Information Bottleneck Based Data Correction in Continual Learning

Shuai Chen[1,2], Mingyi Zhang[1], Junge Zhang[1⋆], and Kaiqi Huang[1,2,3⋆]

[1] CRISE, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology

Abstract. Continual Learning (CL) requires model to retain previously learned knowledge while learning new tasks. Recently, experience replay-based methods have made significant progress in addressing this challenge. These methods primarily select data from old tasks and store them in a buffer. When learning new task, they train the model using both the current and buffered data. However, the limited number of old data can lead to the model being influenced by new tasks. The repeated replaying of buffer data and the gradual discarding of old task data (unsampled data) also result in a biased estimation of the model towards the old tasks, causing overfitting issues. All these factors can affect the CL performance. Therefore, we propose a data correction algorithm based on the Information Bottleneck (IBCL) to enhance the performance of the replay-based CL system. This algorithm comprises two components: the *Information Bottleneck Task Agnostic Constraints* (IBTA), which encourages the buffer data to learn task-relevant features related to the old tasks, thereby reducing the impact of new tasks. The *Information Bottleneck Unsampled Data Surrogate* (IBDS), which models the information of the unsampled data in the old tasks to alleviate data bias. Our method can be flexibly combined with most existing experience replay methods. We have verified the effectiveness of our method through a series of experiments, demonstrating its potential for improving the performance of CL algorithms.

Keywords: Continual learning · Incremental learning · Experience replay

1 Introduction

Modern deep learning algorithms have achieved remarkable performance in a diverse range of tasks [48–50]. However, one of their significant limitations is the inability to retain knowledge from previously learned tasks when receiving training on new tasks, known as catastrophic forgetting [18,25]. Addressing this issue has been a subject of extensive research in the field of continual learning (CL) [12, 36, 52, 53]. Among the proposed methods, the experience replay approachs have received significant attention [11,13,31,54]. These methods aim to

⋆ Corresponding author.

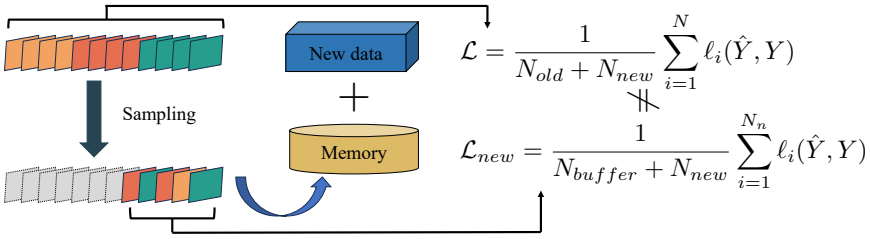

Fig. 1: During the process of CL, when the old task is sampled, there is a consequent discarding of some unsampled data. This leads to bias in the model's estimation of the distribution of the old task.

overcome catastrophic forgetting by selectively storing a subset of data from old tasks in the buffer and then combining buffer data with new task data for joint training. Various replay methods have been investigated from different perspectives to address the challenge of sample selection and balancing the contributions of new and old samples during the replay process [6, 14, 37].

Despite the progress made in reducing catastrophic forgetting through replay methods, certain limitations still exist. A major issue is that the memory of the buffer is too small, and the model cannot store enough data to restore previously learned knowledge [4, 21, 35]. This will give rise to two significant problems. Firstly, the scarcity of old data will result in an imbalance between the number of new and old tasks, leading to a disruption of knowledge from the old task during the learning process of the new task. Secondly, in the CL process, each time the old task data is sampled, some unsampled data will be discarded, introducing bias in the model's estimation of the old task distribution. Moreover, the repeated replaying of buffer data can contribute to overfitting of this specific data. All of these factors can significantly impact the CL performance. Currently, there are some methods to address the above issues, such as selecting more representative samples [20], increasing the diversity of old samples through data expansion [3], and improving the generalization ability of the model [5].

This work aims to tackle the aforementioned challenges from a different perspective. Regarding the first issue, we propose encouraging the data in the buffer to acquire task agnostic features related to the old task, thereby mitigating the impact of the new task. For the second issue, we focus on an important yet under-investigated question: *How can unsampled data effectively contribute to learning new tasks?* In Figure 1, we illustrate the cause of model overfitting, which arises because replay-based methods often utilize only a subset of data from old task, leading to biased results. When training on new tasks, we design the objective function as follows:

$$\mathcal{L}_{new} = \frac{1}{N_{buffer} + N_{new}} \sum_{i=1}^{N_n} \ell_i(\hat{Y}, Y). \tag{1}$$

where $\ell_i(\hat{Y}, Y)$ represents the loss function that takes the true outcome Y and predicts outcome $\hat{Y}$ as inputs. N_{buffer} and N_{new} denote the numbers of buffer and new task data, respectively. N_n is the sum of these two quantities. However, this loss function does not provide an unbiased estimate of the loss function for both new and old tasks, as the model gradually loses some data from past tasks during training. The unbiased loss function for new and old tasks should be formulated as follows:

$$\mathbb{E}[\mathcal{L}_{new}] \neq \mathcal{L} := \frac{1}{N_{old} + N_{new}} \sum_{i=1}^{N} \ell_i(\hat{Y}, Y). \tag{2}$$

where N_{old} and N_{new} represent the numbers of old task and new task data, respectively, and N signifies the sum of these two quantities. To address the bias arising from the discarding of unsampled data from old task, our method aims to estimate the information of the unsampled data using specific strategies. However, unsampled data cannot directly participate in the training of new tasks, we indirectly obtain a surrogate for the information about unsampled data through the relationship between sampled and unsampled data. This information related to the unsampled data serves as a regularization term to constrain the training process of the model, thereby achieving bias correction for the dataset.

