Partitioned Memory Storage Inspired Few-Shot Class-Incremental learning

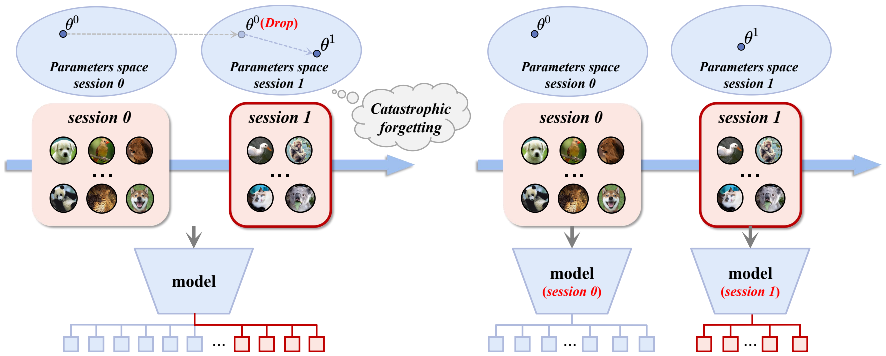
BMVC 2024 Submission # 402

Abstract

Current mainstream deep learning techniques exhibit an over-reliance on extensive training data and a lack of adaptability to the dynamic world, marking a considerable disparity from human intelligence. To bridge this gap, *Few-Shot Class-Incremental Learning* (FSCIL) has emerged, focusing on continuous learning of new categories with limited samples without forgetting old knowledge. Existing FSCIL studies typically use a single model to learn knowledge across all sessions, inevitably leading to the stability-plasticity dilemma. Unlike machines, humans store varied knowledge in different cerebral cortices. Inspired by this characteristic, our paper aims to develop a method that learns independent models for each session. It can inherently prevent catastrophic forgetting. During the testing stage, our method integrates *Uncertainty Quantification* (UQ) for model deployment. Our method provides a fresh viewpoint for FSCIL and demonstrates the state-of-the-art performance on CIFAR-100 and *mini*-ImageNet datasets.

1 Introduction

Deep learning has achieved significant milestones in numerous large-scale computer vision tasks. These approaches generally invoke learning the mapping from samples to corresponding labels using extensive datasets. However, a trained deep neural network usually lacks the ability to generalize to new categories. Recently, *Class Incremental Learning* (CIL) has received widespread attention since it enables trained models to be extended to new categories [3, 8]. CIL strives to allow models to continually learn different categories from data streams instead of a fixed dataset while preserving the capability to recognize previously encountered categories. Most CIL studies focus on letting models learn incrementally when the new class samples are sufficient. However, acquiring samples from new categories proves challenging and resource-intensive in many piratical scenarios [15]. To tackle the challenge posed by the scarcity of samples from new categories, a task of more challenging and practical significance is denoted as FSCIL. In contrast to CIL, FSCIL must incrementally learn in situations with extremely limited labeled samples for new categories. The current FSCIL studies predominantly adhere to the traditional CIL paradigm, wherein a single model assimilates all data throughout the incremental process, sharing identical model parameters and decision boundaries. This learning paradigm presents significant challenges in conserving the model's old memory and deviates from the manner in which the human brain stores memories. As shown in Fig. 1, acquiring knowledge of new categories inevitably alters parameters trained on old ones, causing catastrophic forgetting.



(a) Normal incremental method (b) Parameter isolation method

Figure 1: Comparing ordinary incremental learning method to parameter-isolation method. θ^t represents the model parameter on session t. (a) Model parameter changes trigger catastrophic forgetting (b) The model for each session learns the data independently and saves the parameters separately.

In contrast to this memory-burdened and forgettable learning paradigm, the human brain distributes learned knowledge in different areas of the cerebral cortex [6, 14]. After exposure to a given task, a human can promptly associate it with the corresponding area of the cerebral cortex. This adaptive learning method is particularly well-suited for FSCIL because it can store memories separately. However, from the standpoint of biological and cognitive science, the challenges faced by deep neural networks in emulating the human cerebral cortex within FSCIL are principally manifested in the following two aspects. **(I)** ***How can the model imitate the partitioned memory storage?*** The unavailability of the session label results in the necessity for the model to store the knowledge acquired across all sessions entirely within a unified memory area that cannot be segregated. **(II)** ***How can one establish a mapping from task to memory?*** The session label during the testing process is also missing in the FSCIL. This implies that directly mapping the relevant areas of the cerebral cortex, as the human brain naturally does for a given task, is tricky. This paper is dedicated to bridging the gap between the learning process of FSCIL and the way that the human cerebral cortex stores memories by addressing the aforementioned challenges.

