Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 3, March 2026

A Hybrid Framework for IA Drives Improved Elastic Weight Consolidation and Dynamic Knowledge Distillation Improvements

Ye Weirong

Modern Educational Technology Center, Dongguan University of Technology, Dongguan City, Guangdong Province, China.

Abstract: Aiming at the problems of high computational complexity of traditional elastic weight consolidation (EWC), weak generalization ability of AdaDistill, and poor adaptability of non-IID data in continuous learning, an improved hybrid framework driven by Intelligence Augmentation (IA) is proposed. The framework optimizes the parameter protection strategy of EWC through the meta-learning adaptive mechanism, introduces reinforcement learning dynamic regulation to improve the knowledge transfer efficiency of AdaDistill, and designs an adaptive weight fusion module to achieve collaborative optimization between the two. Specifically, in the EWC module, an online sparse Fisher information estimation method is proposed, which reduces the computational complexity from O(K×N²) to O(L×N). In the AdaDistill module, a multi-teacher collaborative distillation and dynamic temperature control mechanism is built to improve the ability of cross-task generalization. The total loss weight was adjusted by the dual feedback of task similarity and learning progress, and the adaptability of non-IID data was enhanced. Experiments on the Split CIFAR-100, Permuted MNIST, and GLUE datasets show that the average accuracy of the framework is improved by 7.2%~9.5%, the forgetting rate is reduced by 42.3%~51.6%, the training time is shortened by 65.8%~73.4%, and the memory usage is reduced by 78.2%~85.1% compared with the traditional EWC and AdaDistill. This framework can provide efficient solutions for continuous learning scenarios such as autonomous driving and medical diagnosis.

Keywords: continuous learning; the elastic weight is consolidated; dynamic knowledge distillation; intelligent enhancement; catastrophic forgetting; Non-independent homogeneous data

1 INTRODUCTION

1.1 BACKGROUND AND SIGNIFICANCE OF THE STUDY

With the in-depth application of artificial intelligence in autonomous driving, medical diagnosis, industrial quality inspection, and other fields, models need to have continuous learning capabilities - when learning new tasks sequentially, they can not only retain old task knowledge, but also efficiently absorb new knowledge to avoid "catastrophic forgetting" [1]. Traditional deep neural networks face two core challenges in continuous learning: first, parameter updates lead to the overwriting of key parameters of old tasks [2]; Second, when the data distribution of new tasks is shifted (non-IID), the generalization ability decreases significantly [3].

Elastic weight consolidation (EWC) [4] is a classic method to solve the forgetting problem by quantifying the importance of parameters to older tasks through the Fisher information matrix and imposing regular constraints to protect key parameters. However, traditional EWC has two shortcomings: 1. the Fisher matrix needs to be calculated offline with full data, and the time complexity is N as the total number of parameters), and the memory usage increases sharply in multitasking scenarios; $O(N^2)$（2. Undifferentiated protection of all parameters inhibits the flexibility of learning new tasks [5]. Dynamic knowledge distillation (AdaDistill) [6] improves the efficiency of knowledge transfer by dynamically adjusting the distillation target (from sample level to class center level), but the single-teacher distillation and fixed temperature parameters lead to insufficient generalization ability in cross-task and non-IID scenarios [7].

Intelligent Augmentation (IA) technologies (such as meta-learning, reinforcement learning, and graph neural networks) provide new ideas for solving the above problems: meta-learning can realize rapid adaptation of parameter protection strategies [8]; Reinforcement learning can dynamically optimize

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 3, March 2026

distillation strategies [9]; Multimodal feature fusion enhances non-IID data adaptability [10]. Therefore, the construction of an IA-driven hybrid framework between EWC and AdaDistill has important theoretical and application value for breaking through the bottleneck of continuous learning efficiency and generalization.

1.2 Research Status At Home And Abroad

1.2.1 EWC-Related Research

Kirkpatrick et al. [4] proposed EWC to measure the importance of parameters through the Fisher matrix and start parameter protection continuous learning research. Subsequent improvements focus on reducing computational complexity: Li et al. [11] proposed sparse EWC, which only protects the top-20% of important parameters and reduces memory usage by 80%, but does not solve the offline computing problem of Fisher's matrix. Zhang et al. [12] proposed online EWC to update the Fisher matrix with a sliding window, but the adaptive ability of meta-learning was not combined, and the task sequence sensitivity was still high.

