Continual Learning in Artificial Neural Networks:

Addressing Catastrophic forgetting

Author: Shambhavi Danayak
Student Id: 012654513
Contact: sdanayak@csuchico.edu

CSCI 693 - Research Methods in Computer Science

California State University – Chico

Chico, CA 95928

Abstract

Catastrophic forgetting is one of the main challenges in continual learning, where neural networks must learn new information while retaining previously learned knowledge. This study investigates the extent of catastrophic forgetting in a baseline sequential learning model and evaluates two widely studied classical solution approaches, Replay and Elastic Weight Consolidation (EWC)—across five sequential binary classification tasks derived from the MNIST dataset, in a controlled environment. Experiments were conducted using a two-layer feedforward multilayer perceptron trained on five sequential binary classification tasks derived from the MNIST dataset. An identical model architecture was used to ensure fair and direct performance comparisons. Model accuracy and forgetting rate were measured after each task to assess the stability–plasticity trade-off.

Results show that the baseline model experienced substantial forgetting as additional tasks were introduced. Replay method showed the strongest retention across all tasks while Elastics weight consolidation showed carried performance based on hyperparameter tuning. Overall, Replay achieved the best stability vs plasticity balance. The findings also highlight that while Replay requires memory to store samples from previous tasks, EWC becomes less effective as the number of tasks increases due to cumulative regularization constraints. These observations emphasize the need for careful evaluation when selecting continual learning methods for real-world applications.

INTRODUCTION

Context and Background

The ability of an Artificial neural network to learn new tasks and information sequentially from a data stream, integrating new knowledge while retaining previously learned information, is known as Continual Learning. Unlike humans, who can accumulate new knowledge without

forgetting previously learned information and have the ability to perform on all sequential tasks without compromising on the accuracy, neural networks typically overwrite old representations when trained on new tasks. This phenomenon, known as catastrophic forgetting, causes performance on previously learned tasks to deteriorate sharply when new data is introduced.

The human brain demonstrates remarkable resistance to forgetting. According to Neuroscience research, the human memory system includes 2 parts to it; Hippocampus, which supports rapid short-term learning, and Neocortex, which stabilizes long-term knowledge. Inspired by this, AI researchers have proposed strategies such as memory replay and elastic weight consolidation (EWC) to better manage knowledge retention.

Catastrophic forgetting is often framed within the plasticity–stability dilemma, which captures the tension between two competing needs:

- plasticity, the ability to learn new information, and
- stability, the ability to preserve previously learned knowledge.
  Excessive plasticity leads to forgetting, while excessive stability prevents the model from adapting to new tasks. Successfully balancing these two factors is essential for developing systems that can operate in dynamic environments.

Continual learning has become increasingly relevant as machine learning systems transition from isolated applications to real-world environments involving streaming data, personalization, and adaptive decision-making. From robotics and autonomous driving to medical monitoring and AI assistants, modern systems require the ability to learn continuously without retraining from scratch. Because of this, methods that preserve prior knowledge while enabling new learning are central to the advancement of long-term artificial intelligence.

Two major approaches to this challenge are Replay (rehearsal-based memory) and Elastic Weight Consolidation (EWC), which uses regularization inspired by biological synaptic stability. Both methods represent foundational strategies in continual learning research and are widely studied on benchmark datasets such as split-MNIST.

Problem Statement and Research Questions

Despite tremendous progress in continual learning, neural nets still lack reliable mechanisms for preserving previously learned knowledge when dealing with sequential tasks. Standard training procedures overwrite earlier representations, resulting in catastrophic forgetting, where task performance rapidly deteriorates as new data is introduced. This research examines whether classical mitigation techniques can meaningfully reduce forgetting in a controlled experimental environment.

This study investigates the following research questions:

- To what extent catastrophic forgetting occurs in a simple baseline neural network trained sequentially on split-MNIST tasks?
- How does Elastic Weight Consolidation (EWC) perform relative to the baseline model, and how sensitive is its performance to key hyperparameters such as regularization strength and learning rate?
- How effective is the Replay method at reducing catastrophic forgetting?
- Which continual learning method provides strong retention of previously learned knowledge while still allowing the model to learn new tasks in an experimental setting?
- Which method offers a better overall balance between stability (retaining past knowledge) and plasticity (learning new information)?