To tackle the first challenge, we adopt a parameter-isolation strategy for incremental training, training distinct classification models for each session. The conventional FSCIL poses a risk of catastrophic forgetting because all sessions share the same model parameters. In our parameter-isolation strategy for FSCIL, each model exclusively stores knowledge acquired in its respective session, as shown in Fig. 1 (b). This approach effectively retains knowledge of each session and prevents catastrophic forgetting. To address the second challenge, we predict the session label of the samples by UQ. Test samples are fed into models on each session individually to obtain classification results along with corresponding uncertainty values. After, the appropriate classification result is selected based on the uncertainty value. UQ enables the model to express the level of uncertainty associated with a given sample. It normally provides low uncertainty for for the learned categories and high uncertainty for never-encountered categories.

We conduct comprehensive comparative and ablation experiments on multiple benchmark datasets to validate the effectiveness. The contribution is summarized as follows:

- We introduce a novel perspective for FSCIL, mirroring the memory storage mechanism of the human cerebral cortex. Our work pioneers the use of a parameter-isolation method in FSCIL. This innovative approach addresses the challenges and achieves state-of-the-art performance on benchmark datasets.
- We use information entropy to compute the uncertainty of the samples in the testing stage, which enables the prediction of sample session labels.
- We employ a novel data augmentation strategy to generate virtual prototypes. This strategy significantly enhances the feature extraction capability of the backbone.

2 Related Word

FSCIL represents a more challenging task compared to class-incremental learning. FSCIL aims to continuously acquire knowledge of new categories with a limited amount of labeled data while simultaneously preventing the inadvertent forgetting of previously learned categories [13]. The problem setting of FSCIL is firstly proposed in TOPIC [11]. It utilizes the *Neural Gas* (NG) network to stabilize the topological manifold features between new and old categories. CEC [13] uses a graph neural network to update the classifier parameters in each incremental session based on the global knowledge of previous sessions. Simultaneously, it creates pseudo-incremental scenarios to optimize the graph neural network during base training. The pseudo-incremental approach has been used in many studies [15, 16]. Unlike previous approaches that only consider the performance of the current session, [15] uses a method that enhances the scalability of the incremental model by compacting the embedding space through the inversion of virtual prototypes. In this paper, we crop and mix images of different categories to form images of virtual categories and generate virtual prototypes based on semantic information similar to real categories.

