A Study of Biologically Plausible Neural Network: The Role and Interactions of Brain-Inspired Mechanisms in Continual Learning

Anonymous authors
Paper under double-blind review

Abstract

Humans excel at continually acquiring, consolidating, and retaining information from an ever-changing environment, whereas artificial neural networks (ANNs) exhibit catastrophic forgetting. There are considerable differences in the complexity of synapses, the processing of information, and the learning mechanisms in biological neural networks and their artificial counterparts, which may explain the mismatch in performance. We consider a biologically plausible framework that constitutes separate populations of exclusively excitatory and inhibitory neurons that adhere to Dale's principle, and the excitatory pyramidal neurons are augmented with dendritic-like structures for context-dependent processing of stimuli. We then conduct a comprehensive study on the role and interactions of different mechanisms inspired by the brain, including sparse non-overlapping representations, Hebbian learning, synaptic consolidation, and replay of past activations that accompanied the learning event. Our study suggests that the employing of multiple complementary mechanisms in a biologically plausible architecture, similar to the brain, may be effective in enabling continual learning in ANNs. [1]

1 Introduction

The human brain excels at continually learning from a dynamically changing environment, whereas standard artificial neural networks (ANNs) are inherently designed for training from stationary i.i.d. data. Sequential learning of tasks in continual learning (CL) violates this strong assumption, resulting in catastrophic forgetting. Although ANNs are inspired by biological neurons (Fukushima, 1980), they omit numerous details of the design principles and learning mechanisms in the brain. These fundamental differences may account for the mismatch in performance and behavior.

Biological neural networks are characterized by considerably more complex synapses and dynamic context-dependent processing of information. In addition, individual neurons have a specific role. Each presynaptic neuron has an exclusive excitatory or inhibitory impact on its postsynaptic partners, as postulated by Dale's principle (Strata et al., 1999). Furthermore, distal dendritic segments in pyramidal neurons, which comprise the majority of excitatory cells in the neocortex, receive additional context information and enable context-dependent processing of information. This, in conjunction with inhibition, allows the network to learn task-specific patterns and avoid catastrophic forgetting (Yang et al., 2014; Iyer et al., 2022; Barron et al., 2017). Furthermore, replay of non-overlapping and sparse neural activities of previous experiences in the neocortex and hippocampus is considered to play a critical role in memory formation, consolidation, and retrieval (Walker & Stickgold, 2004; McClelland et al., 1995). To protect information from erasure, the brain employs synaptic consolidation, in which plasticity rates are selectively reduced in proportion to strengthened synapses (Cichon & Gan, 2015).

Thus, we study the role and interactions of different mechanisms inspired by the brain in a biologically plausible framework in a CL setup. The underlying model constitutes separate populations of exclusively excitatory

[1] We will make the code available upon acceptance.

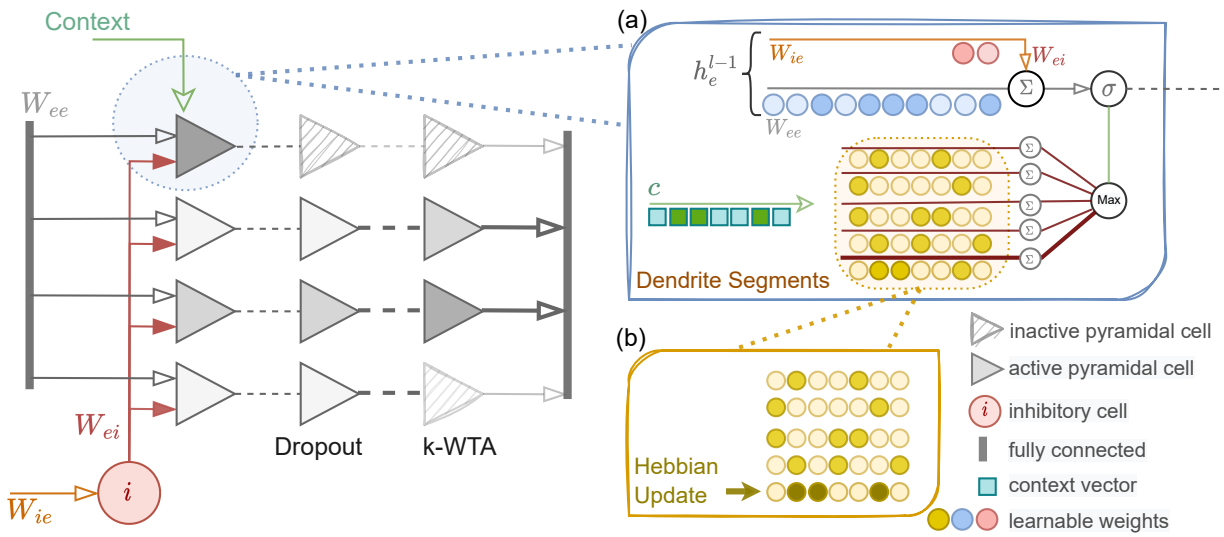

Figure 1: Architecture of one hidden layer in the biologically plausible framework. Each layer consists of separate populations of exclusively excitatory pyramidal cells and inhibitory neurons, which adhere to Dale's principle. The shade indicates the strength of weights or activations, with a darker shade indicating a higher value. (a) The pyramidal cells are augmented with dendritic segments, which receive an additional context signal c and the dendritic segment whose weights are most aligned with the context vector (bottom row) is selected to modulate the output activity of the feedforward neurons for context-dependent processing of information. (b) The Hebbian update step further strengthens the association between the context and the winning dendritic segment with maximum absolute value (indicated with a darker shade in the bottom row). Finally, Heterogeneous dropout keeps the activation count of each pyramidal cell (indicated with the gray shade) and drops the neurons that were most active for the previous task (the darkest shade dropped) to enforce non-overlapping representations. The top-k remaining cells then project to the next layer (increased shade). This provides us with a more biologically plausible framework within which we can study the role of different brain-inspired mechanisms and provide insights for designing new CL methods.

and inhibitory neurons in each layer, which adheres to Dale's principle (Cornford et al., 2020) and excitatory neurons (mimicking pyramidal cells) are augmented with dendrite-like structures for context-dependent processing of information (Iyer et al., 2022). Dendritic segments process an additional context signal encoding task information and subsequently modulate the feedforward activity of the excitatory neuron (Figure 1). We then systematically study the effect of controlling the overlap in representations, employing the "fire together, wire together" learning paradigm and employing experience replay and synaptic consolidation. Our study shows that:

i. An ANN architecture equipped with context-dependent processing of information by dendrites and adhering to Dale's principle can learn effectively in CL setup. Importantly, accounting for the discrepancy in the effect of weight changes in excitatory and inhibitory neurons further reduces forgetting in CL.
ii. Enforcing different levels of activation sparsity in the hidden layers using k-winner-take-all activations and employing a complementary dropout mechanism that encourages the model to use a different set of active neurons for each task can effectively control the overlap in representations, and hence reduce interference while allowing for resusability.
iii. Task similarities need to be considered when enforcing such constraints to allow for a balance between forwarding transfer and interference.
iv. Mimicking the ubiquitous "fire together, wire together" learning rule in the brain through a Hebbian update step on the connection between context signal and dendritic segments, which further strengthens context gating and facilitates the formation of task-specific subnetworks.

v. We show that employing both synaptic consolidation with importance measures adjusted to take into account the discrepancy in the effect of weight changes and a replay mechanism in a context-specific manner is critical for consolidating information across different tasks, especially for challenging CL settings.

Our study suggests that employing multiple complementary mechanisms in a biologically plausible architecture, similar to what is believed to exist in the brain, can be effective in enabling CL in ANNs. To the best of our knowledge, we are the first to provide a comprehensive study of the integration of different brain-inspired mechanisms in a biologically plausible architecture in a CL setting.

2 Biologically Plausible Framework for CL

We provide details of the biologically plausible framework within which we conduct our study.

2.1 Dale's Principle

Biological neural networks differ from their artificial counterparts in the complexity of synapses and the role of individual units. In particular, the majority of neurons in the brain adhere to Dale's principle, which posits that presynaptic neurons can only have an exclusive excitatory or inhibitory impact on their postsynaptic partners (Strata et al., 1999). Several studies show that the balanced dynamics (Murphy & Miller, 2009; Van Vreeswijk & Sompolinsky, 1996) of excitatory and inhibitory populations provide functional advantages, including efficient predictive coding (Boerlin et al., 2013) and pattern learning (Ingrosso & Abbott, 2019). Furthermore, inhibition is hypothesized to play a role in alleviating catastrophic forgetting (Barron et al., 2017). Standard ANNs, however, lack adherence to Dale's principle, as neurons contain both positive and negative output weights, and signs can change while learning.

