Mitigating Catastrophic Forgetting in Continual Learning Using the Gradient-Based Approach: A Literature Review

Haitham Ghallab, Mona Nasr, Hanan Fahmy
Department of Information Systems-Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

***Abstract*—Continual learning, also referred to as lifelong learning, has emerged as a significant advancement for model adaptation and generalization in deep learning with the capability to train models sequentially from a continuous stream of data across multiple tasks while retaining previously acquired knowledge. Continual learning is used to build powerful deep learning models that can efficiently adapt to dynamic environments and fast-shifting preferences by utilizing computational and memory resources, and it can ensure scalability by acquiring new skills over time. Continual learning enables models to train incrementally from an ongoing stream of data by learning new data as it comes while saving old experiences, which eliminates the need to collect new data with old data to be retrained together from scratch, saving time, resources, and effort. However, despite continual learning advantages, it still faces a significant challenge known as catastrophic forgetting. Catastrophic forgetting is a phenomenon in continual learning where a model forgets previously learned knowledge when trained on new tasks, making it challenging to preserve performance on earlier tasks while learning new ones. Catastrophic forgetting is a central challenge in advancing the field of continual learning as it undermines the main goal of continual learning, which is to maintain long-term performance across all encountered tasks. Therefore, several types of research have been proposed recently to address and mitigate the catastrophic forgetting dilemma to unlock the full potential of continual learning. As a result, this research provides a detailed and comprehensive review of one of the state-of-the-art approaches to mitigate catastrophic forgetting in continual learning known as the gradient-based approach. Furthermore, a performance evaluation is conducted for the recent gradient-based models, including the limitations and the promising directions for future research.**

Keywords*—Deep learning; continual learning; model adaptation and generalization; catastrophic forgetting; gradient-based approach

I. Introduction and Problem Definition

Human beings and other species possess the innate ability to learn, adapt, and retain information throughout their existence. This natural capability, termed continuous learning, is supported by neurocognitive mechanisms that enable organisms to dynamically adapt to new experiences while retaining prior knowledge. Neurocognitive mechanisms involve a complex interplay of neurons and synapses that dynamically process, store, and retrieve information. The brain achieves this remarkable feat through processes such as neuroplasticity, which allows neural pathways to adapt in response to new experiences, and consolidation, which stabilizes memories and integrates them with prior knowledge. This enables humans and animals to continuously acquire, refine, and transfer knowledge while retaining previous learning [1]. For example, humans can learn new skills, such as playing a musical instrument, without losing the ability to perform unrelated tasks like speaking or walking. And since deep learning mimics certain aspects of the human brain, particularly how neurons in the brain process and transmit information, then deep learning can use this biological efficiency to incrementally train its models, but unfortunately unlike the human neurocognitive mechanisms, mimicking continual learning in artificial neural networks contrasts sharply with the challenges referred to as catastrophic forgetting—a phenomenon where the acquisition of new knowledge disrupts or erases previously learned information. Addressing this limitation is central to advancing continual learning systems in artificial intelligence [2].

Continual learning, also known as lifelong learning, seeks to emulate the brain's neurocognitive mechanisms in artificial systems. By enabling models to incrementally learn and adapt to new information without forgetting past knowledge, continual learning systems aim to achieve human-like adaptability. These systems have far-reaching implications for applications in dynamic environments, such as robotics, autonomous vehicles, financial forecasting, environmental monitoring, adaptive user interface, and personalized healthcare, where consistent performance across evolving tasks is essential [2], [3]. Deep learning has revolutionized the interaction with technology and process data. By mimicking the way the human brain works, it enables systems to learn from experience, adapt to new information, and perform tasks without explicit programming. This makes deep learning crucial in automating complex processes, improving accuracy, and handling vast amounts of data, which are essential in today's data-driven world.

Before continual learning and other common model adaptation and generalization paradigms, deep learning models used to be trained using fixed datasets, which caused a major challenge especially in dynamic real-time environments where new data arrives continuously and data distribution shifts sharply therefore deep learning models struggled to maintain accuracy.

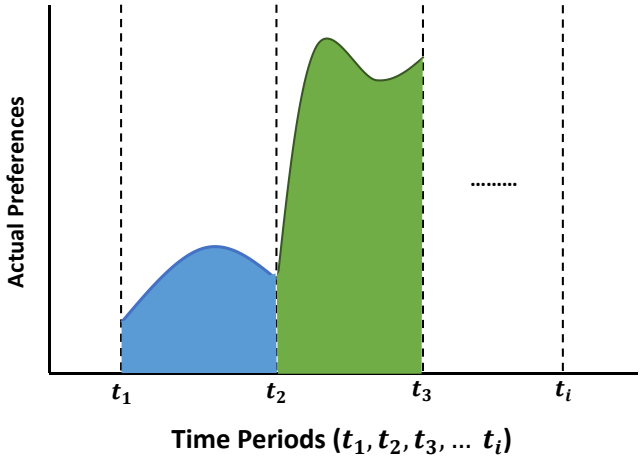

Fig. 1. Data distribution shifts in dynamic environments.

Additionally, deep learning models require significant computational power and large-scale datasets to function effectively, especially for training, as these models often need specialized hardware like GPUs and large amounts of labeled data to produce accurate results [4], which is a challenge especially when building models for real-time multi-task classification purposes in dynamic environments. This is a key problem in deep learning and real-time analysis, where data preferences and patterns are changed rapidly over time as shown in Fig. 1. In this case, traditional deep learning models often use an iterative deployment mechanism to keep up with the changing patterns by collecting the new arriving stream of data with old experiences and training the model from scratch to include the entire set of data and not lose efficiency, and this solution is computationally expensive and inefficient.

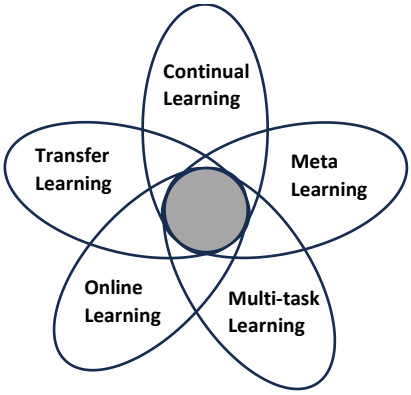

Fig. 2. Model adaptation and generalization paradigms.

As a result, several model adaptation and generalization paradigms have been proposed to address this challenge either by incrementally training deep learning models with new experiences as in continual learning (also referred to as incremental learning, lifelong learning, continuous learning) [5], or by allowing models to adapt to new unseen tasks by utilizing related past experiences using multi-task learning, meta-learning, transfer learning, or online learning [5], [6], [7], [8], [9]. As shown in Fig. 2, although the objectives of these learning paradigms may differ, they may overlap in certain aspects, which sometimes may confuse the researchers. Table I shows the main differences and highlights the focus of each approach.

