AN INCREMENTAL LEARNING APPROACH FOR SUSTAINABLE REGIONAL ISOLATION AND INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans are capable of acquiring new knowledge on a constant basis, while integrating and optimizing old knowledge without forgetting them. This is mainly attributed to the human brain's ability of partitioned learning and memory replay. In this paper, we simulate this ability and propose an incremental learning network of Sustainable Regional Isolation and Integration (SRII). SRII consists of two phases, regional isolation and regional integration, which are iterated to achieve continuous incremental class learning. Regional isolation isolates new learning processes to avoid interfering with existing knowledge, while regional integration introduces knowledge distillation and margin loss regularization term, knowledge distillation to transfer replay knowledge for alleviating catastrophic forgetting, margin loss regularization term to clarify the boundaries of new and old knowledge for alleviating recency bias. and margin loss regularization term to clarify the boundaries of new and old knowledge in regional integration, which alleviates catastrophic forgetting and recency bias. Experimental results on the CIFAR100 and miniImageNet datasets demonstrate that SRII outperforms the state-of-the-arts to avoid catastrophic forgetting. In all 5-stage and 10-stage incremental settings, SRII outperforms the baseline and achieves at least $5.27\%+$ average accuracy improvement. Our source code is available at `https://github.com/Wuziyi123/SRII`.

1 INTRODUCTION

Humans can learn incrementally, acquire new knowledge constantly, and integrate and optimize old knowledge without forgetting it. This extraordinary capacity for continuous learning is linked to the partitioned learning structure of the human brain: the hippocampal system and neocortical system. The hippocampus is usually associated with immediate memory, such as short-term memory, and allows for rapid learning of new information. The neocortex is usually involved with retaining and retrieving distant memories and can form long-term memory (Parisi et al., 2019). The coordination between the hippocampus and neocortex can minimize the interference between knowledge in the process of rapid learning, guaranteeing ongoing learning (O'Reilly et al., 2014).

Human's continuous learning capacity also benefits from the memory replay of the hippocampal system. There are mounting evidences that the hippocampus can rapidly switch between encoding and replay modes, accelerating the creation and consolidation of long-term memories in the neocortex. First, the hippocampus rapidly encodes memories while awake. Subsequently, during offline periods (i.e., sleep), the hippocampus reactivates memory trajectories through memory replay, facilitating the transfer of recent memory to the cortex to enter the long-term storage state (O'Neill et al., 2010).

Inspired by the above-mentioned partitioned learning and memory replay of the human brain, this paper constructs a neural network model named "sustainable regional isolation and integration" (SRII), which can protect long-term memory and mitigate "catastrophic forgetting" (Parisi et al., 2019; Carpenter & Grossberg, 1988; McCloskey & Cohen, 1989; French, 1999; Polikar et al., 2001; Dai et al., 2020) during continuous learning. It is realized through the following two aspects.

Firstly, inspired by the partitioned learning ability of brain, SRII isolates the old knowledge storage region and the new learning region in each incremental session to minimize the interference between old and new knowledge. The new learning region is used to learn new knowledge at each streaming session. In this paper, the dual storage partitions of incremental sessions, i.e., the old knowledge region and the new learning region, are continuously allocated through a two-stage cycle of isolation

and integration, as shown in Figure 1. In the isolation phase, half of the SRII's capacity is devoted to storing old knowledge and the other half to acquiring new knowledge. In the integration phase, the new and old knowledge partitions are merged. Then, it enters the next isolation phase of the cycle to isolate the new dual storage partitions.

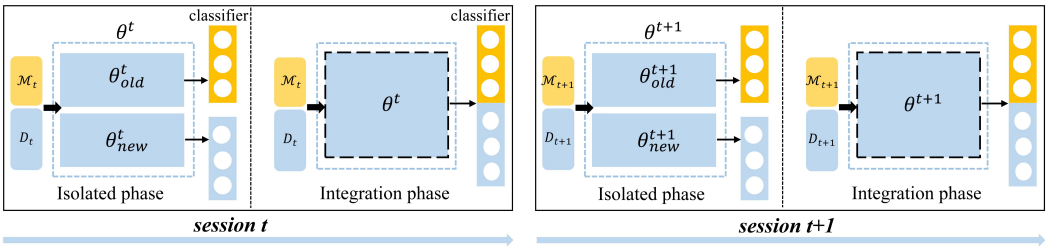

Figure 1: Incremental learning with continuous isolation and integration. t denotes the phase of the session, $\mathcal{M}$ and D denote the samples of old and new knowledge, respectively. θ_{old} stores old knowledge, and θ_{new} is used for training of new knowledge, where $\theta = \{\theta_{old}, \theta_{new}\}$, $\mathcal{M}_{t+1} = \{\mathcal{M}_t, D_t\}$. For incremental session t, in the isolation phase, the new and old knowledge are learned in isolation, and the classifiers of the new and old knowledge output independently. In the integration phase, the regions of old and new knowledge are integrated and the classifier outputs uniformly. Then, the model enters incremental session $t+1$. It will perform isolation and integration again.

Secondly, SRII imitates the hippocampus for memory replay, focusing on two aspects of memory replay. One is to consolidate old knowledge during replay, and the other is to clear the boundaries of new and old knowledge to alleviate "recency bias" (Hou et al., 2019; Smith et al., 2021; Wu et al., 2019). To retain old knowledge, we use neural networks to encode new knowledge and applying knowledge distillation (Hou et al., 2019; Smith et al., 2021; Wang et al., 2020; Mai et al., 2021; Kim & Choi, 2021; Castro et al., 2018; Ramapuram et al., 2020; Liu et al., 2018; 2021) to constrain the transfer of old knowledge. To alleviate the recency bias, SRII uses margin loss to suppress the behavior of mistakenly classifying the old classes into the new classes in the memory playback process, so as to improve the network's ability to distinguish between the new and old class decision boundaries.

In summary, our contributions are as follows:

- Taking ideas from partitioned learning, we came up with a two-stage learning mechanism that includes isolation and integration. The mechanism uses the parameter separation operation to isolate the overall parameter space for learning, and then uses an integration method to form a unified representation of the parameter space again. The two are iterated alternately to achieve continuous class incremental learning.
- To alleviate recency bias, the margin loss is proposed to encourage "inter-class separation and intra-class aggregation", which improves the discriminatory ability of the classifier between new and old class decision boundaries.
- Finally, to validate the effectiveness of SRII, ablation experiments were systematically performed on CIFAR100 (Krizhevsky et al., 2009) and miniImageNet (Vinyals et al., 2016) and compared with the-state-of-the-art incremental learning methods. Our method obtains an average accuracy improvement of more than $5.27\%+$.

This paper is organized as follows: Section 2 introduces the current situation of incremental learning research, summarizes and compares the advantages of various methods as well as their limitations. Section 3 first outlines the general structure of SRII, followed by a detailed description of how SRII works and how it is trained, including region isolation, region integration, network training, and loss function design. Section A is devoted to experiments, which consists of three parts: experimental setup, comparison with the state-of-the-arts, and ablation experiments. Section 4 summarizes the strengths and weaknesses of the work and thoroughly discusses future work.

