Learn More, Forget Less: Cues from Human Brain

Arijit Patra and Tapabrata Chakraborti(✉)

Department of Engineering Science, University of Oxford, Oxford, UK
{arijit.patra,tapabrata.chakraborty}@eng.ox.ac.uk

Abstract. Humans learn new information incrementally while consolidating old information at every stage in a lifelong learning process. While this appears perfectly natural for humans, the same task has proven to be challenging for learning machines. Deep neural networks are still prone to catastrophic forgetting of previously learnt information when presented with information from a sufficiently new distribution. To address this problem, we present NeoNet, a simple yet effective method that is motivated by recent findings in computational neuroscience on the process of long-term memory consolidation in humans. The network relies on a pseudorehearsal strategy to model the working of relevant sections of the brain that are associated with long-term memory consolidation processes. Experiments on benchmark classification tasks achieve state-of-the-art results that demonstrate the potential of the proposed method, with improvements in additions of novel information attained without requiring to store exemplars of past classes.

Keywords: Pseudorehearsal · Continual learning · Catastrophic forgetting

1 Introduction

Humans learn continually throughout life in small steps: we acquire and consolidate new knowledge through abstract representations in the context of existing knowledge. The idea of 'lifelong learning' [1], though natural to humans, has proven difficult to replicate in connectionist architectures like deep networks, where there is a tendency of losing the representation of a learned distribution when presented with data from a different distribution. This issue of 'catastrophic forgetting' is not only encountered when learning a new task, but even with the same task under conditions such as addition of new classes of data [2].

Multiple studies [1,3] have established the stability-plasticity dilemma to be a central tenet of the forgetting problem in both biological and artificial neural networks. The trade-off between stable memories from past learnt information or acquired experiences tend to be in conflict with the desired plasticity towards absorption of new knowledge in neural pathways [4]. Recent experiments

A. Patra and T. Chakraborti—Both authors are equally contributed.

H. Ishikawa et al. (Eds.): ACCV 2020, LNCS 12625, pp. 187–202, 2021.
https://doi.org/10.1007/978-3-030-69538-5_12

in computational neuroscience were able to shed light on this phenomenon, and established that the formation of stable memories in the brain happens without significant conflict with acquisition of new short-term memories by a process called long-term consolidation [5]. In this hypothesis, memory consolidation happens over varying time horizons. It links primarily three regions in the brain - the hippocampus, which deals with the processing of immediate information, which then associates learnt features to a region called prefrontal cortex that consolidates very recent memories ('working memory'), and a third region called the neocortex assembles memories from the prefrontal cortex to form stable long-term reservoirs of learnt knowledge, with the hippocampus being able to independently access the neocortex for matching tasks between novel arrivals of sensory inputs to old stable memories, completing a three-stage closed loop [6].

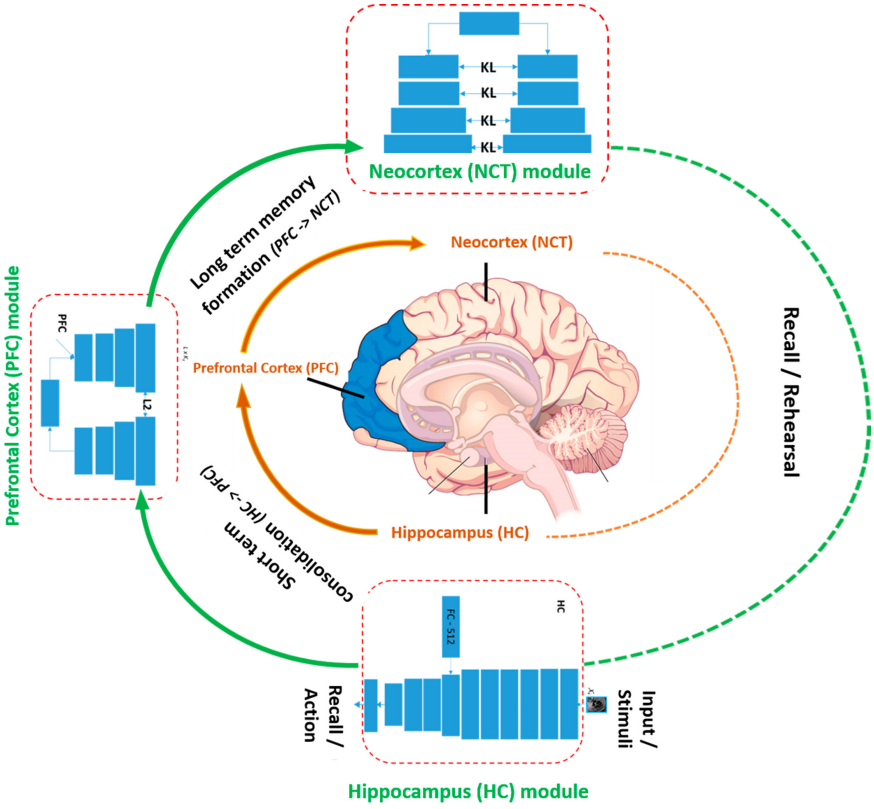

Fig. 1. Schematic diagram of proposed NeoNet and correspondences with brain regions. Structures of the mammalian brain relevant to the study are labelled in orange and the corresponding neural network modules in green. Memory formation pathways are bold arrows (orange for brain pathways, green for the brain-inspired modules); recall pathways are dotted. Detailed diagrams of modules are included in later sections. (Color figure online)