Although, continual learning shows a significant contribution to deep learning, especially in dynamic environments by enabling models to adapt efficiently and maintain accuracy to data distribution shifts and multi-task processes, with minimum and constant computation powers and memory utilization. But continual learning still faces a significant challenge known as catastrophic forgetting [2], [5]. Catastrophic forgetting, also referred to as catastrophic interference, is a phenomenon in continual learning where a model forgets previously learned knowledge when trained on new tasks, making it challenging to preserve performance on earlier tasks while learning new ones. Catastrophic forgetting is a central challenge in advancing the field of continual learning as it undermines the main goal of continual learning, which is to maintain long-term performance across all encountered tasks [2], [5]. However, it is important to highlight that while some model adaptation and generalization paradigms, such as multi-task learning and transfer learning, improve learning efficiency and generalization across tasks, they do not directly address the issue of catastrophic forgetting [6],[9]. Unlike continual learning, which is specifically designed to retain previously learned knowledge while learning new tasks sequentially [5].

TABLE I MODEL ADAPTATION AND GENERALIZATION PARADIGMS

Learning Paradigm	Main Objective
Continual Learning [5]	Enable models to learn from a continuous stream of tasks without forgetting previously learned knowledge, which is very significant in dynamic environments with high data distribution shits over time.
Multi-task Learning [6]	Enable models to solve multiple related tasks simultaneously. Instead of treating each task independently, models leverage shared information across similar tasks to improve the learning efficiency and generalization performance of all tasks.
Meta Learning [7]	Enable models to perform effectively in rare unseen tasks where datasets are currently evolving and not yet available, by utilizing similar experiences from related tasks. It requires diverse task datasets for meta-training, and few-shot or limited data for testing/adaptation.
Online Learning [8]	Enable models to immediate and real-time short-term adaptation to the recent observations and does not inherently address long-term retention of knowledge or ensure that past patterns are preserved, for instance online learning models might update their predictions as new data arrives without revisiting historical data.
Transfer Learning [9]	Enable models to reuse knowledge learned from a source task or domain to improve learning on a target task or domain. Instead of training a model from scratch for every task. It involves pretraining on a large source dataset and fine-tuning on the target task.

So, the primary objective of this research is to address the main issues of catastrophic forgetting in continual deep learning and recent potential solutions. Section II introduces the background and key concepts underlying continual deep learning, establishing a foundational understanding of the topic. Section III presents an overview of the latest gradient-based approaches developed to mitigate catastrophic forgetting in continual deep learning. Finally, Section IV offers a

discussion of the research and Section V presents conclusions, limitations, and directions for future work.

II. Background and Key Concepts

The field of continual learning addresses one of the most significant challenges in artificial intelligence: enabling systems to learn sequentially from non-stationary data while retaining previously acquired knowledge. Unlike traditional deep learning without adaptation and generalization capabilities, where models are trained on static datasets, continual learning operates in dynamic environments where new tasks or patterns continuously emerge. However, this learning paradigm faces two fundamental challenges:

a) Catastrophic forgetting: The tendency of neural networks to lose performance on previously learned tasks when trained on new ones [10].

b) Stability-plasticity dilemma: The need to balance the retention of old knowledge (stability) with the incorporation of new information (plasticity) [11].

These challenges have profound implications for real-world applications, including robotics, autonomous systems, and personalized assistants, where adaptability and knowledge retention are paramount. Below, the research explained the key concepts underpinning continual learning in deep learning tasks, and its associated challenges to establish a foundational understanding of the topic.

A. Continual Learning

Continual Learning, also known as lifelong learning, is one of the most common model adaptation and generalization paradigms, as it refers to the ability of a deep learning model to learn from a continuous stream of tasks without forgetting previously learned knowledge [5], [12]. Unlike traditional deep learning setups, where models are trained on a fixed dataset, continual learning simulates a dynamic learning environment where new tasks emerge sequentially. For instance, continual learning aims to train models on a sequence of N tasks $T_1, T_2, T_3, \ldots, T_N$, where each task T_i is defined by its dataset $D_i = \{(x_j, y_j)\}_{j=1}^{n_i}$ and its objective $L_i(\theta)$, with θ denoting the model parameters, and the purpose of the model is to minimize the cumulative loss across all tasks, as shown in Eq. (1).

$$\theta^* = arg \min_{\theta} \sum_{i=1}^{N} L_i(\theta) \qquad (1)$$

So, the continual learning model main objective is to perform this optimization without degrading performance on earlier tasks, $T_1, T_2, T_3, \ldots, T_{i-1}$ when learning task T_i, but this is challenging because the datasets D_i are typically not accessible once the task T_i is completed, and therefore continual learning model making it hard to preserve performance on earlier tasks while learning new ones and respectively mitigating the famous problem of continual deep learning known as the catastrophic forgetting dilemma [10], as shown in Fig. 3.

For example, Fig. 3 shows the performance of a baseline model and a continual learning model across sequential tasks $T_1, T_2, T_3, \ldots, T_{i-1}$. The baseline model suffers a sharp decline in earlier task performance (catastrophic forgetting), while the continual learning model maintains better stability. Furthermore, continual learning must navigate the delicate balance between maintaining model stability—preserving knowledge from earlier tasks and ensuring sufficient plasticity to adapt and learn new information as it becomes available, which is another challenge in continual learning known as the stability-plasticity dilemma.

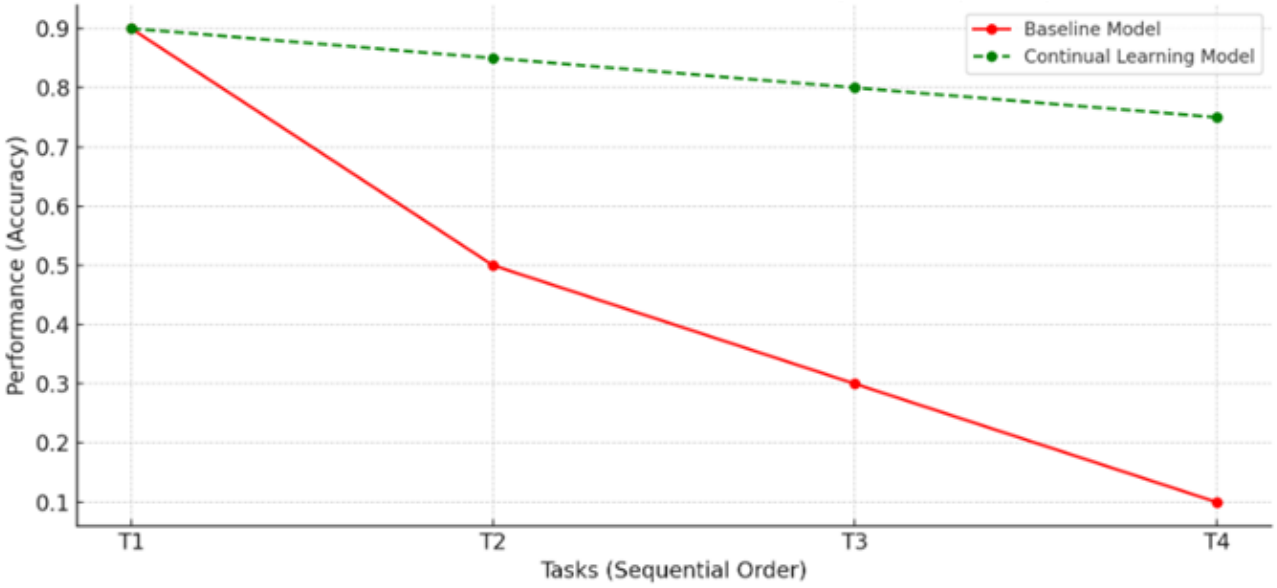

Fig. 3. Task performance over time (Catastrophic forgetting).

