An Energy Sampling Replay-Based Continual Learning Framework

Xingzhong Zhang[1][0009−0000−5564−7387], Joon Huang Chuah(✉)[2][0000−0001−9058−3497], Chu Kiong Loo[3][0000−0001−7867−2665], and Stefan Wermter[4][0000−0003−1343−4775]

[1] Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia
22063370@siswa.um.edu.my
[2] Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia
Faculty of Engineering and Information Technology, Southern University College, Skudai, Johor, Malaysia jhchuah@sc.edu.my
[3] Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia ckloo.um@um.edu.my
[4] Knowledge Technology, Department of Informatics, Universität Hamburg, Germany
stefan.wermter@informatik.uni-hamburg.de

Abstract. Continual Learning represents a significant challenge within the field of computer vision, primarily due to the issue of catastrophic forgetting that arises with sequential learning tasks. Among the array of strategies explored in current continual learning research, replay-based methods have shown notable effectiveness. In this paper, we introduce a novel Energy Sampling Replay-based (ESR) structure for image classification. This framework enhances the selection process of samples for replay by leveraging the energy distribution of the samples, thereby improving the effectiveness of memory samples during the replay phase and increasing accuracy. We have conducted extensive experiments across various continual learning methodologies and datasets. The results demonstrate that our approach effectively mitigates forgetting on CIFAR-10, CIFAR-100 and CIFAR-110 datasets by optimizing the replay strategy.

Keywords: Image classification · Continual learning · Catastrophic forgetting · Energy-based sampling.

1 Introduction

In the rapidly evolving field of machine learning, the concept of Continual Learning (CL) emerges as a crucial paradigm to address the challenge of learning new tasks sequentially without forgetting previously acquired knowledge. This paper focuses on developing an Energy Sampling Replay-based Continual Learning Framework for image classification aimed at enhancing the efficiency and effectiveness of CL models.

CL methodologies can be broadly categorized into three groups of approaches: regularization-based approaches, replay-based approaches, and optimization-based approaches. Regularization-based approaches [14, 11, 20] aim to mitigate

catastrophic forgetting by limiting the variation of learned knowledge, employing techniques like Elastic Weight Consolidation (EWC) [8] and Memory Aware Synapses (MAS) [2] to preserve knowledge of previous tasks while learning new knowledge. Replay-based approaches [5, 19, 24], such as experience replay and generative replay, counteract forgetting by emulating and restoring data distributions of previous tasks, ensuring the model's adaptability and memory retention. Optimization-based approaches [16, 4, 23], including gradient projection and meta-learning strategies, focus on modifying the optimization process to balance the preservation of old knowledge with the incorporation of new insights, thereby fostering a dynamic learning environment.

Energy-based models (EBMs) have become a method that has received increasing attention in recent years. Some studies have already applied it in domains such as domain adaptation and active learning. Among methods in [26, 25, 7], EBMs offer a promising alternative by leveraging the concept of energy functions to model the probability distribution of data. EBMs have the distinct advantage of addressing both probabilistic and non-probabilistic unsupervised learning tasks, making them particularly suitable for CL scenarios. By replacing the conventional softmax layer with an energy-based model classifier, [13, 12, 15] utilize energy scores as a novel output metric, theoretically aligned with the input's probability density and less prone to overconfidence issues. This approach does not only address the limitations of softmax in continual learning tasks but also provides a more flexible framework for managing sequential task learning.

In this paper, we present a novel Energy Sampling Replay-based (ESR) framework for Continual Learning in the context of computer vision, specifically tackling the challenge of catastrophic forgetting that arises when models are trained on sequential learning tasks. Leveraging the principles of energy models, our framework enhances the selection process of memory samples during the replay phase by utilizing the energy distribution of the samples. This approach improves the effectiveness of replay and contributes to increasing the overall accuracy of the model across various tasks. Through extensive experiments conducted across multiple datasets and CL methodologies, our framework demonstrates significant improvements in mitigating forgetting, particularly on CIFAR-10, CIFAR-100, and CIFAR-110 datasets [9]. The key contributions of this paper are as follows:

