Tuned Compositional Feature Replays for Efficient Stream Learning

Morgan B. Talbot, Rushikesh Zawar, Rohil Badkundi, Mengmi Zhang, and Gabriel Kreiman

Abstract—Our brains extract durable, generalizable knowledge from transient experiences of the world. Artificial neural networks learn something similar, but achieve this ability when trained on nonrepeating video frames in temporal order (online stream learning), models that learn well from shuffled datasets catastrophically forget old knowledge upon learning new stimuli. We propose a new continual learning algorithm, compositional replay using memory blocks (CRUMB), which mitigates forgetting by replaying feature maps reconstructed by combining generic parts. CRUMB concatenates trainable and reusable “memory block” vectors to compositionally reconstruct feature map tensors in convolutional neural networks (CNNs). Storing the indices of memory blocks used to reconstruct new stimuli enables memories of the stimuli to be replayed during later tasks. This reconstruction mechanism also primes the neural network to minimize catastrophic forgetting by biasing it toward attending to information about object shapes more than information about image textures and stabilizes the network during stream learning by providing a shared feature-level basis for all training examples. These properties allow CRUMB to outperform an otherwise identical algorithm that stores and replays raw images while occupying only 3.6% as much memory. We stress-tested CRUMB alongside 13 competing methods on seven challenging datasets. To address the limited number of existing online stream learning datasets, we introduce two new benchmarks by adapting existing datasets for stream learning. With only 3.7%–4.1% as much memory and 15%–43% as much runtime, CRUMB mitigates catastrophic forgetting more effectively than the state-of-the-art. Our code is available at https://github.com/MorganBDT/crumb.git

Index Terms—Brain-inspired replay, catastrophic forgetting, deep learning, stream learning.

I. INTRODUCTION

HUMANS adapt to new and changing environments by learning rapidly and continuously. Previously learned skills and experiences are retained even as they are transferred and applied to new tasks, which are learned from a stream of highly temporally correlated stimuli and without direct access to past experiences. In contrast, in standard class-incremental image classification settings in continual learning, neural networks are presented with images that are independently and identically distributed (i.i.d.), with multiple presentations of each image [1], [2], [3]. To better emulate a human learning environment or that of an autonomous robot that must learn in real time, we focus on a challenging and realistic variant of class-incremental learning—online stream learning. Online stream learning has two key characteristics (Fig. 1): 1) the input is in the form of video streams with highly temporally correlated frames and 2) each training example is presented only once: no repeated presentations of old data are allowed.

In online stream learning settings, current machine learning systems tend to fail to retain good performance on previously learned tasks, exhibiting catastrophic forgetting [4], [5], [6]. Catastrophic forgetting is a pervasive problem in continual learning settings for both deep neural networks [7] and other models such as linear regression [8] and self-organizing maps [9], and can also cause neural models to be biased toward more recently encountered training data [10]. One strategy for overcoming catastrophic forgetting is to store a copy of all or most encountered training examples for later replay, effectively converting to an offline learning paradigm [11]. This approach, however, often requires an impractically large amount of memory [12]. Moreover, much of the information in raw images is redundant, with many pixel values needed to represent each feature-level concept relevant to classification. Finally, storing old training data might also be undesirable from a data security or privacy standpoint, such as in hospitals and other healthcare settings [13].

Manuscript received 6 March 2023; revised 25 October 2023 and 24 November 2023; accepted 3 December 2023. This work was supported in part by NIH under Grant R01EY026025; in part by the National Research Foundation, Singapore, through its AI Singapore Programme (AISPG) under Award AISG2-RP-2021-025 and its Singapore National Research Foundation Fellowship (NRFF) under Award NRF-NRFIF-2023-0001; in part by the Center for Brains, Minds and Machines through the NSF Science and Technology Center Award under Grant CCF-1231216; and in part by the National Institute of General Medical Sciences under Award T32GM007753 and Award T32GM144273. (Morgan B. Talbot and Rushikesh Zawar contributed equally to this work.) (Corresponding authors: Mengmi Zhang; Gabriel Kreiman.) Morgan B. Talbot is with the Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139 USA, also with Boston Children’s Hospital (BCH), Boston, MA 02115, USA, and also with the Center for Brains, Minds, and Machines (CBMM), Cambridge, MA 02142 USA. Rushikesh Zawar is with Harvard University, Cambridge, MA 02138 USA, also with Boston Children’s Hospital (BCH), Boston, MA 02115 USA, and also with the Birla Institute of Technology and Science at Pilani (BITS Pilani), Pilani 333031, India. Rohil Badkundi is with Harvard University, Cambridge, MA 02138 USA, also with Boston Children’s Hospital (BCH), Boston, MA 02115 USA, and also with the Center for Brains, Minds, and Machines (CBMM), Cambridge, MA 02142 USA. Mengmi Zhang is with the Center for Brains, Minds, and Machines (CBMM), Cambridge, MA 02142 USA, also with Boston Children’s Hospital (BCH), Boston, MA 02115 USA, also with the Agency for Science, Technology and Research (A*STAR) Center for Frontier AI Research (CFAR), Singapore 138632, also with the A*STAR Institute for Infocomm Research (I2R), Singapore 138632, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: mengmi.zhang@ntu.edu.sg). Gabriel Kreiman is with Boston Children’s Hospital (BCH), Boston, MA 02115, USA, also with the Harvard Medical School, Boston, MA 02115 USA, and also with the Center for Brains, Minds, and Machines (CBMM), Cambridge, MA 02142 USA (e-mail: gabriel.kreiman@tch.harvard.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2023.3344085, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2023.3344085
To address both memory inefficiency and data privacy concerns while achieving state-of-the-art online stream learning performance, we propose a new continual learning approach, compositional replay using memory blocks (CRUMB) (Fig. 2). In our method, each new image is processed by the early layers of a convolutional neural network (CNN) to produce a feature map tensor. The feature map is decomposed by slicing it into chunks, each of which is a vector of feature activations at a specific spatial location. Each chunk is then replaced by the most cosine-similar row (“memory block”) of a trainable “codebook matrix.” This mechanism encodes images as a composition of discrete feature-level concepts, some of which appear to have semantic interpretations. Storage of a complete training example for replay requires keeping only the indices of the memory blocks needed to reconstruct the original feature map, along with the class label, occupying only 3.6% of the memory footprint of a raw image. During replay, feature maps reconstructed via stored indices are fed to the later layers of the CNN such that these layers are trained on both stored and newly encountered images to learn new tasks while retaining previous knowledge.

Our key contributions are given as follows.

1) **Trainable Compositional Replay**: We propose a new compositional feature-level replay algorithm, CRUMB, for online stream learning. The composition mechanism is end-to-end trainable and reusable. CRUMB’s codebook of memory blocks captures the essential components needed for reconstructing feature maps. During the pretraining phase, the memory block mechanism primes the CNN for stream learning with high accuracy and induces a beneficial bias toward object shapes. Using memory blocks as a shared basis for new and recalled examples helps stabilize the network during stream learning.

2) **Reduced Forgetting**: We tested CRUMB on seven continual learning datasets alongside 13 competing methods, showing that CRUMB typically outperforms state-of-the-art approaches by large margins.

3) **Superior Efficiency**: Storing $n$ compositional feature maps for replay prevents catastrophic forgetting substantially more effectively than storing $n$ raw images while only requiring about 3.6% as much memory. In addition, compared with the next most accurate method (REMIND [4]), CRUMB requires only about 15%–43% as much training runtime and occupies only 3.7%–4.1% of REMIND’s peak memory footprint.