Cornford et al. (2020) incorporate Dale's principle into ANNs (referred to as DANNs), which take into account the distinct connectivity patterns of excitatory and inhibitory neurons (Tremblay et al., 2016) and perform comparable to standard ANNs in the benchmark object recognition task. Each layer l comprises of a separate population of excitatory, $h_e^l \in \mathbb{R}_+^{n_e}$, and inhibitory $h_i^l \in \mathbb{R}_+^{n_i}$ neurons, where $n_e \gg n_i$ and synaptic weights are strictly non-negative. Similar to biological networks, while both populations receive excitatory projections from the previous layer (h_e^{l-1}), only excitatory neurons project between layers, whereas inhibitory neurons inhibit the activity of excitatory units of the same layer. Cornford et al. (2020) characterized these properties by three sets of strictly positive weights: excitatory connections between layers $W_{ee}^l \in \mathbb{R}_+^{n_e \times n_e}$, excitatory projection to inhibitory units $W_{ie}^l \in \mathbb{R}_+^{n_i \times n_e}$, and inhibitory projections within the layer $W_{ei}^l \in \mathbb{R}_+^{n_e \times n_i}$. The output of the excitatory units is impacted by the subtractive inhibition from the inhibitory units:

$$z^l = (W_{ee}^l - W_{ei}^l W_{ie}^l) h_e^{l-1} + b^l \tag{1}$$

where $b^l \in \mathbb{R}^{n_e}$ is the bias term. Figure 1 shows the interactions and connectivity between excitatory pyramidal cells (triangle symbol) and inhibitory neurons (denoted by i).

We aim to employ DANNs as feedforward neurons to show that they can also learn in a challenging CL setting and performance comparable to standard ANNs and provide a biologically plausible framework for further studying the role of inhibition in alleviating catastrophic forgetting.

2.2 Active Dendrites

The brain employs specific structures and mechanisms for context-dependent processing and routing of information. The prefrontal cortex, which plays an important role in cognitive control (Miller & Cohen, 2001), receives sensory inputs as well as contextual information, which allows it to choose the most relevant sensory features for the present task to guide actions (Mante et al., 2013; Fuster, 2015; Siegel et al., 2015; Zeng et al., 2019). Of particular interest are pyramidal cells, which represent the most populous members of the excitatory family of neurons in the brain (Bekkers, 2011). The dendritic spines in pyramid cells exhibit highly non-linear integrative properties that are considered important for learning task-specific patterns (Yang et al.,

2014). Pyramidal cells integrate a range of diverse inputs into multiple independent dendritic segments, allowing contextual inputs in active dendrites to modulate the response of a neuron, making it more likely to fire. However, standard ANNs are based on a point neuron model (Lapique, 1907) which is an oversimplified model of biological computations and lacks the sophisticated non-linear and context-dependent behavior of pyramidal cells.

Iyer et al. (2022) model these integrative properties of dendrites by augmenting each neuron with a set of dendritic segments. Multiple dendritic segments receive additional contextual information, which is processed using a separate set of weights. The resultant dendritic output modulates the feedforward activation, which is computed by a linearly weighted sum of the feedforward inputs. This computation results in a neuron where the magnitude of the response to a given stimulus is highly context-dependent. To enable task-specific processing of information, the prototype vector for task τ is evaluated by taking the element-wise mean of the tasks samples, $\mathcal{D}_\tau$ at the beginning of the task and then subsequently provided as context during training.

$$c_\tau = \frac{1}{|\mathcal{D}_\tau|} \sum_{x \in \mathcal{D}_\tau} x \tag{2}$$

During inference, the closest prototype vector to each test sample, x', is selected as the context using the Euclidean distance among all task prototypes, $\boldsymbol{C}$, stored in memory.

$$c' = \arg\min_{c_\tau} \|\boldsymbol{x}' - \boldsymbol{C}_\tau\|_2 \tag{3}$$

Following Iyer et al. (2022), we augment the excitatory units in each layer with dendritic segments (Figure 1 (a)). The feedforward activity of excitatory units is modulated by dendritic segments, which receive a context vector. Given the context vector, each dendritic segment j computes $u_j^T c$, given weight $u_j \in \mathbb{R}^d$ and the context vector $c \in \mathbb{R}^d$ where d is the dimensions of the input image. For excitatory neurons, the dendritic segment with the highest response to the context (maximum absolute value with the sign retained) is selected to modulate output activity.

$$h_e^l = k\text{-}WTA(z_l \times \sigma(u_\kappa^T c)), \qquad \text{where } \kappa = \arg\max_j |u_j^T c| \tag{4}$$

where σ is the sigmoid function (Han & Moraga, 1995) and $k\text{-}WTA(.)$ is the k-Winner-Take-All activation function (Ahmad & Scheinkman, 2019) which propagates only the top k neurons and sets the rest at zero. This provides us with a biologically plausible framework where, similar to biological networks, the feedforward neurons adhere to Dale's principle, and the excitatory neurons mimic the integrative properties of active dendrties for context-dependent processing of stimuli.

3 Continual Learning Settings

To study the role of different components inspired by the brain in a biologically plausible NN for CL and gauge their roles in the performance and characteristics of the model, we conduct all our experiments under uniform settings. Implementation details and experimental setup are provided in Appendix. We evaluate the models on two CL scenarios. **Domain incremental learning (Domain-IL)** refers to the CL scenario in which the classes remain the same in subsequent tasks but the input distribution changes. We consider Rot-MNIST which involves classifying the 10 digits in each task with each digit rotated by an angle between 0 and 180 degrees, and Perm-MNIST which applies a fixed random permutation to the pixels for each task. Importantly, there are different variants of Rot-MNIST with varying difficulties. We incrementally rotate the digits to a fixed degree, i.e. {0, 8, ..., (N-1)*8} for task $\{\tau_1, \tau_2, .., \tau_N\}$ which is substantially more challenging than random sampling rotations. Importantly, the Rot-MNIST dataset captures the notion of similarity in subsequent tasks, where the similarity between two tasks is defined by the difference in their degree of rotation, whereas each task in Perm-MNIST is independent. We also consider the challenging **Class incremental learning (Class-IL)** scenario where new classes are added with each subsequent task and the agent must learn to distinguish not only amongst the classes within the current task but also across all learned tasks. Seq-MNIST divides the MNIST classification into 5 tasks with 2 classes for each task.

Table 1: Effect of each component of the biologically plausible framework on different datasets with varying number of tasks. We first show the effect of utilizing feedforward neurons adhering to Dale's principle in conjunction with *Active Dendrites* to form the framework within which we evaluate the individual effect of brain-inspired mechanisms before combining them all together (along with Heterogeneous Dropout) to forge Bio-ANN. For all the experiments, we set the percentage of active neurons to 5. We provide the average task performance and 1 std of three runs.

Method	Rot-MNIST			Perm-MNIST			Seq-MNIST
	5	10	20	5	10	20	
Active Dendrites	$92.72_{\pm 0.31}$	$71.48_{\pm 0.60}$	$48.13_{\pm 0.73}$	$95.53_{\pm 0.10}$	$94.37_{\pm 0.26}$	$91.76_{\pm 0.39}$	$20.06_{\pm 0.36}$
+ Dale's Principle	$92.28_{\pm 0.27}$	$70.78_{\pm 0.23}$	$48.79_{\pm 0.27}$	$95.77_{\pm 0.33}$	$95.06_{\pm 0.29}$	$92.40_{\pm 0.38}$	$19.81_{\pm 0.03}$
+ Hebbian Update	$92.68_{\pm 0.36}$	$71.42_{\pm 0.94}$	$49.26_{\pm 0.58}$	$95.97_{\pm 0.16}$	$94.96_{\pm 0.14}$	$92.69_{\pm 0.19}$	$19.85_{\pm 0.04}$
+ SC	$93.40_{\pm 0.86}$	$75.87_{\pm 1.35}$	$64.78_{\pm 3.43}$	$96.67_{\pm 0.23}$	$96.36_{\pm 0.10}$	$95.61_{\pm 0.10}$	$20.26_{\pm 0.56}$
+ ER	$95.15_{\pm 0.37}$	$90.86_{\pm 0.52}$	$83.42_{\pm 0.44}$	$96.75_{\pm 0.15}$	$96.01_{\pm 0.14}$	$94.50_{\pm 0.16}$	$86.88_{\pm 0.83}$
+ ER + CR	$96.67_{\pm 0.06}$	$93.85_{\pm 0.24}$	$89.38_{\pm 0.16}$	$\mathbf{97.34}_{\pm 0.03}$	$97.03_{\pm 0.04}$	$96.12_{\pm 0.04}$	$89.23_{\pm 0.48}$
Bio-ANN	$\mathbf{96.82}_{\pm 0.14}$	$\mathbf{94.64}_{\pm 0.23}$	$\mathbf{91.32}_{\pm 0.26}$	$97.33_{\pm 0.04}$	$\mathbf{97.07}_{\pm 0.05}$	$\mathbf{96.51}_{\pm 0.03}$	$\mathbf{89.26}_{\pm 0.42}$

Table 2: Effect of different levels of sparsity in activations on the performance of the model. Columns show the ratio of active neurons (k in k-WTA activation), and rows provide the number of tasks.

k / #Tasks	Rot-MNIST				Perm-MNIST			
	0.05	0.10	0.20	0.50	0.05	0.10	0.20	0.50
5	$92.28_{\pm 0.27}$	$92.26_{\pm 0.31}$	$\mathbf{92.79}_{\pm 0.44}$	$92.26_{\pm 0.65}$	$95.77_{\pm 0.33}$	$\mathbf{96.32}_{\pm 0.20}$	$90.29_{\pm 6.07}$	$74.51_{\pm 13.55}$
10	$70.78_{\pm 0.23}$	$71.95_{\pm 1.54}$	$\mathbf{73.32}_{\pm 0.69}$	$71.61_{\pm 0.76}$	$\mathbf{95.06}_{\pm 0.29}$	$93.45_{\pm 0.92}$	$72.68_{\pm 12.83}$	$41.33_{\pm 6.72}$
20	$\mathbf{48.79}_{\pm 0.27}$	$47.96_{\pm 1.84}$	$48.65_{\pm 0.91}$	$47.71_{\pm 0.91}$	$\mathbf{92.40}_{\pm 0.38}$	$84.28_{\pm 1.35}$	$63.84_{\pm 3.45}$	$20.80_{\pm 0.99}$

4 Empirical Evaluation

To investigate the impact of the different components inspired by the brain, we use the aforementioned biologically plausible framework and study the effect on the performance and characteristics of the model.