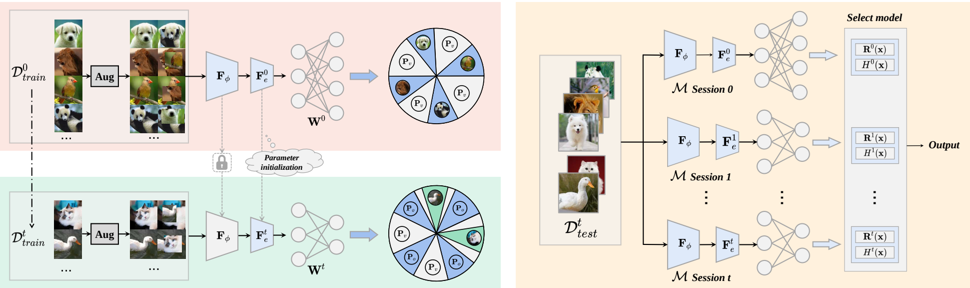
3 Problem Setting

FSCIL aims to incrementally learn new categories in limited labeled samples while retaining old knowledge. In FSCIL, the complete training dataset $\mathcal{D}_{train}$ can be represented as $\{\mathcal{D}^{(0)}_{train}, \mathcal{D}^{(1)}_{train}, \mathcal{D}^{(2)}_{train}, \cdots, \mathcal{D}^{(N)}_{train}\}$, whever $\mathcal{D}^{(t)}_{train}$ represents the training data of t-th session, and N is the total number of sessions. The samples and their labels from different sessions do not overlap. $\mathcal{D}^{(0)}_{train}$ is the base training dataset, which contains abundant training samples from different categories. $\mathcal{D}^{(i)}_{train}(1 \leq i \leq N)$ is the incremental training dataset, which contains only a few samples. It contains N classes, and each class contains K samples, known as N-way K-shot. During the training in the t-th session, the model only interacts with the samples and labels from the current session $\mathcal{D}^{(t)}_{train}$. During the testing stage of the i-session, the evaluation involves all the categories the model has already learned.

4 Methodology

In this paper, we use a CutMix-based data augmentation approach for base session training to generate virtual prototypes to train more robust feature extractors. For incremental sessions, we train separate classification models for each session to avoid the appearance of forgetting.

This is the first parameter isolation method in FSCIL. To prevent the limited training data from being overfitted during incremental learning, we use a branch training strategy to freeze most of the parameters of the base model. In the testing stage, we employ UQ on fed samples by calculating the information entropy derived from the model output to assist the sample in selecting the appropriate classification model. This process enables the samples to choose the proper classification results from multiple models. The overall framework of our proposed method is shown in Fig. 2.



(a) The training phase on base session and incremental sessions

(b) The testing phase

Figure 2: The overview of our proposed humankind memory-inspired FSCIL approach. (a) The CutMix data augmentation method enhances the feature extractor's performance by generating virtual prototypes. For session t, freeze the bulk $\mathbf{F}_\phi$ of the feature extractor, fine-tune the tail $\mathbf{F}_e^t$ and the classifier $\mathbf{W}^t$. (b) In the testing stage, the samples $\mathbf{x}$ are fed to a trained series of models outputting classification results $\mathbf{R}^t(\mathbf{x})$ and uncertainty $H^t(\mathbf{x})$. The reliable result is then selected based on the value of $H^t(\mathbf{x})$.

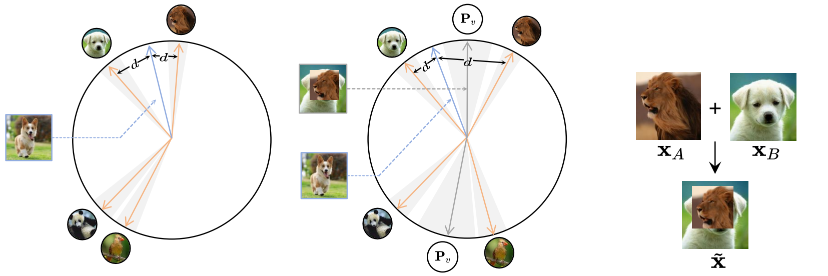
4.1 Generate virtual prototypes

FSCIL typically employs a metric learning approach based on the nearest neighbor concept for the classification of samples. The feature vector is closest to the prototype of the class to which the sample is predicted. However, as shown in Fig. 3 (a), if the proximity between prototypes in the embedding space is excessively close, it results in an elevated error rate during sample classification.

To enhance the separation between prototypes and create a sparser embedding space that exclusively prototypes of genuine categories, [15] employs a method for generating virtual prototypes. This method involves the insertion of generated virtual prototypes into the embedding space. Drawing inspiration from [15], we use virtual prototype generation in the training of the base session to enhance the feature extraction ability of the backbone. Specifically, we employ the CutMix augmentation method to generate virtual samples using samples from true categories and then extract the feature vectors from these virtual samples to synthesize virtual prototypes. A detailed virtual prototype schematic in our method is shown in Fig. 3 (c). Eq. (1) shows the process of cropping and mixing image $\mathbf{x}_B$ onto image $\mathbf{x}_A$, the method called CutMix [12].

$$\tilde{\mathbf{x}} = \mathbf{M} \odot \mathbf{x}_A + (\mathbf{1} - \mathbf{M}) \odot \mathbf{x}_B, \quad \mathbf{M} \in \{0,1\}^{H \times W} \tag{1}$$