2 Related Work

2.1 REGULARIZATION AND KNOWLEDGE DISTILLATION

It uses the redundancy of parameters to regularize the historical information that needs to be retained before learning new data to balance recollection and update. Generally, the old knowledge is protected from being covered by the new knowledge by imposing constraints on the loss function of the new task (Li & Hoiem, 2017; Kirkpatrick et al., 2017). The Learning without forgetting (LwF) (Li & Hoiem, 2017) proposed by Li, Zhizhong et al. introduces the distillation loss (Hinton et al., 2015) of the new model output into the loss function and then trains the model on the new task by fine-tuning (e.g., (Howard & Ruder, 2018)) to avoid performance degradation of the new model on the old task due to the excessive adjustment of the old model's parameters. The feature encoder of Encoder based lifelong learning (EBLL) (Rannen et al., 2017) records key information for each task with training, and the reconstruction loss on this encoder can be used to preserve historical information when new learning occurs. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) and its upgraded versions (Liu et al., 2018; Schwarz et al., 2018) use the Fisher information matrix to constrain network parameters and limit the model's forgetting. Regrettably, the EWC and its enhanced variants are incapable of reconciling new and old tasks that compete for crucial parameters. In 2017, Zenke et al. (2017) developed an IS technique to calculate the importance of weights online, where the weight parameters change dynamically based on its contribution to the loss during training. The Memory Aware Synapses (MAS) (Aljundi et al., 2018) technique estimates the relevance of weights by calculating the effect of parameters change on model output. More efforts (e.g., Learning without memorizing (LwM) (Dhar et al., 2019), Deep Model Consolidation (DMC) (Zhang et al., 2020), Big Incremental learning (Bic) (Wu et al., 2019)) have used distillation loss to prevent forgetting in recent years. As in Dhar et al. (2019), a gradient flow information distillation method based on attention mechanism mapping was presented. In conclusion, the domains of knowledge distillation and incremental learning have generally converged. In this paper we use knowledge distillation to promote knowledge transfer among numerous incremental processes.

2.2 REPLAY

One of the most successful strategies to prevent "catastrophic forgetting" is to remind a model by keeping old knowledge in incremental learning (Kim & Choi, 2021). Therefore, sample replay was applied to a multitude of advanced incremental learning methods, as is the case in this paper. Nonetheless, the research in (Hou et al., 2019; Smith et al., 2021) indicates that there is a substantial "recency bias" problem in the process of memory replay due to the imbalance between old and new data in the sample replay method and the softmax classifier's inadequate ability in the incremental learning scenario. Recent research methodologies, such as Bic (Wu et al., 2019), Learning a unified classifier incrementally via rebalancing (LUCIR) (Hou et al., 2019), and End to end incremental learning (E2E) (Castro et al., 2018), are prone to focusing on "recency bias". Additionally, Supervised contrastive replay (SCR) (Mai et al., 2021) proposed by Zheda Mai also believes that the "recency bias" caused by the commonly used softmax classifier in incremental learning remains an unsolved issue. In our study, firstly we pay close attention to this issue by normalizing the weight vectors of the classifier in order to reduce the bias caused by the size of values[1]. Secondly, a regular term margin loss is created to improve classification performance by learning the characteristics of the current task while avoiding bias towards the feature representation of the nearest task. Unlike Always be dreaming: A new approach for data-free class incremental learning (ABD) (Smith et al., 2021), which distinguishes new and old tasks by computing the local cross-entropy loss on the new task, our approach is inspired by the class margin loss design in LUCIR Hou et al. (2019), which maximizes the spatial distance between the new and old task features.

2.3 DYNAMIC STRUCTURE

As examples of dynamic structural approaches, progressive frameworks have attracted wide attention (Hung et al., 2019; Rusu et al., 2016; Yoon et al., 2017). These studies enable incrementally building deep incremental models by scaling network capacity to some extent while retaining network structure compactness. Considering that the structural complexity of the progressive framework grows linearly, several incremental learning studies in recent years have begun to try to tap the potential of fixed-capacity networks by establishing a pre-defined, continuous extension for future tasks. For instance,

[1]Specifically, the weight vectors of the classifier is normalized with the L2 normalization, as seen in Figure 2.

Bayesian nonparametric weight factorization for continual learning (BNWF) (Mehta et al., 2020) allocates tensor space fairly based on the number of tasks in a fixed-capacity network. Learn-prune-share (LPS) (Wang et al., 2020) uses mask and pruning strategies extensively to isolate task space, but its trainable space for new tasks decreases gradually. Furthermore, the method of partition by workload makes it difficult to integrate the region of LPS. Split-and-Bridge (S&B) (Kim & Choi, 2021) also adopts the dynamic structure of fixed-capacity to resist "catastrophic forgetting", with distillation loss as an aid to migrate old knowledge. However, S&B doesn't take into account recency bias, and the isolation mechanism for tasks is complicated, which isn't good for training.

Our work also strives to maximize the potential of fixed-capacity networks by preserving the isolation mechanism of tasks used in incremental learning methods such as S&B (Kim & Choi, 2021), BNWF (Mehta et al., 2020), and LPS (Wang et al., 2020). By continuously allocating the storage areas of new and old knowledge equally, we isolate the learning environment of new and old knowledge. The continuous integration process is also limited to areas of new and old knowledge, as seen in Figure 1. This two-stage continuous isolation and integration setup imparts scalability to the fixed-capacity network, which not only reduces learning interference and storage consumption, but also allows for the equitable distribution of trainable space for all incremental tasks. Section A.3 confirms SRII's validity experimentally, demonstrating that it outperforms S&B.

3 METHODOLOGY

In this section, we present a two-stage approach to the class incremental learning problem in a single network, namely "regional isolation" and "regional integration". As shown in Figure 2.

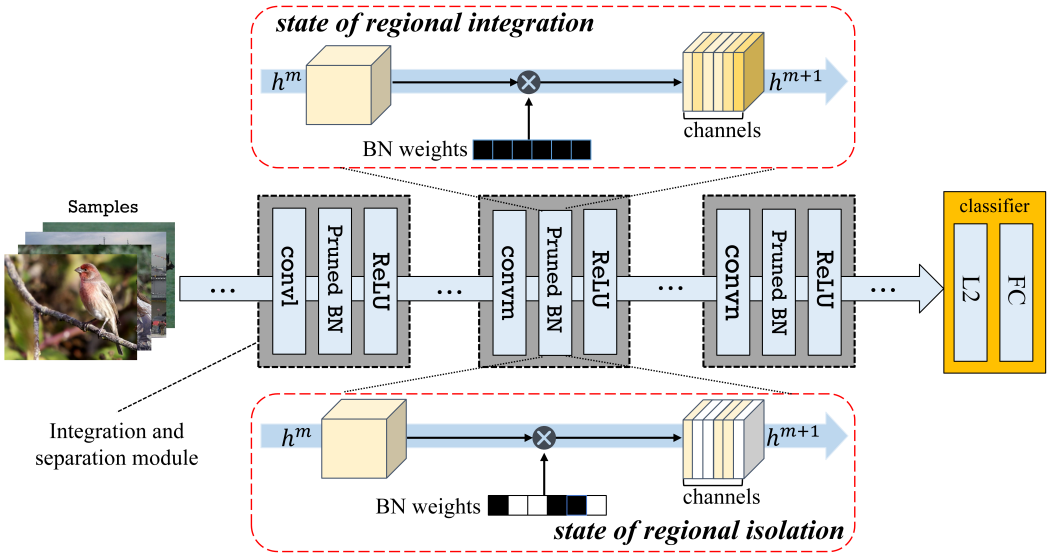

Figure 2: Structure of SRII. L2 stands for L2 normalization, FC denotes the fully connected layer. The backbone of the network is made up of a series of "integration and separation modules" that have two states: "regional isolation" and "regional integration". During the state of regional isolation, the m-th convolutional layer h^m is processed by the weights of the pruned BN layer to appear as channel separation. The yellow channels store the old knowledge, the white channels are free and ready for new learning, and the new knowledge region is isolated from the old knowledge region. During the state of regional integration, all BN weights are available and the network uses all channels for learning.

