Does Alignment help continual learning?

Anurag Daram* ANURAG.DARAM@UTSA.EDU
Dhireesha Kudithipudi
University of Texas at San Antonio

Editors: Manuel Grana, Pawel Ksieniewicz, Leandro Minku, Pawel Zyblewski

Abstract

Backpropagation relies on instantaneous weight transport and global updates, thus questioning its neural plausibility. Continual learning mechanisms that are largely biologically inspired employ backpropagation as the baseline training rule. In this work, the role of learning rules that avoid the weight transport problem in the context of continual learning is examined. We investigate weight estimation approaches that use linear combinations of local and non-local regularization primitives for alignment-based learning. These approaches are coupled with parameter regularization and replay mechanisms to demonstrate robust continual learning capabilities. We show that the layer-wise operations observed in alignment-based learning help to boost performance in complex task-aware and task-free scenarios on multiple image classification datasets. We study the dynamics of representational similarity from the learning rules and provide its mapping to the knowledge preservation capabilities of the models.

Keywords: Life-long learning, Imbalanced data

1. Introduction

Continual learning (CL) represents an area of study that aims to enable AI models to successfully interact and learn in real-world environments. The features include, but are not limited to, the acquisition of new knowledge and skills while keeping prior learned concepts, transfer knowledge across tasks, learn from exposure to few examples and noise tolerance. CL models focus on addressing a fundamental trade-off: the *stability-plasticity dilemma*, whereby a model that emphasizes stability tends to suffer from poor forward transfer and adaptation to new tasks, whereas one that is too plastic is unable to retain previously learned information, a phenomenon commonly known as catastrophic forgetting or interference (McCloskey and Cohen, 1989). Addressing catastrophic forgetting is an important focus of research in deep learning, with several methods proposed to help mitigate forgetting (De Lange et al., 2021; Zenke et al., 2017; Kirkpatrick et al., 2017a). Biological brains seem to have solved this dilemma, being able to learn continuously throughout their lifetime while adapting to novel scenarios and environments. Taking inspiration from neuroscience, researchers have proposed several types of mechanisms to address catastrophic forgetting (Kudithipudi et al., 2022). These can broadly be classified into (i) loss or functional parametric adjustments, i.e. regularization (Kirkpatrick et al., 2017b; Zenke et al., 2017), (ii) dynamic architectures, including neurogenesis (Ebrahimi et al., 2020; Rusu et al., 2016), and (iii) rehearsal or replay of previous experiences (Aljundi et al., 2019; van de Ven et al., 2020).

* Corresponding author.

Several of the biologically inspired mechanisms still employ backpropagation as the baseline learning algorithm, which is the workhorse of modern deep learning. However, any exact implementation of backpropagation requires the instantaneous transport of weights for propagating errors, making it inherently non-local. This scheme is biologically non-plausible since there are no known neural mechanisms for instantaneously coupling synaptic weights (Grossberg, 1987). To alleviate this popularly known "weight transport" problem, researchers have investigated several feedback alignment approaches (Sanfiz and Akrout, 2021) to approximate backpropagation. Furthermore, it has been demonstrated that these alignment techniques can boost the models' resistance to noise and adversarial attacks (Akrout, 2019). Despite showing performance comparable to backpropagation and improving robustness, these learning mechanisms have never been explored in the context of continual learning. Moreover, a number of continual learning mechanism implementations combine backpropagation – a non-local process – with local learning rules (Ostapenko et al., 2021; Arani et al., 2022). The learning dynamics become unstable as a result, since the global loss function overrides changes imposed by the local rules at different layers (Choromanska et al., 2018). This leads us to the question: *Can learning rules created as a substrate of combining different regularization primitives provide a better alternative to backpropagation for continual learning?* Henceforth, we explore simple local and non-local primitives with decoupled alignment-based learning for better representational invariance and robustness in continual learning scenarios.

In this work, we investigate and compare the performance of alignment-based methods with backpropagation for continual learning scenarios. First, we untie the weights into forward and backward components, which are optimized using a loss function that is proportional to the global cost function and the linear sum of local and non-local primitives. Further, we couple this loss with weight- and neuron-based regularization functions and replay-based mechanisms to apply them to continual learning scenarios. Secondly, we investigate how these local and non-local rules formed with regularization primitives enable robust continual learning. Finally, we analyze the efficacy of these algorithms through the lens of semantic similarity across layer-wise activations while learning.

Through the proposed framework, we observe a consistent improvement in CL performance across multiple benchmarks for the weight and neuron-based regularization models combined with alignment-based learning rules. It is important to note that this correlation in performance is not specific to one learning mechanism, but depending on the criteria of importance measurement, respective improvements are demonstrated. However, for replay-based mechanisms, both the local and non-local learning rules fail to demonstrate an improvement in performance. However, we observe that alignment-based learning rules offer improved robustness to noisy updates in comparison to backpropagation across all the continual learning mechanisms. Furthermore, we use the Centered Kernel Alignment (CKA) representational variance metric to better understand the optimization dynamics of these biologically plausible alternatives. We show that these credit assignment strategies are competitive with backpropagation and illustrate generalizability across multiple continual learning models.

2. Related Work

Local and Non-local learning mechanisms. Human brain demonstrates remarkable capability to learn continuously through its many complex plasticity mechanisms (Kudithipudi et al., 2022). There has been evidence of multiple plasticity mechanisms (Abraham and Bear, 1996; Watt and Desai, 2010) and learning systems in the brain (Mcnaughton and Morris, 1987; Ellefsen et al., 2015). These mechanisms are crucial in enabling our learning capabilities. Researchers investigated the role of such plasticity mechanisms for enhancing learning capabilities (Abraham, 2008; Clune, 2019). The work in (Allred and Roy, 2020) demonstrate the use of local learning rules in unsupervised manner to mitigate catastrophic forgetting. Mechanisms such as metaplasticity (plasticity of plasticity) (Abraham, 2008) in synapses were shown to retain previously learned knowledge and enable continual learning capabilities. Moreover, homeostatic mechanisms have shown to boost the computational performance of spiking neural networks (Maass et al., 2002). Additionally, the use of local learning rules with neuromodulation (Miconi et al., 2020; Daram and Yanguas-Gil, 2020; Yanguas-Gil et al., 2019) enabled the networks to learn rapidly in a resource- and energy-efficient manner.

