Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 23–32
DOI: 10.15439/2021F12
ISSN 2300-5963 ACSIS, Vol. 26

State-of-the-Art Techniques in Artificial Intelligence for Continual Learning: A Review

Bukola Salami
School of Computing, University of Eastern Finland, Kuopio, Finland
bukolas@uef.fi

Keijo Haataja
School of Computing, University of Eastern Finland, Kuopio, Finland
keijo.haataja@uef.fi

Pekka Toivanen
School of Computing, University of Eastern Finland, Kuopio, Finland
pekka.toivanen@uef.fi

***Abstract*—Artificial neural networks are used in many state-of-the-art systems for perception, and they thrive at solving classification problems, but they lack the ability to transfer that learning to a new task. Human and animals both have the capability of acquiring knowledge and transfer them continually throughout their lifespan. This term is known as continual learning. Continual learning capabilities are important to ANN in the real world especially with the continuous stream of big data. However, it remains a challenge to be achieved because they are prone to a problem called catastrophic forgetting.. Fixing this problem is critical, so that ANN incrementally learn and improve when deployed to real life situations. In this paper, we did a taxonomy of continual learning in human by introducing plasticity-stability dilemma, hence the Hebbian plasticity and compensatory homeostatic plasticity process of learning and memory formation that occurs in the brain. We also did a state-of-the-art review of three different approaches to continual learning to mitigate catastrophic forgetting.**

***Index Terms*—Artificial Intelligence; Continual Learning; Catastrophic Forgetting; Artificial Neural Networks; Stability–Plasticity Dilemma**

I. Introduction

LEARNING continually has always been the grand goal of any Artificial Intelligent (AI) systems functioning in the real-world scenario because AI systems can be continuous exposed to streams of data and so are required to remember existing tasks when modelled on new stream of data. This has recently attracted much attention in the AI community, especially related to Artificial Neural Networks (ANNs) [1]. Humans and animals have an exceptional ability to learn large number of different skills and tasks but also to select the ones which are useful and relevant without negatively interfering with each other and at the same time being able to recall information when needed on such tasks that were previously learned [2, 3]. The ability to do this is called Continual Learning and can also be referred to as Lifelong Learning or Incremental Learning [1, 2]. AI agents should demonstrate a capability for continual learning [4]. The goal is to gather knowledge across tasks, particularly through model sharing and possibly having only one model that can perform well on all the learned tasks [5]. However, the existing standard for model deployment has a critical flaw: data are dynamic, and this continues to change [1, 6].

With remarkable successes accomplish over the past few years in AI, deep network applications are however restricted to sole, distinct problem. Where every single network has to be trained and re-trained from the beginning every single time a new task is fed into the network and as a result their training remains very challenging to deal with particularly in real-world settings and in situations where data are scarce and/or computation is costly [7]. Furthermore, the sequence of tasks may not be clearly labelled tasks and they may switch randomly, leading to an individual task recuring in long time intervals. Therefore, the main challenge of an AI agent to learn continually is being susceptible to catastrophic forgetting or catastrophic interference [4].

This well-known phenomenon was first recognized by McCloskey and Cohen in [8]. Catastrophic forgetting always leads to a degrade generalization performance or in the worst case, a complete loss of information on an older task that was previously performed because it was simply re-trained on the new task or dataset sequentially [1, 2]. It specifically happens when a network is trained in sequence on several tasks because the weights that are imperative for task A are modified to incorporate the goal of task B, and as a result of these changes to the network, the accuracy on task A can severely reduce after some training updates on task B [4, 9]. This is conceivably, one of the main gap between modern ANNs design and biological neural networks because of the complexity of synapses [10].

To overcome the lack of continual learning in ANNs, recently three main strategies have been proposed: Progressive/Architectural Strategy, Rehearsal Strategy, and Regularization Strategy [2, 5].

A lifelong learning architecture capable of continual learning could guide the field of AI into a period of extraordinary performance, generality, and integration. This architecture could also prevent the need for costly data collection, labelling and retraining that sets constrains on today's state-of-the-art computer systems [9]. In essence, to overcome catastrophic forgetting, an AI system should display the capability to gain new knowledge and simultaneously improve the network on existing tasks based on the continuous stream of data, thereby, preventing significant existing knowledge from being forgotten [1, 2]. This is known as the stability-plasticity dilemma. Plasticity is the ability to integrate new knowledge, while stability is preserving existing knowledge while new stream of data is processed. Although, a model will not be able to gain new knowledge from new training data if they are too stable. Likewise, a model with abundant plasticity can suffer from a great weight change and forget a previously learned task. [11, 12]. One of the effective approaches to plasticity was the one addresses by Stephen Grossberg, articulated in 1980 on the solution to the stability–plasticity dilemma which states that "a system must remain plastic enough to learn important and new information, while also maintaining stability in its memories for information that has already acquired" [9]. Such adaptation and memory formation are what can be observed in biological neurosystems. Humans have remarkable ability to preserve old knowledge and skills learned and it is mainly reliant on how often they are recollected and used. Tasks that are practiced and performed regularly, tend to be unforgettable, unlike the ones that are

This work is supported and funded by Digital Innovation HUB of Northern Savo Region – DigiCenterNS Kuopio, Finland

so old and frequently not used. Strangely, this adaptation and memory formation sometimes happens with little or no form of supervision whatsoever. This process, at the fundamental level, according to Hebbian theory, is the consolidation of neurons connected to synapses, that performs together at the same time, compared to neurons with unrelated performance behavior [5].

Most of the algorithms presented in this paper are based on the current state and advancements in both neurophysiology and computational neuroscience field that are capable of continual learning in AI. [7, 9]

In this paper, we review some of the major works in continual learning both in advanced animals such as humans, whales, dogs, dolphins etc. and AI agent: we focus on how humans and animals acquire new knowledge and memories and at the same time been able to retain the useful ones over time. We also discussed several proposed algorithms for continual learning systems to overcome catastrophic forgetting. The rest of this paper is organized as follows: Section 2 reviews continual learning in humans and animal. Section 3 also reviews few continual learning strategies and algorithms proposed in the last 4 years. section 3.1 introduced the fundamental of continual learning, its desiderata, and the three different strategies. Section 3.2 reviewed some common algorithms proposed for continual learning with their respective mathematical equations and in table 1, a summary of different recent algorithms to continual learning is given. In section 4 we proposed our novel research idea for forgetting in ANN, and in section 5 the conclusion of the paper .