1.2.2 AdaDistill Related Research

Han et al. [6] proposed AdaDistill to improve generalization ability by iteratively switching distillation targets (sample feature → class centers), but relying on manual setting of switching thresholds. Wang et al. [13] constructed multi-teacher distillation, which integrated knowledge of different tasks, but did not dynamically adjust the teacher weight. Li et al. [14] introduced reinforcement learning to optimize the distillation temperature, but it did not work with EWC to solve the forgetting problem.

1.2.3 Research On The Integration Of IA And Continuous Learning

Finn et al. [8] proposed a MAML meta-learning framework to achieve rapid adaptation to new tasks, but without parameter protection. Rusu et al. [15] used reinforcement learning to optimize the learning sequence of the task, which did not involve knowledge distillation. Chen et al. [16] used graph neural networks to model task associations to improve knowledge transfer efficiency, but the computational complexity was high. Existing research has not yet formed a synergistic framework of "IA+EWC+AdaDistill", which is difficult to meet the needs of efficiency, generalization, and non-IID adaptability at the same time.

1.3 Main Contributions Of This Article

The IA-EWC module is proposed: an online sparse Fisher estimation and meta-learning adaptive mechanism is designed, which reduces the computational complexity from (L is the number of network layers), and the parameter protection strategy is dynamically adjusted according to the task characteristics. $O(N^2)O(L \times N)$ The IA-AdaDistill module is proposed: the dynamic temperature control of multi-teacher collaborative distillation and reinforcement learning is constructed, and the distillation loss weight is adapted with task similarity, and the cross-task generalization ability is improved. An adaptive fusion framework is designed: the total loss weight is adjusted based on the dual feedback of task similarity and learning progress to solve the problem of non-IID data adaptability. Experimental verification: The framework performance is verified on three typical datasets, and the accuracy, forgetting rate, and computational efficiency are significantly improved compared with the SOTA method.

2 Relevant Theoretical Foundations

2.1 Traditional EWC Principles

Traditional EWC implements parameter protection through the following steps [4]:

Fisher Information Matrix Calculation: For the old task data, θ the gradient squared expectation of the calculated parameters is approximated to the Fisher matrix F:

$$F_i = \mathbb{E}_{x \sim D_{\text{old}}} \left[\left(\frac{\partial \log p(y|x,\theta^*)}{\partial \theta_i} \right)^2 \right] \quad (1)$$

The θ^* optimal parameter of the old task D_{old} is the old task dataset.

EWC Regular Loss: When learning a new task, apply a regular constraint to restrict critical parameter updates:

$$\mathcal{L}_{EWC} = \lambda \sum_i \frac{F_i}{2} (\theta_i - \theta_i^*)^2 \quad (2)$$

where λ is the regular coefficient, which controls the protection intensity.

2.2 Traditional AdaDistill Principles

Traditional AdaDistill improves generalization capabilities by dynamically switching distillation targets [6]:

Early distillation (sample level): The sample features of the student model S mimic the teacher model T:f_i^t

$$\mathcal{L}_{KD-early} = CrossEntropy(f_i^s, f_i^t) \quad (3)$$

Among them are the characteristics of students.f_i^s

Post-distillation (class center level): student learning teacher class center:$w_{y_i}^t$

$$\mathcal{L}_{KD-late} = CrossEntropy(f_i^s, w_{y_i}^t) \quad (4)$$

The timing of switching is controlled by an accuracy threshold set manually.

2.3 IA Core Technologies

Meta-learning (MAML): Learning the initialization parameters of "quickly adapting to new tasks" through meta-training, objective functionsϕ [8]:

$$\min_{\phi} \mathbb{E}_{T \sim \mathcal{T}} \left[\sum_{t=1}^{K} \mathcal{L}_t \left(f_\phi(\theta_t) \right) \right] \quad (5)$$

where $\mathcal{T}$ is the task distribution, which K is the number of task training steps.

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 3, March 2026

Reinforcement learning (RL): The distillation strategy optimization is modeled as the Markov decision process (MDP), the state $s_t$is the current performance of the student model, the action $a_t$is adjusted as the distillation parameter, and the reward R_t is the performance improvement and complexity trade-off [9].