While neural networks used in vision are not exact replica of actual neural pathways by a long shot, there have been analogous design choices over the years. The stability-plasticity balance in artificial neural networks has often been described as analogous to that hypothesized for mammalian brains [7] as discovered experimentally for the latter in memory modelling experiments [4,8]. Can models based on discoveries in neuroscience regarding memory functions of the human brain help to design neural networks better and mitigate the problem of forgetting during continual learning? This is not a new problem in machine vision, but still an open problem. Can the theories of complimentary learning in the neocortex and the pre-frontal cortex help achieve this goal? The present work explores this possibility by modelling multi-stage recall mechanisms along with the learning tasks, similar to the three-stage model for long term consolidation [6]. For the first stage where new information is obtained and decisions are made on sensory inputs, the hippocampus is primarily responsible in the brain, with information on past knowledge through memory recall from the prefrontal cortex and the neocortex. We model the hippocampus with a classifier that has access to the incoming inputs, and is responsible for the classification task. The memory modules are modelled as autoencoders that can be trained to generate prior representations to serve as 'snapshots' of previously seen information. This analogous design is summarized in Fig. 1.

1.1 Related Work

In neuroscience studies, the dual-memory theory looks at the hippocampus and the cortex as key units towards knowledge absorption from the environment. This has been studied in memory retention evaluations in patients of anterograde and retrograde amnesia [9]. Studies on long-term recall from the cortex [10,11] propose a REM sleep driven consolidation mechanism [12], with learnt knowledge being subsequently overwritten from the hippocampus. Advances in neuroscience regarding the functioning of the human memory in context of information acquisition, consolidation, storage and retrieval, has influenced the design of recent bio-inspired neural network solutions to forgetting. GeppNet [13] and FearNet [14] are two such examples.

GeppNet introduced by Gepperth and Karaoguz in 2016 [13], it is a bio-inspired network that reorganizes the input onto a two-dimensional lattice via a self-organising map (SOM) to form a "long term memory", which is then used by a linear classifier. GeppNet performs rehearsal on all previous training data, plus if sufficiently new data is presented at any stage, the SOM is updated accordingly, otherwise left as is, thus avoiding forgetting older data easily. GeppNet+STM is a variant that employs a memory buffer to store new samples, such that the oldest sample is removed when presented with a new sample. The main difference between the two models is that GeppNet+STM only re-trains in specific time intervals, and in between those intervals any new labeled incoming data is stored in the buffer. Thus GeppNet+STM is better at storing old data, since the original GeppNet is updated whenever new labeled data come in.

FearNet, introduced by Kemker and Kanan in 2017 [14] and published in 2018, draws inspiration from fear conditioning in mice [8] and presents a pseudorehearsal scheme to approximate the recall of memories from the median pre-frontal cortex (mPFC). They design a module inspired by the hippocampus to perform novelty detection on incoming data [15], and encode learnt representations into the mPFC inspired autoencoder during 'sleep' stages. FearNet considers only feature representations and thus all models are fully-connected and use feature representations in the form of ResNet-50 embeddings directly.

1.2 Research Gap and Solution Approach

Limitation of Existing Models. The main constraint of FearNet (2018) is that the entire memory consolidation occurs in the PFC module modelled on pre-frontal cortex of the brain. However, recent studies on memory consolidation of the brain suggest that it occurs in multiple-stages: first, there is a plastic storage of short term memory at the pre-frontal cortex, but there is a further consolidation of long term memory in the neocortex. The latter part is not modelled in FearNet, thus not taking the advantage of plasticity-stability balance in the two sections of the brain, which if suitably exploited, can be expected to yield higher robustness against forgetting during incremental learning. GeppNet (2016) also suffers from similar drawbacks. The use of a single module inspired by the pre-frontal cortex causes the stability-plasticity balance to be governed with a single module. Thus, the representations learnt during initial learning sessions are substituted over time. In the absence of a more stable retention module, the quality of rehearsal representations declines over successive iterations, causing diminishing performance over longer horizons of incremental class addition. We find that inclusion of a separate module for stable long-term knowledge retention, inspired by the stable memory consolidation in the Neocortex (NCT), addresses the issue of diminishing retention performance over longer incremental task sets.