Based on the intuition presented above, we propose an Information Bottleneck based data correction algorithm (IBCL). The information bottleneck method is a technique in information theory used to find the best compromise between accuracy and complexity [41]. It is regarded as the theoretical foundation of deep learning and has been applied in many fields, such as robust or invariant representation learning, causal reasoning, and so on [17]. This theory suggests that the model gradually eliminates irrelevant noise information during the learning process, retaining the most relevant feature information to the task. Borrowing the ability of this method, our algorithm mainly consists of two parts: Information Bottleneck Task Agnostic Constraints (IBTA): During the update process of new tasks, we introduce a constraint on the learning process of the buffer data. This constraint encourages the model to learn task agnostic features that are relevant to the old tasks, effectively preventing interference from the new tasks. Information Bottleneck Unsampled Data Surrogate (IBDS): This module is designed to model the information of the unsampled data. As the unsampled data cannot be directly trained, our method decouples the features of the unsampled data from the sampled data after the completion of learning old tasks. This decoupling allows model to leverage the relationship between the two types of data to estimate the surrogate influence of unsampled data. Consequently, this achieves an indirect implementation of the impact of unsampled data on learning new tasks.

Overall, our work makes the following contributions:

- We propose a data correction framework based on information bottleneck to mitigate model bias when the number of old task samples is small.

- To mitigate the problem of the influence of new tasks on old sample features during the learning process, we propose IBTA to encourage the model to learn task agnostic features.
- To counter the issue of overfitting caused by the repeated replay of buffer data and gradual discarding of unsampled data, we propose IBDS to incorporate the information from unsampled data to assist in the training of the new model.
- Our method can be combined with most replay based methods to improve their performance.

2 Related Work

2.1 Experience Replay Continual Learning

In the context of CL, there are three main approaches to retaining knowledge of past tasks: using a fixed-sized buffer for replaying past samples (replay-based approach) [2, 3, 7, 8, 43], regulating model parameter changes through learning (regularization-based approach) [16, 40, 42], or dynamically expanding model architecture as needed (expansion-based approach) [15, 28, 47]. Despite its simplicity, the replay-based approach has shown promising results in CL settings. Experience Replay (ER) [19] is a widely-used technique, which maintains a fixed-sized buffer to store and replay selected data, helping to mitigate the forgetting of previously learned knowledge. Incremental Classifier and Representation Learning (iCaRL) [37] uses the herding strategy to select a few representative samples. Rainbow Memory [3] proposes a diversity sampling strategy based on per-sample classification uncertainty. However, these methods primarily concentrate on how to choose which old data to store in the buffer, while neglecting the unsampled data. We argue that the unsampled data can also be beneficial in mitigating deviation and should be utilized. To this end, we propose a data correction method that employs the information bottleneck theory to model the information of the unsampled data in the old task, thereby enhancing the performance of several state-of-the-art methods.

2.2 Information Bottleneck

Information bottleneck (IB) is a classical concept in information theory, which is used to explain the behavior of deep learning [41]. Furthermore, IB has been extensively utilized in diverse applications, including but not limited to, representation learning decoupling [23, 33], recommendation systems [29, 45], and enhancing model robustness [17]. Recently, there have been some other works that apply information theory to CL, such as [27, 46], but there are fewer works based on the IB theory. The fundamental construct of IB is based on mutual information. In a previous work [20], the maximization of mutual information is employed to learn adequate feature expression. In contrast, our proposed method concentrates on modeling the unsampled data in the previous task.

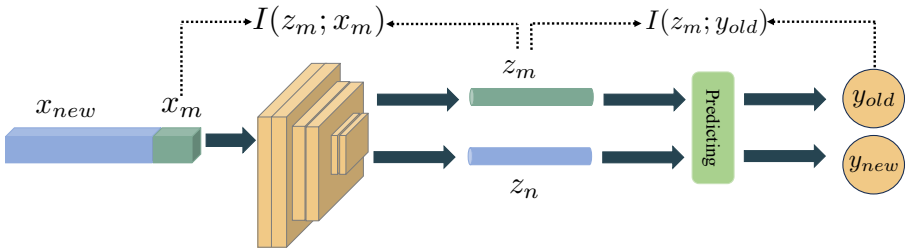

Fig. 2: During the updating of new tasks, we constrain the learning process of buffer data, encouraging the model to learn task agnostic features related to the old tasks, thus avoiding the interference of new tasks.

3 Method

Our method consists of two key processes: the learning process of new tasks and the processing process after learning old tasks. We introduce IBTA and IBDS separately.

3.1 Problem Formulation

In CL scenarios, tasks are introduced one after another and handled separately. We consider a total of T tasks. Each task $D_t = \{x_i^t, y_i^t\}_{i=1}^{n_t}$ comprises n_t input-target pairs (x_i^t, y_i^t), which are assumed to be independent and identically distributed (i.i.d.). The aim for each task is to train a model $f_\theta : X \to Y$ with parameters θ, capable of predicting the class label y for a new image x. In CL algorithms that utilize experience replay, the methods typically store some data from previous task data x_{old} in a buffer. This stored data is referred to as sampled data x_m, with its features denoted as z_m. The data not selected for storage is called unsampled data x_u, and its features are represented as z_u.