Previous studies have demonstrated the efficacy of local learning rules and computations in enhancing learning capabilities. However, local learning rules are very sensitive to the selected hyperparameters, and are thus very selective within the constrained parametric space. Only a small subset of the explored configurations produce favorable gresults (Baldi and Sadowski, 2016). Since they are limited to a local context, in lifelong learning scenarios, they struggle with adaptation to new tasks with lower similarity to previous ones. Weight estimation approaches, on the other hand have shown to address the performance bottleneck (Frenkel et al., 2021) incurred by purely local learning mechanisms.

Weight estimation techniques. Backpropagation enables multi-layered networks to learn complex tasks. However, the implementation of backpropagation requires instantaneous weight transport (using the same set of weights to propagate the forward activations and the gradients), which is non-local in nature (Grossberg, 1987). Moreover, this non-locality leads to significant increase in movement of weights while training (Rojas and Rojas, 1996), which is not suitable for solutions requiring learning on the edge. Researchers have introduced approaches such as feedback alignment (Lillicrap et al., 2016) and weight mirror (Akrout et al., 2019) that seek to approximate backpropagation while circumventing the weight transport problem. A challenge with these techniques, lies in the requirement of a complex hyperparameter tuning regime to achieve good performance during learning (Bartunov et al., 2018). The work in (Kunin et al., 2020; Kornblith et al., 2019) addresses this issue by formulating a more general framework of credit assignment strategies without weight symmetry that are more robust and scalable while performing on-par with backpropagation.

These novel credit assignment approaches have never been explored in the context of continual learning. Additionally, these weight estimation techniques use local and non-local primitives with strong geometric interpretations that could be linearly combined to form complex credit assignment strategies. The strategies are defined by layer-wise regularization functions, which attempt to align based on the symmetry of representations, weights, or activations. When we look at regularization mechanisms for continual learning, they identify

importance and penalize based on either the activations, tracking changes in weights or observing the shift in representations or loss while shifting between tasks. This creates possibilities for the mapping of regularization-based mechanisms for continual learning with alignment-based techniques.

Continual Learning mechanisms. Among the broad desiderata of continual learning (Kudithipudi et al., 2022, 2023), several approaches have focused on addressing catastrophic forgetting. Among these, regularization-based and replay-based are the most common ones. Regularization mechanisms (Kirkpatrick et al., 2017c; Zenke et al., 2017; Schug. Simon, 2020) select important parameters and penalize their updates as a way to retain task-specific knowledge through the lifetime. Replay methods (Chaudhry et al., 2019; Aljundi et al., 2019) on the other hand, either store or generate previously seen experiences and interleave them with the current task to alleviate forgetting while enabling positive transfer of knowledge. Albeit a few approaches (Soures et al., July 2021; Madireddy et al., 2023), most of these models employ backpropagation as the central learning rule and for propagating gradients along the networks. In this work, we select multiple of these mechanisms, couple them with other strategies in place of backpropagation and highlight potential benefits and limitations.

3. Methodology

Continual Learning Problem Formulation. We define the continual learning problem $\mathscr{F}$ as the ability to learn tasks sequentially, while optimizing the performance across the entire distribution of tasks $\mathbb{D}$. The neural network $f_\theta(x)$ with learnable parameters θ attempts to solve this problem without suffering severe performance loss on previously learned tasks while learning new ones. Formally, we consider the distribution of tasks $\mathbb{D} = \{\mathbb{T}^1, \mathbb{T}^2 ... \mathbb{T}^N\}$ for $N \in \mathbb{Z}^+$, wherein each task $\mathbb{T}^k$ is a set $(\mathcal{X}_k, \mathcal{Y}_k)$ of ordered pairs of input data points and their corresponding class labels. Performance on the problem $\mathscr{F}$ is evaluated by first training the network on the tasks $\{\mathbb{T}^i\}$, $i \in [1, k]$ sequentially, then measuring the mean performance $\Psi(\mathbb{D}) \equiv \mu(\{\psi(\mathbb{T}^1), \psi(\mathbb{T}^2), ..., \psi(\mathbb{T}^k)\})$, where $\psi(\mathbb{T}^k)$ represents the performance on task $\mathbb{T}^k$. The objective is to learn a mapping $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ that maximizes the performance across all the samples seen throughout the lifetime.

Learning framework with regularization primitives

In the proposed framework for continual learning, we consider a credit assignment strategy that is parameterized by two sets of weight parameters (θ), namely the forward weights θ_f and backward weights θ_b. The net loss $\mathcal{L}$ in the model is defined by the sum of a global loss function $\mathcal{G}$ for θ_f and a local layer-wise regularization function $\mathcal{P}$ for θ_b,

$$\mathcal{L}(\theta_f, \theta_b) = \mathcal{G}(\theta_f) + \mathcal{P}(\theta_b) \tag{1}$$

As shown in Eq. 1, every training step consists of two sub-parts: i) updating the forward weights using the error signal transported via the backward weights (similar to feedback alignment), and ii) updating the backward weights as a function of the primitives. These primitives can be combined to derive different local and non-local learning rules that impact the dynamics of the backward weights. These primitives, for any layer l are derived as a

function of the backward weights $B_l \in \theta_b$, the forward weights $W_l \in \theta_f$, forward activation of the layer a_l and the forward activations of the next layer a_{l+1}. In this scenario, the forward weights are outgoing while the backward weights are incoming to a layer. Based on the parameters being used, the primitives define the locality of the learning rules. We follow the definitions proposed in (Whittington and Bogacz, 2017) to determine the locality of a primitive, wherein a) *local computation* is defined as those which involve synaptic weights acting on their associated inputs and b) *local plasticity* implies that modifications to weights depend only on the pre- and post-synaptic activations.

There are multiple primitives which can be both local and non-local in nature, as introduced in (Akrout et al., 2019; Kunin et al., 2020).

- *Decay primitive* (P_l^{decay}) - The decay primitive focuses on penalizing the euclidean norm of the backward weights (B_l). The primitive is represented as $\frac{1}{2}||B_l||^2$. As a result, while taking the derivative of this primitive with respect to the backward weights gives B_l.

