Lifelong Neural Predictive Coding: Learning Cumulatively Online without Forgetting

Anonymous Author(s)
Affiliation
Address
email

Abstract

In lifelong learning systems based on artificial neural networks, one of the biggest obstacles is the inability to retain old knowledge as new information is encountered. This phenomenon is known as catastrophic forgetting. In this paper, we propose a new kind of connectionist architecture, the Sequential Neural Coding Network, that is robust to forgetting when learning from streams of data points and, unlike networks of today, does not learn via the popular back-propagation of errors. Grounded in the neurocognitive theory of predictive processing, our model adapts synapses in a biologically-plausible fashion while another neural system learns to direct and control this cortex-like structure, mimicking some of the task-executive control functionality of the basal ganglia. In our experiments, we demonstrate that our self-organizing system experiences significantly less forgetting compared to standard neural models, outperforming a swath of previously proposed methods, including rehearsal/data buffer-based methods, on both standard (SplitMNIST, Split Fashion MNIST, etc.) and custom benchmarks even though it is trained in a stream-like fashion. Our work offers evidence that emulating mechanisms in real neuronal systems, e.g., local learning, lateral competition, can yield new directions and possibilities for tackling the grand challenge of lifelong machine learning.

1 Introduction

Lifelong learning is a part of machine learning and artificial intelligence research with the goal of developing a computational agent that can learn over time and continually accumulate new knowledge while retaining its previously learned experiences [1, 2]. For example, suppose an agent needs to learn to classify digits, then types of clothing, and then sketches of objects. As each new task arrives the agent is expected to process the accompanying data and learn the new task but still remember how to complete the old tasks, at least without significant degradation in performance or loss of generalization. Modern-day connectionist systems are typically trained on a fixed pool of data samples, collected in controlled environments, in isolation and random order, and then evaluated on a separate validation data pool. This is a far cry from what we really desire from learning machines.

When we look to humans or other animals, we see that they are more than capable of learning in a continual manner, making decisions based on sensorimotor input throughout their lifespans [3]. This ability to incrementally acquire and refine knowledge over long periods of time is driven by cognitive processes that come together to create the experience-driven specialization of motor and perceptual skills [4, 3]. Thus, evaluating how neural systems generalize on task sequences, as opposed to single, isolated tasks, proves to be a far greater challenge. In order to continually adapt, the brain must retain specific memories of prior tasks. In working towards the challenge of lifelong machine learning, this paper makes the following contributions: 1) we propose the sequential neural coding network, an interactive generative model that jointly reconstructs input and predicts its label, and an algorithm for

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

updating its synapses, 2) we show that memory retention is vastly improved by integrating our model's multi-step nature with lateral competition that is driven by a task selection function inspired by the basal ganglia [5], and 3) we compare our model's performance against state-of-the-art baselines, including both regularization and rehearsal/replay-based methods, on publicly-available benchmarks.

2 Related Work

It is well-known that when artificial neural networks (ANNs) are trained on more than one task sequentially, the new information contained in subsequent tasks leads to catastrophic interference (a.k.a. catastrophic forgetting) with the information acquired in earlier tasks [6, 7, 8]. This happens in connectionist systems when the new data instances to be learned are significantly different from previously observed ones. This causes the new information to overwrite knowledge currently encoded in the system's synaptic weights, due to the sharing of neural representations over tasks [9, 6] (this is known as the representational overlap problem). In isolated, task-specific (offline) learning (though it still occurs [10]), this type of weight overwriting occurs to a lesser degree because the patterns are presented to the agent pseudo-randomly and multiple times, i.e., via multiple epochs.

Over the decades, there have been many approaches proposed to mitigate catastrophic forgetting in neural systems. Some of the earliest attempts proposed memory systems where prior data points were stored and regularly replayed to the network in a process called "rehearsal", which involved interleaving these data points with samples drawn from new datasets [11, 12, 13, 14, 15]. Though effective, the main drawback of these approaches is that they require explicitly storing old data. Such a mechanism is not known to exist in the brain and, as a practical matter, this leads to exploding working (hardware) memory requirements (inefficiency). In addition, rehearsal-based approaches do not offer any mechanisms to preserve consolidated knowledge in the face of acquiring new information [4]. Other approaches attempt to allocate additional neural "resources", i.e., growing the networks when required [16, 17], motivated by earlier findings [18]. However, this leads to dramatically increasing computational requirements over time as the networks grow larger. To compound these issues further, systems with growing capacity cannot know how many resources to allocate at a given time since the number of tasks and samples are not known a priori (without imposing strong assumptions on the input distribution). Other approaches try to block old information from being overwritten through regularization [19]. From this vast collection of research, each approach bearing strengths and weaknesses, three suggested remedies have emerged: 1) allocate additional neural resources to accommodate new knowledge, 2) use non-overlapping representations (or semi-distributed ones [20]) if resources are fixed, and 3) interleave old patterns with new ones as new information is acquired.

In this work, we consider the setting where the space available to the agent grows slowly compared to the rate of new tasks being presented. This means that we cannot just create a new, separate network for every task observed in the stream. Furthermore, storing, reshuffling, and re-presenting the data in the stream is not feasible in this setting. Concretely, our approach could be classified as class incremental learning (Class-IL) [21] given that we **do not task descriptors/identifiers** at both training and test time.In addition to addressing the above, our contribution to lifelong learning is motivated by premise that developing algorithms [22, 23, 24, 25, 26] that serve as alternatives to back-propagation will lead to the creation of promising architectures with mechanisms better equipped to tackle problems like catastrophic interference.

3 Cumulative Learning with Neural Coding

Notation: We start by defining key notation.$\odot$ indicates a Hadamard product while $\cdot$ denotes matrix/vector multiplication. $(\mathbf{v})^T$ is the transpose of $\mathbf{v}$. Matrices/vectors are depicted in bold-font, e.g., matrix $\mathbf{M}$ or vector $\mathbf{v}$ (scalars shown in italic). Finally, $||\mathbf{v}||_2$ denotes the Euclidean norm of $\mathbf{v}$.

3.1 Sequential Learning and the Data Continuum

This work focuses on learning a neural system in the context of sequential learning. Starting from an early definition [27] of this form of learning, we assume there is a sequence of tasks $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \ldots$ (with potentially no end to the sequence) that are presented to a system in order. When faced with the $(N+1)$th task, the system should use the knowledge it has gained from the previous N tasks to aid in learning and performing the current task. The knowledge of a system is stored in a knowledge

base (KB), e.g., the synapses of a neural model. Each task $\mathcal{T}_i$ has its own corresponding dataset $D_i = \{(\mathbf{y}_1, \mathbf{x}_1, t_i) \dots (\mathbf{y}_{n_i}, \mathbf{x}_{n_i}, t_i)\}$ with n_i examples. Here $\mathbf{x}_j \in \mathcal{R}^{J_x \times 1}$ represents the feature vector of the jth example (J_x is its dimensionality), $\mathbf{y}_j \in \{0,1\}^{J_y \times 1}$ is the target (label), and t_i is the task descriptor that identifies $(\mathbf{y}_j, \mathbf{x}_j)$ as being a data point from task i. The task descriptor is one-hot encoded as $\mathbf{t} \in \{0,1\}^{(N+1)\times 1}$ and, when a new task is encountered, the size of the one-hot vector is increased by one – thus the network adds an extra randomly initialized input node if it accepts t_i as an extra input. Note that while we present the task data continuum with t_i explicitly depicted, we will generalize our models to not depend on t_i (making this problem task descriptor-free).