II. Continual Learning In Advanced Animals

New skills and knowledge can easily be acquired and transferred across domains in advanced animals to complete tasks, while artificial neural systems are still in the early stages regarding transfer learning, which is prone to catastrophic forgetting [1]. Likewise, humans and animals can learn in a continual way, but it has been somewhat challenging for an AI system to do the same [5].

Evidence found recently suggests that the human and animal brain can avoid forgetting by shielding previously learnt knowledge and skills in the neocortical circuits. The brain significantly benefits from the integration of multisensory information, which provide the means for an effective communication. Furthermore, in conditions of sensory hesitation with respect to the predominant tendency to train ANNs on uni-sensory information, such as audio or visual information [1]. For example, when a mouse learns a new task/skill, a part of its excitatory synapse is reinforced, and this leads to an increase in the capacity of individual dendritic spines of the mouse brain neurons [5]. Afterward, these increased dendritic spines persevere in spite of learning some other skills alongside the old one, and it results to retention of such skill after a few months later. When some of these dendritic spines are selected and cleared up, the matching skill is forgotten. This gives a fundamental evidence that neural mechanisms for supporting the protection of these synapses are important to retention of task performance. The results obtained with the mouse experiment alongside with some other neurobiological models suggested that continual learning in the neocortex depends on task-specific synaptic consolidation, by which knowledge is strongly encoded by reducing the plasticity of synapses that are vital to previously learned tasks and therefore stable over a long timeframe [13].

The principal core idea is that learning is associated with persistent and experience-driven changes to the brain, as given with the mouse example, that help them in the effective performance of vital tasks, such as the acquisition of necessities like food and shelter while avoiding the unpleasantness that accompanies injury or predation [14]. This is the inspiration behind autonomous embodied agents research on multisensory features for early development and sensorimotor specialization in human brain [1].

A. Stability–Plasticity Dilemma

The human brain experiences neural plastic changes across its lifespan both in healthy conditions and also after brain lesions. The process where the brain adapts to environmental challenges and disease is referred to as plasticity [15, 16]. This process was first demonstrated by neuroanatomist Michele Vincenzo Malacarne in 1783 when he intensively trained one in each pair of two birds and two dogs from the same clutch of eggs and litter respectively [15]. The external environment surrounding animals can be considered as static for a short period of time but will become dynamic over a long time. Animals essentially learn quickly about new stimuli to adapt to such environments when it changes, so also, the plasticity that occurs at the neural pathways and continuously changes with respect to internal and external stimuli [17, 18].

Plasticity is an important part for neural malleability at the cells and circuits level in the brain. Neural plasticity can serve multiple functions, such as been *homeostatic* in nature for excitement within a network, it could also be *mnemonic* to form the basis of the memory and lastly, been *metaplasticity* [18].

One important form of plasticity is indicated across sensory modalities, however, a large part of the human brain neurons are present at birth, therefore plasticity and associated learning are expected to occur early in life. The brain needs to be plastic enough to acquire new knowledge and memories but stable enough to retain them over time. This balance is known as the plasticity-stability dilemma [19, 20]. Humans have amazing ability to adapt by efficiently gaining new skills, transforming them to new experiences, and recalling and transferring them across several areas where they are needed. It is also true that humans have the capacity to forget gradually some previously learnt information at some point when they get older. Therefore, learning of new information rarely affect consolidated knowledge in human [1, 21]. Stability-plasticity dilemma is the degree whereby a system must be inclined to integrate and learn novel skills and, most of all, how these learning processes can be rewarded by internal mechanisms which stabilize and modulate neural activity just to avert catastrophic forgetting [1]. Artificial neural networks gain their principal structure by sensorimotor experiences, from the imitation of human brain which is mainly plastic during the crucial phase of early development [16]. Sensorimotor skill learning, like any other form of learning, happens through the general mechanism of experience-dependent synaptic plasticity. When new skill is learned via general training, synapses in the brain are revised to form a lasting motor memory of that skill learned [22].

Stability-Plasticity positioned at multiple brain areas are regulated by the mechanisms of neurosynaptic plasticity. Neurosynaptic plasticity mechanisms is such that it protects knowledge about previously learned tasks from forgetting, by decreasing the rates of synaptic plasticity. However, there are two types of plasticity needed for a stable continual process: Hebbian Plasticity [23] and Compensatory Homeostatic Plasticity [24]. When used together, both Hebbian learning and Compensatory Homeostatic Plasticity stabilize neural cells to shape the optimal patterns of experience-driven connectivity, integration, and functionality in a network [1, 16, 24].

Neurosynaptic plasticity is an important attribute in the brain because it produces physical changes in the neural structure and allows us to learn, remember, and adapt to any changing environments [16] as well as activity-dependent synaptic plasticity in learning and memory formation. Synaptic plasticity was first discovered in the hippocampus of the human brain in the early 1970s. It was concluded that an increase in the strength of the synaptic input of the stimulated connections only is produced by repeated, near-synchronous activation of both pre- and post-synaptic neurons [14] and this process is known as Long-Term Potentiation (LTP). These characteristics of synaptic plasticity suggests its role in learning new skills as well as being an information storage device [25].

However, memories may not be properly stabilized if synapses are easily bendable and in such state of perpetual flux, old learning can easily be overwritten by new learning. Hence, for any learning system, there is essentially constraint between the competing requirements of stability and plasticity [22].

B. The Hebbian Synaptic Plasticity

The brain can adapt to a changing environment and as well as providing important insights into the shape of cortex's connectivity and function. It has been shown that while fundamental designs of connectivity in the visual system are noticeable at early development, normal visual input is essential for the accurate development of the visual cortex [26]. Donald Hebb in 1949 was the first to propose the theory describing and explaining the mechanisms of synaptic plasticity in the adaptation of neurons to external stimuli. Hebb postulated that the connection between two neurons is strengthened, when one neuron pilots the activity of another neuron [27]. In the following years, Hebb's idea has been interpreted to the weight changes among nodes of a single layer perceptron in ANNs based on coincidence or the product of pre- and postsynaptic activity mimicked from the brain neurons, thereby altering the connection of neurons into changes relative to the coactivity of the input and output nodes in ANNs [14]. Thus, considering Hebb's theory from an ANN's standpoint, after a network has been trained using backpropagation successfully, the synapses between neurons that synchronous fires a given input are made stronger for as long as it takes, to maintain and improve its outputs [27, 28].