3.2 Information Bottleneck Task Agnostic Constraints

During the learning of new tasks, the model focus on the new task due to the limited data available in the buffer. In order to mitigate the impact of new tasks, our method encourages the buffer data to focus on learning the task-specific features, thus minimizing their influence. In order to compress the information in the dataset into the features of the sampled data as much as possible, the mutual information between the features z_m of the sampled data and the buffer dataset x_m should be minimized; In order to make the feature be able to distinguish well, the feature z_m of the sample needs to predict the label y_{old} as accurately as possible. The weighting parameters α is used to balance the contributions of different components in the overall objective function.

$$\mathcal{L}_{IBTA} := \min \underbrace{\alpha I(z_m; x_{\mathrm{m}})}_{(1)} - \underbrace{I(z_m; y_{\mathrm{old}})}_{(2)}. \tag{3}$$

In the given framework, the expression (1) represents a compression component that characterizes the mutual information between the buffer dataset x_m and the sampled data features z_m. This component aims to capture the relevance of z_m to the old dataset x_m. The expression (2) corresponds to an accuracy component that measures the effectiveness of z_m in representing the underlying distribution of the dataset. It aims to ensure that the sampled features z_m preserve sufficient information about the original data. Next, we will proceed to optimize the compression term $I(z_m; x_m)$. This term can be quantified using the Kullback-Leibler (KL) divergence as follows:

$$\begin{aligned} I(z_m; x_{\mathrm{m}}) &= \mathbb{E}_x \left[\mathrm{KL}(p(z_m|x_{\mathrm{m}}) \parallel p(z_m))\right] \\ &= \mathbb{E}_x \left[\int p(z_m|x_{\mathrm{m}}) \log p(z_m|x_{\mathrm{m}})\, dz - \int p(z_m|x_{\mathrm{m}}) \log p(z_m)\, dz\right]. \end{aligned} \tag{4}$$

Nonetheless, due to the computational complexity involved in calculating $p(z_m) = \int p(z_m|x_m)p(x_m)dx$, we resort to the utilization of a variational approximation approach. Instead of using $p(z_m)$, we adopt a variational distribution denoted as $q(z_m)$. Leveraging Gibbs' inequality, we recognize the non-negativity of the Kullback-Leibler (KL) divergence. Thus, we can establish an upper bound for Equation (4),

$$\begin{aligned} &- \int p(z_m|x_{\mathrm{m}}) \log p(z_m) dz \leq - \int p(z_m|x_{\mathrm{m}}) \log q(z_m) dz \\ &\Rightarrow \mathrm{KL}(p(z_m|x_{\mathrm{m}}) \parallel p(z_m)) \leq \mathrm{KL}(p(z_m|x_{\mathrm{m}}) \parallel q(z_m)). \end{aligned} \tag{5}$$

Similar to prior research [1, 29, 34, 44, 45], we can make the assumption that the posterior distribution $p(z_m \mid x_m)$ follows a Gaussian distribution and set it $\mathcal{N}\left(e(x_m), \mathrm{diag}(\sigma(x_m))\right)$. In this context, $e(x_m)$ represents the encoded embedding of the variables x_m, and $\mathrm{diag}(\sigma(x_m))$ indicates the variance in a diagonal matrix form. By employing the reparameterization trick, the embedding z_m can be generated through the formula $z_m = e(x_m) + \epsilon \odot \sigma(x_m)$, where $\epsilon \sim \mathcal{N}(0, I)$. Notably, if we set $\sigma(x_m)$ as an all-zero matrix, z_m will effectively become a deterministic embedding. Conversely, the prior distribution $q(z_m)$ is assumed to adhere to a standard Gaussian variational distribution, namely $q(z_m) = \mathcal{N}(0, I)$. Ultimately, the previously mentioned upper bound can be rephrased as follows:

$$\mathrm{KL}(p(z_m \mid x_{\mathrm{m}}) \| q(z_m)) = \|e(x_{\mathrm{m}})\|_2^2 + \sum_d \left(\sigma_d - \frac{1}{2}\log \sigma_d - 1\right). \tag{6}$$

where σ_d corresponds to an element within diag $\{\sigma(x_m)\}$. This signifies that for a deterministic embedding z_m, we can enhance the optimization of this upper bound by directly implementing ℓ_2-norm regularization on the embedding vector z_m. This process is equivalent to refining the compression term $I(z_m; x_m)$. When considering the mutual information $I(z_m; y_{old})$ within Equation (3), we can express it as $I(z_m; y_{old}) = H(y_{old}) - H(y_{old} \mid z_m)$. As $H(y_{old})$ is a positive constant that can be disregarded, the following inequality holds:

$$I(z_m; y_{old}) \geq -H(y_{old} \mid z_m). \tag{7}$$

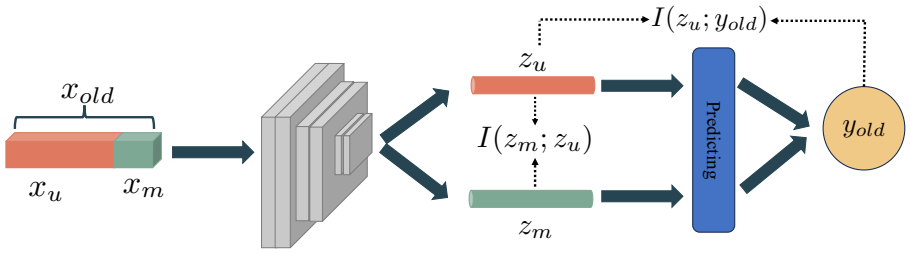

Fig. 3: This module is used to model the information of unsampled data. Because unsampled data cannot be trained directly, our method decouples the features of unsampled data and sampled data after the completion of the old task learning, so that it is convenient to estimate the surrogate loss of unsampled data in the new task by using the relationship between them, and indirectly realize the influence of unsampled data on the new task learning.