4) **New Benchmarks**: We adapted two datasets, Toybox [14] and iLab [15], to introduce new online stream learning benchmarks. All benchmark details along with source code, results, and data are available at https://github.com/MorganBDT/crumb.git.

II. RELATED WORK

A. **Weight Regularization**

Weight regularization methods typically store weights trained on previous tasks and impose constraints on subsequent weight updates to minimize catastrophic forgetting [12], [16], [17], [18], [19], [20], [21]. However, storing the importance of the millions of parameters required by state-of-the-art recognition models across all previous tasks is costly [12], [22]. Moreover, empirical comparisons suggest that weight regularization methods typically do not mitigate catastrophic forgetting as effectively as architecture adaptation and replay methods [23].

B. **Architecture Adaptation**

Architecture adaptation methods expand or reorganize the structure of their neural networks to accommodate new tasks to be learned. Approaches include adding groups of new neurons (which does not always scale well) [12], [16], [18], [19], [20], [24], isolating parts of a larger neural network for each task [25], [26], [27], [28], [29], compressing parameters in a consolidation phase [30], and pruning neurons or weights for later reuse. Pruning approaches include $L1$ regularization and activity threshold-based sparsification of neurons [31], and combining pruning of weights with parameter importance-based regularization [32]. Neuron pruning/reuse can also be combined with the addition of new neurons to improve performance and enable increased flexibility [33]. All of these approaches add significant complexity, and some require explicit labeling of task identities, which is not feasible in many online learning applications.

C. **Image and Feature Replay**

In replay methods, images or features from previous tasks are stored and later retrieved or regenerated to be shown to the model to prevent forgetting [17], [34], [35], [36], [37], [38], [39]. Replay can be highly effective but comes with some caveats. Relying on replaying limited sets of stored examples can lead to overfitting. Storing a large number of raw images for replay is memory-intensive. To limit memory requirements, generative replay systems combine data from new tasks with...
synthetic data produced by generative models to resemble previously encountered stimuli [40], [41], [42], [43], [44], [45], [46]. However, the generative models needed to create adequate synthetic data remain large, memory-intensive, and difficult to train [22].

Other replay methods save memory by storing raw or compressed feature maps from intermediate CNN layers [4], [47] or generate synthetic examples by sampling from simple feature-level probability distributions for each class [48]. REMIND [4] achieves high performance in online stream learning by compressing feature maps using a product quantizer [49]. However, the product quantizer is trained by performing k-means clustering on a large subset of training data stored in memory, a process that scales poorly for increasingly large datasets. In contrast, CRUMB’s differentiable codebook is trained by backpropagation alongside other network parameters, dramatically reducing memory requirements for codebook initialization.

CRUMB’s feature-based replay mechanism is inspired by biological replay observed in the hippocampus and other brain areas [50], [51], [52] and by complementary learning systems theory [53]. Recent work has explored modeling the hippocampus and neocortex as separate neural networks that interact via distillation losses and other mechanisms [54]. In contrast to storing knowledge implicitly in a short-term memory network, CRUMB’s memory blocks and replay buffer store short-term memories that represent individual training examples and interact with the CNN via replay to facilitate long-term memory storage and consolidation.

III. METHODS

A. Online Stream Learning Benchmarks

1) Training Protocols: We consider two online class-incremental settings: class instance and class-i.i.d. [4] (Fig. 1).

a) Class instance: Each task contains short video clips of different objects from two or more classes, and the video clips are presented one after another in random order within each task without repetition. An ideal learning algorithm in this setting would be stable enough to remember prior tasks while being sufficiently plastic to learn generalizable class boundaries for new classification tasks, despite encountering many highly correlated images of each object before moving on to the next.

b) Class-i.i.d.: Images/video frames are randomly shuffled within each task but not interspersed among tasks and are shown only once in class instance. This is a less challenging protocol and should not be considered stream learning in the strictest sense because the shuffling of images in each task destroys any temporal structure among images.

In both settings, our model and all competing baseline models are allowed to train for many epochs on the first task but are restricted to viewing each image from subsequent tasks only once. This emulates real-time acquisition of training data that cannot be stored except in a limited-capacity replay buffer.

2) Stream Learning Benchmark Datasets: We evaluated our model on five video datasets (class-instance and class-i.i.d. protocols) and two image datasets (class-i.i.d. only). For all datasets, we used different task and example orderings across training runs. A global holdout test set of images/frame sequences was used for all runs. To help address the limited number of online stream learning benchmarks, we adapted two datasets designed for studying object transformations, Toybox [14] and iLab [15], for online stream learning.

The CORE50 video dataset [55] contains images of 50 objects in ten classes. Each object has 11 instances, which are 15-s video clips of the object under particular conditions and poses. We followed [4] for the training and testing data split and sampled each video at 1 frame per second (FPS).

The Toybox video dataset [14] contains videos of toy objects from 12 classes. We used a subset of the dataset containing 348 toy objects, each of which has ten instances containing different patterns of object motion. We sampled
each instance at 1 FPS, resulting in 15 images per instance per object. We chose three of the ten instances for our test set, leaving seven instances for training.

The iLab (iLab-2M-Light) video dataset [15] contains videos of toy vehicles from 14 classes. We used a subset of the dataset containing 392 vehicles, with eight instances (backgrounds) per object and 15 images per instance. We chose two of the eight instances for our test set.

The iCub (iCubWorld Transformations) video dataset [56] contains videos taken by the iCub robot of 20 classes of household objects undergoing viewpoint transformations. We used isolated rotation, scaling, and background transformations as our training set and the provided “MIX” sequence (which combines all transformations) as our test set.

The iLab + CORE50 video dataset combines iLab and CORE50 to create a stream learning benchmark with 24 distinct classes. All iLab classes are learned before CORE50, introducing a mild domain shift. We uniformly subsample iLab to balance the number of images per class with CORE50.

To evaluate our model in long-range online class-incremental learning with many more classes than the video datasets described above, we also include results on two image datasets. The standard Online-CIFAR100 image dataset [57] is split into 20 tasks with five classes each, while the standard Online-ImageNet image dataset [58] is split into ten tasks with 100 classes each. Class-instance training is not applicable to image datasets because they do not consist of videos.

3) Baseline Algorithms for Comparison: All baseline algorithms use the same training protocols as CRUMB. CRUMB and most baselines use a SqueezeNet CNN pretrained on ImageNet [59], but due to implementation constraints, adaptive aggregation network (AAN) [60], CoPE [61], GSS [36], LwF [16], RM [39], and Stable SGD [62] use non-pretrained ResNet models [63]. We reimplemented some methods due to varying code availability. CRUMB and all baselines are implemented using the PyTorch library [64].

a) Weight regularization: We compared against elastic weight consolidation (EWC) [12], synaptic intelligence (SI) [19], memory aware synapses (MAS) [65], learning without forgetting (LwF) [16], and Stable SGD [62].

b) Memory distillation and replay: We compared against gradient episodic memory (GEM) [38], incremental classifier and representation learner (iCARL) [35], bias correction (BiC) [34], gradient sample selection (GSS) [36], continual prototype evolution (CoPE) [61], AAN [60], REMIND [4], and rainbow memory (RM) [39].

The lower bound is trained sequentially over all tasks without any measures to avoid catastrophic forgetting.