Where $\tilde{\mathbf{x}}$ denotes the image with the virtual category resulting from the fusion of $\mathbf{x}_A$ and $\mathbf{x}_B$, $\mathbf{M}$ is a binary mask designed to drop image information, $\odot$ represents the pixel-by-pixel



(a) A normal embedding space

(b) A normal embedding space based on virtical protytypes generation

(c) CutMix-based augmentation

Figure 3: Placement of virtual prototypes in embedding space. (a) The nearest neighbor-based algorithm can encounter challenges in accurate recognition when prototypes of real categories in close proximity within the embedding space. (b) Inserting the virtual prototypes increases the separation between real prototypes and increases the sparsity of the embedding space. (c) Procedure for virtual sample generation.

multiplication of the two images. In our work, we restrict the size of $\mathbf{M}$ to half the size of the original image. This constraint is implemented to guarantee the comprehensive integration of semantic features from both categories into the virtual class. The virtual samples generated by the CutMix method eliminate non-pixel information, allowing the virtual prototypes to closely align with the real categories in embedding space, further amplifying the separation between real prototypes and enabling the feature extractor to extract more discriminative features between classes.

4.2 Branch training in incremental sessions

Many FSCIL studies adopt a strategy of freezing the parameters of the feature extractor following the training on the base class to suppress catastrophic forgetting for deep neural networks in incremental stage [7, 13]. It effectively suppresses forgetting but limits the feature extractor's capacity to learn new categories. To facilitate the adaption of the model's parameters to the new class while preserving its feature extraction capabilities obtained on the base categories, we partition the feature extractor into two parts, $\mathbf{F}_\phi$ and $\mathbf{F}_e$. $\mathbf{F}_\phi$ is the main component of the network, and $\mathbf{F}_e$ is the lesser layer at the tail end of the network. The entire classification model $\mathcal{M}$ in session i consist of $\{\mathbf{F}_\phi, \mathbf{F}_e^t, \mathbf{W}^t\}$, $\mathbf{W}^t$ is the classifier of session i. The feature vector $\mathbf{p}_i$ extracted from the image $\mathbf{x}_i$ by the feature extractor is shown in Eq. (2). The feature extractor extracts all training samples of class k to generate feature vectors and computes the mean to obtain the prototype $\overline{\mathbf{p}^k}$ for that class.

$$\mathbf{p}_i = \mathbf{F}_e(\mathbf{F}_\phi(\mathbf{x}_i)) \tag{2}$$

Where $|\mathcal{C}(k)|$ denotes the total number of training samples of class k. With sufficient data in the base session, we train $\mathbf{F}_\phi$, $\mathbf{F}_e^0$, and $\mathbf{W}^0$. In incremental sessions, the $\mathbf{F}_e^t$ and $\mathbf{W}^t$ $(t \geq 1)$ undergo fine-tuning, and the parameters of $\mathbf{F}_\phi^t$ is frozen. FSCIL differs from conventional CIL in that the new categories in FSCIL comprise only a limited amount of data, so adjusting $\mathbf{F}_\phi$ during incremental training can result in severe overfitting to new categories. Hence, the parameters of $\mathbf{F}_\phi$ are shared among all session models. This training strategy suppresses the forgetting phenomenon, enabling the model to better adapt to new knowledge.

4.3 Model selection based on uncertainty qualification

During FSCIL testing, without access to session labels, mapping samples to the correct model poses a critical challenge. During testing stage, we have a feature extractor backbone $\mathbf{F}_\phi$, a sequence of feature extractor tails $\{\mathbf{F}_e^0, \mathbf{F}_e^1, \ldots, \mathbf{F}_e^n\}$ and a sequence of classifier $\{\mathbf{W}^0, \mathbf{W}^1, \ldots, \mathbf{W}^n\}$. Choosing the right one from a range of models for a test sample is critical. UQ can measure the model's certainty level with fed samples, producing uncertainty values for classification. In this study, we use information entropy to measure the uncertainty of the fed samples. The information entropy is calculated as shown in Eq. (3).

$$H^t(\mathbf{x}) = -\sum_{c=1}^{|\mathcal{C}^t|} p(o^c)\log p(o^c) \tag{3}$$

Where $H^t(\mathbf{x})$ denotes the information entropy of the model in session t for the fed sample $\mathbf{x}$, $|\mathcal{C}^t|$ denotes the total number of categories in session t, $p(o^c)$ represents the probability of class c for fed sample $\mathbf{x}$. During the testing stage, the samples are fed into all session models. Simultaneously, each model computes the uncertainty of the fed samples based on the Eq. (3). $p(o^c)$ is obtained by mapping the output of the neural network to a probability. Finally, the model with the minimum uncertainty is chosen as the classification model for the given sample, and the predicted class from that model is accepted.