This term can be seen as a cross entropy loss function. Therefore, we have implemented the constraint process for learning buffer data. However, the loss function obtained in this way is still biased, as there is missing unsampled data, which means that the contribution to the old task label y_{old} should be $I(z; y) = I(z_m, z_u; y)$, where z_u is unsampled data feature. During the learning process of the new task, this item cannot be directly calculated. We will calculate it indirectly in the following section.

3.3 Information Bottleneck Unsampled Data Surrogate

To estimate the mutual information $I(z; y) = I(z_m, z_u; y)$ during the learning process of new task, it becomes essential to obtain z_u. Given that x_u denotes unsampled data, we cannot directly calculate z_u from x_u when learning new task. This necessitates an indirect estimation of z_u from old task model. As shown in Fig.3, during the learning stage of old task, both sampled x_m and unsampled x_u data are involved in the training process. After training completion, the features corresponding to these two parts z_m and z_u are tightly coupled. Consequently, a specific operation is needed after training ends to decouple these two features. If we can decouple the features of sampled and unsampled data by optimizing $I(z_m; z_u)$, we can get $I(z; y) = I(z_m; y) + I(z_u; y)$. In this case, the estimate of $I(z_u; y)$ can be used as a surrogate for the unsampled data information during learning new task. The feature z_m of the sampled data and the feature z_u of the unsampled data must maintain as much independence from one another as possible to obtain better discrimination. Drawing on principles from information theory, we are inspired to derive a minimization objective function based on the aforementioned analysis:

$$\mathcal{L}_{IBDS} := \min I(z_m; z_u). \tag{8}$$

represents a decoupling component, capturing the level of interdependence between z_u and z_m. It aims to minimize the dependence of z_u on z_m, ensuring that z_u provides additional information beyond what is already captured by z_m.

Inspired by previous research work [1, 29, 34, 44, 45], utilizing the chain rule of mutual information, we derive the subsequent equation pertaining to the decoupling component:

$$I(z_m; z_u) = I(z_m; y_{old}) - I(z_m; y_{old} \mid z_u) + I(z_m; z_u \mid y_{old}). \tag{9}$$

We proceed to analyze the expression $I(z_m; z_u \mid y_{old})$ in greater detail. Given that the distribution of z_m exclusively hinges on the variables x_m, and the latter is influenced by y_{old}, we establish $H(z_m \mid y_{old}, z_u) = H(z_m \mid y_{old})$, where $H(\cdot \mid \cdot)$ signifies conditional entropy. By combining the attributes of mutual information, we arrive at the following relationship:

$$\begin{aligned} I(z_m; z_u \mid y_{\text{old}}) &= H(z_m \mid y_{\text{old}}) - H(z_m \mid y_{\text{old}}, z_u) \\ &= H(z_m \mid y_{\text{old}}) - H(z_m \mid y_{\text{old}}) = 0. \end{aligned} \tag{10}$$

By substituting Eq.(9), we have,

$$I(z_m; z_u) = I(z_m; y_{old}) - I(z_m; y_{old} \mid z_u) \tag{11}$$

Since the term $I(z_m; y_{old} \mid z_u)$ in Eq.(11) is still difficult to be calculated, we simplify it further by using the form of conditional entropy:

$$\begin{aligned} I(z_m; y_{\text{old}} \mid z_u) &= H(z_m \mid z_u) + H(y_{\text{old}} \mid z_u) - I(z_m; y_{\text{old}} \mid z_u) \\ &= H(z_m \mid z_u) - H(z_m \mid y_{\text{old}}, z_u) \\ &= H(y_{\text{old}} \mid z_u) - H(y_{\text{old}} \mid z_m, z_u) \end{aligned} \tag{12}$$

$$I(z_m; z_u) = I(z_m; y_{old}) - H(y_{old} \mid z_u) + H(y_{old} \mid z_m, z_u). \tag{13}$$

Finally, combining Eq.(7) and Eq.(13), we can rewrite the objective function $\mathcal{L}_{IBDS}$ in Eq.(8) as follows,

$$\mathcal{L}_{IBDS} = -H(y_{old} \mid z_m) - H(y_{old} \mid z_u) + H(y_{old} \mid z_m, z_u). \tag{14}$$

By optimizing $\mathcal{L}_{IBDS}$, we get the decoupled expression z_u. Now we can compute $I(z_u; y)$, which acts as a surrogate loss to compensate for missing weights due to unsampled data.

During the training process of neural network, the model will encounter the problem of semantic drift [22,24,51], which means that the feature distribution of sampled data will change with the learning of new tasks. This concept highlights a significant challenge: using the fixed features of unsampled data as constraints in the learning of new task becomes problematic. As the model trains on new tasks, the estimation of unsampled data features gradually deviates from their true values, which not only affects the effectiveness of bias correction but also, to some extent, impacts the overall learning process. Just as the "no free lunch"

theorem advocated in machine learning. Once samples are lost, it becomes difficult to perfectly estimate the distribution of unsampled data, the best one can do is to minimize the model's bias when addressing old tasks. To mitigate this impact, it's essential to consider the characteristics of neural networks in CL [9]. Early in the learning of new tasks, gradients undergo significant changes, and the model frequently encounters considerable biases. A viable strategy involves imposing stricter constraints on the new task learning process during the initial stages of model training. This approach instructs the model in reducing biases when learning new tasks. As the training advances, these constraints are gradually relaxed, permitting the model to better adjust to the learning requirements of new tasks during the slower update phases. To address the processing of unsampled data features, a dynamic decay strategy is employed to balance the learning conflicts between old and new tasks, thereby optimizing the model's overall performance. This method allows the model to discover a more balanced learning point between preserving old knowledge and acquiring new tasks, enhancing its learning efficiency and effectiveness. Specifically, inspired by [30], we propose a dynamic cosine annealing factor η for constraint terms when updating unsampled data.

$$\eta = \left(\frac{1}{2}(1 + cos(\frac{e_{cur}}{e_{max}})\pi)\right) \tag{15}$$

where e_{cur} and e_{max} represent current and maximum number of epochs.

