PNAS RESEARCH ARTICLE | NEUROSCIENCE OPEN ACCESS

Learning in deep neural networks and brains with similarity-weighted interleaved learning

Rajat Saxena[a], Justin L. Shobe[a], and Bruce L. McNaughton[a,b,1]

Edited by György Buzsáki, New York University Grossman School of Medicine, New York, NY; received September 29, 2021; accepted May 24, 2022

Understanding how the brain learns throughout a lifetime remains a long-standing challenge. In artificial neural networks (ANNs), incorporating novel information too rapidly results in catastrophic interference, i.e., abrupt loss of previously acquired knowledge. Complementary Learning Systems Theory (CLST) suggests that new memories can be gradually integrated into the neocortex by interleaving new memories with existing knowledge. This approach, however, has been assumed to require interleaving all existing knowledge every time something new is learned, which is implausible because it is time-consuming and requires a large amount of data. We show that deep, nonlinear ANNs can learn new information by interleaving only a subset of old items that share substantial representational similarity with the new information. By using such similarity-weighted interleaved learning (SWIL), ANNs can learn new information rapidly with a similar accuracy level and minimal interference, while using a much smaller number of old items presented per epoch (fast and data-efficient). SWIL is shown to work with various standard classification datasets (Fashion-MNIST, CIFAR10, and CIFAR100), deep neural network architectures, and in sequential learning frameworks. We show that data efficiency and speedup in learning new items are increased roughly proportionally to the number of nonoverlapping classes stored in the network, which implies an enormous possible speedup in human brains, which encode a high number of separate categories. Finally, we propose a theoretical model of how SWIL might be implemented in the brain.

complementary learning systems | learning | memory | neural networks | memory consolidation

Significance

Unlike humans, artificial neural networks rapidly forget previously learned information when learning something new and must be retrained by interleaving the new and old items; however, interleaving all old items is time-consuming and might be unnecessary. It might be sufficient to interleave only old items having substantial similarity to new ones. We show that training with similarity-weighted interleaving of old items with new ones allows deep networks to learn new items rapidly without forgetting, while using substantially less data. We hypothesize how similarity-weighted interleaving might be implemented in the brain using persistent excitability traces on recently active neurons and attractor dynamics. These findings may advance both neuroscience and machine learning.

Artificial neural networks (ANNs) tend to lose previously acquired knowledge abruptly when new information is incorporated too quickly ("catastrophic interference") (1, 2). Successful lifelong learners (e.g., humans) do not suffer from this problem, potentially by using mechanisms suggested in the Complementary Learning Systems Theory (CLST) (3) (see also ref. 4). CLST states that the brain relies on complementary learning systems: the hippocampus (HC) for rapid acquisition of new memories and the neocortex (NC) for the gradual incorporation of the new data into context-independent structured knowledge. During "offline periods," such as sleep and quiet awake rest, the HC triggers replay of recent experiences in the NC, while the NC spontaneously retrieves and interleaves representations of existing classes (5–7). The interleaved replay allows gradual adjustment of NC synaptic weights, in a gradient-descent manner, to create context-independent category representations, thereby gracefully integrating new memories and overcoming catastrophic interference. Numerous studies have since successfully used interleaved replay to achieve lifelong learning in neural networks (8, 9).

In practice, however, the CLST raises two significant issues. First, how can the brain possibly perform a comprehensive interleaving when it does not have access to all the old data? One potential solution is "Pseudorehearsal" (10), where random inputs can elicit generative replay of internal representations without requiring explicit access to previously learned examples. Attractor-like dynamics may allow the brain to accomplish pseudorehearsal, but it is unclear what to pseudorehearse. Thus, the second problem is that there is not enough time to interleave all of the previously learned information after each new learning event. "Similarity Weighted Interleaved Learning" (SWIL) was proposed as a solution to this second problem, suggesting that it may be sufficient to interleave only old items with substantial representational similarity to new items (11). Empirical behavioral studies showed that highly consistent new items could be rapidly integrated into NC structured knowledge with little or no interference (12, 13). This indicates that the speed of integrating new information depends on its consistency with the prior knowledge (14). Inspired by this behavioral result, and by a reexamination of the distribution of catastrophic interference among previously acquired classes, which is described below, McClelland et al. (11) demonstrated that SWIL allowed learning new information using 2.5x fewer item presentations per epoch in a simple dataset with two superordinate

Author affiliations: [a]Department of Neurobiology and Behavior, University of California, Irvine, CA 92697; and [b]Canadian Centre for Behavioural Neuroscience, The University of Lethbridge, Lethbridge, Alberta T1K 3M4, Canada

Author contributions: R.S. and B.L.M. designed research; R.S. performed research and analyzed data with assistance from J.L.S.; and R.S. and B.L.M. wrote the paper with inputs from J.L.S.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: brucemcn@uci.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2115229119/-/DCSupplemental.

Published June 27, 2022.

categories and achieved the same performance as training the network on the entire data. However, the authors did not find a similar effect when using more complex datasets, raising concerns about the algorithm's scalability.

The current study has overcome these limitations by modifying the SWIL algorithm to work with Convolutional Neural Networks (CNNs) on traditional classification datasets (Fashion-MNIST, CIFAR10, and CIFAR100). We exploit the hierarchical structure of existing knowledge to selectively interleave only the old items that have higher representational similarity to new items. With this strategy, we can reach performance levels comparable to that achieved by using the entire training dataset, thereby substantially reducing the amount of data required (data-efficient) and learning time (speedup). We then show that SWIL can also be used in a sequential learning framework. Additionally, we show that learning a new class can be extremely data-efficient—i.e., a much smaller number of old items being presented—if it shares similarities with far fewer previously learned classes, which is likely the case in human learning. Finally, we present a theoretical model of how SWIL might be implemented in the brain using previously stored attractors with an excitability bias proportional to their overlap with new items.

Learning Dynamics in Deep Neural Networks for Image-Classification Dataset

McClelland et al. showed that, in a deep linear network with one hidden layer, SWIL allows learning a new class similarly to fully interleaved learning (FIL)—i.e., interleaving the entire old classes with the new class—but using 40% fewer items (11). However, the network was trained on a very simple dataset, with only two superordinate categories, raising questions regarding the algorithm's scalability. We started by exploring how learning on different classes evolves in a deep linear neural network with one hidden layer (*SI Appendix*, Fig. 1*A*) on a more complex dataset: Fashion-MNIST (15). The model was first trained to 87% total test accuracy on 8 of the 10 classes, omitting the "boot" and "bag" classes (*SI Appendix*, Fig. 1*B*). We then retrained the model to learn the (new) "boot" class under two different conditions, with 10 repetitions per condition: 1) focused learning (FoL)—only new "boot" class presented—and 2) FIL—all the classes (new + previously learned) presented with equal probability. A total of 180 images were presented per epoch for both conditions (same images in each epoch). The network was tested on a total of 9,000 previously unseen images (test dataset; 1,000 images per class), excluding the bag class. The training was stopped when the network's performance reached asymptote. As expected, FoL caused interference with old classes, which was overcome with FIL (Fig. 1, second column). As alluded to above, interference with the old data in FoL varied across classes, which was part of the original inspiration for SWIL, and suggests a graded similarity relationship between the new "boot" class and the old classes. For example, recall on the "sneaker" and "sandals" falls off faster than the "trouser," perhaps because integrating the new "boot" class would selectively change synaptic weights representing the "sneaker" and "sandals" class, causing more interference.