B. Catastrophic Forgetting

Catastrophic forgetting is a main challenge in building continual deep learning models. It refers to the drastic decline in a neural network's performance on previously learned tasks when it is trained on new tasks [2], [10]. This phenomenon occurs because deep learning models typically share a common set of parameters θ across all tasks. When trained on a new task T_{i+1}, the optimization process updates θ to minimize the loss for the new task, inadvertently overwriting information critical to earlier tasks. For instance, consider a model trained sequentially on two tasks:

1) Task 1 (T_1): Loss function $L_1(\theta)$ with dataset $D_1 = \{(x_1, y_1)\}$

2) Task 2 (T_2): Loss function $L_2(\theta)$ with dataset $D_2 = \{(x_2, y_2)\}$

During training on T_1, the model learns parameters θ^* by minimizing $L_1(\theta)$, as shown in Eq. (2).

$$\theta_{T1}^* = arg \min_{\theta} L_1(\theta) \quad (2)$$

When training begins on T_2, the model minimizes $L_2(\theta)$, as shown in Eq. (3).

$$\theta_{T2}^* = arg \min_{\theta} L_2(\theta) \quad (3)$$

However, the gradients $\nabla L_2(\theta)$ used to optimize T_2 often conflict with $\nabla L_1(\theta)$. This results in updates to θ that degrade the performance on T_1, i.e., $L_1(\theta_{T_2}^*) > L_1(\theta_{T_1}^*)$. As a result, catastrophic forgetting become a central challenge in advancing the field of continual learning in deep learning tasks as it undermines the main goal of continual learning, which is to maintain long-term performance across all encountered tasks. Therefore, the next section will address several types of research that have been proposed recently to address and mitigate the catastrophic forgetting dilemma to unlock the full potential of continual deep learning [2], [10].

C. Stability-Plasticity Dilemma

The stability-plasticity dilemma is a core challenge in continual deep learning. It refers to the trade-off between:

1) Stability: The ability of a model to retain and preserve knowledge from previously learned tasks [11].

2) Plasticity: The ability of a model to adapt to new tasks and incorporate new knowledge effectively [11].

In continual deep learning, achieving a balance between these two opposing forces is critical. Excessive stability can lead to rigidity, where the model fails to adapt to new tasks. On the other hand, excessive plasticity can cause catastrophic forgetting, where new learning overwrites previously acquired knowledge [13]. For instance, consider a sequence of *N* tasks $T_1, T_2, T_3, \ldots, T_N$, where each task T_i has its dataset D_i and loss function $L_i(\theta)$, with θ denoting the shared model parameters. The objective in continual learning is to minimize cumulative loss, as shown in Eq. (4).

$$L(\theta) = \sum_{i=1}^{N} L_i(\theta) \quad (4)$$

To address the stability-plasticity dilemma, the following constraints are defined:

3) Stability constraint: For previously learned tasks $T_j (j < i)$, the loss should not increase beyond a threshold ϵ, as shown in Eq. (5).

$$L_j(\theta) \leq L_j\left(\theta_{T_j}^*\right) + \epsilon \quad (5)$$

where, $\theta_{T_j}^*$ is the parameter configuration after training on task T_j.

4) Plasticity constraint: The model must minimize the loss for the current task T_i, as shown in Eq. (6).

$$\theta = arg \min_{\theta} L_i(\theta) \quad (6)$$

where, the gradient updates for θ often result in interference between tasks. If the gradients for T_i conflict with T_j, the performance on earlier tasks degrades.

As a result, maintaining a balance between these two forces (stability and plasticity) ensures that the model's parameter space retains important information for older tasks, typically through regularization or rehearsal, and it also enable the model to adapt to new tasks and incorporate new knowledge effectively [11], [13].

III. Mitigating Catastrophic Forgetting Approaches

As mentioned above, continual learning in deep neural networks faces two intertwined challenges: maintaining a balance between stability and plasticity and addressing catastrophic forgetting. Several approaches have been proposed to tackle these two challenges as shown in Fig. 4 categorized broadly into regularization-based, knowledge-distillation-based, Bayesian-based, gradient-based, architecture-based, replay-based, and other hybrid methods.

In this research, the main objective is to focus on the gradient-based approach including its definition, strengths, weakness points, and its recent models including Gradient Episodic Memory (GEM) [14], Averaged Gradient Episodic Memory (A-GEM) [15] and Orthogonal Gradient Descent (OGD) [16]. Additionally, the research presents how the gradient-based approach differs from the other approaches based on factors such as the core idea, memory requirements, computational efficiency, and flexibility.