Dynamic Output Units: The output nodes in our setting get re-used for each new task. For example, output node 1 of the network could represent a prediction for the digit "1" in task $\mathcal{T}_1$ (e.g., digit recognition), while in task $\mathcal{T}_2$ (e.g., clothing recognition) the same node could represent a prediction for "pants". If a new task has more classes than previous tasks, we add output nodes with randomly initialized weights. For example, if prior tasks were binary and the new task has 4 targets, we add 2 more outputs to the network. This is a difficult form of cumulative learning [28].

Context Units: For the model that we develop in this study, when task t is encountered, for each layer ℓ in the network, we make use of a task embedding vector $\mathbf{g}_t^\ell$ (this new memory is much smaller than creating a new network for task t, which would require new weight matrices per layer rather than an extra vector). All $\mathbf{g}_t^\ell \in \mathcal{R}^{J_\ell \times 1}$ (J_ℓ is the number of units in layer ℓ) are stored in a memory matrix $\mathbf{M}^\ell \in \mathcal{R}^{(N+1)\times J_\ell}$ ($\mathcal{M} = \{\mathbf{M}^1, ...\mathbf{M}^L\}$), where a context can be retrieved using a one-hot encoding of the task descriptor, i.e., $\mathbf{M}^\ell \cdot \mathbf{t}$ (note that this will be produced by another system (see Section 3.3).

3.2 The Interactive Generative Model

The sequential neural coding network (S-NCN) is designed to make flexible conditional predictions – e.g., predicting $\mathbf{y}$ given $\mathbf{x}$, predicting both $\mathbf{y}$ and $\mathbf{x}$, predicting the missing parts of $\mathbf{x}$ given $\mathbf{y}$ and the observed parts of $\mathbf{x}$, etc. In order to do so, it treats inputs and outputs in a non-standard way. The input to the model is a task descriptor t_i and the output units represent $(\mathbf{y}, \mathbf{x})$. To predict $\mathbf{y}_i$ given $\mathbf{x}_i$, we clamp the output nodes that are responsible for predicting $\mathbf{x}$ and force their output to be $\mathbf{x}_i$.[1] During training, outputs are clamped to both $\mathbf{y}_i$ and $\mathbf{x}_i$, forcing the S-NCN to update latent states and synapses. The S-NCN can operate as a probabilistic generative network by feeding in a random noise vector as input, but we leave this extension to future work (our focus here is predicting $\mathbf{y}$ given $\mathbf{x}$).

The full computational process of the S-NCN is defined by three key steps: 1) layer-wise hypothesis generation, 2) state error-correction, and 3) synaptic weight evolution. The first two steps iteratively predict and correct the representations of the model for the input and target values of the task. After K iterations, model weights and the current task descriptor memory are updated. In this section, we provide details of the above steps and then describe the objective function that our model optimizes.

3.2.1 Inference: Predicting and Correcting States

Layer-wise State Prediction. The architecture of the S-NCN can be viewed as a stack of parallel, stateful neural-based prediction layers $P_1, ..., P_\ell, ..., P_m$, where the goal of each predictor is to guess the internal state of the predictor in the layer below (i.e., the S-NCN is not a feedforward network). The state of P_ℓ (layer ℓ) is represented by the (zero-initialized) vector $\mathbf{z}^\ell \in \mathcal{R}^{J_\ell \times 1}$ (J_ℓ is the number of units in layer ℓ). P_ℓ makes a prediction $\mathbf{z}_\mu^{\ell-1}$ about the current state $\mathbf{z}^{\ell-1}$ of $P_{\ell-1}$ (layer $\ell - 1$). Furthermore, we let $\mathbf{z}_x^0$ and $\mathbf{z}_y^0$ denote clamped outputs. That is, if we want to predict the label $\mathbf{y}_i$ given the features $\mathbf{x}_i$, we set $\mathbf{z}_x^0 = \mathbf{x}_i$ and if we want to train, we set both $\mathbf{z}_x^0 = \mathbf{x}_i$ and $\mathbf{z}_y^0 = \mathbf{y}_i$. The values predicted by layer 1 are denoted by $\mathbf{z}_{\mu,x}^0$ and $\mathbf{z}_{\mu,y}^0$. Note that $\mathbf{z}^0 = [\mathbf{z}_x^0, \mathbf{z}_y^0]$, where the two are concatenated (if either is missing, the SNCN completes the required values – see Appendix).

With respect to neural structure, the parallel predictors of the S-NCN are locally connected through (forward) generative weights $\mathbf{W}^\ell \in \mathcal{R}^{J_\ell \times J_{\ell+1}}$ and error (feedback) weights $\mathbf{E}^\ell \in \mathcal{R}^{J_{\ell+1} \times J_\ell}$, which work to transmit error information to relevant regions of neural processing elements, effectively coordinating all of the model's predictors. Formally, a predictor, with state $\mathbf{z}^{\ell+1}$, that guesses the

[1] Similarly, in the case of missing data, we can ask the network to predict the $\mathbf{y}_i$ and the missing parts of $\mathbf{x}_i$ given the observed parts of $\mathbf{x}_i$ by clamping outputs to only the observed parts of $\mathbf{x}_i$.

state of $\mathbf{z}^\ell$, takes on the form (given matrix $\mathbf{W}^{\ell+1}$ and activation $\phi^{\ell+1}$):

$$\mathbf{z}^\ell_\mu = \mathbf{W}^{\ell+1} \cdot \phi^{\ell+1}(\mathbf{z}^{\ell+1}), \quad \mathbf{e}^\ell = (\mathbf{z}^\ell - \mathbf{z}^\ell_\mu) \tag{1}$$

where $\mathbf{e}^\ell$ is a block of error units. Error units are paired with each predictor. Their task is to compute the disagreement or mismatch between the predictor's output and the target activity pattern $\mathbf{z}^\ell$. The error unit vector $\mathbf{e}^\ell$ can also be derived from the total discrepancy reduction presented in Equation 5. Note that for layer 0, $\mathbf{e}^0 = [\mathbf{e}^0_x, \mathbf{e}^0_y]$ (there are error neurons for $\mathbf{x}_i$ as well as others for $\mathbf{y}_i$, the concatenation of which makes up the bottom-most prediction error signal).