4.1 Effect of Inhibitory Neurons

We first study whether feedforward networks with separate populations of excitatory and inhibitory units can work well in the CL setting. Importantly, we note that when learning a sequence of tasks with inhibitory neurons, it is beneficial to take into account the disparities in the degree to which updates to different parameters affect the layer's output distribution (Cornford et al., 2020) and hence forgetting. Specifically, since W^l_{ie} and W^l_{ei} affect the output distribution to a higher degree than W^l_{ee}, we reduce the learning rate for these weights after the first task (see Appendix).

Table 1 shows that models with feedforward neurons adhering to Dale's principle perform as well as the standard neurons and can also further mitigate forgetting in some cases. Note that this gain comes with considerably fewer parameters and context-dependent processing, as we keep the number of neurons in each layer the same, and only excitatory neurons ($\sim$90%) are augmented with dendritic segments. For 20 tasks, Active Dendrite with Dale's principle reduces the parameters from $\sim$70M to less than $\sim$64M parameters. We hypothesize that having separate populations within a layer enables them to play a specialized role. In particular, inhibitory neurons can selectively inhibit certain excitatory neurons based on stimulus, which can further facilitate the formation of task-specific subnetworks and complement the context-dependent processing of information by dendritic segments.

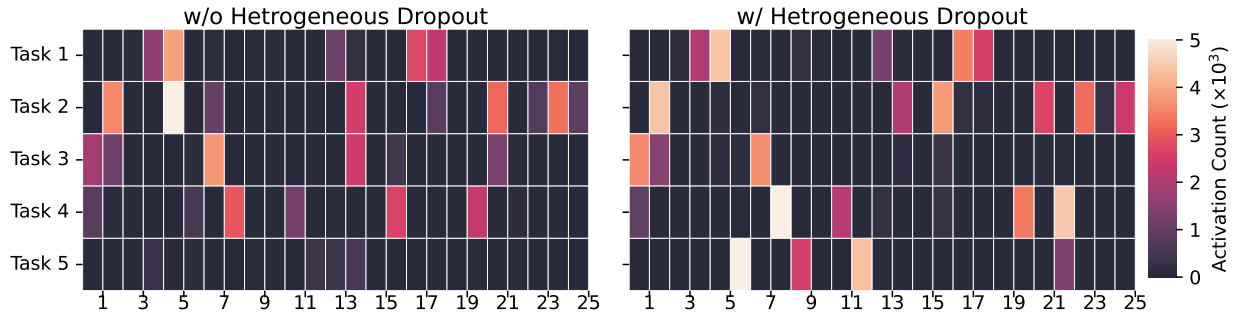

Figure 2: Total activation counts for the test set of each task (y-axis) for a random set of 25 units in the second hidden layer of the model. Heterogeneous dropout reduces the overlap in activations and facilitates the formation of task-specific subnetworks.

Table 3: Effect of Heterogeneous dropout with increasing ρ values on different datasets with varying number of tasks.

Dataset	# Tasks	w/o Dropout	Dropout parameter (ρ)				
			0.1	0.3	0.5	0.7	1.0
Rot-MNIST	5	92.28±0.20	91.79±0.53	92.53±0.11	92.74±0.38	93.19±0.32	**93.42**±0.25
	10	70.78±0.23	71.53±1.07	72.38±1.44	73.63±1.00	74.20±0.78	**75.50**±0.74
	20	48.79±0.27	48.57±0.90	48.91±0.65	49.84±0.59	51.03±0.31	**51.11**±0.76
Perm-MNIST	5	95.77±0.33	95.70±0.29	95.97±0.44	96.40±0.28	**96.58**±0.17	96.48±0.26
	10	95.06±0.29	95.23±0.04	95.65±0.20	95.54±0.26	95.74±0.22	**95.94**±0.24
	20	92.40±0.38	92.83±0.42	**93.20**±0.32	92.82±0.06	93.09±0.47	91.77±0.30

4.2 Sparse Activations Facilitate the Formation of Subnetworks

Neocortical circuits are characterized by high levels of sparsity in neural activations (Barth & Poulet, 2012; Graham & Field, 2006). There is further evidence suggesting that neuronal coding of natural sensory stimuli should be sparse (Barth & Poulet, 2012; Tolhurst et al., 2009). This is in stark contrast to the dense and highly entangled connectivity in standard ANNs. Particularly for CL, sparsity provides several advantages: sparse non-overlapping representations can reduce interference between tasks (Abbasi et al., 2022; Iyer et al., 2022; Aljundi et al., 2019), can lead to the natural emergence of task-specific modules (Hadsell et al., 2020).

We study the effect of different levels of activation sparsity by varying the ratio of active neurons in k-winner-take-all (k-WTA) activations (Ahmad & Scheinkman, 2019). Each hidden layer of our model has a constant sparsity in its connections (randomly 50% of weights are set to 0 at initialization) and propagates only the top-k activations (in Figure 1, k-WTA layer). Table 2 shows that sparsity plays a critical role in enabling CL in DNNs. Sparsity in activations effectively reduces interference by reducing the overlap in representations. Interestingly, the stark difference in the effect of different levels of sparse activations on Rot-MNIST and Perm-MNIST highlights the importance of considering task similarity in the design of CL methods. As the tasks in Perm-MNIST are independent of each other, having fewer active neurons (5%) enables the network to learn non-overlapping representations for each task, while the high task similarity in Rot-MNIST can benefit from overlapping representations, which allows for the reusability of features across the tasks. The number of tasks the agent has to learn also has an effect on the optimal sparsity level. In Appendix, we show that having different levels of sparsity in different layers can further improve performance. As the earlier layers learn general features, having a higher ratio of active neurons can enable higher resuability and forward transfer. For the later layers, a smaller ratio of active neurons can reduce interference between task-specific features.

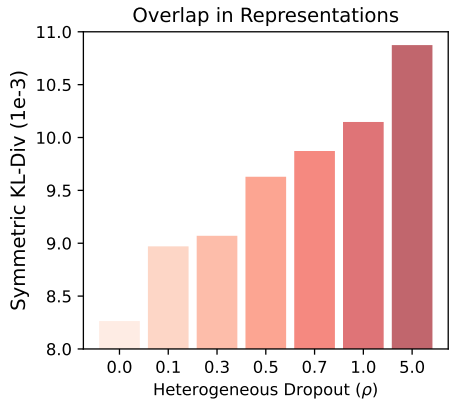

Figure 3: Effect of dropout ρ on the overlap between the distributions of layer two activation counts for each task in Perm-MNIST with 2 tasks. Higher ρ reduces the overlap.

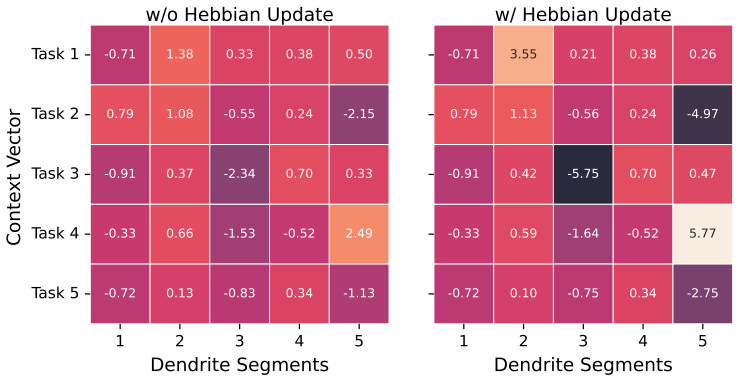

Figure 4: Dendritic segment activations of a unit in layer 1 for the context vectors of each task for a model trained on Perm-MNIST with 5 tasks. Hebbian update strengthens the association between the context and the dendritic segments, increasing the magnitude of the modulating signal.