- *Amp Primitive* (P_l^{amp}) - The amp primitive encourages the alignment of the pre-synaptic activations (a_l) with the backward reconstruction ($B_l a_{l+1}$). The primitive is represented by the negative trace of inner product of the activations, with the reconstruction as $-tr(a_l^T B_l a_{l+1})$. The update of the backward weight using this primitive is Hebbian in nature, similar to the rule presented in weight mirror (Akrout et al., 2019).

- *Null primitive* (P_l^{null}) - The null primitive applies a Euclidean norm penalty on the backward reconstruction of the pre-synaptic activations. The null primitive is represented by $\frac{1}{2}||B_l a_{l+1}||^2$, thereby adding sparsity to the layer wise activity. The derivative sums to $B_l a_{l+1} a_{l+1}^T$ thereby increasing the separability of the activations by imposing a quadratic penalty of activations.

- *Sparse primitive* (P_l^{sparse}) - The sparse primitive penalizes the Euclidean norm of the alternative construction of the post-synaptic activations by taking a product of the pre-synaptic activations with the backward weights ($\frac{1}{2}||a_l B_l||^2$). Unlike the aforementioned primitives, this one is non-local in nature since the connections from $l+1$ layer to its predecessor cannot act on a_l.

- *Self primitive* (P_l^{self}) - The self primitive aligns the forward and backward weights by directly promoting their inner product ($-tr(B_l W_l)$). The update directly subtracts the forward weights ($-W_l$) from the backward weights, thereby attempting to increase the orthogonality with respect to each other. This primitive is also non-local in nature since the assumption is that the gradient of backward weights can directly access the forward weights.

We can derive different learning and credit assignment strategies through a linear combination of the aforementioned primitives. When considering these primitives in the context of regularization-based continual learning mechanisms, we can observe that alignments of either activations or weights can form strong correlations with the importance measurement and penalizing techniques.

Alignment based local and non-local learning

The linear combination of the primitives addresses instability issues with singular update mechanisms such as Hebbian update observed in P_l^{amp}, or the direct application of forward weight penalty in P_l^{self}. Therefore, these primitives are combined to form a set of local and non-local learning rules that can be used as alternatives to backpropagation. Moreover, these combinations include credit assignment techniques that encompass both local and global updates in the network, which can benefit continual learning. Although multiple combinations of these primitives are possible, three of those combinations which focus either on aligning activations, weights or representations work effectively in bridging the performance gap with backpropagation.

Information Alignment (IA) : This is a local learning rule defined by the combination of three local primitives.

$$\mathcal{P}_{IA} = \sum_{l \in \text{layers}} \left(\beta_1(- \text{tr}(a_l^T B_l a_{l+1})) + \beta_2(\frac{1}{2}||B_l||^2) + \beta_3(\frac{1}{2}||B_l a_{l+1}||^2) \right) \tag{2}$$

where $\beta_1, \beta_2, \beta_3$ represent tunable parameters. The gradient of $\mathcal{P}_{IA}$ is proportional to the gradient of a quadratically regularized linear autoencoder (Kunin et al., 2019). The autoencoder can be represented by $\frac{1}{2}||a_l - B_l W_l a_l||^2 + \frac{\beta_2}{2}(||W_l||^2 + ||B_l||^2)$, and the gradient with respect to B_l, attempts to achieve symmetry of the encoder and the decoder between the forward and the backward construction of the pre-synaptic activation a_l for a given layer l. The information alignment rule is local in nature and captures layer-specific features without high instability, but in deeper networks, it is unable to capture downstream dependencies leading to a drop in performance on complex tasks. However, in a continual learning context, the layer specific features can be compartmentalized to improve separability of task-specific representations, which could help improve performance in gating or sparsity based approaches (Masse et al., 2018).

Symmetric Alignment (SA) : It is defined by a combination of the *self* and *decay* primitives.

$$\mathcal{P}_{SA} = \sum_{l \in layers} \left(\alpha_1(\frac{1}{2}||B_l||^2) + \alpha_2(- tr(B_l W_l)) \right) \tag{3}$$

The gradient of $\mathcal{P}_{SA}$ is proportional to the gradient with respect to B_l of $\frac{1}{2}||W_l - B_l||$ and thereby encourages the symmetry of weights. Despite being non-local in nature, symmetric alignment still mitigates the need for instantaneous weight transport. It effectively optimizes the framework of decoupled forward-backward weight updates, wherein the backward weights are eventually encouraged to become the transpose of the forward counterparts. Moreover, aligning based on the weights can be critical in identification of importance parameters for several weight regularization-based methods for continual learning.

Activation Alignment (AA) : This rule is defined by the sum of *amp* and *sparse* primitives.

$$\mathcal{P}_{AA} = \sum_{l \in layers} \left(\gamma_1(- tr(a_l^T B_l a_{l+1})) + \gamma_2(\frac{1}{2}||a_l B_l||^2) \right), \tag{4}$$

where γ_1 and γ_2 are tunable parameters. When $\gamma_1 = \gamma_2$, and $a_{l+1} = W_l a_l$, then the gradient of $\mathcal{P}_{AA}$ is proportional to the gradient with respect to B_l of $\frac{1}{2}||W_l a_l - B_l^T a_l||^2$. This update mechanism encourages the alignment of the activations, i.e. the post-synaptic forward activations of layer l with the pre-synaptic backward activations. This function, similar to SA is non-local in nature and explicitly imposes the backward weights to align to the transpose of forward weights. However, the optimization based on the activations can play an important role in increasing the separability of important and unimportant neurons for neuron-activation based regularization techniques.

These learning mechanisms operate on a principle of "weight estimation" avoiding instantaneous weight transport required in backpropagation. Additionally, these weight estimation techniques are observed to be more robust to noisy updates in comparison to approaches using backpropagation (Guerguiev et al., 2019). The next section covers how these mechanisms can be incorporated in the context of continual learning and how the primitives in presence of noise can help in achieving robust continual learning.