The upper bound is trained off-line on shuffled images from both the current and all previous tasks over multiple epochs.

Chance predicts class labels by randomly choosing one out of the total of $C_t$ classes seen in or before current task $t$.

B. Proposed Algorithm: CRUMB

We propose a new continual learning algorithm, CRUMBs. CRUMB consists of a 2-D CNN augmented by an $n \times d$ codebook matrix $B$. A schematic of CRUMB is shown in Fig. 2, with further details described in algorithm 1. CRUMB extracts a feature map from each given image using the early layers of a pretrained CNN and stores a subset of the feature maps in a buffer. When CRUMB later encounters a new task, it avoids catastrophic forgetting of previous tasks by replaying stored feature maps of images from those tasks to the later layers of the network. To reduce memory requirements, CRUMB uses its codebook matrix $B$ to reconstruct each feature map. Rows of $B$ (“memory blocks”) are initially pretrained on ImageNet using standard methods but also leads to image datasets because they do not consist of videos.

```
Algorithm 1 CRUMB at Task $t$

Input: training images $I$, from new classes, stored codebook matrix $B$, replay buffer $X$ of stored memory block indices and their class labels (maximum number $n_X$ of total stored examples in $X$ varies by dataset).

Training:
for batch in $I$, do
  Reconstruct feature map $Z$ as $\tilde{Z}$ for each image in batch by concatenating memory blocks (rows of $B$)
  Train $P(\cdot)$ using loss $L(P(\tilde{Z}), P(\tilde{Z}), y_c)$, with $\alpha = 0$ in streaming
  Train memory blocks in $B$ that form part of $\tilde{Z}$, using $L_{CE}(P(\tilde{Z}), y_c)$
if $t > 1$ then
  Randomly sample images $x$ out of $X$ to form a replay batch
  Reconstruct $\tilde{Z}$ for each $x$ by concatenating memory blocks
  Train $P(\cdot)$ using loss $L_{CE}(P(\tilde{Z}), y_c)$
  Train memory blocks in $B$ that are part of $\tilde{Z}$ via backpropagation
end if
end for

Testing:
for batch in testing images do
  Compute predictions $p = P(F(\cdot))$ on test images using $Z$ only
end for
```

Authorized licensed use limited to: Harvard Library. Downloaded on December 27,2023 at 02:15:01 UTC from IEEE Xplore. Restrictions apply.
undergo additional ImageNet pretraining with an objective incorporating predictions on both $Z$ and $\tilde{Z}$. Prior to stream learning, only the final layer of $P(\cdot)$ is randomly reinitialized to reflect the number of classes to be learned during streaming.

2) Reconstructing Feature Maps From Memory: CRUMB produces reconstructed feature map $\tilde{Z}$ using only $Z$ and the contents of its $n \times d$ codebook matrix $B$, where each of the $n$ rows $B_k$ is a “memory block” vector. Hyperparameters $n$ and $d$ are determined empirically (see Sections IV-D4 and IV-D5). $Z$ is first partitioned evenly along its feature dimension into $s/d$ tensors, with each tensor $Z_{f,y}$ of size $d \times w \times h$. Each tensor $Z_f$ is further partitioned by spatial location into $w \cdot h$ vectors, denoted $Z_{f,x,y} \in \mathbb{R}^{d}$, where $d$ is also the dimension of each row $B_k$ in the matrix $B$. For each vector $Z_{f,x,y}$ in $Z$, a similarity score $\gamma_k$ is calculated between it and each memory block $B_k$ as follows:

$$\gamma_{f,x,y,k} = \left( \frac{Z_{f,x,y} \cdot B_k}{\|B_k\|_2} \right)$$

where $\langle \cdot , \cdot \rangle$ is the dot product and $\| \cdot \|_2$ is the $L2$-norm. Because $B_k$ is normalized, $\gamma_{f,x,y,k}$ is highest for the memory block most similar in vector direction to the given $Z_{f,x,y}$. The memory block $B_k$ with the highest $\gamma$ similarity value replaces $Z_{f,x,y}$ at its corresponding location in $\tilde{Z}$ as follows:

$$\tilde{Z}_{f,x,y} \leftarrow B_{k_{f,x,y}} \text{, where } k_{f,x,y} = \text{argmax}(\gamma_{f,x,y,k})$$

Because $\tilde{Z}$ is reconstructed entirely from memory blocks $B_k$, we can save all information needed to reconstruct $\tilde{Z}$ again later by storing both $B$ and the values of $k$ at each $f$, $x$, and $y$ location in $Z$. Thus, the feature map for the $t$th training image can be stored as $m_t = (k_{1,1,1}, \ldots, k_{f,x,y}, \ldots, k_{j/d,w,h})$.

For example, in our main implementation, $Z$ is a $512 \times 13 \times 13$ tensor. $d = 16$ so that $Z$ is split into $32 \times 13 \times 13 = 5408$ vectors of length $16$, which are each replaced in $\tilde{Z}$ by a $16$-D memory block from a $256 \times 16$ matrix $B$. The memory blocks themselves occupy a near-negligible amount of memory: $256 \times 16 = 4096$ floating-point values, compared to the 5408 integers required to store a single compressed training example in this implementation.

3) Training: During training, both $Z$ and $\tilde{Z}$ are passed separately through the classifier $P(\cdot)$ to obtain two classification probability vectors $p = P(Z)$ and $\tilde{p} = P(\tilde{Z})$, where the dimension of $p_i$ and $\tilde{p}_i$ is equal to the total number of classes $C_t$ that have been seen in or before the current task $t$. The loss function $L$ used for training is a weighted sum of the cross-entropy losses $L_{CE}$ derived from $p$ and $\tilde{p}$. With $y_c$ defined as the ground-truth class label of a given image

$$L(p, \tilde{p}, y_c) = \alpha L_{CE}(p, y_c) + \beta L_{CE}(\tilde{p}, y_c)$$

Larger values of $\alpha$ penalize “direct” prediction errors from $P(Z)$, while larger values of $\beta$ penalize “codebook-out” prediction errors from $P(\tilde{Z})$. Although our model generates class predictions based on both $Z$ and $\tilde{Z}$, we use the empirically more accurate predictions from $Z$ during inference on the test set. Empirically, the best performance was achieved by including both direct and codebook-out predictions in the loss function for pretraining ($\alpha = 1$ and $\beta = 1$) and then removing the direct loss for stream learning ($\alpha = 0$ and $\beta = 1$) (see Section IV-D6) for analysis. Setting $\alpha$ to 0 makes the loss function for new batches of images more similar to that used for replay, where only $\tilde{Z}$ is available. Replacement of $Z$ by the reconstructed version $\tilde{Z}$ can be viewed as both a means to efficiently mitigate catastrophic forgetting and a regularization technique to prevent overfitting and stabilize the CNN.

Although values in CRUMB’s memory blocks play the role of activation values in their reconstruction of $\tilde{Z}$, they are trainable parameters of the model. Backpropagation from $\tilde{Z}$-based predictions generates gradients for the values in each memory block used for reconstruction, and stochastic gradient descent modifies the memory blocks toward minimizing the same training objective used for the network weights (cross-entropy loss).

4) Initializing the Codebook Matrix and CNN: CRUMB’s performance benefits from targeted initialization and pretraining of its CNN and the memory blocks in its codebook matrix, especially in the class-instance setting. The values in the codebook matrix $B$ directly replace those in “natural” feature maps derived from images during training—accordingly, $B$ is initialized using a simple univariate distribution designed to match that of natural feature maps from a pretrained network.