Computing Similarity between Different Classes

The reduction in performance was higher for similar old classes on learning new items for FoL. This relation between learning and similarity between multiple class attributes was explored previously (11), and it was shown that a deep linear network could acquire already-known consistent attributes rapidly. In contrast, the inconsistent attributes that needed the addition of a new branch in the existing class hierarchy required slow, gradual, interleaved learning. In the current work, we computed similarity at the feature level using published methods (16, 17). Briefly, we computed cosine similarity between the average per-class activation vectors for existing- and new-class items for a target hidden layer (typically, the penultimate layer in these simulations; *Materials and Methods*). Fig. 2*A* shows the similarity matrix calculated from the penultimate-layer activations for the pretrained network on new "boot" and old classes of the Fashion-MNIST dataset. The similarity between classes is consistent with our visual perception of objects. For example, a higher similarity between the "boot" class and the "sneaker" and "sandal" classes and between "shirt" and "t-shirt" classes, etc., can be observed in the hierarchical clustering plots (Fig. 2*B*). The similarity matrix corresponds strongly to the confusion matrix generated at the end of FIL from the previous section (Fig. 2*C*). Higher similarity leads to more confusion; for example, "shirt" class images get confused with "t-shirt," "pullover," and "coat" classes, suggesting that our similarity

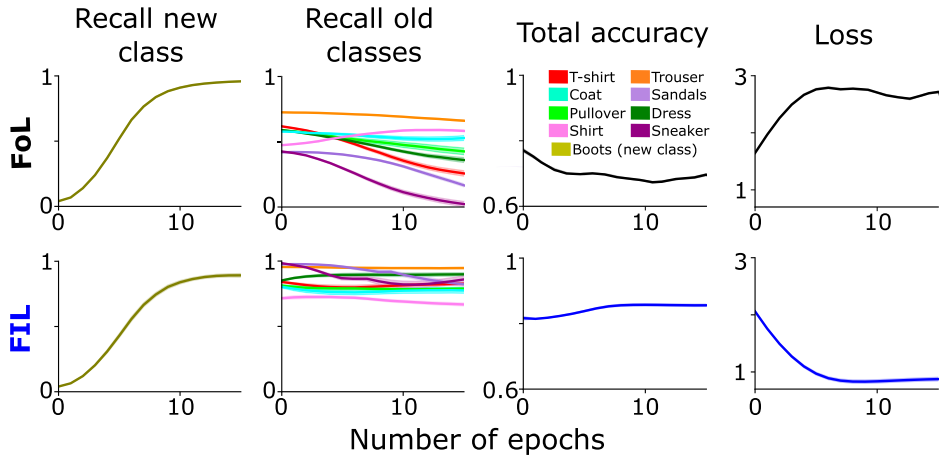

Fig. 1. Pretrained network's performance on learning new "boot" class in two conditions: FoL (*Upper*) and FIL (*Lower*). Recall on new "boot" class (olive green), recall on the existing classes (plotted in different colors), total accuracy (a high score means low error), and cross-entropy loss (a measure of total error), respectively, are shown as a function of the number of epochs on the held-out test dataset. Each plot shows the mean over 10 repetitions; shaded areas are ±1 SEM.

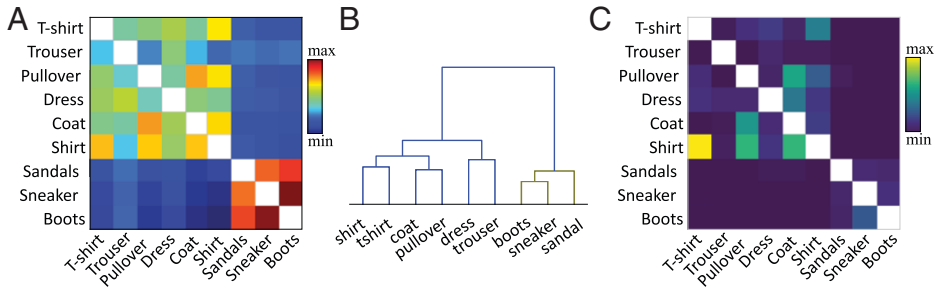

Fig. 2. (*A*) Similarity matrix for existing classes and new "boot" class of the pretrained network using the penultimate-layer activations. Diagonal values (same-class similarity plotted as white) were removed. (*B*) Agglomerative hierarchical clustering applied on the similarity matrix in *A*. (*C*) Confusion matrix for FIL after training to learn the "boot" class. Diagonal values were removed for scaling clarity.

measure predicts the learning dynamics of the neural network. A similar class-similarity profile is present in the recall curve of old classes in the FoL results described in the previous section (Fig. 1; recall similar old classes). FoL of the "boot" class leads to rapid forgetting of similar old classes ("sneaker" and "sandal") compared to different old classes ("trouser," etc.).

Rapid and Data-Efficient Learning of New Items in Deep Linear Neural Networks

Next, we examined novel class learning dynamics in three new conditions, along with the two previous ones, with 10 repetitions per condition: 1) FoL (total $n = 6{,}000$ images per epoch); 2) FIL (total $n = 54{,}000$ images per epoch, 6,000 images/class); 3) partial interleaved learning (PIL)—a much smaller subset of images (total $n = 350$ images per epoch, ~39 images/class) with images from each class (new + existing) presented with equal probability; 4) SWIL—retrain with the same total number of images per epoch as PIL, but the existing class images were weighted according to similarity with the (new) "boot" class; and 5) Equally Weighted Interleaved Learning (EqWIL)—retrain with the same number of "boot" class images as SWIL, but the existing class images were weighted equally (Fig. 3*A*). The same held-out test dataset (total $n = 9{,}000$ images) described above was used. The training was stopped when the network's performance reached asymptote for each condition. New "boot" class accuracy takes longer to asymptote and reaches a lower value using PIL than FIL ($H = 7.27$, $P < 0.05$) (Fig. 3*B*, first column; Table 1, "New Class" column), although fewer items (1/150x) were presented at each epoch. For SWIL, the similarity calculation was used to determine the proportion of existing old-class items to be interleaved. Based on this, we randomly sampled input images with weighted probabilities from each old class. This led to a higher number of "sneaker" and "sandal" class images (most similar) being interleaved compared to other classes (Fig. 3*A*). We will refer to the "sneaker" and "sandal" classes as similar old classes and the rest of the old classes as different old classes, based on the dendrogram (Fig. 2*B*). With SWIL, the model learned the new "boot" class faster and with similar interference with the existing classes compared to PIL ($H = 5.44$; $P < 0.05$). Moreover, the recall on the new class, total accuracy, and loss for SWIL were comparable to FIL (Fig. 3*B*, first column; $H = 0.056$, $P > 0.05$; Table 1, "New Class" column). The learning on the new "boot" class in EqWIL was the same as SWIL, but there was a greater

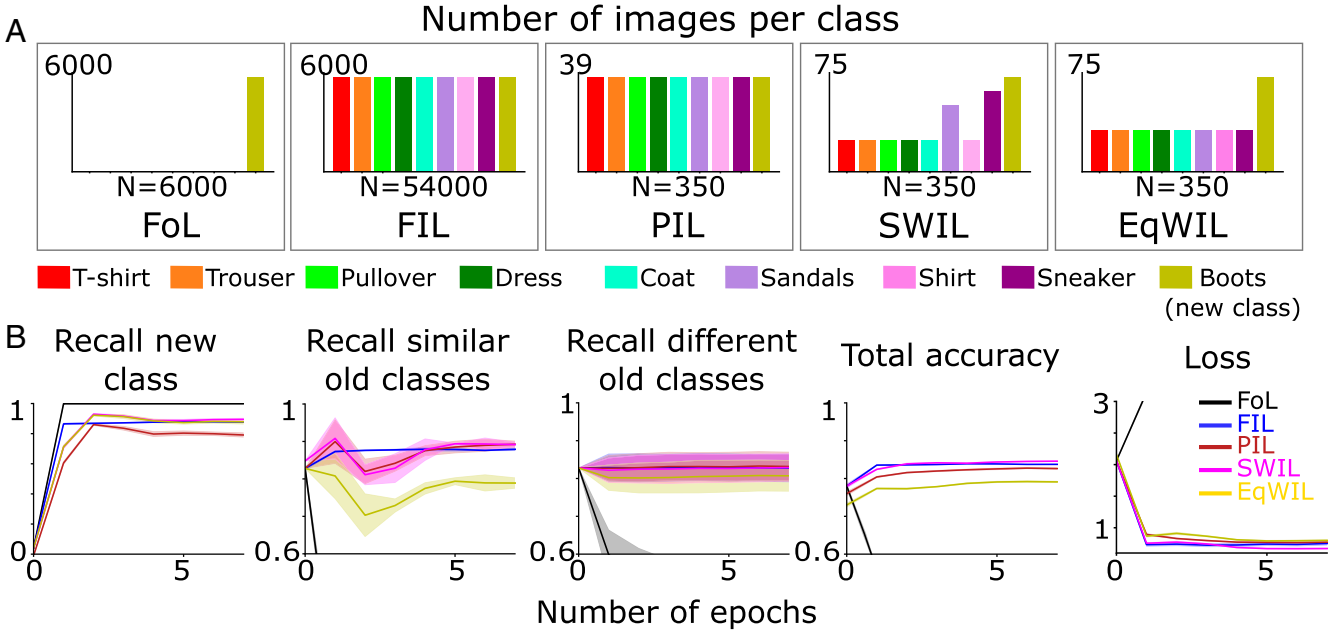

Fig. 3. (*A*) The pretrained neural network was trained to learn a new "boot" class (olive green) in five different learning conditions until the performance asymptoted: 1) FoL (total $n = 6{,}000$ images/epoch), 2) FIL (total $n = 54{,}000$ images/epoch), 3) PIL (total $n = 350$ images/epoch), 4) SWIL (total $n = 350$ images/epoch), and 5) EqWIL (total $n = 350$ images/epoch). (*B*) Recall on the new class, recall on similar old classes ("sneaker" and "sandals"), recall on different old classes, total accuracy on all classes, and cross-entropy loss for the held-out test dataset as a function of the number of epochs for FoL (black), FIL (blue), PIL (brown), SWIL (magenta), and EqWIL (gold). Each plot shows the mean over 10 repetitions; shaded areas are ±1 SEM.

