Remember the Past: Distilling Datasets into Addressable Memories for Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose an algorithm that compresses the critical information of a large dataset into compact addressable memories. These memories can then be recalled to quickly re-train a neural network and recover the performance (instead of storing and re-training on the full original dataset).

Building upon the dataset distillation framework, we make a key observation that a *shared common representation* allows for more efficient and effective distillation. Concretely, we learn a set of bases (aka "memories") which are shared between classes and combined through learned flexible addressing functions to generate a diverse set of training examples. This leads to several benefits: 1) the size of compressed data does not necessarily grow linearly with the number of classes; 2) an overall higher compression rate with more effective distillation is achieved; and 3) more generalized queries are allowed beyond recalling the original classes.

We demonstrate state-of-the-art results on the dataset distillation task across five benchmarks, including up to 16.5% and 9.7% accuracy improvement when distilling CIFAR10 and CIFAR100 respectively. We then leverage our framework to perform continual learning, achieving state-of-the-art results on four benchmarks, with 23.2% accuracy improvement on MANY.

1 Introduction

Compressing a large amount of information into a small memory storage space is one of the key components in human intelligence [1, 2, 3] – a person can retrieve the memories from the past and fastly recover the corresponding skills. Deep learning methods have made large strides in building task-specific models, but are shown to easily forget the past knowledge when learning new tasks [4, 5].

To equip neural network learners with the memorizing ability, dataset distillation [6] is proposed as a potential solution, which performs pragmatic compression and prioritize information that affects the model training for compression. Concretely, a compressed set of examples (memories) is learned to summarize the key information in a dataset that affects model training; these examples can then be used to quickly retrain models and recover the corresponding skills. This differs from the standard reconstruction-based compression algorithms [7, 8, 9] and shows strong performance [10, 11, 12, 13].

A critical question in building powerful compressed memories is: what structures and representations should we use to build the memories? An effective structure and organization of memories can lead to different fundamental assumptions about data and affect the compression and learning behaviours. Existing works [10, 11, 14, 12, 15, 13, 6] follow a simple representation, where a set of learnable examples is assigned for each class. However, under this assumption, the size of the memories can linearly grow with the number of classes, making the distillation of dataset with large number of classes challenging. Naturally, this can potentially lead to redundancies in the learned memories,

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

due to the separation of data among classes. Furthermore, this representation is less generalizable to continuous label space, where infinite number of label values exists.

In our paper, we make the observation that there are information sharing among classes, and hypothesize that a common and compact representation exists for all classes. Following the hypothesis, we propose to formulate the problem as a memory addressing process, where the memories store a common set of basis shared by all labels, and the recombination of basis is performed through the addressing function. This decomposition between memories and addressing functions enables the possibility that all common information is stored in one part of the representation, and the accessing of the common information depends on the specific labels and is handled through an extra function. We find that this formulation can significantly improve both the compression rate and the performance.

We adopt the back-propagation through time learning framework to train the memories and addressing functions, and identify several critical factors that can improve the performance. Specifically, we find that adopting the momentum term, and performing long unrolls in the inner optimization loop, can drastically improve the performance. This differs from the common usage of bi-level optimization algorithm on this task [6, 16], and leads to strong performances outperforming single-step gradient matching methods [10, 11] even with the simple data representation.

In the experiments, we extensively evaluate our algorithm on five benchmarks of the Dataset Distillation task, and show that it consistently outperforms previous state-of-the-art by a significant margin. For example, we achieve 66.4% accuracy on CIFAR10 with the storage space of 1 image per class, improving over the previous state-of-the-art KIP method [12, 13] by 16.5%. We further demonstrate our method on the continual learning tasks, and show that a simple "compress-then-recall" method using our framework leads to state-of-the-art results on four datasets. For example, we outperform all prior methods by 23.2% in retained accuracy on the challenging MANY [17] benchmark. Finally, we demonstrate the generality of our approach by extending to image-based (rather than label-based) memory recall, and synthesizing new classifiers (unseen during training) from our distilled memories.

2 Related works

Dataset Distillation. The task of dataset distillation is fundamentally a compression problem, with a different prioritization on the information contained in data. There have been mainly several lines of methods, developed with different criteria to prioritize information. *Generalization loss* with bi-level optimization framework [18, 19, 20] has been widely adopted in various tasks and is adopted in the early works of dataset distillation [6, 16]. It emphasizes the loss at the final optimization state. *Gradient-matching or score-matching* methods [10, 11, 15] are adopted to directly match the induced gradients from synthetic data. If ideally matched over the gradient field, the compressed dataset can naturally lead to the same model parameters with gradient descent. *Kernel method* [13, 12] shows that with the connection to Gaussian processes, a kernel inducing points method can be adopted to achieve strong performance, but large computation costs. These are also connected with the recent progress on pragmatic compression methods [21, 22], which compress or match distributions based on a decision process (in dataset distillation's case, the gradient descent search process).

Continual learning. Broadening the learning paradigms, continual learning problem aims to build agents that learn through a stream of tasks and accrue knowledge along the process. "Catastrophic forgetting" [5, 23, 4] is a well-known phenomenon in this setting, where the neural network forgets previous skills when learning new ones. Various methods [24, 25, 26, 27, 17, 28, 29, 30, 31, 32, 33, 34, 35] on regularization, replay, or dynamic model, have been proposed to alleviate the issue and address the "stability-plasticity dilema" [36, 37]. Memory buffer has been a critical component in the past methods [24, 25, 26, 27, 17, 28, 38, 39], but mainly relies on random selection of real samples with different strategies. Recently, several works extend the usage of memory to storing random basis [39] (online setting) or SVD bases [38] (offline setting).