3.4 Algorithm

By combining above two parts, we can get the total loss function as:

$$\begin{aligned} \mathcal{L}_{CIB} &= \alpha I(z_m; x_{\mathrm{m}}) - I(z; y_{\mathrm{old}}) \\ &= \alpha I(z_m; x_{\mathrm{m}}) - \beta I(z_u; y) - \gamma I(z_m; y) \\ &\leq \alpha \|e(x_{\mathrm{m}})\|_2^2 + \beta \cdot \eta H(y_{\mathrm{old}} \mid z_u) + \gamma H(y_{\mathrm{old}} \mid z_m). \end{aligned} \tag{16}$$

The weighting parameters α, β, γ are used to balance the contributions of different components in the overall objective function.

The final objective function comprises three terms: term (1) represents a regularization constraint on the input data, term (2) represents surrogate loss to reflects the influence of the unsampled data; term (3) denotes the cross-entropy between z_m and y_{old}, this can be seen as the cross-entropy loss function. The aim is to make the learned feature expression fully predict the final label.

Discussion This approach comprises two primary modules: the IBTA module facilitates the learning of new tasks, whereas the IBDS module is utilized post-learning of old tasks to derive features from unsampled data. These features are then integrated as constraints within the training of new tasks, thereby shaping the overall optimization loss function. Our method can be combined with most replay-based strategies. Taking the naive ER method as an example, it randomly selects some samples from the old task and stores them in a buffer, which are trained together with the samples from the new task. Upon finalizing

Table 1: Comparison of the average accuracy (the higher the better) of four split datasets with different buffer sizes.

Method	Split CIFAR-10			Split CIFAR-100			Split ImageNet-100			Split miniImageNet		
Joint	92.38			73.29			80.23			53.55		
FT	19.67			9.29			8.68			9.52		
Buffer size	100	200	500	200	500	2000	200	500	2000	1000	2000	5000
ER	36.39	44.79	57.74	14.35	19.66	36.76	13.63	18.37	34.25	8.37	16.49	24.17
+IBCL	45.43	51.91	63.12	23.98	28.32	42.01	22.72	28.93	43.11	14.79	22.72	27.69
ER-ACE	53.90	63.41	70.53	26.28	36.48	48.41	24.23	37.12	49.55	17.95	22.60	27.92
+IBCL	63.85	70.97	74.82	32.12	40.94	51.89	30.72	41.81	52.62	24.77	27.63	31.02
DER++	57.65	64.88	72.70	25.11	37.13	52.08	26.50	43.65	58.05	18.02	23.44	30.43
+IBCL	66.41	72.52	76.61	33.82	44.21	54.89	32.63	46.57	59.92	27.72	31.41	35.79
X-DER	59.29	65.19	68.10	35.34	44.62	54.44	33.21	46.72	55.23	25.24	26.38	29.91
+IBCL	67.91	73.82	74.14	43.86	48.72	55.91	43.25	49.62	56.59	29.90	30.21	32.66

the learning of old task, the sampled data is processed through the old model to yield z_m, and both sampled and unsampled data are processed to extract coupled features z. Subsequently, by optimizing IBDS, features z_u for each class of unsampled data are obtained. During the training of the new task, IBTA and z_u are combined to optimize the total loss function for learning the new task. Detailed pseudocode of this procedure is provided in the appendix.

4 Experiments

In this section, we will present the datasets and evaluation metrics used in the experiment, and describe the implementation details. Then we describe the experimental results of our method on multiple standard benchmarks. Finally, we conduct an ablation study to illustrate the effectiveness of each module of our method.

4.1 Dataset

We validate our results on four distinct datasets: CIFAR-10, CIFAR-100, miniImageNet, and ImageNet-100. CIFAR-10 and CIFAR-100 each consist of 50,000 training samples and 10,000 testing samples, with 10 and 100 classes respectively. miniImageNet and ImageNet-100 are subsets of ImageNet, both containing 100 classes. miniImageNet includes 500 images per category for training, while ImageNet-100 includes 1300 images per category for training.

4.2 Details

To better compare with other methods, we follow the settings of [46] and [5]. For the CIFAR-10/100 and ImageNet-100 datasets, we adopt ResNet-18, and for the miniImageNet dataset, we adopt EfficientNet-B2. For CIFAR-10/100, we

Table 2: Comparison of our method with other regularization methods on three datasets, we use two typical methods DER++ and ER-ACE as benchmarks for comparison.