Continual Learning strategies with alignment

The weight estimation alignment techniques presented above add a two-factor loss to the model (shown in Eq. 1, with the global loss modifying the forward weights and the layer-specific loss updating the backward weights. In the context of continual learning, with a focus on regularization-based techniques, an additional loss is added to the model to regularize synaptic updates based on task-specific information. The net loss across the entire distribution of tasks $\mathbb{D}$ can thereby be modelled by

$$\mathcal{L}_{\mathbb{D}}(\theta) = \sum_{t=1}^{N-1} \left(\mathcal{L}(\theta_f^t, \theta_b^t) + \mathcal{L}_{reg}(\theta^t)\right) \tag{5}$$

With alignment-based learning mechanisms, the regularization-based schemes for continual learning can be applied to either θ_f, θ_b, or even both. In this manuscript, we evaluate these rules as the base learning mechanism for multiple weight and neuron regularization based approaches. Moreover, we employ these learning mechanisms with replay-based approaches as a test for generalizability.

Regularization mechanisms. We evaluate these rules as the base learning method for multiple weight and neuron-based regularization mechanisms for continual learning. We select multiple regularization mechanisms, namely 1) online Elastic Weight Consolidation (oEWC) (Schwarz et al., 2018) that applies a quadratic penalty term for each previously learned task, whereby each task's term penalizes the parameters for how different they are compared to their value directly after finishing the training on that task; 2) Synaptic Intelligence (SI) (Zenke et al., 2017) that consists of only one quadratic term that penalizes changes to important parameters which are identified by tracking each synapse's credit assignment during the task. The importance parameter is measured by computing the per parameter contribution to the change of loss for the current task and thus strongly contributing parameters are heavily penalized in subsequent tasks; 3) Learning without Forgetting (LwF) (Li, 2017) is a distillation-based approach towards continual learning, wherein previous model outputs are used as soft labels for previous tasks; 4) Stochastic

Synapses (SS) (Schug. Simon, 2020) is a regularization approach that implements Bernoulli transmission probabilities and magnitudes with synapses to measure their importance and applies a regularization factor based on the probability factor; 5) Uncertainty-regularized continual learning (UCL) (Jung et al., 2020) is a neuron importance based regularization mechanisms. The importance is measured by means of an 'uncertainty' factor that is computed from the variability during training of each neuron's incoming weights. The idea is that weights that are important for a task tend to vary less during training. Thus, a neuron's importance for a task can be measured based on the stability of its incoming weights during training of that task; 6) Neuron state-dependent mechanisms for continual learning (NEO) (Daram and Kudithipudi, 2023) is a neuron importance based mechanism coupled with state-dependent selective learning rules to mitigate catastrophic forgetting.

Replay mechanisms. For replay-based schemes, we select: 1) Episodic Replay (ER) (Chaudhry et al., 2019) which uses random sampling for retrieval from memory and a reservoir sampling technique with a ring buffer to update the replay memory. 2) Averaged Gradient Episodic Memory (A-GEM) (Chaudhry et al., 2018) uses episodic memory as an optimization constraint to avoid catastrophic forgetting. The sample handling of A-GEM avoids solving a quadratic optimization problem for retrieval of samples in the buffer and, instead uses the mean gradient of such samples from the buffer. 2) iCaRL (Rebuffi et al., 2017) uses a neural network for feature extraction and performs classification based on a nearest-class-mean rule, where the class means is retrieved from stored data with a special form of distillation. 3) Gradient-based Sample Selection (GSS) (Aljundi et al., 2019) is a sample selection strategy for a setup without task boundaries or at least knowledge about these. Each seen sample is regarded as an individual constraint, to which every following sample must be compatible. Sample selection in this context is identical to a constraint reduction problem, which is solved by a greedy strategy. This strategy selects n random samples from the buffer and calculates the cosine-similarity between the gradient of the current sample and the gradients of the selected samples.

4. Results and Discussion

The following section will cover the performance and analysis of the alignment-based rules on different continual learning benchmarks in both task-aware and task-agnostic scenarios. We evaluate on three scenarios, i.e., task-incremental, domain-incremental and class-incremental (van de Ven and Tolias, 2019). In the case of task-incremental learning (Task-IL), the model is aware of the task identity during training and inference, whereas in domain-incremental learning (Domain-IL) scenario, tasks share the same output layer while the model is unaware of task identity. And for class-incremental learning (Class-IL), it expands upon Domain-IL scenario, wherein the output layer is not shared between tasks and the output head increases with the number of tasks.

Datasets and Tasks

To study the characteristics of catastrophic forgetting with alignment based rules, we define experiments on extensively adopted continual learning benchmarks. We evaluate on both

Table 1: Comparison of mean accuracy (MA%) of regularization mechanisms on Domain-IL and Task-IL scenarios for the Split MNIST dataset

	Method	Domain-IL(%)	Task-IL (%)
Non-CL Baselines	SGD	61.23 ± 0.66	84.32 ± 0.99
	IA	60.62 ± 0.11	83.82 ± 1.20
	AA	63.32 ± 0.48	85.47 ± 0.56
	SA	62.65 ± 0.26	84.24 ± 0.85
CL Baselines	oEWC	65.42 ± 1.6	99.12 ± 0.11
	SI	65.36 ± 1.57	99.09 ± 0.15
	LwF	71.50 ± 1.63	99.57 ± 0.02
	SS*	82.9 ± 0.01	-
	NEO	78.14 ± 2.23	99.04 ± 0.14
	UCL	69.72 ± 2.53	99.32 ± 0.05
Information Alignment	oEWC + IA	63.28 ± 2.43	98.15 ± 0.46
	SI + IA	64.56 ± 1.24	98.34 ± 0.26
	LwF + IA	68.82 ± 1.63	98.22 ± 0.32
	SS + IA	80.11 ± 0.54	-
	NEO + IA	76.92 ± 1.37	98.18 ± 0.45
	UCL + IA	66.78 ± 3.10	98.02 ± 0.66
Symmetric Alignment	oEWC + SA	66.13 ± 0.88	99.10 ± 0.08
	SI + SA	65.30 ± 1.35	99.11 ± 0.04
	LwF + SA	70.22 ± 0.94	99.14 ± 0.01
	SS + SA	82.45 ± 0.14	-
	NEO + SA	77.69 ± 1.08	99.02 ± 0.11
	UCL + SA	70.00 ± 1.44	99.26 ± 0.04
Activation Alignment	oEWC + AA	64.69 ± 0.86	98.74 ± 0.23
	SI + AA	64.77 ± 1.57	99.01 ± 0.15
	LwF + AA	69.93 ± 1.26	98.89 ± 0.21
	SS + AA	83.3 ± 0.05	-
	NEO + AA	79.9 ± 1.05	99.19 ± 0.04
	UCL + AA	69.72 ± 2.53	99.32 ± 0.05

* SS cannot leverage task-awareness, and therefore is missing values in the task-IL column.