These outcomes will be used to determine whether classical continual learning methods effectively mitigate catastrophic forgetting while maintaining a balanced trade-off between stability and plasticity.

Significance of the Study

Understanding catastrophic forgetting is essential for building machine learning systems capable of long-term, adaptive behavior. This study is significant for several reasons:

- It provides insight into how neural network parameters evolve during sequential learning and clarifies the mechanisms through which Replay and EWC either preserve or

overwrite prior knowledge. This helps deepen the understanding of stability–plasticity trade-offs in continual learning.

- It provides a baseline to implement more advanced continual learning algorithms.This is valuable for real-world applications that require incremental updates — such as personalized AI, real-time monitoring systems, robotics, or adaptive user modeling — where retraining from scratch is impractical.
- The findings can guide the design of neural architectures that support continual learning under varying computational constraints. Understanding how methods behave in small-scale settings helps inform decisions about memory usage, regularization strength, and model capacity when building scalable, resource-aware learning systems.

Scope and Organization

The study aims to evaluate continual learning techniques within the domain of supervised classification sequential tasks. The experiment will use the MNIST dataset, divided into sequential binary tasks (e.g. {0,1}, {2-3}...) to simulate task-based learning. The scope of this study is to implement Elastic Weight Consolidation, Replay and a baseline sequential model all using the same network architecture. Performance of each method will be accessed using metrics such as accuracy, forgetting rate, and overall model stability. The study is intentionally limited to small-scale datasets and controlled experimental conditions to allow clear measurement and analysis of catastrophic forgetting under constrained computational resources.

The remainder of the paper is organized as follows: the Literature Review summarizes key research on catastrophic forgetting and continual learning strategies; the Methodology describes the dataset, model architecture, and implementation details for each technique; the Results section reports accuracy and forgetting outcomes for all models; the Discussion interprets these findings; and the Conclusion highlights the implications of the work and identifies directions for future research.

LITERATURE REVIEW

Overview of the Existing Research

Catastrophic forgetting has been widely documented as a fundamental limitation of neural networks trained on sequential tasks. As the network updates its parameters to learn new information, it tends to overwrite representations required for earlier tasks. This challenge has motivated a substantial body of research on continual learning.

A major study of continual learning methods is provided in the IEEE Continual Learning Survey: Defying Forgetting in Classification Tasks (2022), which categorizes solutions as Regularization-based methods such as Elastic Weight Consolidation, Replay methods, and dynamic architectural approaches such as progressive networks.

One of the most influential works in regularization-based continual learning is the paper Overcoming Catastrophic Forgetting in Neural Networks (Kirkpatrick et al., 2017), which introduces Elastic Weight Consolidation (EWC). EWC slows forgetting by penalizing changes to parameters deemed important for previous tasks, determined through an approximation of the Fisher Information Matrix.

Additionally, the IBM think article provides an industrial idea of how this challenge is being addressed and what work is currently going on.

Finally, open-source educational materials such as the GitHub repository Intro to Continual Learning provide practical demonstrations of catastrophic forgetting and walk through implementations of EWC and replay. These hands-on resources bridge the gap between theoretical understanding and practical application of continual learning algorithms.

Collectively, these sources provide a good foundation for researching catastrophic forgetting and understanding how continual learning methods aim to balance stability and plasticity.

Critical Evaluation of Literature

Although continual learning has advanced significantly, the existing research still faces several limitations and unresolved challenges. Many studies focus primarily on simple classification

tasks and controlled benchmarks, leaving open questions about how these methods scale to more complex, real-world problems involving evolving or non-stationary data streams.

Elastic Weight consolidation is one of the classical solutions for continual learning, it is very sensitive to hyperparameters such as lambda which determines how strongly the model penalizes changes to important weights. Moreover, EWC also becomes increasingly memory-intensive as the number of tasks grows, since it must store the Fisher information matrix and the optimal parameter values from previous tasks. This makes the method less practical when dealing with a large number of sequential tasks.

Replay methods usually perform better, but they come with their own drawbacks. They require storing samples from previous tasks or generating them later, which may not always be realistic or efficient, especially for large datasets or systems with limited memory.

Another common issue in the literature is that many studies rely on simple benchmark datasets such as MNIST or small subsets of CIFAR-10. While these datasets are useful for basic evaluation, they do not capture the complexity of real-world, continuously evolving data.