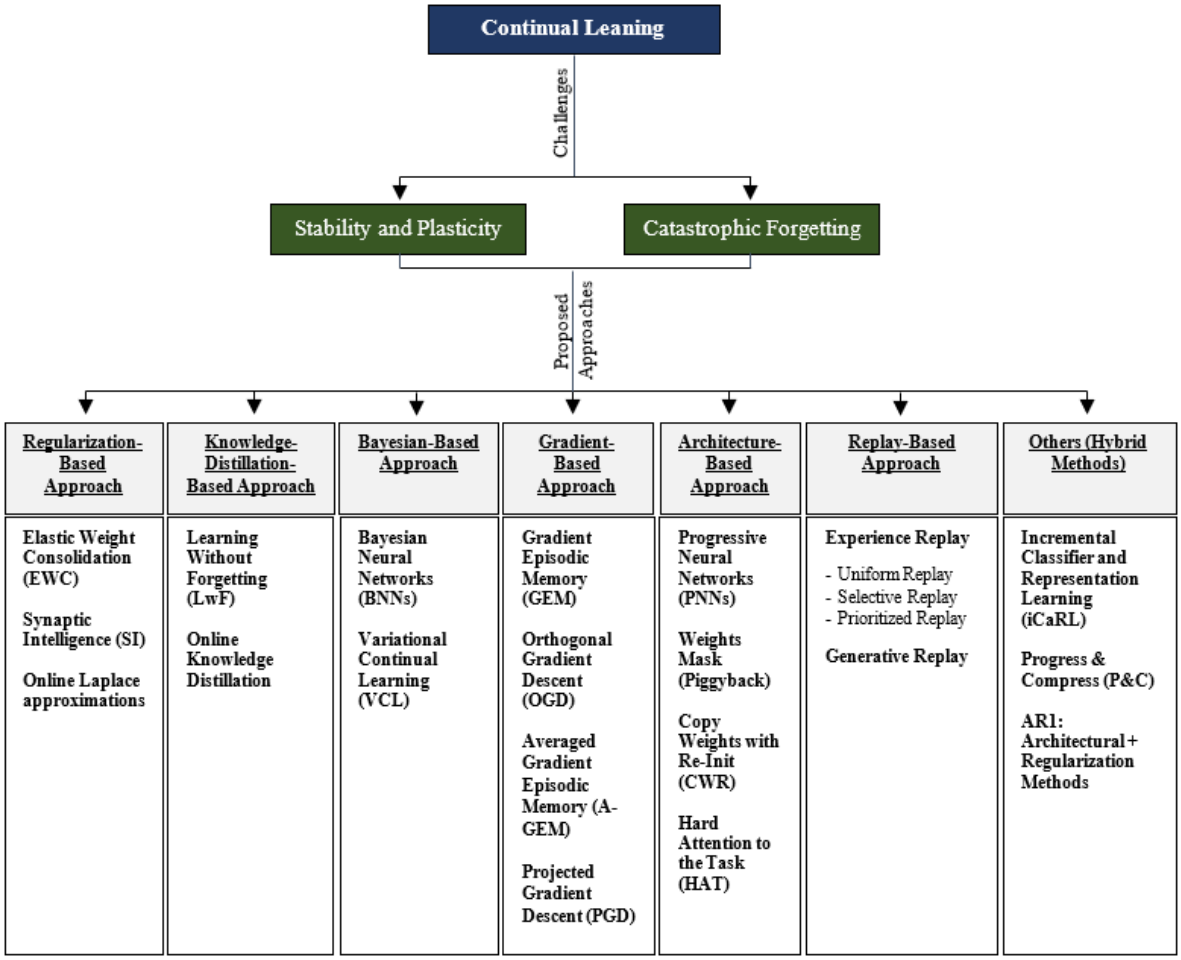

Fig. 4. Mitigating catastrophic forgetting approaches taxonomy.

A. Gradient-Based Approach

The gradient-based approach is a prominent methodology in continual deep learning, designed to address the challenge of catastrophic forgetting by carefully adjusting the gradient updates when a model learns new tasks. The key idea is to modify the gradient direction to minimize interference with previously learned tasks, ensuring a balance between retaining old knowledge and learning new information [14]. This approach operates within the optimization process, ensuring that parameter updates for new tasks do not disrupt the knowledge gained from previous ones. By focusing on the dynamics of gradients during training, it provides a flexible and efficient framework for tackling forgetting without relying on extensive memory storage or architectural changes. For instance, if a neural network first trained on Task A, where the goal is to classify points into two categories (e.g. red and blue) based on their positions in a 2D plane. After mastering Task A, the network is asked to learn Task B, which involves classifying points into two different categories (e.g. green and yellow) based on a new data distribution. Without careful control, when the network learns Task B, the gradients calculated for this task might overwrite what was learned for Task A, causing catastrophic forgetting. Gradient-based approaches solve this by modifying how the network updates its parameters. Let's illustrate this with Gradient Episodic Memory (GEM):

1) Memory of task A: GEM keeps a small memory buffer of examples from Task A (e.g. a few red and blue points). These points represent the knowledge of Task A that the model should not forget.

2) Gradient check: When the model computes the gradient to learn Task B, GEM checks if this gradient would increase the loss on the stored Task A examples. If it does, GEM modifies the gradient to ensure that the loss on Task A examples does not worsen.

3) Adjusted gradient update: The model updates its parameters using the adjusted gradient, which balances learning for Task B while preserving performance on Task A.

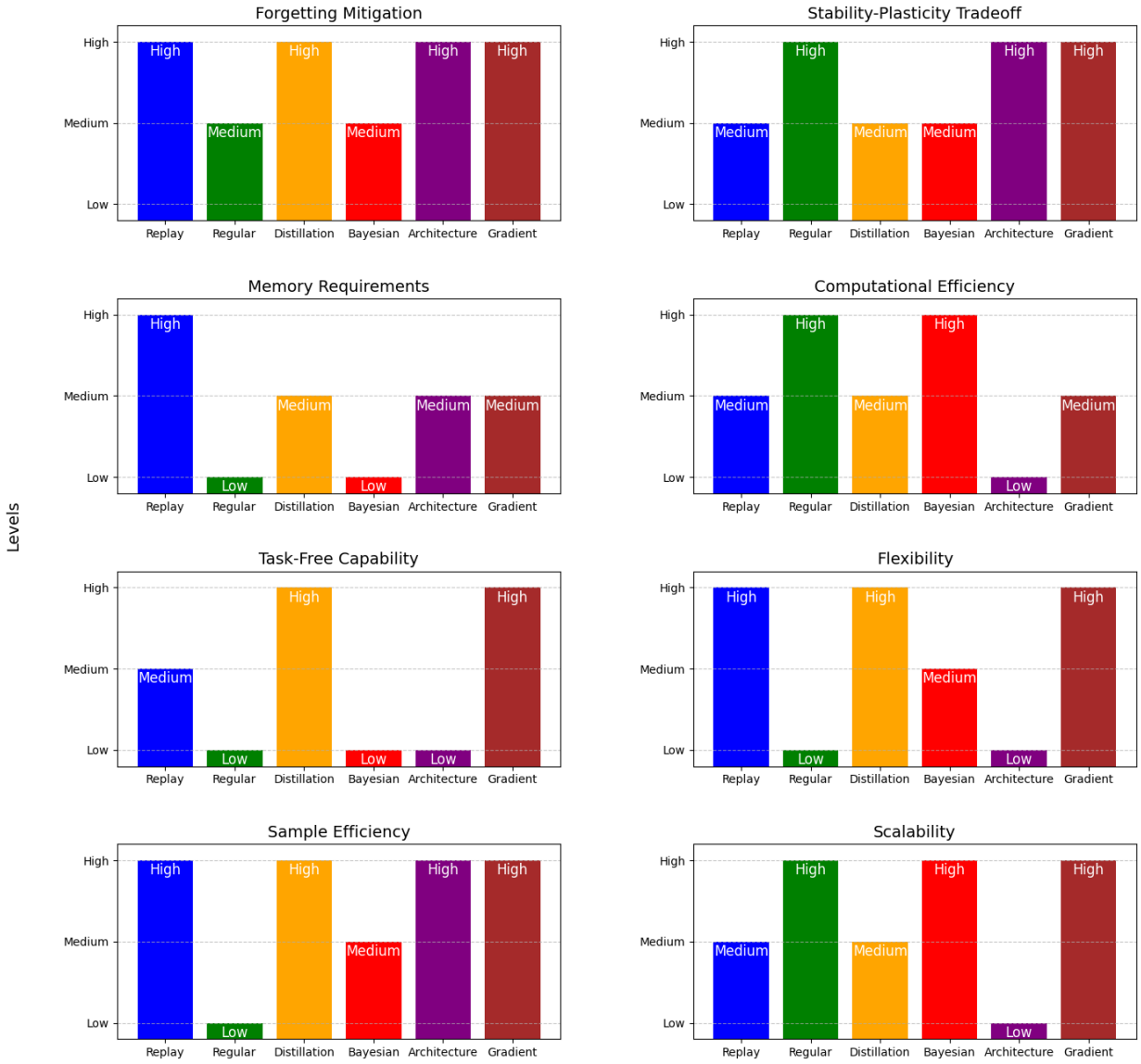

Fig. 5. Key factors for comparing continual learning approaches.