Table 2: Comparison of mean accuracy (MA%) of continual learning mechanisms on Class-IL scenarios for the Split CIFAR-10 and Split CIFAR-100 datasets

	Method	**Split CIFAR-10**				**Split CIFAR-100**			
		SGD(%)	IA (%)	SA(%)	AA (%)	SGD(%)	IA (%)	SA(%)	AA (%)
Regularization Mechanisms	oEWC	19.49 ± 0.12	19.12 ± 1.23	$\mathbf{19.84 \pm 0.08}$	19.37 ± 0.16	8.12 ± 0.35	7.36 ± 0.74	$\mathbf{8.64 \pm 0.27}$	8.05 ± 0.42
	SI	19.48 ± 0.17	19.46 ± 0.22	$\mathbf{20.02 \pm 0.13}$	19.42 ± 0.21	8.10 ± 0.24	6.97 ± 0.72	$\mathbf{8.32 \pm 0.13}$	8.22 ± 0.45
	LwF	$\mathbf{19.61 \pm 0.05}$	18.10 ± 1.63	19.23 ± 0.50	18.84 ± 0.77	$\mathbf{15.93 \pm 0.87}$	13.22 ± 1.49	14.72 ± 0.25	15.04 ± 0.62
	SS	28.13 ± 0.04	$\mathbf{29.22 \pm 0.09}$	28.01 ± 0.11	28.87 ± 0.21	9.82 ± 0.16	9.68 ± 0.28	$\mathbf{10.21 \pm 0.36}$	9.88 ± 0.64
	NEO	25.61 ± 0.05	23.31 ± 1.03	24.95 ± 0.09	$\mathbf{27.72 \pm 0.35}$	8.42 ± 0.12	6.20 ± 0.42	8.13 ± 0.18	$\mathbf{8.48 \pm 0.02}$
	UCL	17.63 ± 0.08	15.85 ± 0.58	$\mathbf{18.04 \pm 0.12}$	17.77 ± 0.04	7.35 ± 0.18	7.04 ± 0.18	$\mathbf{7.86 \pm 0.07}$	7.22 ± 0.13
Replay Mechanisms	ER	$\mathbf{44.79 \pm 1.86}$	41.26 ± 2.45	43.48 ± 0.62	42.10 ± 1.10	37.57 ± 0.21	32.08 ± 2.04	$\mathbf{38.54 \pm 0.12}$	37.23 ± 0.95
	iCARL	$\mathbf{47.55 \pm 3.95}$	43.16 ± 2.03	45.80 ± 1.20	46.62 ± 0.47	$\mathbf{37.83 \pm 0.21}$	35.50 ± 1.22	37.04 ± 0.18	36.45 ± 1.39
	A-GEM	22.67 ± 0.57	20.72 ± 0.89	21.90 ± 0.17	$\mathbf{23.04 \pm 0.10}$	20.38 ± 1.45	19.15 ± 1.88	21.87 ± 0.54	$\mathbf{22.70 \pm 0.05}$
	GSS	49.73 ± 4.78	45.26 ± 2.15	$\mathbf{50.10 \pm 1.63}$	48.20 ± 2.78	40.20 ± 1.40	38.16 ± 0.92	$\mathbf{41.36 \pm 0.88}$	39.44 ± 0.45

task aware and agnostic scenarios on the Split MNIST, Split CIFAR-10 and Split CIFAR-100 datasets.

Split MNIST task consists of splitting the standard ten-class image classification of MNIST digits into five 2-class classification tasks. The model incrementally sees the five

Table 3: Comparison of mean accuracy (MA%) of continual learning mechanisms on Task-IL scenarios for the Split CIFAR-10 and Split CIFAR-100 datasets

	Method	Split CIFAR-10				Split CIFAR-100			
		SGD(%)	IA (%)	SA(%)	AA (%)	SGD(%)	IA (%)	SA(%)	AA (%)
Regularization Mechanisms	oEWC	85.78 ± 1.2	81.26 ± 3.39	**86.45 ± 0.74**	85.83 ± 0.50	62.86 ± 1.70	59.23 ± 4.20	**63.37 ± 0.31**	63.24 ± 0.45
	SI	85.61 ± 1.51	83.46 ± 1.22	**88.10 ± 0.54**	87.42 ± 1.36	60.74 ± 0.39	59.88 ± 0.75	**61.76 ± 0.39**	60.20 ± 1.16
	LwF	**88.42 ± 0.68**	84.34 ± 2.75	87.68 ± 1.12	87.38 ± 1.20	**68.28 ± 1.19**	65.30 ± 3.24	67.20 ± 0.64	67.47 ± 1.34
	NEO	86.4 ± 1.74	83.52 ± 1.43	85.50 ± 1.10	**88.20 ± 1.85**	58.36 ± 1.82	55.21 ± 1.23	57.40 ± 0.68	**59.12 ± 0.77**
	UCL	**86.72 ± 1.65**	84.04 ± 2.20	85.28 ± 1.36	86.37 ± 1.40	63.62 ± 2.13	61.03 ± 3.21	**63.85 ± 2.13**	63.62 ± 0.80
Replay Mechanisms	ER	91.19 ± 0.94	87.70 ± 2.85	90.04 ± 1.25	**91.74 ± 1.63**	68.43 ± 0.24	66.20 ± 1.16	67.50 ± 0.08	**68.89 ± 0.15**
	iCARL	**88.99 ± 2.13**	84.57 ± 0.97	87.20 ± 0.53	87.03 ± 0.92	67.23 ± 1.75	64.50 ± 0.91	67.04 ± 0.71	**67.45 ± 0.59**
	A-GEM	83.88 ± 1.49	81.16 ± 2.35	**83.92 ± 1.18**	82.18 ± 1.45	61.35 ± 1.27	60.33 ± 2.65	61.55 ± 1.04	**62.70 ± 1.01**
	GSS	**88.80 ± 2.89**	87.12 ± 2.12	88.44 ± 0.94	88.29 ± 1.12	69.57 ± 1.68	66.59 ± 1.48	69.03 ± 0.36	**70.20 ± 0.25**

tasks over time. We evaluate on the MNIST dataset for Task-IL and Domain-IL scenarios of continual learning.

