Mutual information optimization for mitigating catastrophic forgetting in continual learning: An information-theoretic approach

Ngo Huu Phuc*, Vi Bao Ngoc, Phan Hai Hong, Nguyen Chi Cong

Institute of Information and Communication Technology, Military Technical Academy, 236 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam.
*Corresponding author: phucnh@lqdtu.edu.vn
Received 14 May 2025; Revised 4 Jul. 2025; Accepted 20 Sep. 2025; Published 2 Oct. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.106.2025.129-136

ABSTRACT

Continual learning systems encounter the critical challenge of catastrophic forgetting, where neural networks lose previously acquired knowledge when adapting to new tasks. In this paper, we propose Continual Mutual Information Preservation (CMIP), an information-theoretic approach that leverages Mutual Information (MI) optimization and entropy regularization to retain prior knowledge while learning compact and informative latent representations. CMIP uses an auxiliary network to estimate MI and a replay memory, in which each mini-batch comprises 50% current-task samples and 50% samples replayed from previous tasks. Experiments are conducted on the MNIST-Split and CIFAR-100-Split datasets for the class-incremental learning (Class-IL) setting. On MNIST-Split, CMIP achieves 90.97% accuracy with an 8.81% forgetting rate, outperforming EWC (20.64% accuracy, ~77% forgetting) and GEM (65.1% accuracy, ~33% forgetting). This method is applicable to real-world scenarios, such as robotic perception and real-time data streams.

Keywords: Continual learning; Catastrophic forgetting; Mutual information; Information theory; Neural networks; Memory replay.

1. INTRODUCTION

Continual learning enables models to sequentially acquire knowledge from a series of tasks T_i with $i=1,2,...,n$ without compromising previously learned capabilities - a fundamental requirement for practical applications, including robotic learning (where systems must recognize novel objects while retaining the ability to recognize previously encountered ones) or real-time data processing. However, the phenomenon of "catastrophic forgetting" poses a significant challenge, as neural networks tend to overwrite previously acquired knowledge when exposed to new task distributions [1].

Existing methodologies, including Elastic Weight Consolidation (EWC) [2, 9, 10] and Gradient Episodic Memory (GEM) [3, 9, 10], have demonstrated partial success in addressing this challenge. EWC uses weight regularization to protect parameters for previous tasks, but its effectiveness diminishes when task distributions differ significantly. Conversely, GEM necessitates extensive memory storage, posing challenges for embedded-system deployment.

Information theory provides a rigorous framework for optimizing latent representations. It encompasses concepts such as mutual information (MI) and entropy. MI quantifies statistical dependencies between variables, offering valuable insights for knowledge preservation [[5]]. Entropy facilitates the elimination of redundant information, thereby enhancing model generalization capabilities [[7]]. Recent advances, including MI optimization in self-supervised learning [[6]] and information flow analysis in neural architectures [[7]], demonstrate the significant potential of information-theoretic principles.

We introduce Continual Mutual Information Preservation (CMIP), an enhanced information-theoretic framework that extends Information Bottleneck (IB) [[4]] and Mutual Information Neural Estimation (MINE) [[5]] principles to continual learning contexts. CMIP leverages MI

optimization to preserve information from previously learned tasks while incorporating entropy regularization to reduce redundant information content. Our primary contributions include:

1. The CMIP algorithm which integrates MI optimization, entropy regularization, and replay memory with mixed mini-batch training to mitigate catastrophic forgetting.
2. Theoretical justification demonstrating how MI optimization between current representations and previous task labels prevents catastrophic forgetting while entropy regularization eliminates redundant information.
3. Empirical evaluation on MNIST-Split and CIFAR-100-Split datasets within Class-IL scenarios.

We focus on the Class-IL setting, which is widely regarded as the most challenging scenario in continual learning [10]. While Task-IL and Domain-IL are also important, prior studies [10] have shown that many methods - including simple baselines - already achieve near-perfect performance on these settings, especially on datasets like MNIST. Therefore, we prioritize Class-IL to better assess the effectiveness of CMIP under more realistic and demanding conditions.

To ensure fair and direct comparison with prior work, such as EWC [2,10] and GEM [3,10], we conduct experiments on widely used benchmark datasets: MNIST-Split and CIFAR-100-Split. These datasets, while relatively simple, are standard in the continual learning literature and allow for reproducible evaluation under controlled conditions. Future work may extend CMIP to more complex and real-world datasets such as TinyImageNet, CORe50, or robotic/IoT data streams.