Method	Split CIFAR-10			Split CIFAR-100			Split miniImageNet		
Buffer size	100	200	500	200	500	2000	1000	2000	5000
ER-ACE	53.90	63.41	70.53	26.28	36.48	48.41	17.95	22.60	27.92
+ sSGD	56.26	64.73	71.45	28.07	39.59	49.70	18.11	22.43	24.12
+ oEwC	52.36	61.09	68.70	24.93	35.06	45.59	19.04	24.32	29.46
+ oLAP	52.76	63.19	70.32	26.42	36.58	47.66	18.34	23.19	28.77
+ OCM	57.18	64.65	70.86	28.18	37.74	49.03	20.32	24.32	28.57
+ LiDER	56.08	65.32	71.75	27.94	38.43	50.32	19.69	24.13	30.00
+ DualHSIC	60.52	68.08	73.78	29.08	38.94	50.55	22.33	25.41	30.12
+ Our	63.85	70.97	74.82	32.12	40.94	51.89	24.77	27.63	31.02
DER++	57.65	64.88	72.70	25.11	37.13	52.08	18.02	23.44	30.43
+ sSGD	55.81	64.44	72.05	24.76	38.48	50.74	16.31	19.29	24.24
+ oEwC	55.78	63.02	71.64	24.51	35.22	51.53	18.87	24.53	31.91
+ oLAP	54.86	62.54	71.38	23.26	34.48	50.80	18.91	25.02	32.78
+ OCM	59.25	65.81	73.53	27.46	38.94	52.25	20.93	24.75	31.16
+ LiDER	58.43	66.02	73.39	27.32	39.25	53.27	21.58	28.33	35.04
+ DualHSIC	64.98	70.28	75.94	31.46	41.86	53.53	24.78	29.37	34.98
+ Our	66.41	72.52	76.61	33.82	44.21	54.89	27.72	31.41	35.79

adopt a training epoch of 50 and set the batch size to 32 and 64, respectively. For MiniImageNet and ImageNet-100, we adopt training epochs of 80 and 120, with batch size set to 128 and 64, respectively. In order to better compare the experimental results, we adopt the experimental results reported in [46] and [5], or the hyperparametric design in the original article.

4.3 Evaluation Metrics

We mainly evaluate on class incremental learning scenario. In this setup, each sequential task introduces a distinct set of classes that do not overlap with those from previous tasks. The challenge lies in training a classifier that can effectively classify all classes encountered up to the current stage, without the need for explicit task ids. We evaluate the final performance of the method by average accuracy.

$$ACC = \frac{1}{T}\sum_{i=1}^{T} R_{T,i}, \tag{17}$$

where $R_{T,i}$ denotes the test accuracy on task i after the model has finished task T.

Table 3: Abalation study of the average accuracy (the higher the better) of split ImageNet-100 and split CIFAR-100 with different buffer sizes.

Method	Split CIFAR-100			Split ImageNet-100		
Buffer size	200	500	2000	200	500	2000
ER-ACE	26.28	36.48	48.41	24.23	37.12	49.55
+ IBTA	27.36	37.49	49.12	25.63	37.92	50.12
+ IBDS	30.44	39.52	50.91	29.83	40.71	51.85
+IBCL	32.12	40.94	51.89	30.72	41.81	52.62
DER++	25.11	37.13	52.08	26.50	43.65	58.05
+IBTA	27.88	38.74	53.96	27.39	45.72	59.33
+ IBDS	31.53	42.59	54.33	29.58	46.15	59.70
+IBCL	33.82	44.21	54.89	32.63	46.57	59.92

4.4 Baseline Models

We conduct experiments on multiple state-of-the-art experience replay CL models: **ER** [10], **DER++** [8], **X-DER** [6], and **ER-ACE** [9]. In addition to the original loss function based on replay, our method can be considered as a new regularization method. Therefore, we also compare our method with existing regularization-based techniques,including **sSGD** [32], **oEWC** [39], **oLAP** [38], and more recent SOTA methods, **OCM** [20], **LiDER** [5] and **DualHSIC** [46].

4.5 Results

The comparison results are reported in Table 1. For the method of listing several baselines, our proposed method has improved the model's performance in various datasets and buffer sizes. In addition, we have the following main observations: 1) There has been a significant improvement in basic methods such as ER and DER++, as these two methods adopt more basic methods for sampling. Our method can reduce the bias caused by the sampling process to a certain extent. 2) Our methods have also shown some effectiveness on X-DER and ER-ACE, which take additional measures to alleviate sample imbalance and improve sample utilization, resulting in improved model performance. Our method further enhances the model's capabilities on top of this.

To further demonstrate the effectiveness of our method, we compare our method with other regularization methods, and it can be found that we achieve the best performance, recent methods such as OCM, LiDER and DualHSIC achieve significant progress compared to baseline when the number of buffer is small, and our method further improves the performance when the buffer data is small by considering modeling for the unsampled data.

4.6 Ablation Study

In order to understand the effect of each of the items in the loss function on the performance of the algorithm, an ablation study has been carried out and the

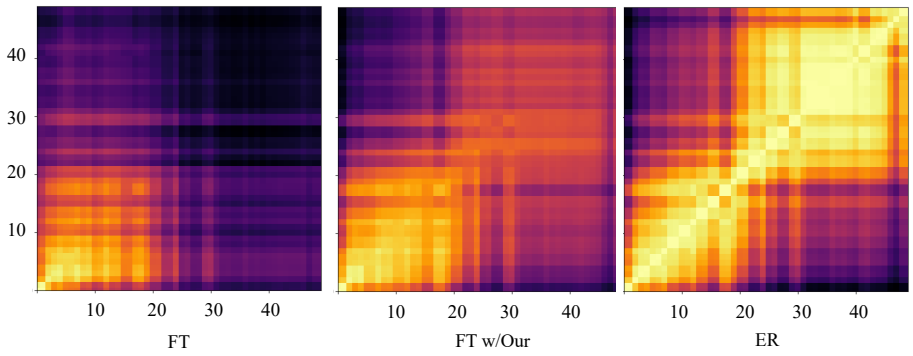

Fig. 4: This figure shows the CKA matrix based on the ER algorithm, FT and incorporating our algorithm. It can be observed from it that due to our modeling of unsampled data, the model shows more information about the old task.

results are shown in Table 3. It is observed that in most of the cases, increasing any of the components has an impact on the performance of the algorithm. However, increasing the IBDS gives a large performance improvement, which indicates that it plays a vital role in the performance of the algorithm. This component represents the loss of substitution of unsampled data, which is necessary for training on new tasks using unsampled data features. Therefore, if this aspect is not well learned, it may negatively affect the subsequent performance of the algorithm. IBTA also has an impact on the performance of the algorithm, the involves regularization constraints on the input features in order to make the model learn enough robust features.