Table 1. Performance on test-set at asymptote for Fashion-MNIST dataset

Learning condition	N	Recall			Total accuracy	Loss
		New class	Similar old class	Different old class		
FoL	6,000	1 ± 1e-5	0.013 ± 0.056	0.03 ± 0.189	0.13 ± 0.002	15.28 ± 0.7
FIL	54,000	0.933 ± 0.006	0.882 ± 1e-4	0.837 ± 0.038	0.852 ± 0.001	0.763 ± 0.025
PIL	350	0.822 ± 0.009	0.905 ± 0.005	0.838 ± 0.04	0.828 ± 0.002	0.782 ± 0.013
SWIL	350	0.936 ± 0.006	0.906 ± 0.004	0.835 ± 0.041	0.864 ± 0.001	0.731 ± 0.011
EqWIL	350	0.932 ± 0.008	0.785 ± 0.007	0.804 ± 0.042	0.805 ± 0.002	0.794 ± 0.011

Displayed are the means ±1 SEM over 10 repetitions for each condition.

degree of interference with the similar old classes ($H = 10.99$, $P < 0.05$) (Fig. 3*B*, second column; Table 1, "Similar Old Class" column). The following two measures were used to compare SWIL and FIL: 1) MemRed = ratio of the number of images stored in FIL and SWIL, signifying the reduction in the amount of data stored; and 2) Speedup = ratio of the total number of items presented in FIL and SWIL required to reach saturation accuracy for new-class recall, indicating the reduction in item presentations (time) needed to learn a new class. SWIL allowed learning the new item with reduced data demand, MemRed = 154.3x (54,000/350), and much faster, Speedup = 77.1x (54,000/350 × 2). Even with a smaller number of items, the model achieved the same performance by exploiting the hierarchical structure of the prior knowledge of the model using SWIL. SWIL provides a middle ground between PIL and EqWIL, allowing for the integration of a new class and minimal interference with the existing classes (both similar and different).

Learning a New Class in CNNs Using SWIL on CIFAR10

Next, to test whether SWIL would work in a more complex setting, we trained a six-layer nonlinear CNN with a fully-connected output layer (Fig. 4*A*) to recognize images from 8 different classes (except "cat" and "car") out of 10 classes in CIFAR10 (18). We retrained the model to learn the "cat" class in the five different training conditions—FoL, FIL, PIL, SWIL, and EqWIL—defined previously. Fig. 4*C* shows the distribution of images per class for the five conditions. $n = 2{,}400$ total images per epoch were presented for SWIL, PIL, and EqWIL conditions, compared to $n = 45{,}000$ and $n = 5{,}000$ images per epoch for FIL and FoL, respectively. The network was trained for each condition until the performance asymptoted. The model was tested on a total of 9,000 (held-out test dataset; 1,000 images/class, excluding the "car" class) previously unseen images. Fig. 4*B* shows the similarity matrix calculated for the CIFAR10 classes. The "cat" class is more similar to the "dog" and other animal classes falling under the same branch (Fig. 4 *B*, *Left*). We will refer to the "truck," "ship," and "plane" classes as different old classes and the rest of the old animal classes as similar old classes, based on the dendrogram (Fig. 4*B*). In FoL, the model learned the new "cat" class, but forgot the old classes. Similar to the Fashion-MNIST results, there was a gradient of interference with the "dog" class (maximum similarity with the "cat" class) and the "truck" class (minimum similarity), exhibiting maximum and minimum forgetting, respectively. As expected, FIL overcame catastrophic interference during the new "cat" class learning (Fig. 4*D*). In PIL, the model learned the new "cat" class using 18.75x fewer item presentations at each epoch, but recall for "cat" class asymptoted at a lower value than FIL ($H = 5.72$, $P < 0.05$). The recall on the new class and similar and different old classes, total accuracy, and loss for SWIL were comparable to FIL ($H = 0.42$, $P > 0.05$; Table 2; Fig. 4*D*). The recall on the new "cat" class using SWIL was higher than PIL ($H = 7.89$, $P < 0.05$). In EqWIL, learning on the new "cat" class was similar to SWIL and FIL, but there was higher interference with the similar old classes ($H = 24.77$, $P < 0.05$; Table 2). The performance on the different old classes was comparable for the FIL, PIL, SWIL, and EqWIL conditions ($H = 0.6$, $P > 0.05$). SWIL resulted in better integration of the new "cat" class than PIL and helped overcome the interference observed in EqWIL. Learning of a new-class item was much faster using SWIL than FIL; Speedup = 31.25x (45,000 × 10/2,400 × 6), while using significantly less data (MemRed = 18.75x). These results confirmed that SWIL could learn new-class items, even in nonlinear CNNs and on a more realistic dataset.

Effect of Consistency of New Items with Old Classes on Learning Time and Data Required

A new item is called consistent if it could be added to the previously learned classes without requiring large changes to the network (11). Based on this framework, learning a new class that interferes with fewer existing classes (higher consistency) can be integrated more easily into the network than a new class that interferes with multiple existing classes (lower consistency). To test this, we used the pretrained CNN from the previous section to learn a new "car" class in all five learning conditions described earlier. Fig. 5*A* shows the similarity matrix for the "car" class, which is more similar to the "truck," "ship," and "plane" classes (under the same hierarchical node) compared to the other existing classes. To further confirm, we performed *t*-distributed stochastic neighbor embedding (t-SNE) (19) on the penultimate-layer activations used for similarity calculation (Fig. 5*B*). The "car" class overlaps significantly with the other vehicle classes ("truck," "plane," and "ship"), whereas the "cat" class (from the previous section) overlaps with the other animal classes ("dog," "frog," "horse," "bird," and "deer"). As expected, FoL on the "car" class causes catastrophic interference, with the similar old classes exhibiting higher interference, which was overcome using FIL (Fig. 5*D*). A total of $n = 2{,}000$ images were presented per epoch for PIL, SWIL, and EqWIL (Fig. 5*C*). As in the previous sections, the maximal new-class recall reached is lower in PIL than FIL ($H = 12.37$, $P < 0.05$, Table 3). In SWIL, the model learned the new "car" class at similar accuracy with minimal interference with the existing classes (on both similar and different) compared to FIL ($H = 0.79$, $P > 0.05$, Table 3). Using EqWIL, the model learned the new "car" class the same as SWIL, but there was a higher degree of interference with the other similar classes, such as the "truck" ($H = 53.81$, $P < 0.05$; Table 3; Fig. 5*D*, second