Contributions of Proposed Model. The proposed NeoNet adds the model of the neocortex along with associated changes in the training protocol to enforce a division of primary responsibility towards plasticity and stability between the two generative components (unlike in prior art related to pseudorehearsal and generative approaches) through two main contributions as follows:

1. A *Neocortex (NCT) Module* is added to take into account that human memory consolidation occurs in stages in different parts of the brain: short term memory in pre-frontal cortex and long term memory in neocortex. We hypothesize that in incremental learning scenarios where new class data can arrive over extended time in a large number of increments, knowledge retention and a consequent generation of suitable rehearsal exemplars will become progressively challenging. Thus, we need stable memory components that are relatively unaffected when adapting to more recent information.

2. A *Multi-stage Pseudo-Rehearsal* process is proposed to improve the training regime. This helps to strike a balance between accommodating new knowledge while maintaining previous knowledge: plasticity in the pre-frontal cortex and stability in the neocortex.

2 Methodology

In this section, we explain first the main new functionalities of the proposed network and the benefits thereof. Then we go on to describe the technical details of the main modules that constitute the network architecture.

2.1 Main Functionalities of the Proposed Model

1) Short-long term memory balance

Model components are motivated by the process of memory consolidation from short-term working memory to long-term memory and subsequent reconsolidation processes. In mammals, such systems-level consolidation has been observed to occur in a circuit beginning with immediate processing in the hippocampus, followed by transfer of information to the pre-frontal cortex which acts as the primary reservoir of working memory. Finally, long-term consolidation occurs over extended time by transferring memories to the synaptically stable neocortex from the plastic prefrontal cortex [5]. This biological process serves as an analogy for the proposed machine learning method.

We consider the hippocampal processing to represent the detection of novel classes in input data streams (Step 1, Fig. 2). The storage of temporary working memory (which occurs to a lesser degree in the hippocampus as well, but is confined to the PFC in our task for simplicity) is implemented in an autoencoder module inspired by the pre-frontal cortex. The encoding of seen classes happens immediately after the HC training for the session is completed as class exemplar representations are extracted in the form of first fully-connected logits and treated as inputs and target outputs for the PFC autoencoder (Step 2, Fig. 2). Over the next sessions, when the HC module is being adapted to data from new classes, the PFC decoder utilizes the stored class mean and the diagonalized covariance matrices to generate class feature representations which are used to encode stable long term memories into the encoder-decoder architecture representative of the neocortex (NCT) (Step 3, Fig. 2). This second encoding process is carried out in parallel with the HC training stages beyond the first session.

2) Stability-plasticity balance through multi-stage pseudo-rehearsal

The intuition behind having two stages of pseudo-rehearsal is that the transfer and consolidation of information over several stages from plastic to stable memories [5]. Computationally, such a pipeline enables us to implement a distributed stability-plasticity balance. The neocortex (NCT) module imposes a relatively strict regularisation on its parameters in terms of the encoding-decoding processes and thus represents stability in memory storage. The pre-frontal cortex

(PFC) module allows for a more flexible encoding of exemplar features over multiple sessions with relatively weak guarantees of parameter preservation resulting in a comparatively relaxed regime, thus representing plasticity in memory storage.

Sequential learning is performed as a sequence of training sessions over which the architecture needs to demonstrate competitive accuracy while retaining past knowledge. There are N sessions considered, each with K classes consisting of a variable number of instances. Feature representations from each class in a session are used to compute class means and covariance matrices that are utilised for pseudo generative replay in later stages.

1. *HC module:* In a training session M $(1 < M \leq N)$, the HC module learns a classification task on available classes. Note that a secondary input branch is used after the convolutional layers of the HC network during incremental training to enable the introduction of generated embeddings for previously seen classes. In subsequent incremental learning stages, these embeddings are generated through the dual pseudo-rehearsal scheme.
2. *PFC module:* The learning in the HC stage is followed by extraction of representations over validation set instances per class considered. These exemplar feature representations are then used to train the autoencoder based PFC module where a reconstruction error is coupled with the classification loss to learn reproducible class reconstructions
3. *NCT module:* For long-term stable memory storage, the representations so generated are encoded on to the NCT module inspired by the neocortex. The long-term storage interval is approximated within L successive training sessions, and thus, when the HC network is trained on the M^{th} session, the encodings from the PFC are written on to the NCT, which thus far would have been exposed to representations up to the $(M - L)^{th}$ session only.