5 Characteristic Analysis

We analyze various characteristics of the model to gain insight into the factors contributing to the performance enhancements achieved by our approach.

5.1 CKA Matrix Analysis

We use Central Kernel Alignment (CKA) [26] in Fig.4, which is an effective tool to measure the similarity of network representations and evaluate the relationships between these backbones. We send the same batch of instances to the ER (2000 buffer size) algorithm model and fine-tuning based model respectively to obtain corresponding feature maps, thus intuitively demonstrating the similarity of algorithm feature maps. Our algorithm can be used in conjunction with the experience replay method to model unsampled data. When the sampled data is 0, i.e. fine-tuning, our method can also get some benefits.

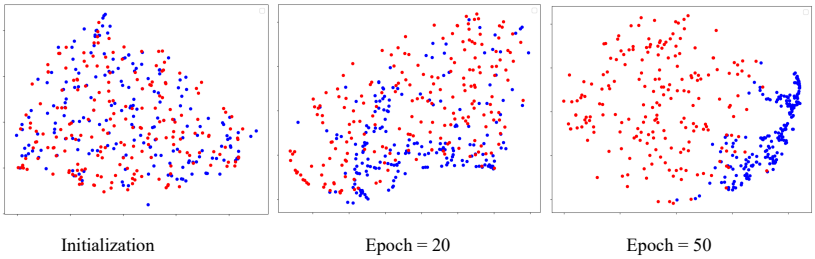

Fig. 5: This image uses TSNE to display the changes in the characteristics of sampled and unsampled data with the number of training epochs of the model. We have shown one class of situation, with a sample data size of 200 and an unsampled data size of 1100.

5.2 Feature Decoupling Analysis

A significant advance of our method is the decoupling of sampled and unsampled data features. We demonstrate this in Fig.5, where we use TSNE technology to visualize the changes in these two features with the increase of network training epochs. We test our method, the ER algorithm, on 200 sampled data and 1100 unsampled data from the ImageNet-100 dataset. As the network training epochs increase, we can observe that the two sets of features begin to separate from their initial coupling, which indicates that our method has achieved decoupling of the two data.

6 Conclusion

In conclusion, our proposed method is designed to tackle the issue of data bias in continual learning through the application of experience replay. Drawing inspiration from information bottleneck theory, we steer the incorporation of unsampled data information, thereby bolstering the model's capacity to preserve past knowledge while accommodating novel tasks. Our conducted experiments offer empirical validation of the advantages conferred by our approach. While we have presented a method for computing unsampled data, it's important to acknowledge the existence of several alternative strategies for estimating these. This opens up exciting possibilities for continued exploration and innovation in this field.

Acknowledgements

This work was supported by National Science and Technology Major Project, Grant No.2022ZD0116403, the Youth Innovation Promotion Association CAS and Strategic Priority Research Program of Chinese Academy of Sciences XDA27010300.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv (2016)
2. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. In: Adv. Neural Inform. Process. Syst. (2019)
3. Bang, J., Kim, H., Yoo, Y., Ha, J.W., Choi, J.: Rainbow memory: Continual learning with a memory of diverse samples. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
4. Biesialska, M., Biesialska, K., Costa-Jussa, M.R.: Continual lifelong learning in natural language processing: A survey. arXiv (2020)
5. Bonicelli, L., Boschini, M., Porrello, A., Spampinato, C., Calderara, S.: On the effectiveness of lipschitz-driven rehearsal in continual learning. In: Adv. Neural Inform. Process. Syst. (2022)
6. Boschini, M., Bonicelli, L., Buzzega, P., Porrello, A., Calderara, S.: Class-incremental continual learning into the extended der-verse. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
7. Boschini, M., Buzzega, P., Bonicelli, L., Porrello, A., Calderara, S.: Continual semi-supervised learning through contrastive interpolation consistency. Pattern Recognition Letters **162**, 9–14 (2022)
8. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. In: Adv. Neural Inform. Process. Syst. (2020)
9. Caccia, L., Aljundi, R., Asadi, N., Tuytelaars, T., Pineau, J., Belilovsky, E.: New insights on reducing abrupt representation change in online continual learning. In: Int. Conf. Learn. Represent. (2022)
10. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H.S., Ranzato, M.: Continual learning with tiny episodic memories. Arxiv (2019)
11. Chen, S., Zhang, M., Zhang, J., Huang, K.: Exemplar-based continual learning via contrastive learning. IEEE Transactions on Artificial Intelligence **5**(7), 3313–3324 (2024)
12. Cossu, A., Carta, A., Lomonaco, V., Bacciu, D.: Continual learning for recurrent neural networks: an empirical evaluation. Neural Networks (2021)
13. Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
14. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: PODNet: Pooled Outputs Distillation for Small-Tasks Incremental Learning. In: Eur. Conf. Comput. Vis. (2020)
15. Douillard, A., Ram, A., Couairon, G., Cord, M.: Dytox: Transformers for continual learning with dynamic token expansion. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
16. Ebrahimi, S., Elhoseiny, M., Darrell, T., Rohrbach, M.: Uncertainty-guided continual learning with bayesian neural networks. In: Int. Conf. Learn. Represent. (2020)
17. Federici, M., Dutta, A., Forré, P., Kushman, N., Akata, Z.: Learning robust representations via multi-view information bottleneck. In: Int. Conf. Learn. Represent. (2020)
18. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences (1999)

19. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv (2015)
20. Guo, Y., Liu, B., Zhao, D.: Online continual learning through mutual information maximization. In: Int. Conf. Mach. Learn. (2022)
21. Hadsell, R., Rao, D., Rusu, A.A., Pascanu, R.: Embracing change: Continual learning in deep neural networks. Trends in cognitive sciences (2020)
22. Iscen, A., Zhang, J., Lazebnik, S., Schmid, C.: Memory-efficient incremental learning through feature adaptation. In: Eur. Conf. Comput. Vis. (2020)
23. Jaiswal, A., Brekelmans, R., Moyer, D., Steeg, G.V., AbdAlmageed, W., Natarajan, P.: Discovery and separation of features for invariant representation learning. arXiv (2019)
24. Jie, S., Deng, Z.H., Li, Z.: Alleviating representational shift for continual fine-tuning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
25. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proc. of the national academy of sciences (2017)
26. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.E.: Similarity of neural network representations revisited. In: Int. Conf. Mach. Learn. (2019)
27. Li, D., Wang, T., Chen, J., Ren, Q., Kawaguchi, K., Zeng, Z.: Towards continual learning desiderata via hsic-bottleneck orthogonalization and equiangular embedding. In: AAAI (2024)
28. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: Int. Conf. Mach. Learn. (2019)
29. Liu, D., Cheng, P., Zhu, H., Dong, Z., He, X., Pan, W., Ming, Z.: Mitigating confounding bias in recommendation via information bottleneck. In: Proceedings of the 15th ACM Conference on Recommender Systems (2021)
30. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv (2016)
31. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., van de Weijer, J.: Class-incremental learning: Survey and performance evaluation on image classification. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
32. Mirzadeh, S.I., Farajtabar, M., Pascanu, R., Ghasemzadeh, H.: Understanding the role of training regimes in continual learning. Arxiv (2020)
33. Moyer, D., Gao, S., Brekelmans, R., Steeg, G.V., Galstyan, A.: Evading the adversary in invariant representation. CoRR (2018)
34. Pan, Z., Niu, L., Zhang, J., Zhang, L.: Disentangled information bottleneck. In: AAAI (2021)
35. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks (2019)
36. Qu, H., Rahmani, H., Xu, L., Williams, B., Liu, J.: Recent advances of continual learning in computer vision: An overview. arXiv (2021)
37. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
38. Ritter, H., Botev, A., Barber, D.: Online structured laplace approximations for overcoming catastrophic forgetting. In: Adv. Neural Inform. Process. Syst. (2018)
39. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)

40. Swaroop, S., Nguyen, C.V., Bui, T.D., Turner, R.E.: Improving and understanding variational continual learning. ArXiv (2019)
41. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method (2000)
42. Titsias, M.K., Schwarz, J., Matthews, A.G.d.G., Pascanu, R., Teh, Y.W.: Functional regularisation for continual learning with gaussian processes. In: Int. Conf. Learn. Represent. (2019)
43. Von Oswald, J., Zhao, D., Kobayashi, S., Schug, S., Caccia, M., Zucchet, N., Sacramento, J.: Learning where to learn: Gradient sparsity in meta and continual learning. In: Adv. Neural Inform. Process. Syst. (2021)
44. Wang, Y., Rudner, T.G., Wilson, A.G.: Visual explanations of image-text representations via multi-modal information bottleneck attribution. In: Adv. Neural Inform. Process. Syst. (2024)
45. Wang, Z., Chen, X., Wen, R., Huang, S.L., Kuruoglu, E.E., Zheng, Y.: Information theoretic counterfactual learning from missing-not-at-random feedback. In: Adv. Neural Inform. Process. Syst. (2020)
46. Wang, Z., Zhan, Z., Gong, Y., Shao, Y., Ioannidis, S., Wang, Y., Dy, J.: Dualhsic: Hsic-bottleneck and alignment for continual learning. In: Int. Conf. Mach. Learn. (2023)
47. Wortsman, M., Ramanujan, V., Liu, R., Kembhavi, A., Rastegari, M., Yosinski, J., Farhadi, A.: Supermasks in superposition. In: Adv. Neural Inform. Process. Syst. (2020)
48. Xu, P., Zhang, J., Huang, K.: Exploration via joint policy diversity for sparse-reward multi-agent tasks. In: IJCAI (2023)
49. Xu, P., Zhang, J., Huang, K.: Population-based diverse exploration for sparse-reward multi-agent tasks. In: IJCAI (2024)
50. Xu, P., Zhang, J., Yin, Q., Yu, C., Yang, Y., Huang, K.: Subspace-aware exploration for sparse-reward multi-agent tasks. In: AAAI (2023)
51. Yu, L., Twardowski, B., Liu, X., Herranz, L., Wang, K., Cheng, Y., Jui, S., Weijer, J.v.d.: Semantic drift compensation for class-incremental learning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
52. Zhang, X., Yi, J., Tao, J., Wang, C., Zhang, C.Y.: Do you remember? overcoming catastrophic forgetting for fake audio detection. In: Int. Conf. Mach. Learn. (2023)
53. Zhang, X., Yi, J., Wang, C., Zhang, C.Y., Zeng, S., Tao, J.: What to remember: Self-adaptive continual learning for audio deepfake detection. In: AAAI (2024)
54. Zhou, D.W., Wang, Q.W., Qi, Z.H., Ye, H.J., Zhan, D.C., Liu, Z.: Deep class-incremental learning: A survey. arXiv (2023)