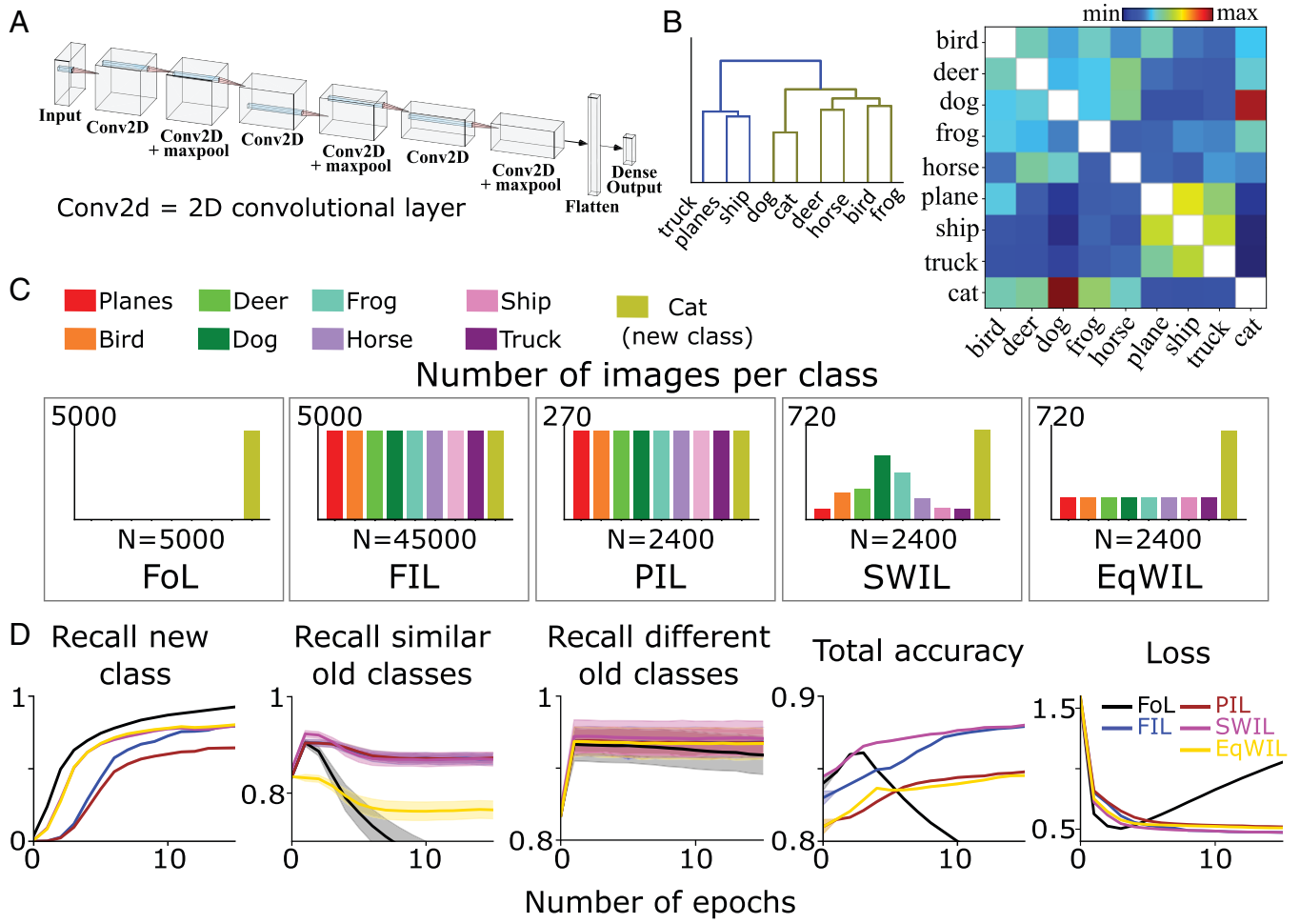

Fig. 4. (*A*) A six-layer nonlinear CNN with a fully connected output layer was used for learning eight classes of the CIFAR10 dataset. (*B*) The similarity matrix (*Right*) was calculated from the last convolution layer's activation after presenting the new "cat" class. Agglomerative hierarchical clustering (*B*, *Left*) applied to the similarity matrix showing the grouping of animal (olive green) and vehicle (blue) superclasses in the dendrogram. (*C*) The pretrained CNN was then trained to learn the "cat" class (olive green) under five different conditions, until the performance asymptoted: 1) FoL (total n = 5,000 images/epoch), 2) FIL (total n = 45,000 images/epoch), 3) PIL (total n = 2,400 images/epoch), 4) SWIL (total n = 2,400 images/epoch), and 5) EqWIL (total n = 2,400 images/epoch). The simulation for each condition was repeated 10 times. (*D*) Recall on the new class, recall on similar old classes (other animal classes in the CIFAR10 dataset), recall on different old classes ("plane," "ship," and "truck"), total accuracy on all classes, and cross-entropy loss for the held-out test dataset as a function of the number of epochs for FoL (black), FIL (blue), PIL (brown), SWIL (magenta), and EqWIL (gold). Each plot shows the mean over 10 repetitions; shaded areas are ±1 SEM.

column). SWIL allowed learning new items faster, Speedup = 48.75x (45,000 × 12/2,000 × 6), with reduced memory demand than FIL, MemRed = 22.5x. The "car" class could be learned faster and by interleaving fewer classes ("truck," "ship," and "plane") than the "cat" class (48.75x vs. 31.25x), which overlaps with a higher number of classes ("dog," "frog," "horse," "frog," and "deer"). These simulations indicate that the amount of old-class data required to interleave and speedup for learning a new class depends on the consistency of the new information with prior knowledge.

Sequential Learning Using SWIL

Next, we tested whether SWIL can be used to learn new items presented in a sequence (sequential learning framework). To do this, we took the trained CNN models from Fig. 4, the CIFAR10 "cat" class (task 1) section for both the FIL and SWIL conditions (trained on 9 out of 10 classes), and then trained the model from each condition to learn a new "car" class (task 2). Fig. 6, first column shows the distribution of individual class items used in SWIL to learn the "car" classes

Table 2. Performance on test-set at asymptote for the CIFAR10 cat dataset

		Recall				
Learning condition	N	New class	Similar old class	Different old class	Total accuracy	Loss
FoL	5,000	0.93 ± 0.3e-3	0.59 ± 0.038	0.907 ± 0.026	0.75 ± 3.5e-3	1.1 ± 1.1–3
FIL	45,000	0.789 ± 3.3e-3	0.872 ± 0.046	0.935 ± 0.021	0.880 ± 2.5e-3	0.469 ± 1.4e-3
PIL	2,400	0.64 ± 6.9e-3	0.873 ± 0.046	0.934 ± 0.021	0.847 ± 4.3e-3	0.52 ± 1.8e-3
SWIL	2,400	0.792 ± 6e-3	0.873 ± 0.054	0.936 ± 0.024	0.883 ± 2.9e-3	0.467 ± 1.6e-3
EqWIL	2,400	0.795 ± 2.7e-3	0.756 ± 0.072	0.933 ± 0.022	0.846 ± 5.6e-3	0.51 ± 1.3e-3

Displayed are the means ±1 SEM over 10 repetitions for each condition.