- We propose an energy-based sampling strategy that significantly improves the selection of memory samples for replay by analyzing their energy distribution, leading to more effective learning processes.
- We introduces a novel approach that combines random sampling with a Minimum versus Second-Minimum strategy. This hybrid sampling technique enables the selection of samples from the memory buffer that exhibit greater uncertainty and representativeness, enhancing the model's ability to handle diverse and dynamic data distributions effectively.
- The framework's effectiveness and efficiency are validated across a variety of datasets, demonstrating adaptability to different visual tasks and environments. By optimizing with energy-based sampling, our method improves model accuracy, offering substantial advantages for continual learning.

2 Related work

Continual Learning methods are crucial for addressing catastrophic forgetting in computer vision, primarily employing regularization, replay, and optimization strategies. Regularization methods, such as Elastic Weight Consolidation (EWC) [8] and Memory Aware Synapses (MAS) [2], aim to maintain the integrity of previously learned information by imposing constraints on the model's parameters, assessing the significance of each through mechanisms like the Fisher Information Matrix or unsupervised online assessments. These methods, while effective in preserving old knowledge, demand meticulous balancing to avoid overfitting on new tasks or eroding prior learning outcomes. Replay strategies [5, 19, 24], including experience, generative, and feature replay, focus on reconstructing past data distributions to bolster memory retention. The GEM [16] tackles catastrophic forgetting by leveraging episodic memory to reduce impacts on prior tasks and facilitate beneficial knowledge transfer, yet it requires further development in task descriptor utilization, memory management, and computational efficiency. The A-GEM [4] method refines GEM by averaging episodic memory losses, significantly boosting computational and memory efficiency, at the cost of some task-specific performance to gain broader applicability and simplicity. CLEAR [21] effectively mitigates catastrophic forgetting by combining on-policy learning with off-policy replay and behavioral cloning, thus enhancing stability and plasticity without needing detailed task knowledge; however, its extensive memory needs for storing past experiences pose limitations. Techniques such as the AQM [3] for experience replay or generative models for feature replay address issues like data imbalance or representation shifts, aiming for resource-efficient learning across tasks. Optimization-based methods [16, 4, 23] complement these by adjusting the optimization process to balance new and old information, ensuring dynamic adaptation and learning efficiency.

Researchers have significantly advanced the field of energy-based models, moving from the foundational Boltzmann machines [1, 22] to more sophisticated frameworks that suitable for deep learning architectures [10, 17, 18]. This progression highlights their efforts to provide a versatile approach to addressing unsupervised learning challenges, including clustering and feature extraction. EBMs, by defining an energy function that represents an unnormalized probability distribution, allow for a nuanced handling of data occurrence probabilities. Notably, in the context of CL [25, 7], EBMs have been explored for their potential to minimize interference between new and existing knowledge, providing an alternative to the softmax classifier's limitations. For instance, approaches like Energy-Aware Domain Adaptation (EADA) [26] leverage energy distributions to address domain adaptation, while novel methods [6] interpret classifier logits as energy functions, facilitating a seamless integration of data and label distributions. This advancement in EBMs showcases their effectiveness in mitigating overconfidence issues prevalent in softmax-based classifiers and enhancing model adaptability and generalization across continuous learning tasks.