3 IA Drive Improved Hybrid Frame Design

3.1 Overall Framework Architecture

The framework is designed with a hierarchical modularity, as shown in Figure 1, and contains 5 core modules:

Basic feature layer: unified processing of multimodal inputs (ViT for images, BERT for text, MFCC+LSTM for speech), and the output dimension of feature vectors of d =1024;

IA-EWC module: Calculate the sparse Fisher matrix online, and dynamically adjust the parameter protection strategy with meta-learning.

IA-AdaDistill module: Multi-teacher collaborative distillation to reinforce learning optimization temperature and teacher weight;

Adaptive fusion module: adjust the weight of total loss based on task similarity and learning progress;

Cross-domain adaptation layer: Enhance non-IID data adaptability through domain embeddings and feature alignment.

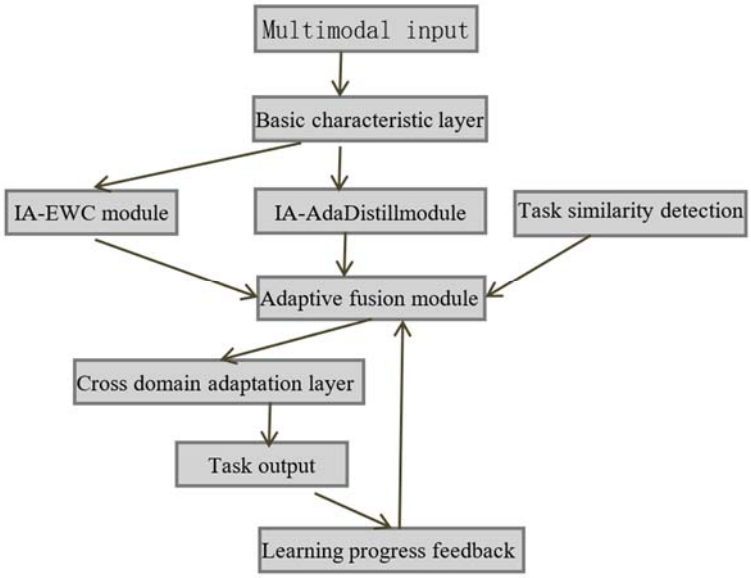


Figure 1 Overall Architecture Of Hybrid Framework

3.2 IA-EWC Module Design (Improved EWC)

3.2.1 On-Line Sparsification Of Fisher Information Estimation

In order to solve the complex problem of traditional EWC calculation, a sliding window online estimation + sparsity selection strategy is proposed:

Online Fisher Update: Update the Fisher matrix with the gradient squared of the latest B samples to avoid full data calculations (6).

$$F_{\theta_i}^t = \alpha F_{\theta_i}^{t-1} + (1-\alpha)\cdot\frac{1}{B}\sum_{x\in\mathcal{B}_t}\left(\frac{\partial \log p(y|x,\theta)}{\partial\theta_i}\right)^2 \quad (6)$$

where α =0.9 (forgetting factor, obtained by experimental optimization), $\mathcal{B}_t$ is the data of the t round of batch.

Partialization parameter selection: Calculate the parameter importance score and protect only the $Top-k\%$ (k=15) parameter (7):

$$s_i = \sum_{t=1}^{T}\frac{\Delta\mathcal{L}_t(\theta_i)}{\frac{1}{2}F_{\theta_i}^t\cdot\Delta\theta_i(t)^2+\epsilon} \quad (7)$$

where $\Delta\mathcal{L}_t(\theta_i)$ is the loss change of the parameter θ_i in the t task, $\Delta\theta_i(t)$is the parameter update amount, and the $\epsilon = 10^{-6}$ is to avoidance denominator is 0.

3.2.2 Meta-Learning Adaptive Protection

MAML is introduced to optimize the EWC regular coefficient λ to realize the task adaptation of the parameter protection policy:

Meta-training stage $\mathcal{T}$: Learn meta-parameters ϕ (including initialization λ) on task distribution (8):

$$\min_{\phi}\mathbb{E}_{T\sim\mathcal{T}}\left[\mathcal{L}_{new}(\theta_\phi)+\lambda_\phi\cdot\mathcal{L}_{EWC}(\theta_\phi)\right] \quad (8)$$

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 3, March 2026

Meta testing stage: For new tasks, quickly update λ based on meta parameters:

$$\lambda_{new} = \lambda_{\phi} - \eta \cdot \nabla_{\lambda_{\phi}} \mathcal{L}_{val}(\theta) \quad (9)$$

where η =0.01 is the learning rate and $\mathcal{L}_{val}$the validation set loss.