A simple formula for Hebbian plasticity considers a change in the synaptic weight w and it is updated as the product of the activities in pre-synaptic x and post-synaptic y with learning rate η is given as [1]:

$$\Delta w = x \cdot y \cdot \eta \tag{1}$$

Yet, Hebbian plasticity is unstable while alone, but depended on and requires compensatory mechanisms to stabilize its learning process. This is attainable by enhancing Hebbian plasticity with some constraints like upper limits on specific synaptic weights or regular neural activity, which can only be done by homeostatic plasticity [29, 27]. Homeostasis plasticity is also referred to as a compensatory process that stabilizes the neural firing rates in the brain [24]

III. Overcoming Catastrophic Forgetting With Continual Learning Algorithm

Catastrophic forgetting problem can occur in different ways. One way is between mini-batches when using stochastic gradient descent methods during the general training processes. Another way is the degradation of the generalization performance of a network [12, 30]. Similar to the continual learning methods, in Stochastic Gradient Descent (SGD) optimization, every mini-batch can be thought of as a mini-task offered sequentially to the network. In this context, the interest is describing the changes in the learning of the neural networks by analysing examples of forgetting events [14]. This happens when task that have been learned and correctly classified at some time t in the optimization process are afterward misclassified at a time $t' > t$ [31]. It should also be noted that catastrophic forgetting occurs to ANN models including SOMs as well as Deep Neural Networks, for example Transfer Learning in DNN [32].

Typically, the current approaches to overcome catastrophic forgetting in ANN have concurrently made data available from tasks during training. By passing in data from several tasks while training and learning, forgetting is prevented. This is attributed to the fact that the weights of the network can be mutually optimized for high performance on all training tasks. This case is frequently referred to as the multitask learning, and a good example can be seen in reinforcement learning method where a successfully trained single agent can be used to play many Atari games effectively. If data are introduced to the network sequentially, multitask learning can only be used if tasks are recorded by an episodic memory system and replayed during training to the network [19, 33]. However, this method can be impractical when dealing and learning a large number of tasks, as large number of memories would be required to stored and replayed, likewise been related to number of tasks [4, 14, 29].

A. Continual Learning Basics

Continual Learning is the basic step towards AI, because it permits an intelligent agent to continuously adapt to changes that occur in data and tasks. Nevertheless, there are some consequences during learning for both supervised and unsupervised learning. For example, when data are not properly represented or there is a mistake in the input distribution, a model can overfits the recently seen data, which is something continual learning systems aim to address [34, 35].

A series of desiderata are used to defined Continual Learning in practice which includes Firstly, online learning meaning that learning can occur at every moment, with no permanent tasks or datasets and with no clear boundaries/restrictions between tasks. Secondly,

forward/backward transfer of model from existing tasks to new tasks with the possibility of the new task improving the performance of older tasks. Furthermore, resistance to catastrophic forgetting, that is, new learning task does not degrade the performance on previous data, and lastly, there should be no direct access to previous tasks but be able to retain it [34, 35].

An infinite sequence of data is considered for a general continual learning setting, where at each timestep t the network accepts a new data $\{x_t, y_t\}$ to draw a non independent and identically distributed. from an existing distribution P that could by itself experience some rapid or gradual changes. The key goal is to learn a function F parameterized by θ that can minimize a predefined loss $\mathscr{L}$ on the new data without interfering on existing tasks and also with the possibility of improving on the tasks that were learned previously [34]:

$$\theta^t = \underset{\theta,\xi}{\operatorname{argmin}}\, \mathcal{L}(F(x_t;\theta), y_t) + \sum \xi_i \qquad (2)$$

Such that:

$$\mathcal{L}(F(x_t;\theta), y_t) \leq \mathcal{L}(F(x_t;\theta^{t-1}), y_t) + \xi_i, \qquad (3)$$

$$\xi_i \geq 0\,;\ \forall i \in [0..t-1] \qquad (4)$$

Where x_t is the input, y_t is the ouput and $\xi = \{\xi_i\}$ is the slack variable that allows some constraints to be violated like small increase in loss from previous tasks [34].

Some strategies have been designed for continual learning, which are: Firstly, the progressive/architectural strategy. Architectural strategy can be used to incrementally builds a network's structure for every single task being processed. In addition, it also tries to copy and re-use as much as possible the attributes of the previous model in the process. The second strategy is known as rehearsal methods, since it keeps a memory of data analyzed on previous tasks and continues to retrain the network on this memory to maintain its performance. And the third approach is regularization. Regularization strategy tries to re-use a single neural network, which is by including a few regularization penalties to alleviate the behaviour of the network with respect to previous tasks [1, 20]. Usually, rehearsal and progressive strategies, performs very well but always declines as the number of tasks increase, and might require a high computational power. With some differences from the first two approaches, the implementation of regularization strategy is quite simple, they require little memory, but it performance might not be up to that of rehearsal methods [7]. One main problem encountered when applying regularization strategy is determining what task best represents the behaviour of the network and, this can lead to the form of regularization penalty that would be taken [7].

Recently, a lot of attention has been shifted to the idea of using regularization function to fit the existing task for learning a new task in a network. This method can be understood as an approximation of sequential Bayesian. Some distinctive examples of this regularization approach include the elastic weight consolidation [4] and learning without forgetting [21].

B. A Review of Some Popular Continual Learning Strategies

Several algorithms have been proposed so far to mitigate catastrophic forgetting in neural networks and few are reviewed in this paper:

[4] proposed an algorithm that performs operation like synaptic consolidation used on the brain on ANNs by constraining some important parameters to stay close to their old values. This algorithm is known as Elastic Weight Consolidation (EWC).

In EWC, the performance in task A is protects by constraining its parameters to stay in a region of low error just for task A to be positioned mainly around θ_A^*. This constraint is implemented as a quadratic penalty and can exist as a spring anchoring the parameters to the previous solution, hence been called elastic. However, all parameters should not have the same stiffness of this spring, but it must be larger for parameters that are very much affected by the performance in task A.