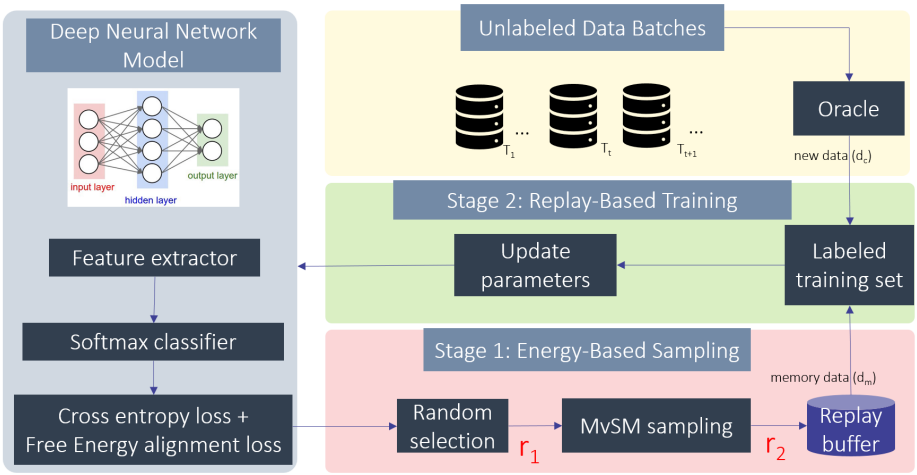

Fig. 1. Proposed framework of ESR.

3 Methodology

The CL framework proposed in this article is illustrated in Fig. 1. A dataset that is divided into several tasks, is fed into the network model for training in consecutive batches. T_1 represents the data for the first task, T_t represents the data for the t_{th} task, and $T_t + 1$ for the data of the $(t + 1)_{th}$ task. Each round of learning corresponding to a task is called an experience (E), and in the context of image classification tasks, each experience typically includes training samples from several classes. After an Oracle annotates the data at time T_t, it becomes new current data (d_c) for training E_t. Subsequently, a portion of samples from the previous round of training is randomly selected, and then, through a Minimum versus second-minimum method, samples are chosen to be stored in the replay buffer. This part of the data, referred to as memory data (d_m), and the previous current data (d_c) together form the labeled training set for the current experience. In stage two, the neural network model is trained using a replay-based approach, and the parameters are updated.

3.1 Replay-based method

The replay-based method helps consolidate knowledge learned from previous tasks and improve the stability and performance of learning by retaining a subset of training samples from earlier tasks in a memory buffer and retraining these samples in subsequent training processes. Therefore, when selecting samples to store in the memory buffer, it's essential to consider the balance between memory data and current data, allowing the model to learn new tasks while losing as little knowledge as possible from old tasks.

Random sampling for selecting training samples in replay-based CL methods encompasses numerous pitfalls, such as producing a biased and unrepresentative selection, especially from imbalanced datasets, leading to a model bias towards over-represented classes. This randomness may exclude diverse and informative samples crucial for a comprehensive understanding of previously learned tasks, undermining the memory buffer's role in effectively supporting new task learning or old task recall. Additionally, this approach may inefficiently allocate buffer space to less informative samples, increasing computational demands and diminishing learning efficiency. Performance-wise, indiscriminate sampling fosters learning imbalances, causing the model to disproportionately forget under-represented tasks and negatively impacting performance across a variety of tasks, a situation exacerbated in dynamic environments where data distributions evolve, and randomly selected samples swiftly become outdated. Therefore, we utilize an energy-based sampling method to improve upon this aspect.

3.2 Energy-based sampling

Within energy-based methods, the energy function outlines an unnormalized probability distribution, where lower energy levels indicate higher probabilities of data occurrence. Leveraging this property, we introduce an Energy Alignment Loss to address the issue of the model's inability to distinguish between old classes and new classes. Suppose the three shapes in Fig. 2 represent two old classes (rectangle and triangle) and one new class (circle), and their samples have biases as well as overlapping sections in the feature space. By setting a regularization term, samples on the feature domain boundaries of each class can be filtered out, similar to identifying samples that reflect domain divergence in domain adaptation tasks.

In the energy-based loss used for memory data sampling, the concept of free energy is pivotal for understanding the distribution and likelihood of input data. The free energy, denoted as $F(x)$, quantifies the "energy" or likelihood of an input instance x, with lower values indicating higher probabilities or more favorable states according to the model. Here, x represents an input instance, a feature vector derived from the dataset. The formula for calculating free energy is given by:

$$\mathcal{F}(x) = -\log \sum_{y \in \mathcal{Y}} \exp\left(-E(x, y)\right) \tag{1}$$

Within this formula, Y stands for the set of all possible labels in the classification task, and $E(x, y)$ is the energy function that assigns a scalar value representing the energy associated with the input x having a label y. This energy function is designed to yield lower energies for configurations of x and y that are more probable or correct, based on the model's training. The summation aggregates the exponentiated negative energies over all possible labels y, which is then transformed by the negative logarithm into the free energy $F(x)$. This transformation ensures that the free energy reflects a probabilistic measure, indicating the likelihood of the input x within the model's learned energy landscape. Utilizing

free energy in this way allows ESR to effectively assess and select informative unlabeled data from the old data, crucial for memory data sampling strategies.