4.3 Heterogeneous Dropout for Non-overlapping Activations and Subnetworks

Information in the brain is encoded by the strong activation of a relatively small set of neurons that form a sparse coding. A different subset of neurons is utilized to represent different types of stimuli Graham & Field (2006). Furthermore, there is evidence of non-overlapping representations in the brain. To mimic this, we employ Heterogeneous dropout (Abbasi et al., 2022) which in conjunction with context-dependent processing of information, can effectively reduce the overlap of representations, leading to less interference between tasks and, thereby, less forgetting. During training, we track the frequency of activations for each neuron in a layer for a given task, and in the subsequent tasks, the probability of a neuron being dropped is inversely proportional to its activation counts. This encourages the model to learn the new task using neurons that have been less active for previous tasks. Figure 1 shows that neurons that have been more active (darker shade) are more likely to be dropped before k-WTA is applied. Specifically, let $[a_t^l]_j$ denote the activation counter of the neuron j in the layer l after learning t tasks. For the learning task $t+1$, the probability that this neuron is retained is given by:

$$[p_{t+1}^l]_j = exp(\frac{-[a_t^l]_j}{\max_j [a_t^l]_j}\rho) \tag{5}$$

where ρ controls the strength of enforcement of non-overlapping representations, with larger values leading to less overlap. This provides us with an efficient mechanism for controlling the degree of overlap between the representations of different tasks and, hence, the degree of forward transfer and interference based on the task similarities.

Table 3 shows that employing Heterogeneous dropout can further improve the performance of the model. We also analyze the effect of the ρ parameter on the activation counts and the overlap in the representations. Figure 2 shows that Heterogeneous dropout can facilitate the formation of task-specific subnetworks and Figure 3 shows the symmetric KL-divergence between the distribution of activation counts on the test set of Task 1 and Task 2 on the model trained with different ρ values on Perm-MNIST with two tasks. As we increase the ρ parameter, the activations in each layer become increasingly dissimilar. Heterogeneous dropout provides a simple mechanism for balancing the reusability and interference of features depending on the similarity of tasks.

Table 4: Effect of layer-wise dropout ρ on Rot-MNIST with 5 tasks with varying degrees of incremental rotation (θ_{inc}) in each subsequent task. Row 0 shows (ρ^{l_1}, ρ^{l_2}) the ρ values for the first and second hidden layers, respectively.

ρ^{l_1}, ρ^{l_2}	Task Similarity (θ_{inc})					
	2	4	8	16	24	32
0.1, 0.1	$97.60_{\pm 0.12}$	$96.74_{\pm 0.16}$	$91.79_{\pm 0.53}$	$74.99_{\pm 1.16}$	$63.33_{\pm 1.15}$	$57.39_{\pm 2.36}$
0.1, 0.5	$97.77_{\pm 0.08}$	$97.02_{\pm 0.11}$	$92.39_{\pm 0.39}$	$75.56_{\pm 1.46}$	$64.18_{\pm 1.79}$	$57.05_{\pm 2.13}$
0.5, 0.5	$97.88_{\pm 0.12}$	$97.22_{\pm 0.11}$	$92.74_{\pm 0.38}$	$76.73_{\pm 1.01}$	$64.18_{\pm 1.42}$	$58.35_{\pm 0.73}$
0.5, 1.0	$97.88_{\pm 0.04}$	$\mathbf{97.25}_{\pm 0.11}$	$92.87_{\pm 0.39}$	$76.87_{\pm 0.39}$	$64.84_{\pm 0.65}$	$59.40_{\pm 2.31}$
1.0, 1.0	$\mathbf{97.89}_{\pm 0.09}$	$97.19_{\pm 0.09}$	$93.42_{\pm 0.25}$	$77.48_{\pm 0.94}$	$66.33_{\pm 1.62}$	$61.35_{\pm 1.90}$
1.0, 2.0	$97.68_{\pm 0.09}$	$97.00_{\pm 0.23}$	$93.46_{\pm 0.78}$	$79.07_{\pm 0.67}$	$68.20_{\pm 2.34}$	$63.08_{\pm 0.86}$
2.0, 2.0	$97.42_{\pm 0.17}$	$97.00_{\pm 0.11}$	$\mathbf{93.53}_{\pm 0.53}$	$80.03_{\pm 0.62}$	$69.99_{\pm 1.97}$	$65.74_{\pm 1.21}$
2.0, 5.0	$97.39_{\pm 0.03}$	$96.54_{\pm 0.15}$	$92.95_{\pm 0.01}$	$\mathbf{80.55}_{\pm 0.89}$	$\mathbf{73.74}_{\pm 0.21}$	$69.46_{\pm 2.66}$
5.0, 5.0	$96.86_{\pm 0.11}$	$96.12_{\pm 0.08}$	$92.33_{\pm 0.18}$	$79.53_{\pm 0.42}$	$72.47_{\pm 1.23}$	$\mathbf{70.77}_{\pm 2.11}$
w/o Dropout	$97.72_{\pm 0.29}$	$96.93_{\pm 0.40}$	$92.31_{\pm 0.56}$	$75.67_{\pm 1.40}$	$63.68_{\pm 1.36}$	$56.49_{\pm 2.86}$

4.4 Layerwise Heterogeneous Dropout and Task Similarity

For an effective CL agent, it is important to maintain a balance between forward transfer and interference across tasks. As the earlier layers learn general features, a higher portion of the features can be reused to learn the new task, which can facilitate forward transfer, whereas the later layers learn more task-specific features, which can cause interference. Heterogeneous dropout provides us with an efficient mechanism for controlling the degree of overlap between the activations, and hence the features of each layer. Here, we investigate whether having different levels of sparsity (controlled with the ρ parameter) in different layers can further improve performance. As the earlier layers learn general features, having higher overlap (smaller ρ) between the set of active neurons can enable higher resuability and forward transfer. For the later layers, a lesser overlap between the activations (higher ρ) can reduce interference between task-specific features.

To study the effect of Heterogeneous dropout in relation to task similarity, we vary the incremental rotation, θ_{inc}, in each subsequent task for Rot-MNIST setting with 5 tasks. The rotation of task τ is given by $(\tau - 1)\theta_{inc}$. Table 4.4 shows the performance of the model for different layerwise ρ values. Generally, heterogeneous dropout consistently improves the performance of the model, especially when the task similarity is low. For $\theta_{inc} = 32$, it provides $\sim$25% improvement. As task similarity decreases (θ_{inc} increases), higher values of ρ are more effective. Furthermore, we see that having different ρ values for each layer can provide additional gains in performance.

4.5 Hebbian Learning Strengthens Context Gating

For a biologically plausible ANN, it is important to incorporate not only the design elements of biological neurons but also the learning mechanisms it employs. Lifetime plasticity in the brain generally follows the Hebbian principle: a neuron that consistently contributes to the firing of another neuron will build a stronger connection to that neuron (Hebb, 2005).

Therefore, we follow the approach in Flesch et al. (2023) to complement error-based learning with the Hebbian update to strengthen the connections between contextual information and dendritic segments (Figure 1(b)). Each supervised parameter update with backpropagation is followed by a Hebbian update step on the dendritic segments to strengthen the connections between the context input and the corresponding dendritic segment, which is activated. To constrain the parameters, we use Oja's rule, which adds weight decay to Hebbian learning (Oja, 1982),

$$u_\kappa \leftarrow u_\kappa + \eta_h d(c - du_\kappa) \tag{6}$$

where η_h is the learning rate, κ is the index of the winning dendrite with weight u_κ and the modulating signal $d = u_\kappa^T c$ for the context signal c. Figure 4 shows that the Hebbian update step increases the magnitude of

the modulating signal from the dendrites on the feedforward activity, which can further strengthen context-dependent gating and facilitate the formation of task-specific subnetworks. Table 1 shows that this results in a consistent improvement in performance.

4.6 Synaptic Consolidation Further Mitigates Forgetting

In addition to their integrative properties, dendrites also play a key role in retaining information and providing protection against erasure (Cichon & Gan, 2015; Yang et al., 2009). New spines that are formed on different sets of dendritic branches in response to learning different tasks are protected from being eliminated through mediation of synaptic plasticity and structural changes that persist when learning a new task (Yang et al., 2009).

We employ synaptic consolidation by incorporating *Synaptic Intelligence* (Zenke et al., 2017) (details in Appendix) which maintains an importance estimate of each synapse in an online manner during training and subsequently reduces the plasticity of synapses which are considered important for learned tasks. In particular, we adjust the importance estimate to account for the disparity in the degree to which updates to different parameters affect the output of the layer due to the inhibitory interneuron architecture in the DANN layers (Cornford et al., 2020). The importance estimates of the excitatory connections to the inhibitory units and the intra-layer inhibitory connections are upscaled to further penalize changes to these weights. Table 1 shows that employing Synaptic Intelligence (+SC) in this manner further mitigates forgetting. Particularly for Rot-MNIST with 20 tasks, it provides considerable performance improvement.