The IEEE survey also notes that no single continual learning method works best in all situations. Performance often depends on the specific task, dataset, and experimental setup.

Overall, these limitations show the need for studies that compare methods in a controlled and consistent way. Such experiments can help clarify how and why certain strategies succeed or fail, and provide clearer insights into their strengths and weaknesses.

Theoretical Framework

Continual learning research is guided by several conceptual models that parallel how the human brain operates. The brain maintains a balance between learning new information and retaining past experiences. This balance is formalized in machine learning as the stability–plasticity dilemma, which states that a learning system must remain plastic enough to acquire new

knowledge while stable enough to preserve prior knowledge. Catastrophic forgetting occurs when plasticity overwhelms stability, causing older representations to be overwritten.
Elastic Weight Consolidation (EWC)- where previously learned tasks serve as a prior distribution and the Fisher Information Matrix estimates which parameters are most important to retain. This mirrors the biological idea that certain synaptic connections become “consolidated” when they carry essential long-term information.
Humans strengthen memories during rest or sleep by reactivating neural patterns associated with earlier experiences. Similarly, artificial replay methods preserve prior knowledge by reintroducing samples from earlier tasks during training on new tasks.

Research Gap and Rationale

While prior work has demonstrated the effectiveness of EWC and Replay in mitigating catastrophic forgetting, several gaps remain. First, there are relatively few studies that analyze the sensitivity of EWC to hyperparameter tuning in simple neural architectures, despite its known dependence on regularization strength. Second, many studies report task accuracy without explicitly quantifying forgetting rates, limiting understanding of how performance degrades across sequential tasks. Finally, there is a lack of comparative analysis in controlled experimental baselines, making it difficult to differentiate continual learning methods based on performance or resource requirements.
This study addresses these gaps by conducting a controlled experimental comparison of a baseline sequential model, EWC, and Replay using a consistent neural architecture. By incorporating hyperparameter tuning, measuring forgetting rates, and evaluating stability across tasks, the study provides clearer insight into the trade-offs that characterize classical continual learning approaches.

METHODOLOGY

Research Design

This study uses an experimental research design to evaluate how different continual learning techniques affect a neural network's ability to retain knowledge across sequential tasks. The experiment involves training a multilayer perceptron (MLP) model with Relu activation function on 5 consecutive classification tasks ((e.g. {0-1}, {2-3}, …) derived from the MNIST dataset and comparing three learning strategies:

1. Baseline sequential learning
2. Replay-based continual learning
3. Elastic Weight Consolidation (EWC)

Participants and Settings

- The participants in the study are image samples from the MNIST handwritten digits dataset, consisting of 70,000 grayscale images. For continual learning taste are split as following:
    - Task 1: 0 vs 1
    - Task 2: 2 vs 3
    - Task 3: 4 vs 5
    - Task 4: 6 vs 7
    - Task 5: 8 vs 9
- MNIST dataset is split into Training, Validation, and test set.
- Sampling Method: A convenience sampling approach is used, as MNIST is a publicly available dataset typically used for academic machine learning experimentation.
- Research settings:
    - All experiments were conducted in Google colab notebook and saved on google drive for easy access. (Continual Learning Colab) using:
        - PyTorch
        - GPU acceleration (Free version of Tesla T4 GPU)
        - Python 3.12.12 version

```
import torch, platform
print("Python version:", platform.python_version())
print("Torch version:", torch.__version__)
print("CUDA available:", torch.cuda.is_available())
if torch.cuda.is_available():
    print("GPU:", torch.cuda.get_device_name(0))
else:
    print("GPU not detected")
```

```
Python version: 3.12.12
Torch version: 2.9.0+cu126
CUDA available: True
GPU: Tesla T4
```