2.2 Design of Components

1) Hippocampus (HC) Module. This is designed for immediate knowledge absorption upon arrival of labeled data by means of a convolutional model enforcing a Bergman distance based classification scheme, inspired by the probabilistic framework of [15] with the minimization objective formulated as:

$$P(c|x) = \frac{z_c}{\sum_{c_1} z_{c_1}} \tag{1}$$

Where,

$$z_c = 1/\delta + \min_j |x - w_{c,j}|^2 \tag{2}$$

δ is a small correction factor to ensure boundedness, $W_{c,j}$ is the j^{th} stored example for class c, and x is the incoming sample. The architecture is schematically represented in Fig. 1, and implements a sequence of convolutional operations, followed by fully-connected layers. The network is so implemented as to be

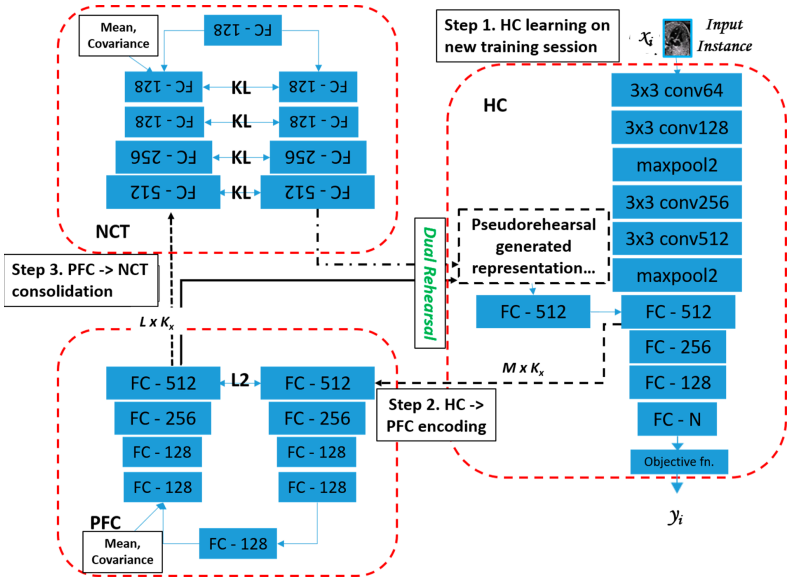

Fig. 2. Three modules of the architecture, HC (for immediate processing), PFC (for short-term storage), NCT (long-term stable storage of learnt representations) shown with encoding and rehearsal pathways.

able to take dual input: the auxiliary input layer is appended directly to the fully connected set, allowing 1-D feature representations. This allows generated exemplars of a past task to be a feature representation from the autoencoder modules instead of stored past exemplars. This is analogous to the mammalian ability to encode salient information about surroundings with very few exemplars and the ability to form associations with prior knowledge.

2) Pre-frontal Cortex (PFC) Module. Mimicking the function of its namesake in the human brain, the PFC module encodes memories of the current task while the data is still available after the HC training, with the decoder arm learning to reconstruct these exemplars with a high degree of fidelity. The encoder and decoder branches are constructed as fully-connected autoencoder layers being downsampled and upsampled respectively. To allow adaptation to novel data and allow for encoding steps to occur with a sufficient plasticity (modeling short term memory handling in human brain), the reconstruction error is used only between the finally generated representations and the input without intermediate regularization. Class mean features and class specific covariance matrices are retained in the memory and used to sample representations to be used as input to the PFC to generate pseudo-exemplars through the decoder. Such attempts at pseudo-rehearsal, proposed by [16] and revived by [14] allow generating past representations without requiring actual storage of exemplars.

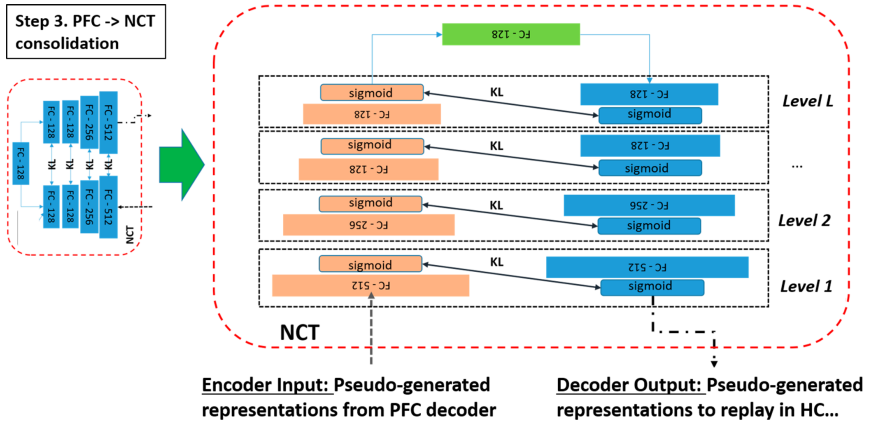

Fig. 3. Long term memory consolidation in the proposed Neocortex (NCT) module

Algorithm 1: NeoNet Training Protocol

Define parameters
- K = no. of classes, p = no. of classes per incremental learning step.
- S = max no. of increments = $K/2/p + 1$.
- L = no. of increments between HC to PFC and between PFC to NCT.

Perform base class training

- K = Base train HC with $K/2$ classes.
- S = Fine-tune HC $\rightarrow$ PFC, PFC $\rightarrow$ NCT.
- L = Retain base μ, σ for incremental learning.