After learning Task A, the optimal parameter region for Task A is identified (a "safe zone" in the parameter space). When learning Task B, the new gradient points toward the optimal parameters for Task B. However, if this update would move the parameters out of the "safe zone" for Task A, GEM adjusts the gradient direction to stay within a region that satisfies both tasks [14]. This adjustment can be mathematically represented as a projection, as shown in Eq. (7).

$$g_{adjusted} = g - \frac{g^{\top} m}{m^{\top} m} m \qquad (7)$$

where, g is the gradient for Task B, m is the gradient from memory examples of Task A. And finally, $g_{adjusted}$ is the new gradient that minimizes interference with Task A. Think of it like walking a path (Task B) while staying within a marked boundary (Task A). Instead of walking straight ahead (ignoring the boundary), GEM adjusts your step to ensure you remain within the boundary while moving forward. By modifying the gradient updates this way, the model learns Task B without significantly forgetting Task A, achieving a balance between stability (retaining old knowledge) and plasticity (learning new knowledge). Gradient Episodic Memory (GEM) is a key example of this approach. It adjusts gradients during training to ensure that the loss on stored samples from previous tasks does not increase, preserving past knowledge [14]. Averaged Gradient Episodic Memory (A-GEM) simplifies GEM by using averaged gradients across stored samples, reducing computational overhead while maintaining performance [15]. Orthogonal Gradient Descent (OGD) further refines this by projecting the gradients of new tasks onto spaces orthogonal to gradients of old tasks, minimizing interference [16].

Furthermore, to better understand the position of gradient-based methods, it is essential to compare them with other approaches, including replay-based [17], regularization-based [18], knowledge-distillation-based [19], Bayesian-based [20], architecture-based [21], and hybrid methods [22], across

multiple factors as shown in Fig. 5. One of the most important factors is catastrophic forgetting mitigation, which measures how well an approach retains previously learned knowledge while integrating new information. Replay-based, knowledge-distillation-based, and gradient-based methods are particularly effective in this aspect. Another critical factor is the stability-plasticity tradeoff, which reflects an approach's ability to balance learning new tasks while preserving old ones. Regularization-based and architecture-based methods focus more on stability, while gradient-based and replay-based methods offer a better balance between stability and plasticity. Memory requirements vary significantly among continual learning approaches. Replay-based methods require high memory usage since they store past examples, whereas regularization-based and Bayesian-based methods demand far less memory. Gradient-based approaches generally have moderate memory requirements. Similarly, computational efficiency plays a crucial role, as some methods require extensive processing power.

Regularization-based and Bayesian-based approaches are generally efficient, while replay-based and architecture-based methods tend to have higher computational costs due to data storage or network expansion. Another important consideration is task-free capability, which determines whether a method can learn continuously without predefined task boundaries. Knowledge-distillation-based and gradient-based approaches perform well in this aspect, while regularization-based and Bayesian-based approaches typically struggle with task-free learning. Closely related is flexibility, which measures how well an approach adapts to different continual learning settings, such as class-incremental or domain-incremental learning. Replay-based, knowledge-distillation-based, and gradient-based approaches are highly flexible, whereas regularization-based and architecture-based approaches tend to be more constrained. Sample efficiency is another key factor that determines how well a method can learn from a limited number of training examples. Replay-based and gradient-based methods perform well in this regard, as they either revisit stored data or adjust learning strategies dynamically. Finally, scalability is crucial for applying continual learning to large datasets and real-world applications. Regularization-based, Bayesian-based, and gradient-based methods tend to scale better, whereas architecture-based methods struggle due to high computational costs and network expansion constraints. Each of these factors plays a crucial role in determining the best continual learning approach for a given application. Some methods excel in retaining past knowledge, while others prioritize efficiency or adaptability. The ideal approach often depends on the specific constraints of the task, whether it requires high memory efficiency, task-free learning, or the ability to scale to large datasets.

B. Gradient Episodic Memory (GEM)

Gradient Episodic Memory (GEM) is a continual learning model designed to mitigate catastrophic forgetting. Its core component is an episodic memory M_t, which retains a subset of previously encountered examples from task t. For ease of implementation, integer task descriptors are used to index the episodic memory. Since integer task descriptors do not inherently support strong forward transfer (i.e., zero-shot learning), GEM instead prioritizes minimizing negative backward transfer by efficiently utilizing memory storage. In practice, the learner has a fixed memory capacity of M. If the total number of tasks T, is known in advance, memory can be evenly distributed across tasks, allocating $m = M/T$ slots per task. However, if T is unknown, m can be gradually reduced as new tasks are introduced. A simple strategy for memory management involves storing the most recent m examples from each task, though more sophisticated techniques, such as constructing a coreset per task, could improve efficiency [14]. The model parameters denoted as $\theta \in R^p$, define the predictor f_θ, and the loss function is evaluated on the stored examples from task k, as shown in Eq. (8).

$$L(f_\theta, M_k) = \frac{1}{|M_k|} \sum_{(x_i,k,y_i)\in M_k} L(f_\theta(x_i, k), y_i) \quad (8)$$

The performance of the Gradient Episodic Memory (GEM) model is evaluated using three benchmark datasets (MNIST Permutations, MNIST Rotations, and Incremental CIFAR100), and the results highlight GEM's strong performance compared to state-of-the-art methods. However, there are three key areas for potential improvement. First, GEM does not utilize structured task descriptors, which could facilitate positive forward transfer and enable zero-shot learning. Second, advanced memory management strategies, such as constructing task-specific coresets, were not explored in this study. Third, GEM requires a separate backward pass for each task during training, leading to increased computational overhead. Addressing these limitations presents promising research opportunities for extending learning models [14].

C. Averaged Gradient Episodic Memory (A-GEM)

Although GEM demonstrates strong effectiveness in a single epoch setting, its performance improvements come at the cost of significant computational overhead during training. At each update step, GEM constructs the matrix G using all stored samples from the episodic memory, making the inner loop optimization computationally expensive, particularly when the memory size M and the number of tasks increase. To address this efficiency challenge, a more computationally feasible variant of GEM, known as Averaged GEM (A-GEM) is introduced [15]. Unlike GEM, which ensures that the loss for each previous task—approximated using episodic memory samples—does not increase at each training step, A-GEM instead seeks to maintain a non-increasing average episodic memory loss across all prior tasks. The objective of A-GEM is shown in Eq. (9).

$$minimize_\theta\ l(f_\theta, D_t)\ \ s.t\ \ l(f_\theta, M) \le l(f_\theta^{t-1}, M) \quad (9)$$

The performance of the Averaged Gradient Episodic Memory (A-GEM) model is evaluated using four datasets stream (MNIST Permutations, CUB Split, AWA Split, and CIFAR). The experimental results shows that A-GEM offers the best balance between final average accuracy and computational/memory efficiency. It is approximately 100 times faster and requires 10 times less memory than GEM while outperforming regularization-based approaches in accuracy. Additionally, leveraging compositional task descriptors enhances few-shot learning across all methods, with A-GEM often achieving the best results. However,

experiments reveal a notable performance gap between lifelong learning (LLL) methods, including A-GEM, trained sequentially and the same model trained in a non-sequential multi-task setting, despite exposure to the same data. While task descriptors improve few-shot learning, the limited cross-task transferability among methods suggests that eliminating forgetting alone is insufficient for effective knowledge transfer. Future research will focus on addressing these challenges [15].