3 Background: dataset distillation

The task of Dataset Distillation [6] is proposed to compress the key information of a large-scale training dataset into a small amount of learned data, which can be stored using limited memory space and retrieved through label indices or task information to recover the performance of a model.

Problem Setting. Formally, given a large dataset $\mathcal{D}_{tr} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ containing N pairs of training data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, where $\boldsymbol{x}_i$ is an image and $\boldsymbol{y}_i$ is the corresponding label in C classes, a small dataset

$\mathcal{D}_s = \{(\boldsymbol{x}'_i, \boldsymbol{y}_i)\}_{i=1}^{N'}, N' \ll N$, can be synthesized or distilled, such that a model trained on $\mathcal{D}_s$ can have the same generalization ability with ones trained on $\mathcal{D}_{tr}$:

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_{te}}\left[\mathbf{m}(f(\boldsymbol{x};\boldsymbol{\theta}^{(*)}),\boldsymbol{y})\right] \simeq \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_{te}}\left[\mathbf{m}(f(\boldsymbol{x};\boldsymbol{\theta}'^{(*)}),\boldsymbol{y})\right] \tag{1}$$

where $\boldsymbol{\theta}^{(*)}$ and $\boldsymbol{\theta}'^{(*)}$ are the optimized parameters using $\mathcal{D}_{tr}$ and $\mathcal{D}_s$ respectively, $\mathbf{m}$ is a metric, e.g., accuracy, and $\mathcal{D}_{te}$ is the test dataset. The model is often a neural network classifier $f(\cdot;\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$ and can be trained with a loss function $\ell(f(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{y})$.

Synthetic dataset representations. The synthetic dataset contains the core parameters that need to be learned in dataset distillation. The representation of the synthetic data affects the compactness and effectiveness of the distillation process. In the existing methods [6, 10, 11, 14, 12, 15], the dataset $\mathcal{D}_s$ is defined as a collection of learnable data samples $(\boldsymbol{x}', \boldsymbol{y})$, and the number of samples is separately and equally distributed across classes. This representation has several disadvantages: firstly, the number of synthetic data samples needed for a dataset linearly grows with the number of classes, leading to limited applicability when the number of classes is large or undefined (e.g., language or other continuous labels); secondly, the potentially shared and common information across classes is ignored - this results in a less compact representation of the distilled information and lower compression rate; lastly, the representation is not able to generalize to new classes or tasks, due to the lack of common representation learned across classes.

The underperforming back-propagation through time algorithm (BPTT). Bi-level optimization using generalization loss to back-propagate gradients through inner loops is widely adopted across different tasks [20, 18, 40]. However, albeit used in the original dataset distillation framework [6], it has been shown to consistently underperform other competing algorithms, such as gradient matching methods [10, 11]. In our paper, we point out several critical aspects that affect the performance and show that BPTT can outperform single-step gradient matching methods by a large margin, which contrasts with the current common observations in this community.

4 Model

Overview. In this section, firstly, we present a new perspective of the problem, where the Dataset Distillation problem is formulated as a *memory addressing process* - instead of learning synthetic images separately for each class - we construct and learn a common memory representation that can be addressed through an addressing matrix to construct synthetic datasets. Under this formulation, the number of synthetic images doesn't have to grow linearly with the number of classes, the shared information among classes can be exploited to reduce redundancies and improve compression rate, and datasets can be distilled with respect to more generic queries. Secondly, we further show several critical empirical facets of back-propagation through time framework, which lead to drastic improvements on the performance and outperform the single-step gradient matching methods. This is in contrast with the current common observation that gradient matching outperforms back-propagation through time framework on dataset distillation tasks.

In the following, we present the two core components of our method, (1) the new formulation of dataset distillation, with memories and addressing matrices in Sec. 4.1, and (2) the learning framework under back-propagation through time in Sec. 4.2.

4.1 Dataset distillation as memory addressing

Problem formulation. Given a task-specific dataset $\mathcal{D}_{tr} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$, we aim to learn a single compact representation $\mathcal{M}$, referred to as memories, that can be accessed through a learned addressing function $\mathcal{A}(\cdot)$. which takes all possible values of $\boldsymbol{y}$ as input and recall the corresponding synthetic data. With a set of $\{\boldsymbol{y}_i\}$, a synthetic dataset recalled using the above process can train a model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ from scratch and obtain the same generalization ability as trained on $\mathcal{D}_{tr}$.

Ideally, the memories $\mathcal{M}$ and the addressing function $\mathcal{A}$ can jointly capture the critical information that defines the task mapping from $\boldsymbol{x} \in \mathcal{X}$ to $\boldsymbol{y} \in \mathcal{Y}$, such that, when training a model using the synthetic dataset collected through accessing memories with a set of $\{\boldsymbol{y}_i\}$, the trained model can obtain similar or same generalization ability with ones trained on $\mathcal{D}_{tr}$. For example, in the standard classification tasks, we can enumerate all possible values in the label space (discrete) and address the memories, to construct the synthetic dataset.