Perform incremental learning
Initialise counters $M = 0; M' = 0$;
while $M < S$ **do**
  $M = M + 1$;
  while $M' < M$ **do**
    $M' = M' + 1$;
    Short consolidation: fine-tune HC $\rightarrow$ PFC;
    Long consolidation: **if** $M' < M - L$ **then** fine-tune PFC $\rightarrow$ NCT;
  end
  Update μ, σ;
  Rehearsal: Feeback NCT representation into HC;
  Adaptation: Adapt HC on incremental classes;
end

3) Neocortex (NCT) Module. In the human brain, the long term potentiation of memories (synaptic consolidation) occurs over extended time, and manifests as a transfer of relevant working memories from the pre-frontal cortex into the neocortex. The neocortex inspired NCT module is treated as a

standard autoencoder deriving reconstructed class exemplars from the PFC module to create more stable encodings. The schematic details of the NCT module along with the long term memory consolidation process therein is presented in Fig. 3 and also represented in Algorithm 1. The design considerations are as follows:

- Contrary to the relaxed regularization requirements in the PFC reconstruction to ensure plasticity, the NCT module attempts to facilitate stability in the learnt representation memories, and hence requires more specific regularization. For the choice of the intermediate regularization, we focus on the possibility to include a formulation that inherently prioritizes the preservation of the most salient learnt knowledge and ensure minimal disruption to the task specific parameter importance. Such a formulation should ideally be expressible as an information content measure.
- Because of the possible interpretation of intermediate representations as normalized values for feature salience probabilities, the corresponding stages of the encoder and decoder are mutually regularized with a KL divergence measure. This allows implicit consolidation of parameters most pertinent to such encoding generation as the KL divergence is used between the decoder arm optimized for an immediate previous set of classes and the encoder arm being exposed to the new arriving representations.
- The usage of a KL divergence approach enables an implicit measure of parameter importance because of the correspondence with the Fisher information matrix. The second derivative of the KL divergence that yields a Fisher metric in a product form with parameter perturbation squared [17].
- As the objective function minimized in this module is essentially a summation over the KL divergences computed across corresponding normalized encoder and decoder layers, the gradient descent algorithm optimizes the derivative of this summation of KL divergences.

3 Experiments and Results

3.1 Experimental Setup

Model Setup: The model architectures are shown in Fig. 2 and Fig. 3. All layers in all the modules are initialized with a Xavier scheme [18]. The number of fully-connected units at the encoder input/decoder output is kept the same for the PFC and NCT modules and equal to the first fully-connected layer in the HC to ensure compatibility of feature representations. The specific dimensions of the features were established through a grid search on the number of units, checking for suitability from 128 to 1024 units. The suitable number of layers used in the encoders and decoders and in the HC module were found for each dataset and the finally chosen configurations were the ones that performed the best on average across datasets with 50% of classes considered. The training, validation and test split was maintained at 60:20:20 in all cases.

Training: Initial training of the HC is carried for 500 epochs, followed by PFC consolidation over 200 epochs. The learning rate in the encoder is kept the same as the HC model and that of the decoder is initially kept at 1/10th of the HC learning rate and is decreased by a factor of 10 every 50 epochs. This is to ensure that the decoder arm efficiently learns to regenerate prior representations. In the NCT module, the learning rate regime is kept similar to the PFC but training is carried out for 250 epochs to ensure a stable consolidation of representations. Replay settings involve representations generated from both the PFC and NCT modules supplied together to the auxiliary input of the HC.

Baselines: Considering the class-incremental focus here, we consider the multi-class adaptation of learning without forgetting (LwF.MC), iCaRL [19], LwM [20] and incremental rebalancing classifier (ICR) [21]. Also, due to our neuroscience motivations, we also compare with FearNet [14] and GeppNet+STM (Gepp-STM) [13]. Since our test performances are evaluated using the HC module (with pseudorehearsal inputs), we modify architectures of iCaRL, LwM, ICR and LWF.MC to have the same sequence of convolutional and fully-connected layers while implementing their methods for exemplar storage and replay (details in Appendix) for a fair comparison. A cosine distance based distillation loss is implemented for the ICR baseline in line with the setting in the original paper.

Datasets: Incremental learning tasks are performed on the CIFAR 100 dataset [22], CalTech 256 [23] and CUB-200 [24] datasets. CIFAR-100, consisting of 100 classes of images of 32×32 pixels. CUB-200 has 200 classes of images of birds, curated for a fine-grained classification task. We use the 2011 version of CUB-200. CalTech-256 provides 256 categories of at least 80 images per category, including images with clutter and an overall increased difficulty compared to CIFAR-100. The datasets are chosen to evaluate the model on progressively difficult image classification tasks.