D. Orthogonal Gradient Descent (OGD)

Catastrophic forgetting occurs in neural networks when gradient updates for a new task modify the model without preserving knowledge from previous tasks. To address this issue, the Orthogonal Gradient Descent (OGD) method is introduced, which adjusts the update direction to retain crucial information from earlier tasks. The key principle of OGD, as illustrated in Fig. 4, is to constrain parameter updates to remain within the orthogonal subspace of past task gradients, thereby mitigating interference and preserving learned representations [16].

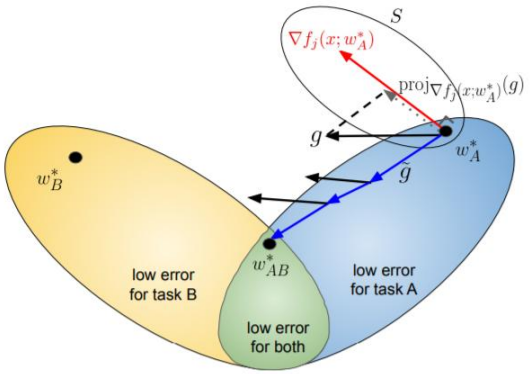

Fig. 6. The Key principle of OGD [16].

An illustration in Fig. 6 demonstrates how Orthogonal Gradient Descent (OGD) adjusts gradient directions to prevent interference between tasks. Here, g represents the original gradient computed for task B, while $\tilde{g}$ is its projection onto the orthogonal subspace relative to the gradient $\nabla f_j(x;\ w_A^*)$ from task A. Constraining updates within this orthogonal subspace (depicted in blue) enables the model parameters to move toward a region of lower error (shown in green) that benefits both tasks [16].

E. Recent Models

As shown in Table II, several types of research have been proposed to improve performance in overcoming and mitigating catastrophic forgetting in continual deep learning using a gradient-based approach.

Utility-based Perturbed Gradient Descent (UPGD), introduced by Elsayed et al. (2023), is an online learning algorithm designed for continual learning agents. It mitigates forgetting by preserving important weights and features while selectively perturbing less critical ones based on their utility. Empirical results demonstrate that UPGD effectively reduces forgetting and maintains network plasticity, allowing modern representation learning techniques to function efficiently in a continual learning setting. This novel approach enables learning agents to operate over extended periods by implementing utility-aware update rules that safeguard essential parameters while adjusting less significant ones. These rules help address key challenges in continual learning, such as catastrophic forgetting and declining plasticity. Experimental evaluations confirm that UPGD enhances network adaptability and facilitates the reuse of previously learned features, making it particularly suited for environments requiring rapid adaptation to evolving tasks [23].

Adversarial Augmentation with Gradient Episodic Memory (Adv-GEM), showed by Wu et al. (2024), enhances data diversity by leveraging gradient episodic memory. This method strengthens existing continual reinforcement learning (RL) algorithms, improving their average performance, reducing catastrophic forgetting, and facilitating forward transfer in robot control tasks. The framework is designed for easy expansion, allowing for further enhancements. Future research will aim to optimize augmentation efficiency, validate the approach across various real-world scenarios, and develop adaptive strategies to handle different task complexities effectively [24].

Asymmetric Gradient Distance (AGD) metric and Maximum Discrepancy Optimization (MaxDO) strategy, proposed by Lyu et al. (2023), are used in Parallel Continual Learning (PCL) effectively to reduce training conflicts and suppresses forgetting of completed tasks. PCL involves training multiple tasks simultaneously with unpredictable start and end times, leading to challenges such as training conflicts and catastrophic forgetting. These issues arise due to discrepancies in the direction and magnitude of gradients from different tasks. To address this, PCL is formulated as a minimum distance optimization problem among gradients, and an Asymmetric Gradient Distance (AGD) metric is introduced to measure gradient discrepancies. AGD accounts for both gradient magnitudes and directions while allowing a tolerance for minor conflicting gradients, thereby mitigating imbalances in parallel training. Additionally, a Maximum Discrepancy Optimization (MaxDO) strategy is proposed to minimize the largest gradient discrepancy across tasks. Extensive experiments on three image recognition datasets demonstrate the effectiveness of this approach in both task-incremental and class-incremental PCL settings [25].

Unified Gradient Projection with Flatter Sharpness for Continual Learning (UniGrad-FS), proposed by Li et al. (2024), enhances CL performance. The core idea is to apply efficient gradient projection in regions with minimal gradient conflicts, making the method widely compatible with gradient-based optimizers. For evaluation, UniGrad and UniGrad-FS are integrated into two state-of-the-art baselines, WA and MEMO, leading to performance improvements of +2.09 per cent and +1.72 per cent, respectively, in a 20-step CIFAR100 benchmark. Further experiments on CIFAR100 and Tiny-ImageNet confirm the method's effectiveness and simplicity across various settings, demonstrating its potential as a general solution for CL [26].

TABLE II GRADIENT BASED MODELS PERFORMANCE

Models	Results						Datasets
	Accuracy (%) / Tasks (T_i)						
	T_1	T_2	T_3	T_4	T_5	Average Accuracy (ACC) $P(t) := \frac{1}{N}\sum_{i=1}^{N} p_i(t)$	
GEM [14]	89%	83%	79%	86%	84%	84%	MNIST Permutations
	88%	89%	82%	80%	74%	82%	MNIST Rotations
	71%	68%	52%	57%	65%	63%	Incremental CIFAR100
A-GEM [15]	99%	97%	93%	90%	87%	93%	Permuted MNIST
	69%	57%	60%	63%	61%	62%	Split CIFAR
Adv-GEM [24]	80%	72%	76%	70%	72%	74%	MW4 (EWC + Adv-GEM)
	90%	85%	92%	87%	86%	88%	MW4 (PackNet + Adv-GEM)
	75%	70%	68%	74%	73%	72%	CW10 (EWC + Adv-GEM)
	90%	88%	91%	87%	89%	89%	CW10 (PackNet + Adv-GEM)
OGD [16]	90%	87%	92%	90%	86%	89%	Permuted MNIST
	91%	82%	79%	73%	63%	77%	Rotated MNIST
	98%	99%	98%	98%	99%	98%	Split MNIST
UPGD [23]	80%	75%	78%	74%	78%	77%	MNIST
	78%	72%	74%	76%	75%	75%	EMNIST
	60%	66%	62%	68%	64%	64%	CIFAR10
GradMA [33]	99%	97%	98%	97%	99%	98%	MNIST
	80%	78%	81%	77%	79%	79%	CIFAR10
	66%	62%	65%	60%	62%	63%	CIFAR100
	52%	50%	45%	55%	48%	50%	Tiny-ImageNet
RWM [30]	93%	92%	93%	94%	95%	93%	CLEAR
TS-ACL [31]	90%	87%	85%	89%	89%	88%	UCI-HAR
	94%	90%	93%	91%	92%	92%	UWave
	99%	97%	98%	99%	97%	98%	DSA
	55%	60%	58%	55%	57%	57%	GRABMyo
	85%	83%	86%	82%	84%	84%	WISDM
SharpSeq (SS) [32]	56%	59%	64%	62%	63%	60%	ACE2005
	62%	61%	62%	61%	60%	61%	MAVEN