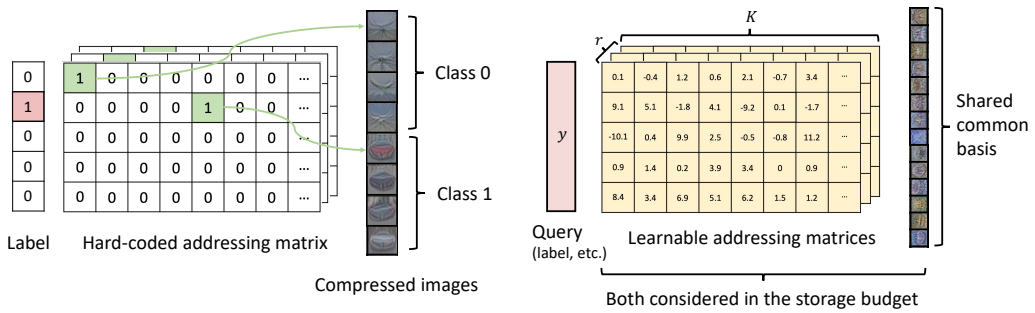

Figure 1: Distill a large-scale dataset into compressed memories. Left: the standard dataset distillation task under the formulation of memory addressing. Right: learnable addressing matrix with shared common bases. The queries can be generalized to any vector representation, besides one-hot labels, i.e. for a general dataset from $\mathcal{X}$ to $\mathcal{Y}$, it can be distilled into memories for recall and model re-training.

Under this definition, since we are learning a single, shared and accessible representation for all $\boldsymbol{y}$s, the size of memories can be defined flexibly regardless of the number of classes, removing the linear growth limitation in the standard distillation settings. There is also no constraint on the form of $\boldsymbol{y}$s, which can be either discrete or continuous. The *storage budget* or *compression rate* is calculated by considering the storage space of parameters in both memories and addressing functions, which should be as compact as possible.

Memory representation. We use a set of basis to store in the memories $\mathcal{M} = \{\boldsymbol{b}_0, ..., \boldsymbol{b}_{K-1}\}$, where each vector $\boldsymbol{b}_k \in \mathbb{R}^d$ has the same dimension as $\boldsymbol{x} \in \mathbb{R}^d$, and all vectors collectively define the intrinsic components in a dataset that characterize the task mapping from $\mathcal{X}$ to $\mathcal{Y}$. Through re-using the basis, we can produce a desired synthetic dataset for model re-training. *Spatial redundancies in images:* note that, as a special case, images can contain redundancies spatially and be stored in a downsampled version to improve the compression rate. The downsampled image basis can be passed via a deterministic upsampling process (e.g., bilinear interpolation) to recover the original resolution.

Memory addressing. For each query $\boldsymbol{y}$, we use a parameterized function $\mathcal{A}(\boldsymbol{y})$ to re-combine the basis in the memories $\mathcal{M}$. Similar to previous methods on accessing memories, we use $\mathcal{A}$ to produce a set of coefficients, and linearly combine the basis to produce synthetic data. Formally, to retrieve r synthetic data for each $\boldsymbol{y}$, we define a set of matrices $\{\boldsymbol{A}_0, ..., \boldsymbol{A}_{r-1}\}$, $\boldsymbol{A}_i \in \mathbb{R}^{d_y \times K}$, where d_y is the dimension size of $\boldsymbol{y}$ and r is the number of data samples that can be retrieved. With the memories $\mathcal{M} = \{\boldsymbol{b}_0, ..., \boldsymbol{b}_{K-1}\}$, we define:

$$\boldsymbol{x}'_i = \boldsymbol{y}^T \boldsymbol{A}_i [\boldsymbol{b}_0; ...; \boldsymbol{b}_{K-1}]^T, \boldsymbol{x}' \in \mathbb{R}^d \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^{d_y \times 1}$ is in a vectorized form, such one-hot encoding of categorical labels, and $\boldsymbol{v} = \boldsymbol{y}^T \boldsymbol{A}_i$ corresponds to a coefficient vector $\boldsymbol{v}$ that combines the basis. The produced synthetic data $\boldsymbol{x}'$ is paired with $\boldsymbol{y}$ as the corresponding label. Our model is shown on the right of figure 1.

Constructing a dataset. To construct a synthetic dataset $\mathcal{D}_s$ for model re-training, we are often given a set of samples $\{\boldsymbol{y}_i\}$, or can enumerate all possible values of $\boldsymbol{y}$ (if discrete). With the set $\{\boldsymbol{y}_i\}$, we use eqn.2 to address and retrieve synthetic dataset $\mathcal{D}_s^{\boldsymbol{y}=\boldsymbol{y}_i} = \{(\boldsymbol{x}'_j, \boldsymbol{y}_i)\}_{j=1}^r$ for each $\boldsymbol{y}_i$. The final dataset is the union of all $\mathcal{D}_s = \bigcup \mathcal{D}_s^{\boldsymbol{y}=\boldsymbol{y}_i}$. The dataset $\mathcal{D}_s$ can be used in either minibatch forms for stochastic gradient descent, or as a whole for batch gradient descent.

Generalized possibilities of queries. Another advantage of our formulation, besides a compact and shared representation, is the various possibilities of queries $\boldsymbol{y}$. In principle, under this formulation, the label $\boldsymbol{y}$ can be flexibly defined as other forms, such as language, audio, of which the representation resides in continuous space or follow a distribution $p(h(\boldsymbol{y}))$ defined by a feature extractor $h(\cdot)$. The set of $\{\boldsymbol{y}_i\}$ then can be sampled from the distribution p instead of having to enumerate all possible values. This provides a general way of compressing or distilling a large dataset without constraint on the forms of labels.