To further explain the optimal choice of constraint and weights, the neural network training is considered from a probabilistic viewpoint using Bayes' rule and also noting that the log probability of the data $\mathcal{D}$ given the parameters log p $(|\theta)$ from the Bayes' rule equation is simply the negative of the loss function $-\mathscr{L}(\theta)$ [4]:

$$\log p(\theta\,|\,\mathcal{D}) = \log p(\mathcal{D}_B\,|\,\theta) + \log p(\theta\,|\,\mathcal{D}_A) - \log p(\mathcal{D}_B). \qquad (5)$$

The key to implement EWC is that all the information about task A, must have been accepted into the posterior distribution $p(\theta|\mathcal{D}_A)$. The true posterior probability is inflexible, so the posterior distribution was approximated as a gaussian distribution with average specified by parameters θ_A^* and a diagonal precision specified by the diagonal of the Fisher information matrix F. This matrix F is used because it has three key characteristics: Firstly, it is equivalent to the second derivative of the loss near a minimum. Secondly, it can be computed from first-order derivation alone and it is quite easy to compute even for big models. Thirdly, a positive semidefinite is guaranteed. Therefore, the loss $\mathscr{L}$ minimized in EWC is computed as:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \qquad (6)$$

where $\mathcal{L}_B(\theta)$ is the loss on task B only, λ determines how important the existing task is compared with the new task, and i gives labels to each parameter. However, when considering a third task C, the EWC algorithm might try to maintain the network parameters value close to the learned parameters of previous tasks A and B. This can be imposed either with two separate penalties or as one by observing that the sum of two quadratic penalties is itself a quadratic penalty [4]. Because computing over the diagonal of fisher requires summation of all possible outputs, thus EWC has complexity linear in the number of outputs, limiting its application to low-dimensional output spaces [10].

A simple structural regularizer that can be computed online was introduced by [10] and also implemented locally at each synapse/parameter (weights and biases). The authors developed an algorithm which can keep track of an *importance measure* ω_k^{μ}.

Considering the change in loss function $\mathscr{L}$ for an infinitesimal parameter update $\delta(t)$ at time t, where $\theta(t)$ is the trajectory in parameter space between task A and task B, and g is the gradient can be written as [10]:

$$\mathcal{L}(\theta(t) + \delta(t)) - \mathcal{L}(\theta(t) \approx \sum_k g_k(t)\delta_k(t) \quad (7)$$

However, to calculate the change in the loss over the whole trajectory, all the infinitesimal, and the changes are summed over, which amount to computing the path integral of the gradient vector from the start time t_0 to the end time t_1 and also the loss between the end and the start point $\mathscr{L}(\theta(t_1)) - \mathscr{L}(\theta(t_0)$ [10]):

$$\int_{t^{\mu-1}}^{t^{\mu}} g\big(\theta(t)\big) \cdot \theta'(t)dt = \sum_k \int_{t^{\mu-1}}^{t^{\mu}} g_k(\theta(t))\theta_k'(t)dt \equiv -\sum_k \omega_k^{\mu} \quad (8)$$

The authors tried to solve the problem of minimizing the total loss function summated on all tasks, $\mathscr{L} = \sum_\mu \mathcal{L}\mu$, with no contact to the loss function $\mathscr{L}\mu$ of the past training except the new task μ at any given time but with this minimization come catastrophic forgetting which led to a drastic weight changes between the old task and the new task (ν < μ) while training task μ. To avoid this problem, they introduced quadratic surrogate loss which approximates the summed loss function of old task $\mathscr{L}\nu$ (ν < μ). The implication of using the quadratic surrogate loss for training instead of the actual loss function, is that the final parameters will remain the same and change in loss during the training process [10]:

$$\tilde{\mathcal{L}}_\mu = \mathcal{L}\mu + c\sum_k \Omega_k^{\mu}(\tilde{\theta}_k - \theta_k)^2 \quad (9)$$

Where c is the dimensionless strength parameter, $\tilde{\theta}_k$ is the reference weight at the end of previous task, and Ω_k^{μ} is the per-parameter regularization strength. The equation 9 can only achieve two tasks.

Although [10] algorithm is similar to EWC in [4] in that more importance synapses are strongly directed towards the reference weight, however, the method computes the importance measure online including all the learning trajectory [10], considering that, EWC is about the point estimate of the diagonal of the Fisher information matrix at the final synapse values, that has to be calculated during a separate stage at the end of each task [4].

Inspired by Hebbian learning in neuroplasticity, [5] proposed Memory Aware Synapses (MASes). Unlike previous proposed research on synapses, their continual learning method can learn using unlabelled data and in online manner. The sensitivity of the output function was the main focus and not the loss while estimating importance weights for the network parameters. After the model has been trained on the approximation F of the true function $\bar{F}$, the function F output was preserved, and its sensitivity was measured for changes. A small perturbation δ in the parameters θ results in a change in the function output that can be approximated by

$$F(x_k;\theta + \delta) - F(x_k;\theta) \approx \sum_{i,j} g_{ij}(x_k)\delta_{ij} \quad (10)$$

Where $g_{ij}(x_k)$ is the gradient of the function learned and δ_{ij} is the change in parameter θ_{ij}. But the goal is to preserve the prediction of F. To do this, the gradients of all data point were accumulated to obtain importance weight Ω_{ij} [5]

$$\Omega_{ij} = \frac{1}{N}\sum_{k=1}^{N} \|g_{ij}(x_k)\| \quad (11)$$

Where N is the total number of data points. However, when function F is multi-dimensional, the gradients for each output can be computed by using the squared l_2 norm the learned function output. The importance is measured by the sensitivity of the squared l_2 norm over learned function output. To learn a new task, a new loss $\mathcal{L}_n(\theta)$, and a regularizer for penalty to change important parameters (high Ω_{ij})

$$\mathcal{L}(\theta) = \mathcal{L}_n(\theta) + \lambda\sum_{i,j} \Omega_{ij}\big(\theta_{ij} - \theta_{ij}^*\big)^2 \quad (12)$$

Where λ is the regularizer's hyperparameter and θ_{ij}^* is the previous network parameter [5].

[36] explicitly address the diagonal assumption made by EWC algorithm in [4]. They assumed that if the Fisher Information Matrix is not diagonal, EWC might fail to stop the network from drifting from "good parameter space". The proposed Rotated Elastic Weight Consolidation approach is based on rotating the parameter space of a network, that is, re-parameterization of the parameter space θ, in a way that the output of the forward pass is not changed, while the computed Fisher Information Matric from the gradients during the backpropagation is approximately diagonal [4]. To obtain reparameterization, the rotation matrix is computed using Singular Value Decomposition (SVD) even though computing SVD on a very large matrices is quite expensive. Applying chain rule on FIM and computing using SVD, the following equation was obtained where θ' is W' the new rotated weight matrix [36]:

$$E_{x\sim\pi}[XX^T] = U_1 S_1 V_1^T \quad (13)$$

$$E_{\substack{x\sim\pi \\ y\sim p}}\left[\left(\frac{\partial L}{\partial y}\right)\left(\frac{\partial L}{\partial y}\right)^T\right] = U_2 S_2 V_2^T \quad (14)$$

$$W' = U_2^T W U_2^T \quad (15)$$

The results obtained with rotated EWC is outstandingly more real at overcoming catastrophic forgetting in sequential task learning problems [36].