4.3.1 Fine tune with real categories

The generation of virtual samples enhances the feature extraction capability of the backbone. However, it concurrently introduces non-existent categories to the classification system. These non-existent categories have an impact on the uncertainty values calculated by the model for the sample. To enhance the trustworthiness of UQ results, we employ real categories to fine-tune the model. Following the training of a robust backbone with samples containing virtual categories, we adjust the number of neurons in the tail of the fully connected layer to match the count of true categories for the present session. Subsequently, we fine-tune the model using training samples exclusively composed of true categories.

4.3.2 Categories imbalance in uncertainty qualification

To assess sample uncertainty across different classification systems, ensuring uniformity in their target categories amount is essential. This standardization is crucial to render the results of the UQ meaningful and comparable for analysis. However, in FSCIL's setting, base sessions comprise many categories, while incremental sessions feature fewer. To address the class imbalance problem for UQ in base and incremental sessions, we further refine the UQ of the base session model. To maintain consistency in target categories between the base session model and incremental session models during UQ, we partition the probabilities generated by the base session model to N_{sub} sub-results, N_{sub} as shown in Eq. 4.

$$N_{sub} = \frac{|\mathcal{C}^0|}{|\mathcal{C}^i|}, \quad |\mathcal{C}^i| = |\mathcal{C}^j|(1 \le i, j \le n) \tag{4}$$

$|\mathcal{C}^0|$ and $|\mathcal{C}^i|$ represent the number of categories in the base session and the incremental session, respectively. The output of the base session model has a total of N_{sub} sub-results, where each sub-results contains the probability of $|\mathcal{C}^i|$ categories. We quantify the uncertainty for each sub-result and select the smallest of them as the UQ result for the whole base session model. In detail, during the testing stage, each sample $\mathbf{x}_i$ is fed into the models on all sessions, resulting in a set of uncertainty values as shown in Eq. (5).

$$\mathbf{U} = \{H^0(\mathbf{x}_i), H^1(\mathbf{x}_i), \cdots, H^n(\mathbf{x}_i)\} \tag{5}$$

Nonetheless, to ensure uniform participation of the target categories in UQ, the N_{sub} sub-results of the base session model are involved in the UQ, rather than $H^0(\mathbf{x}_i)$. The sub-results scheme ensures that the category number of the base session model aligns with $|\mathcal{C}^i|$. Hence, the set of uncertainties obtained into all models is expressed as shown in Eq. (6).

$$\mathbf{U} = \{H_1^0(\mathbf{x}_i), \cdots, H_{N_{sub}}^0(\mathbf{x}_i), H^1(\mathbf{x}_i), \cdots, H^n(\mathbf{x}_i)\} \tag{6}$$

Where $\mathbf{U}_{\text{base}} = \{H_1^0(\mathbf{x}_i), \cdots, H_{N_{sub}}^0(\mathbf{x}_i)\}$ denotes the set of all sub-results from the base session model. The minimum value in $\mathbf{U}_{\text{base}}$ as the uncertainty value of the base session model participates in the UQ containing all session models. This sub-results partitioning strategy for the base session model effectively tackles the category imbalance between the base classes and the incremental classes. This is achieved by aligning the categories in the sub-results of the base session with the incremental class.

5 Experiments

5.1 Dataset

CIFAR-100 contains 100 classes with 600 32×32 RGB images per class, of which 500 are used for training and 100 for testing, and the entire dataset has 600,000 images. 60 classes are used as base classes, 40 classes are used as incremental classes, and there are eight sessions in the incremental stage. The data in each incremental session appears as the 5-way 5-shot. *mini*-ImageNet is a subset of ImageNet that contains a total of 100 classes, each containing 600 84×84 RGB images. 60 classes are used as base classes, and 40 classes are used as incremental classes. The 40 classes are evenly distributed into eight sessions, with five classes per session. The data in each incremental session appears as the 5-way 5-shot.