The following subsections describe the key processes of SRII, regional isolation and regional integration, followed by the network training process and loss function.

3.1 SUSTAINABLE REGIONAL ISOLATION

The objective of regional isolation is to learn new knowledge as independently as possible based on retaining original knowledge, as shown in Figure 3. It consists of **Regional isolation with BN sparsity** and **Forming new knowledge**. The former is to reserve channels for new knowledge to avoid confusion with old knowledge, while the latter is to acquire new knowledge with the reserved channels.

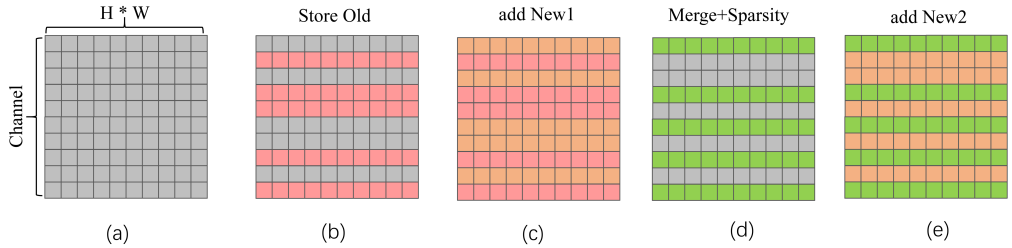

Figure 3: Continuous learning by regional isolation. The grey regions reflect the unoccupied channels accessible for new learning. (a) is the initialization state, all idle. After training, 50% of the channels in (b) are selected to store old knowledge, while the remaining 50% of the grey channels are idle. (c) use the idle area of (b) to learn new knowledge **New1** and store it in the orange channels. Then, Under the constraint of distillation loss, a synchronous sparse training (the same process as (a) - (b)) is carried out while fusing the isolated new and old knowledge in (c), so (d) is obtained with half of the idle channels. Finally, learn the **New2** to get (e). (d)-(e) is consistent with (b)-(c). In other words, the (c)-(d) procedure is iterative to support class incremental learning.

3.1.1 REGIONAL ISOLATION WITH BN SPARSITY

We would like to continuously open up isolated learning regions for new knowledge, so that the parameter updating of new learning does not have an impact on the memory of old knowledge. A natural idea is to separate the old knowledge region and the new learning region from the network to learn the respective parameters separately. For this purpose, we adopt structured channel pruning (Liu et al., 2017), which performs regional isolation based on the credit assignment of each channel to the output. In this paper, the contribution of the channel is represented by the BN layer multiplication weight factors γ immediately after it. The idea behind it is that the smaller the weight factors γ by which the channel features are multiplied, the less contributions the channels are with respect to the output, as in (1). Therefore, the channels with lower γ values can be removed while the channels with higher γ values are kept as old knowledge region without affecting the representation of the original knowledge. The removed channels with low γ values can be given a new role, i.e., used as a new learning region.

$$\hat{h} = \frac{h^m - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}};\ h^{m+1} = \gamma\hat{h} + \beta \tag{1}$$

The BN layer normalizes the channel input h^m using the statistical characteristic values μ_B and σ_B^2, and then restores the data characteristics before normalization through the weight γ and bias β learned in the BN to obtain the channel output h^{m+1}.

The credit assignment of channels features is controlled by adjusting the sparsity of the distribution of γ. Based on this idea, the BN weight factors γ are first trained to be sparse, and then the weight factors γ of all channels in all convolutional layers are involved in the global ranking. The channels corresponding to the first 50% γ values are set as the region of old knowledge. The channels corresponding to the latter 50% γ values are no longer used for old knowledge representation and can be used to adapt to new learning, as shown in Figure 4.

3.1.2 FORMING NEW KNOWLEDGE

For each incremental session, the new learning regions provided by regional isolation can be used to form the new knowledge. Forming new knowledge means learning new knowledge in the new knowledge channels obtained by regional isolation. Since our goal is to learn the new knowledge

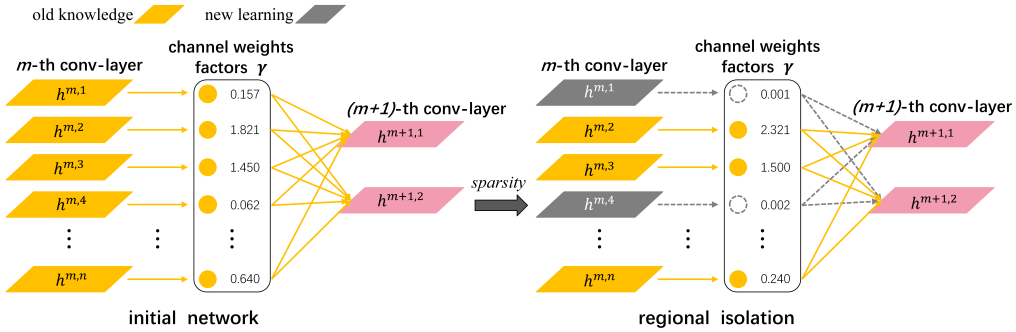

Figure 4: Regional isolation with channel sparsity. The input channel h^m of the $m\text{-}th$ convolutional layer is multiplied with the channel weight factors γ to obtain the output channel h^{m+1}. After sparse training, partial factors γ of the initialized network is close to 0, and the input channels connected to it, such as $h^{m,1}$ and $h^{m,4}$, are used for new learning. The remaining orange channels preserve the old knowledge and regional isolation is completed.

independently without disturbing the old knowledge, we propose to fix the parameters of the old knowledge region and use the cross-entropy loss to update the parameters of the new learning region.

To fix the parameters of the old knowledge region, the weight freezing method is adopted. Specifically, a weight freezing constraint is imposed on the objective function. The weight freezing constraint sets the gradient of the convolutional kernel weight parameters assigned to the old knowledge channels to zero. It means that the channel features with zero grad won't be changed by the parameter update, and the old knowledge region won't be changed either. As a result, the model won't lose its ability to perform old tasks.

Accordingly, the objective function with constraint is defined as (2).

$$\begin{aligned} &\arg\min_{\theta^t} L_{cls}(\widetilde{y_i^t}, y_i^t; \theta^t),\ y_i^t \in C_{new}^t \\ &s.t.\ grad(\theta_{C_{old}}^t) = 0,\ \theta_{C_{old}}^t \in \theta^t \end{aligned} \tag{2}$$

For incremental learning session $t(t > 0)$, we define the full set of parameters trainable at session t as θ^t, where the convolutional kernel parameters for the channels of old knowledge are $\theta_{c_{old}}^t$. The channels of old knowledge are determined by the value of the channel weight factors γ, as seen in Figure 4. L_{cls} denotes the cross-entropy loss function, C_{new} denotes the classes of the new session t for incremental learning, y denotes the ground truth of the new classes's samples, and $\tilde{y}$ denotes the predicted values of the samples of the new classes. $grad(\cdot) = \theta$ denotes the gradient of the convolution kernel weights is set to 0. (2) iterates until the loss is no longer decreasing. After the training is completed, the new knowledge is stored into the new learning region.