Continual Relation Extraction via Sequential Multi-task Learning (CREST), introduced by Le et al. (2024), designed to mitigate catastrophic forgetting in continual relation extraction (CRE) using a customized multi-task learning framework. CREST addresses the challenge of differing gradient magnitudes across objectives, effectively bridging the gap between multi-task learning and continual learning. Extensive experiments on multiple datasets show that CREST significantly enhances CRE performance and outperforms existing state-of-the-art multi-task learning frameworks. These results highlight its potential as a promising solution for continual learning in relation extraction [27].

Continual Flatness (C-Flat) method, proposed by Bian et al. (2025), is designed to balance the trade-off between sensitivity to new tasks and stability in preserving memory, addressing catastrophic forgetting in continual learning (CL). It achieves this by promoting a flatter loss landscape optimized for CL. C-Flat is a plug-and-play approach that can be seamlessly integrated into any CL method with minimal implementation effort [28].

VERSE, proposed by Banerjee et al. (2024), introduces a novel streaming approach that processes each training example only once, requires a single data pass, supports class-incremental learning, and enables real-time evaluation. The method relies on virtual gradients to adapt to new examples while preserving generalization to past data, mitigating catastrophic forgetting. Additionally, an exponential moving average-based semantic memory is incorporated to enhance performance. Experimental results on diverse datasets with temporally correlated observations confirm the method's effectiveness, demonstrating superior performance compared to existing approaches [29].

Radian Weight Modification (RWM), presented by Zhang et al. (2024), a continual learning approach for audio deepfake detection. RWM categorizes classes into two groups: genuine audio, which exhibits compact feature distributions across tasks, and fake audio, which has more dispersed distributions. These distinctions are quantified by using in-class cosine distance, guiding RWM in applying a trainable gradient modification direction tailored for different data types. Experimental comparisons with mainstream continual learning methods demonstrate that RWM excels in both knowledge retention and mitigating forgetting in deepfake detection [30].

TS-ACL, introduced by Fan et al. (2024), is an analytical continual learning framework designed for time series class-incremental pattern recognition, addressing catastrophic forgetting through gradient-free recursive regression learning. This approach not only enhances learning efficiency but also ensures privacy preservation. Experimental evaluations across five benchmark datasets demonstrate that TS-ACL surpasses existing methods, achieving an optimal balance between stability and plasticity. Additionally, it maintains both the non-forgetting and weight-invariant properties, making it a highly robust solution. Its efficiency and minimal computational requirements make TS-ACL particularly well-suited for resource-constrained environments such as edge computing [31].

SharpSeq (SS), proposed by Le et al. (2024), is a novel framework designed to seamlessly integrate state-of-the-art gradient-based multi-objective optimization methods into continual event detection systems. It effectively tackles challenges such as imbalanced training data and the unique constraints of continual learning, leading to significant performance improvements in event detection over time. Comprehensive empirical benchmarks confirm SharpSeq's effectiveness and adaptability, demonstrating its applicability beyond event detection to a wide range of continual learning tasks across various domains. This work establishes a strong foundation for future research, highlighting the potential of multi-objective optimization in advancing continual learning methodologies [32].

GradMA (Gradient-Memory-based Accelerated), presented by Luo et al. (2023), is a method designed to mitigate catastrophic forgetting in federated learning (FL), particularly in scenarios with data heterogeneity and partial participation. It achieves this by simultaneously refining the update directions of both the server and workers. On the worker side, GradMA utilizes the gradients from the previous local model, the centralized model, and the parameter differences between the current local model and the centralized model as constraints in a quadratic programming (QP) formulation, enabling adaptive correction of the local model's update direction. Meanwhile, on the server side, GradMA integrates memorized accumulated gradients from all workers as QP constraints to enhance the centralized model's update direction. Additionally, theoretical convergence analysis is provided under a smooth non-convex setting, and extensive experiments validate the effectiveness of GradMA in reducing forgetting while improving FL performance [33].

IV. Discussion

Through the analysis of gradient-based continual learning approaches, it becomes evident that while these methods offer significant progress toward mitigating catastrophic forgetting, they are not without trade-offs. A recurring challenge is balancing computational efficiency with memory usage, particularly when episodic memory buffers are employed. Moreover, the performance of many models in highly dynamic, non-stationary environments remains inconsistent. In practice, real-world continual learning applications such as autonomous agents, real-time surveillance, and personalized healthcare demand models that are both scalable and resilient to noisy or imbalanced data. The research also highlights that no single approach fully resolves the stability-plasticity dilemma, and that hybrid strategies integrating gradient projection with rehearsal, regularization, or adaptive memory may be necessary. We believe future progress lies in the development of lightweight, task-agnostic architectures that can dynamically adapt while maintaining a strong capacity for long-term retention and generalization.

V. Conclusion

The research presents a comprehensive review of gradient-based approach for mitigating catastrophic forgetting in continual learning. Through an in-depth analysis of key concepts such as continual learning (CL), catastrophic forgetting challenge, and stability and plasticity dilemma. Next, the research highlights the strengths, limitations, and comparative performance of the most common gradient-based models including Gradient Episodic Memory (GEM), Averaged Gradient Episodic Memory (A-GEM), and Orthogonal Gradient Descent (OGD). The findings confirm that gradient-based methods effectively reduce forgetting by strategically adjusting model updates to preserve prior knowledge while integrating new information.

Despite the strong potential of gradient-based approaches in continual learning, they come with notable limitations. First, many of these methods (e.g., GEM, A-GEM, OGD) rely on storing samples from previous tasks, which increases memory requirements and may not be scaled efficiently in memory-constrained environments. Second, their performance may degrade in real-world scenarios where data distributions are non-stationary, unpredictable, or imbalanced. These environments require high robustness, which some gradient-based models currently lack. Third, there is a growing need for novel and hybrid approaches that combine the strengths of gradient projection with adaptive techniques such as attention mechanisms, reinforcement learning, or dynamic memory

allocation to better handle varying task complexities and improve scalability.

Furthermore, despite the progress in continual learning, challenges remain in achieving an optimal balance between stability and plasticity, improving computational efficiency, and enhancing scalability to real-world applications. Future research should explore hybrid approaches that integrate gradient-based learning with replay-based and regularization-based methods, optimize memory utilization, and investigate new architectures that promote long-term knowledge retention without excessive computational costs. By addressing these challenges, continual learning can unlock its full potential, enabling deep learning models to adapt efficiently in dynamic and evolving environments.