Split CIFAR-10 task is analogous to the Split-MNIST task. However, we evaluate this task on Task-IL and Class-IL scenarios only.

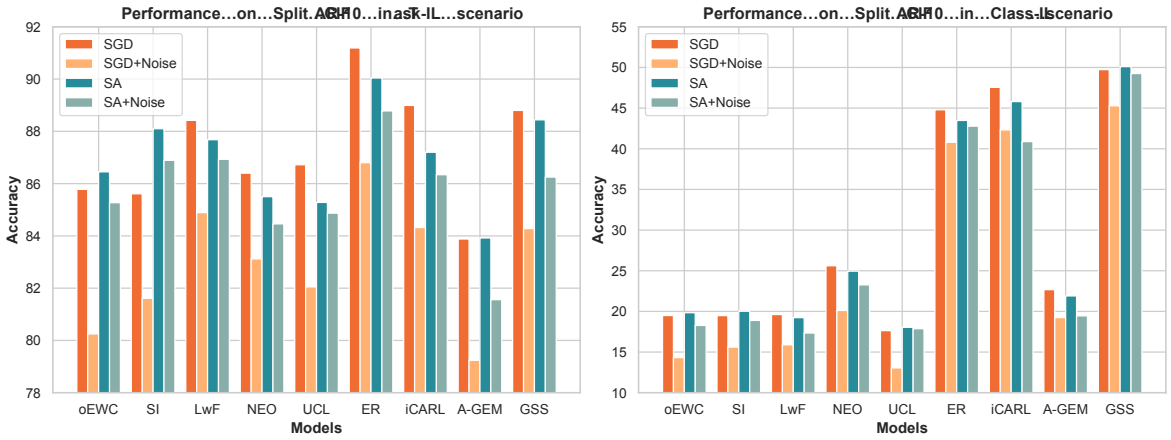
Split CIFAR-100 task consists of 100 classes, which could be split into either 20 tasks of 5 classes each or 10 tasks of 10 classes each. Our construction of this task uses the latter configuration of 10 tasks (Chen et al., 2020).

For the Split MNIST task, we use a multilayer perceptron with two layers of 400 neurons each. As for the Split CIFAR-10 and Split CIFAR-100 task, we use a ResNet-18 architecture with an additional final classifier layer.

Evaluating learning rules

We evaluate the learning rules on two different architectures based on the datasets. Table 1 shows the mean accuracy of the neuron and weight regularization models on the Split MNIST task for Domain-IL and Task-IL scenarios. We observe that for the non-CL baselines, the activation alignment learning rule performs consistently better across both task scenarios. The gradient of AA, as shown in Eq. 4, attempts to align the layer-wise activation response, which in turn impacts the representational similarity across tasks. This preservation of similarity leads to lower forgetting in the network. While evaluating with continual learning models, fully local information alignment is not able to perform in comparison to SGD and other non-local alignment rules. No singular alignment rule works best across the multiple regularization mechanisms and network models. For instance, online-EWC, SI, and UCL, which rely on synapse dynamics (importance measurement dependent on the movement in θ_f across tasks and the shift in their gradients across tasks), perform well with the symmetric alignment rule, which primarily encourages aligning the backward and forward weights while training. On the other hand, methods like NEO and SS, which primarily rely on activation-based importance, perform better under the constraints of the activation-alignment learning rule. LwF, which utilizes soft labels with a distillation loss, performs better with backpropagation itself. This observation points us towards the potential challenge of alignment-based learning mechanisms with distillation. However, consistently better performance on the Split MNIST problem was achieved when alignment-based rules were used.

One problem when considering alignment-based mechanisms is their metaparameter stability and baseline performance while scaling to complex tasks (Sanfiz and Akrout, 2021). Therefore, for the next set of experiments, we evaluate the CL algorithms (including both regularization and replay-based approaches) on Split CIFAR-10 and Split CIFAR-100 datasets with both Task-IL and Class-IL scenarios. Tables 2 and 3 demonstrate the mean accuracy of the aforementioned regularization and replay-based models with alignment-based mechanisms. We notice that as the complexity of the model and task increased, the purely local Information Alignment learning rule observed a higher drop in performance. This can be attributed to the implicit symmetry of weights not being able to generate pseudogradients as close to the exact ones. However, both non-local learning rules encourage symmetry, and both performed at least as well as backpropagation for complex tasks. Even in these scenarios, the correlation between the performance of regularization mechanisms is analogous to the behavior shown in the Split MNIST task. The replay mechanisms, on the other hand, did not show any noticeable improvement in performance when coupled with non-local alignment rules. The benefits of these rules with replay are demonstrated when training with noisy perturbations in weight updates. The next section addresses the robustness and variance analysis of the proposed learning rules in both task-aware and agnostic scenarios.