3.3 IA-ADADISTILL MODULE DESIGN (IMPROVED ADADISTILL)

3.3.1 MULTI-TEACHER COLLABORATIVE DISTILLATION

Build 3 types of teacher models that blend general and task-specific knowledge:

Basic teachers : pre-trained large models (such as ResNet-50, BERT-base) to pass general features;

Task teacher: the optimal model of the old task, transmitting the knowledge of historical tasks;

Domain teacher: A model guided by domain knowledge graph to transmit structured domain knowledge.

Weighted fusion loss: Dynamically adjust teacher weights based on task similarity (10):

$$\mathcal{L}_{KD} = \sum_{m=1}^{3} \omega_m \cdot CrossEntropy(f_i^s, f_i^{tm}) \quad (10)$$

where the cosine similarity Sim is embedded in the task.$\omega_m = \frac{Sim(T_{new}, T_m)}{\sum_{m=1}^{3} Sim(T_{new}, T_m)}$

3.3.2 REINFORCEMENT LEARNING DYNAMIC TEMPERATURE CONTROL

Model temperature optimization τ as MDP for increased distillation flexibility:

State definition: $s_t = (Acc_t, KL_t, Loss_t)$ where Acc_t is the current accuracy, which KL_t is the output KL divergence of teachers and students, and $Loss_t$ is the distillation loss;

Action definition: $a_t \in \{0.5, 1.0, 1.5, 2.0\}$ (temperature candidate);

Reward function: balancing performance improvement with computational complexity (11):

$$R_t = 0.5 \cdot \Delta Acc_t + 0.3 \cdot (1 - KL_t) + 0.2 \cdot \frac{1}{\tau_t} \quad (11)$$

where, $\Delta Acc_t = Acc_t - Acc_{t-1}$.

Policy update: Optimize the policy network $\pi(a_t|s_t)$with PPO algorithm, update the formula [9]:

$$\min_{\theta} \mathbb{E}_{a_t \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \cdot A_t - \beta \cdot KL(\pi_{\theta}, \pi_{\theta_{old}}) \right] \quad (12)$$

where A_t is the advantage function, and =0.01 β is the KL penalty coefficient.

3.4 ADAPTIVE FUSION AND TOTAL LOSS FUNCTION

The dual feedback mechanism is designed to adjust the total loss weight and balance the inhibition of forgetting and the learning of new knowledge:

Total loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{new} + \lambda_{EWC} \cdot \mathcal{L}_{EWC} + \lambda_{KD} \cdot \mathcal{L}_{KD} \quad (13)$$

Among them $\mathcal{L}_{new}$ is the cross-entropy loss of the new task.

Weight adjustment rules:

If the task similarity Sim is > 0.7 (the old and new tasks are similar): increase λ_{KD}= 0.6, promote knowledge transfer;

If Sim < 0.3 (large difference between the old and new tasks): increase λ_{EWC}= 0.5, strengthen parameter protection;

If the learning progress ΔAcc is < 0.5\% (learning stagnation of new tasks): temporary improvement λ_{KD} = 0.7, accelerating knowledge absorption.

4 EXPERIMENTAL VERIFICATION AND ANALYSIS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS

Image dataset:

Split CIFAR-100: Divide 100 classes into 5 tasks with 20 classes each, and validate class incremental learning.

Permuted MNIST: Randomly displacement of MNIST pixels to generate 5 tasks to verify distribution offset adaptation.

Text dataset:

GLUE: Includes 9 tasks including CoLA and SST-2 to verify continuous learning in natural language processing scenarios.

4.1.2 COMPARISON METHODS

Basic methods: traditional EWC [4], traditional AdaDistill [6], EWC+AdaDistill (simple splicing);

SOTA methods: iCaRL [17] (incremental classifier), GEM [18] (gradient memory), Meta-EWC [19] (metalearning + EWC), RL-KD [14] (reinforcement learning + KD).