REFERENCES

[1] Rudroff, Thorsten, Oona Rainio, and Riku Klén. 2024. "Neuroplasticity Meets Artificial Intelligence: A Hippocampus-Inspired Approach to the Stability–Plasticity Dilemma" Brain Sciences 14, no. 11: 1111. doi: 10.3390/brainsci14111111

[2] Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou et al. "An empirical study of catastrophic forgetting in large language models during continual fine-tuning." arXiv preprint arXiv:2308.08747 (2023).

[3] Z. Wang, E. Yang, L. Shen and H. Huang, "A Comprehensive Survey of Forgetting in Deep Learning Beyond Continual Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 3, pp. 1464-1483, March 2025, doi: 10.1109/TPAMI.2024.3498346.

[4] Menghani, Gaurav. "Efficient deep learning: A survey on making deep learning models smaller, faster, and better." ACM Computing Surveys 55, no. 12 (2023): 1-37.

[5] L. Wang, X. Zhang, H. Su and J. Zhu, "A Comprehensive Survey of Continual Learning: Theory, Method and Application," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362-5383, Aug. 2024, doi: 10.1109/TPAMI.2024.3367329.

[6] Chen, Shijie, Yu Zhang, and Qiang Yang. "Multi-task learning in natural language processing: An overview." ACM Computing Surveys 56, no. 12 (2024): 1-32.

[7] Tian, Yingjie, Xiaoxi Zhao, and Wei Huang. "Meta-learning approaches for learning-to-learn in deep learning: A survey." Neurocomputing 494 (2022): 203-223.

[8] Sahoo, Doyen, Quang Pham, Jing Lu, and Steven CH Hoi. "Online deep learning: Learning deep neural networks on the fly." arXiv preprint arXiv:1711.03705 (2017), doi: 10.48550/arXiv.1711.03705.

[9] Iman, Mohammadreza, Hamid Reza Arabnia, and Khaled Rasheed. "A review of deep transfer learning and recent advancements." Technologies 11, no. 2 (2023): 40.

[10] Aleixo, Everton L., Juan G. Colonna, Marco Cristo, and Everlandio Fernandes. "Catastrophic forgetting in deep learning: a comprehensive taxonomy." arXiv preprint arXiv:2312.10549 (2023). doi:10.48550/arXiv.2312.10549

[11] Kim, Dongwan, and Bohyung Han. "On the stability-plasticity dilemma of class-incremental learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20196-20204. 2023.

[12] Hadsell, Raia, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. "Embracing change: Continual learning in deep neural networks." Trends in cognitive sciences 24, no. 12 (2020): 1028-1040.

[13] Sui, Qingya, Qiong Fu, Yuki Todo, Jun Tang, and Shangce Gao. "Addressing the Stability-Plasticity Dilemma in Continual Learning through Dynamic Training Strategies." In 2024 International Conference on Networking, Sensing and Control (ICNSC), pp. 1-6. IEEE, 2024.

[14] Lopez-Paz, David, and Marc'Aurelio Ranzato. "Gradient episodic memory for continual learning." Advances in neural information processing systems 30 (2017).

[15] Chaudhry, Arslan, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. "Efficient lifelong learning with a-gem." arXiv preprint arXiv:1812.00420 (2018). doi: 10.48550/arXiv.1812.00420

[16] Farajtabar, Mehrdad, Navid Azizan, Alex Mott, and Ang Li. "Orthogonal gradient descent for continual learning." In International conference on artificial intelligence and statistics, pp. 3762-3773. PMLR, 2020.

[17] Rolnick, David, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. "Experience replay for continual learning." Advances in neural information processing systems 32 (2019).

[18] Zhao, Xuyang, Huiyuan Wang, Weiran Huang, and Wei Lin. "A statistical theory of regularization-based continual learning." arXiv preprint arXiv:2406.06213 (2024).

[19] Li, Songze, Tonghua Su, Xuyao Zhang, and Zhongjie Wang. "Continual Learning With Knowledge Distillation: A Survey." IEEE Transactions on Neural Networks and Learning Systems (2024).

[20] Lee, Soochan, Hyeonseong Jeon, Jaehyeon Son, and Gunhee Kim. "Learning to continually learn with the Bayesian principle." arXiv preprint arXiv:2405.18758 (2024).

[21] Rusu, Andrei A., Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick et al. "Progressive neural networks." arXiv preprint arXiv:1606.04671 (2016).

[22] Van de Ven, Gido M., and Andreas S. Tolias. "Three scenarios for continual learning." arXiv preprint arXiv:1904.07734 (2019).

[23] Elsayed, Mohamed, and A. Rupam Mahmood. "Utility-based perturbed gradient descent: An optimizer for continual learning." arXiv preprint arXiv:2302.03281 (2023).

[24] Wu, Sihao, Xingyu Zhao, and Xiaowei Huang. "Data Augmentation for Continual RL via Adversarial Gradient Episodic Memory." arXiv preprint arXiv:2408.13452 (2024).

[25] Lyu, Fan, Qing Sun, Fanhua Shang, Liang Wan, and Wei Feng. "Measuring asymmetric gradient discrepancy in parallel continual learning." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11411-11420. 2023.

[26] Li, Wei, Tao Feng, Hangjie Yuan, Ang Bian, Guodong Du et al. "Unigrad-fs: Unified gradient projection with flatter sharpness for continual learning." IEEE Transactions on Industrial Informatics (2024).

[27] Le, Thanh-Thien, Manh Nguyen, Tung Thanh Nguyen, Linh Ngo Van, and Thien Huu Nguyen. "Continual relation extraction via sequential multi-task learning." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, pp. 18444-18452. 2024.

[28] Bian, Ang, Wei Li, Hangjie Yuan, Mang Wang, Zixiang Zhao et al. "Make continual learning stronger via C-flat." Advances in Neural Information Processing Systems 37 (2025): 7608-7630.

[29] Banerjee, Soumya, Vinay K. Verma, Avideep Mukherjee, Deepak Gupta, Vinay P. Namboodiri et al. "Verse: Virtual-gradient aware streaming lifelong learning with anytime inference." In 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 493-500. IEEE, 2024.

[30] Zhang, Xiaohui, Jiangyan Yi, Chenglong Wang, Chu Yuan Zhang, Siding Zeng et al. "What to remember: Self-adaptive continual learning for audio deepfake detection." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, pp. 19569-19577. 2024.

[31] Fan, Kejia, Jiaxu Li, Songning Lai, Linpu Lv, Anfeng Liu et al. "TS-ACL: A Time Series Analytic Continual Learning Framework for Privacy-Preserving and Class-Incremental Pattern Recognition." arXiv preprint arXiv:2410.15954 (2024).

[32] Le, Thanh-Thien, Viet Dao, Linh Nguyen, Thi-Nhung Nguyen, Linh Ngo et al. "Sharpseq: Empowering continual event detection through sharpness-aware sequential-task learning." In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 3632-3644. 2024.

[33] Luo, Kangyang, Xiang Li, Yunshi Lan, and Ming Gao. "Gradma: A gradient-memory-based accelerated federated learning with alleviated catastrophic forgetting." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3708-3717. 2023.