In early experiments, we tried initializing the codebook using $k$-means cluster centers from feature vectors of CIFAR100 images [57], but this did not improve the performance.

Stream learning performance was substantially improved by pretraining CRUMB on ImageNet [58] classification with 1000 classes, compared to applying CRUMB to stream learning with a CNN pretrained by standard methods. Pretraining tunes the values in the memory blocks and also regularizes the CNN in preparation for stream learning by training it to make predictions using lossy reconstructions of feature maps.

5) Replay to Mitigate Catastrophic Forgetting: In online stream learning (see Section III-A1), the model is presented with images $I_t$ from new classes $c^{new}$ in task $t$, where $c^{new}$ are drawn from the subset of classes in the dataset that the model has not seen in previous tasks.

Replay of examples from previous tasks is a proven strategy to mitigate catastrophic forgetting in class-incremental settings [17], [34], [35], [36], and feature-level replay can be considerably more memory-efficient than storing raw images [4]. We store compressed representations of feature maps from images in each task and then replay a batch of stored feature maps after each batch of new images during later tasks to mitigate forgetting.

CRUMB stores up to $n_X$ pairs of labels and tensors $(y_i, m_i)$, corresponding to images from old classes $c^{old}$ of previous tasks. Depending on the number of seen classes $C_t$, the storage for each old class contains $n_X/C_{t-1}$ pairs. $n_X$ is chosen for each dataset depending roughly on the total number of classes. Some algorithms select representative image examples to store and replay based on different scoring functions [67], [68], [69], [70]. However, random sampling uniformly across classes yields outstanding performance in continual learning tasks [22], and we adopt this strategy to select examples from the buffer for replay. To choose training examples for storage
in the replay buffer, CRUMB keeps every \( j \)th image in each batch, where \( j \) is calculated by dividing the number of training images in each task by the replay buffer capacity \( n_x \) such that the buffer is filled near the end of the current task’s training epoch. In the class-instance setting, this maximizes sample diversity by minimizing the number of frames sampled from within the same video clip and further avoiding sampling frames from the same clip that are in close temporal proximity.

For replay-based baseline methods (iCARL, REMIND, and so on), we limit the number of examples that can be stored in the buffer to fit within a memory budget that is fixed across all methods (see details in Supplementary Section S5). Aside from \( n_x \) and the training batch size (which is smaller for video datasets), CRUMB uses the same hyperparameters for all datasets, including \( n, d, \alpha, \beta \), learning rate, and initialization and pretraining protocols.

**IV. RESULTS**

**A. Stream Learning on Video Datasets**

A naive CNN trained on stream learning benchmarks learns each task effectively but rapidly and catastrophically forgets all prior tasks in doing so. In contrast, a brute-force approach to overcoming catastrophic forgetting that achieves excellent performance in a stream learning setting is to store all encountered images and corresponding class labels, shuffle them, and exhaustively retrain on the resulting dataset in an offline, i.i.d. fashion. This renders the benchmark equivalent to offline class-incremental learning (“upper bound” in Fig. 3) [11]. By storing a subset of old examples and using a compositional strategy to compress these examples, CRUMB allows CNNs to approach the performance of a brute-force approach with roughly an order-of-magnitude reduction in training time and
CRUMB outperforms state-of-the-art algorithms on most benchmarks. Each number is the mean top-1 accuracy on all tasks/classes after the completion of stream learning. Values are averaged from ten (CORe50, Toybox, iLab, iCub, iLab + CORe50, and Online-CIFAR100) or five (Online-ImageNet) independent runs. The highest accuracy in each column (excluding the offline upper bound) is in bold, while the second-highest is italicized. Algorithm name abbreviations can be found in Section III-A3. Class instance, in which video frames are presented in temporal order, is only applicable to video datasets CORe50, Toybox, iLab, iCub, and iLab + CORe50. Due to resource constraints, for Online-ImageNet, iCub, and CORe50 + iLab, We tested a subset of baseline algorithms. Class-instance and class-i.i.d. settings are abbreviated as “inst” and “iid,” respectively. Results are grouped vertically with memory distillation and replay methods at the top (ours—CoPe), weight regularizations in the middle (EWC—LwF), and lower/upper bounds at the bottom. Data preparation methods are detailed in Supplementary Section S4.A.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>inst</td>
<td>iid</td>
<td>inst</td>
<td>iid</td>
<td>inst</td>
<td>iid</td>
<td>iid</td>
</tr>
<tr>
<td>Ours</td>
<td>78.5</td>
<td>81.2</td>
<td>74.9</td>
<td>75.7</td>
<td>77.9</td>
<td>79.5</td>
<td>60.0</td>
</tr>
<tr>
<td>REMIND [4]</td>
<td>77.0</td>
<td>76.0</td>
<td>66.2</td>
<td>84.1</td>
<td>68.1</td>
<td>81.0</td>
<td>33.2</td>
</tr>
<tr>
<td>iCARL [35]</td>
<td>27.0</td>
<td>28.5</td>
<td>27.3</td>
<td>26.5</td>
<td>15.6</td>
<td>23.6</td>
<td>23.4</td>
</tr>
<tr>
<td>GEM [38]</td>
<td>11.9</td>
<td>13.5</td>
<td>14.3</td>
<td>15.7</td>
<td>13.0</td>
<td>12.8</td>
<td>5.2</td>
</tr>
<tr>
<td>RM [39]</td>
<td>12.0</td>
<td>12.4</td>
<td>9.8</td>
<td>20.8</td>
<td>18.2</td>
<td>9.3</td>
<td>-</td>
</tr>
<tr>
<td>AAN [60]</td>
<td>14.0</td>
<td>15.6</td>
<td>13.2</td>
<td>17.6</td>
<td>10.6</td>
<td>15.0</td>
<td>-</td>
</tr>
<tr>
<td>GSS [56]</td>
<td>15.0</td>
<td>16.6</td>
<td>14.7</td>
<td>15.0</td>
<td>13.0</td>
<td>12.8</td>
<td>-</td>
</tr>
<tr>
<td>BiC [34]</td>
<td>10.2</td>
<td>11.8</td>
<td>11.0</td>
<td>10.2</td>
<td>11.2</td>
<td>10.9</td>
<td>-</td>
</tr>
<tr>
<td>CoPe [61]</td>
<td>16.6</td>
<td>16.3</td>
<td>21.7</td>
<td>22.4</td>
<td>17.6</td>
<td>18.0</td>
<td>-</td>
</tr>
<tr>
<td>EWC [12]</td>
<td>12.2</td>
<td>12.4</td>
<td>14.3</td>
<td>15.7</td>
<td>13.5</td>
<td>12.0</td>
<td>7.3</td>
</tr>
<tr>
<td>MAS [65]</td>
<td>14.4</td>
<td>17.4</td>
<td>18.9</td>
<td>19.2</td>
<td>20.5</td>
<td>22.1</td>
<td>4.8</td>
</tr>
<tr>
<td>SI [19]</td>
<td>12.0</td>
<td>12.9</td>
<td>14.3</td>
<td>15.5</td>
<td>12.8</td>
<td>13.0</td>
<td>5.3</td>
</tr>
<tr>
<td>Stable SGD [62]</td>
<td>13.7</td>
<td>13.2</td>
<td>13.5</td>
<td>13.8</td>
<td>9.8</td>
<td>6.9</td>
<td>-</td>
</tr>
<tr>
<td>LwF [16]</td>
<td>12.5</td>
<td>12.9</td>
<td>21.9</td>
<td>20.9</td>
<td>10.5</td>
<td>11.9</td>
<td>-</td>
</tr>
<tr>
<td>Lower bound</td>
<td>12.1</td>
<td>12.8</td>
<td>15.5</td>
<td>16.9</td>
<td>12.8</td>
<td>16.4</td>
<td>5.8</td>
</tr>
<tr>
<td>Upper bound</td>
<td>85.3</td>
<td>84.6</td>
<td>91.0</td>
<td>92.0</td>
<td>91.3</td>
<td>91.4</td>
<td>76.9</td>
</tr>
</tbody>
</table>