Latent State Correction and Context Updating. Once each layer ℓ has made a prediction about the layer below it and error units have been activated, the state of layer ℓ can be adjusted to take into account the local top-down and bottom-up error information. Using its current state and the error nodes (along with its task embedding vector $\mathbf{g}^\ell_t$), layer ℓ in the S-NCN adjusts its state $\mathbf{z}^\ell$ as follows:

$$\mathbf{z}^\ell(k) = f^\ell(\mathbf{z}^\ell(k-1) + \beta \mathbf{d}^\ell, \mathbf{g}^\ell_t), \quad \text{where } \mathbf{d}^\ell = -\mathbf{e}^\ell + \mathbf{E}^\ell \cdot \mathbf{e}^{\ell-1} \tag{2}$$

where β is the state correction rate and k marks one step of the S-NCN's K-step inference process.Note that $\mathbf{d}^\ell$ is the perturbation that adjusts the values of the hidden states – it combines a top-down expectation of layer ℓ w/ a bottom-up pressure from layer $\ell - 1$). The error feedback weights $\mathbf{E}^\ell$ are parameters that play a crucial role in this calculation, as they are responsible for transmitting the error for ℓ back up to layer $\ell + 1$. Notably, part of the state-correction requires competition among the individual units in a given layer through the function $f^\ell(\mathbf{z}^\ell, \mathbf{g}^\ell_t)$. There are various ways in which this function can be implemented and, in this study, we implemented it as:

$$f^\ell(\mathbf{z}^\ell, \mathbf{g}^\ell_t) = \left(\mathbf{I} \odot \mathbf{V}^\ell\right) \cdot \mathbf{z}^\ell, \text{where, } \mathbf{V}^\ell = \text{BKWTA}(\mathbf{g}^\ell_t, K) \cdot \text{BKWTA}((\mathbf{g}^\ell_t)^T, K)$$

where $\text{BKWTA}(\mathbf{v}, K)$ is the binarized K winners-take-all function, yielding a binary vector with 1 at the index of each of the K winning units, or formally:

$$\text{BKWTA}(\mathbf{v}, K) = \left\{1 \text{ if } v_j \in \{K \text{ largest elements of } \mathbf{v}\} \text{ and } \phi(\mathbf{v}) = 0 \text{ otherwise}\right\}.$$

In the Appendix, we study other forms of the competition function (and find that the above performed best, so we report this in the main paper). Note that this lateral inhibition is a function of evolving context $\mathbf{g}^\ell$, triggered by a task pointer t_i (produced by a task selector model, which we define later).

In real neural systems, competition between units within a layer is thought to facilitate contextual processing [29], where only some neuronal signals are strengthened while the activity of others is suppressed. Moreover, lateral competition, often modeled in classical models with anti-Hebbian learning [30], encourages the formation of sparse codes [31, 32]. Since the S-NCN is an interactive/iterative model, incorporating lateral competition (in the form of a task-driven memory matrix) is natural and computed online, highlighting the model's flexibility compared to an ANN – to incorporate lateral synaptic activity, recurrent weights could be introduced to an ANN but training it would requiring unrolling via backprop through time. More importantly, a lateral mechanism in the S-NCN works as an inductive bias that leads it to acquire task-dependent representations, which can aid in memory retention since the system has to store information on multiple, disjoint tasks. In a sense, this task specialization that we build into the neural dynamics is similar in spirit to activation sharpening [33].

3.2.2 Updating Synaptic Parameters

Given that the S-NCN is an interactive network [34], inferring its states requires running Equations 1 and 2 K times, where the model alternates between making predictions and then correcting states once error units have been computed. Once latent states have been inferred, the S-NCN is then able to adjust its synaptic values. The synaptic updates take the form of Hebbian-like rules:

$$\Delta \mathbf{W}^\ell = \mathbf{e}^\ell \cdot (\phi^\ell(\mathbf{z}^{\ell+1}))^T \odot \mathbf{S}^\ell_W, \text{ and, } \Delta \mathbf{E}^\ell = \alpha\left(\mathbf{d}^{\ell+1} \cdot (\mathbf{e}^\ell)^T\right) \odot \mathbf{S}^\ell_E \tag{3}$$

where α is a scaling factor, usually set to < 1.0, that makes the error feedback weights change at a slower rate than the forward weights (this improves convergence [35]). $\mathbf{S}^\ell_W \in \mathcal{R}^{J_\ell \times J_{\ell+1}}$ and $\mathbf{S}^\ell_E \in \mathcal{R}^{J_{\ell+1} \times J_\ell}$ are modulation factors that provide stability to the weight updates (see Appendix).

An important property of the above weight update rules is that they are local – to compute changes in the synapses, all we require is the information immediately available to the neuron(s) of interest (making these rules function similarly to classical Hebbian updates [36], although there are important differences to them, as discussed in [35]). Since a neuron is able to immediately generate a hypothesis

Algorithm 1 State inference procedure.

Input: sample $(\mathbf{y}, \mathbf{x}, \mathbf{t})$, β, $\mathcal{M}$, & Θ
function INFERSTATES($(\mathbf{y}, \mathbf{x}, \mathbf{t}), \Theta$)
  $(\mathbf{g}^1, ..., \mathbf{g}^\ell, ..., \mathbf{g}^L) \leftarrow getContexts(\mathbf{t}, \mathcal{M})$
  Set $\mathbf{z}^1, ..., \mathbf{z}^\ell, ..., \mathbf{z}^L$ to $\mathbf{0}$
  $\mathbf{e}^L = \mathbf{0}$, $\mathbf{z}^0_x = \mathbf{x}$, $\mathbf{z}^0_y = \mathbf{y}$
  for $k = 1$ to K **do**
    for $\ell = L$ to 1 **do** ▷ Run layerwise predictors
      if $\ell > 1$ **then** ▷ Latent prediction layer
        $\mathbf{z}^{\ell-1}_\mu = \mathbf{W}^\ell \cdot \phi^\ell(\mathbf{z}^\ell)$
        $\mathbf{e}^{\ell-1} = (\mathbf{z}^{\ell-1} - \mathbf{z}^{\ell-1}_\mu)$
      else ▷ Sensory prediction layer
        $\mathbf{z}^{\ell-1}_{\mu,x} = \mathbf{W}^\ell_x \cdot \phi^\ell(\mathbf{z}^\ell)$
        $\mathbf{e}^{\ell-1}_x = (\mathbf{z}^{\ell-1}_x - \mathbf{z}^{\ell-1}_{\mu,x})$
        $\mathbf{z}^{\ell-1}_{\mu,y} = \mathbf{W}^\ell_y \cdot \phi^\ell(\mathbf{z}^\ell)$
        $\mathbf{e}^{\ell-1}_y = (\mathbf{z}^{\ell-1}_y - \mathbf{z}^{\ell-1}_{\mu,y})$
    for $\ell = L$ to 1 **do** ▷ Correct internal states
      if $\ell == 1$ **then**
        $\mathbf{d}^\ell = -\mathbf{e} + \mathbf{E}^\ell_x \cdot \mathbf{e}^{\ell-1}_x + \mathbf{E}^\ell_y \cdot \mathbf{e}^{\ell-1}_y$
      else
        $\mathbf{d}^\ell = -\mathbf{e} + \mathbf{E}^\ell \cdot \mathbf{e}^{\ell-1}$
      $\mathbf{z}^\ell = f^\ell\Big(\phi^\ell(\mathbf{z}^\ell) + \beta \mathbf{d}^\ell, \mathbf{g}^\ell\Big)$
  $\Lambda = (\mathbf{z}^1, ..., \mathbf{z}^\ell, ..., \mathbf{z}^L, \mathbf{d}^1, ..., \mathbf{d}^\ell, ..., \mathbf{d}^L,$
    $\mathbf{e}^0, ..., \mathbf{e}^\ell, ..., \mathbf{e}^{L-1})$
  Return Λ

Algorithm 2 Weight update computation.