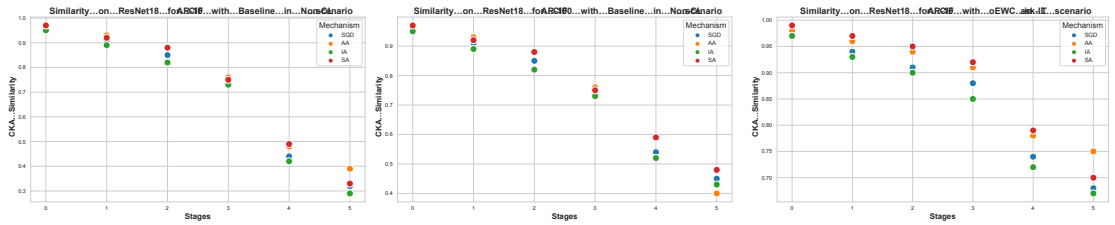


(*a*) Task-IL scenario (*b*) Class-IL scenario

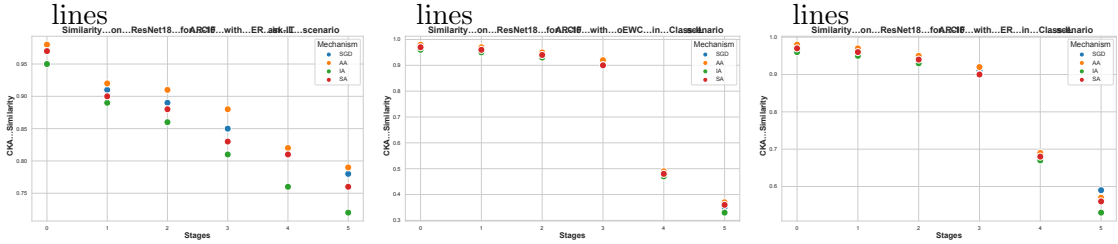
Figure 1: Mean accuracy of the continual learning mechanisms on the Split CIFAR-10 dataset in (a) task-IL scenarios and (b) task-agnostic scenarios. The models were trained with and without noise added to the gradients. The log variance of applied Gaussian noise was set to -5. Models trained using Symmetric Alignment are observed to be more robust than backpropagation.

Robustness Analysis

As a proxy to determine our conclusions about the robustness of these rules under the context of continual learning, we model this uncertainty by adding Gaussian noise to the



(a) Split CIFAR-10 Baselines (b) Split CIFAR-100 Baselines (c) Split CIFAR-100 Baselines



(d) ER on Split CIFAR-10 (Task-IL) (e) oEWC on Split CIFAR-10 (Class-IL) (f) ER on CIFAR-10 with (Class-IL)

Figure 2: Visualization of the representational similarity across different stages using ResNet-18 as the prototypical model for different learning scenarios. The y-axis represents the CKA score between activations of block of convolutions (stages) before and after training on the second task. In general, both activation and symmetric alignment enable better preservation of representations from the prior task. Symmetric alignment with weight regularization mechanisms such as oEWC show better retention capabilities in deeper layers.

backward updates during learning. We modify the update rule by,

$$\Delta\theta_b \propto \frac{\partial \mathcal{P}}{\partial \theta_b} + \mathcal{N}(0, \sigma^2) \tag{6}$$

The forward weights ($\Delta\theta_f \propto \tilde{\nabla}\mathcal{G}$) change based on the modification of the backward weights. We add similar noise to the gradients for backpropagation as well. Figure 1, shows the performance of the CL algorithms on Split CIFAR-10 benchmark with the log variance of Gaussian noise (σ^2) at -5. It can be seen that Symmetric alignment is more robust than backpropagation and other rules towards noisy updates. Prior work has shown how noise in the presence of spiking discontinuities was useful for weight estimation (Guerguiev et al., 2019). This behavior was translated to rate coded networks as well. Therefore, employing pseudogradients from weight estimation techniques with noisy updates leads to increased robust learning in comparison to using exact noisy gradients.

Representational Variance Analysis

In this section, we investigate the changes in semantic representations of the activations in the network while training sequentially. To understand these properties, we utilize the Centered Kernel Alignment metric proposed in (Kornblith et al., 2019). This metric measures the similarity between two representations of the same dataset which is invariant to orthogonal transformation. For instance, given the data from a task $\mathbb{T}_k$ with n examples, we compare two representations H and S, with m_h and m_s features such that. $H \in \mathbb{R}^{n \times m_h}$ and $S \in \mathbb{R}^{n \times m_s}$. In this case the size of features is the same i.e. $m_h = m_s$, since we compare features from the same layers in the network. The CKA similarity between two representations is given by,

$$CKA(H, S) = \frac{||H^T S||}{||H^T H||_F^2 ||S^T S||_F^2} \tag{7}$$

In our investigation, we measure the CKA similarity between representations of the same data across different tasks, i.e H represents the layer activations for task k and S represents the layer activations for task $k+1$. Figure 2 compares the CKA similarity for CL algorithms trained using the alignment learning rules and backpropagation. The lower layers in the network do not change much while training later tasks, however the major change is observed in the deeper layers. In general, those with higher representational similarity demonstrate better capabilities in preserving previous knowledge, thereby showing an improved response to catastrophic forgetting. It can be noted that the CKA similarity is higher for the alignment based mechanisms, demonstrating the impact of local interactions towards learning tasks sequentially. These interactions work in conjunction with importance measurement techniques to mitigate forgetting. Moreover, interpolating this similarity analysis shows the efficacy of the learning algorithms in the context of continual learning.

5. Conclusion

In this paper, we investigated the role of multiple alignment based learning rules for continual learning. We raise a question regarding why should alignment mechanisms be employed

in lieu of backpropagation? This concern is addressed due to its two-fold benefits. One, there is a positive correlation between the local primitives and the importance measurement techniques, making them viable for exploring CL models that leverage locality. We show that the layer-wise alignment of weights, activations or representations can benefit synapse or neuron regularization mechanisms for alleviating catastrophic forgetting. Two, the decoupled loss function with weight estimation techniques is observed to be more robust to noise, which is a necessary feature for continual learning. We observe that employing alignment-based rules instead of backpropagation enhances the models' resilience to noisy updates. By using representational similarity techniques, we demonstrate how the local regularization primitives are better at preserving prior tasks' knowledge. Overall, this work provides an insight into the potential benefits and limitations of biologically plausible alternatives to backpropagation that are generalizable across several continual learning models.

Acknowledgments

We would like to thank the Neuromorphic AI lab members at UTSA and Dr. William Severa for the feedback on the paper. This effort is partially supported by NSF EFRI BRAID Award #2317706 and NSF NAIAD Award #2332744.

References

Wickliffe C. Abraham. Metaplasticity: Tuning synapses and networks for plasticity. *Nature Reviews Neuroscience*, 9(5):387–399, May 2008. ISSN 1471-0048. doi: 10.1038/nrn2356.

Wickliffe C. Abraham and Mark F. Bear. Metaplasticity: The plasticity of synaptic plasticity. *Trends in Neurosciences*, 19(4):126–130, 1996. ISSN 0166-2236. doi: 10.1016/S0166-2236(96)80018-X. URL http://www.sciencedirect.com/science/article/pii/S016622369680018X.

Mohamed Akrout. On the adversarial robustness of neural networks without weight transport. *arXiv preprint arXiv:1908.03560*, 2019.