Variational continual learning [37] and Continual Learning with Adaptive Weights (CLAWs) [38] are another regularization strategy [37]. In [37], Bayesian inference provides a fundamental framework for continual learning with its algorithm where the posterior of the model parameters is learned and updated continually from a sequence of datasets. To achieve this algorithm, online Variational Inference (VI) was merged with Monte Carlo VI for neural networks to produce Variational Continual Learning (VCL). In addition, VCL was enhanced to contain a small episodic memory by the combination VI with the coreset data summarization process. The coreset can be compared to an episodic memory that holds important information from previous tasks, where the algorithm can go

back so as to refresh its memory of these important information.

Similar to [37], in [38], their approach is based on probabilistic modelling and variational inference [37]. But rather than strictly dividing the architecture into shared and task-specific parts, the approach adapts the contributions of each neuron using Gaussian distribution for the adaptation as the probabilistic model and afterwards the adaptation parameters are included within the variational parameter in Monte Carlo VI.

[39] is another form of regularization approach to continual learning. The method presents the option to control the stability and compactness of the learned task. This makes this method also agreeable for network compression applications and online learning. They proposed a task-based hard attention mechanism that can preserves learning from an existing task without affecting the learning of a new task. As well as learning tasks with the binary attention. A task can also be learned over gated task embeddings, using backpropagation and minibatch SGD. Some attributes of hard attention task are: It can store, as well as maintain a lightweight structure. Secondly, the task is learned instead of a heuristic approach or rule-driven. Thirdly, the mask is not necessarily binary, and this might be useful if the weights need to be re-used for learning other tasks, i.e., to overcome catastrophic forgetting.

[21] proposes Learning without Forgetting method which compose of Convolutional Neural Networks (CNN) and this approach can be perceived as combination of Distillation Networks (transfer of information from a large to a small model) [40] and fine-tuning. The main idea here is only used on new task data for training the network. The network learns from parameters that works fine on old task and uses this information to train the new tasks without the use of data from previous tasks.

To achieve this, the responses y_o on each new task object from the original network for outputs on the old tasks (defined by shared parameters θ_s and task-specific θ_o) were recorded, then the network was trained for the loss to be minimize for all tasks and regularization R by using SGD. To define the loss for a new task, the output $\hat{y}_n$ was merged with the one-hot ground truth y_n [21]:

$$L_{new}(y_n, \hat{y}_n) = -y_n \cdot log\ \hat{y}_n \tag{16}$$

To transfer the known, knowledge distillation loss must be introduced to the network [21]

$$L_{new}(y_n, \hat{y}_n) = -H(y_o', \hat{y}_o') \tag{17}$$
$$= -\sum_{i=1} y_o'^{(i)} \log \hat{y}_o'^{(i)}$$

[41] provided an architectural strategy algorithm called Reinforced Continual Learning (RCL). It comprises of three networks: *value network*, *controller*, and *task network*. The controller is executed as a Long Short-Term Memory network (LSTM) or as Recurrent Neural network to generate policies and determine how many filters/nodes will be added to each task. The value network was designed as a multilayer perceptrons/fully-connected network, that approximates the value of the state [41]:

$$\pi(a_{1:m}|s;\theta_c) = \prod_{i=1}^{m} p_{t,i,a_i'} \tag{18}$$

Where θ_c the controller network's parameter.

However, the task network, on the other hand, can be any network of interest for solving any task, for example object detection or image classification. Furthermore, RCL adaptively expands the network when a new task arrives, while using stochastic gradient descent with η as the learning rate [41]:

$$\min_{W_t/W_{t-1}^a} L_t(W_t; D_t) \tag{19}$$

$$W_t/W_{t-1}^a \leftarrow W_t/W_{t-1}^a - \eta \nabla W_t/W_{t-1}^a L_t \tag{20}$$

[42] propose Incremental Moment Matching (IMM) framework from Bayesian Neural networks, Here moments of posterior distribution which are trained on old and new task are matched together in an incremental way using Gaussian distribution. Considering that the objective is to determine the ideal parameter $\mu_{1:k}^*$ and $\Sigma_{1:k}^*$ of the gaussian approximation function $q_{1:k}$ from the posterior parameter of the kth task (μ_k, Σ_k), two different moment match method can be used: mean-IMM and mode-IMM. Mean-IMM finds the average of the parameters of two networks for both old and new task [42]:

$$\mu_{1:k}^*, \Sigma_{1:k}^* = \underset{\mu_{1:k}/\Sigma_{1:k}}{\text{argmin}} \sum_k^K \alpha_k\, KL(q_k \| q_{1:k}) \tag{21}$$

Mode-IMM is an alternative form of mean-IMM. It merges the parameter of old and new network using Laplacian approximation of the posterior of gaussian distribution [42]:

$$\log q_{1:k} \approx \sum_k^K \alpha_k \log q_k + C = -\frac{1}{2}\theta^T \left(\sum_k^K \alpha_k\, \Sigma_k^{-1}\right)\theta + \left(\sum_k^K \alpha_k\, \Sigma_k^{-1}\mu_k\right)\theta + C' \tag{22}$$

The result obtained from the experimental with both IMM on shows that Mode-IMM performed better than mean-IMM and other comparative models in the various dataset. The limitation is that IMM performance decreases with more complex dataset.

[43] introduced a model architecture called Progressive Neural Network (PNN) that support transfer of knowledge across sequence of tasks particularly in reinforcement learning. Progressive network makes use of transfer learning by retaining a pool of knowledge through training of an agent from a previous task, and also having the ability to transfer that knowledge to another agent to improve convergence speed. After PNN finishes training of a previous task, its parameter θ' is frozen when switching to the second task, after which another parameter θ is instantiated [43]:

$$h_i^k = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)}\right) \tag{23}$$

Where $W_i^{(k)}$ is weight matrix, $U_i^{(k:j)}$ in the lateral connection, and f is an element-wise non-linearity.

PNN is robust to harmful features learned in incompatible tasks by the RL agent. A major downside of PNN is the growth in number of parameters with the number of tasks.