Input: Λ, λ, γ, $\mathcal{M}$, & Θ
function UPDATEWEIGHTS(Λ, Θ)
  // Calculate weight displacements
  for $\ell = L$ to 1 **do**
    $\Delta\mathbf{W}^\ell = \Big(\mathbf{e}^{\ell-1} \cdot (\phi^\ell(\mathbf{z}^\ell))^T\Big) \odot \mathbf{S}^\ell_W$
    $\Delta\mathbf{E}^\ell = \gamma(\mathbf{d}^\ell \cdot (\mathbf{e}^{\ell-1})^T) \odot \mathbf{S}^\ell_E$
  $\Delta\mathbf{W}^1_x = \Big(\mathbf{e}^0_x \cdot (\phi^1(\mathbf{z}^1))^T\Big) \odot \mathbf{S}^1_{W,x}$
  $\Delta\mathbf{E}^1_x = \gamma(\mathbf{d}^1 \cdot (\mathbf{e}^0_x)^T) \odot \mathbf{S}^1_{E,x}$
  $\Delta\mathbf{W}^1_y = \Big(\mathbf{e}^0_y \cdot (\phi^1(\mathbf{z}^1))^T\Big) \odot \mathbf{S}^1_{W,y}$
  $\Delta\mathbf{E}^1_y = \gamma(\mathbf{d}^1 \cdot (\mathbf{e}^0_y)^T) \odot \mathbf{S}^1_{E,y}$

  // Update current weights
  for $\ell = L$ to 1 **do**
    $\mathbf{W}^\ell = \mathbf{W}^\ell + \lambda\Delta\mathbf{W}^\ell$
    $\mathbf{E}^\ell = \mathbf{E}^\ell + \lambda\Delta\mathbf{E}^\ell$
  $\mathbf{W}^1_x = \mathbf{W}^1_x + \lambda\Delta\mathbf{W}^1_x$
  $\mathbf{E}^1_x = \mathbf{E}^1_x + \lambda\Delta\mathbf{E}^1_x$
  $\mathbf{W}^1_y = \mathbf{W}^1_y + \lambda\Delta\mathbf{W}^1_y$
  $\mathbf{E}^1_y = \mathbf{E}^1_y + \lambda\Delta\mathbf{E}^1_y$
  Update $(\mathbf{g}^1, \cdots, \mathbf{g}^L, \mathcal{M})$ via Eqn. 4
  // Return new weights
  $\Theta = \{\mathbf{W}^1_x, \mathbf{W}^1_y, ..., \mathbf{W}^\ell ... \mathbf{W}^L,$
    $\mathbf{E}^1_x, \mathbf{E}^1_y, ..., \mathbf{E}^\ell, ..., \mathbf{E}^L\}$
  Return Θ

given its own internal state, without requiring the active generation of other predictors, and its error can be readily computed after prediction by comparing to the current state of the target neuron state, the weight updates of any predictor layer may be computed in parallel to others. This would allow us to allocate dedicated computing cores to particular predictors, or "pieces", of the S-NCN. Observe that the S-NCN does not require activation derivatives in any of its computations (this neurobiologically more realistic and favorable for specialized hardware implementations).

Furthermore, during the learning process, each context vector $\mathbf{g}^\ell_t$ and the corresponding memory matrix is adjusted according to the following simple contrastive rule:

$$\mathbf{g}^\ell_{t+1} = \mathbf{g}^\ell_t + \eta_e \mathbf{d}^\ell - \eta_g(\mathbf{g}^\ell_t - \frac{1}{t-1}\sum_{j=1}^{t-1}(\mathbf{g}^\ell_j)), \text{ and, } \Delta\mathbf{M}^\ell = \mathbf{g}^\ell_{t+1} * (\mathbf{t}_j)^T \tag{4}$$

where η_e modulates a long-term memory update using the current perturbation to be applied to layer ℓ. η_g controls the repulsion term, which "pushes" context codes away from each other (for diversity). These adapted codes, which influence inter-neuronal competition in a task-sensitive manner, could be viewed as a simplification of distributed temporal context [37], where contiguity, i.e., recall/generation of one item is influenced by the presence of another, is introduced into S-NCN distributed processing.

The pseudocode illustrating how the elements described so far are combined in an S-NCN system is presented in Algorithms 1 and 2. The transmission of bottom-up and top-down errors in the S-NCN is motivated by the theory of predictive processing [38, 39, 40, 41, 42, 43, 44] and classical work on interactive networks [34, 45, 46], where models undergo a settling process to process input stimuli more than once (see Appendix for a discussion on motivations). Though this requires extra computation, the process endows the network with desirable properties, e.g., the ability to conduct constraint satisfaction [47, 48]. By using a multi-step processing, laterally-competitive processing scheme, the S-NCN is able to "select" subnetworks (portions of neurons) for specific tasks, reducing representational overlap and, ultimately, forgetting. This selection is driven by the task pointer, produced by the task selector, $\mathcal{T}(\mathbf{x})$, the final piece of the S-NCN system, which we describe next.

3.2.3 Objective Function

During training, when presented with stimulus $(\mathbf{y}_i, \mathbf{x}_i, \mathbf{t}_i)$, the S-NCN adjusts its internal states and synapses so that the output $\mathbf{z}^0_\mu$ of layer 1 (P_1) is as close as possible to $(\mathbf{y}_i, \mathbf{x}_i)$. It does this by minimizing *total discrepancy* [49] – a measure of its total internal disorder (and one that can

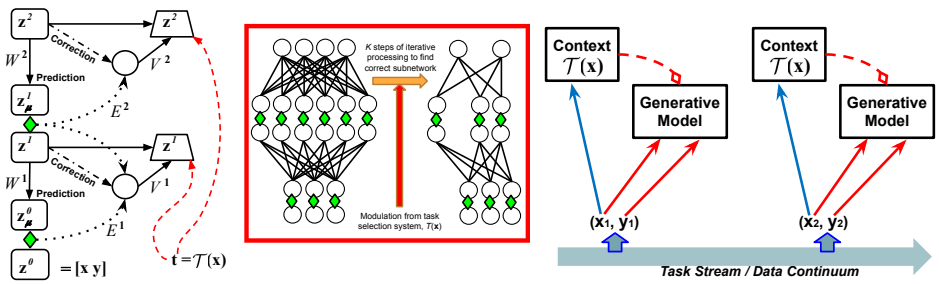

Figure 1: (Left) The sequential neural coding network is shown processing an input $(\mathbf{y}, \mathbf{x})$ over one step, given a task context $\mathbf{t}$ (produced by the task selection model $\mathcal{T}(\mathbf{x})$), overriding the current $\mathbf{z}^\ell$ after error-correction and application of lateral inhibition matrices $\{\mathbf{V}^1, \mathbf{V}^2\}$. Inside the red box is shown one possible emergent subnetwork. Green diamonds indicate error units. (Right) The full complementary neural system shown processing patterns from a data continuum.

be shown to approximate free energy [50]), which is the sum of all mismatches between layerwise guesses and actual states. In its general form, total discrepancy for an S-NCN is formally:

$$\mathcal{L}(\Theta) = \sum_{\ell=0}^{L-1} \mathcal{L}^\ell(\Theta^\ell), \text{ where } \mathcal{L}^\ell(\Theta^\ell) = \frac{1}{2}(||\mathbf{z}^\ell - \mathbf{z}^\ell_\mu||_2)^2. \tag{5}$$

Θ contains all of the synaptic parameters, i.e., $\Theta = \{\mathbf{W}^1, \mathbf{E}^1, \cdots, \mathbf{W}^L, \mathbf{E}^L\}$ and $\Theta^\ell = \{\mathbf{W}^\ell, \mathbf{E}^\ell\}$.