3.2 Regional Integration

After regional isolation is completed, the isolated knowledge of new and old should form a unified representation. Two issues should be taken into account in the integration. The first is to maintain the discrimination of the old classes. We propose a regional integration method that accepts both new and old knowledge samples as input and uses knowledge distillation to transfer the old knowledge. Secondly, we need to mitigate recency bias. Since incremental learning places too much emphasis on recent memory at the expense of old knowledge when new learning occurs, we designed margin loss to mitigate the recency bias.

3.2.1 Knowledge Distillation

For incremental learning sessions $t(t > 0)$, we remember and transfer old knowledge with knowledge distillation in order to avoid forgetting old classes. Knowledge distillation was first adopted by Hinton (Howard & Ruder, 2018) and is commonly used for model compression and knowledge transfer. In this paper, the network trained in $t-1$ session is referred to as θ^{t-1} and the network trained in the t

session is referred to as θ^t. When learning a new task, distillation from θ^{t-1} to θ^t is introduced to retain the knowledge learned from the old classes. The distillation loss is defined as L_{DS}.

L_{DS} is calculated by weighting L_{soft} and L_{hard} and the calculation is (3).

$$
\begin{gathered}
L_{soft} = -\sum_{j}^{N} p_j^T log(q_j^T) \qquad\qquad L_{hard} = -\sum_{j}^{N} c_j log(q_j^1) \\
where:\ p_i^T = \frac{exp(v_i/T)}{\sum_k^N exp(v_k/T)},\ q_i^T = \frac{exp(z_i/T)}{\sum_k^N exp(z_k/T)},\ v_i = \theta_{logit}^{t-1}(x),\ z_i = \theta_{logit}^{t}(x) \\
The\ distillation\ loss\ \ L_{DS}\ \ is\ given\ by: \quad L_{DS} = \alpha L_{soft} + \beta L_{hard}
\end{gathered}
\tag{3}
$$

θ is the set of model parameters. c_j represents the ground truth value in the j-th class, $c_j \in \{0, 1\}$, which is one for positive labels and zero for negative labels. v_i represents the logit output of the input sample on θ^{t-1}, and z_i represents the logit output of the input sample on θ^t. p_i^T and q_i^T denote the values of the distillation outputs of the input samples x on model θ^{t-1} and θ^t in the i-th class, respectively. The distillation temperature T is generally set to 2. When $T = 1$, the distillation output changes to the standard output of softmax. α and β are balance coefficients, where $\alpha = 1.0$ and $\beta = 0.24$.

3.2.2 Alleviating Recency Bias

The regional integration also introduces the margin loss regular term as a soft constraint to alleviate the recency bias problem. It further improves the classification accuracy of SRII by making the gap between old and new classes bigger in the probability space to suppress the destruction of historical memory with new learning.

margin loss is short for L_M. L_M aims to suppress the recency bias by separating the decision boundaries of the old and new classes, and the expression is (4).

$$
\begin{gathered}
margin = max(clafi_{old\_num:new\_num}(x_i, y_i; \theta)) - max(clafi_{:old\_num}(x_i, y_i; \theta)),\ x_i, y_i \in C_{old} \\
L_M = log(s_0 + mean(topK(margin, k))/b),\ k, b \in N_+
\end{gathered}
\tag{4}
$$

where x_i denotes sample, y_i denotes label, C_{old} denotes old knowledge category, and $clafi$ is the softmax classifier. The number s_0 ensures that the definition field of the log function is a positive number. The vector $clafi_{old\_num:new\_num}$ is the probability that the classifier classifies the samples into new knowledge categories, and the vector $clafi_{:old_{num}}$ is the probability that the classifier classifies the samples into old knowledge categories. The $margin$ represents the difference between the probability of mis-predicting an old class sample as a new class sample and the probability of being correctly classified. In order to simplify the calculation, the average value of the top k maximum $margins$ in every mini-batch is used to calculate L_M, and b is the scaling factor. With the goal of improving the accuracy of incremental recognition, using the grid search method, we obtain the empirical parameter k as $1/5$ of the batch size and b as 20. In addition, $L2$ normalization is applied in this research to clean the input data of the classification layer in advance.

3.3 Network Training

The training for SRII is divided into a single base training and a series of incremental training, as shown in Figure 5, see Section 3.3.1, Section 3.3.2 and Algorithm 1 for details. In Algorithm 1, lines 4-7 represent the base training of SRII, and lines 8-15 represent the incremental training of SRII. D_t denotes the new knowledge and $\mathcal{M}_t$ is the old knowledge. t denotes the task stage. List V stores channel identifier.

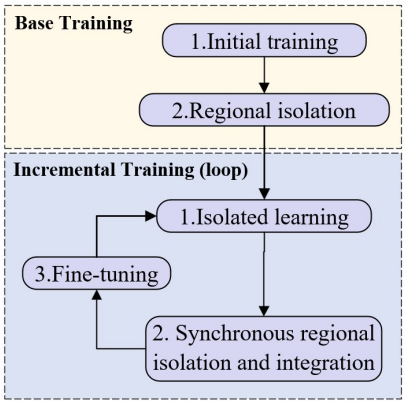

Figure 5: Training for SRII.

Algorithm 1 Pseudo code of SRII

Require: data $D_t^{tr} = \{D_t, \mathcal{M}_t\},\ \theta$
Output: θ
for steps $\leftarrow 1$ to $|D_t^{tr}|$ **do**
  if in the case of Initial training, **then**
    train θ by minimizing Eq. (5) in $\{D_t\}$.
  elif in the case of Regional integration, **then**
    train θ by minimizing Eq. (6) in $\{D_t\}$.
  elif in the case of Isolated learning, **then**
    if first run this, **then**
      store channel id of old knowledge in V.
    end if
    train θ by minimizing Eq. (7) in $\{D_t\}$.
    freeze weights of channels recorded by V.
  else
    train θ by minimizing Eq. (8) in $\{D_t, \mathcal{M}_t\}$.
  end if
end for

3.3.1 BASE TRAINING

Basic training consists of "**Initial training**" and "**Regional isolation**". Initial training is performed for the first new learning, and regional isolation compresses the storage space occupied by the first new learning and reserves space for subsequent learning.

The loss of **initial training** is defined as L^{Init}. Initial training uses cross-entropy loss to extract sample information directly, formalized as (5). c_i represents the ground truth value in the i-th class, $c_i \in \{0, 1\}$, which is one for positive labels and zero for negative labels. q_i represents the value of the softmax output of the new learning sample in the i-th class.

$$L^{Init} = \sum_{i}^{N} c_i log(q_i) \tag{5}$$

The **regional isolation** loss, L^{Split}, is defined as (6). As described in Section 3.1.1, regional isolation is achieved by training the sparsity of channels on the BN weight factors γ.

$$\begin{aligned} L^{Split} &= \sum_{i}^{N} c_i log(q_i) + \delta g(\gamma) \\ s.t.\ g(\gamma) &= \sum_{k=1}^{K} |r_k|,\ r_k \in \gamma \end{aligned} \tag{6}$$

$g(\gamma)$ is the L1-norm[2] of the weight factors γ, which is adopted to achieve channel sparsity. δ balances two terms and is set to 1e-4, which follows the Parameter-Setting of the (Liu et al., 2017) for channel sparsity. r_k is a factor in the set of weights factor γ, and the total number of weight factors contained in the set γ is K.