5.2 Implementation Details

Model Architecture Many FSCIL studies use ResNet [4] as the backbone for classification [7, 13, 17]. We employ ResNet-18 as the backbone network to validate the performance of our proposed method on the benchmark datasets. We follow the setting of [13, 17] for CIFAR-100 and *mini*-ImageNet, and we use randomly initialized model parameters. The entire feature extraction section of ResNet-18 is partitioned into four blocks, with each block consisting of three convolution layers. In our branch training strategy, the first three blocks are $\mathbf{F}_\phi$, where the parameters are frozen after training on the base session. The last block in the tail is $\mathbf{F}_e$, and it continuously updates the parameters during incremental sessions.

Baseline We use these recent methods as baselines in comparative experiments: iCaRL [9], EEIL [1], LUCIR [5], TOPIC [11], CEC [13], F2M [10], Entropy-reg [7], MetaFSCIL [2], GKEAL [17]. The performance of these baselines is mostly sourced from [17] for fair and consistent comparisons. In ablation experiments, the absence of branch training implies that only the classifier undergoes fine-tuning during the incremental process while the parameters of the feature extractor remain frozen.

Training details Our experiment is implemented using the Python version of PyTorch. The optimizer is SGD, with the learning rate set to 0.1 for the base session and 0.05 for incremental sessions. Momentum in SGD is set to 0.9. For CIFAR-100, The training epochs are 200 in the base session and 20 in incremental sessions. For *mini*-ImageNet, the training epochs are 300 in the base session and 30 in incremental sessions.

5.3 Comparison with the state-of-the-art methods

To validate the state-of-the-art performance of our method, we conduct comparative experiments with contemporary approaches on two benchmark datasets: CIFAR-100 and *mini*-ImageNet. The comparative performance of our method compared to other FSCIL methods is shown in Fig. 4. And the detailed performance for CIFAR-100 is shown in Tab. 1.

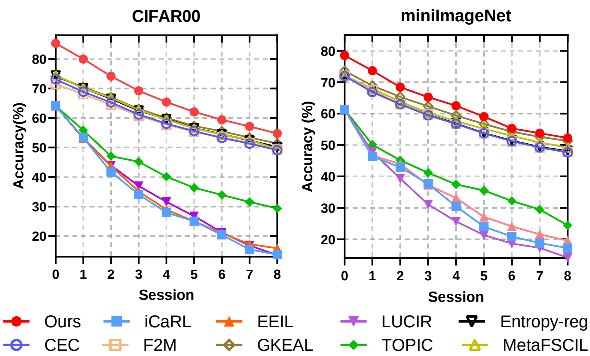


Figure 4: Average accuracy on each session of two benchmark datasets: CIFAR-100, *mini*-ImageNet.

Table 1: The accuracy of each session among the compared methods on CIFAR-100.

Method	Accuracy in each session (%)									AA
	0	1	2	3	4	5	6	7	8	
iCaRL [9]	64.10	53.28	41.69	34.13	27.93	25.06	20.41	15.48	13.73	32.86
EEIL [1]	64.10	53.11	43.71	35.15	28.96	24.98	21.01	17.26	15.85	33.79
LUCIR [5]	64.10	53.05	43.96	36.97	31.61	26.73	21.23	16.78	13.54	34.21
TOPIC [11]	64.10	55.88	47.07	45.16	40.11	36.38	33.96	31.55	29.37	42.62
CEC [13]	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	59.52
F2M [10]	71.45	68.10	64.43	60.80	57.76	55.26	53.53	51.57	49.35	59.13
MetaFSCIL [2]	74.50	70.10	66.84	62.77	59.48	56.52	54.36	52.56	49.47	60.73
Entropy-reg [7]	74.40	70.20	66.54	62.51	59.71	56.58	54.52	52.39	50.14	60.77
GKEAL [17]	74.01	70.45	67.01	63.08	60.01	57.30	55.50	53.39	51.40	61.35
Ours	**85.30**	**79.94**	**74.16**	**69.19**	**65.40**	**62.09**	**59.40**	**57.16**	**54.73**	**67.44**