The remainder of the paper is structured as follows: Section 2 presents theoretical foundations; Section 3 details the CMIP methodology; Section 4 describes experimental configurations and results; Section 5 provides a discussion; Section 6 concludes our findings.

2. THEORETICAL FOUNDATIONS

2.1. Continual learning and catastrophic forgetting

Continual learning requires models to sequentially master tasks T_i with *i=1,2,...,n*, each defined by a distinct data distribution $D_i = \{(x_i, y_i)\}$. For instance, MNIST-Split scenarios T_1 involve digit classification for classes 0-1, while T_2 involve digit classification for classes 2-3. Catastrophic forgetting manifests when training on task T_i substantially degrades performance on previously learned tasks T_{i-1} [[1]]. EWC protects weights using Fisher information, achieving approximately 20.64% average accuracy with ~77% forgetting on MNIST-Split. GEM requires substantial memory resources, attaining ~65.1% average accuracy with ~33% catastrophic forgetting on MNIST-Split. However, EWC faces limitations when task distributions exhibit significant divergence, while GEM requires considerable memory resources, posing challenges for embedded-system deployment.

2.2. Information-theoretic foundations

The principal information-theoretic tools employed in our approach include:

- Entropy *H(X)*: Quantifies uncertainty within variable *X*:

  For discrete variables:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{1}$$

  For continuous variables:

$$H(X) = -\int_X p(x) \log p(x) dx \tag{2}$$

- Mutual information (MI) *I(Z;Y)*: measures the statistical dependence between representation *Z* and labels *Y*:

$$I(Z;Y) = \sum_{z,y} p(z,y) \log \frac{p(z,y)}{p(z)\,p(y)} \tag{3}$$

• Information bottleneck (IB) [[4]]: an optimization principle that compresses representation Z (minimizing $I(X;Z)$ while retaining $I(Z;Y)$):

$$\min_{Z} I(X;Z) - \beta I(Z;Y) \tag{4}$$

where, β represents the balance coefficient.

• Mutual information neural estimation (MINE) [[5]]: employs an auxiliary neural network T_θ to estimate MI through:

$$I(Z;Y) \geq E_{p(z,y)}\left[T_\theta(z,y)\right] - \log E_{p(z)p(y)}\left[e^{T_\theta(z,y)}\right] \tag{5}$$

MINE proves particularly suitable for continual learning due to its capability to estimate MI on complex data distributions.

2.3. Information theory in continual learning

Our central hypothesis involves optimizing MI between the current task representation Z and previous task labels Y_{i-1} to preserve prior knowledge, while simultaneously reducing entropy $H(Z)$ to eliminate redundant information. These components integrate into the overall loss function, ensuring models effectively learn new data while retaining information from previous tasks.

3. PROPOSED METHOD: CONTINUAL MUTUAL INFORMATION PRESERVATION (CMIP)

3.1. Core concepts

CMIP addresses catastrophic forgetting through three fundamental components:

1. Cross-entropy loss minimization L_{CE} for current task performance.

2. MI loss optimization L_{MI} to preserve information from previous tasks by estimating MI between latent representation Z and labels Y_{i-1} through an auxiliary network T_θ.

3. Entropy regularization L_{LEN} to compress representation Z, reducing noise and enhancing generalization capabilities.

3.2. Loss function formulation

We define the CMIP loss as:

$$L = L_{CE} + \lambda_1 L_{MI} + \lambda_2 L_{ENT} \tag{6}$$

where the terms are

• L_{CE} represents cross-entropy loss, measuring classification error:

$$L_{CE} = -\sum_{y_i \in Y_i} p(y_i) \log \hat{p}(y_i)$$

with Y_i denoting current task labels and $\hat{p}(y_i)$ representing predicted probabilities.

• L_{MI} is computed according to the MINE-based formulation:

$$L_{MI} = -\left(E_{p(z,y)}\left[T_\theta(z,y)\right] - \log E_{p(z)p(y)}\left[e^{T_\theta(z,y)}\right]\right)$$

where: $z \in Z = f_\phi(X)$ represents latent representation from primary model f_ϕ, X denotes input. set, $y \in Y$ represents task labels, T_θ constitutes auxiliary MI network with parameters θ.

• L_{ENT} represents entropy loss of latent representation:

$$L_{ENT} = -\sum_{z} p(z) \log p(z)$$

• $\lambda_1, \lambda_2 \in [0,1]$ are balance coefficients (e.g., $\lambda_1 = 0.1,\ \lambda_2 = 0.01$).