Connection with standard Dataset Distillation. In the standard setting, the task is defined for classification tasks with discrete labels and each label owns its unique set of synthetic data. We can

Algorithm 1

hyperparameters: Momentum rate $\beta_{\boldsymbol{\theta}}, \beta_{\boldsymbol{\phi}}$, learning rate $\alpha_{\boldsymbol{\theta}}, \alpha_{\boldsymbol{\phi}}$ for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ respectively.
input: Dataset $\mathcal{D}_{tr}$, memories $\mathcal{M}$, addressing function $\mathcal{A}$, loss function $\ell(\cdot,\cdot)$
repeat
  Sample a subset of labels $\mathcal{Y}'$
  Address the memories $\mathcal{M}$ and obtain synthetic dataset $\mathcal{D}_s^{\mathcal{Y}'}$ with eqn. 2
  Randomly initialize model parameters $\boldsymbol{\theta}_0$
  Initialize momentum $\boldsymbol{m}_0 = 0$
  for $t = 1$ **to** T **do**
    Sample a minibatch $\boldsymbol{B}_s = \{(\boldsymbol{x}'_i, \boldsymbol{y}_i)\}$ from $\mathcal{D}_s^{\mathcal{Y}'}$
    Compute $\mathcal{L} = \frac{1}{|\boldsymbol{B}_s|}\sum_{i=1}^{|\boldsymbol{B}_s|} \ell(f_{\boldsymbol{\theta}_{t-1}}(\boldsymbol{x}'_i), y_i)$
    Update momentum $\boldsymbol{m}_t = \beta_{\boldsymbol{\theta}}\boldsymbol{m}_{t-1} + \frac{d\mathcal{L}}{d\boldsymbol{\theta}_{t-1}}$
    Update $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha_{\boldsymbol{\theta}}\boldsymbol{m}_t$
  end for
  Sample a minibatch $B = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ from $\mathcal{D}_{tr}$ with labels in $\mathcal{Y}'$
  Compute $J(\boldsymbol{\phi}) = \frac{1}{|\boldsymbol{B}|}\sum_{i=1}^{|\boldsymbol{B}|} \ell(f_{\boldsymbol{\theta}_T}(\boldsymbol{x}_i), \boldsymbol{y}_i)$ Update $\boldsymbol{\phi} \leftarrow$ opt-alg$(J(\boldsymbol{\phi}), \alpha_{\boldsymbol{\phi}}, \beta_{\boldsymbol{\phi}})$
until Converge

show that this is a special case of our formulation: if the basis is constructed as the collection of those label-specific synthetic data ($K = N'$), and the addressing matrix $\boldsymbol{A}_i$ is defined in space $\{0,1\}^{C\times N'}$ with hardcoded 1s at position $c+i$ of each row c, the "retrieval" process can also be defined as eqn. 2.

4.2 Learning framework: back-propagation through time

In this section, we build upon the back-propagation through time algorithm and discuss in detail the learning framework that performs the distillation process from a dataset to memories and addressing functions. Specifically, we identify several critical aspects in the algorithm that can boost the performance of the task.

Starting from notations, we define the parameters contained in both memories and addressing functions as $\boldsymbol{\phi}$, which are collectively optimized. A loss function is $\ell(\cdot,\cdot)$ is defined on the task-specific dataset $\mathcal{D}_{tr}$. We denote an optimization algorithm as opt$(\cdot,\cdot;\alpha,\beta)$, where α and β are the learning rate and the momentum rate, respectively.

To learn the parameters $\boldsymbol{\phi} = \{\mathcal{M}, \mathcal{A}\}$, we follow a standard bi-level optimization framework with back-propagation through time (BPTT), where the inner-loop uses the synthetic dataset $\mathcal{D}_s$ to train a randomly initialized model starting from scratch, and a generalization loss is computed using a minibatch $\boldsymbol{B} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ sampled from $\mathcal{D}_{tr}$:

$$\begin{aligned} \min J(\boldsymbol{\phi}) &= \frac{1}{|\boldsymbol{B}|}\sum_{i=1}^{|\boldsymbol{B}|} \ell(f(\boldsymbol{x}_i;\boldsymbol{\theta}^*), \boldsymbol{y}_i), \\ \text{subject to} \quad \boldsymbol{\theta}^* &= \text{opt}\big(\boldsymbol{\theta}_0, \mathcal{D}_s; \alpha_{\boldsymbol{\theta}}, \beta_{\boldsymbol{\theta}}\big) \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_0$ represents the initializing parameters and $\boldsymbol{\theta}^*$ is the optimized model parameters in the inner optimization loop. In practice, we use a subsampled version $\mathcal{Y}'$ of all $\boldsymbol{y}$s and retrieve the corresponding subset $\mathcal{D}_s^{\mathcal{Y}'}$. This reduces the computation cost in inner loops. We empirically observe that equivalent results can be achieved with faster speed. The algorithm is summarized in Alg. 1.

Critical factors in BPTT. Although a natural choice in performing dataset distillation, the BPTT framework has shown to underperform other algorithms, such as gradient matching methods, in various works and benchmarks. In our work, we empirically identify several aspects that can drastically improve the performance.

Long unrolled trajectories. In previous data distillation works [6], the number of inner loop optimization steps is set to relative small values, potentially due to the common practice in meta-learning and few-shot learning works [28, 18, 41]. Instead we find that unrolling the trajectories long enough (e.g. 200 steps) can potentially produce $\boldsymbol{\theta}^*$ that better summarizes the information contained in memories and produce more effective gradients.