4.7 Experience Replay is Essential for Enabling CL in Challenging Scenarios

Replay of past neural activation patterns in the brain is considered to play a critical role in memory formation, consolidation, and retrieval (Walker & Stickgold, 2004; McClelland et al., 1995). The replay mechanism in the hippocampus (Kumaran et al., 2016) has inspired a series of rehearsal-based approaches (Li & Hoiem, 2017; Chaudhry et al., 2019; Lopez-Paz & Ranzato, 2017; Arani et al., 2022) that have been proven to be effective in challenging continual learning scenarios (Farquhar & Gal, 2018; Hadsell et al., 2020). Therefore, to replay samples from previous tasks, we utilize a small episodic memory buffer that is maintained through *Reservoir sampling* (Vitter, 1985). It attempts to approximately match the distribution of the incoming stream by assigning equal probabilities to each new sample to be represented in the buffer. During training, samples from the current task, $(x_b, y_b) \sim \mathcal{D}_\tau$, are interleaved with memory buffer samples, $(x_m, y_m) \sim \mathcal{M}$ to approximate the joint distribution of tasks seen so far. Furthermore, to mimic the replay of the activation patterns that accompanied the learning event in the brain, we also save the output logits, z_m, across the training trajectory and enforce consistency loss when replaying samples from episodic memory. Concretely, the loss is given by:

$$\mathcal{L} = \mathcal{L}_{cls}(f(x_b;\theta), y_b) + \alpha\mathcal{L}_{cls}(f(x_m;\theta), y_m) + \beta(f(x_m;\theta) - z_m)^2 \tag{7}$$

where $f(.;\theta)$ is the model parameterized by θ, $\mathcal{L}_{cls}$ is the standard cross-entropy loss, and α and β controls the strength of interleaved training and the consistency constraint, respectively.

Table 1 shows that experience replay (+ER) complements context-dependent information processing and enables the model to learn the joint distribution well in varying challenging settings. In particular, the failure of the model to avoid forgetting in the Class-IL setting (Seq-MNIST) without experience replay suggests that context-dependent processing of information alone does not suffice, and experience replay might be essential. Adding consistency regularization (+CR) further improves performance as the model receives additional relational information about the structural similarity of classes, which facilitates the consolidation of information.

4.8 Combining the individual components

Having shown the individual effect of each of the brain-inspired components in the biologically plausible framework, here we look at their combined effect. The resultant model is referred to as Bio-ANN. Table 1 shows that the different components complement each other and consistently improve the performance of

the model. Our empirical results suggest that employing multiple complementary components and learning mechanisms, similar to the brain, may be an effective approach to enable continual learning in ANNs.

5 Discussion

Continual learning is a hallmark of intelligence, and the human brain constitutes the most efficient learning agent capable of CL. Therefore, incorporating the different components and mechanisms employed in the brain and studying their interactions can provide valuable insights for the design of ANNs suitable for CL. While there are several studies that are inspired by the brain, they focus primarily on one aspect. Since the brain employs all these different components in tandem, it stands to reason that their interactions, or complementary nature, are what enable effective learning instead of one component alone. Furthermore, the underlying framework within which these components are employed and the learning mechanisms might also be critical. The effort to close the gap between current AI and human intelligence could benefit from our enhanced understanding of the brain and incorporating similar mechanisms in ANNs. This is the fundamental question we aimed to study and bring to the attention of the research community at large.

We conducted a study on the effect of different brain-inspired mechanisms under a biologically plausible framework in the CL setting. The underlying model incorporates several key components of the design principles and learning mechanisms in the brain: each layer constitutes separate populations of exclusively excitatory and inhibitory units, which adheres to Dale's principle, and the excitatory pyramidal neurons are augmented with dendritic segments for context-dependent processing of information. We first showed that, equipped with the integrative properties of dendrites, the feedforward network adhering to Dale's principle not only performs as well as standard ANNs, but also provides gains. Then we studied the individual role of different components. We showed that controlling the sparsity in activations using k-WTA activations and Heterogeneous dropout mechanism that encourages the model to use a different set of neurons for each task is an effective approach for maintaining a balance between reusability of features and interference, which is critical for enabling CL. We further showed that complementing error-based learning with the "fire together, wire together" learning paradigm can further strengthen the association between the context signal and dendritic segments that process them and facilitate context-dependent gating. To further mitigate forgetting, we incorporated synaptic consolidation in conjunction with experience replay and showed their effectiveness in challenging CL settings. Finally, the combined effect of these components suggests that, similar to the brain, employing multiple complementary mechanisms in a biologically plausible architecture is an effective approach to enable CL in ANN. It also provides a framework for further study of the role of inhibition in mitigating catastrophic forgetting.

However, there are several limitations and potential avenues for future research. In particular, as dendritic segments provide an effective mechanism for studying the effect of encoding different information in the context signal, they provide an interesting research avenue as to what information is useful for the sequential learning of tasks and the effect of different context signals. Neuroscience studies suggest that multiple brain regions are involved in processing a stimulus and, while there is evidence that active dendritic segments receive contextual information that is different from the input received by the proximal segments, it is unclear what information is encoded in the contextual information and how it is extracted. Here, we used the context signal as in (Iyer et al., 2022), which aims to encode the identity of the task by taking the average input image of all the samples in the task. Although this approach empirically works well in the Perm-MNIST setting, it is important to consider its utility and limitations under different CL settings. Given the specific design of Perm-MNIST, binary-centered digits, and the independent nature of the permutations in each task, the average input image can provide a good approximation of the applied permutation, and hence efficiently encode the task identity. However, this is not straightforward for Rot-MNIST where the task similarities are higher and even more challenging for natural images, where averaging the input image does not provide a meaningful signal. More importantly, it does not seem biologically plausible to encode task information alone as the context signal and ignore the similarity of classes occurring in different tasks. For instance, it seems more reasonable to process slight rotations of the same digits similarly (as in Rot-MNIST) rather than processing them through different subnetworks. Ideally, we would like the context signal for different rotations of a digit to be highly similar. It is, however, quite challenging to design context signals that can capture a wide range of complexities in sequential learning of tasks. Furthermore, instead of

hand engineering the context signal to bias learning towards certain types of task, an effective approach for learning the context signal in an end-to-end training framework is an interesting direction for future search.

In general, our study presents a compelling case for incorporating the design principles and learning machinery of the brain into ANNs and provides credence to the argument that distilling the details of the learning machinery of the brain can bring us closer to human intelligence (Hassabis et al., 2017; Hayes et al., 2021). Furthermore, deep learning is increasingly being used in neuroscience research to model and analyze brain data (Richards et al., 2019). The utility of the model for such research depends on two critical aspects: the performance of the model and how close the architecture is to the brain (Cornford et al., 2020; Schrimpf et al., 2020). The biologically plausible framework in our study incorporates several design components and learning mechanisms of the brain and performs well in a (continual learning) task that is closer to human learning. Therefore, we believe that this work may also be useful for the neuroscience community in evaluating and guiding computational neuroscience. Studying the properties of ANNs with higher similarity to the brain may provide insight into the mechanisms of brain functions. We believe that the fields of artificial intelligence and neuroscience are intricately intertwined, and progress in one can drive the other as well.

References

Ali Abbasi, Parsa Nooralinejad, Vladimir Braverman, Hamed Pirsiavash, and Soheil Kolouri. Sparsity and heterogeneous dropout for continual learning in the null space of neural activations. In *Conference on Lifelong Learning Agents*, pp. 617–628. PMLR, 2022.

Subutai Ahmad and Luiz Scheinkman. How can we be so dense? the benefits of using highly sparse representations. *arXiv preprint arXiv:1903.11257*, 2019.

Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. Selfless sequential learning. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkxbrn0cYX`.

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=uxxFrDwrE7Y`.

Helen C Barron, Tim P Vogels, Timothy E Behrens, and Mani Ramaswami. Inhibitory engrams in perception and memory. *Proceedings of the National Academy of Sciences*, 114(26):6666–6674, 2017.

Alison L Barth and James FA Poulet. Experimental evidence for sparse firing in the neocortex. *Trends in neurosciences*, 35(6):345–355, 2012.

John M Bekkers. Pyramidal neurons. *Current biology*, 21(24):R975, 2011.

Marcus K Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature neuroscience*, 19(12):1697–1706, 2016.

Martin Boerlin, Christian K Machens, and Sophie Denève. Predictive coding of dynamical variables in balanced spiking networks. *PLoS computational biology*, 9(11):e1003258, 2013.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Hkf2_sC5FX`.

Joseph Cichon and Wen-Biao Gan. Branch-specific dendritic ca 2+ spikes cause persistent synaptic plasticity. *Nature*, 520(7546):180–185, 2015.

Jonathan Cornford, Damjan Kalajdzievski, Marco Leite, Amélie Lamarquette, Dimitri Michael Kullmann, and Blake Aaron Richards. Learning to live with dale's principle: Anns with separate excitatory and inhibitory units. In *International Conference on Learning Representations*, 2020.

Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

Timo Flesch, David G Nagy, Andrew Saxe, and Christopher Summerfield. Modelling continual learning in humans with hebbian context gating and exponentially decaying task signals. *PLOS Computational Biology*, 19(1):e1010808, 2023.

Peter Foldiak. Sparse coding in the primate cortex. *The handbook of brain theory and neural networks*, 2003.

Kunihiko Fukushima. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 36:193–202, 1980.

Joaquin Fuster. *The prefrontal cortex*. Academic press, 2015.

Daniel J Graham and David J Field. Sparse coding in the neocortex. *Evolution of nervous systems*, 3: 181–187, 2006.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pp. 195–201. Springer, 1995.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Tyler L Hayes, Giri P Krishnan, Maxim Bazhenov, Hava T Siegelmann, Terrence J Sejnowski, and Christopher Kanan. Replay in deep learning: Current approaches and missing biological elements. *Neural Computation*, 33(11):2908–2950, 2021.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

Carl Holmgren, Tibor Harkany, Björn Svennenfors, and Yuri Zilberter. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *The Journal of physiology*, 551(1):139–153, 2003.