3.3.2 INCREMENTAL TRAINING

Incremental training includes "**Isolated learning**", "**Synchronous regional isolation and integration**", and "**Fine-tuning**". Isolated learning puts newly acquired knowledge into partitioned reserved space to minimize interference with existing knowledge. Synchronous regional isolation and integration fuses and compresses partitioned knowledge to assist the model in learning new and old knowledge within the same domain space as well as to reserve space for future learning. The

[2]L1 regularization causes a large number of parameters to be zero, which results in sparse solutions. Discarding parameters close to zero has little effect on the recognition ability of SRII. We can set aside the channels corresponding to the factors γ close to zero to integrate the new knowledge.

simultaneous regional isolation continues to reserve space for new incremental sessions, continuously maintaining the ability to learn new knowledge. In addition, to enhance the classifier's ability to adapt to the acquired knowledge, we set up a fine-tuning process with low training challenge, which freezes all network layers except the classifier and updates only the classifier's parameters.

The **isolated learning** loss L^{Iso} is exactly the same as the initialized training loss L^{Init}, since only new learning is occurring at this time.

$$L^{Iso} = \sum_{i}^{N} c_i log(q_i) \tag{7}$$

Although the loss of **fine-tuning** is the same as **synchronous regional isolation and integration** and is denoted by $L^{S\&I}$, fine-tuning only updates the parameters of the classifier. Note that "Synchronous regional isolation and integration" is a synchronous operation of "regional isolation" and "regional integration", which requires the regular term $g(\gamma)$ to achieve "regional isolation". L_{DS} and margin loss are used for "regional integration". Accordingly, $L^{S\&I}$ can be expressed as (8).

$$L^{S\&I} = L_{DS} + \lambda * L_M + \delta g(\gamma) \tag{8}$$

where λ = 0.36 and δ = 1e-4 are balance cofficient. The expression of L_{DS} is (3), and the expression of margin loss is (4).

4 Conclusion

In this paper, SRII was proposed to overcome catastrophic forgetting in incremental learning. SRII consists of regional isolation and regional integration, in which regional isolation isolates new learning processes to avoid interfering with existing knowledge, while regional integration establishes a unified, high-precision cognition to adapt to the requirements for single-headed output of class incremental learning. SRII outperforms other baselines throughout the experiment, validating the superiority of our method. A limitation of our method is that it has a certain demand on training time, since the two-stage process of isolation and integration requires more time.

Our future work will continue to focus on the recency bias, with the goal of developing a more lightweight network capable of quickly learning new knowledge from a small number of samples. Finally, although reinforcement learning research in recent years has achieved ultra-high performance beyond human beings in several areas, its agents still suffer from catastrophic forgetting, which is what we will explore also in the next phase of our work.

References

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

Gail A. Carpenter and Stephen Grossberg. The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88, 1988.

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

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

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Xiaoliang Dai, Hongxu Yin, and Niraj K Jha. Incremental learning using a grow-and-prune paradigm with efficient neural networks. *IEEE Transactions on Emerging Topics in Computing*, 2020.

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

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

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jong-Yeong Kim and Dong-Wan Choi. Split-and-bridge: Adaptable class incremental learning within a single neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8137–8145, 2021.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

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

Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2262–2268. IEEE, 2018.

Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2544–2553, 2021.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, pp. 2736–2744, 2017.

Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3589–3599, 2021.

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

Nikhil Mehta, Kevin J Liang, and Lawrence Carin. Bayesian nonparametric weight factorization for continual learning. *arXiv preprint arXiv:2004.10098*, 2020.

Joseph O'Neill, Barty Pleydell-Bouverie, David Dupret, and Jozsef Csicsvari. Play it again: reactivation of waking experience and memory. *Trends in neurosciences*, 33(5):220–229, 2010.

Randall C O'Reilly, Rajan Bhattacharyya, Michael D Howard, and Nicholas Ketz. Complementary learning systems. *Cognitive science*, 38(6):1229–1248, 2014.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Robi Polikar, Lalita Upda, Satish S Upda, and Vasant Honavar. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, 31(4):497–508, 2001.

Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Lifelong generative modeling. *Neurocomputing*, 404:381–400, 2020.

Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1320–1328, 2017.

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

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9374–9384, 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Zifeng Wang, Tong Jian, Kaushik Chowdhury, Yanzhi Wang, Jennifer Dy, and Stratis Ioannidis. Learn-prune-share for lifelong learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 641–650. IEEE, 2020.

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

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

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

Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1131–1140, 2020.

A APPENDIX

A.1 DATASET AND TEST SCHEME

In this paper, All implementations are based on PyTorch and NVIDIA CUDA. The experiments were performed on a RTX-2080 Ti GPU with 12G memory and CPU@14.4 GHz. The experimental datasets are CIFAR100 and miniImageNet.

CIFAR100 dataset. CIFAR100 is a subset of a dataset with 80 million small images collected by Alex Krizhevsky, which are commonly used as benchmark datasets in the field of incremental learning. It contains 60,000 RGB images from over 100 classes, with 500 images per class for training and 100 images per class for testing. The size of each image is 32×32.

miniImageNet dataset. miniImageNet dataset is a subset of ImageNet-1k. It contains 60,000 color images in 100 classes. Each class has 500 images for training and 100 images for testing. The size of each image is 84×84. Compared to CIFAR100, the miniImageNet dataset is more complex, more suitable for prototyping, and is a commonly used benchmark dataset in the field of incremental learning.

Experiment setup. We preprocess images by following the protocol of VGG (Simonyan & Zisserman, 2014). For a fair comparison, the experiments use the same replay sample selection protocol as "Incremental classifier and representation learning" (iCaRL) (Rebuffi et al., 2017) and the maximum exemplar storage capacity of 2000.

All experiments exploit the SGD optimizer and initialize the learning rate to 2.0. We use the same class random transformation seed in all works. There are 100 epochs in each training stage. The learning rate for epoch 0-47, 48-61, 62-79 and 80-99 are 2.0, 0.4, 2/25 and 2/125, respectively. Weight decay is constant at 5e-4 and the batch size is 128. Additionally, we just access each task once and do not optimally adjust the hyper parameters in the entire task set for the reason that tuning the hyper parameters with all available data would result in overfitting and limit the flexibility of new learning, which is contrary to the principle of incremental learning.

A.2 SRII VS BASELINES

Figure 6 reports the comparison of SRII with several incremental learning approaches. SRII achieves leading results throughout the approaches. The average accuracy are shown in Table 1, Table 2, and Table 3 in detail.

As illustrated in Figure 6(a), a 10-stage incremental training was performed on CIFAR100, followed by two series of experiments on the miniImageNet with 10 and 20 classes per incremental step, as shown in Figure 6(b) and (c). Finally, SRII with ResNet-50 achieves a 10-stage Top-1 average accuracy of 76.89 percent for CIFAR100, exceeding ResNet-18 (2.71%) and VGG16 (9.34%), as seen in Figure 6(d).