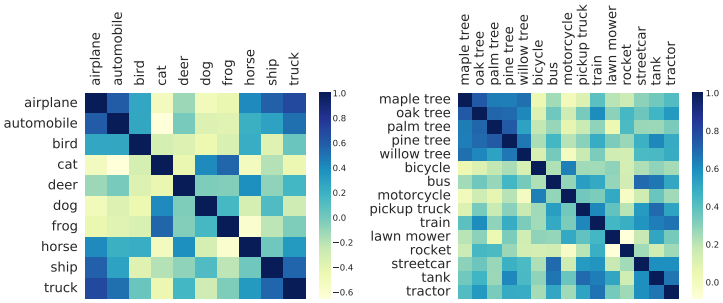

Figure 2: Similarity matrices of learned addressing coefficients for the CIFAR10 dataset (left) and a subset of CIFAR100 classes (right).

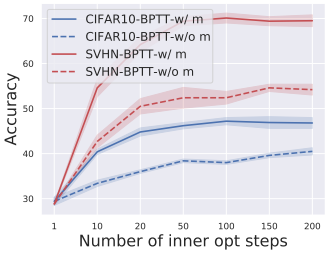

Figure 3: Analysis on BPTT steps and momentums.

Momentum term. Another aspect ignored in dataset distillation and not commonly used in meta-learning is the momentum term. Indeed, adding momentum term in meta-learning can potentially even hurt the performance and lead to less gradient diversity [41]. However, we show that, in dataset distillation, momentum term is critical, even in short inner loop optimization settings. We provide more detailed analysis in the experiment section.

5 Experiments

We thoroughly evaluate our model under various tasks and benchmarks, and demonstrate the benefits of our model over previous methods. First, we show that using a shared representation is core to improve the distillation performance and compression rates. Specifically, we observe that there is strong evidence that there is information re-using across classes. Moreover, we show that the shared representation enables more general queries, which can be continuous. We further show the benefits of our model on continual learning and synthesizing new classifiers. For example, we observe that a simple compress-then-recall method can achieve performance outperforming state-of-the-art continual learning models with complex designs.

5.1 Dataset Distillation

In this section, we follow the standard setting of dataset distillation, and compress a dataset into memories that can be recalled with labels.

Datasets. We test our models on five standard dataset distillation benchmarks: MNIST [42], Fashion-MNIST [43], SVHN [44], CIFAR10 [45], CIFAR100 [45]. MNIST contains 10 classes with 60,000 writing digit images as training and 10,000 as test set. The images are gray-scale with a shape of 28×28. FashionMNIST is a dataset with clothing and shoe images and consists of a training with size 60,000 and a test set with size 10,000. Each image is 28×28 in gray scale, and has a label from 10 classes. SVHN contains street digit images where each image has shape $32 \times 32 \times 3$. The dataset contains 73257 images for training and 26032 images for testing. CIFAR10 and CIFAR100 are color image datasets, with 50,000 training images and 10,000 testing images on each. CIFAR10 has 10 classes with 5,000 images per class, and CIFAR100 has 100 classes with 500 images per class.

Experiment settings. We evaluate our distillations model under three different memory budgets for each dataset: 1/10/50 images per class. We focus on high compression rate scenarios and consider 1 and 10 settings for CIFAR100. Following the previous works [14, 11, 12, 15], the main network architecture used in experiments is a simple convolutional network (ConvNet) with 3×3 filters, InstanceNorm, ReLU and 2×2 average pooling. For the evaluation protocol, each model is evaluated on 20 randomly initialized models, trained for 300 epochs on a synthetic dataset, and tested on a held-out testing dataset. We use one GPU per experiment run.

Memory budget calculation. Since our model uses memories and addressing matrices to store the compressed information, we treat the total number of images as a memory storage budget:

$$\text{size(basis)} + \text{size(addressing matrices)} \approx \text{size(images)} \tag{4}$$

where images is the total number of images for all classes and $\text{size}(\cdot)$ is the total size of a tensor. We assume all tensors are in the format of float numbers.

	I/C	DC [10]	DSA [11]	KIP [12] (NN)	CAFE* [46]	TM [15]	DM [14]	Ours
MNIST [42]	1	91.7±0.5	88.7±0.6	90.1±0.1	93.1±0.3	-	89.7±0.6	**98.7±0.7**
	10	97.4±0.2	97.8±0.1	97.5±0.0	97.5±0.1	-	97.5±0.1	**99.3±0.5**
	50	98.8±0.2	99.2±0.1	98.3±0.1	98.9±0.2	-	98.6±0.1	**99.4±0.4**
F-MNIST [43]	1	70.5±0.6	70.6±0.6	73.5±0.5	77.1±0.9	-	-	**88.5±0.1**
	10	82.3±0.4	84.6±0.3	86.8±0.1	83.0±0.3	-	-	**90.0±0.7**
	50	83.6±0.4	88.7±0.2	88.0±0.1	88.2±0.3	-	-	**91.2±0.3**
SVHN [44]	1	31.2±1.4	27.5±1.4	57.3±0.1	42.9±3.0	-	-	**87.3±0.1**
	10	76.1±0.6	79.2±0.5	75.0±0.1	77.9±0.6	-	-	**89.1±0.2**
	50	82.3±0.3	84.4±0.4	80.5±0.1	82.3±0.4	-	-	**89.5±0.2**
CIFAR10 [45]	1	28.3±0.5	28.8±0.7	49.9±0.2	31.6±0.8	46.3±0.8	26.0±0.8	**66.4±0.4**
	10	44.9±0.5	52.1±0.5	62.7±0.3	50.9±0.5	65.3±0.7	48.9±0.6	**71.2±0.4**
	50	53.9±0.5	60.6±0.5	68.6±0.2	62.3±0.4	71.6±0.2	63.0±0.4	**73.6±0.5**
CIFAR100 [45]	1	12.8±0.3	13.9±0.3	15.7±0.2	14.0±0.3	24.3±0.3	11.4±0.3	**34.0±0.4**
	10	25.2±0.3	32.3±0.3	28.3±0.1	31.5±0.2	40.1±0.4	29.7±0.3	**42.9±0.7**