Evaluation Metrics: We adapt metrics of evaluation proposed by Kemker et al. [25] to our incremental learning performance in terms of prior task accuracy retention, and the present task accuracy. *Normalized accuracy of learning new tasks* is the average validation performance on the test data of all classes in the current session and indicates generalization ability over new data distributions. Mathematically it may be formulated as $E_{new} = \frac{1}{N-1}\sum_{2}^{N}\frac{A_{new}}{A_{all}}$, where A_{new} is the validation accuracy on the new classes seen in the most current session. *Normalized accuracy of retaining old knowledge* is the accuracy on the initial set of classes after all classes have been trained. Mathematically this may be formulated as $E_{init} = \frac{1}{N-1}\sum_{2}^{N}\frac{A_{init}}{A_{all}}$, where A_{init} is the accuracy for the classes considered on the initial training session after all sessions are completed and A_{all} is the accuracy on the validation data of all classes in the dataset. The past performance at current session is an average over all the prior sessions with respect to class-wise mean validation performances.

3.2 Results and Analysis

Quantitative Results: In the first session, a proportion of the classes are used to train the models (initially half of available classes are trained for in the first session, with subsequent ablations of 25% and 75%), followed by increments of 2, 5 and 10 classes in successive sessions. Results for the method in these settings with baselines are shown in Table 1. On the ability to preserve knowledge, the model is seen to resist losing out salient information over prior tasks, post the overall completion. In order to capture the overall dynamics of information retention, the final overall performance obtained by testing the finally obtained HC configuration on validation data across all classes has been used to normalize the metrics (A_{all} in $E_{init}$and E_{new} expressions). Thus, higher the values of E_{new}, better is the generalization on new tasks, and higher values of E_{init} imply superior knowledge retention. So, higher values for both imply the agent is better both in mitigating forgetting and improving generalization.

Table 1. Performance on CIFAR 100, Caltech 256, CUB 200

	CIFAR 100						CALTECH 256					
Model	2 classes		5 classes		10 classes		2 classes		5 classes		10 classes	
	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}
iCaRL	0.912	0.807	0.903	0.795	0.822	0.821	0.830	0.582	0.837	0.608	0.813	0.612
LwF.MC	0.796	0.752	0.813	0.764	0.817	0.832	0.653	0.547	0.681	0.566	0.692	0.580
LwM	0.857	0.705	0.783	0.805	0.876	0.813	0.847	0.661	0.853	0.655	0.825	0.657
FEL	0.801	0.814	0.797	0.820	0.809	0.836	0.773	0.672	0.784	0.658	0.861	0.603
FearNet	0.929	0.820	0.937	0.802	0.941	0.829	0.873	0.673	0.871	0.658	0.896	0.670
ICR	0.861	0.697	0.795	0.812	0.837	0.786	0.851	0.684	0.858	0.634	0.837	0.621
NeoNet	0.935	0.885	0.945	0.874	0.952	0.893	0.922	0.731	0.907	0.756	0.918	0.768

	CUB 200					
Model	2 classes		5 classes		10 classes	
	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}
iCaRL	0.874	0.573	0.792	0.610	0.881	0.601
LwF.MC	0.638	0.471	0.743	0.541	0.675	0.572
LwM	0.858	0.612	0.853	0.632	0.840	0.629
FEL	0.703	0.784	0.710	0.682	0.673	0.791
FearNet	0.879	0.637	0.883	0.677	0.902	0.683
ICR	0.825	0.570	0.794	0.630	0.805	0.675
NeoNet	0.913	0.721	0.923	0.668	0.954	0.775

One of the objectives of proposing a separation between primarily plastic and primary stable generative modules was to ensure that long-term consolidation of the learnt representations can be effectively accomplished. This is necessary for adequate retention when adding a new classes over a large number of incremental

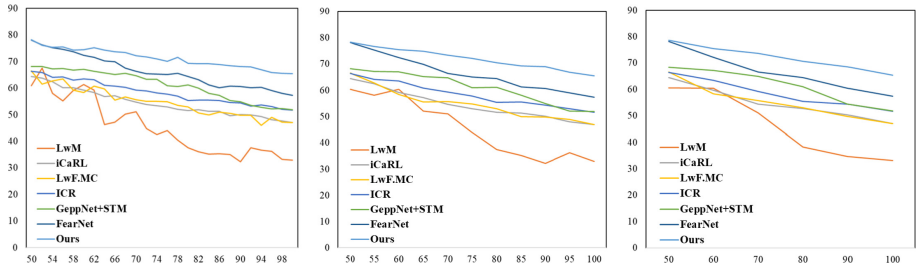

Fig. 4. Mean test accuracy on already learnt classes for the CIFAR 100 experiments, upon new increments of 2 classes (left), 5 classes (middle) and 10 classes (right). Number of base classes is 50. The X-axis shows the number of classes the model has been exposed to at that point, and the Y-axis shows the mean accuracy on all these classes.