Mohamed Akrout, Collin Wilson, Peter Humphreys, Timothy Lillicrap, and Douglas B Tweed. Deep learning without weight transport. *Advances in neural information processing systems*, 32, 2019.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, pages 11816–11825, 2019.

Jason M Allred and Kaushik Roy. Controlled forgetting: Targeted stimulation and dopaminergic plasticity modulation for unsupervised lifelong learning in spiking neural networks. *Frontiers in neuroscience*, 14, 2020.

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. *arXiv preprint arXiv:2201.12604*, 2022.

Pierre Baldi and Peter Sadowski. A theory of local learning, the learning channel, and the optimality of backpropagation. *Neural Networks*, 83:51–74, 2016.

Sergey Bartunov, Adam Santoro, Blake Richards, Luke Marris, Geoffrey E Hinton, and Timothy Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Advances in neural information processing systems*, 31, 2018.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Hung-Jen Chen, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. Mitigating forgetting in online continual learning via instance-aware parameterization. *Advances in Neural Information Processing Systems*, 33:17466–17477, 2020.

Anna Choromanska, E Tandon, Sadhana Kumaravel, Ronny Luss, Irina Rish, Brian Kingsbury, Ravi Tejwani, and Djallel Bouneffouf. Beyond backprop: Alternating minimization with co-activation memory. *stat*, 1050:24, 2018.

Jeff Clune. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*, 2019.

Anurag Daram and Dhireesha Kudithipudi. Neo: Neuron state dependent mechanisms for efficient continual learning. In *Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference*, pages 11–19, 2023.

Anurag Daram and Angel Yanguas-Gil. Exploring neuromodulation for dynamic learning. *Frontiers in Neuroscience*, 14:526929, 2020.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. *arXiv preprint arXiv:2003.09553*, 2020.

Kai Olav Ellefsen, Jean-Baptiste Mouret, and Jeff Clune. Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLoS computational biology*, 11 (4):e1004128, 2015.

Charlotte Frenkel, Martin Lefebvre, and David Bol. Learning without feedback: Fixed random learning signals allow for feedforward training of deep neural networks. *Frontiers in neuroscience*, 15:629892, 2021.

Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive science*, 11(1):23–63, 1987.

Jordan Guerguiev, Konrad Kording, and Blake Richards. Spike-based causal inference for weight alignment. In *International Conference on Learning Representations*, 2019.

Sangwon Jung, Hongjoon Ahn, Sungmin Cha, and Taesup Moon. Continual learning with node-importance based adaptive group sparse regularization. *Advances in Neural Information Processing Systems*, 33:3647–3658, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017a. ISSN 0027-8424. doi: 10.1073/pnas.1611835114.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017b.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017c.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022.

Dhireesha Kudithipudi, Anurag Daram, Abdullah M Zyarah, Fatima Tuz Zohora, James B Aimone, Angel Yanguas-Gil, Nicholas Soures, Emre Neftci, Matthew Mattina, Vincenzo Lomonaco, et al. Design principles for lifelong learning ai accelerators. *Nature Electronics*, 6(11):807–822, 2023.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *International conference on machine learning*, pages 3560–3569. PMLR, 2019.

Daniel Kunin, Aran Nayebi, Javier Sagastuy-Brena, Surya Ganguli, Jonathan Bloom, and Daniel Yamins. Two routes to scalable credit assignment without weight symmetry. In *International Conference on Machine Learning*, pages 5511–5521. PMLR, 2020.

Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):13276, 2016.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002.

Sandeep Madireddy, Angel Yanguas-Gil, and Prasanna Balaprakash. Improving performance in continual learning tasks using bio-inspired architectures. In *Conference on Lifelong Learning Agents*, pages 992–1008. PMLR, 2023.

Nicolas Y Masse, Gregory D Grant, and David J Freedman. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, 115(44):E10467–E10475, 2018.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989. ISBN 0079-7421. URL http://www.sciencedirect.com/science/article/pii/S0079742108605368.

Bl Mcnaughton and Rgm Morris. Hippocampal Synaptic Enhancement and Information-Storage Within a Distributed Memory System. *Trends in Neurosciences*, 10(10):408–415, 1987. ISSN 0166-2236. doi: 10.1016/0166-2236(87)90011-7.

Thomas Miconi, Aditya Rawal, Jeff Clune, and Kenneth O Stanley. Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity. *arXiv preprint arXiv:2002.10585*, 2020.

Oleksiy Ostapenko, Pau Rodriguez, Massimo Caccia, and Laurent Charlin. Continual learning via local module composition. *Advances in Neural Information Processing Systems*, 34:30298–30312, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

Raul Rojas and Raúl Rojas. The backpropagation algorithm. *Neural networks: a systematic introduction*, pages 149–182, 1996.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Albert Jiménez Sanfiz and Mohamed Akrout. Benchmarking the accuracy and robustness of feedback alignment algorithms. *arXiv preprint arXiv:2108.13446*, 2021.

Steger. Angelika Schug. Simon, Benzing. Frederik. Task-agnostic continual learning via stochastic synapses. https://sites.google.com/view/cl-icml/accepted-papers?authuser=0, 2020. (Accessed on 09/21/2020).

Jonathan Schwarz, Jelena Luketina, Wojciech M Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. *arXiv preprint arXiv:1805.06370*, 2018.

Nicholas Soures, Peter Helfer, Anurag Daram, Tej Pandit, and Dhireesha Kudithipudi. Tacos: Task agnostic continual learning in spiking neural networks. In *Theory and Foundation of Continual Learning Workshop at ICML'2021*, July 2021.

Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning. *arXiv:1904.07734 [cs, stat]*, April 2019.

Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14, 2020.

Alanna J Watt and Niraj S Desai. Homeostatic plasticity and stdp: keeping a neuron's cool in a fluctuating world. *Frontiers in synaptic neuroscience*, 2:5, 2010.

James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 29(5):1229–1262, 2017.

Angel Yanguas-Gil, Anil Mane, Jeffrey W Elam, Felix Wang, William Severa, Anurag Reddy Daram, and Dhireesha Kudithipudi. The insect brain as a model system for low power electronics and edge processing applications. In *2019 IEEE Space Computing Conference (SCC)*, pages 60–66. IEEE, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3987–3995. JMLR. org, 2017.