Table 1: We compare our method with previous works on ConvNet performance recovery. Our algorithm consistently outperforms all previous methods and achieves state-of-the-art. *Note that we selected the best results from baseline model variants. I/C is images per class (storage budget eqn. 4).

	I/C	Single-step GM	Ours^{BPTT}	Ours$^{\text{BPTT+ds}}$	Ours$^{\text{Full w/o Aug.}}$	Ours^{Full}
CIFAR10	1	28.8±0.7	49.1±0.6	55.2±0.5	64.2±0.6	**66.4±0.4**
	10	52.1±0.5	62.4±0.4	65.9±0.4	70.9±0.4	**71.2±0.4**
	50	60.6±0.5	70.5±0.4	71.1±0.5	72.1±0.5	**73.8±0.4**
CIFAR100	1	13.9±0.3	21.3±0.6	25.9±0.4	33.5±0.2	**34.0±0.4**
	10	32.3±0.3	34.7±0.5	36.5±0.4	40.6±0.3	**42.9±0.7**

Table 2: Ablation studies of every component and comparison with single-step gradient matching [10].

Model details. We use basis with size downsampled by a factor of 2 from the standard image size based on datasets. All models are trained for 50k iterations with SGD optimizer. For the inner loop optimization, we set the momentum rate as 0.9, and use 150 steps for small memory budgets 1 and 10, and 200 steps for budget 50. The selection of number of basis to use is based on a held-out validation set (10% of training set), shown in figure 4. Number of addressing matrices is calculated with eqn. 4. More details in appendix.

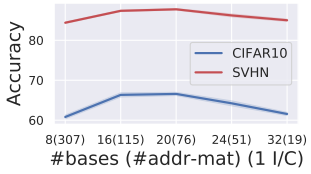

Figure 4: Validation set

Result 1: state-of-the-art accuracy. We compare our model with previous methods: Dataset Condensation (DC) [10], Differentiable Siamese Augmentation (DSA) [11], Kernel Inducing Points (KIP) [12], Distribution Matching (DM) [14], Aligning Features (CAFE) [46] and Trajectory Matching (TM) [15]. The results are summarized in table 1. Following previous methods [11, 12, 15], we adopt simple data augmentations, flip and rotation, on CIFAR10 and CIFAR100 datasets, and use ZCA on SVHN, CIFAR10 and CIFAR100. As shown in the table, our model consistently outperforms previous methods, especially under high compression rate cases where 1 image is allowed per class.

Analysis: information sharing across classes. The core observation in our method is that a common representation can enable information sharing across classes and reduce redundancies. To verify this, we calculate the average coefficients $\bar{\boldsymbol{v}} = \frac{1}{r}\sum_{i=0}^{r-1} \boldsymbol{y}^T \boldsymbol{A}_i$ for each class $\boldsymbol{y}$ and visualize the cosine similarities of $\bar{\boldsymbol{v}}$ from two classes. The visualizations are shown in fig. 2. Higher cosine similarity scores indicate that two classes are utilizing similar basis components in the memories to produce synthetic images. For example, classes maple, oak, palm, pine and willow trees have strong sharing, while lawn mower and rocket are distinct from each other. Similar patterns can be found in CIFAR10.

Result 2: back-propagation through time is a strong baseline. In figure 3 and table 2, we show that BPTT is already a strong baseline that can outperform previous single-step gradient methods by up to 40% (on SVHN 1 image per class). Results shown in the figure are obtained through standard synthetic dataset organization without downsampling and memory-addressing framework. This contrasts to previous observation that single-step gradient matching methods always outperform BPTT with generalization loss [10, 11, 6]. The effects of momentum and long unrolls in BPTT are shown in figure 3. It is interesting to observe that even for shorter unrolls, momentum term is still important and can significantly improve the performance.

Ablation studies. To analyze the effects of each component, we perform ablation studies on CIFAR10 and CIFAR100, summarized in table 2. We show that the observations on BPTT, spatial redundancies, and shared representations can all improve the results and compression rate. See more in the appendix.

5.2 Continual Learning

One of the key usage of memories is to prevent forgetting when a model continually learns through tasks. In this section, we evaluate our algorithm on the standard continual learning benchmarks and show that, due to the strong performance, a simple "compress-then-recall" method with our model can already rival with previous state-of-the-arts with complex designs.

Efficient lifelong learning. Following [47], we work with the problem where all tasks are streamed in mini-batches and learned in a *single pass*. A learner is allowed to be equipped with a small memory buffer. The data samples after seen will not be available unless stored in the buffer. We use mini-batch size 10 to stream the data, following previous works [28, 24].

Evaluation. The learner's performance after learning on the task stream is evaluated under two metrics: retained accuracy (RA) and backward-transfer and interference (BTI). RA is the average accuracy of the final trained model on all tasks, and BTI measures the performance difference between after it was learned and after the full training process.