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.
C. Memory and Runtime Efficiency

CRUMB’s closest competitor in top-1 accuracy is REMIND [4]. Both models require specific pretraining procedures. REMIND trains a product quantizer using k-means clustering of feature vectors, which requires a large portion of training data to be held in memory simultaneously. In contrast, CRUMB’s codebook matrix is trained by backpropagation in tandem with CNN parameter updates. This approach requires only 3.7%–4.1% of the peak RAM usage of REMIND for large datasets such as Online-CIFAR100 and Online-ImageNet. CRUMB also has a runtime only about 15%–43% as long as REMIND’s (Table II).

D. Model Analysis

To elucidate the importance of CRUMB’s various components, we performed a series of ablation studies and experiments with altered training procedures. Accuracy results on CORe50 in both class-instance and class-i.i.d. settings are included for each experiment in Tables III and IV, but throughout the text in this section, we discuss class-instance results except where otherwise stated. Experiment names are in bold throughout this section. We conducted statistical significance tests for each experiment (see Supplementary Section S4.B).  

1) Replay: n CRUMB Feature Maps Beat n Images: Feature-level replay is the main mechanism by which CRUMB prevents catastrophic forgetting. Removing replay dramatically reduces accuracy by 64.9%. However, CRUMB does not require storing a large number of feature maps to mitigate forgetting: reducing buffer size \(n_x\) from 200 (Ours) to 100 (Half capacity) and to 50 (Quarter capacity) had a relatively small impact, with 4.7% and 13.3% accuracy drops, respectively.

The quality of stored replay examples is also important. Ours, which stores memory block indices to compositionally reconstruct up to \(n_x\) feature maps, had dramatically higher accuracy than storing the same number \(n_x\) of entire raw images and training the network without any feature map reconstruction (Image replay), even though CRUMB’s reconstruction of feature maps inevitably discards information and uses only 3.6% as much memory. As shown in Table III, Ours attains 10.4% higher accuracy than Image replay on CORe50 (buffer size \(n_x = 200\)), 4.2% higher on Toybox, 15.4% higher on iLab, 7.8% higher on iCub, and 2.8% higher on iLab + CORe50. This result appears to hold only for the five video streaming datasets; however: Ours attained accuracy 0.16% higher than Image replay on Online-CIFAR100 (not statistically significant) and 18.3% lower than Image replay on Online-ImageNet, in the class-i.i.d. setting. Ours uses only 3.6% as much memory as Image replay: when the amount of memory usage is held constant (e.g., in gigabytes) instead of the maximum number of replay buffer items \(n_x\), CRUMB outperforms image replay methods such as iCARL by very large margins on all datasets (see Table I).

Table III also shows that CRUMB continues to outperform image replay on CORe50 when memory resources for the replay buffer are not constrained. Even at \(n_x = 6400\), where algorithms can store the entire CORe50 training set in the replay buffer, Ours outperforms Image replay by 9.8% on class instance and 3.2% on class-i.i.d.. The reduced performance of Image replay relative to Ours is partly rescued by adding CRUMB pretraining (Ours p.t. + im. rep., 3.7% and 2.2% below Ours in class instance and class-i.i.d., respectively), even though the memory blocks play no role in either of the two image replay conditions during streaming (see Table IV).

Replaying high-level features also contributed to CRUMB’s performance. Storing \(n_x\) low-level feature maps from layer 3 instead of layer 12 (Early feature replay versus Ours) reduced performance by 16.6%. CRUMB effectively stores memories with a higher level of abstraction than both Image replay and Early feature replay and comes with both accuracy and memory efficiency improvements. The memory recall (MeRec) method [48] uses a further level of abstraction by storing only the elementwise mean and standard deviation of feature activations for each learned class and generating examples for replay by sampling from a Gaussian distribution.
parameterized by these values. In MeRec replay, we implemented MeRec’s replay mechanism as a drop-in replacement for CRUMB’s memory block reconstruction at the same network layer and observed the accuracy of 56.9% and 41.3% below Ours for class instance and class-i.i.d., respectively, but 8.1% and 28.1% above No replay. MeRec stores an amount of data equivalent to two complete feature maps (mean and standard deviation) per class. In the implementation of CRUMB used for Ours, this is equivalent to 16 compressed feature maps per class or \( N = 160 \) in total, which is fewer than Ours at \( n_c = 200 \) but more than Half capacity at \( n_c = 100 \).

2) CRUMB Pretraining Primes CNN Weights for Streaming: Our model’s performance is maximized by using CRUMB to pretrain the CNN and memory blocks on ImageNet prior to stream learning, particularly in the class-instance condition. Using randomly initialized memory blocks and a CNN pretrained without CRUMB (Vanilla pretrain) instead of CRUMB pretraining (Ours) reduced performance by 25.5% and 3.2% on class instance and class-i.i.d., respectively. Our results also indicate that the benefit of CRUMB pretraining is attributable primarily to changes in the CNN weights, rather than changes to the memory blocks. Starting stream learning with CRUMB-pretrained weights and randomly reinitialized memory blocks (Pretrain weights) performs only 1.1% and 0.5% worse than Ours, while starting with vanilla-pretrained weights and CRUMB-pretrained memory blocks (Pretrain mem. blocks) is 23.7% and 2.8% worse than Ours, a marginal improvement over vanilla pretraining. As explained in Section IV-D1, CRUMB pretraining also improves stream learning performance when raw images are used for replay.

CRUMB pretraining using the smaller CIFAR100 dataset (100 classes) instead of ImageNet (1000 classes) (CIFAR100 pretrain) decreases accuracy by 11.6%.

In Freeze memory, no updates to memory blocks were allowed after pretraining. This had no statistically significant effect on accuracy, indicating that fine-tuning the memory blocks was unnecessary for stream learning on CORoSe50.

3) CRUMB Pretraining Induces a Shape Bias in the CNN: We hypothesized that CRUMB’s pretraining procedure induces a bias toward shape information over texture information by training the CNN to make class predictions using lossy feature representations with unperturbed spatial distributions. A bias toward shape information has been shown to help mitigate forgetting by flattening the local minima for each task in the loss landscape [72]. To gauge CRUMB’s degree of reliance on shape and texture information, we evaluated its performance on test set examples with three different perturbations.