The Free Energy Alignment (FEA) loss is employed to address the challenge of feature confusion by aligning the free energy distributions of memory data and current data. The calculation of FEA loss begins with the evaluation of the model's output on a batch of previous data to obtain the previous energy output. Meanwhile, the current data energy, which represents the free energy of samples from the current data, is used to calculate the current data batch's mean free energy. This mean acts as a reference (global mean) for aligning the free energy of the memory data.

The FEA loss itself is computed using a custom loss criterion, the Free Energy Alignment Loss, applied to the current energy output of the current data with respect to the global mean free energy. Mathematically, it is defined as:

$$\mathcal{L}_{fea}\left(x;\theta\right) = \max\left(0, \mathcal{F}\left(x;\theta\right) - \delta\right) \tag{2}$$

where $\mathcal{F}\left(x;\theta\right)$ is the energy output of the current data batch, and δ is the dynamically updated global mean free energy of memory samples. This approach effectively reduces the free energy bias between classes. It selects samples near the overlap of class features, which promotes more effective knowledge transfer from memory data to current data.

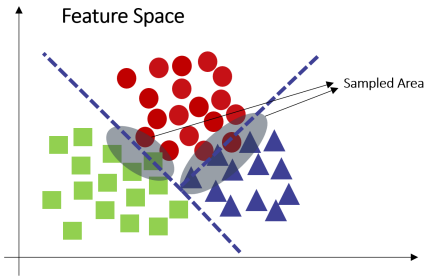

Fig. 2. Feature space representation of three classes.

3.3 Minimum versus second-minimum strategy

The selection process of stage one in Fig. 1 is divided into two steps: First, a certain number of samples, denoted as r_1 are randomly selected from the trained data. Then, from this subset, r_2 samples are selected using a Minimum versus Second-Minimum (MvSM) sampling strategy designed based on the energy distribution. During this process, the ratio between the two rounds of selection is defined as the parameter α, which determines the relationship between α and

the number of selected samples as follows:

$$\alpha = \frac{r_2}{r_1} \tag{3}$$

The selection strategy integrates a two-step approach beginning with the MvSM method to gauge the certainty of the model's predictions by calculating the difference between the two lowest energy scores for each sample. The equation is:

$$U(x) = E(x, y^*; \theta) - E(x, y'; \theta) \tag{4}$$

where y^* and y' are the lowest and the second-lowest energy output from the model. Samples with larger differences are seen as having clearer classification boundaries, making them prime candidates for initial selection. This approach first filters samples based on their free energy, prioritizing those with lower energy as more critical or representative, and selects a subset based on a predetermined ratio. The process then refines this selection by arranging the chosen samples according to their MvSM uncertainty values, with a preference for higher values to ensure training focuses on samples where the model is most confident. This strategy optimizes the learning process by carefully balancing exposure to both informative and challenging samples, enhancing the effectiveness of training within a CL setup.

4 Experiments and analysis

4.1 Experimental settings

In this section, we will detail the experimental setup and results analysis of our study. The ESR method proposed in this paper was implemented as a plugin within the Avalanche Continual Learning Library framework, and added to the training strategy. Since our method builds upon the basic Replay Plugin, the experimental results of the Replay Plugin were used as the baseline for comparison with our method in sections 4.2 and 4.4. The datasets were divided into several experiences for sequential training. The model was evaluated after each experience.

Dataset We evaluate the ESR Continual Learning Framework using CIFAR-10, CIFAR-100, and CIFAR-110 datasets. CIFAR-10 and CIFAR-100 are popular datasets for machine learning, featuring images across 10 and 100 classes, respectively. CIFAR-110 combines both to create a unique Continual Learning benchmark, starting with CIFAR-10 and incrementally introducing the diverse classes of CIFAR-100.