Benchmarks. We evaluate our method on three tasks widely used in previous Continual Learning works. MNIST Rotations [24] contains 20 tasks with 1,000 samples in each. Every task consists of images rotated by a fixed angle from 0 to 180 degrees. MNIST Permutations [48] has 20 tasks, and each task contains 1,000 images generated through shuffling the image pixels by a fixed permutation. Many Permutations [17] is a longer variant with 100 tasks in total and 200 samples in each. Incremental CIFAR-100 [29, 24] splits the CIFAR100 dataset into 20 5-way classification tasks for learning.

Our model. Based on our distillation method, we adopt a simple framework to perform continual learning: "compress then recall". During the training phase, we do not perform learning on neural networks, instead, the dataset of each task is distilled to memories and the paired addressing matrices. During test phase, we simply fetch the corresponding memories and addressing matrices for each task, and train a new model from scratch to perform classification. *Memory buffer designs.* When a new task starts, we use the full remaining memory buffer to store the samples and perform distillation with both buffer samples and streamed samples. After a task ends, the distilled memories and addressing matrices are stored in the buffer, taking $1/T$ of the space. Namely, the buffer size keeps shrinking when more compressed representation of tasks is stored. Note that we compare our model with previous methods under the *exact same memory sizes for fair comparisons*. See appendix.

Results. We show that this simple method is already a strong baseline that outperforms prior arts on four benchmarks, summarized in table 3. Our method is compared with: Online, EWC [48], GEM [24], MER [17], C-MAML [28], La-MAML [28] and Sparse-LaMAML [30]. For example, we can obtain a 23% boost on MNIST MANY benchmark.

We further compare our model with previous works Kernel Continual Learning [39] and Stable SGD [37] following their settings. We list the results here due to space limit: our model achieves $87.34^{\pm 0.92}$ and $88.25^{\pm 0.58}$ on Permutated MNIST and Rotated MNIST with 60,000 samples for each

	Rotations		Permutations		MANY		CIFAR-100	
	RA↑	BTI↓	RA↑	BTI↓	RA↑	BTI↓	RA↑	BTI↓
ONLINE	$53.38^{\pm 1.53}$	$-5.44^{\pm 1.70}$	$55.42^{\pm 0.65}$	$-13.76^{\pm 1.19}$	$32.62^{\pm 0.43}$	$-19.06^{\pm 0.86}$	$32.62^{\pm 0.43}$	$-19.06^{\pm 0.86}$
EWC [48]	$57.96^{\pm 1.33}$	$-20.42^{\pm 1.60}$	$62.32^{\pm 1.34}$	$-13.32^{\pm 2.24}$	$33.10^{\pm 0.14}$	$-18.50^{\pm 0.91}$	-	-
GEM [24]	$67.38^{\pm 1.75}$	$-18.02^{\pm 1.99}$	$55.42^{\pm 1.10}$	$-24.42^{\pm 1.10}$	$39.50^{\pm 0.62}$	$-17.50^{\pm 0.41}$	$48.27^{\pm 1.10}$	$-13.7^{\pm 0.70}$
MER [17]	$77.42^{\pm 0.78}$	$-5.60^{\pm 0.70}$	$73.46^{\pm 0.45}$	$-9.96^{\pm 0.45}$	$51.00^{\pm 0.54}$	$-13.57^{\pm 0.45}$	$51.38^{\pm 1.05}$	$-12.83^{\pm 1.44}$
La-M [28]	$77.42^{\pm 0.65}$	$-8.64^{\pm 0.40}$	$74.34^{\pm 0.67}$	$-7.60^{\pm 0.51}$	$50.43^{\pm 0.21}$	$-10.00^{\pm 0.36}$	$61.18^{\pm 1.44}$	$-9.00^{\pm 0.2}$
sp-La [30]	$77.77^{\pm 0.58}$	$-8.16^{\pm 0.61}$	$76.88^{\pm 0.72}$	$-8.39^{\pm 0.63}$	$50.81^{\pm 0.79}$	$-13.73^{\pm 0.73}$	-	-
Ours	$\mathbf{80.32^{\pm 0.28}}$	-	$\mathbf{78.48^{\pm 0.76}}$	-	$\mathbf{74.07^{\pm 0.51}}$	-	$\mathbf{62.58^{\pm 1.1}}$	-

Table 3: We show that "compress-then-recall" is a strong baseline that outperforms previous methods on four continual learning benchmarks. Baseline numbers are from [28] or obtained from public official repos.

task, outperforming both KCL and Stable SGD. Interestingly, our results also are higher than the multitask upperbound ($86.5^{\pm 0.21}$ and $87.3^{\pm 0.47}$), potentially due to that there are task interference in joint training, which can be naturally avoided in our method.

5.3 Synthesizing new classifiers after learning

If we want to memorize the past, what is the benefit of storing the compressed representation rather than a trained model? In this section, we show that our compressed representation can enable flexible synthesis on new classifiers after the learning. Specifically, we demonstrate on extrapolating between tasks to train new models, and performing memory recall with images instead of labels, showing the generalizability of our framework on other query forms.