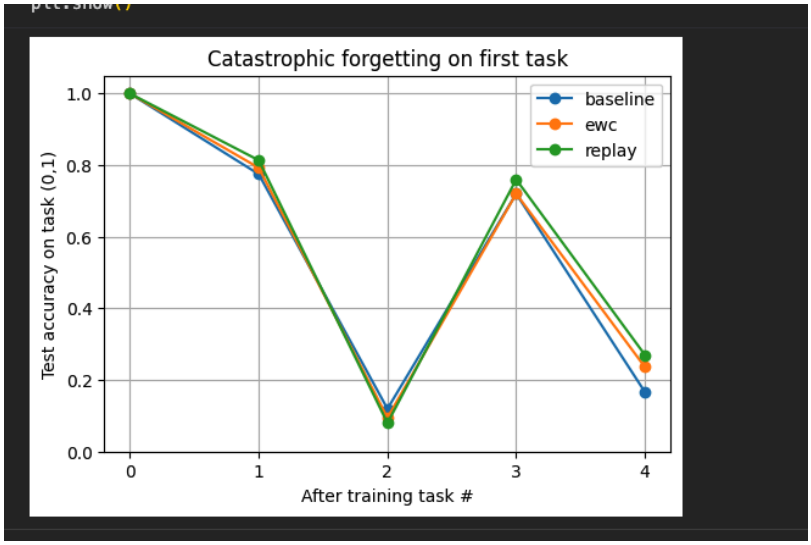
Data Collection Methods

- Dataset: MNIST Handwritten digits dataset (Kaggle)
- Data preparation:
    - Torchvision used to load the dataset
    - Project directory mounted to google drive for saving experimental work such as models, results, plots etc.
    - Split dataset for task 1 to 5 ({0 vs 1}, {2 vs 3}...)
- Model architecture:
    - A fixed 2 layer feedforward MLP with 256 hidden layers
    - Relu activation function
    - Adam optimizer
    - Cross Entropy loss
- Training Procedures
    - Each model (Baseline, EWC, Replay) will be trained sequentially across the MNIST tasks. After each task accuracy will be measured on all previously learned tasks to quantify catastrophic forgetting. Evaluation Metrics include:
        - Accuracy across all tasks (average)
        - Forgetting rate (drop in accuracy on earlier tasks)
        - Efficiency, learning rate, training time, memory usage.
- Reliability and validity
    - Using the same network architecture across all methods maintains internal validity, ensuring differences arise from the continual learning technique rather than model complexity.

- MNIST is a benchmark dataset providing external validity

Data Analysis Techniques

- For every learning method, accuracy was recorded to track model's performance trajectory as more tasks are added:
  - Immediately after learning each task
  - After training on all subsequent tasks
- Forgetting rate: how much accuracy the model loses on Task t from the best point in time to the end of training.
- Hyperparameter tuning:
  - EWC is sensitive to lambda value (low lambda value means high forgetting while high lambda value means strong regularization).
- Curves: better understand the rate of forgetting across all 3 methods.



Ethical Considerations

This study does not involve human subjects, personal information, or sensitive data. The MNIST dataset is publicly available and widely used for educational and research purposes.

RESULTS AND FINDINGS

Overview

Sequential learning experiments conducted using a baseline neural network, Elastic Weight Consolidation (EWC), and Replay-based continual learning. Model performance was evaluated across five sequential binary classification tasks derived from the MNIST dataset. Results are organized by learning method and focus on accuracy trends, forgetting behavior, and the effect of hyperparameter tuning.

Baseline Sequential Learning Performance

The baseline sequential learning model showed substantial performance degradation as new tasks were introduced (Fig1). After training on each task, accuracy on previously learned tasks declines sharply. By the completion of the final task, performance on the earliest tasks had dropped significantly, indicating severe catastrophic forgetting.
Figure 1 shows the baseline model's average accuracy across tasks as a function of training progression. The results show a consistent downward trend in retained performance, confirming that standard sequential training without continual learning methods is insufficient for preserving prior knowledge.
Figure 1

saved: /content/drive/MyDrive/continual_learning_project/results

	method	after_task	eval_on	val_acc	test_acc
10	baseline	4	(0, 1)	0.168317	0.166335
11	baseline	4	(2, 3)	0.694199	0.663415
12	baseline	4	(4, 5)	0.203463	0.195789
13	baseline	4	(6, 7)	0.853266	0.855701
14	baseline	4	(8, 9)	0.993704	0.993204

Elastic Weight Consolidation (EWC) Performance

Elastic Weight Consolidation demonstrated improved knowledge retention compared to the baseline model. The model's effectiveness varied depending on hyperparameter tuning. Elastic Weight Consolidation demonstrated improved knowledge retention compared to the baseline model.