Network We adopt the ResNet50 backbone for feature extraction, modifying its initial convolutional layer to accommodate different input channels. A bottleneck linear layer reduces feature dimensionality, enhanced by batch normalization.

Class predictions were made through a linear classifier mapping the compressed features to their respective classes.

Our training employs a composite loss function that combines the FEA loss, as detailed in Section 3.2, with CrossEntropyLoss, essential for classification tasks. The total loss function $\mathcal{L}$ for our model, incorporating these components, is formally defined as:

$$\mathcal{L}(\theta; \mathcal{D}) = \sum_{i \in \mathcal{D}} \mathcal{L}_{fea}(x_i; \theta) + \sum_{i \in \mathcal{D}} -\log \left(\frac{\exp(o_{y_i})}{\sum_j \exp(o_j)} \right) \tag{5}$$

In this equation, θ denotes the parameters of our model, and $\mathcal{D}$ represents the dataset used for training. Each instance i in $\mathcal{D}$ contributes to the loss through the Free Energy Alignment loss $\mathcal{L}_{fea}(x_i; \theta)$ and the CrossEntropyLoss. The latter is calculated by taking the negative logarithm of the predicted probability for the true class y_i, normalized by the sum of exponential scores of all class logits o_j. This mechanism pushes the model to fine-tune its parameters to increase the probability of the actual class label while decreasing that of the incorrect labels.

After experimenting with various configurations, we ultimately selected specific parameters for the Stochastic Gradient Descent (SGD) optimizer that yielded the best experimental results. The final configuration for SGD that we employed uses a learning rate of 0.001 and a momentum of 0.9. These parameters were found to be optimal in achieving efficient convergence and robust training outcomes in our continual learning tasks.

Evaluation Metrics In our project, we utilize specific metrics from the Avalanche library to evaluate our CL model, including the Forgetting metric which is particularly crucial for assessing how well the model retains previously learned knowledge while acquiring new tasks. The formula for the Forgetting metric for a particular task k is simplified as follows:

$$\text{Forgetting}(k) = \left(\frac{C_{\text{init}}(k)}{N(k)} \right) - \left(\frac{C_{\text{sub}}(k)}{N(k)} \right) \tag{6}$$

Here, $C_{\text{init}}(k)$ represents the number of correct predictions immediately after the model is first trained on task k, capturing the initial mastery of the task. $C_{\text{sub}}(k)$ denotes the number of correct predictions after the model has been trained on subsequent tasks, indicating the retention of task k abilities amidst new learning experiences. $N(k)$ is the total number of predictions made for task k during assessments, ensuring that accuracy measurements are comparable.

This metric quantifies the decline in task performance, aiming for minimal forgetting to ensure effective knowledge retention across different learning tasks. By maintaining low values of Forgetting(k), the model demonstrates its capability to handle new information without significant loss of performance on previously learned tasks, an essential feature for continual learning models.

Additionally, the Average Mean Class Accuracy (AMCA) is utilized to evaluate the model's confidence and accuracy in predictions. The AMCA is computed as

follows:

$$AMCA = \frac{1}{C}\sum_{i=1}^{C} \text{Accuracy}_i \tag{7}$$

where C is the number of classes in the task, and Accuracy_i is the accuracy for the i^{th} class. This metric provides insights into the model's predictive confidence, with a higher AMCA value indicating better performance and higher prediction certainty.

Ideally, we aim for the model to retain as much knowledge of previous tasks as possible while learning new information, hence a lower forgetting metric is preferred. We desire a model not only to make accurate predictions but also to have high confidence in its predictions, making a higher AMCA value more desirable.