Kevin Lee Hunter, Lawrence Spracklen, and Subutai Ahmad. Two sparsities are better than one: Unlocking the performance benefits of sparse-sparse networks. *arXiv preprint arXiv:2112.13896*, 2021.

Alessandro Ingrosso and LF Abbott. Training dynamically balanced excitatory-inhibitory networks. *PloS one*, 14(8):e0220547, 2019.

Abhiram Iyer, Karan Grewal, Akash Velu, Lucas Oliveira Souza, Jeremy Forest, and Subutai Ahmad. Avoiding catastrophe: Active dendrites enable multi-task learning in dynamic environments. *Frontiers in neurorobotics*, 16, 2022.

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, 2017.

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.

Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.

Louis Lapique. Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization. *Journal of Physiology and Patholoigy*, 9:620–635, 1907.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in neural information processing systems*, pp. 6467–6476, 2017.

Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 67–82, 2018.

Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15 (3):267–273, 1982.

Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2020.

Markus Siegel, Timothy J Buschman, and Earl K Miller. Cortical information flow during flexible sensorimotor decisions. *Science*, 348(6241):1352–1355, 2015.

Piergiorgio Strata, Robin Harvey, et al. Dale's principle. *Brain research bulletin*, 50(5):349–350, 1999.

David J Tolhurst, Darragh Smyth, and Ian D Thompson. The sparseness of neuronal responses in ferret primary visual cortex. *Journal of Neuroscience*, 29(8):2355–2370, 2009.

Robin Tremblay, Soohyun Lee, and Bernardo Rudy. Gabaergic interneurons in the neocortex: from cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

Carl Van Vreeswijk and Haim Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996.

Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985.

Matthew P Walker and Robert Stickgold. Sleep-dependent learning and memory consolidation. *Neuron*, 44 (1):121–133, 2004.

Guang Yang, Feng Pan, and Wen-Biao Gan. Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924, 2009.

Guang Yang, Cora Sau Wan Lai, Joseph Cichon, Lei Ma, Wei Li, and Wen-Biao Gan. Sleep promotes branch-specific formation of dendritic spines after learning. *Science*, 344(6188):1173–1178, 2014.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

A Appendix

A.1 Related Work - Biological Inspired AI

The human brain has long been a source of inspiration for ANNs design (Hassabis et al., 2017; Kudithipudi et al., 2022). However, we have failed to take full advantage of our enhanced understanding of the brain, and there are fundamental differences between the design principles and learning mechanisms employed in the brain and ANNs. These differences may account for the huge gap in performance and behavior.

From an architecture design perspective, standard ANNs are predominantly based on the point neuron model (Lapique, 1907) which is an outdated and oversimplified model of biological computations that lacks sophisticated and context-dependent processing in the brain. Furthermore, neurons in standard ANNs lack adherence to Dale's principle (Strata et al., 1999) to which most neurons in the brain adhere. Unlike the brain where presynaptic neurons have an exclusively excitatory or inhibitory impact on their postsynaptic partners, neurons in standard ANNs contain both positive and negative output weights, and signs can change while learning. These constitute two of the major fundamental differences in the underlying design principle of ANNs and the brain.

Two recent studies attempt to address this gap. Cornford et al. (2020) incorporated Dale's principle into ANNs (DANNs) in a more biologically plausible manner and show that with certain initialization and regularization considerations, DANNs can perform comparable to standard ANNs in object recognition task, which earlier attempts failed to do so. Our study extends DANNs to the more challenging CL setting and shows that accounting for the discrepancy in the effect of weight changes in excitatory and inhibitory neurons can further reduce forgetting in CL. Iyer et al. (2022) propose an alternative to the point neuron model and provide an algorithmic abstraction of pyramidal neurons in the neocortex. Each neuron is augmented with dendritic segments which receive an additional context signal and the output of the dendrite segment modulates the activity of the neuron, allowing context-dependent processing of information. Our study builds upon their work and provides a biologically plausible architecture characterized with both adherence to Dale's principle and the context-dependent processing of pyramidal neurons. This provides us with a framework to study the role of brain-inspired mechanisms and allows us to study the role of inhibition in the challenging continual learning setting, which is closer to human learning.

From a learning perspective, several approaches have been inspired by the brain, particularly for CL (Kudithipudi et al., 2022).The replay mechanism in the hippocampus (Kumaran et al., 2016) has inspired a series of rehearsal-based approaches (Hayes et al., 2021; Hadsell et al., 2020) which have proven to be effective in challenging continual learning scenarios (Farquhar & Gal, 2018). Another popular approach for continual learning, regularization-based approaches Zenke et al. (2017); Kirkpatrick et al. (2017), has been inspired by neurobiological models that suggest that CL in the neocortex is based on a task-specific synaptic consolidation process that involves rendering a proportion of synapses less plastic and, therefore, stable over long timescales (Benna & Fusi, 2016; Yang et al., 2009). While both approaches are inspired by the brain, researches have mostly discounted the fact that the brain employs both of them in conjunction to consolidate information rather than in isolation. Therefore, the research in both of these methods has been orthogonal. Furthermore, they have been applied on top of standard ANNs which are not representative of the complexities of the neuron in the brain. Our study employs replay and synaptic consolidation together in a more biologically plausible architecture and shows that they complement each other to improve performance.

Furthermore, our framework employs several techniques to mimic the characteristics of activations in the brain. As yyramidal neurons in the neocortex have highly sparse connectivity to each other (Hunter et al., 2021; Holmgren et al., 2003) and only a small percentage ($<2\%$) of neurons are active for given stimuli neurons (Barth & Poulet, 2012), we apply k-winner-take-all (k-WTA) activations (Ahmad & Scheinkman, 2019) to mimic activation sparsity. Several studies have shown the benefits of sparsity in CL (Abbasi et al., 2022; Mallya et al., 2018; Aljundi et al., 2019), they do not consider that the brain not only utilizes sparsity, it does so in an efficient manner to encode information. Information in the brain is encoded by the strong activation of a relatively small set of neurons, forming sparse coding. A different subset of neurons is utilized to represent different types of stimuli (Foldiak, 2003) and semantically similar stimuli activate a similar set of neurons. Heterogeneous dropout (Abbasi et al., 2022) coupled with k-WTA activations aims to mimic these

characteristics by encouraging the new task to utilize a new set of neurons for learning. Finally, we argue that it is important not only to incorporate the design elements of biological neurons but also the learning mechanisms that they employ. Lifetime plasticity in the brain generally follows the Hebbian principle (Hebb, 2005). Therefore, we follow the approach in Flesch et al. (2023) to complement error-based learning with Hebbian update to strengthen the connections between contextual information and dendritic segments and show that it strengthens context gating.

Our study provides a biologically plausible framework with the underlying architecture with the context-dependent processing of information and adherence to Dale's principle. Additionally, it employs the learning mechanisms (experience replay, synaptic consolidation, and Hebbian update) and characteristics (sparse non-overlapping activations and task-specific subnetworks) of the brain. To the best of our knowledge, we are the first to provide a comprehensive study of the integration of different brain-inspired mechanisms in a biologically plausible architecture in a CL setting.

A.2 Additional Results

Additionally, we conducted experiments on Fashion-MNIST, which is more challenging than the MNIST datasets. We considered both the Class-IL (Seq-FMNIST) and Domain-IL (Rot-FMNIST) setting. Seq-FMNIST divides the classification into 5 tasks with 2 classes each, while Rot-FMNIST involves classifying the 10 classes in each task with the samples rotated by i.e. {0, 8, ..., (N-1)*8} for task $\{\tau_1, \tau_2, .., \tau_N\}$.

For brevity, we refer to Active Dendrites + Dale's principle as ActiveDANN. To show the effect of different components better (ActiveDANN without ER fails in the class-IL setting), we consider ActiveDann + ER as the baseline upon which we add the other components. Empirical results in Table A.2 show that the findings on MNIST settings also translate to Fashion-MNIST and each component leads to performance improvement.

Table 5: Effect of each component of the biologically plausible framework on different Seq-FMNIST and Rot-FMNIST. For all experiments, we use a memory budget of 500 samples. HD refers to heterogeneous dropout. We provide the average task performance and 1 std of 5 runs.

