± 1.2	7.8 ± 1.3
Representation Replay	70.8 ± 1.2	**83.6** ± 1.5	77.2 ± 1.4	−8.2 ± 1.3	8.3 ± 1.5
L2 Regularization	75.2 ± 0.6	79.1 ± 0.8	77.3 ± 0.7	−2.2 ± 0.9	2.2 ± 0.9
Random Replay(3)	**83.5** ± 1.8	73.5 ± 1.7	78.7 ± 1.4	−4.5 ± 0.8	4.5 ± 0.9
GPCC Replay (6)	83.0 ± 0.7	77.1 ± 0.4	**79.8** ± 0.6	−2.1 ± 0.5	**2.1** ± 0.5

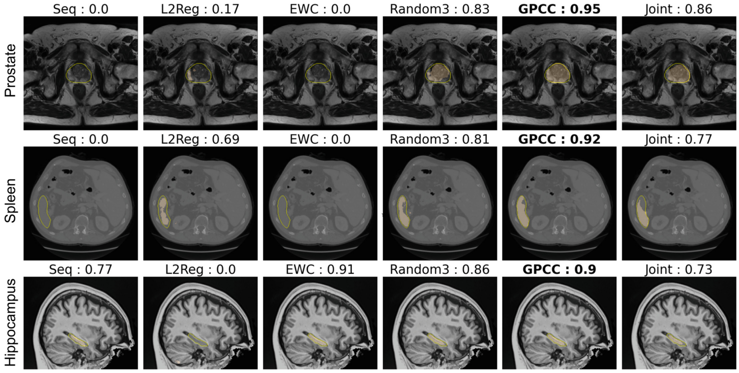

Fig. 2. Qualitative results for task incremental segmentation of prostate, spleen, and hippocampus using the methods in Table 4. Ground truths are in yellow borders and predictions are in peach. The bolded method has the highest DSC score.

retain previously learned knowledge, which suggests limitations in their ability to perform continual training of the backbone.

We finally assess the performance of our method on an even more challenging task of incremental learning segmentation. This involves continuously training a single model to accurately segment various organs while incorporating new organs as segmentation targets during each episode. Utilizing a single model for segmenting multiple organs offers potential advantages, particularly when there are constraints such as limited annotation data that hinder joint training. Task or class incremental learning becomes invaluable in such scenarios, as it allows us to incorporate new organs as segmentation targets without requiring a complete retraining process. To investigate the feasibility of this concept, we dedicate a section in our study to experimental analysis. The results of this analysis are presented in Table 4. We find that GPCC replay(12) shows substantial improvement by increasing the ACC score by **45.4**% over Sequential(SGD) and achieves

Table 4. Comparison of Baselines for Task Incremental Segmentation. The best DSC (in continual), ACC and AFGT are presented in bold.

Learning Method	Dataset wise DSC scores (%)			CL Metrics		
	Promise12 Prostate (↑)	MSD Spleen(↑)	Drayd Hippocampus(↑)	ACC (↑)	BWT (↑)	AFGT (↓)
Individual	81.5 ± 2.1	91.5 ± 1.6	79.1 ± 1.1	-	-	-
Joint	80.5 ± 0.7	80.5 ± 0.8	79.1 ± 0.6	-	-	-
Sequential(SGD)	0.0 ± 0.0	0.0 ± 0.0	67.5 ± 2.4	37.9 ± 1.7	−80 ± 1.8	80 ± 1.7
EWC	0.0 ± 0.0	0.0 ± 0.0	**80.8** ± 1.4	42.6 ± 1.3	−87.3 ± 1.6	87.2 ± 1.7
Representation Replay	0.0 ± 0.0	0.0 ± 0.0	80.1 ± 2.4	42.2 ± 1.3	−86.3 ± 1.2	86.3 ± 1.0
L2 Regularization	49.7 ± 3.2	56.7 ± 2.5	15.2 ± 3.5	49.4 ± 3.3	−17.2 ± 2.6	17.2 ± 2.6
Random Replay(3)	66.5 ± 2.5	83.7 ± 2.3	80.1 ± 1.5	78.3 ± 1.3	−13.1 ± 1.1	13.1 ± 1.1
GPCC Replay (12)	**78.4** ± 1.6	**87.9** ± 1.4	80.5 ± 1.5	**83.3** ± 1.3	−4.2 ± 1.2	**4.2** ± 1.2