“Spatial perturbation” and “feature perturbation” modify the \( 13 \times 13 \times 512 \) feature map at the same level where it is reconstructed using memory blocks. In “spatial perturbation,” the positions of all of the 512-D feature vectors are randomly shuffled in the \( 13 \times 13 \) spatial grid, destroying global shape information but leaving feature information intact. In “feature perturbation,” for each image independently, we set a random selection of 50% of the features (i.e., 256 out of 512 total features) to zero, perturbing feature information but leaving coarse shape information intact. Spatial and feature perturbations do not directly interact with CRUMB’s reconstruction mechanism because CRUMB uses the original, nonreconstructed feature map to make class predictions. The “feature perturbation” strategy assumes that feature information is more related to texture information than shape information. Therefore, we also include “style perturbation” for ImageNet by testing on the Stylized-ImageNet dataset, which consists of images with heavily distorted local textures but largely intact global object shapes [71]. We would expect the performance of a relatively shape-biased network to be more severely reduced by spatial perturbation and less severely affected by feature and style perturbations than a control network. Fig. 4 shows CRUMB’s “relative accuracy advantage” for each perturbation across several datasets. The relative accuracy advantage is calculated by dividing the accuracy drop (unperturbed accuracy minus perturbed accuracy) for each perturbation by the network’s unperturbed accuracy and subtracting this value for CRUMB from the corresponding value for a control network. The “ImageNet (pretraining)” condition in Fig. 4 shows CRUMB’s shape bias on ImageNet following pretraining on ImageNet, with lower resilience against spatial perturbations (red bars with diagonal lines) and higher

**TABLE IV**

<table>
<thead>
<tr>
<th>Category</th>
<th>Experiment name</th>
<th>Class-inst. % avg. accuracy</th>
<th>Class-iid. % avg. accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unablated</td>
<td>Ours</td>
<td>78.22</td>
<td>79.93</td>
</tr>
<tr>
<td>Replay format</td>
<td>Image replay</td>
<td>67.80*</td>
<td>75.63*</td>
</tr>
<tr>
<td></td>
<td>Ours p.t. + im. rep.</td>
<td>73.45*</td>
<td>77.72*</td>
</tr>
<tr>
<td></td>
<td>Early feature replay</td>
<td>61.64*</td>
<td>64.28*</td>
</tr>
<tr>
<td></td>
<td>MeRec replay [18]</td>
<td>21.35*</td>
<td>38.55*</td>
</tr>
<tr>
<td>Pretraining ablation</td>
<td>Half capacity</td>
<td>73.55*</td>
<td>75.14*</td>
</tr>
<tr>
<td></td>
<td>Quarter capacity</td>
<td>64.90*</td>
<td>67.80*</td>
</tr>
<tr>
<td></td>
<td>No replay</td>
<td>13.28*</td>
<td>10.58*</td>
</tr>
<tr>
<td>Memory block init.</td>
<td>Vanilla pretrain</td>
<td>52.70*</td>
<td>66.60*</td>
</tr>
<tr>
<td></td>
<td>Pretrain weights</td>
<td>77.05*</td>
<td>79.40*</td>
</tr>
<tr>
<td></td>
<td>Pretrain mem. blocks</td>
<td>54.54*</td>
<td>77.18*</td>
</tr>
<tr>
<td></td>
<td>CIFAR100 pretrain</td>
<td>66.64*</td>
<td>74.96*</td>
</tr>
<tr>
<td>Freeze memory</td>
<td>Freeze memory</td>
<td>78.06</td>
<td>80.44</td>
</tr>
<tr>
<td>Memory block size Number of memory blocks</td>
<td>Uniform init.</td>
<td>39.84*</td>
<td>64.82*</td>
</tr>
<tr>
<td>256 blocks (Ours)</td>
<td>Dense matched init</td>
<td>77.24</td>
<td>78.64*</td>
</tr>
<tr>
<td>512 blocks</td>
<td>9.60*</td>
<td>9.47*</td>
<td></td>
</tr>
<tr>
<td>2 blocks</td>
<td>64.25*</td>
<td>71.23*</td>
<td></td>
</tr>
<tr>
<td>4 blocks</td>
<td>70.10*</td>
<td>75.77*</td>
<td></td>
</tr>
<tr>
<td>8 blocks</td>
<td>74.60*</td>
<td>79.96*</td>
<td></td>
</tr>
<tr>
<td>16 blocks</td>
<td>77.70</td>
<td>80.30</td>
<td></td>
</tr>
<tr>
<td>Memory block size</td>
<td>4-dim. blocks</td>
<td>74.21*</td>
<td>77.55*</td>
</tr>
<tr>
<td>8-dim. blocks (Ours)</td>
<td>78.22</td>
<td>79.93</td>
<td></td>
</tr>
<tr>
<td>16-dim. blocks</td>
<td>79.05*</td>
<td>81.02*</td>
<td></td>
</tr>
<tr>
<td>32-dim. blocks</td>
<td>78.17</td>
<td>79.93</td>
<td></td>
</tr>
<tr>
<td>16-dim. blocks adj.</td>
<td>79.09*</td>
<td>82.36*</td>
<td></td>
</tr>
<tr>
<td>32-dim. blocks adj.</td>
<td>80.31*</td>
<td>81.54*</td>
<td></td>
</tr>
<tr>
<td>Loss functions</td>
<td>Ours - direct loss</td>
<td>73.82*</td>
<td>78.11*</td>
</tr>
<tr>
<td></td>
<td>Ours + direct loss</td>
<td>65.40*</td>
<td>69.20*</td>
</tr>
<tr>
<td></td>
<td>Direct loss</td>
<td>48.12*</td>
<td>50.07*</td>
</tr>
<tr>
<td>CNN Architecture</td>
<td>MobileNetv2 CNN</td>
<td>76.15*</td>
<td>79.27</td>
</tr>
</tbody>
</table>
4) CRUMB Can Learn With Very Few Memory Blocks: CRUMB’s performance did not change dramatically with changes to the number of memory blocks. Reducing the memory block count from 256 blocks to as few as 16 blocks, which effectively shrinks the library of feature combinations available to reconstruct feature maps, did not significantly decrease accuracy. Reducing the count further to eight blocks decreased accuracy by 3.6%, and reducing to four or two blocks decreased accuracy by 8.1% and 13.9%, respectively. Increasing to 512 blocks did not significantly increase accuracy. This suggests a saturation effect, where a relatively small number of memory blocks are sufficient to reconstruct a wide variety of feature maps.

5) CRUMB Is Robust to Different Memory Block Sizes, and Memory Block Size Affects Memory Efficiency: CRUMB performs well with a range of memory block sizes. Decreasing the number of elements in each memory block from 8 to 4 (4-dim. blocks) results in a modest decrease in performance, 4.0% and 2.4% on class instance and class-i.i.d., respectively. Increasing the number of elements from 8 to 16 or 32 (16-dim. blocks and 32-dim. blocks), which arguably makes accurate reconstruction of feature maps more challenging because higher dimensional vectors must be replaced by discrete choices of memory blocks, had negligible impact on performance (see Table IV).

The maximum number of examples stored in CRUMB’s replay buffer ($n_x$) was held constant for the memory block size perturbations above. However, increasing the dimension of the memory blocks from 8 to 16 or 32 means that only half- or one-quarter as many blocks, respectively, are needed to reconstruct each feature map, so only half-one-quarter as many indices need to be stored in the replay buffer per image. This allows double/quadruple the number of examples to be stored in the replay buffer within the same memory budget. When we allowed the maximum number of examples stored in the buffer to change accordingly ($2n_x$ for 16-dim. blocks adj. and $4n_x$ for 32-dim. blocks adj.), we observed accuracy improvements: 16-dim. blocks adj. achieves 1.5% and 2.4% higher accuracy than Ours ($n_x$ with 8-D blocks) on class instance and class-i.i.d., respectively, and 32-dim. blocks adj. achieves 2.1% and 1.7% higher accuracy. During hyperparameter tuning for our main results, we observed that 16-D memory blocks maximized testing accuracy.