Table 1 shows that on the CIFAR100 dataset, SRII not only outperforms LwF (Li & Hoiem, 2017) by 29.06%, EWC (Kirkpatrick et al., 2017) by 23.65%, and Averaged-Gradient Episodic Memory (A-GEM) (Chaudhry et al., 2018) by 28.42%, which do not have regional isolation but also outperforms S&B (Kim & Choi, 2021) by 6%, which focuses on regional isolation and its performance is closest to SRII. In comparison to other approaches, the minimal difference (6.12%) between the SRII of 74.18% and the top limit of 80.3% demonstrates our method's superiority. Owing to the higher variance and interference among class incremental tasks, the A-GEM (Chaudhry et al., 2018) and Episodic memories replay (EMR) (Chaudhry et al., 2019) methods commonly used for task increment performed poorly in this experiment. LUCIR (Hou et al., 2019) and ABD (Smith et al., 2021) outperformed LwF, EWC, and A-GEM, which could be attributed to their efforts to overcome the recency bias. Additionally, even compared with the challenges of memory replay methods iCaRL (Rebuffi et al., 2017) and SCR (Mai et al., 2021), SRII still provides 10.94% and 8.20% advantages, respectively. Memory replay remains an excellent means of avoiding forgetting. The early iCaRL ranks so highly, probably due to its NCM (Nearest-Class-Mean) classifier relying on memory exemplars (representative samples) rather than the fully connected layer for measurement classification, which avoids the structural changes of old class connections and thus has a certain resistance to catastrophic forgetting.

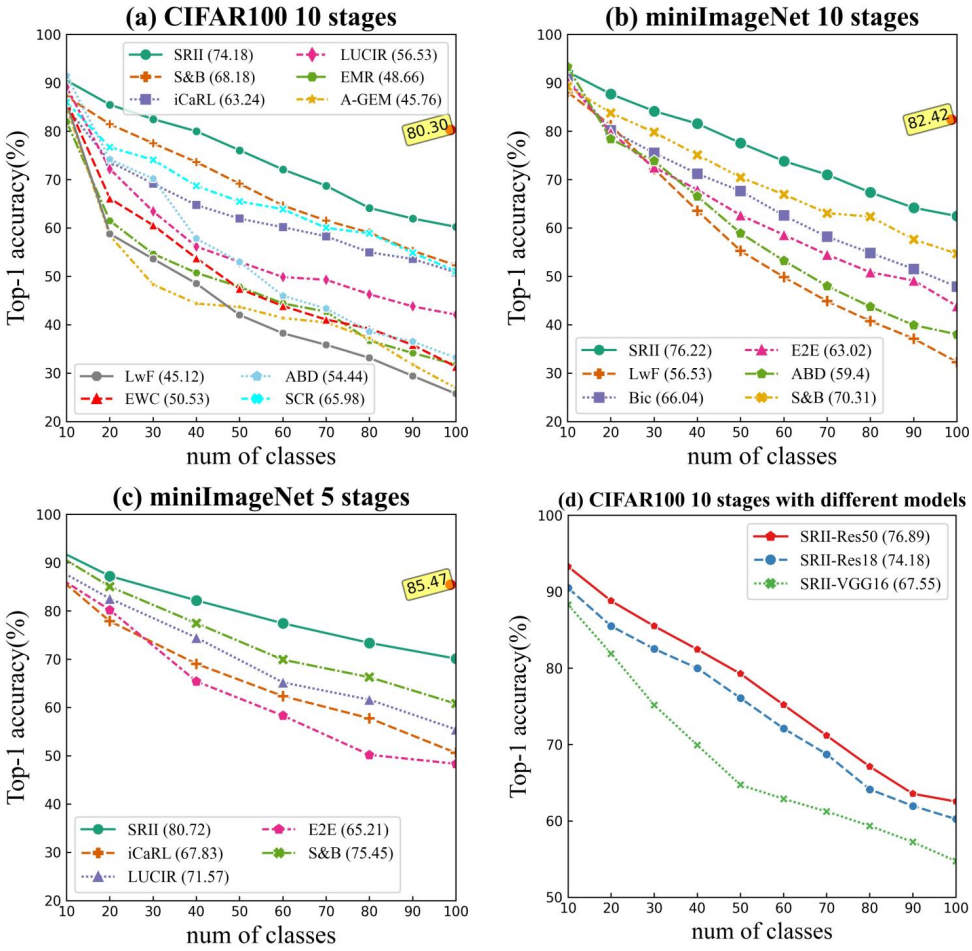

Figure 6: SRII achieves state-of-the-art results. The mean values of Top-1 accuracy are indicated in parentheses. The yellow text boxes at the right border of the figure mark the joint training upper bound averages, i.e., the upper bound in Table 1. For a fair comparison, the experiments follow the replay sample selection protocol of (Chaudhry et al., 2019) and the experimental setup in A.1.

Table 1: Recognition accuracy of 10-stage on CIFAR100. "Avg." denotes the average accuracy over 10 stages, UP is short for Upper Bound. Best results are marked in bold.

Method	Number of classes										
	10	20	30	40	50	60	70	80	90	100	Avg.
A-GEM	85.0	58.42	48.3	44.39	43.7	41.4	40.48	37.26	31.74	26.91	45.76
EMR	82	61.5	54.67	50.74	47.83	44.42	42.71	36.73	34.17	31.78	48.66
iCaRL	84.9	73.7	69.17	64.75	61.94	60.17	58.3	54.99	53.6	50.83	63.24
LUCIR	89.1	72.2	63.43	56.17	53	49.87	49.3	46.31	43.81	42.09	56.53
LwF	85.8	58.8	53.62	48.52	42.02	38.24	35.86	33.16	29.43	25.74	45.12
EWC	86.1	66.1	60.57	53.75	47.42	43.88	41.07	39.24	35.83	31.33	50.53
ABD	**91.5**	74.2	70.2	57.8	52.98	46	43.36	38.59	36.52	33.2	54.44
SCR	86	76.7	74.1	68.7	65.5	63.9	60.03	58.9	54.91	51.08	65.98
S&B	87.2	81.47	77.52	73.64	69.15	64.66	61.55	59.05	55.31	52.29	68.18
UP	87.31	84.57	82.43	81.59	79.74	78.64	78.42	77.11	76.85	76.32	**80.30**
SRII	90.5	**85.5**	**82.52**	**80**	**76.08**	**72.11**	**68.74**	**64.14**	**61.97**	**60.25**	**74.18**

In Table 2 and Figure 6(b), SRII and S&B achieve the best performance in miniImageNet's 10-stage operation. In the last incremental step, SRII outperforms LwF, Bic, E2E, ABD, and S&B by 30.18%,

14.6%, 18.6%, 24.45%, and 7.76%, respectively, with an average gap of 6.2% from the upper bound. ABD (Smith et al., 2021) is 6.64% worse than Bic (Wu et al., 2019), which is consistent with ABD findings. Bic is close to E2E and greater than LwF, as evidenced by Bic's research. It can be seen that the recognition accuracy in miniImageNet is generally slightly higher than CIFAR100, probably because miniImageNet contains more detailed pixel information for recognition. Considering the low pixels of the CIFAR100 dataset, the first pooling layer in the ResNet-18 (He et al., 2016) structure is removed in this experiment. Additionally, CIFAR100 and miniImageNet employ various normalization factors on their respective datasets to improve identification accuracy, all of which may have a modest effect.