The above loss decomposes the problem of credit assignment in the S-NCN into several sub-problems that each focuses on the comparison between the prediction $\mathbf{z}^\ell_\mu$ made by layer $\ell + 1$ and the actual state value $\mathbf{z}^\ell$ of layer ℓ. The resulting updates to each state $\mathbf{z}^\ell$, as well as the relevant parameters, will then depend on a bottom-up transmitted error signal and the top-down influence of the mismatch with the expectation of the predictor immediately above [49, 35]. Note that, while we motivate aspects of our model from a neuro-cognitive perspective, the error units and weight updates can be derived from the total discrepancy function above [35] (and cast as approximately minimizing free energy).

3.3 The Neural Task Selection Model

In the last section, we described the S-NCN as taking in a task signal that drives the choice of the right internal task context memory $\mathbf{g}^\ell$ for layer ℓ. To create this task pointer, we design a second neural model that we call the task selector $\mathcal{T}(\mathbf{x})$ – this will allow the S-NCN to automatically decide when a new task has been encountered and to determine, at test time, what task data points belong to. The motivation behind $\mathcal{T}(\mathbf{x})$ comes from the neuroscientific idea that the basal ganglia plays the role of information routing (among other roles), which serves as a form of task selection [51, 52]). In other words, it selects or enables various cognitive programs stored in other cortical regions [5]. As a result, we develop a sort of complementary learning system (CLS) (different than that of [53], which sets up an interaction between a hippocampus and a neocortex model), which pairs our cortex-like model with a basal ganglia-like model. We refer to our task selector $\mathcal{T}(\mathbf{x})$ as the "functional neural basal ganglia" (FNBG) in order to emphasize the fact that the actual basal ganglia in the human brain is far more complex and does far more than what our simple computational model does. Our FNBG model, built with competitive learning, has two roles: 1) task shift detection - deciding whether data from an input stream is indicating the presence of a new task, and 2) task recognition - deciding whether incoming data requires switching to an existent task context or creating a new context knowledge.

Task Shift Detection: In order to detect the occurrence of a new task while processing data from the pattern stream, $\mathcal{T}(\mathbf{x})$ utilizes the output error neurons of the generative model described earlier to detect spikes in their values that might indicate distributional/context drift. Specifically, $\mathcal{T}(\mathbf{x})$ maintains an exponential running mean $\mu_\mathcal{L}$ and variance $\sigma^2_\mathcal{L}$ of the norm of the cortex model's label error neurons $\mathbf{e}^0_y$. The necessary statistics are calculated as follows:

$$\Delta = ||\mathbf{e}^0_y||^2_2 - \mu_\mathcal{L}(t-1) \tag{6}$$

$$\mu_\mathcal{L}(t) = \mu_\mathcal{L}(t-1) + \eta\Delta \tag{7}$$

$$\sigma^2_\mathcal{L}(t) = (1-\eta)\sigma^2_\mathcal{L}(t-1) + \eta(\Delta)^2 \tag{8}$$

with $\eta = 0.1$ (determined by preliminary experiments). Using the above dynamic statistics, a shift is detected by determining if the following inequality evaluates to true (repeatedly for a series of 5 consecutive batches): $\mu_{\mathcal{L}}(t) > \mu_{\mathcal{L}}(t-1) + 2\sqrt{\sigma^2_{\mathcal{L}}(t-1)}$. Upon detection of a boundary, we suppress the check until 1000 samples have been seen after the last detected shift (creating a refractory period to allow the competitive learning model, described next, to acquire enough data to learn).

Task Recognition through Competitive Learning: To conduct task recognition, $\mathcal{T}(\mathbf{x})$ first randomly projects the input $\mathbf{x}$ to a low-dimensional space ("key") $\mathbf{k} = \mathbf{R} \cdot \mathbf{x}$ ($\mathbf{R}$ is initialized from a Gaussian distribution). We then update a rolling average estimate of the streaming input using this generated key as follows: $\mathbf{k}_\mu = (1-\tau)\mathbf{k}_\mu + \tau\mathbf{k}$ (with $\tau = 0.65$). Finally, with the matrix $\mathbf{Q}$, FNBG maps this rolling representation $\mathbf{k}_\mu$ to a decision as to which task context the S-NCN generative cortex is to utilize, i.e., $\hat{\mathbf{t}} = \text{BKWTA}(\mathbf{Q} \cdot (\mathbf{k}_\mu / ||\mathbf{k}_\mu)||_2, K = 1)$.The task pointer produced for the S-NCN model is then $t_i = \arg\max(\hat{\mathbf{t}})$ ($\arg\max(\hat{\mathbf{t}})$ returns the index of the maximum value in $\hat{\mathbf{t}}$).

To update the FNBG weights, while also avoiding catastrophic interference in $\mathcal{T}(\mathbf{x})$ itself, we propose a biologically-inspired learning rule based on competitive learning. Specifically, we develop what we call "guided competitive learning", since during the act of processing a stream of certain samples from the task we know we are operating on, we know which neuron out of a set of $T-1$ task output neurons should be selected. This leads to the following update rule formally defined as:

$$\Delta\mathbf{Q} = -(\rho\mathbf{t}) \cdot (\mathbf{k}_\mu / ||\mathbf{k}_\mu||_2)^T \tag{9}$$

where ρ is the competitive weight adjustment factor (a value we found works well in the range of $[0.5, 1)$). Note that $\mathbf{t} = \hat{\mathbf{t}}$ recovers an unsupervised classical competitive Hebbian learning. However, we force the model to a specific task pointer value by using $\mathbf{t}$, the one-hot encoding of the dynamic integer variable t maintained by the FNBG itself, initialized as $t = 0$. Every time a task shift is detected according to Equations 6-8, this dynamic variable is incremented by one, i.e., $t \leftarrow t + 1$.

For both task recognition and the FNBG's synaptic update, note that the rolling representation $\mathbf{k}_\mu$ is normalized by its Euclidean norm so that we may utilize a vectorizable form of competitive learning based on dot products (taking advantage of GPU-based matrix multiplication). In short, we take the (normalized) rolling average representation of the input stream for a given task, compute its dot-product simultaneously with all currently-available task weight vectors, and choose the dot product with maximal value as the winner. Finally, we re-normalize the matrix Q by its column Euclidean norms after each update, i.e., $Q[:,i] = (Q[:,i] + \Delta Q[:,i])/||Q[:,i]||_2$ where $Q[:,i]$ indicates extracting all values in column i from Q (this normalization is similar to that of adaptive resonance theory [54]).