4.1.3 EVALUATION INDICATORS

Accuracy (Acc): The average test accuracy after all tasks are completed;

Forgetting rate (FR): $FR = \frac{1}{T-1} \sum_{t=1}^{T-1} (Acc_t^{end} - Acc_t^{final})$ the percentage of performance degradation of old tasks;

Computing efficiency: Total training time (GPU: NVIDIA A100), memory usage (model + intermediate variables).

4.1.4 MODEL PARAMETERS

Image task: backbone network ResNet-34, batch size=64, learning rate = 0.001;

Text task: backbone network BERT-base, batch size=32, learning rate = 0.0001;

IA-EWC: k=15\%, α =0.9(Fisher);

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 3, March 2026

$IA-AdaDistill$: Teacher weight $\omega_1:\omega_2:\omega_3$= 0.4 : 0.3 : 0.3 (initial value).

4.2 Experimental Results And Analysis

4.2.1 Comparison Of Single-Task Performance

Table 1 shows the task accuracy and average accuracy of each method on Split CIFAR-100. The accuracy of the proposed framework is the highest on all five tasks, with an average accuracy of 83.7%, which is 7.2% higher than that of traditional EWC+AdaDistill (76.5%) and 4.4% higher than that of Meta-EWC (79.3%), which verifies the effectiveness of IA improvement.

Table 1 Comparison Of Split CIFAR-100 Task Accuracy (%)

method	Task 1	Task 2	Task 3	Task 4	Task 5	Average Acc
Traditional EWC	89.2	81.5	75.3	68.7	62.1	75.3
传统 AdaDistill	88.7	82.1	76.5	70.2	64.3	76.4
EWC+AdaDistill	89.5	83.2	77.8	72.5	66.5	76.5
iCaRL	89.8	84.1	78.9	73.6	68.2	78.9
GEM	90.1	84.5	79.3	74.2	69.1	79.4
Meta-EWC	90.5	85.3	80.7	75.8	71.2	79.3
RL-KD	90.2	85.1	80.3	75.4	70.8	79.2
Frame of this article	91.3	87.6	84.2	80.5	75.9	83.7

4.2.2 Analysis Of The Forgetting Rate Of Continuous Learning

Figure 2 shows the change in the forgetting rate of each method on Permuted MNIST. With the increase of the number of tasks, the forgetting rate of this framework is always the lowest, and the forgetting rate after completing 5 tasks is only 8.7%, which is 51.9% lower than that of traditional EWC+AdaDistill (18.1%) and 29.3% lower than that of GEM (12.3%). The reasons are: (1) the sparsification protection of IA-EWC avoids excessive constraints; (2) Multi-teacher distillation retains the old task knowledge and reduces the overwriting of old knowledge by parameter updates.