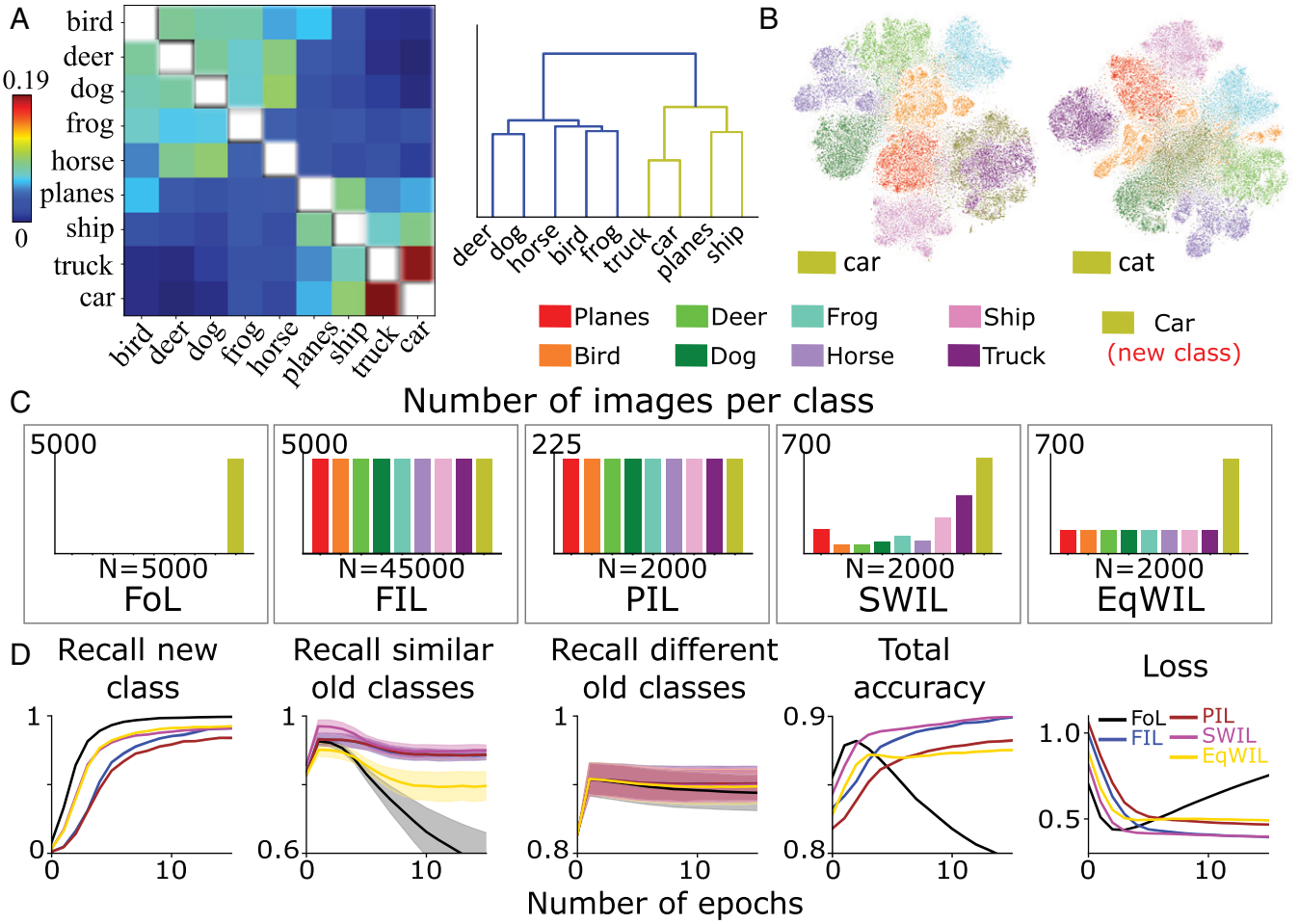

Fig. 5. (*A*) Similarity matrix (*Left*) calculated from the penultimate-layer activations and the result of agglomerative hierarchical clustering (*Right*) on the similarity matrix after the presentation of the new "car" class. (*B*) t-SNE applied on the last convolution layer activations for the "car" class (*Left*; from simulation in Fig. 4) and the "cat" class (*Right*) learning. (*C*) The pretrained CNN was trained to learn the new "car" class (olive green) in five different learning conditions until the performance asymptoted: 1) FoL (total n = 5,000 images/epoch), 2) FIL (total n = 45,000 images/epoch), 3) PIL (total n = 2,000 images/epoch), 4) SWIL (total n = 2,000 images/epoch), and 5) EqWIL (total n = 2,000 images/epoch). (*D*) Recall on the new class, recall on similar old classes ("plane," "ship," and "truck"), recall on different old classes (other animal classes), total accuracy, and cross-entropy loss for the held-out test dataset as a function of the number of epochs for FoL (black), FIL (blue), PIL (brown), SWIL (magenta), and EqWIL (gold) conditions. Each plot shows the mean over 10 repetitions; shaded areas are ±1 SEM.

(total n = 2,500 images per epoch compared to n = 50,000 images per epoch). Note that the "cat" class was also interleaved to learn the new "car" class. The SWIL results were compared only to FIL, since that provides the best performance. SWIL achieved the same performance on new and old classes as FIL (Fig. 6; $H = 14.3$, $P > 0.05$). The new "car" class was learned much faster by using SWIL; Speedup = 45x (50,000 × 20/2,500 × 8), while a total of 20x fewer items (MemRed) were presented per epoch than FIL. For both "cat" and "car" class results, we presented a smaller number of images per epoch in

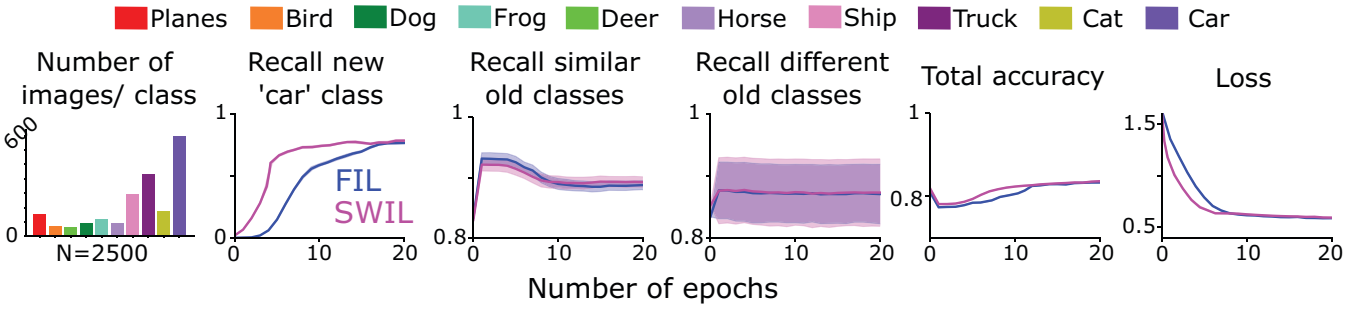

Fig. 6. The six-layer CNN was trained to learn the new "cat" class (task 1) followed by learning the "car" (task 2) class until the performance asymptoted in both conditions: 1) FIL: all old classes (plotted in different colors) + new class ("cat"/"car") image presented with equal probability; and 2) SWIL: old-class examples weighted and presented in proportion to their similarity with a new class ("cat"/"car"). Notice that we included the "cat" class learned in task 1 weighted according to the similarity for learning the "car" class in task 2. Columns represent the distribution of images presented per class per epoch, recall on the new class, recall on similar old classes, recall on different old classes, total accuracy, and cross-entropy loss for the held-out test dataset as a function of the number of epochs for FIL (blue) and SWIL (magenta) conditions for learning "cat" and "car" classes. Each plot shows the mean over 10 repetitions; shaded areas are ±1 SEM.

SWIL (18.75x and 20x) than the entire dataset presented per epoch in FIL, and the network still rapidly learned the new-class items (31.25x and 45x). Extending this idea, one can expect a multiple-fold reduction in learning time and data storage to learn new-class items with an increasing number of learned classes, which might be the case in human brains. The results demonstrate that SWIL can be used to integrate multiple new classes in a sequential learning framework, allowing a neural network to learn continually without interference.

Reduced Learning Time and Data Required with Increasing Distance across Classes Using SWIL

We wanted to test the general applicability of the SWIL algorithm and test whether it can be used for a dataset with many more classes and more complex network architecture. We trained a complex CNN: VGG19 (20) (19 layers) on 90 out of 100 classes of the CIFAR100 dataset (500 training images/class and 100 test images/class). The network was then trained to learn a (new) "train" class. Fig. 7*A* shows the similarity matrix computed from the activations of the penultimate layers on the CIFAR100 dataset. The new "train" class was similar to many existing "vehicle" superclasses (Fig. 7*B*; VGG19, "bus," "streetcar," "tractor," etc.). SWIL allowed learning new items much faster (Speedup = 95.45x [45,500 × 6/1,430 × 2]) and with significantly smaller data (MemRed = 31.8x) than FIL with no difference in the performance ($H = 8.21$, $P > 0.05$). As expected, SWIL overcame the lower recall on the new class using PIL ($H = 10.34$, $P < 0.05$) and the higher level of interference with EqWIL ($H = 24.77$, $P < 0.05$) (Fig. 7*C* and Table 4). Next, we wondered whether a large distance between different class representations underlies the faster speedup observed here. To check this, we trained two more neural network architectures: 1) six-layer CNN (same as Figs. 4 and 5 from CIFAR10); and 2) VGG11 (11 layers) on 90 classes of the CIFAR100 dataset, followed by training on a new "train" class in only two conditions: FIL and SWIL. There was a higher overlap between the new "train" class and the "vehicle" superclass for both new network architectures, but individual