6) Loss From Reconstructed Features Is Sufficient: CRUMB’s performance is affected by the choice of components in its loss function. The loss function [see (3)] is the weighted sum of two terms, “direct loss” and “codebook-out loss.” Our experiments show that the best performance is achieved when both direct loss and codebook-out loss are included in pretraining, but only codebook-out loss is included during stream learning. Removing direct loss from pretraining (“Ours — direct loss”) results in a 4.4% drop in accuracy in the later stream learning tasks—learning from only reconstructed feature maps from start to finish, including during pretraining, is sufficient for decent performance. Including only codebook-out loss (“Ours”) in stream learning yields a dramatic 30.1% gain in accuracy compared to using only direct loss (“Direct loss”) and a gain of 12.8% compared to using a weighted sum of direct loss and codebook-out loss (“Ours + direct loss”), despite the fact that only the direct, nonreconstructed feature map is used for inference on the test set.

7) Initialization of the Memory Blocks Matters: CRUMB’s performance is somewhat sensitive to the initialization of the values in the memory blocks. CRUMB trains its memory blocks in tandem with network weights after initialization and concatenates them in different combinations to reconstruct feature maps produced by an intermediate network layer. We compared the stream learning performance of four memory block initialization strategies, including initializing with values drawn from: 1) a standard normal distribution (Normal init.);
2) a uniform distribution on the interval [0, 1] (Uniform init.); 3) a distribution designed to match that of the nonzero values in the feature maps to be reconstructed, with 64% of all values reset to zero to approximately match the sparsity of typical feature maps (Ours); and 4) the same as 3), but with no values set to zero (Dense matched init.). Accuracy for Normal init. was 30.5% and 5.7% lower than Ours for class-instance and class-i.i.d. protocols, respectively, accuracy for Uniform init. was 38.4% and 15.1% lower, and accuracy for Dense matched init. was 1.0% (p = 0.045 > 0.01) and 1.1% lower (see Table IV). It appears that drawing initial values for the memory blocks from a similar distribution to that of natural feature maps improves the performance. When applying CRUMB to new network architectures, a simple alternative procedure to initialize the memory blocks would be to obtain feature maps from a batch of images, pool all values from all feature maps into one long vector, and initialize each memory block value by randomly sampling a value from this vector.

8) CRUMB Is Applicable Across CNN Architectures: In MobileNetV2 CNN, we implement CRUMB using the MobileNetV2 [73] CNN backbone instead of SqueezeNet [59]. Here, CRUMB reconstructs the 14 × 14 × 64 input to MobileNetV2’s sixth layer, using 256-D memory blocks as in Ours. With the total memory usage of the replay buffer held constant, the performance of MobileNetV2 CNN is comparable to Ours with only a 2.1% accuracy drop in class instance and a 0.7% (not significant) drop in class-i.i.d.

9) Some Memory Blocks Are Coarsely Interpretable: Visualizations of image locations where specific memory blocks are activated (Fig. 5) show that some memory blocks appear to be human-interpretable. Some blocks responded to features seen in images of one specific class or of a subset of classes, and others responded to features that are likely irrelevant to classification. In addition to the blocks visualized in Fig. 5, we found blocks that tend to respond to vertical lines, crosshatch patterns on balls and cups, pure white backgrounds, vegetation backgrounds, and wooden floor backgrounds, each of which can be interpreted as a semantic, compositional part of various test set images. Given the observations earlier in this section that either randomly reinitializing memory blocks prior to stream learning (Pretrain weights) or freezing memory blocks during stream learning (Freeze memory) has minimal effects on performance, it is not necessarily the case that these interpretable associations indicate learned representations within the memory blocks themselves. Another possibility is that a sufficient diversity of memory blocks allows useful associations between memory blocks and features or classes to be learned by the CNN through changes to the network weights.

The procedure for generating the visualizations in Fig. 5(a) can be understood as follows. For this analysis, we used a CRUMB model trained on CORE50 in the class-instance setting. CORE50 test set images are first passed through the early layers of the CNN to produce a feature map, which CRUMB then reconstructs by concatenating memory block vectors to produce an approximated version of the original feature map (see Section III-B2). In this study, each feature map is of size 13 × 13 × 512, meaning spatial dimensions of 13 × 13 with 512 features at each spatial location. Each memory block is one of the 256 row vectors in the 256 × 8 codebook matrix used for this analysis. The memory blocks are 8-D vectors, so each spatial location in the feature map’s 13 × 13 grid is represented by a 512-D vector formed by concatenating 512/8 = 64 memory blocks end-to-end. Color coding in Fig. 5(a) shows at most one block per spatial location, the one activated by the first eight features in the 512-D feature vector, even though 64 memory blocks are activated at each location in total. We focus on the first eight features for visualization purposes because it is not necessarily the case that blocks activated by the first set of eight features encode the same image features as they might when activated by the kth set of eight features (where 2 ≤ k ≤ 64). To produce the images in Fig. 5(a), each test set image is divided into a square 13 × 13 grid. Image grid locations are overlaid with colored squares such that the color of each square depends on the memory block activated by the first eight features at the corresponding spatial location in the feature map reconstructed by CRUMB. We only assigned colors to a handful of memory blocks with interesting properties and the same color to sets of memory blocks that seemed to respond to very similar features. Fig. 5(b) shows the sorted distribution of the frequencies with which each of the 256 memory blocks was used to reconstruct feature maps from the CORE50 test set, with color and memory block index-coded arrows indicating the memory blocks visualized in Fig. 5(a).

V. CONCLUSION AND DISCUSSION
We developed a novel compositional replay strategy to tackle the problem of online stream learning, in which
algorithms must learn tasks incrementally from nonrepeating, temporally correlated inputs. Our algorithm, CRUMB, learns a set of “memory blocks” that are selected via cosine similarity and concatenated to reconstruct feature maps from an intermediate CNN layer. The indices of selected memory blocks are stored for a subset of training images, enabling memory-efficient replay of feature maps to mitigate catastrophic forgetting. CRUMB achieves state-of-the-art online stream learning accuracy across seven datasets. Furthermore, CRUMB outperforms replay of an equal number of raw images by large accuracy margins across five video datasets, despite using only 3.6% as much memory as image replay.

Several factors seem to make important contributions to CRUMB’s high performance. As shown in Fig. 4, pretraining with memory block reconstruction biases the CNN toward attending to object shapes rather than textures, an effect that typically endures throughout stream learning and which has been demonstrated to reduce catastrophic forgetting by flattening the loss minima for each task [72]. This could explain why CRUMB pretraining improves performance even if raw image replay is used during stream learning (“Ours p.t. + im. rep.” in Table IV), in which case feature map reconstruction plays no role whatsoever during stream learning.

Backpropagation updates to memory blocks during pretraining and stream learning appear to be less important for performance than updates to the CNN weights. Randomly reinitializing the memory blocks after pretraining has a small negative effect on performance only in the class-instance setting while keeping only the pretrained memory blocks but resetting to the original “vanilla” pretrained CNN weights before stream learning has a much larger negative impact (“pretrain weights” versus “pretrain mem. blocks” in Table IV). Furthermore, freezing the memory blocks during streaming has no discernible effect (“freeze memory” in Table IV). However, “normal init.” and “uniform init.” in Table IV show that the choice of probability distribution used to randomlyinitialize the memory blocks can have a dramatic impact on performance, with best performance attained by matching the univariate distribution of the memory blocks to that of natural feature maps. It seems to be important for stream learning that the later layers of the network receive feature maps with consistent univariate statistics, whether they are natural feature maps from the feature extractor layers or reconstructed feature maps from memory block concatenation.