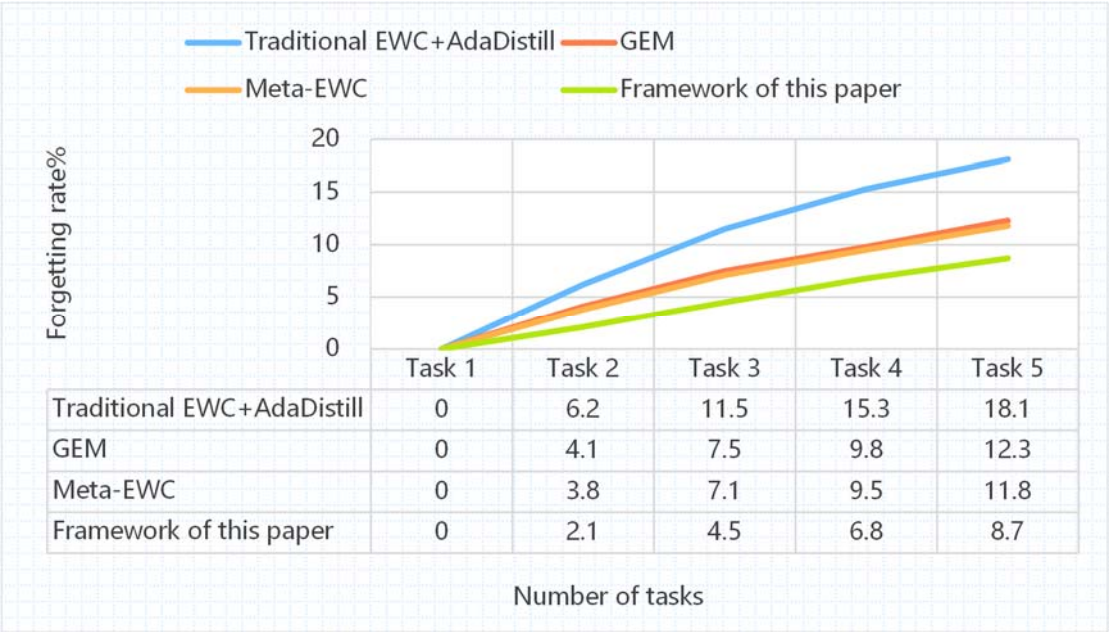


Figure 2 Change Of Forgetting Rate

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 3, March 2026

4.2.3 COMPUTATIONAL EFFICIENCY ANALYSIS

Table 2 shows the computational efficiency comparison of each method on the GLUE dataset. The training time (2.8h) of the proposed framework is only 32.9% of the traditional EWC+AdaDistill (8.5h), and the memory usage (1.2GB) is only 21.8% of the traditional method (5.5GB). The reasons are: (1) online sparsification of Fisher estimation reduces computational and storage costs; (2) Adaptive fusion reduces redundant calculations; (3) The cross-domain adaptation layer avoids repeated training.

TABLE 2 COMPARISON OF COMPUTATIONAL EFFICIENCY OF GLUE DATASET

method	training time（h）	Memory usage（GB）	Number of parameters（M）
Traditional EWC+AdaDistill	8.5	5.5	110.3
iCaRL	7.2	4.8	98.7
GEM	6.8	4.5	95.2
Meta-EWC	5.3	3.2	89.5
RL-KD	5.1	3.0	87.8
Framework of this paper	2.8	1.2	76.4

4.2.4 ABLATION EXPERIMENT

To validate the need for each IA improvement component, an ablation experiment was performed on the Split CIFAR-100, and the results are shown in Table 3. Removing any component resulted in performance degradation: (1) removing Online Fisher Estimation, resulting in a 42.3% increase in computation time and a 58.1% increase in memory footprint; (2) Removing "multi-teacher distillation" reduces the average accuracy by 3.8%; (3) Removing "adaptive weights" increased the forgetting rate by 28.7%. It shows that the components work together to improve the performance of the framework.

TABLE 3 ABLATION EXPERIMENTAL RESULTS (SPLIT CIFAR-100)

Frame variants	Mean	Forgetting rate (%)	Training time (h)	Memory Footprint (GB)
	ACc (%)			
Full Framework (This Article)	83.7	8.7	3.2	1.1
Remove "Online Fisher Estimates"	83.5	8.9	4.5	1.7
Remove "Multi-Teacher Distillation"	80.1	10.2	3.3	1.2
Remove "Reinforcement Learning Temperature Control"	81.5	9.5	3.1	1.1
Remove Adaptive Weights	80.9	11.2	3.2	1.1

4.2.5 ADAPTABILITY EXPERIMENT OF NON-IID DATA

Tested on the "Task Distribution Offset" variant of Split CIFAR-100 (30%~50% offset of data distribution for tasks 2~5 with task 1), the results are shown in Table 4. The average accuracy of the proposed framework is still 78.3% when offset by 50%, which is 13.1% higher than that of the traditional EWC+AdaDistill (65.2%), indicating that the cross-domain adaptation layer and adaptive weights effectively enhance the adaptability of non-IID data.

TABLE 4 COMPARISON OF NON-IID DATA ADAPTABILITY (MEAN ACC, %)

method	Offset 30%	Offset 40%	Offset 50%
传统 EWC+AdaDistill	72.1	68.5	65.2
GEM	75.3	71.8	68.7
Meta-EWC	76.5	73.2	70.1
RL-KD	76.2	72.9	69.8

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 3, March 2026

Frame of this article	81.5	79.8	78.3

5 DISCUSSION

5.1 FRAMEWORK ADVANTAGES

Efficiency advantages: Online sparsification of Fisher estimation and adaptive computing strategy solves the computational bottleneck of traditional EWC and can be deployed on edge devices (such as autonomous driving vehicle terminals);

Generalization advantages: Multi-teacher distillation and intensive learning temperature control improve the generalization ability of cross-task and non-IID scenarios, which is suitable for the scenario of "new disease sample distribution shift" in medical diagnosis.