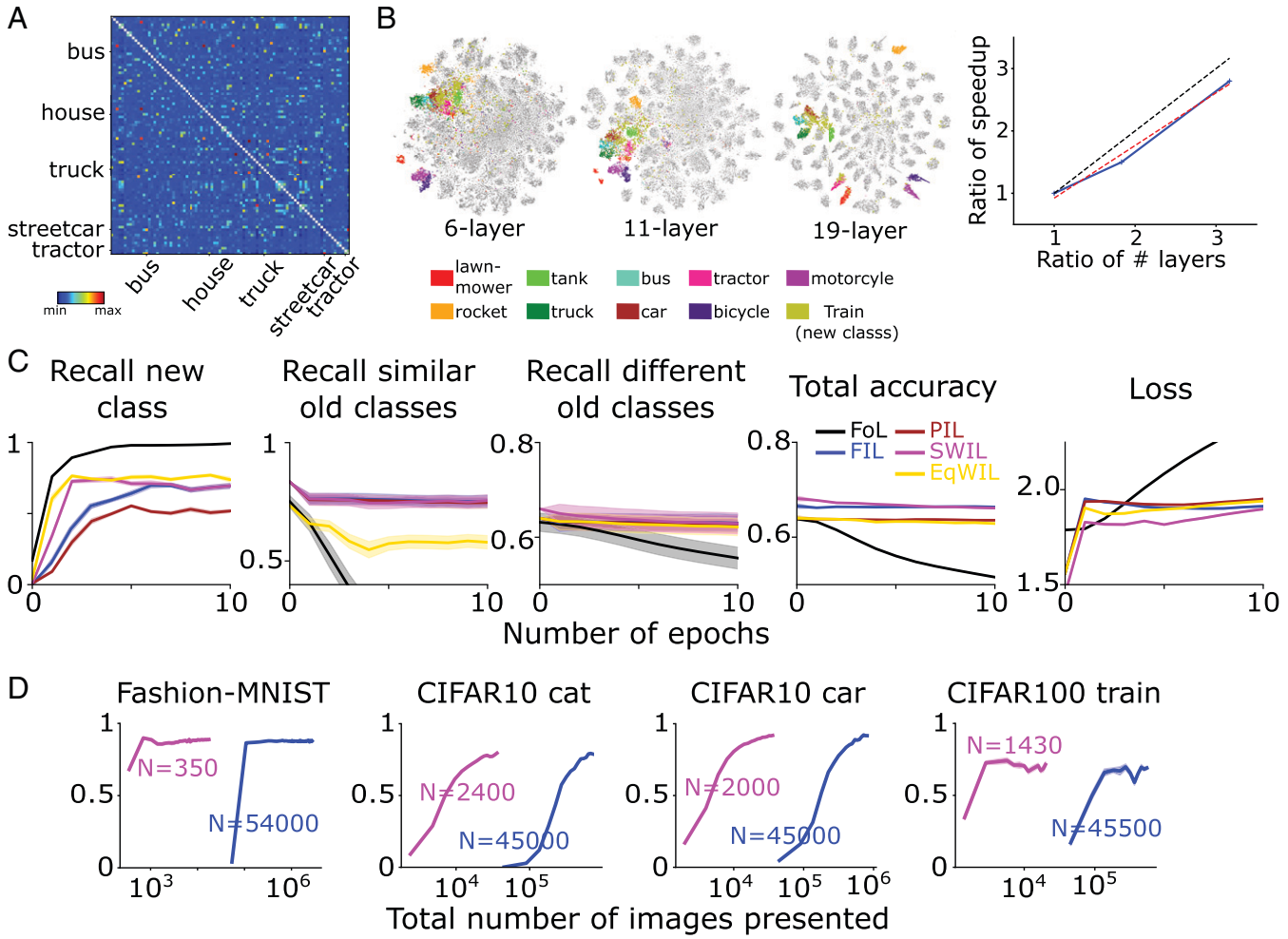

Fig. 7. (*A*) Similarity matrix from the penultimate layer's activation after presenting the new "train" class for VGG19. Five classes—"truck," "streetcar," "bus," "house," and "tractor"—sharing the maximum similarity with the "train" class are indicated. The diagonal elements (similarity = 1) are excluded from the similarity matrix. (*B*, *Left*) *Top* two dimensions of the t-SNE applied to the penultimate-layer activations of the six-layer CNN, VGG11, and VGG19 networks. (*B*, *Right*) Ratio of speedup $\left(\frac{FIL}{SWIL}\right)$ observed (*y* axis) vs. ratio of the number of layers for the three different networks with respect to six-layer CNN. Black dashed line, red dashed line, and solid blue line represent slope = 1 line, best-fit line, and simulation results, respectively. (*C*) Learning dynamics using VGG19 (20): Recall on new "train" class, recall on similar old classes ("vehicles" superclass), recall on different old classes (everything except "vehicles" superclass), total accuracy, and cross-entropy loss for the held-out test dataset as a function of the number of epochs for FoL (black), FIL (blue), PIL (brown), SWIL (magenta), and EqWIL (gold) conditions. Each plot shows the mean over 10 repetitions; shaded areas are ±1 SEM. (*D*) Columns represent Recall for the Fashion-MNIST "boot" class (Fig. 3), CIFAR10 "cat" class (Fig. 4), CIFAR10 "car" class (Fig. 5), and CIFAR100 VGG19 "train" class as a function of the total number of images presented (log-scale) for SWIL (magenta) and FIL (blue). "*N*" represents the total number of images presented per epoch (old + new class) for each learning condition. Note that the *x* axis in *D* starts from epoch #1, showing recall after the network has been trained on one epoch.

Table 3. Performance on test-set at asymptote for the CIFAR10 car dataset

Learning condition	*N*	Recall			Total accuracy	Loss
		New class	Similar old class	Different old class		
FoL	5,000	0.994 ± 0.5e-2	0.516 ± 0.213	0.875 ± 0.025	0.78 ± 0.5e-3	0.816 ± 1.6e-3
FIL	45,000	0.917 ± 3.5e-3	0.885 ± 0.013	0.902 ± 0.024	0.900 ± 0.2e-3	0.389 ± 0.9e-3
PIL	2,000	0.858 ± 6.1e-3	0.887 ± 0.012	0.902 ± 0.025	0.88 ± 0.3e-3	0.461 ± 1.2e-3
SWIL	2,000	0.918 ± 2.5e-3	0.89 ± 0.013	0.902 ± 0.025	0.901 ± 0.4e-3	0.388 ± 1.1e-3
EqWIL	2,000	0.923 ± 3.3e-3	0.821 ± 0.024	0.89 ± 0.025	0.876 ± 0.6e-3	0.487 ± 1.5e-3

Displayed are the means ±1 SEM over 10 repetitions for each condition.

classes were less separated compared to VGG19 simulations (Fig. 7 *B*, *Left*). The Speedup in learning new items with SWIL with respect to FIL scaled roughly linearly with the increase in the number of layers (slope = 0.84; Fig. 7 *B*, *Right*, *SI Appendix*, Fig. 3 and Table 1). This result shows that increased representational distance across classes (penultimate layer) can lead to faster learning (Speedup) and reduced memory load (MemRed).

Next, we wanted to check whether speedup would increase even further if the network is trained on many more nonoverlapping classes, with a larger distance between their representations. To do this, we took a deep linear network (used in the Fashion-MNIST examples in Figs. 1–3) and trained it to learn a combined dataset consisting of 8 Fashion-MNIST classes (excluding "bags" and "boot") and 10 Digit-MNIST classes. The network was then trained to learn a new "boot" class. As expected, the "boot" class was more similar to "sandals" and "sneaker" (similar old classes), followed by the remaining Fashion-MNIST classes, and, finally, Digit-MNIST classes (*SI Appendix*, Fig. 4 *A* and *B*). Based on this, we interleaved more similar old-class exemplars followed by Fashion-MNIST and Digit-MNIST class exemplars for the SWIL (total $n = 350$ images per epoch; *SI Appendix*, Fig. 4*C*). The simulations showed that SWIL allows rapid learning of new-class items without interference, similarly to FIL, but using a much smaller subset of data, MemRed = 325.7x (114,000/350) and Speedup = 162.85x (228,000/1,400) (*SI Appendix*, Fig. 4*D*). The speedup observed in the current result is 2.1x (162.85/77.1), with a 2.25x (18/8) increase in the number of classes compared to Fashion-MNIST results. The results from this section helps to establish that SWIL works even for a more complex dataset (CIFAR100) and neural network architecture (VGG19), proving the general applicability of the algorithm. Additionally, we demonstrated that increased internal distance across classes or increasing the number of nonoverlapping classes could lead to faster learning (Speedup) and reduced memory load (MemRed).

Discussion

Summary. ANNs face a major challenge in continual learning, often exhibiting catastrophic interference (1, 2). To overcome this problem, numerous studies have used an FIL, i.e., joint training the network on new and previously learned items (8, 9). FIL requires interleaving all the existing information every time there is new information, making it a biologically implausible and time-consuming process. Recently, it was shown that FIL might not be required, and interleaving only old items with substantial representational similarity to new items (SWIL) could enable the same performance (11). However, there were concerns raised regarding the scalability of SWIL. We extended the SWIL algorithm and tested it on different datasets—Fashion-MNIST, CIFAR10, and CIFAR100—and neural network architectures—deep linear networks and CNNs. Across all conditions, SWIL and EqWIL perform better in learning new classes compared to PIL. This is expected, as we have increased the relative frequency of the new class compared to old classes. We also demonstrated that carefully selecting and interleaving similar items (SWIL) reduced catastrophic interference with the similar old classes compared to equally subsampling existing classes (EqWIL). SWIL was sufficient to perform similarly to FIL on both new and existing classes, thus providing significant speedup in learning new items (Fig. 7*D*), while substantially reducing the required training data. SWIL allowed learning new classes in a sequential learning framework, further proving its general applicability. Finally, we showed that a new class with lower overlap with previously learned classes (larger distance) could be integrated much more quickly (reduced time) and with even fewer items stored (more data-efficient) than a new class that shares similarities with many old classes. Overall, our results provide a possible insight into how the brain actually may solve one of the main failings of the original CLST model—unrealistic training time.

Comparison with Other Approaches and Potential Issues. Recent brain-inspired approaches to alleviate catastrophic interference can be categorized into 1) regularization-based and 2) generative-replay–based methods. Regularization-based methods, such as Elastic Weight Consolidation (EWC) (21), Learning without Forgetting (22), and Synaptic Intelligence (23), typically involve measuring the importance of each parameter and adding

Table 4. Performance on test-set at asymptote for CIFAR100 dataset using VGG19

Learning condition	*N*	Recall			Total accuracy	Loss
		New class	Similar old class	Different old class		
FoL	500	0.997 ± 1.5e-3	0.076 ± 0.05	0.543 ± 0.023	0.501 ± 0.7e-3	2.48 ± 3.6e-3
FIL	45,500	0.696 ± 0.010	0.755 ± 0.05	0.631 ± 0.018	0.642 ± 0.4e-3	1.926 ± 1.4e-3
PIL	1,430	0.558 ± 0.018	0.751 ± 0.046	0.625 ± 0.018	0.609 ± 0.6e-3	1.982 ± 4.5e-3
SWIL	1,430	0.704 ± 0.016	0.753 ± 0.05	0.628 ± 0.022	0.641 ± 0.5e-3	1.923 ± 1.8e-3
EqWIL	1,430	0.723 ± 0.012	0.583 ± 0.054	0.622 ± 0.018	0.603 ± 0.1e-3	1.973 ± 2.7e-3

Displayed are the means ±1 SEM over 10 repetitions for each condition.