Table 2: Recognition accuracy of 10-stage on miniImageNet

Method	**Number of classes**										
	10	20	30	40	50	60	70	80	90	100	Avg.
S&B	89.26	83.77	79.8	75.12	70.46	66.93	63.08	62.35	57.63	54.72	70.31
LwF	88.1	81.2	72.2	63.57	55.3	49.85	44.86	40.77	37.14	32.3	56.53
Bic	90.8	80.18	75.53	71.23	67.65	62.58	58.19	54.86	51.54	47.88	66.04
E2E	90.45	79.68	72.53	67.93	62.65	58.58	54.49	50.86	49.14	43.88	63.02
ABD	93.3	78.39	73.86	66.54	58.94	53.24	48.01	43.73	39.93	38.03	59.4
UP	89.5	86.71	84.13	83.25	81.56	82.14	80.57	79.11	78.88	78.32	**82.42**
SRII	**92.3**	**87.69**	**84.18**	**81.57**	**77.57**	**73.84**	**71.04**	**67.36**	**64.19**	**62.48**	**76.22**

Table 3: Recognition accuracy of 5-stage on miniImageNet

Method	**Number of classes (5 Stages)**						
	5	20	40	60	80	100	Avg.
S&B	93.2	85.05	77.47	69.91	66.29	60.80	75.45
iCaRL	89.24	77.89	69.06	62.37	57.78	50.62	67.83
E2E	88.76	80.21	65.43	58.32	50.21	48.35	65.21
LUCIR	90.03	82.57	74.48	65.21	61.63	55.48	71.57
UP	93.2	88.81	85.69	83.92	81.46	79.75	**85.47**
SRII	**93.9**	**87.24**	**82.2**	**77.45**	**73.40**	**70.14**	**80.72**

A.3 Ablation Study

In order to verify the performance improvement of SRII with sustainable regional isolation and the margin loss that can alleviate recency bias during regional integration, we performed ablation studies as follows.

Table 4: Retention of old knowledge in SRII. The network uses the free space provided by the previous regional isolation process to learn new knowledge during the isolation learning period. Then the feature extraction part (all parts before the fully connected layer) is frozen, and the old knowledge exemplars on the classifier (fully connected layer) for fine-tuning are replayed. The degree of retention of the old knowledge in the feature extraction part is observed regardless of the recognition ability of new knowledge. It can be seen that more complete old knowledge is still retained at the moment when the isolated learning has just been completed.

Nets for feature extraction	The recognition accuracy of old knowledge (%)	
	Joint training (%)	Fine-tuning recovery (%)
ResNet-18	86.4	85.3 (1.1↓)
ResNet-34	87.8	86.4 (1.4↓)
VGG16	88.9	87.2 (1.7↓)

Sustainable regional isolation ablation experiment. In Table 4 we verify that our sustainable regional isolation approach proposed above can retain more old knowledge while learning new

knowledge. By looking at the three rows of data in the table, we notice that SRII using ResNet-18, ResNet-50, and VGG16 as feature extraction layers indeed stores relatively complete information about old knowledge (e.g., for ResNet-18 there is less than 1.2% decay of old knowledge recognition accuracy).

According to Table 4, the reason for the network's ability to retain old knowledge is that SRII uses region isolation to independent the process of new learning to suppress the interference and destruction of old knowledge by new learning. In particular, since the regional integration has not been performed in Table 4, the model has not yet formed a unified expression competence for new and old knowledge, and a certain degree of knowledge loss will subsequently occur in the regional integration training. Nevertheless, it provides a good start to the fusion, which is valuable. The results in Figure 7 also validate that, as regional isolation increases the class incremental recognition accuracy.

Figure 7 confirms that regional isolation can improve recognition of SRII. With ResNet-18 as the detection backbone and CIFAR100 data set, 1-stage basic training and 9-stage incremental training were performed. It can be seen that SRII with regional isolation brings significant accuracy benefits compared to the baseline network without regional isolation. Figure 8 is a refinement of the two incremental processes for 30 and 50 classes, which confirms that regional isolation improves classification ability. All four plots in Figure 8 indicate that newly arrived classes (i.e., a few misclassified scatters in the lower right corner of the plots) are accurately classified and that the network frequently misclassifies old classes as new classes (i.e., more scatters exist in the upper right corner of the plots). We notice that (b) and (d) have fewer misclassifications. This is due to the fact that regional isolation keeps the training environments of new and old classes separate, which makes the interference with the old classes less when training the new classes, so the network maintains a higher recognition ability than (a) and (c).

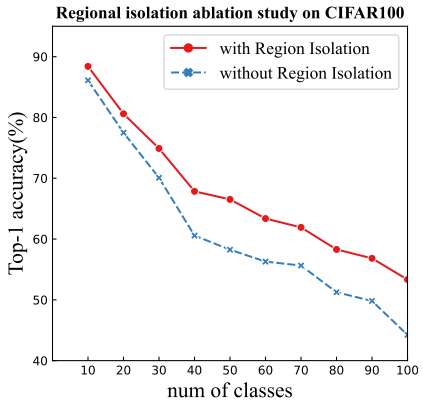

Figure 7: Accuracy of regional isolation ablation.

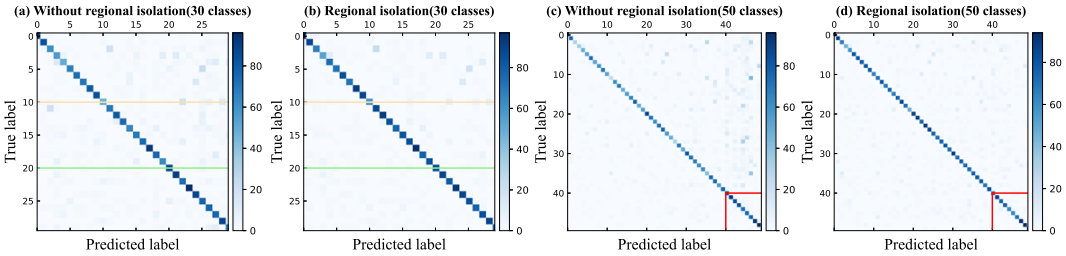

Figure 8: Confusion matrix map with and without regional isolation on miniImageNet. The size of the sample replay region is 2000. Recent classes have fewer misclassifications (as seen in (a)(b)), yellow and green lines assist in viewing. The last incremental task classification performed well (as seen in (c)(d)), with red lines aiding the view. Best view in color.

Margin loss ablation experiment. We separate the margin loss from SRII as a baseline and look at the value of this component on its own.

Figure 9 presents the accuracy improvement process relying on margin loss, and the model with margin loss always wins in each learning stage. The dataset is equally divided into 5 learning stages, and each is trained 100 times. Margin loss improves the upper limit of accuracy. Since there is only new knowledge in the initial stage, Stage 1 is not plotted. Additionally, the loss comparison of Stage 3 is specially added to reflect that margin loss can further train the model.