In EWC, parameter importance was estimated after each task using the Fisher Information Matrix, which assigns higher importance values to parameters that contribute more significantly to task performance. During training on subsequent tasks, updates to these parameters were penalized in proportion to their estimated importance, thereby constraining changes to weights deemed critical for previously learned tasks.

After each task, parameter importance was estimated using the diagonal of the Fisher Information Matrix.

$$F_i = \mathbb{E}_{(x,y)\sim D_t}\left[\left(\frac{\partial}{\partial \theta_i}\log p(y \mid x, \theta)\right)^2\right]$$

The regularization strength λ (lambda) controls the trade-off between learning new tasks and preserving previously learned knowledge.

$$\mathcal{L}_{\text{EWC}} = \sum_i F_i \left(\theta_i - \theta_i^*\right)^2$$

Where:

- θ_i^* = parameter value after learning previous task
- F_i = Fisher importance for parameter i

Hyperparameter tuning on EWC

A grid search was conducted over key parameters including the regularization strength (λ), learning rate, number of training epochs, and the number of batches used to estimate the Fisher Information Matrix. Performance was evaluated using the final mean validation accuracy after completion of all five sequential tasks (Fig 2).

Figure 2 summarizes the top-performing hyperparameter configurations.The highest-performing configuration achieved a final mean validation accuracy of approximately 0.63, corresponding to λ = 1000, a learning rate of 0.0005, two training epochs, and three Fisher estimation batches. Configurations with lower regularization strength (e.g., λ = 200 or λ = 500) consistently produced lower final accuracy values. These findings demonstrate that EWC effectiveness is strongly influenced by hyperparameter selection, with suboptimal configurations resulting in increased forgetting and reduced task retention.

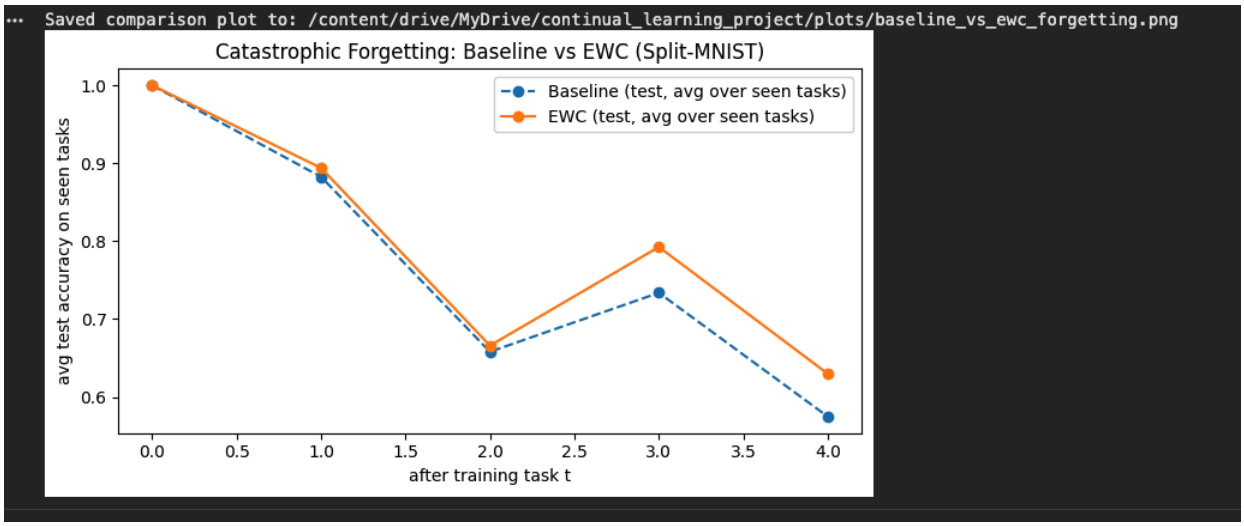
Figure 2

Best cfg (lambda, lr, epochs, fisher_batches) = (1000, 0.0005, 2, 3) score =

	lambda	lr	epochs	fisher_batches	final_mean_val_acc
38	1000	0.0005	2	3	0.627456
39	1000	0.0005	2	5	0.626534
36	1000	0.0005	1	3	0.625921
35	1000	0.0010	2	5	0.625656
37	1000	0.0005	1	5	0.622945
34	1000	0.0010	2	3	0.620431
31	500	0.0005	2	5	0.619202
28	500	0.0005	1	3	0.618857
29	500	0.0005	1	5	0.618429
23	200	0.0005	2	5	0.618411

Figure 3 presents a comparison between baseline sequential learning and Elastic Weight Consolidation using a fixed regularization strength (λ = 1000, batch size=64 and 5 epochs).