Based on the experimental results, our approach attains satisfactory performance for both CIFAR-100 and *mini*-ImageNet, demonstrating excellent in base session and incremental sessions. In CIFAR-100, our method outperforms GKEAL [17] by 3.73% in the final session, with an average accuracy improvement of 6.09%. The experimental results unequivocally demonstrate that our method outperforms other state-of-the-art methods for FSCIL datasets. Our proposed parameter-isolation approach in FSCIL effectively mitigates the issue of catastrophic forgetting. Simultaneously, the experimental results substantiate that the UQ technique can construct a mapping from samples to models. This learning method offers a fresh perspective for the FSCIL field.

5.4 Ablation Study

To assess the efficacy of our methods, we conduct ablation experiments on CIFAR-100. The detailed results are provided in Tab. 2. Post application of the virtual prototype generation method, accuracy rose to 82.93%. This suggests the method effectively enhances backbone feature extraction. However, deploying branch training with model selection yielded unsatisfactory performance due to category imbalance in UQ. We address this with the SR strategy during testing, resulting in AA increasing from 58.38% to 65.27%. This underscores the importance of addressing category imbalance in UQ within the FSCIL domain. Further, fine-tuning the model with real categories after virtual prototype enhancement increased AA by 2.17%. In summary, our method raised the last session accuracy from 49.04% to 54.73% and AA from 62.88 to 67.44. This indicates our method effectively resists forgetting and generalizes to new classes.

Table 2: The comprehensive ablation experiment on CIFAR-100. For abbreviations, **AA** represents the average accuracy. **CM** represents the generation methods for virtual prototypes. **BR** represents the branch training strategy. **MS** represents the model selection based on UQ **SR** represents the sub-results of the classification model in the base session. **FT** represents the fine-tuning for the the categories number of classifier.

CM	BR	MS	SR	FT	Accuracy in each session (%)									AA
					0	1	2	3	4	5	6	7	8	
					80.92	75.86	69.81	65.45	61.13	57.39	54.41	51.99	49.04	62.88
✓					82.93	76.35	71.22	66.32	61.88	58.65	55.71	52.71	49.37	63.90
✓	✓	✓			82.93	72.06	64.45	59.03	55.51	51.49	49.08	46.95	43.98	58.38
✓	✓	✓	✓		82.93	77.60	71.90	67.02	63.62	59.92	56.94	54.73	51.79	65.27
✓	✓	✓	✓	✓	85.30	79.94	74.16	69.19	65.40	62.09	59.40	57.16	54.73	67.44

6 Conclusion

In this paper, we propose a FSCIL method that emulates memory storage in the human cerebral cortex. This is the first parameter isolation method in FSCIL. We train distinct classification models for each session to uphold the performance of the entire classification system on old categories. During the testing, we quantify the uncertainty of the samples in deriving session label, facilitating the selection of the appropriate classification model. By emulating the memory storage mechanism of the cerebral cortex, our approach renders the learning process of FSCIL more aligned with the human learning style, offering a novel perspective to the field of FSCIL. In the future, we aim to develop UQ methods that are better suited for limited samples and meticulously consider detailed feature differences.

References

[1] 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), pages 233–248, 2018.

[2] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14166–14175, 2022.

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

[4] 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*, pages 770–778, 2016.

[5] 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*, pages 831–839, 2019.

[6] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022.

[7] Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[8] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

[9] 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*, pages 2001–2010, 2017.

[10] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in neural information processing systems*, 34:6747–6761, 2021.

[11] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12183–12192, 2020.

[12] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[13] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2021.

[14] Jinghua Zhang, Li Liu, Olli Silven, Matti Pietikäinen, and Dewen Hu. Few-shot class-incremental learning: A survey. *arXiv preprint arXiv:2308.06764*, 2023.

[15] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9046–9056, 2022.

[16] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6801–6810, 2021.

[17] Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. Gkeal: Gaussian kernel embedded analytic learning for few-shot class incremental task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2023.