3.4 Putting It All Together: A Complementary System

At a high level, given the above, the full S-NCN complementary system, depicted in Figure 1 (Right), consists of:1) a task selection model (inspired by the information routing/executive control behavior of the basal ganglia [5]) which creates the task contexts that laterally inhibit/gate the activities of the generative S-NCN, and 2) a generative model that learns to predict inputs/labels given a task context. In essence, the FNBG $\mathcal{T}(\mathbf{x})$ takes in $\mathbf{x}_j$ to produce $\hat{\mathbf{t}}$ (a one-hot encoding of the task pointer t_i) which is then fed into the generative S-NCN (along with $\mathbf{x}$ and $\mathbf{y}$) to compute predictions. [2]

4 Experiments

Simulation Setup: In our experiments, we train models with three hidden layers, whether they be multilayer perceptrons (MLPs) or S-NCNs and compare against baselines from the literature. All models were restricted to contain (a maximum of) 500 units per layer. For the S-NCN, weights were initialized from a Gaussian distribution scaled by each layer's fan-in and were optimized using stochastic gradient descent with learning rate of $\lambda = 0.01$. Baseline models were trained on each task for 40 epochs while the S-NCN only made a single pass. The output layer for each MLP was a maximum entropy classifier and the objective was to minimize Categorical cross entropy – in the S-NCN, this was encoded in its label error neurons $\mathbf{e}^0_y$. (See Appendix for experimental details, computing infrastructure, hyper-parameter details, and code details.)

[2] Please see the Appendix where we summarize symbols, notation, and abbreviation used in this work.

Table 1: Generalization metrics (10 trials) for sequence orderings # 1 & #2 (higher values are better).

	Ordering #1: $\{M1, M2, GD1, FM1, FM2, GD2\}$ **(High Color Sim.)**			
	Equal		**Unequal**	
	ACC	**BWT**	**ACC**	**BWT**
Backprop	0.241 ± 0.050	-0.759 ± 0.030	0.185 ± 0.048	-0.791 ± 0.048
EWC	0.280 ± 0.023	-0.714 ± 0.030	0.185 ± 0.046	-0.726 ± 0.039
Md-IMM	0.521 ± 0.027	-0.392 ± 0.023	0.480 ± 0.039	-0.240 ± 0.040
DT+Md-IMM	0.530 ± 0.024	-0.387 ± 0.021	0.551 ± 0.042	-0.220 ± 0.042
L2+DT+Md-IMM	0.532 ± 0.025	-0.237 ± 0.027	0.520 ± 0.040	-0.240 ± 0.045
HAT	0.550 ± 0.019	-0.211 ± 0.020	0.492 ± 0.031	-0.231 ± 0.036
S-NCN (ours)	$\mathbf{0.716 \pm 0.013}$	$\mathbf{-0.031 \pm 0.017}$	$\mathbf{0.713 \pm 0.011}$	$\mathbf{-0.041 \pm 0.012}$
	Ordering #2: $\{GD2, M1, FM2, M2, GD1, FM1\}$ **(Low Color Sim.)**			
Backprop	0.303 ± 0.030	-0.644 ± 0.037	0.287 ± 0.043	-0.671 ± 0.043
EWC	0.303 ± 0.031	-0.643 ± 0.033	0.291 ± 0.039	-0.663 ± 0.047
Md-IMM	0.584 ± 0.027	-0.091 ± 0.030	0.533 ± 0.034	-0.230 ± 0.036
DT+Md-IMM	0.591 ± 0.020	-0.088 ± 0.032	0.528 ± 0.036	-0.211 ± 0.039
L2+DT+Md-IMM	0.630 ± 0.029	-0.076 ± 0.030	0.551 ± 0.037	-0.201 ± 0.041
HAT	0.596 ± 0.026	-0.114 ± 0.029	0.563 ± 0.031	-0.210 ± 0.044
S-NCN (ours)	$\mathbf{0.721 \pm 0.014}$	$\mathbf{-0.042 \pm 0.013}$	$\mathbf{0.667 \pm 0.011}$	$\mathbf{-0.097 \pm 0.013}$

Table 2: Generalization metrics (10 trials) for Split MNIST, Split Fashion MNIST (FMNIST) and Not-MNIST benchmarks. Above dashed line are multi-head models & below are single-head models.

	MNIST		**Fashion MNIST**		**NotMNIST**	
	ACC	**BWT**	**ACC**	**BWT**	**ACC**	**BWT**
EWC	0.760 ± 0.030	-0.210	0.739 ± 0.020	-0.201	0.790 ± 0.020	-0.176
VCL	0.980 ± 0.210	-0.003	0.980 ± 0.20	-0.002	0.953 ± 0.003	-0.004
IMM	0.951 ± 0.018	-0.007	0.950 ± 0.013	-0.005	0.925 ± 0.011	-0.006
HAT	0.972 ± 0.011	-0.040	0.968 ± 0.011	-0.004	0.942 ± 0.009	-0.005
GEM	0.922 ± 0.110	$\mathbf{+0.001}$	0.930 ± 0.12	$+0.001$	0.919 ± 0.021	-0.003
A-GEM	0.950 ± 0.09	$\mathbf{+0.001}$	0.955 ± 0.11	$+0.001$	0.925 ± 0.020	-0.002
ER	0.938 ± 0.06	$\mathbf{-0.002}$	0.945 ± 0.10	-0.004	0.927 ± 0.016	-0.004
EWC	0.190 ± 0.030	-0.357	0.199 ± 0.06	-0.350	0.186 ± 0.020	-0.361
NR+M	0.950 ± 0.470	-0.100	0.948 ± 0.380	-0.090	0.880 ± 0.028	-0.103
SI	0.197 ± 0.110	-0.367	0.198 ± 0.100	-0.370	0.161 ± 0.030	-0.370
MAS	0.195 ± 0.290	-0.340	0.180 ± 0.250	-0.340	0.178 ± 0.060	-0.341
Lwf	0.846 ± 0.340	-0.120	0.875 ± 0.300	-0.130	0.626 ± 0.091	-0.130
ICarl	0.940 ± 0.410	-0.100	0.960 ± 0.400	-0.110	0.887 ± 0.102	-0.109
Lucir	0.940 ± 0.310	-0.103	0.950 ± 0.340	-0.110	0.935 ± 0.093	-0.101
GDumb	0.978 ± 0.09	$\mathbf{-0.005}$	0.973 ± 0.09	-0.006	0.940 ± 0.080	-0.004
Mnem	0.960 ± 0.320	-0.091	0.968 ± 0.300	$\mathbf{+0.007}$	0.950 ± 0.071	-0.011
S-NCN	$\mathbf{0.981 \pm 0.300}$	-0.005	$\mathbf{0.982 \pm 0.400}$	-0.003	$\mathbf{0.957 \pm 0.400}$	$\mathbf{-0.004}$

Evaluation Metrics: To measure model generalization over the sequence of tasks, we make use of the resulting task matrix R (as in [55]), an $N \times N$ matrix of task accuracy scores (normalized to $[0, 1]$), where in this study $N = T$. We measure average accuracy (ACC) (mean performance across tasks) and backward transfer (BWT). BWT measures the influence that learning a task T_t has on the performance of task $T_k < T_t$. A positive BWT indicates that a learning task T_t increases performance on a preceding task T_k. As such, a higher BWT is better and a strongly negative BWT means there is stronger (more catastrophic) forgetting. Mean and standard deviation (10 trials) are reported for ACC and mean (10 trials) for BWT (see Appendix for standard deviation). The formulas for ACC, BWT, and for a new set of metrics created to analyze memory retention, are provided in the Appendix.