Figure 3



Replay-Based Continual Learning Performance

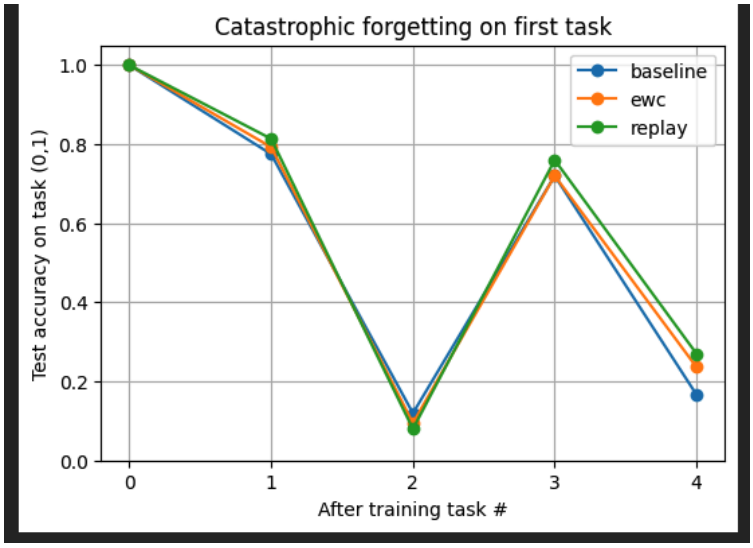
Replay-based continual learning achieved the strongest performance among all evaluated methods. By incorporating samples from previously learned tasks during training on new tasks, the Replay method maintained high accuracy across all tasks and exhibited minimal forgetting. Figure 4 shows that Replay consistently preserved performance on earlier tasks even after training on subsequent tasks.

Figure 4

	method	after_task	eval_on	val_acc	test_acc
0	baseline	0	(0, 1)	0.999100	1.000000
1	baseline	1	(0, 1)	0.731773	0.774900
2	baseline	1	(2, 3)	0.989184	0.990244
3	baseline	2	(0, 1)	0.104410	0.119522
4	baseline	2	(2, 3)	0.858407	0.855610

Figure 5 Compared to both the baseline and EWC models, Replay maintained the highest average accuracy and the lowest forgetting rate throughout the task sequence.

Figure 5



Forgetting Rate Comparison

Forgetting rates were computed by measuring the decrease in accuracy on each task relative to its peak performance during training.
The baseline model exhibited the highest forgetting rates across all tasks.Replay achieved the lowest forgetting rates. EWC reduces forgetting relative to the baseline, particularly when tuned with higher regularization strength. (Fig 5)

Summary of Key Findings

The main findings from this study include:

- Baseline sequential learning resulted in severe catastrophic forgetting.
- EWC improved retention compared to the baseline but remained sensitive to hyperparameter selection.
- Replay-based continual learning consistently achieved the highest task accuracy and lowest forgetting rates.
- Hyperparameter tuning played a critical role in determining EWC effectiveness.

DISCUSSION

Interpretation of Results

This study investigated the extent of catastrophic forgetting in sequential neural network training and evaluated the effectiveness of Elastic Weight Consolidation (EWC) and Replay-based continual learning strategies. The results show that a baseline neural network trained sequentially without mitigation mechanisms experiences severe catastrophic forgetting, with performance on earlier tasks deteriorating rapidly as new tasks are introduced.
Elastic Weight Consolidation reduced forgetting relative to the baseline model, confirming that constraining updates to important parameters can help preserve previously learned knowledge. Additionally, EWC performance was highly sensitive to hyperparameter tuning, particularly the

regularization strength λ. Higher λ values resulted in improved retention but also limited the model's flexibility to adapt to new tasks. These findings align with Kirkpatrick et al. (2017), who noted that EWC requires careful tuning to balance stability and plasticity.