Margin loss can improve accuracy because it enables the model to shift from biasing towards new knowledge to more accurately distinguishing new knowledge from old knowledge, alleviating recency

bias. We verify that margin loss enhances the network's ability to distinguish between new and old knowledge by visualizing the high-dimensional features extracted from SRII, as seen in Figure 10. For a fair comparison, we uniformly trained 10 base classes and 10 incremental classes using the ResNet-18 network as the backbone in combination with the CIFAR100 dataset to observe the ability of SRII with and without margin loss to discriminate between new and old knowledge. As the last convolutional layer of the deep neural network contains the richest spatial and semantic information and is often used for classification visualization, we extract the high-dimensional features of the last convolutional layer of the SRII to demonstrate the network's discrimination between new and old knowledge based on the PCA dimensionality reduction, as seen in Figure 10. Without margin loss, the distribution space of the old class is confused with the distribution space of the new class in Figure 10(a)(b). In addition, Figure 10(c) shows the suppression of "recency bias" by margin loss, which is achieved by separating the decision boundaries of new and old classes to achieve "inter-class separation and intra-class aggregation". Compared to the baseline without margin loss, the network with margin loss better separates the samples of new and old classes, and the samples within each class are more compact.

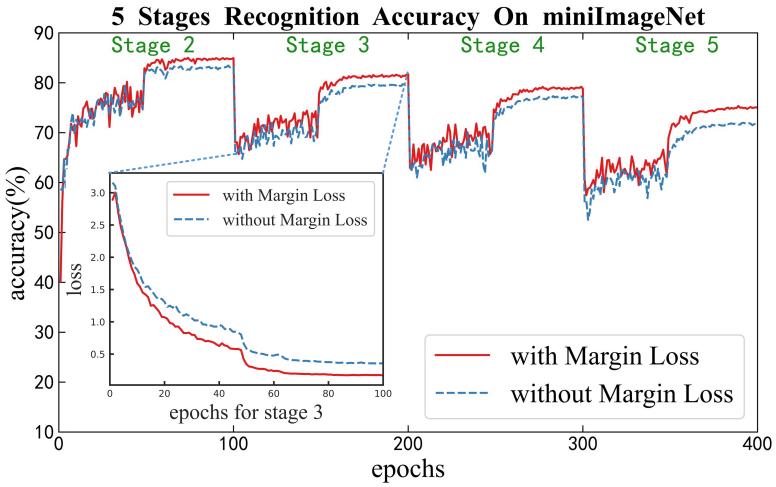

Figure 9: Ablation study of margin loss.

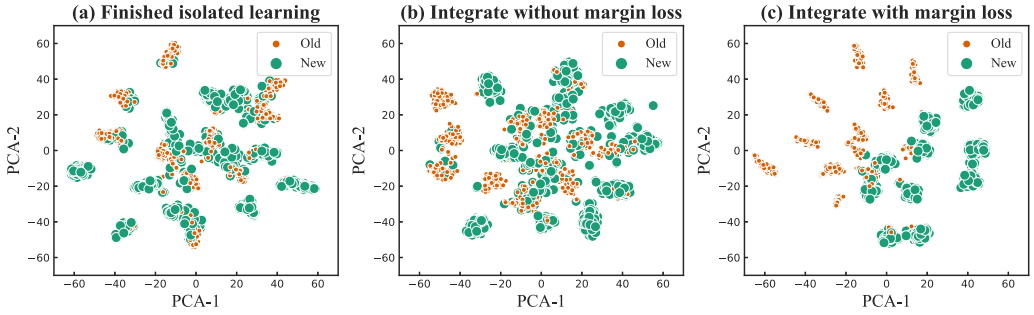

Figure 10: Distribution of knowledge in three cases. PCA-1 is the x-axis and PCA-2 is the y-axis. Figure (a) shows the initial distribution of new and old knowledge on the CIFAR100 dataset that has just completed the phase of isolated learning (green denotes new knowledge, orange denotes old knowledge, and there are 10 categories for new and old knowledge). Figure (b) is the new and old knowledge distribution state after integration training without margin loss constraint. Figure (c) is the integration result after training with margin loss constraint. Figure (c) separates better than Figure (b), with "inter-class separation and intra-class aggregation". PCA is used for visualization.

In Table 5 and Table 6, we discussed the ablation experiments on margin loss. As a baseline, the margin loss will be taken out of the whole SRII. Then, 10 stages of incremental experiments will be run with the CIFAR100 and miniImageNet datasets to see how it affects the performance of SRII.

SRII($-L_M$) denotes the absence of margin loss. "Avg." denotes the average accuracy of 10 stages with CIFAR100 or miniImageNet.

Table 5: Ablation study accuracy of margin loss on CIFAR100

Method	**Number of classes(10 Stages)**										
	10	20	30	40	50	60	70	80	90	100	Avg.
SRII($-L_M$)	88.9	82.45	79.15	76.73	72.50	68.21	63.69	59.74	56.27	54.52	70.22
SRII	90.5	**85.5**	**82.52**	**80**	**76.08**	**72.11**	**68.74**	**64.14**	**61.97**	**60.25**	**74.18**
Gains	1.6↑	3.05↑	3.37↑	3.27↑	3.58↑	3.9↑	5.05↑	4.4↑	5.7↑	5.73↑	3.96↑

Table 6: Ablation study accuracy of margin loss on miniImageNet

Method	**Number of classes(10 Stages)**										
	10	20	30	40	50	60	70	80	90	100	Avg.
SRII($-L_M$)	91.9	83.3	80.53	77.63	73.44	69.15	64.45	60.51	57.1	54.87	71.29
SRII	**92.3**	**87.69**	**84.18**	**81.57**	**77.57**	**73.84**	**71.04**	**67.36**	**64.19**	**62.48**	**76.22**
Gains	0.4↑	4.39↑	3.65↑	3.94↑	4.13↑	4.69↑	6.59↑	6.85↑	7.09↑	7.61↑	4.93↑

Random increase and decrease ablation study. To better analyze and evaluate the impact of the SRII components, the table of random increase and decrease ablation study with two components of sustainable regional isolation (RI) and regular term L_M is supplied, as shown in Table 7.

Table 7: Random increase and decrease ablation study of each component on CIFAR100. RI denotes regional isolation, L_M denotes margin loss, and Final Acc.↓ denotes the decline of recognition accuracy in the last stage compared with baseline SRII. The best results are marked in bold, and the worst results are marked in red.

Variations		**Number of classes(10 Stages)**										Avg	Final
RI	L_M	10	20	30	40	50	60	70	80	90	100	Acc.	Acc.↓
		88.6	78.60	72.66	67.55	63.94	61.25	58.4	54.36	52.35	49.68	64.74	10.57
	✓	88.5	81.94	76.73	70.74	68.29	66.51	62.33	59.21	55.65	52.43	68.23	7.82
✓		88.9	82.45	79.15	76.73	72.50	68.21	63.69	59.74	56.27	54.52	70.22	5.73
✓	✓	90.5	**85.5**	**82.52**	**80**	**76.08**	**72.11**	**68.74**	**64.14**	**61.97**	**60.25**	**74.18**	—

Through the ablation study of components to analyze the impact of each component. It is primarily necessary to underline that performance degradation emerges regardless of the missing components, which substantiates that each component of SRII is necessary. Table 7 reflects the importance of RI and L_M in that removing RI or L_M brings about 7.82% and 5.73% of the final performance drop as well as 5.95% and 5.95% average performance drop, respectively. Additionally, the elimination of RI and L_M at the same time puts the model under the most severe loss, 10.57%. In one word, all components of SRII play a positive role, among them regional isolation is an important link to alleviate "catastrophic forgetting".