4.1 The Multi-Dataset Task Stream

To start, we tested the S-NCN model on a complex task sequence composed of several learning benchmarks.We create task sequences by breaking apart MNIST (*M*), Fashion MNIST (*FM*), and Google Draw (*GD*) each into two "sub-tasks" (e.g., for MNIST, *M1* and *M2*), or portions of data with a particular subset of the original dataset's classes. See the Appendix for details on sub-tasks/task sequence creation. In Table 1, we present two task orderings (Ordering #1 is "High Color Sim." and Ordering #2 is "Low Color Sim.") each under two conditions: sub-tasks that have 1) an equal number

of classes (5 each), and 2) an unequal number of classes (number of classes was chosen randomly, omitting 5 as an option). The number of classes was sampled once and held constant for all trials.

We evaluate our proposed S-NCN system (as well as four variations of it in the Appendix) – hyperparameters were $\beta = 0.05$, $K = 10$, $\eta_g = 0.9$, $\eta_e = 0.01$, $\alpha = 0.98$). The baselines include an MLP trained only with backprop (Backprop), Elastic Weight Consolidation (EWC) [19], the Mode-IMM method [56], the Md-IMM method combined with either DropTransfer (DT+Md-IMM) or both L2-Transfer and DropTransfer (L2-DT-Mode-IMM) [56], and the competitive model, hard attention to task (HAT) [57]. For each baseline, we tuned hyper-parameters based on their accuracy on each task's development set. See Appendix for extra baseline results.

Discussion: Results are reported in Table 1 (see Appendix for more results). Each simulation was run 10 times (each trial used a unique seed) – we report both mean and standard deviation. As observed in our results, **we see that our S-NCN model outperforms all baselines consistently, in terms of ACC and BWT, exhibiting improved memory retention over the baselines, such as backprop, and more notably, EWC** (and, in the Appendix, our expanded results show that the FNBG-driven lateral inhibition is key to improving memory retention the most). This result is robust across both task sequences and equal/unequal class settings. Meta-parameters for the S-NCNs were only tweaked minorly, with the same values used in all scenarios. The observation that lateral inhibition improves the neural computation (in our model) also corroborates the result of [48].

4.2 Continual Learning Benchmarks

To connect our model to current lifelong learning results, we experimented with a wide swatch of approaches on three benchmarks – Split MNIST, Split NotMNIST, and Split Fashion MNIST (FMNIST). Furthermore, we compare to multi-head (below dashed line in Table 2) and single-head models (above dashed line). We compare the S-NCN to replay/rehearsal and non-replay methods: naïve rehearsal with memory (NR+M), EWC, synaptic intelligence (SI), MAS [58], Lwf [59], GEM [60], ICarl [61], Lucir [62], and Mnemonics [63] (additional baselines can be found in the Appendix).

Discussion: In Table 2, we report ACC and BWT, averaged over 10 trials, offering an extensive, comprehensive comparison of methods and demonstrating that, for all three benchmarks, **the proposed S-NCN outperforms all of them in terms of ACC**, and nearly all in terms of BWT (and on par with GEM, the difference in BWT being negligible) demonstrating the power afforded by designing models with stronger grounding in neurobiology (see Appendix for a discussion of limitations). Furthermore, it is promising to see that the S-NCN outperforms/matches performance with not only the single-head models but also with the multi-head models (except GEM), which enjoy an easier version of the problem since they can utilize a different classifier per task. Finally, note that the S-NCN, due to the FNBG-driven lateral competition, learns to compose task contexts in a data-dependent manner. Desirably, the S-NCN is a single-head model, meaning that it does not grow out a separate softmax classifier for each task (as in multi-headed approaches, e.g., HAT, IMM, GEM), which means that the S-NCN is tackling the harder form of the forgetting problem, learning representations that preserve knowledge across disjoint tasks. Furthermore, note that our model is online, whereas models such as IMM or SI require multiple passes per task dataset, and does not require validation data in order to run an expensive neural architecture search (NAS) as in [64].

5 Conclusion

In this paper, we proposed the sequential neural coding network (S-NCN), an interactive generative model, and its local learning procedure for lifelong machine learning. As demonstrated on several benchmarks/setups, this model retains the knowledge acquired from prior tasks when learning new ones in task data streams, primarily when lateral inhibition, driven by a self-organizing task selection model, sharpens neural activities within its layers. In terms of negative social impact, S-NCN-based lifelong learning models could lead to the development of better-performing systems, i.e., robotic systems (drones), that might result in the loss of life (see Appendix for an expanded discussion). To safeguard against this, it will be important to develop an ethical framework to guide the design and training of general intelligent systems to ensure safe integration into human society.

References

[1] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and autonomous systems*, vol. 15, no. 1-2, pp. 25–46, 1995.

[2] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *arXiv preprint arXiv:1802.07569*, 2018.

[3] J. D. Power and B. L. Schlaggar, "Neural plasticity across the lifespan," *Wiley Interdisciplinary Reviews: Developmental Biology*, vol. 6, no. 1, 2017.

[4] F. Zenke and W. Gerstner, "Hebbian plasticity requires compensatory processes on multiple timescales," *Phil. Trans. R. Soc. B*, vol. 372, no. 1715, p. 20160259, 2017.

[5] G. Leisman, O. Braun-Benjamin, and R. Melillo, "Cognitive-motor interactions of the basal ganglia in development," *Frontiers in systems neuroscience*, vol. 8, p. 16, 2014.

[6] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *The psychology of learning and motivation*, vol. 24, no. 109, p. 92, 1989.

[7] R. Ratcliff, "Connectionist models of recognition memory: constraints imposed by learning and forgetting functions." *Psychological review*, vol. 97, no. 2, p. 285, 1990.

[8] S. Lewandowsky, "On the relation between catastrophic interference and generalization in connectionist networks," *Journal of Biological Systems*, vol. 2, no. 03, pp. 307–333, 1994.

[9] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[10] M. Toneva, A. Sordoni, R. T. d. Combes, A. Trischler, Y. Bengio, and G. J. Gordon, "An empirical study of example forgetting during deep neural network learning," *arXiv preprint arXiv:1812.05159*, 2018.

[11] A. Robins, "Catastrophic forgetting in neural networks: the role of rehearsal mechanisms," in *Artificial Neural Networks and Expert Systems, 1993. Proceedings., First New Zealand International Two-Stream Conference on*. IEEE, 1993, pp. 65–68.

[12] ——, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[13] ——, "Consolidation in neural networks and in the sleeping brain," *Connection Science*, vol. 8, no. 2, pp. 259–276, 1996.

[14] A. Gepperth and C. Karaoguz, "A bio-inspired incremental learning architecture for applied perceptual problems," *Cognitive Computation*, vol. 8, no. 5, pp. 924–934, 2016.

[15] S.-A. Rebuffi, A. Kolesnikov, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proc. CVPR*, 2017.