Replay-based continual learning consistently achieved the strongest performance across all tasks, exhibiting minimal forgetting and maintaining high accuracy throughout the task sequence. This outcome supports prior research showing that rehearsal-based methods are among the most effective approaches for mitigating catastrophic forgetting on benchmark datasets such as split-MNIST. By reintroducing samples from previous tasks during training, Replay directly reinforces earlier representations and avoids many of the constraints imposed by regularization-based methods.

Implications

This study findings highlights the trade-offs that must be considered when deploying continual learning systems in real-world environments. Applications with sufficient memory resources may benefit most from replay-based strategies, whereas memory-constrained systems may require carefully tuned regularization-based approaches or hybrid solutions.

Additionally, the study also demonstrates the importance of hyperparameter sensitivity analysis in continual learning research. Without tuning, EWC performance may be underestimated, leading to misleading conclusions about its effectiveness.

Limitations

There were several limitations that affected study findings and interpretations. First, the experiments were conducted on the MNIST dataset, which represents a relatively simple and well-structured classification problem. While MNIST helped with controlled evaluation, it does not capture the complexity of real-world data streams. Second, the study used a simple feedforward neural architecture, which may limit the generalizability of the findings to deeper or more complex models. Finally, computational constraints restricted the scope of hyperparameter exploration and task sequence length, which may influence scalability conclusions.

Suggestions for Future

Based on the research results, several follow-up studies can be conducted, such as:

- Expand to a more complex dataset, **e**valuate methods on datasets such as CIFAR-10, Fashion-MNIST etc. to test scalability.
- Explore Hybrid Approaches, combine replay with EWC or other regularization strategies to investigate whether hybrid models can balance memory efficiency and performance.
- Investigate Alternative Regularization Methods, such as Synaptic Intelligence, memory aware Synapses etc.

CONCLUSION

This study examined catastrophic forgetting in neural networks across 5 sequential MNIST tasks. The study evaluated 3 methods, baseline model, Replay and Elastic Weight Consolidation. The Baseline model results clearly highlight the challenge of catastrophic forgetting that standard neural network model faces. EWC reduced forgetting compared to the baseline, with performance highly dependent on the choice of λ. Replay consistently achieved the strongest retention of past knowledge while still supporting effective learning of new tasks. Overall, Replay achieved the best balance between stability and plasticity in this experimental setting.

Research Contribution

This research addresses a key gap in continual learning research by providing a controlled, multi task experimental comparison using a fixed environment. The study provides a cleaner and more in-depth understanding of classical continual learning methods in a controlled environment.

Closing Thoughts

Continual learning remains one of the challenges in developing adaptive and autonomous AI systems. The findings of this study highlight the importance of memory-based strategies for achieving long-term stability and show that even foundational methods like Replay can

significantly mitigate forgetting when applied effectively. This study provides a very strong foundation for future research.

REFERENCES

1. M. De Lange et al., "A Continual Learning Survey: Defying Forgetting in Classification Tasks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3366-3385, 1 July 2022, doi: 10.1109/TPAMI.2021.3057446.
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, Mar. 2017. doi:10.1073/pnas.1611835114
3. M. Serra-Perello and A. Ortiz, “Incremental learning methodologies for addressing catastrophic forgetting: Analysis and Experimental Evaluation,” Journal of Artificial Intelligence Research, vol. 83, Aug. 2025. doi:10.1613/jair.1.18405
4. I. Belcic and C. Stryker, “What is catastrophic forgetting?,” IBM, https://www.ibm.com/think/topics/catastrophic-forgetting (accessed Sep. 26, 2025).
5. clam004, “Clam004/intro_continual_learning: This is a tutorial to connect the fundamental mathematics to a practical implementation addressing the continual learning problem of Artificial Intelligence,” GitHub, https://github.com/clam004/intro_continual_learning (accessed Sep. 26, 2025).
6. “Continual Learning AI,” YouTube, https://www.youtube.com/watch?v=z9DDg2CJjeE&list=PLm6QXeaB-XkBfM5RgQP6wC R7Jegdg51Px (accessed Oct. 10, 2025).
7. G. van de Ven, D. Kudithipudi, and K. Leuven, “Home,” Cognitive Computational Neuroscience, https://2024.ccneuro.org/k-and-t-continual-learning/#:~:text=continually,the%20brain%20 that%20underlie%20the (accessed Oct. 10, 2025).