3.3. Replay memory and mini-batch construction

To mitigate catastrophic forgetting, we store N representative samples in replay memory

following each task completion (e.g., *N=50*). During training, each mini-batch comprises: 50% samples from the current task dataset and 50% samples retrieved from the replay memory of previous tasks. This approach enables models to "rehearse" information from previous tasks while optimizing for new data.

3.4. CMIP algorithm

Algorithm: Continual mutual information preservation (CMIP):

Input: Tasks T_i with data $D_i = \{(x_i, y_i)\}$, primary model f_ϕ (e.g., CNN LeNet-like architecture for image data), auxiliary network T_θ for MI estimation, and hyperparameters λ_1, λ_2, η, N.

Output: Continuously trained model f_ϕ.

Procedure:

1. Initialize parameters ϕ and θ randomly.
2. For each task T_i:
   - Sample data from D_i and train a model using mixed mini-batches comprising 50% current samples and 50% replay samples (if available).
   - Compute:
     - $z_i = f_\phi(x_i)$,
     - L_{CE} from classification predictions,
     - If $i = 1$: Train using standard cross-entropy loss only,
     - If $i > 1$: Utilize replay samples to compute L_{MI} between previous sample representations z_{i-1} and their labels y_{i-1},
     - L_{ENT} from z_i (post-softmax).
   - Update θ according to [[5]].
   - Update ϕ using:

$$\phi \leftarrow \phi - \eta \nabla_\phi \left(L_{CE} + \lambda_1 L_{MI} + \lambda_2 L_{ENT} \right)$$

   - Update the replay memory for the current task with N selected samples.
3. Return f_ϕ.

3.5. Theoretical analysis

Within the CMIP framework, the primary objective involves maintaining useful information from previously learned tasks to mitigate catastrophic forgetting during new data training. Key factors include:

1. MI optimization: MI quantifies the dependency between latent representation *Z* and labels *Y*. The MI definition between Z and Y follows equation (4) and is demonstrated in [[5]]. In CMIP, maximizing $I(Z; Y_{i-1})$ between current task representations and previous task labels ensures critical information from the past remains preserved. The auxiliary MI network estimates this value using MINE methodology, effectively "rehearsing" previously learned task information.

2. Entropy regularization: Representation Z entropy is computed using equation (1).

$$H(Z) = -\sum_z p(z) \log p(z)$$

Reducing $H(Z)$ (eliminating redundant information) compresses latent representations and removes noise, thereby improving model generalization capabilities. The combination of MI optimization and entropy reduction creates a mechanism analogous to information bottleneck (IB)

principles expressed in equation (4). In CMIP, the coefficient β is replaced by balance coefficients λ_1 (for MI loss) and λ_2 (for entropy loss), maintaining reasonable trade-offs between retaining prior knowledge and learning new data efficiently.

3. Compared to previous methods like EWC (parameter protection only) or GEM (based on sample storage and reuse), CMIP focuses on representation optimization through MI and entropy regularization. Consequently, models not only learn current tasks effectively but also maintain critical characteristics from previous tasks, significantly reducing catastrophic forgetting to substantial levels (average ~8.81% on MNIST-Split data with Class-IL problems).

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

We evaluate CMIP on MNIST-Split and CIFAR-100-Split, which are standard benchmarks in continual learning. These datasets are selected to enable direct comparison with prior methods and to validate the core effectiveness of our approach under controlled settings.

MNIST-Split for Class-IL problems, containing 60,000 training samples; 10,000 test samples; with size 28x28; divided into 5 tasks from T_1 to T_5; each task ~6,000 samples; normalized with mean = 0.1307, std = 0.3081, and CIFAR-100-Split for Class-IL problems, containing 50,000 training and 10,000 test images of size 32×32; divided into 10 tasks of 10 classes each (approximately 500 samples per class). We compare against fine-tuning (cross-entropy only), EWC [2, 10], and GEM [3, 10] (100 replay samples per task).

Metrics include average accuracy across all tasks

$$T_i \ \ (i = 1,2,\ldots,n):\ Acc_{avg} = \frac{1}{n}\sum_{i=1}^{n} Acc_i$$

where, Acc_i represents model accuracy on task T_i after completing the entire learning process, and Acc_{avg} denotes the overall model average accuracy.