TABLE 1: SUMMARY OF SOME OTHER DIFFERENT APPROACHES TO ALLEVIATE CATASTROPHIC FORGETTING

Authors, Year, and Country:	Proposed Methods/Algorithms and Strategies/Approaches:	Important Note:	Limitation
[4] 2018, United State of America	Elastic Weight Consolidation (EWC), Regularization	The EWC uses only one network with static network capacity and nominal computational overhead which has a low computational cost.	EWC are very sensitive to the diagonal approximation of the FIM used in practice because of the large size of a full FIM and costly to compute weights for regularization penalty
[10] 2017, Australia	Synaptic Intelligence using Quadratic Surrogate Loss SI, Regularization	The method computes the per-synapse consolidation strength in an online manner and over an entire learning trajectory in parameter space and individual synapses act as higher dimensional dynamical systems.	SI can only learn importance weights during training, which leads to lack of adaptation to some particular subset.
[5] 2018, Germany	Memory Aware Synapses (MAS), Regularization	The important parameters (high Ω_{ij}) can be reused, through model sharing, which is only possible with a penalty when changing the parameters.	It is limited by brittleness caused by representation drift mostly common to regularization methods.
[36] 2018, Italy	Rotated Elastic Weight Consolidation, Regularization	The evaluation of the experiment on various learning tasks shows that the approach performed well compared to the standard EWC.	Rotated EWC can also suffer from brittleness caused by representation drift.
[39] 2018, Sweden	Hard Attention to Task (HAT), Regularization	HAT presents the option to monitor the used network capacity throughout different tasks and layers and it has only two hyperparameters, and are both referred to as the stability and compactness of the learned task.	HAT gradually declines in classification accuracy during training with no signs and hope of ever increasing
[21] 2016, Netherlands	Learning without Forgetting, Regularization	The method is only proposed for convolutional neural networks. It is a hybrid of knowledge distillation and fine-tuning.	Additional memory and computation are needed in LFL to compare activations
[44] 2020, Italy	Embedding Regularization for continual learning, Regularization	ER develops an efficient way to regularize the behaviour of the network by acting on its internal embeddings, i.e., the activations of one or more layers closer to the exit.	In ER, when the memory grows, the required training time also increases.
[45] 2020, Germany	Bayesian Neural Networks for Non-Stationary Data, Regularization	It makes use of Bayesian forgetting and a Gaussian diffusion process for adaptation to non-stationary data, leading to a better predictive performance	Bayesian neural networks with a uni-modal approximate posterior often find poor local minima if the dataset is small and models are complex, which is especially challenging in situation where data are streamed
[46] 2018, Canada	FearNet, Architectural	The basolateral amygdala is used to determine which memory system to use for recalling task and it is more memory efficient.	FearNet can suffer from recall when the number of classes to learn is high.
[41] 2019, Germany	Reinforced Continual Learning (RCL), Architectural	RCL explores the best neural network architecture for each upcoming task.	The training time of RCL is particularly important and high for large networks with more layers
[47] 2020, France	Move-to-Data: Incremental learning approach, Architectural	This approach does not require gradient based optimization	The Move-to-Data method is limited to only one fully connected layers
[33] 2017,	Gradient Episodic Memory (GEM),	The advanced memory management was not investigated, and the iteration	It may be less scalable and require many observations and complex generative model to

United State of America	Regularization, Rehearsal	requires one backward pass per task, which increases the computational time.	represent realistic tasks, and the Effective prioritized replay remains an unsolved problem
[48] 2019, United States of America	Average Gradient Episodic Memory (A-GEM), Regularization, Rehearsal	It is about 100 times faster and memory is 10 times less required; compared to regularization-based approaches, it achieved a significantly high average accuracy	The model is plausibly a little incremental over GEM
[49] 2017, United State of America	Incremental Classifier and Representation Learning (iCaRL), Regularization, Rehearsal	It comprises of 3 major components: a nearest-mean-of-exemplars, a herding to prioritize exemplars, and a representation learning step, and It learns strong classifiers and data representation at the same time	iCaRL's performance is still lower than what other systems achieve when trained in a batch setting,
[50] 2020, Virtual Conference	Functional Regularised Continual Learning (FRCL): Gaussian processes, Regularization, Rehearsal	When viewed from the regularisation perspective, it regularises the functional outputs of the neural network, while when viewed from a rehearsal method perspective, a principled way is provided for compressing data from previous task, by optimizing the selection of inducing points.	It suffers from a fixed memory buffer in which case the summaries of all the previous seen tasks need to be compressed one needs into a single summary.
[38] 2020, United State of America	Continual Learning with Adaptive Weights (CLAW), Regularization, Rehearsal and Architectural	It is based on variational inference from VCL.	This approach did not actually compare their result to other to VCL, every other.
[37] 2018, Canada	Variational Continual Learning, Regularization, Rehearsal and Architectural	VCL is most suitable for efficient model fine-tuning in sequential decision-making problems, and can be applied to generative model and discriminative model.	VCL also suffers from brittleness caused by representation drift

IV. OUR NOVEL RESEARCH PROPOSAL

According to [51], the human brain act as information filters. From the inward region of the brain (hippocampus), when new information is taken in, old irrelevant information is filtered out and the updated information are stored for long term retrieval and decision making. The unused pieces are however deleted to create space. It is called forgetting in neuroscience. Forgetting occurs when the synaptic connection between neurons weakens and are eliminated over time [51]. To effectively adapt, humans need to strategically forget, so also the need to forget in ANN for a successful continual learning.

The previously discussed works have approached the problem of catastrophic forgetting, specifically the continual learning with valuable strategies and algorithms. Most of works, tackled ways to achieve continual learning, but left out the aspect of forgetting. Forgetting some older knowledge is essential to accommodate information from new data. The novel idea here is to build a network to deploy learning and the same network will be re-purposed to learn a new task, forgetting some specific information that is irrelevant. Self-Organizing Map (SOM) will be used for this purpose of forgetting. The algorithm will be in such a way that the network learns and update in the opposite direction which will lead to forgetting in the SOM. In addition, we will measure 3 different performance metrics, which are: The Average Accuracy, The Backward Transfer, The Cumulative Backward Transfer Scores of Forgetting.

Forgetting can be beneficial in some cases: 1) it prevents overfitting to specific features and can improve generalization, 2) forgetting outdated data can enhance flexibility of decision made from learning with new data. [51]The main goal of this novel research is to control the forgetting process during learning to protect some vital information and in the process minimizing accuracy loss.