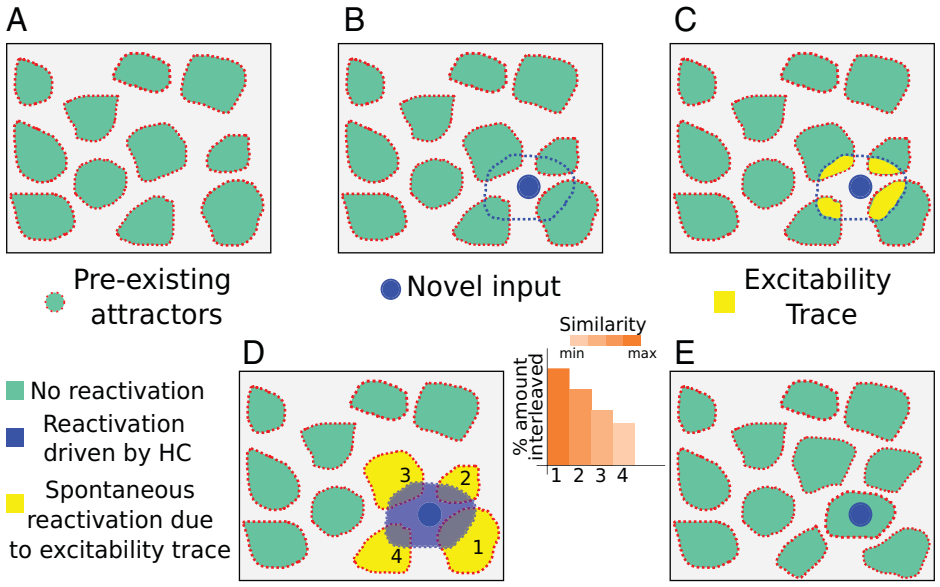

Fig. 8. (*A*) Attractor landscape of already-existing NC schemas with individual cortical attractors shown as a teal-colored basin. (*B*) A novel input (blue dot and contour) experienced during behavior overlaps to various degrees with a few already-existing attractors and leaves an excitability trace (yellow) (*C*) on the overlapping neurons. Thus, previously formed attractors are biased in proportion to their overlap with the new event. (*D*) During sleep, the reactivation of new events is triggered by the HC (blue). HC-driven reactivations are interleaved with spontaneous NC reactivations of the preexisting attractors, which occur in proportion to their overlap with a new event (29). The intensity of orange represents the degree of similarity with the novel input attractor. (*E*) Competitive synaptic plasticity (e.g., BCM or triplet Spike-timing dependent plasticity) results in local synaptic adjustments, creating an attractor for the novel event, with minimal interference with the overlapping preexisting attractors and virtually no change in the attractors representing unrelated events.

a regularization term that penalizes changes in the most relevant parameters or mapping function of the network. These approaches usually suffer when there is a need to learn many new classes incrementally (24, 25). The replay-based approaches (25, 26), motivated by the HC-replay literature, consist of deep generative and task-solver networks. During relearning, new-class exemplars are interleaved with generated pseudodata (captures representational statistics of previously learned information). The generative-replay approaches overcome the first issue of CLST, i.e., not having access to the old data. However, the problem is shifted toward implementing an improved generator (an important and hard problem), limiting the performance to the generator's effectiveness. These approaches typically interleave the existing data equally while learning new items, similar to EqWIL, which might not be required and might even be detrimental, given the current study results. SWIL addresses one of the issues with the CLST, i.e., not enough time to interleave all the existing data. After determining the similarity in SWIL using each class's average activation, we sampled the old classes from the training data. This still requires the storage of a large amount of data in memory. This problem might be resolved by combining SWIL with generative replay (10, 25, 26) and testing performance after interleaving generated old items with new items in proportion to similarity. After combining SWIL with generative replay, we should have to store only average activation maps for each class, thus significantly reducing the memory footprint. Indeed, using the pseudorehearsal approach, it might not be necessary to store any old data (10) (*Biological Implementation of SWIL*). We understand that SWIL is not a perfect solution to the lifelong-learning problem, but, rather, is complementary to most past approaches. Future studies should combine SWIL with generative replay or EWC. These combinations might give a better overall solution to a range of task settings, overcoming the shortcomings of each approach.

There are a couple of points to be noted before using SWIL for a learning scenario. The number of new-class images (for example, 720 images for the "cat" class) used for interleaving in SWIL were about ∼1.25 times the images used for the most similar existing class (575 images for the "dog" class). The number of new-class images to interleave can be treated as a hyperparameter to tune in future work. In the current study, we used 1.25 to 1.5 times the most similar old-class images as the value of the number of new-class-images hyperparameter, which worked well across all the presented conditions. Typical transfer-learning approaches freeze initial layers of the networks during new-class learning. We wanted to study the effect of freezing earlier layers in our simulations. To do this, we ran CIFAR10-cat, CIFAR10-car, and CIFAR100 with earlier layers frozen and saw that SWIL still performs comparably to FIL. But, the speedup observed with SWIL was reduced for each dataset (*SI Appendix*, Figs. 5–7), perhaps because of the reduced dimensionality of the frozen network relative to the unfrozen network. We also performed stress testing on the SWIL algorithm for the CIFAR10 dataset (for both the "cat" and "car" classes) by varying the total number of images presented from $n = 15{,}000$ to $n = 500$ (*SI Appendix*, Fig. 8) and found that new-class recall and overall accuracy reach the same value as FIL with similar performance on existing classes. SWIL still did better in learning new classes than PIL and showed less catastrophic interference than EqWIL at all values of total training data size.

We calculated the similarity between a target-layer average activation for a new-class item with the previously learned classes. Since we focused on the learning dynamics at a class level, rather than individual attributes across classes, it would be interesting to look at the learning dynamics of individual feature maps. Similarity can also be computed at different levels: pixel-wide or functional; our approach of looking at the activation-map similarity resembles the ventral visual-stream similarity (27). We computed the similarity typically in the last layers because forgetting first happens at the top layers to output layers. So, the initial layers' features might remain preserved, whereas mapping might change with training. It is possible that the brain may implement a different similarity-calculation

function, and determining the most optimal function for computing similarity is out of the scope of current work. We want to emphasize how the brain and an ANN may learn new information by exploiting the hierarchical distribution of existing knowledge. Learning a new item might be faster with a network pretrained on a dataset with many classes, since there is a high probability of new information being consistent across many more dimensions with the existing classes and features. This is similar to how adults generally learn new information faster than a child, perhaps because they have an extensive repertoire of features developed over their lifetime (28).

Biological Implementation of SWIL. How might SWIL be implemented in the brain? Let's assume that the NC has multiple preexisting attractors for different features (Fig. 8*A*) organized in an energy landscape. The animal is presented with a novel input that overlaps to varying degrees with some, but not all, of the existing attractors (Fig. 8*B*). This overlap is itself a measure of representational similarity. During the novel experience, NC cells from overlapping existing attractors might have increased activation. We assume that there is some mechanism of tagging these cells, such that they would have a persistently increased excitability (30, 31) for some period (Fig. 8*C*). In the posttask sleep, sharp-wave ripple events in the HC (32, 33) will trigger the replay of novel input, and the NC will show bias by spontaneously reactivating and interleaving only those existing attractors in which some cells express the excitability tag. In such an attractor scenario, the spontaneous reactivation of the existing attractors would be proportional (on average) to the overlap with the novel stimuli (i.e., due to the proportion of neurons in the attractor having the excitability tag), making the reactivation probability proportional to similarity, as shown previously in simulations of hippocampal attractor dynamics (33) (Fig. 8*D*). After multiple iterations of interleaved reactivation and local synaptic adjustments, the novel input attractor could be gradually integrated into the energy landscape with minimal disruption of existing attractors (Fig. 8*E*). The learning-induced changes in gK_{Ca} observed in the HC and NC (30, 31, 34) might provide the hypothetical similarity bias in the reactivation probability of existing attractors via increased excitability. Another possibility is that, once the NC has computed similarity between the novel input and existing knowledge, different learning mechanisms might be triggered based on uncertainty levels: mismatch vs. poor similarity vs. surprise, etc., as described in the Adaptive Resonance Theory (35). These learning mechanisms can be mediated by neuromodulators such as acetylcholine, released in response to different uncertainty levels, and triggering plasticity in multiple brain regions (36). For example, lower levels of acetylcholine during sleep might allow a larger spread of activity both from the HC to the NC and within the NC, by releasing cholinergic suppression on excitatory feedback synapses. Future work should explore the in vivo manipulation of excitability during behavior and sleep to understand replay dynamics and memory consolidation.