4.2 Comparison with state-of-the-art methods

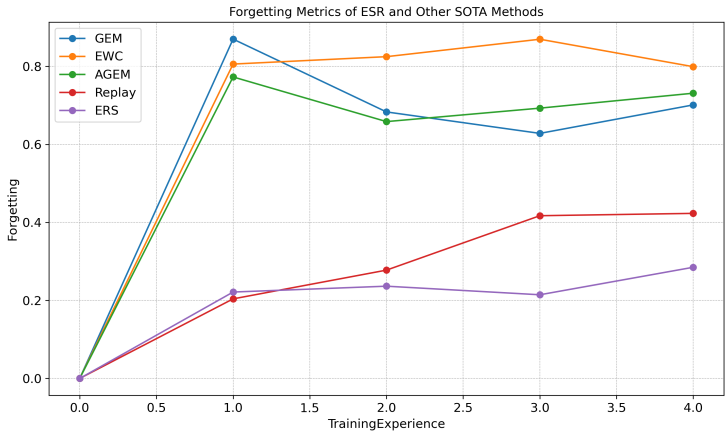

Fig. 3. Forgetting metrics of ESR and other state-of-the-art methods.

In this section, we compare the ESR method with four other state-of-the-art methods (EWC[8], GEM[16], AGEM[4], and Replay[21]) implemented as plugins within the Avalanche Continual Learning Library framework, using the SplitCIFAR10 dataset for our experiments. We set the number of experiences to 5, with CIFAR-10 comprising a total of 10 classes, meaning each experience involves learning two new classes. Figure 3 displays the forgetting metrics for these five methods. It is noticeable that the Forgetting values for the ESR (purple) and Replay (red) methods are lower compared to the other three methods, with our ESR method achieving a Forgetting value of 0.285. Thus, in this round of experiments, ESR demonstrates good performance in mitigating the model's

forgetting of old knowledge. As seen in Figure 4, the purple line representing the ESR method scores higher on the AMCA metric compared to the other four methods. This indicates that ESR can improve the quality of selected memory data based on the Replay method, thereby achieving greater confidence of the model in its classification of a sample.

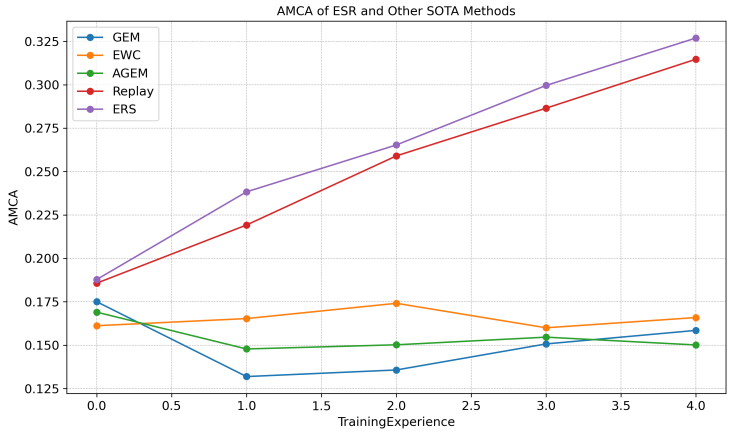

Fig. 4. AMCA of ESR and other state-of-the-art methods.

4.3 Effect of α in sampling memory data

In this section, we analyze the impact of the ratio of energy-based sampling to random sampling, denoted as α, on the experimental results. The selection of specific α values was based on two primary considerations:

1. **Sampling Density and Efficiency:** It was essential to ensure that the number of samples in the random sampling step was sufficient to maintain a dispersed representation across the feature space. This dispersion is critical for capturing the diversity of the dataset while enhancing the efficiency of the sample selection process by reducing the number of samples to a manageable size.
2. **Effectiveness of MvSM Sampling:** The chosen α values needed to allow the effectiveness of the MvSM to be clearly demonstrated. Values of α higher than 1/2 (0.5) yield samples that are not diverse enough to showcase the uncertainty of classes within the feature space. Conversely, values below 1/6 (0.1667) result in too few samples, diminishing their representativeness and the ability to generalize the sampling method's effectiveness.

Based on these considerations, the following α values were selected: 1/2 (0.5), 1/3 (0.3333), 1/4 (0.25), 1/5 (0.2), and 1/6 (0.1667). These values provide a

Table 1. The results of ESR with different α.