Method	Seq-FMNIST	Rot-FMNIST
Joint	$94.33_{\pm 0.51}$	$98.15_{\pm 0.09}$
SGD	$19.83_{\pm 0.04}$	$51.89_{\pm 0.27}$
ActiveDANN + ER	$77.56_{\pm 0.27}$	$80.99_{\pm 0.53}$
+ Hebb	$78.02_{\pm 0.38}$	$82.16_{\pm 0.26}$
+ SC	$78.05_{\pm 0.61}$	$82.55_{\pm 0.37}$
+ HD	$78.74_{\pm 0.38}$	$83.97_{\pm 0.46}$
Bio-ANN	$\mathbf{79.28}_{\pm 0.42}$	$\mathbf{89.22}_{\pm 0.21}$

A.3 Experimental Setup

To study the role of the different components inspired by the brain in a biologically plausible NN for CL and to gauge their roles in the performance and characteristics of the model, we conduct all our experiments under uniform settings. Unless otherwise stated, we use a multi-layer perception (MLP) with two hidden layers with 2048 units and k-WTA activations. Each neuron is augmented with N dendritic segments where N corresponds to the number of tasks and the dimensions correspond to the dimensions of the context vector that correspond to the input image size (784 for all MNIST-based settings). The model is trained using an SGD optimizer with 0.3 learning rate and a batch size of 128 for 3 epochs on each task. We set the weight sparsity to 0 and set the percentage of active neurons to 5%. For our experiments involving Dale's principle, we maintain the same number of total units in each layer divided into 1844 excitatory and 204 inhibitory units. Only the excitatory units are augmented with dendritic segments. Importantly, we use the initialization strategy and corrections for the SGD update as posited in Cornford et al. (2020) to account for the disparities in the degree to which updates to different parameters affect

the layer output distribution. Updates to inhibitory unit parameters are scaled down relative to update of excitatory parameters. Concretely, the gradient updates to W_{ie} were scaled by $\sqrt{n_e}^{-1}$ and W_{ei} by d^{-1}, where n_e are the number of excitatory neurons in the layer and d is the input dimension of the layer. Furthermore, to select the hyperparameters for different settings, we use a small validation set. Note that, as the goal was not to achieve the best possible accuracy, but rather to show the effect of each component, we did not conduct an extensive hyperparameter search. Table A.3 provides the selected hyperparameters for the effect of individual component experiments in Table 1 and Table A.3 provides the selected hyperparameter for Bio-ANN experiments. We report the mean accuracy over all tasks and 1 std over three different random seeds.

Table 6: The selected hyperparamneters for the experiments showing the individual effect of each component (Table 1). The base learning rate for all the experiments is 0.3 and the individual components use the same learning rates for W_{ie} and W_{ei} as (+ Dale's principle). For + SC experiments, we use $\lambda_{W_{ie}}$=10 and $\lambda_{W_{ei}}$=10. For ER experiments, we use a memory budget of 500 samples.

Dataset	#Tasks	+ Dale's Principle	+ Hebbian Update	+ SC	+ ER	+ ER + CR
Rot-MNIST	5	$\eta_{W_{ie}}$=3e-2, $\eta_{W_{ei}}$=3e-3	n_h=3e-10	λ=0.25	α=1, β=0	α=1, β=0.50
	10	$\eta_{W_{ie}}$=3e-2, $\eta_{W_{ei}}$=3e-3	n_h=3e-08	λ=0.25	α=1, β=0	α=1, β=0.50
	20	$\eta_{W_{ie}}$=3e-3, $\eta_{W_{ei}}$=3e-4	n_h=3e-10	λ=1.00	α=1, β=0	α=1, β=0.50
Perm-MNIST	5	$\eta_{W_{ie}}$=3e-2, $\eta_{W_{ei}}$=3e-2	n_h=3e-09	λ=0.10	α=1, β=0	α=1, β=0.50
	10	$\eta_{W_{ie}}$=3e-2, $\eta_{W_{ei}}$=3e-2	n_h=3e-06	λ=0.25	α=1, β=0	α=1, β=0.50
	20	$\eta_{W_{ie}}$=3e-2, $\eta_{W_{ei}}$=3e-3	n_h=3e-09	λ=0.10	α=1, β=0	α=1, β=0.50
Seq-MNIST	5	$\eta_{W_{ie}}$=3e-2, $\eta_{W_{ei}}$=3e-3	n_h=3e-07	λ=0.25	α=1, β=0	α=1, β=0.25

Table 7: The selected hyperparamneters for Bio-ANN experiments in Table 1. We use the same learning rate for each setting as + Dale's principle (Table A.3).

Dataset	#Tasks	n_h	λ	ρ	α	β
Rot-MNIST	5	3e-8	0.25	0.1	1	0.5
	10	3e-8	0.1	0.3	1	0.5
	20	3e-8	0.1	0.3	1	0.5
Perm-MNIST	5	3e-6	0.1	0.1	1	0.5
	10	3e-8	0.1	0.3	1	0.5
	20	3e-8	0.1	0.3	1	0.5
Seq-MNIST	5	3e-6	0.1	0.1	1	0.25

A.4 Effect of adjusting for the inhibitory weights

The inhibitory interneuron architecture of DANN layers introduces disparities in the degree to which updates to different parameters affect the layer's output distribution, e.g. if a single element of W_{ie} is updated, this has an effect on each element of the layer's output. An inhibitory weight update of δ to W_{ie} changes the model distribution approximately n_e times more than an excitatory weight update of δ to W_{ee} (Cornford et al., 2020). The effect of these disparities would be even more pronounced in a CL setting as large changes to the output distribution when learning a new task can cause more forgetting of previous tasks. To account for these, we further reduce the learning rate of W_ie and W_ei after learning the first task. Table A.4 shows that accounting for the higher effect of inhibitory neurons can further improve the performance of the model in majority of the settings. It would be interesting to explore better approaches to account for the aforementioned disparities, which are tailored for CL and consider the effect on forgetting.

Table 8: Effect of adjusting the learning rates of W_{ie} and W_{ei} at the end of the first task on different datasets with a varying number of tasks.

η_{Wie}	η_{Wei}	Rot-MNIST			Perm-MNIST		
		5	10	20	5	10	20
3e-1	3e-1	92.12±0.34	70.86±0.44	46.30±1.03	95.78±0.19	94.73±0.36	**92.67**±0.61
3e-2	3e-2	92.23±0.53	70.23±1.12	47.53±1.79	95.77±0.33	**95.06**±0.29	91.63±0.39
3e-2	3e-3	92.28±0.27	70.78±0.23	47.32±1.43	95.68±0.14	94.96±0.49	92.40±0.38
3e-3	3e-3	**92.34**±0.51	**71.27**±1.69	47.81±1.10	95.70±0.29	94.44±0.70	92.02±0.19
3e-3	3e-4	92.03±0.09	70.79±1.75	**48.79**±0.27	**95.90**±0.13	94.19±0.47	91.04±0.34

A.5 Effect of adjusting for scaling the SI importance estimate for inhibitory weights

Similar to adjusting the learning rate of inhibitory weights, we check whether scaling up the importance estimate of inhibitory neurons can further improve the effectiveness of synaptic consolidation in reducing forgetting. Table A.5 shows that scaling the importance estimate in accordance with the degree to which inhibitory weights affect the output distribution, and hence forgetting, further improves the performance in majority of cases, especially for a lower number of tasks. This suggests that regularization methods designed specifically for networks with inhibitory neurons is a promising research direction.

Table 9: Effect of scaling the importance estimate for W_{ie} and W_{ei} to reduce the parameter shift in the inhibitory weights on Rot-MNIST and Perm-MNIST datasets with varying number of tasks.

λ	$\lambda_{W_{ie}}$	$\lambda_{W_{ei}}$	Rot-MNIST			Perm-MNIST		
			5	10	20	5	10	20
0.1	1	1	92.92±0.22	75.53±1.46	52.16±1.17	96.59±0.17	96.20±0.13	**95.69**±0.10
	10	10	92.77±0.27	74.70±1.05	51.94±1.05	96.57±0.34	96.18±0.18	95.64±0.15
	10	100	92.50±0.65	74.67±1.27	52.55±1.12	**96.67**±0.23	96.26±0.26	95.61±0.10
0.25	1	1	92.77±0.69	**76.30**±0.77	55.05±2.47	96.62±0.28	96.07±0.02	95.36±0.24
	10	10	**93.54**±0.79	75.44±0.81	54.93±2.35	96.54±0.22	96.23±0.16	95.03±0.13
	10	100	93.40±0.86	75.87±1.35	55.06±1.77	96.65±0.25	**96.36**±0.10	95.18±0.29
0.5	1	1	93.23±0.71	75.24±0.62	**60.24**±2.02	96.22±0.37	95.97±0.06	92.94±0.85
	10	10	93.13±0.24	75.95±0.38	59.35±1.53	96.24±0.51	95.81±0.10	92.86±0.87
	10	100	92.85±0.42	75.85±1.18	58.94±2.15	96.36±0.42	96.02±0.27	93.48±0.57
w/o SC			92.28±0.27	70.78±0.23	48.79±0.27	95.77±0.33	95.06±0.29	92.40±0.38

A.6 Effect of Sparsity

To further study the effect of different levels of sparsity in activations and connections, we vary the number of weights randomly set to zero at initialization ($S_W \in \{0, 0.25, 0.50\}$) and the ratio of active neurons (k^{l1}, k^{l2}) in each hidden layer. Table A.6 shows that sparsity in activation plays a critical role in enabling CL in ANNs. Interestingly, sparsity in connections play a considerable role in Perm-MNIST with higher levels of active neurons (≥ 0.2). Furthermore, exploring finer differences in activation sparsity of different layers may further improve the performance. Similar to heterogeneous dropout, we show the effect of activation sparsity in relation to task similarity in Table A.6. Similar tasks (lower θ_{inc}) benefit from a higher number of active neurons that can increase forward transfer, while dissimilar tasks (higher θ_{inc}) perform better with higher activation sparsity that can reduce the overlap in representations.