[16] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[17] G. I. Parisi and S. Wermter, "Lifelong learning of action representations with deep neural self-organization," in *The AAAI 2017 Spring Symposium on Science of Intelligence: Computational Principles of Natural and Artificial Intelligence, Standford, US*, 2017, pp. 608–612.

[18] E. Meier, L. Hertz, and A. Schousboe, "Neurotransmitters as developmental signals," *Neurochemistry international*, vol. 19, no. 1-2, pp. 1–15, 1991.

[19] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[20] R. M. French, "Semi-distributed representations and catastrophic forgetting in connectionist networks," *Connection Science*, vol. 4, no. 3-4, pp. 365–377, 1992.

[21] G. M. Van de Ven and A. S. Tolias, "Three scenarios for continual learning," *arXiv preprint arXiv:1904.07734*, 2019.

[22] J. R. Movellan, "Contrastive hebbian learning in the continuous hopfield model," in *Connectionist Models*. Elsevier, 1991, pp. 10–17.

[23] B. Scellier and Y. Bengio, "Equilibrium propagation: Bridging the gap between energy-based models and backpropagation," *Frontiers in computational neuroscience*, vol. 11, p. 24, 2017.

[24] D.-H. Lee, S. Zhang, A. Fischer, and Y. Bengio, "Difference target propagation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 498–515.

[25] A. G. Ororbia and A. Mali, "Biologically motivated algorithms for propagating local target representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4651–4658.

[26] S. Bartunov, A. Santoro, B. Richards, L. Marris, G. E. Hinton, and T. Lillicrap, "Assessing the scalability of biologically-motivated deep learning algorithms and architectures," in *Advances in Neural Information Processing Systems*, 2018, pp. 9390–9400.

[27] S. Thrun, "Is learning the n-th thing any easier than learning the first?" in *Advances in neural information processing systems*, 1996, pp. 640–646.

[28] G. Fei, S. Wang, and B. Liu, "Learning cumulatively to become more knowledgeable," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1565–1574.

[29] H. Adesnik and M. Scanziani, "Lateral competition for cortical space by layer-specific horizontal circuits," *Nature*, vol. 464, no. 7292, p. 1155, 2010.

[30] P. Földiak, "Forming sparse representations by local anti-hebbian learning," *Biological cybernetics*, vol. 64, no. 2, pp. 165–170, 1990.

[31] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[32] A. D. Szlam, K. Gregor, and Y. L. Cun, "Structured sparse coding via lateral inhibition," in *Advances in Neural Information Processing Systems*, 2011, pp. 1116–1124.

[33] R. M. French, "Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks," in *Proceedings of the 13th annual cognitive science society conference*. Erlbaum, 1991, pp. 173–178.

[34] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception: I. an account of basic findings." *Psychological review*, vol. 88, no. 5, p. 375, 1981.

[35] A. Ororbia, A. Mali, C. L. Giles, and D. Kifer, "Continual learning of recurrent neural networks by locally aligning distributed representations," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[36] D. O. Hebb *et al.*, "The organization of behavior," 1949.

[37] M. W. Howard and M. J. Kahana, "A distributed representation of temporal context," *Journal of Mathematical Psychology*, vol. 46, no. 3, pp. 269–299, 2002.

[38] R. P. Rao and D. H. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex," *Neural computation*, vol. 9, no. 4, pp. 721–763, 1997.

[39] ——, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." *Nature neuroscience*, vol. 2, no. 1, 1999.

[40] Y. Huang and R. P. Rao, "Predictive coding," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 2, no. 5, pp. 580–593, 2011.

[41] C. Wacongne, J.-P. Changeux, and S. Dehaene, "A neuronal model of predictive coding accounting for the mismatch negativity," *Journal of Neuroscience*, vol. 32, no. 11, pp. 3665–3678, 2012.

[42] R. Chalasani and J. C. Principe, "Deep predictive coding networks," *arXiv preprint arXiv:1301.3541*, 2013.

[43] A. Clark, *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, 2015.

[44] E. Santana, M. S. Emigh, P. Zegers, and J. C. Principe, "Exploiting spatio-temporal structure with recurrent winner-take-all networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2017.

[45] J. L. McClelland, "The grain model: A framework for modeling the dynamics of information processing," *Attention and Performance (Volc. XIV): Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience.*, 1993.

[46] R. C. O'Reilly, "Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm," *Neural computation*, vol. 8, no. 5, pp. 895–938, 1996.

[47] ——, "Six principles for biologically based computational models of cortical cognition," *Trends in cognitive sciences*, vol. 2, no. 11, pp. 455–462, 1998.

[48] ——, "Generalization in interactive networks: The benefits of inhibitory competition and hebbian learning," *Neural Computation*, vol. 13, no. 6, pp. 1199–1241, 2001.

[49] A. G. Ororbia II, P. Haffner, D. Reitter, and C. L. Giles, "Learning to adapt by minimizing discrepancy," *arXiv preprint arXiv:1711.11542*, 2017.

[50] K. Friston, "The free-energy principle: a rough guide to the brain?" *Trends in cognitive sciences*, vol. 13, no. 7, pp. 293–301, 2009.

[51] E. Yehene, N. Meiran, and N. Soroker, "Basal ganglia play a unique role in task switching within the frontal-subcortical circuits: evidence from patients with focal lesions," *Journal of Cognitive Neuroscience*, vol. 20, no. 6, pp. 1079–1093, 2008.

[52] T. J. Buschman and E. K. Miller, "Goal-direction and top-down control," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1655, p. 20130471, 2014.

[53] J. L. McClelland, B. L. McNaughton, and R. 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*, vol. 102, no. 3, p. 419, 1995.

[54] S. Grossberg, "Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world," *Neural networks*, vol. 37, pp. 1–47, 2013.

[55] D. Lopez-Paz *et al.*, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6470–6479.

[56] S. Lee, J. Kim, J. Ha, and B. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," *CoRR*, vol. abs/1703.08475, 2017. [Online]. Available: http://arxiv.org/abs/1703.08475

[57] J. Serrà, D. Surís, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," *CoRR*, vol. abs/1801.01423, 2018. [Online]. Available: http://arxiv.org/abs/1801.01423

[58] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[59] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[60] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *arXiv preprint arXiv:1706.08840*, 2017.

[61] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[62] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.

[63] Y. Liu, Y. Su, A.-A. Liu, B. Schiele, and Q. Sun, "Mnemonics training: Multi-class incremental learning without forgetting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 245–12 254.

[64] X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong, "Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting," *arXiv preprint arXiv:1904.00310*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations are discussed in the Appendix/Supplementary Material.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Potential negative societal impacts of the work are discussed in the conclusion and in the Appendix/Supplementary Material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code and the data are proprietary. We will provide complete codebase to support this paper upon acceptance. We do provide detailed pesudocode of our model, describing overall flow which can be used to re-create our model.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report one standard deviation of uncertainty (10 trials).
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Yes in appendix we provide details about the compute used to run all these experiments.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Yes, we do cite all the existing assests in our paper as well as in our codebase.
   (b) Did you mention the license of the assets? [Yes] We do mention licensing in our codebase.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] In the appendix, we provide details for creating the special challenge benchmark of this paper.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]