Conclusions. Overcoming catastrophic interference is of utmost importance to achieve lifelong learning in neural networks. Even though current lifelong-learning models are still far from capturing the wide extent of dynamics exhibited by the mammalian brain, they provide a framework for designing future studies targeting specific questions. SWIL provides insight into this problem, by showing that similarity-weighted replay of old memories can allow learning new items much faster. We hope that the ideas presented here lead to improved understanding of novel-class learning and memory consolidation.

Materials and Methods

Similarity Matrix Calculation. We computed similarity at the feature level by computing the cosine similarity between the activations of existing- and new-class items using the methods described below (16, 17):

1. For a trained neural network, calculate the target-layer (typically penultimate layer) activations for all existing classes and new-class items in the first epoch without updating weights or backpropagation. The output activations A' will be $A' = \{(A_1,y_1), \ldots (A_n,y_n)\}$ where A_i is the class activation, and y_i is the corresponding labels.
2. Perform linear discriminant analysis (LDA) on the output activations A' to find a basis set that maximizes the interclass variance. Project the activations A' into the transformed space AL' calculated by using LDA.
3. Compute the average per-class activation vector (v_i) from the projected samples AL'.
4. Calculate the similarity distribution matrix SM ($n \times n$ size for "n" input classes). The similarity between class i and j, $SM_{i,j}$, is calculated in the following way:

$$SM_{i,j} = \begin{cases} \dfrac{s(v_i, v_j)}{\sum_j^n s(v_i, v_j)} & i \neq j \\ 0 & i = j \end{cases},$$

where,

$$s(v_i, v_j) = \frac{1}{1 + e^{d\langle v_i, v_j \rangle}} \text{ if } i \neq j$$

$$d(v_i, v_j) = 1 - \frac{\langle v_i | v_j \rangle}{||v_i||\ ||v_j||}.$$

$\langle . | . \rangle$ represents the vector dot product, and $s(.,.)$ defines a similarity measure between the average class-activation vectors.

Hyperparameter: Learning Rate and Number of Steps per Epoch. At the start of training, the learning rate was set to 0.001 (Fashion-MNIST, CIFAR10, and CIFAR100) or 0.005 (CIFAR100) with exponential decay of 0.0001 for each learning condition. The learning rate was not optimized for the five different learning conditions to better compare learning dynamics across these conditions. Each learning epoch for different conditions consisted of the same number of steps per epoch, i.e., the number of batch iterations before a training epoch is considered complete. The same number of steps per epoch allows us to directly compare the number of items presented per epoch across learning conditions.

Data Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We thank Srishti Tomar, Zaneta Navratilova, and Scott Kilianski for technical assistance and advice; and James L. McClelland, Terrence J. Sejnowski, Francesco P. Battaglia, Giri P. Krishnan, and Artur Luczak for their comments on the manuscript. This work was supported by Defense Advanced Research Projects Agency Grant HR0011-18-2-0021 and NIH Grant R01 NS121764 (to B.L.M.).

1. M. McCloskey, N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem" in *Psychology of Learning and Motivation*, G. H. Bower, Ed. (Academic Press, New York, 1989), vol. **24**, pp. 109–165.
2. R. Ratcliff, Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychol. Rev.* **97**, 285–308 (1990).
3. J. L. McClelland, B. L. McNaughton, 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. *Psychol. Rev.* **102**, 419–457 (1995).
4. D. Marr, Simple memory: A theory for archicortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **262**, 23–81 (1971).
5. B. L. McNaughton, Cortical hierarchies, sleep, and the extraction of knowledge from memory. *Artif. Intell.* **174**, 205–214 (2010).
6. C. D. Schwindel, B. L. McNaughton, "Hippocampal–cortical interactions and the dynamics of memory trace reactivation" in *Slow Brain Oscillations of Sleep, Resting State and Vigilance*, E. J. W. Van Someren, Y. D. Van Der Werf, P. R. Roelfsema, H. D. Mansvelder, F. H. Lopes Da Silva, Eds. (Progress in Brain Research, Elsevier, Amsterdam, 2011), vol. **193**, pp. 163–177.
7. T. J. Teyler, P. DiScenna, The hippocampal memory indexing theory. *Behav. Neurosci.* **100**, 147–154 (1986).
8. A. Gepperth, C. Karaoguz, A bio-inspired incremental learning architecture for applied perceptual problems. *Cognit. Comput.* **8**, 924–934 (2016).

9. R. Kemker, C. Kanan, FearNet: Brain-inspired model for incremental learning. arXiv [Preprint] (2018). https://arxiv.org/abs/1711.10563. Accessed 23 February 2018.
10. A. Robins, Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.* **7**, 123–146 (1995).
11. J. L. McClelland, B. L. McNaughton, A. K. Lampinen, Integration of new information in memory: New insights from a complementary learning systems perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190637 (2020).
12. D. Tse *et al.*, Schemas and memory consolidation. *Science* **316**, 76–82 (2007).
13. D. Tse *et al.*, Schema-dependent gene activation and memory encoding in neocortex. *Science* **333**, 891–895 (2011).
14. J. L. McClelland, Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *J. Exp. Psychol. Gen.* **142**, 1190–1210 (2013).
15. H. Xiao, K. Rasul, R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv [Preprint] (2017). https://arxiv.org/abs/1708.07747. Accessed 15 September 2017.
16. P. S. Negi, D. Chan, M. Mahoor, Leveraging class similarity to improve deep neural network robustness. arXiv [Preprint] (2018). https://arxiv.org/abs/1812.09744. Accessed 27 December 2018.
17. K. Park, D.-H. Kim, Accelerating image classification using feature map similarity in convolutional neural networks. *Appl. Sci. (Basel)* **9**, 108 (2019).
18. A. Krizhevsky, Learning multiple layers of features from tiny images. Technical Report TR-2009 (University of Toronto, Toronto, ON, CA, 2009).
19. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
20. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv [Preprint] (2015). https://arxiv.org/abs/1409.1556. Accessed 10 April 2015.
21. J. Kirkpatrick *et al.*, Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3521–3526 (2017).
22. Z. Li, D. Hoiem, Learning without forgetting. *Neuron*. **73**, 415–434 (2012).
23. F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence. arXiv [Preprint] (2017). https://arxiv.org/abs/1703.04200. Accessed 12 June 2017.
24. R. Kemker, M. McClure, A. Abitino, T. Hayes, C. Kanan, Measuring catastrophic forgetting in neural networks. arXiv [Preprint] (2017). https://arxiv.org/abs/1708.02072. Accessed 9 November 2017.
25. G. M. van de Ven, H. T. Siegelmann, A. S. Tolias, Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* **11**, 4069 (2020).
26. H. Shin, J. K. Lee, J. Kim, J. Kim, Continual learning with deep generative replay. arXiv [Preprint] (2017). https://arxiv.org/abs/1705.08690. Accessed 12 December 2017.
27. J. J. DiCarlo, D. Zoccolan, N. C. Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
28. A. M. Saxe, J. L. McClelland, S. Ganguli, A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11537–11546 (2019).
29. B. Shen, B. L. McNaughton, Modeling the spontaneous reactivation of experience-specific hippocampal cell assembles during sleep. *Hippocampus* **6**, 685–692 (1996).
30. M. C. de Jonge, J. Black, R. A. Deyo, J. F. Disterhoft, Learning-induced afterhyperpolarization reductions in hippocampus are specific for cell type and potassium conductance. *Exp. Brain Res.* **80**, 456–462 (1990).
31. J. R. Moyer Jr., L. T. Thompson, J. F. Disterhoft, Trace eyeblink conditioning increases CA1 excitability in a transient and learning-specific manner. *J. Neurosci.* **16**, 5536–5546 (1996).
32. G. Buzsáki, Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus* **25**, 1073–1188 (2015).
33. M. A. Wilson, B. L. McNaughton, Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).
34. C. D. Woody, E. Gruen, D. Birt, Changes in membrane currents during Pavlovian conditioning of single cortical neurons. *Brain Res.* **539**, 76–84 (1991).
35. S. Grossberg, How does a brain build a cognitive code? *Psychol. Rev.* **87**, 1–51 (1980).
36. M. E. Hasselmo, B. P. Wyble, Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behav. Brain Res.* **89**, 1–34 (1997).