Average forgetting measure:

For task T_i (with $i \geq 2$), we have:

- $Acc_{i,i}$: Best accuracy immediately after completing task T_i.
- $Acc_{i,n}$: Accuracy of task T_i after learning all tasks (final model task).

Forgetting for each task T_i is computed as: $Forget_i = max\left(0, Acc_{i,i} - Acc_{i,n}\right)$

The average forgetting is calculated as:

$$Forget_{avg} = \frac{1}{n-1}\sum_{i=1}^{n-1} Forget_i$$

For CMIP on MNIST-Split Class-IL problems, the primary model f_ϕ employs LeNet-like CNN architecture optimized for MNIST data (28×28 grayscale images) (Convolution layer 1: 1 input channel; 32 filters; kernel size: 5; padding: 2; activation: ReLU; MaxPool 2×2. Convolution layer 2: 32 input channels; 64 filters; kernel size: 5; padding: 2; activation: ReLU; MaxPool 2×2. Fully connected layer 1: input (post-pooling) 7×7×64; 256 neurons, dropout: 0.2, activation: ReLU. Fully connected layer 2: 128 neurons (generating latent representation *z*), dropout: 0.2, activation: ReLU. Classification layer: 2 output neurons. Auxiliary MI Net T_θ architecture contains 2 layers (64, 32 neurons). Parameters: λ_1=0.1, λ_2=0.01, Adam optimizer $\eta = 0.001$, batch size 50, 5 epochs/task, memory 50 samples/task. CMIP experiments run on CPU (6 cores, 32 GB RAM); EWC, GEM data from [[2], [3], [9], [10]] and re-experimentation.

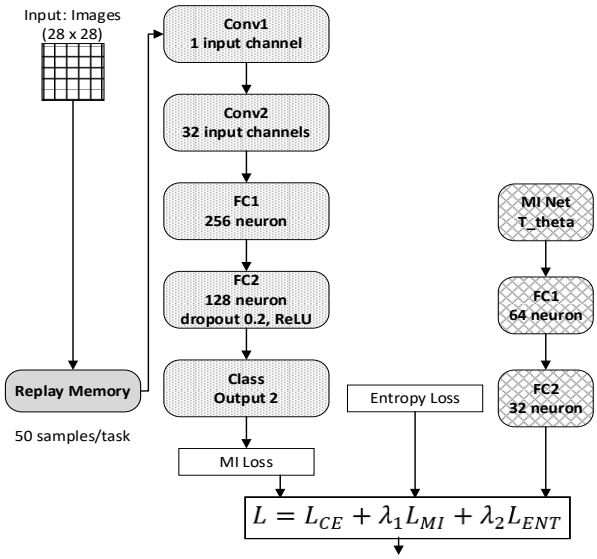

Figure 1. *Proposed CMIP method architecture.*

4.2. Results

Table 1 demonstrates CMIP's superior performance across all datasets. On MNIST-Split, CMIP achieves 90.97% accuracy with 8.81% forgetting, surpassing EWC (20.64%, ~77%) and GEM (65.1%, ~33%). On CIFAR-100-Split, CMIP attains 31.29% accuracy with 20.63% forgetting, outperforming GEM (20.38%) and EWC (8.24%).

Table 1. *Average accuracy and catastrophic forgetting. Data marked with '*' represents our experimental results on the corresponding datasets.*

Method	Class-IL with MNIST-Split		Class-IL with CIFAR-100-Split	
	Accuracy	**Forget**	**Accuracy**	**Forget**
Fine-tune	19.80% *	99.79% *	~3.7% [[9], [10]]	-
EWC	20.64% [[2], [10]]	~77% [[2], [10]]	~8.24% [[10]]	-
GEM	65.1% [[3], [10]]	~33% [[3], [10]]	~20.38% [[3], [10]]	-
CMIP (proposed)	**90.97%**	**8.81%**	**31.29%**	**20.63%**

The results summarized in table 2 - obtained using 10 training epochs per task - highlight CMIP's strong computational efficiency. On the Split-MNIST Class-IL benchmark, CMIP completes all five tasks in just 2.3 minutes of wall-clock time, only slightly slower than GEM (1.7 minutes) and EWC (1.9 minutes), while being over four times faster than fine-tuning (9.4 minutes). Its peak RAM usage remains below 850 MB (compared to 797 MB for GEM and 823 MB for EWC), and its maximum GPU memory allocation is approximately 25 MB. Notably, CMIP achieves these results using only 250 replay samples (~ 0.75 MB) - half the buffer size employed by GEM - without incurring additional training cost. These findings demonstrate that the integration of mutual information and entropy regularization in CMIP introduces negligible computational overhead, establishing it as a lightweight yet highly effective solution for continual learning under the Class-IL setting.