a **5**% increase in ACC compared to Random Replay(3), outperforming all the baselines. In this case, GPCC Replay is also lighter in memory consumption by up to **32**%.

Visually analyzing the predictions for a test sample in Fig. 2 from the task of incremental segmentation, L2 regression and Random Replay(3) are seen to produce only partial segmentation of RoI. On the other hand, GPCC predictions outperform joint learning and are very close to Ground Truths, with a DSC score $\geq$ **90**%. More visual comparison among different methods is given in the supplementary.

5 Conclusion

This paper proposes a novel approach to address the challenge of catastrophic forgetting in medical image segmentation using continual learning. The paper presents a memory replay-based continual learning paradigm that enables the model to learn continuously from a stream of incoming data without the need to retrain from scratch. The proposed algorithm includes an effective image selection method that ranks and selects images based on their contribution to the learning process and faithful representation of the task. The study evaluates the proposed algorithm on three different problems and demonstrates significant performance improvements compared to several relevant baselines.

References

1. Adams, L.C., et al.: Prostate158-an expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. Comput. Biol. Med. **148**, 105817 (2022)
2. Agarwal, C., D'souza, D., Hooker, S.: Estimating example difficulty using variance of gradients. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10368–10378 (2022)

3. NCI-ISBI 2013 challenge: automated segmentation of prostate structures. https://wiki.cancerimagingarchive.net/display/Public/NCI-ISBI+2013+Challenge+-+Automated+Segmentation+of+Prostate+Structures
4. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
5. Antonelli, M., et al.: The medical segmentation decathlon. Nat. Commun. **13**(1), 4128 (2022)
6. Balaji, Y., Farajtabar, M., Yin, D., Mott, A., Li, A.: The effectiveness of memory replay in large scale continual learning. arXiv preprint arXiv:2010.02418 (2020)
7. Boccardi, M., et al.: Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimer's Dement. **11**(2), 175–183 (2015)
8. Denovellis, E., et al.: Data from: hippocampal replay of experience at real-world speeds (2021). https://doi.org/10.7272/Q61N7ZC3
9. Díaz-Rodríguez, N., Lomonaco, V., Filliat, D., Maltoni, D.: Don't forget, there is more than forgetting: new metrics for continual learning. arXiv preprint arXiv:1810.13166 (2018)
10. Hsu, Y.C., Liu, Y.C., Ramasamy, A., Kira, Z.: Re-evaluating continual learning scenarios: a categorization and case for strong baselines. arXiv preprint arXiv:1810.12488 (2018)
11. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A.: Left-ventricle quantification using residual U-Net. In: Pop, M., et al. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 371–380. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_40
12. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. **114**(13), 3521–3526 (2017)
13. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2935–2947 (2017)
14. Litjens, G., et al.: Evaluation of prostate segmentation algorithms for MRI: the promise12 challenge. Med. Image Anal. **18**(2), 359–373 (2014)
15. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
16. Pellegrini, L., Graffieti, G., Lomonaco, V., Maltoni, D.: Latent replay for real-time continual learning. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10203–10209 (2020). https://doi.org/10.1109/IROS45743.2020.9341460
17. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
18. Tiwari, R., Killamsetty, K., Iyer, R., Shenoy, P.: GCR: gradient coreset based replay buffer selection for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 99–108 (2022)
19. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International Conference on Machine Learning, pp. 3987–3995. PMLR (2017)