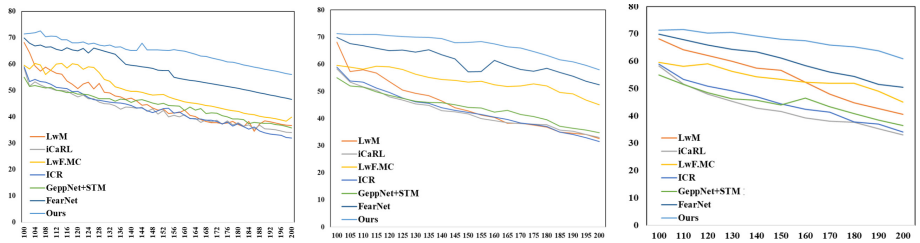

Fig. 5. Mean test accuracy on already learnt classes for the CUB 200 dataset, upon new increments of 2 classes (left), 5 classes (middle) and 10 classes (right). Number of base classes is 100. X-axis shows the number of classes the model has been exposed to at that point, Y-axis shows mean accuracy on all these classes. The decline in mean accuracy is much slower for the NeoNet in the 2-class and 5-class long-range incremental additions, due to the separation of stable and plastic generative components.

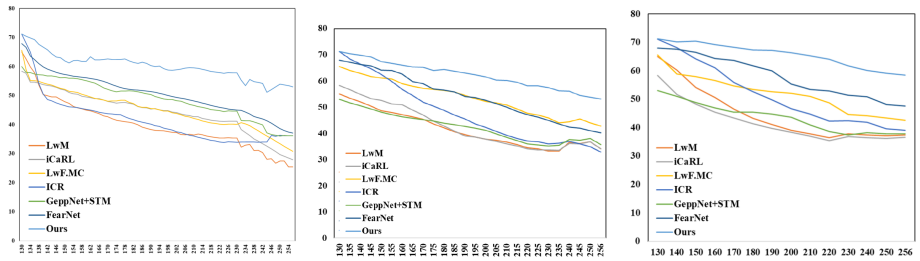

Fig. 6. Mean test accuracy on already learnt classes for Caltech 256 dataset, upon new increments of 2 classes (left), 5 classes (middle) and 10 classes (right). Number of base classes is 112. The decline in mean accuracy is much slower for the NeoNet in the 2-class and 5-class long-range incremental additions, due to the separation of the stable and plastic generative components.

sessions, such as the cases where 2 classes per stage are added beyond the base class training (leading to 25 incremental stages for the CIFAR 100, 50 for CUB-200, 64 for Caltech 256 when starting with base model pre-trained on 50% of classes). In such long incremental horizons, we find our model to have a much better retention of knowledge as seen by the mean class accuracies over these incremental stages (Fig. 4, 5 and 6) even with relatively difficult datasets such as CUB-200 and CalTech 256. Compared to other brain-inspired architectures that explicitly relied on a memory consolidation module, our incorporation of an NCT module to mimic long term memory formation leads to a more gradual decline in overall performance. Thus, a more robust pseudorehearsal strategy is formed by utilising concepts from long-term memory consolidation research.

Table 2. Effect of variation in PFC to NCT consolidation volumes

	CIFAR 100		CUB 200		CALTECH 256	
	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}
L = 0.25M	0.922	0.823	0.902	0.653	0.899	0.677
L = 0.50M	0.947	0.828	0.910	0.649	0.907	0.672
L = 0.75M	0.949	0.826	0.917	0.649	0.910	0.671

Table 3. Effect of variation in NCT regularization extent

	CIFAR 100		CUB 200		CALTECH 256	
	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}
KL (1+2+3+4)	0.947	0.828	0.910	0.649	0.907	0.672
KL (1+2+3)	0.902	0.813	0.869	0.637	0.871	0.671
KL (1+2)	0.887	0.814	0.843	0.629	0.835	0.672
KL (2)	0.801	0.809	0.782	0.626	0.779	0.668

Table 4. Effect of variation in initial session base class volume

	CIFAR 100		CUB 200		CALTECH 256	
	E_{init}	E_{new}	E_{init}	E_{new}	E_{init}	E_{new}
Base = 25%	0.912	0.819	0.892	0.636	0.903	0.668
Base = 50%	0.946	0.825	0.910	0.643	0.902	0.672
Base = 75%	0.948	0.837	0.913	0.648	0.908	0.689

Ablation Studies:

- *Table 2:* We evaluate the balance in plasticity and stability between the PFC and NCT modules by varying the number of classes L to be written in a session from the PFC to the NCT between 1 to M, which is the total number of classes available in a session. In the 10-stage class increment experiment with half of classes available at the initial training, we consider L to be 25%,50% and 75% of M ($M = 10$ here) and report the results in Table 2. The differential consolidation implemented by varying L impacts the overall knowledge retention of the model as seen by the decline in E_{init} being inversely proportional to the percentage of classes per session that is directly transferred from PFC to NCT. Over multiple sessions new classes arrive and the PFC being relatively plastic preferentially adapts parameters to new distributions. So its decoder is relatively impaired in its ability to generate representations of previous classes as compared to those generated right after the original encoding session. This causes the efficiency of rehearsal to diminish over multiple sessions. New class accuracies are practically unaffected by this change as the HC module is directly validated on the data without reliance on external exemplars.
- *Table 3:* In the NCT module, varying the extent of KL divergence measure, can affect the quality of the generated representations and the ability of the model to retain important weights close to optimal values. This affects the overall ability to preserve prior information and is evident in the alterations in performance on recall of past tasks. For this experiment, we consider the setting of 50% classes being available at initial training and L set at 50% on 10-class increment stages. There are four levels of the KL regularization corresponding to the four sets of encoder decoder layers. It is observed in Table 3 that the post completion base knowledge accuracy shows a steady decline with the removal of KL divergence regularization stages in NCT, as evident from the steadily declining E_{init}values. Adaptations to new data remain relatively unaffected as the rehearsal only sporadically impacts new representation learning except in cases where the joint training with past exemplars leads to particularly optimal initializations allowing for improvements in learning on new data.
- *Table 4:* We show the overall effects of changing the proportion of classes trained for in the initial training session. Unlike in Table 2, where the assumption was that 50% of the classes in a dataset are trained for in the first session itself, here we consider initial availability of 25% and 75% with the remaining subjected to 10-class incremental stages. The last stage accounts for the remaining classes, which may be less than 10, thus the discrepancy in the 10-class stage metrics for the otherwise similar experimental state in Table 2. The inclusion of more classes in the initial session is seen to improve the final knowledge retention. The learning regime is seen to prioritize initial sessions more than subsequent replays and a more voluminous base knowledge translates into higher final accuracies.

4 Conclusion

We show that incorporating recent understanding of multi-stage short-long term human memory consolidation into deep incremental/continual learning helps in limiting forgetting previously learnt information, especially when presented with incrementally arriving significant numbers of new class sessions. We do this by designing modules of a deep architecture based on three sections of the brain: Hippocampus for initial processing of incoming data and working memory, Prefrontal cortex for short term memory with plasticity and Neocortex for long term memory with stability. These modules work in tandem and improved results are obtained in standard incremental learning experiments against benchmark methods on public datasets.

References

1. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: a review. Neural Netw. 113, 54–71 (2019)
2. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv:1312.6211 (2013)
3. Zhang, L., et al.: A simplified computational memory model from information processing. Sci. Rep. **6**, 37470 (2016)
4. French, R.M.: Catastrophic forgetting in connectionist networks. Trends Cogn. Sci. **3**, 4 (1999)
5. Fiebig, F., Lansner, A.: Memory consolidation from seconds to weeks: a three-stage neural network model with autonomous reinstatement dynamics. Front. Comput. Neurosci. **8**, 64 (2014)
6. Manohar, S.G., Zokaei, N., Fallon, S.J., Vogels, T., Husain, M.: Neural mechanisms of attending to items in working memory. Neurosci. Biobehav. Rev. 101, 1–12 (2019)
7. Mermillod, M., Bugaiska, A., Bonin, P.: The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. Front. Psychol. **4**, 504 (2013)
8. Kitamura, T., et al.: Engrams and circuits crucial for systems consolidation of a memory. Science **356**, 6333 (2017)
9. Marslen-Wilson, W.D., Teuber, H.L.: Memory for remote events in anterograde amnesia: recognition of public figures from newsphotographs. Neuropsychologia **13**, 353–364 (1975)
10. Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., Miyashita, Y.: Top-down signal from prefrontal cortex in executive control of memory retrieval. Nature **401**, 699 (1999)
11. Maddock, R.J., Garrett, A.S., Buonocore, M.H.: Remembering familiar people: the posterior cingulate cortex and autobiographical memory retrieval. Neuroscience **104**, 667–676 (2001)
12. Siegel, J.M.: The rem sleep-memory consolidation hypothesis. Science **294**, 1058–1063 (2001)
13. Gepperth, A., Karaoguz, C.: A bio-inspired incremental learning architecture for applied perceptual problems. Cogn. Comput. **8**, 5 (2016)

14. Kemker, R., Kanan, C.: FearNet: brain-inspired model for incremental learning. arXiv:1711.10563 (2017)
15. Specht, D.F.: Probabilistic neural networks. Neural Netw. **3**, 109–118 (1990)
16. Robins, A.: Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Sci. **7**, 123–146 (1995)
17. Dabak, A.G., Johnson, D.H.: Relations between Kullback-Leibler distance and Fisher information. Technical report (2002)
18. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
19. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: incremental classifier and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2001–2010 (2017)
20. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5138–5146 (2019)
21. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 831–839 (2019)
22. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
23. Griffin, G., Holub, A., Perona, P.: Caltech-256 Object Category Dataset. California Institute of Technology (2007)
24. Welinder, P., et al.: Caltech-UCSD birds 200, California institute of technology. CNS-TR- 2010–001 (2010)
25. Kemker, R., McClure, M., Abitino, A., Hayes, T., Kanan, C.: Measuring catastrophic forgetting in neural networks. In: AAAI Conference on Artificial Intelligence (2018)