To assess the influence of the regularization coefficients in CMIP, we conducted a grid search over a range of values for λ_1 and λ_2. Each configuration was evaluated on the Split-MNIST Class-IL benchmark. The results are summarized in table 3, where each cell reports the final global accuracy (%) and average forgetting (%) for a specific combination of λ_1 and λ_2. This experimental setup follows the information-theoretic framework proposed in [2, 3, 5, 10], which emphasizes the importance of balancing information retention and compression in continual learning.

Table 2. *Computational cost comparison for EWC, GEM, and CMIP on MNIST-Split (Class-IL): training time, RAM, GPU usage, and replay memory size.*

Method	Total Training Time (min)	Peak RAM Usage (MB)	GPU Memory Usage (MB)	Replay Memory (samples / MB)
Fine-tune	9,36	774	28,15	N/A
EWC	1,89	823,2	56,95	N/A
GEM	1,7	797,1	22,75	500/1,5
CMIP	2.28	841,99	25,35	250/0.75

Table 3. *Final accuracy (%) / average forgetting (%) for different combinations of λ_1 (columns) and λ_2 (rows) on Split-MNIST Class-IL.*

λ_2 \ λ_1	0.005	0.010	0.050	0.100	0.200
0.000	87.05 / 11.58	85.07 / 12.83	88.08 / 10.85	87.63 / 11.12	88.30 / 10.76
0.001	85.02 / 13.01	90.88 / 8.94	90.68 / 9.19	88.94 / 10.25	85.44 / 12.55
0.005	87.80 / 10.79	87.60 / 11.10	89.85 / 9.59	88.39 / 10.62	87.48 / 11.05
0.010	87.79 / 11.02	76.90 / 18.04	88.18 / 10.76	**90.97 / 8.81**	87.33 / 11.31
0.100	86.62 / 11.68	89.93 / 9.45	86.95 / 11.42	89.41 / 9.99	89.33 / 10.02

The results indicate that CMIP is generally robust across a wide range of regularization strengths. Among all configurations, the setting λ_1=0.1 and λ_2=0.01 yields the best trade-off, achieving 90.97% final accuracy with only 8.81% forgetting. This choice aligns well with prior findings in [2, 3, 5, 10], which advocate for moderate mutual information preservation and entropy regularization to enhance stability and mitigate forgetting in continual learning.

5. DISCUSSION

CMIP reduces catastrophic forgetting to 8.81% on MNIST-Split, substantially lower than GEM (~33%) and EWC (~77%), while achieving 90.97% accuracy, surpassing EWC (20.64%) and GEM (65.1%). However, parameters λ_1 and λ_2 exhibit sensitivity, requiring careful tuning through experimentation. CMIP extends IB [[4]] and MINE [[5]] by integrating entropy regularization. This demonstrates that combining MI loss with entropy effectively preserves useful information from previous tasks without noise from redundant information.

These results suggest potential applications in tasks such as robotic learning (learning new tasks without forgetting previous ones) and IoT systems (suitable for memory-constrained environments).

Future work includes optimizing the MINE estimator [[5]], developing automatic schemes for selecting λ_1 and λ_2, and extending CMIP to real-time data processing [[8]].

6. CONCLUSIONS

This paper proposes CMIP - an approach based on mutual information optimization and entropy regularization in continual learning. Experimental results demonstrate significantly superior performance compared to commonly used methods like EWC and GEM.

Future research may focus on:

- Optimizing replay memory capacity and sample-selection strategies;
- Automatic tuning of λ_1 and λ_2;
- Extending CMIP to multimodal data and memory-constrained embedded platforms.