Strategy Name	r_1	r_2	α	Accuracy	AMCA	Forgetting
ESR-0.1667	4800	800	0.1667	0.4265	0.3084	0.295
ESR-0.2	4000	800	0.2	0.4304	0.3069	0.3264
ESR-0.25	3200	800	0.25	0.4635	0.3193	0.31
ESR-0.3333	**2400**	**800**	**0.3333**	**0.4967**	**0.3329**	**0.2695**
ESR-0.5	1600	800	0.5	0.4447	0.3283	0.343

balanced range from a lower to a higher preference for energy-based sampling over random sampling, allowing us to explore the influence of varying degrees of bias towards energy-efficient samples on learning dynamics.

From Table 1, it can be seen that when α equals 0.3333, the values of the Forgetting metrics are lower compared to the other four experiments after the last experience was trained. In the assessment of classification accuracy and confidence, the accuracy and AMCA at α equal to 0.3333 surpassed the other four experiments. This indicates that at this point in the training process, the energy-based sampling strategy can better balance the learning efficiency of new tasks and the consolidation of knowledge from old tasks.

4.4 Ablation studies

In order to investigate the impact of the ESR method on the replay-based approach, we systematically compared the approach incorporating the energy-based sampling strategy with the baseline replay-based method, which solely uses random sampling, across three datasets: CIFAR-10, CIFAR-100, and CIFAR-110. As shown in section 4.2, the ESR employing energy-based sampling method outperforms the Replay Plugin on the CIFAR-10 dataset.

Table 2. The results of ESR and Replay on the CIFAR-100 and CIFAR-110 datasets.

Strategy Name	Evaluation Metrics	Exp-3	Exp-7	Exp-11	Exp-15	Exp-19
ESR-CIFAR100	accuracy	**0.0969**	**0.1149**	**0.1368**	**0.1511**	**0.1897**
	forgetting	**0.3413**	**0.4226**	**0.4367**	**0.4703**	**0.4582**
Replay-CIFAR100	accuracy	0.0799	0.1053	0.1162	0.1167	0.1295
	forgetting	0.3733	0.58	0.6358	0.6856	0.6957
		Exp-2	Exp-4	Exp-6	Exp-8	Exp-10
ESR-CIFAR110	accuracy	0.2408	0.1663	0.1474	**0.1696**	**0.1598**
	forgetting	**0.3124**	**0.3892**	**0.4163**	**0.4403**	**0.4702**
Replay-CIFAR110	accuracy	**0.2591**	**0.2026**	**0.1661**	0.1227	0.1187
	forgetting	0.3848	0.51	0.5797	0.6535	0.7064

Table 2 presents the accuracy and forgetting metrics for experiments conducted on the CIFAR-100 and CIFAR-110 datasets. In this experiment, we utilized an α value of 0.25. It's notable that although the proposed sampling strategy may

slightly reduce the classification accuracy in the early stages of training, the advantages of employing energy-based sampling gradually become apparent as tasks accumulate, and the gap between this method and the baseline widens.

4.5 Discussion

The introduction of our ESR method marks a significant advancement in enhancing the information content of the samples selected for replay by utilizing their energy distribution. However, this strategy incurs increased computational overhead during the data loading phase. This involves computing the energy for each candidate sample and sorting these samples based on their energy levels. Although this meticulous selection process ensures the quality and representativeness of the samples, it also leads to a substantial increase in time consumption. Specifically, when training on the CIFAR-10 dataset using the ESR method, preparing for each round of experience requires approximately three times more time to select the replay data compared to a basic replay approach that utilizes random sampling.

To effectively balance sampling speed with the quality of samples, we conducted extensive experiments with different settings of the parameter α, as outlined in section 4.3. These experiments are critical for understanding how variations in α affect the efficiency and effectiveness of the learning process. While these studies demonstrate that adjusting α enables fine-tuning the trade-off between operational efficiency and sample quality, optimizing the CL process to meet various constraints and learning goals, it is important to acknowledge the limitations in the choice of α values. The range of α tested was limited, and there is a lack of deeper exploration into the optimal balancing points for sampling efficiency. This limitation points to a promising direction for future research. Further investigations could refine the balance between exploring new knowledge and exploiting learned experiences, potentially through adaptive mechanisms that dynamically adjust energy thresholds based on evolving data distributions. Such advancements could enhance the model's applicability and performance across diverse CL scenarios.