V. CONCLUSION AND FUTURE WORK

Continual Learning is the fundamental step towards AI, because it permits an intelligent agent to continuously adapt to a dynamic environment, a distinctive characteristic of natural intelligence. The goal for continual learning is to acquire knowledge across tasks particularly through model sharing and having a single model that can perform well on all the tasks, however, there is one challenge to achieving this, which is catastrophic forgetting of previous task learned, in the process of learning new task. In this paper, we presented continual learning in advanced biological animals and Artificial intelligent agents. We discussed plasticity-plasticity dilemma and taking it a little further, we talked about Hebbian plasticity and compensatory homeostatic plasticity process of learning and memory formation that occurs in the brain. Despite significant advancement, most of the currently proposed algorithms for continual learning are still far from providing a robust,

flexible, and scalable approach displayed in biological animals. However, we presented a state-of-the-art overview of several algorithms, from the most popular and recent literature on continual learning, where some significant progress has been made to tackle catastrophic forgetting in ANN. On top, we used a table to summarize these algorithms which included: the type of strategy/approach, the dataset used for the performance evaluation and some key notes about these algorithms. In addition, we introduced our novel research proposal on intentional forgetting, which is such that an intelligent system will chose to forget some irrelevant or old information when learning a new task. Evaluating with different dataset, we will measure different performance metrics with the new proposed algorithm.

REFERENCES

[1] German, P., Ronald, K., Jose, P., Christopher, K., & Stefan, W. (2019). Continual lifelong learning with neural networks: A review. *ScienceDirect- Neural Networks, 113*, 54-71. doi:10.1016/j.neunet.2019.01.012

[2] Z. Chen and B. Liu. (2018). *Continual Learning and Catastrophic Forgetting.* Morgan & Claypool Publishers. doi:10.2200/S00832ED1V01Y201802AIM037

[3] Nicolas, M., Gregory, G., & David, F. (2019). Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences, 115*(44), E10467-E10475. doi:10.1073/pnas.1803839115

[4] Kirkpatrick, J., Pascanu, R., Rabinowitza, N., Veness, J., Desjardins, G., Rusu, A. A., . . . Hadsell, R. (2018). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America, 114*(13), 3521–3526. doi:10.1073/pnas.1611835114

[5] Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory Aware Synapses: Learning what (not) to forget. *15th European Conference on Computer Vision ECCV'18*. doi:10.1007/978-3-030-01219-9_9

[6] Vincenzo, L., Davide, M., & Lorenzo, P. (2019). Fine-Grained Continual Learning. *Cornell University: Arxiv.org*, 1-12. Retrieved from https://arxiv.org/abs/1907.03799

[7] Pomponi, J., Scardapane, S., Lomonaco, V., & Uncini, A. (2020). Efficient Continual Learning in Neural Networks with Embedding Regularization. *ScienceDirect - NeuroComputing, 297*, 139-148. doi:10.1016/j.neucom.2020.01.093

[8] Michael, M., & Neal, C. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *ScienceDirect- The Psychology of Learning and Motivation, 24*, 109-165. doi:10.1016/S0079-7421(08)60536-8

[9] Andrew, P., Ryan, C., Patrick, M., Stephen, B., Renee, E., & MarioAguilar-Simon. (2019). Uncertainty-based modulation for lifelong learning. *ScienceDirect - Neural Networks, 120*, 129-142. doi:10.1016/j.neunet.2019.09.011

[10] Zenke, F., Poole, B., & Ganguli, S. (2017). Continual Learning Through Synaptic Intelligence. *Proceedings of the 34 th International Conference on Machine Learning, PMLR 70, 70*, pp. 3987–3995. Sydney, Australia. Doi:10.5555/3305890.3306093

[11] De, L. M., Rahaf, A., Marc, M., Sarah, P., Xu, J., Ales, L., . . . Tinne, T. (2019). Continual learning: A comparative study on how to defy forgetting in classification tasks. *Cornell University: arxiv.org*, 26. doi:10.1109/TPAMI.2021.3057446

[12] Heechul, J., Jeongwoo, J., Minju, J., & Junmo, K. (2016). Less-forgetting Learning in Deep Neural Networks. *IEEE*, 1-5. Retrieved from arXiv:1607.00122

[13] M.Stark, S., & E.L.Stark, C. (2016). *Chapter 67 - Introduction to Memory.* Academic Press. doi:10.1016/B978-0-12-407794-2.00067-5

[14] Magee, J. C., & Grienberger, C. (2020). Synaptic Plasticity Forms and Functions. *Annual Review of Neuroscience, 43*, 95-117. doi:10.1146/annurev-neuro-090919-022842

[15] Quentin, R., Awosika, O., & Leonardo, G. C. (2019). Plasticity and recovery of function. *ScienceDirect: Handbook of Clinical Neurology, 163*, 473-483. doi:10.1016/B978-0-12-804281-6.00025-2.

[16] Wickliffe, C. A., & Robins, A. (2005). Memory retention – the synaptic stability versus plasticity dilemma. *ScienceDirect*. doi:10.1016/j.tins.2004.12.003

[17] Junichiro, H., Junichiro, Y., & Shin, I. (2006). Balancing Plasticity and Stability of On-Line Learning Based on Hierarchical Bayesian Adaptation of Forgetting Factors. *ScienceDirect- NeuroComputing, 69*(16-18), 1954-1961. doi:10.1016/j.neucom.2005.11.020

[18] Sehgal, M., Song, C., L.Ehlers, V., & R.MoyerJr., J. (2013). Learning to learn – Intrinsic plasticity as a metaplasticity mechanism for memory formation. *Neurobiology of Learning and Memory, 105*, 186-199. doi:10.1016/j.nlm.2013.07.008

[19] Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., & Ranzato, M. (2019). On Tiny Episodic Memories in Continual Learning. *Cornell University*, 1-15. Retrieved from arXiv:1902.10486

[20] Lomonaco, V. (2019). *Continual Learning with Deep Architectures.* Bologna: Department of Computer Science and Engineering, University of Bologna.