Experiments in Table IV also show that the design of CRUMB’s loss function is important. When training on new images, using only “codebook-out loss” from classification on reconstructed feature maps leads to much less forgetting than using “direct loss” from natural feature maps, either together with codebook-out loss (“ours + direct loss”) or in isolation (“direct loss”). Only codebook-out loss is available when replaying feature maps reconstructed from memory blocks: using only codebook-out loss for new images means that only codebook-out loss is used throughout stream learning, rather than switching between direct loss for new examples and codebook-out loss for replayed examples. It appears that CRUMB’s memory blocks form a shared, discretized basis for encoding training examples in feature space, which has a stabilizing effect on CNN during stream learning. It is also notable in this context that, unlike during streaming, the inclusion of direct loss is important for gradient updates during pretraining (“Ours — direct loss” has lower accuracy than “Ours” in Table IV). Combined with the observation that memory blocks should ideally be initialized from a univariate distribution approximating that of natural feature maps, this suggests that the pretraining process helps the CNN align its processing of natural and reconstructed feature maps, enabling the network to learn only from reconstructed feature maps during streaming even while continuing to make its most accurate predictions using natural feature maps instead.

The hypothesis that CRUMB’s memory blocks provide a shared feature-level basis that stabilizes the CNN is consistent with the observation that, although CRUMB pretraining improves the performance of raw image replay (“Ours p.t. + im. rep.” in Table IV), it still does not match CRUMB’s performance with the replay buffer size held constant: we speculate that redundant pixel-level information in raw images introduces additional noisy variation into the distribution of feature maps, which affects network stability. Another observation consistent with this hypothesis is that CRUMB only outperforms raw image replay on the five video datasets (with constant n_s, not constant memory usage), where network instability is likely to be more problematic due to temporal correlations within video clips and sudden transitions between them during training.

In both CRUMB and our raw image replay ablation experiments, the early “feature extractor” layers of the CNN are frozen. When we apply CRUMB reconstruction to feature maps from an earlier layer (“early feature replay” in Table IV) and correspondingly allow more layers after this point to have their parameters updated during stream learning, we observe performance worse than both CRUMB and image replay. One interpretation of this is that more unfrozen layers means that more network parameters are exposed to gradient updates and, consequently, to catastrophic forgetting. It is also possible that CRUMB’s reconstruction mechanism is best suited to representing abstract, high-level features that are more likely to be found in later CNN layers, and that too much information is lost if CRUMB attempts to represent lower level information that is interpreted by later layers in more granular ways.

Given the apparent necessity of preserving the information in feature maps during reconstruction, a surprisingly small codebook of memory blocks is sufficient. As few as 16 memory blocks are needed for optimal performance on CORRe50, CRUMB still performs remarkably well with only two memory blocks (see “number of memory blocks” experiments in Table IV). Although CRUMB is already highly memory-efficient with the memory blocks themselves occupying negligible space, reducing the number of memory blocks may enable further CPU memory usage optimizations (e.g., 4-bit integers as indices for 16 memory blocks) and also lowers GPU memory usage. Computational and memory
efficiency is presumably critical in biological memory systems. Indeed, replay of neuronal activity patterns has been observed to help reinforce and consolidate memories in multiple brain areas across different mammalian species [50], [51], [52]. It is unlikely that neural circuits in the brain use replay mechanisms that preserve as much low-level information as pixel-level replay. Instead, it is interesting to speculate that one of the mechanisms by which brains avoid catastrophic forgetting is by replaying compositions of abstract, high-level features in a manner analogous to CRUMB’s replay mechanism.

CRUMB’s superior memory and runtime efficiency makes it ideally suited for settings with limited computational resources. Potential applications include edge computing in mobile devices and autonomous robots that learn continuously from otherwise unmanageable amounts of incoming sensor data, while they explore their surroundings. CRUMB could also be used in federated learning contexts, enabling highly effective replay of previously unseen data points via perhaps unrecognizably lossy representations, thereby minimizing both catastrophic forgetting and data security risks. The interpretable qualities of a subset of memory blocks, however, raise the possibility of identifying weak associations with certain generic features contained in a given training example, for a person with unauthorized access to CNN weights, memory blocks, encoded memories of interest, and a reference dataset to discover memory block interpretations. However, it would still be impossible for such a person to completely reconstruct CRUMB’s memories in their original encodings (e.g., in pixels).

CRUMB is implemented here for SqueezeNet [59] and MobileNetV2 [73] CNNs but could be used to mitigate forgetting across different neural network architectures in the future. For example, memory blocks could be used to efficiently reconstruct and replay vector outputs of self-attention heads at intermediate layers in transformer models [74], [75].

Updating CRUMB’s memory blocks using backpropagation in tandem with network weights is highly efficient and also raises the possibility of tuning memory blocks for shifting domains on the fly. Although updates to the memory blocks beyond pretraining do not appear important for stream learning on CORe50, it is possible that fine-tuning may be necessary for tasks with substantial nonstationarity. In addition, in this study, CRUMB does not adopt the early “feature extractor” layers of the CNN during stream learning. However, the early layers could theoretically be trained using the direct prediction loss while the late layers and memory blocks are trained using codebook-out loss or a combination of these two losses: this approach could enable additional flexibility for domain adaptation. Future studies could apply CRUMB to stream learning or reinforcement learning tasks with shifting domains, emulating humans or robots in continuously changing environments.

ACKNOWLEDGMENT

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. The authors would like to thank Trenton Bricken, Spandan Madan, and Warren Sunada-Wong for sharing their helpful insights into their work at various times during this project.

REFERENCES


Morgan B. Talbot is currently pursuing the M.D. and Ph.D. degrees with the Harvard-MIT Program in Health Sciences and Technology. His research interests include learning and memory in humans and machines, as well as applications of artificial intelligence to psychiatry and mental health.

Rushikesh Zawar received the B.E. degree in computer science and the M.Sc. degree in biological sciences from the Birla Institute of Technology and Science at Pilani (BITS Pilani), Pilani, India. He completed his thesis at Harvard University, Cambridge, MA, USA. His research interests include computer vision, reinforcement learning, neuroscience, and the intersection of artificial intelligence and biology.

Rohil Badkundri is currently pursuing the bachelor’s degree in applied math with Harvard University, Cambridge, MA, USA. He is broadly interested in applications at the intersection of machine learning and biology.

Mengni Zhang is currently an Assistant Professor and a Principal Investigator of the Deep NeuroCognition Laboratory, Nanyang Technological University (NTU), Singapore, and the Agency for Science, Technology and Research (A*STAR), Singapore. Her research background is at the intersection of artificial intelligence and computational neuroscience. She has made contributions to understanding gaze anticipation, zero-shot visual search, contextual reasoning, and continual learning.

Gabriel Kreiman is currently a Professor at the Harvard Medical School, Boston, MA, USA, and Boston Children’s Hospital, Boston, and leads the Executive Function/Memory Module at the Center for Brains, Minds and Machines. His research group combines computational models, behavioral measurements, and neurophysiological recordings to study visual cognition, learning, and memory.