5 Conclusion

In this work, we propose an energy-based sampling strategy applied to the replay-based approach, which can effectively filter samples from old tasks that have a similar energy distribution to current data as memory data. The trained model is better at distinguishing classes between old and new tasks. The Energy Sampling Replay-based method outperforms several state-of-the-art methods in mitigating forgetting on the CIFAR-10 dataset. This approach also exceeds the performance of replay-based methods that do not utilize this strategy on the CIFAR-100 and CIFAR-110 datasets, demonstrating a strong ability to learn new tasks without forgetting the knowledge of old tasks. In our future work, we will attempt to apply the energy-based sampling method to more tasks, such as active learning.

6 Acknowledgement

We express our profound gratitude to Universiti Malaya for awarding the Universiti Malaya Scholarship Scheme as a Graduate Research Assistant under the government grant, MOHE - Kementerian Pendidikan Tinggi (KPT). This support has been crucial for the research (PPRN001A-2023) conducted at the Department of Artificial Intelligence, Faculty of Computer Science & Information Technology, under the guidance of Professor Dr. Loo Chu Kiong.

References

1. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for boltzmann machines. Cognitive science **9**(1), 147–169 (1985)
2. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Proceedings of the European conference on computer vision (ECCV). pp. 139–154 (2018)
3. Caccia, L., Belilovsky, E., Caccia, M., Pineau, J.: Online learned continual compression with adaptive quantization modules. In: International conference on machine learning. pp. 1240–1250. PMLR (2020)
4. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. arXiv preprint arXiv:1812.00420 (2018)
5. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H., Ranzato, M.: On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019)
6. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263 (2019)
7. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5830–5840 (2021)
8. 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. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
9. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
10. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
11. Lee, S.W., Kim, J.H., Jun, J., Ha, J.W., Zhang, B.T.: Overcoming catastrophic forgetting by incremental moment matching. Advances in neural information processing systems **30** (2017)
12. Li, J., Chen, P., He, Z., Yu, S., Liu, S., Jia, J.: Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11578–11589 (2023)
13. Li, S., Du, Y., van de Ven, G., Mordatch, I.: Energy-based models for continual learning. In: Conference on Lifelong Learning Agents. pp. 1–22. PMLR (2022)
14. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)

15. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
16. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)
17. Ranzato, M., Poultney, C., Chopra, S., Cun, Y.: Efficient learning of sparse representations with an energy-based model. Advances in neural information processing systems **19** (2006)
18. Ranzato, M., Boureau, Y.L., Chopra, S., LeCun, Y.: A unified energy-based framework for unsupervised learning. In: Artificial Intelligence and Statistics. pp. 371–379. PMLR (2007)
19. Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., Almahairi, A.: Progressive prompts: Continual learning for language models. arXiv preprint arXiv:2301.12314 (2023)
20. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
21. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. Advances in neural information processing systems **32** (2019)
22. Salakhutdinov, R., Larochelle, H.: Efficient learning of deep boltzmann machines. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 693–700. JMLR Workshop and Conference Proceedings (2010)
23. Tang, S., Chen, D., Zhu, J., Yu, S., Ouyang, W.: Layerwise optimization by gradient decomposition for continual learning. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 9634–9643 (2021)
24. Vitter, J.S.: Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS) **11**(1), 37–57 (1985)
25. Wang, Y., Li, B., Che, T., Zhou, K., Liu, Z., Li, D.: Energy-based open-world uncertainty modeling for confidence calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9302–9311 (2021)
26. Xie, B., Yuan, L., Li, S., Liu, C.H., Cheng, X., Wang, G.: Active learning for domain adaptation: An energy-based approach. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 8708–8716 (2022)