Table 10: Effect of different levels of sparsity in activations (ratio of active neurons, (k^{l1}, k^{l2}) in 1st and 2nd hidden layer respectively) and connections (ratio of zero weights, S_W) on Rot-MNIST and Perm-MNNIST with increasing number of tasks. The best performance across the different sparsity levels for each task is in bold.

	#Tasks	S_W	Activation Sparsity (k^{l1}, k^{l2})						
			(0.05, 0.05)	(0.1, 0.05)	(0.1, 0.1)	(0.2, 0.1)	(0.2, 0.2)	(0.5, 0.2)	(0.5, 0.5)
Rot-MNIST	5	0.00	92.28±0.27	92.63±0.46	92.26±0.31	92.25±0.71	**92.79**±0.44	92.51±0.50	92.26±0.65
		0.25	91.22±0.58	91.50±0.36	92.33±0.26	91.75±0.30	92.60±0.19	91.23±1.63	91.66±0.46
		0.50	90.78±0.11	91.15±0.34	91.25±0.14	91.25±0.92	90.67±0.73	90.75±0.86	90.41±1.48
	10	0.00	70.78±0.23	71.16±0.51	71.95±1.54	72.22±0.63	**73.32**±0.69	71.89±0.62	71.61±0.76
		0.25	70.23±0.76	71.42±0.98	71.84±0.72	73.22±1.34	72.58±0.64	72.23±0.71	71.23±0.84
		0.50	69.61±0.50	70.59±0.62	70.94±0.71	72.05±0.34	72.25±1.40	71.37±0.65	70.85±0.67
	20	0.00	**48.79**±0.27	48.01±0.58	47.96±1.84	48.33±1.23	48.65±0.91	48.19±0.14	47.71±0.91
		0.25	47.72±0.83	48.61±0.30	48.41±0.84	48.53±1.77	48.30±0.87	48.29±1.59	47.11±0.44
		0.50	46.20±0.26	47.15±1.37	48.02±1.10	48.17±1.42	48.30±1.41	47.66±1.53	47.73±0.51
Perm-MNIST	5	0.00	95.77±0.33	95.55±0.27	**96.43**±0.10	95.85±0.29	90.29±6.07	88.18±8.86	74.51±13.55
		0.25	95.45±0.25	95.14±0.27	95.65±0.22	95.75±0.32	93.73±1.29	87.49±2.33	75.97±5.61
		0.50	93.95±0.65	94.19±0.41	94.90±0.22	94.22±1.19	94.05±0.81	91.02±3.71	83.94±9.58
	10	0.00	**95.06**±0.29	94.08±0.95	94.38±0.73	89.54±2.27	78.91±5.26	76.44±7.93	35.86±2.04
		0.25	94.51±0.12	93.52±0.01	93.62±0.64	88.58±2.17	84.94±3.58	69.32±6.24	59.53±9.57
		0.50	92.12±0.62	90.31±2.01	92.29±0.73	88.48±0.22	82.57±2.40	74.42±2.84	69.31±5.66
	20	0.00	**92.40**±0.38	89.41±1.73	84.28±1.35	73.29±2.60	63.84±3.45	58.85±5.00	20.80±0.99
		0.25	90.75±0.63	89.22±0.94	84.29±2.17	75.82±6.68	67.26±1.22	63.96±1.19	41.60±2.64
		0.50	87.31±0.82	84.90±1.99	83.72±0.63	69.62±5.98	66.27±2.74	64.42±2.33	51.92±3.79

Table 11: Effect of different levels of activation sparsity on Rot-MNIST with 5 tasks with varying degrees of incremental rotation (θ_{inc}) in each subsequent task. Row 0 shows (k^{l_1}, k^{l_2}) the ratio of active neurons in the 1st and 2nd hidden layers, respectively.

(k^{l_1}, k^{l_2})	Task Similarity (θ_{inc})					
	2	4	8	16	24	32
0.05, 0.05	97.54±0.06	96.56±0.14	91.95±0.54	75.13±0.83	63.14±0.52	57.01±0.89
0.1, 0.05	97.57±0.32	96.84±0.18	92.49±0.59	76.15±1.28	64.03±1.12	58.56±1.17
0.1, 0.1	97.81±0.08	96.84±0.30	92.28±0.31	**76.80**±1.20	**64.91**±1.17	**58.62**±1.70
0.2, 0.1	97.44±0.74	96.88±0.45	92.47±0.69	75.79±1.26	64.38±1.58	58.34±1.54
0.2, 0.2	**97.88**±0.11	**97.27**±0.15	**92.79**±0.50	76.22±1.46	64.30±1.38	57.22±1.06
0.5, 0.2	97.67±0.64	96.92±0.78	92.61±0.65	75.66±0.95	63.86±0.61	56.11±1.20
0.5, 0.5	97.67±0.55	97.03±0.53	92.29±0.58	74.55±1.01	62.37±0.62	53.13±3.62

Algorithm 1 Bio-ANN: A biologically plausible framework for CL.

Input: Data stream $\mathcal{D}$; Learning rates η, $\eta_{W_{ie}}$, $\eta_{W_{ei}}$; Hebbian learning rate η_h; Heterogeneous dropout ρ; Synaptic consolidation weights λ, $\lambda_{W_{ie}}$, $\lambda_{W_{ei}}$, γ; Experience replay weights α, β

Initialize:

Model weights θ, Reference weights $\theta_c = \{\}$, Task prototypes $C_\tau = \{\}$

Heterogeneous dropout: Overall activation counts $A_\tau = 0$, Keep probabilities $P_\tau = 1$

Memory buffer $\mathcal{M} \leftarrow \{\}$

Synaptic Intelligence: $\omega = 0$, $\Omega = 0$

▷ Sample task from data stream

for $\mathcal{D}_\tau \in \{\mathcal{D}_1, \mathcal{D}_2, .., \mathcal{D}_T\}$ **do**

▷ Task context

Evaluate context vector (Eq. 2):

$c_\tau = \frac{1}{|\mathcal{D}_\tau|} \sum_{x \in \mathcal{D}_\tau} x$

Update the set of prototypes:

$C_\tau \leftarrow \{C_\tau, c_\tau\}$

▷ Train on task τ

while Training **do**

Sample data: $(x_b, y_b) \sim \mathcal{D}_\tau$ and $(x_m, y_m, z_m) \sim \mathcal{M}$

▷ Task specific loss

Get the model output and activation counts on the current task batch:

$z_b, a_b = F(x_b, c_\tau; \theta, P_\tau)$ # Apply Heterogeneous dropout

Calculate task loss:

$\mathcal{L}_\tau = \mathcal{L}_{cls}(z_b, y_b)$

Update overall activation counts:

$A_\tau \leftarrow \text{UpdateActivationCounts}(a_t)$

▷ Experience replay

Infer context for buffer samples (Eq. 3):

$c_m = \arg\min_{c_\tau} \|\boldsymbol{x}' - \boldsymbol{C}_\tau\|_2$

Get model output on buffer samples:

$z = F(x_m, c_m; \theta)$ # Disable Heterogeneous dropout

Calculate replay loss:

$\mathcal{L}_{er} = \alpha \mathcal{L}_{cls}(z, y_m) + \beta (z - z_m)^2$

▷ Synaptic regularization

Calculate SI loss:

$\mathcal{L}_{sc} = \Omega_{adj}(\theta - \theta_c)^2$

Calculate overall loss and clip the gradient between 0 and 1:

$\mathcal{L} = \mathcal{L}_\tau + \mathcal{L}_{er} + \mathcal{L}_{sc}$

$\nabla_\theta \mathcal{L} = \text{Clip}(\nabla_\theta \mathcal{L}, 0, 1)$

▷ Update Models

SGD update: $\theta = \text{UpdateModel}(\nabla_\theta \mathcal{L}, \eta, \eta_{W_{ie}}, \eta_{W_{ei}})$

Hebbian update on dendritic segments: $U = \text{HebbianStep}(\{c_\tau, c_m\}, U)$

Update small omega: $\omega = \omega + \eta \nabla_\theta^2 \mathcal{L})$ ▷ Update SI parameter

$\mathcal{M} \leftarrow Reservoir(\mathcal{M}, (x_b, y_b, z_b))$ ▷ Update memory buffer (Algorithm 2)

▷ Task Boundary

Update keep Probabilities (Eq 5): $P_\tau = exp(\frac{-A_\tau}{\max A_\tau} \rho)$

Update SI Omega and reference weights and reset small omega:

$\Omega = \Omega + \frac{\omega}{(\theta - \theta_c)^2 + \gamma}$

$\omega = 0$

$\theta_c = \theta$

Scale up importance for inhibitory weights

$\Omega_{adj} = \text{ScaleUpInhib}(\Omega, \lambda_{W_{ie}}, \lambda_{W_{ei}})$

return θ

Algorithm 2 Reservoir Sampling

Input: Memory Buffer $\mathcal{M}$, Buffer Size $\mathcal{B}$, Number of examples seen so far N, data points (x, y, z)

if $\mathcal{B} > N$ **then** ▷ Check if memory is full
  $\mathcal{M}[N] \leftarrow (x, y, z)$
else ▷ Select a sample to replace
  n = SampleRandomInteger(min=0, max=N)
  if $n < \mathcal{B}$ **then**
    $\mathcal{M}[n] \leftarrow (x, y, z)$
return $\mathcal{M}$