[21] Li, Z., & Hoiem, D. (2016). Learning without Forgetting. *The 14th European Conference on Computer Vision ECCV2016.* doi:10.1109/TPAMI.2017.2773081

[22] Ajemiana, R., D'Ausilio, A., Moorman, H., & Bizzi, E. (2013). A theory for how sensorimotor skills are learned and retained in noisy and nonstationary neural circuits. *Proceeding of the National Academy of Sciences of the United States of America*, 5078-5087. doi:10.1073/pnas.1320116110

[23] D.O., Hebbs. (1949). *The organization of behavior; a neuropsychological theory.* Psychology Press. doi:10.1007/978-3-642-70911-1_15

[24] Zenke, F., & Gerstner, W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. *Philosophical Transactions of The Royal Society B Biological Sciences, 372*(1715). doi:10.1098/rstb.2016.0259

[25] Martin, S. J., Grimwood, P. D., & Morris, R. G. (2000). Synaptic Plasticity and Memory: An Evaluation of the Hypothesis. *Annual Review of Neuroscience, 23*, 649-711. doi:10.1146/annurev.neuro.23.1.649

[26] Nicolas Y. Masse, Gregory D. Grant, and David J. Freedman. (2019). Alleviating Catastrophic Forgetting using Context-Dependent Gating and Synaptic Stabilization. *Cornell University - arxiv.org*. doi:10.1073/pnas.1803839115

[27] Steven J.Cooper. (2005). Donald O. Hebb's synapse and learning rule: a history and commentary. *Neuroscience and Biobehavioral Reviews, 28*, 851-874. doi:10.1016/j.neubiorev.2004.09.009

[28] Abraham, W. C., Jones, O. D., & Glanzman, D. L. (2019). Is plasticity of synapses the mechanism of long-term memory storage? *Nature Partner Journal- Science of Learning, 4*, 9. doi:10.1038/s41539-019-0048-y

[29] German, P., Ronald, K., Jose, P., Christopher, K., & Stefan, W. (2019). Continual lifelong learning with neural networks: A review. *ScienceDirect- Neural Networks, 113*, 54-71. doi:10.1016/j.neunet.2019.01.012

[30] Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., & Zhang, B.-T. (2017). Overcoming Catastrophic Forgetting by Incremental Moment Matching. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, Califonia, USA. doi:10.5555/3294996.3295218

[31] Toneva, M., Sordoni, A., Tachet, d. C., Trischler, A., Bengio, Y., & Geoffrey, J. G. (2019). An Empirical Study of Example Forgetting During Deep Neural Network Learning. *The International Conference on Learning Representations (ICLR) 2019*. Retrieved from arXiv:1812.05159

[32] Fiona M Richardson, Michael S C Thomas(2008). Critical periods and catastrophic interference effects in the development of self-organizing feature maps. *Developmental Science*, 371–389. doi:10.1111/j.1467-7687.2008.00682.x

[33] Lopez-Paz, D., & Ranzato, M. (2016). Gradient Episodic Memory for Continual Learning. *Facebook Artificial Intelligence Research*, 1-17. doi:10.5555/3295222.3295393

[34] Rahaf, A. (2019). *Continual Learning in Neural Networks.* Leuvan, Belgium: KU Leuven – Faculty of Engineering Science:. Retrieved from arXiv:1910.02718v2

[35] Pascanu, R., Teh, Y., Pickett, M., & Ring, M. (2018). Continual Learning. *Conference on Neural Information Processing Systems.* Montréal, Canada: NeurIPS.

[36] Liu, X., Masana, M., Herranz, L., Weijer, J. V., Lopez, A. M., & Bagdanov, A. D. (2018). Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting. *International Conference on Pattern Recognition'18.* doi:10.1109/ICPR.2018.8545895

[37] Nguyen, C. V., Li, Y., Bui, T. D., & Turner, R. E. (2018). Variational Continual Learning. *International Conference on Learning Representations (ICLR).* doi:10.17863/CAM.35471

[38] Adel, T., Zhao, H., & Turner, R. E. (2020). Continual Learning with Adaptive Weights. *The International Conference on Learning Representations (ICLR).* Retrieved from https://openreview.net/forum?id=Hklso24Kwr

[39] Serrà, J., Surís, D., Miron, M., & Karatzoglou, A. (2018). Overcoming Catastrophic Forgetting with Hard Attention to the Task. *International Conference on Machine Learning (ICML 2018).* Retrieved from arXiv:1801.01423

[40] Hinton, G., Vinyals, O., & Dean, J. (2014). Distilling the Knowledge in a Neural Network. *NIPS 2014 Deep Learning Workshop: Neural and Evolutionary Computing.* Retrieved from https://arxiv.org/abs/1503.02531

[41] Ju, X., & Zhanxing, Z. (2019). Reinforced Continual Learning. *Cornell University*, 1-10. doi:10.5555/3326943.3327027

[42] Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., & Zhang, B.-T. (2017). Overcoming Catastrophic Forgetting by Incremental Moment Matching. *31st Conference on Neural Information Processing Systems (NIPS 2017).* Long Beach, Califonia, USA. doi:10.5555/3294996.3295218

[43] Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., . . . Hadsell, R. (2016). Progressive Neural Network. *Google DeepMind: arXiv:1606.04671*, 1-14. Retrieved from https://arxiv.org/abs/1606.04671

[44] Jary, P., Simone, S., Vincenzo, L., & Aurelio, U. (2020). Efficient Continual Learning in Neural Networks with Embedding Regularization. *ScienceDirect- Neurocomputing, 397*, 139-148. doi:10.1016/j.neucom.2020.01.093

[45] Richard, K., Botond, C., Alexej, K., der, S. P., & Stephan, G. (2020). Continual Learning with Bayesian Neural Networks for Non-Stationary Data. *International Conference on Learning Representations.* Virtual Conference. Retrieved from https://arxiv.org/abs/1910.04112

[46] Kemker, R., & Kanan, C. (2018). FearNet: Brain-Inspired Model for Incremental Learning. *The Sixth International Conference on Learning Representations.* Vancouver, Canada. Retrieved from https://arxiv.org/abs/1711.10563

[47] Miltiadis, P., Jenny, B.-P., Akka, Z., Boris, M., & de, R. A. (2020). Move to-Data: A new Continual Learning approach with Deep CNNs, Application for image-class recognition. *hal-02865878v1f.* Retrieved from https://arxiv.org/abs/2006.07152

[48] Arslan, C., Marc'Aurelio, R., Marcus, R., & Mohamed, E. (2019). Efficient Lifelong Learning with A-GEM. *International Conference on Learning Representations (ICLR).* New Orleans. Retrieved from https://arxiv.org/abs/1812.00420

[49] Rebuffi, S.-A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). iCaRL: Incremental Classifier and Representation Learning. *Conference on Computer Vision and Pattern Recognition.* Honolulu, Hawaii. doi:iCaRL: Incremental Classifier and Representation Learning

[50] Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, Yee Whye Teh(2020). Functional Regularisation for Continual Learning with Gaussian Processes. *International Conference on Learning Representations.* Virtual Conference. Retrieved from https://arxiv.org/abs/1901.11356

[51] Richards, B. A., & Frankland, P. W. (2017). The Persistence and Transience of Memory. *Cell Press journal, 94*(6), 1071-1084. doi:10.1016/j.neuron.2017.04.037