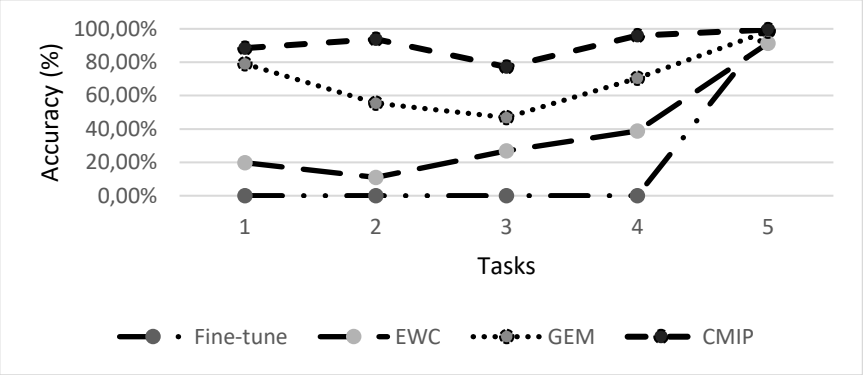

Figure 2. *Individual task accuracy on MNIST-Split Class-IL data through experimentation.*

REFERENCES

[1]. R. M. French, *"Catastrophic forgetting in connectionist networks,"* Trends in Cognitive Sciences, Vol. 3, No. 4, pp. 128–135, (1999). https://doi.org/10.1016/S1364-6613(99)01294-2

[2]. J. Kirkpatrick et al., *"Overcoming catastrophic forgetting in neural networks,"* Proceedings of the National Academy of Sciences, Vol. 114, No. 13, pp. 3521–3526, (2017).

[3]. D. Lopez-Paz and M. Ranzato, *"Gradient episodic memory for continual learning,"* Advances in Neural Information Processing Systems, Vol. 30, pp. 6467–6476, (2017).

[4]. N. Tishby, F. C. Pereira, and W. Bialek, *"The information bottleneck method,"* arXiv preprint, physics/0004057, (2000). https://arxiv.org/abs/physics/0004057

[5]. M. I. Belghazi et al., *"Mutual Information Neural Estimation,"* International Conference on Machine Learning (ICML), (2020). https://arxiv.org/abs/1801.04062

[6]. T. Chen et al., *"A simple framework for contrastive learning of visual representations,"* Proceedings of the 37th International Conference on Machine Learning, Vol. 119, pp. 1597–1607, (2020).

[7]. Y. Polyanskiy and Y. Wu, *"Information theory and deep learning: A modern perspective,"* Annual Review of Statistics and Its Application, Vol. 11, pp. 101–125, (2024).

[8]. T. Hospedales et al., *"Meta-learning in neural networks: A survey,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, No. 9, pp. 5149–5169, (2022).

[9]. Z. Mai et al., *"Online Continual Learning in Image Classification: An Empirical Survey,"* Neurocomputing, Vol. 512, pp. 177–196, (2022). https://doi.org/10.1016/j.neucom.2021.8.811

[10]. G. M. van de Ven, T. Tuytelaars, and A. S. Tolias, *"Three types of incremental learning,"* Nature Machine Intelligence, (2022). https://doi.org/10.1038/s42256-022-00568-3

TÓM TẮT

Tối ưu hóa thông tin lẫn nhau để giảm thiểu tình trạng quên thảm khốc trong học liên tục: Cách tiếp cận lý thuyết thông tin

Học liên tục đối mặt với thách thức quên thảm khốc, khi mô hình mất "kiến thức" từ các tác vụ trước khi học tác vụ mới. Trong bài báo này, chúng tôi đề xuất CMIP, một phương pháp dựa trên lý thuyết thông tin, sử dụng tối ưu hóa thông tin lẫn nhau và chính quy hóa entropy nhằm duy trì kiến thức cũ, tạo biểu diễn tiềm ẩn nén tối ưu. CMIP tích hợp một mạng neuron phụ để ước lượng MI, áp dụng chiến lược replay memory, trong đó mỗi mini-batch huấn luyện được xây dựng với tỷ lệ 50% mẫu của tác vụ hiện tại và 50% mẫu được lấy từ bộ nhớ của tác vụ trước. Thực nghiệm được thực hiện trên tập dữ liệu MNIST-Split và CIFAR-100-Split cho bài toán Class-incremental learning (Class-IL). Trên MNIST-Split, CMIP đạt được độ chính xác trung bình 90.97% và mức quên trung bình chỉ 8.81%, vượt trội so với các phương pháp hiện đại như Elastic Weight Consolidation (EWC) và Gradient Episodic Memory (GEM). Phương pháp có thể áp dụng cho học liên tục trong các ứng dụng thực tế như robot học và xử lý dữ liệu thời gian thực.

Từ khóa: Học liên tục; Lý thuyết thông tin; Thông tin lẫn nhau; Entropy; Quên thảm khốc.