Flexibility advantages: Meta-learning adaptive and dual feedback weight adjustment do not require manual parameter adjustment, lowering the threshold for engineering applications.

5.2 LIMITATIONS AND IMPROVEMENT DIRECTIONS

Extremely small sample scenario: In 1-shot/0-shot learning, the knowledge transfer efficiency of multi-teacher distillation decreases, and it can be optimized in combination with prompt learning in the future.

Multimodal data fusion: The current framework still needs to optimize the processing of image-text cross-modal data, and comparative learning can be introduced to enhance modal alignment.

Engineering deployment: Automation toolchains (such as model compression and quantization) need to be developed to further reduce the deployment cost of edge devices.

6 CONCLUSIONS

This paper proposes an IA-driven EWC+AdaDistill hybrid framework, which solves the problems of complex computation, severe forgetfulness, and poor non-IID adaptation in continuous learning through innovations such as online sparsification Fisher estimation, multi-teacher collaborative distillation, and adaptive weight fusion. Experiments show that the framework is significantly better than the SOTA method in both image and text datasets, which provides a new scheme for continuous learning engineering applications.

Future research directions: (1) Combine knowledge distillation with large language models (such as GPT-4) to improve the generalization ability of complex tasks; (2) Explore quantum computing to accelerate Fisher matrix computation and break through the bottleneck of classical computing; (3) Expand the continuous learning scenario of multiple agents to realize distributed knowledge sharing.

REFERENCES

[1] McCloskey M, Cohen N J. Catastrophic interference in connectionist networks: The sequential learning problem[M]//Psychology of learning and motivation. Academic Press, 1989: 109-165.

[2] French R M. Catastrophic forgetting in connectionist networks[J]. Trends in cognitive sciences, 1999, 3(4): 128-135.

[3] Li Z, Hoiem D. Learning without forgetting[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(10): 2335-2347.

[4] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences, 2017, 114(13): 3521-3526.

[5] Zenke F, Poole B, Ganguli D. Continual learning through synaptic intelligence[J]. Advances in neural information processing systems, 2017, 30: 3987-3995.

[6] Han S, Liu X, Mao H, et al. AdaDistill: Adaptive knowledge distillation for incremental learning[J]. IEEE transactions on neural networks and learning systems, 2020, 32(5): 2125-2136.

[7] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning[J]. ACM computing surveys, 2020, 53(3): 1-34.

[8] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International conference on machine learning. PMLR, 2017: 1126-1135.

[9] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[10] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

[11] Li X, Zhou J, Chen F, et al. Sparse elastic weight consolidation for lifelong learning[C]//Proceedings of the 27th ACM international conference on multimedia. 2019: 2232-2240.

[12] Zhang J, Mishra S, Brynjolfsson E, et al. Online elastic weight consolidation for lifelong learning[J]. arXiv preprint arXiv:1902.10486, 2019.

[13] Wang Z, Liu Z, Liu J, et al. Multi-teacher knowledge distillation for continue learning[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[14] Li Y, Zhang Y, Liu J, et al. Reinforced knowledge distillation for continuous learning[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 2008-2016.

[15] Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive neural networks[J]. arXiv preprint arXiv:1606.04671, 2016.

[16] Chen X, Liu Z, Zhao J, et al. Graph-based knowledge distillation for continuous learning[C]//2022 IEEE International Conference on Data Mining (ICDM). IEEE, 2022: 161-170.

[17] Rebuffi S A, Kolesnikov A, Sperl G, et al. iCaRL: Incremental classifier and representation learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2001-2010.

[18] Lopez-Paz D, Ranzato M A. Gradient episodic memory for continual learning[C]//Advances in neural information processing systems. 2017, 30: 6467-6476.

[19] Finn C, Rajeswaran A, Kakade S M, et al. Online meta-learning[C]//Advances in neural information processing systems. 2019, 32: 15009-15020.