5.3.1 Extrapolating between tasks

In real world, tasks often do not come together and a learner therefore cannot observe all tasks at once. In current machine learning paradigms, when models are separately trained for disjoint tasks, it has difficulty extrapolating between tasks to build new classifiers. This is different from human learning. We show that, if storing our compressed representation, it enables a learner to extrapolate and synthesize new classifiers after learning separately on each task. Specifically, we separate CIFAR100 into 20 disjoint 5-way classification tasks. Independently, we distill the 20 datasets into corresponding memories and test the possibility of generating new classifiers on new class combinations unseen during training. We find that the compressed data can indeed train classifiers on new combinations, for example, we can achieve 72.53% on 2-way classification, and 46.54% on 5-way classification, with 1 image per class storage budget. The upperbound with full real dataset is 92.23% and 82.72%.

5.3.2 Dataset Distillation extension – recall the past with images

We extend the standard setting to recall the past with images: when the label information and task scopes are missing, but a few visual observations can be made, we would like to build classifiers based on the visual data. For example, when we see a bear image and a deer image, but cannot recall the exact word or category, can we recall the memories with images and build a classifier? This is possible with our problem formulation, where the forms of queries are not constraint to labels and we can *distill a dataset to memories addressable by images*.

Formally, after observing a training dataset $\mathcal{D}_{tr}$ with $\mathcal{Y} = \{0, ..., C-1\}$, we would like to flexibly build classifiers for a subtask $\mathcal{Y}_g \subset \mathcal{Y}$ based on visual observations $\mathcal{X}_g$ from $\mathcal{Y}_g$, when the actual information of $\mathcal{Y}_g$ is unknown. We work with 1-shot and 5-shot observation cases. As a baseline, we build a nearest neighbor classifier, which is pretrained on $\mathcal{D}_{tr}$ and takes features of few-shot data to classify test images. As a model variant, we could also "classify-then-recall", using a classifier trained on $\mathcal{D}_{tr}$ to map the image shots into labels and turning into the standard setup. Benefiting from the general design of our formulation, we show that a model can directly perform "image addressing", where a feature network can provide query vectors $\boldsymbol{y}$ in fig.1. The feature network, memories and addressing matrices can be jointly trained on $\mathcal{D}_{tr}$.

Methods	1 shot	5 shot
Nearest neighbor	48.55	61.72
classify-then-recall	50.58	58.46
image addressing	**55.74**	**71.20**

Table 4: Few-shot perf. recovery.

6 Conclusion and limitations

In this paper, we propose a framework that distills a large dataset into compact addressable memories. This framework introduce several benefits, including removing the linear growth contraint on the compressed data size, allowing more general queries beside categorical labels, and most importantly, high compression rate with strong re-training performance which outperforms previous state-of-the-arts. We also demonstrate a "compress-then-recall" method using our framework, leading to new state-of-the-arts on four datasets. Our full model has potential limitation on the costly inner optimization loop, which might be time consuming on gigantic models or on much larger datasets. This limitation might be solved through combining the memory formulation with a different learning framework. We do not observe obvious negative societal impacts in our proposed method.

References

[1] Timothy F Brady, Talia Konkle, and George A Alvarez. Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, 138(4):487, 2009.

[2] Geoffrey R Loftus and Elizabeth F Loftus. *Human memory: The processing of information*. Psychology Press, 2019.

[3] John R Anderson and Gordon H Bower. *Human associative memory*. Psychology press, 2014.

[4] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[5] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[6] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[8] Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. Eigengame: Pca as a nash equilibrium. In *International Conference on Learning Representations*, 2020.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[10] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021.

[11] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, 2021.

[12] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *International Conference on Learning Representations*, 2021.

[14] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. *arXiv preprint arXiv:2110.04181*, 2021.

[15] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. *arXiv preprint arXiv:2203.11932*, 2022.

[16] Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. *arXiv preprint arXiv:1910.02551*, 2019.

[17] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[19] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.

[20] Aniruddh Raghu, Maithra Raghu, Simon Kornblith, David Duvenaud, and Geoffrey Hinton. Teaching with commentaries. In *International Conference on Learning Representations*, 2020.

[21] Sid Reddy, Anca Dragan, and Sergey Levine. Pragmatic image compression for human-in-the-loop decision-making. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, and Stefano Ermon. Comparing distributions by measuring differences that affect decision making. In *International Conference on Learning Representations*, 2021.

[23] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

[24] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30:6467–6476, 2017.

[25] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[26] Tom Veniat, Ludovic Denoyer, and Marc'Aurelio Ranzato. Efficient continual learning with modular networks and task-driven priors. *arXiv preprint arXiv:2012.12631*, 2020.

[27] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

[28] Gunshi Gupta, Karmesh Yadav, and Liam Paull. La-maml: Look-ahead meta learning for continual learning. *arXiv preprint arXiv:2007.13904*, 2020.

[29] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[30] Johannes Von Oswald, Dominic Zhao, Seijin Kobayashi, Simon Schug, Massimo Caccia, Nicolas Zucchet, and João Sacramento. Learning where to learn: Gradient sparsity in meta and continual learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pages 524–540. Springer, 2020.

[32] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2(6), 2019.

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

[34] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[35] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018.

[36] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in psychology*, 4:504, 2013.

[37] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. *Advances in Neural Information Processing Systems*, 33:7308–7320, 2020.

[38] Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2020.

[39] Mohammad Mahdi Derakhshani, Xiantong Zhen, Ling Shao, and Cees Snoek. Kernel continual learning. In *International Conference on Machine Learning*, pages 2621–2631. PMLR, 2021.

[40] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.

[41] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

[42] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[43] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[44] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[45] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[46] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe learning to condense dataset by aligning features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022*, 2022.

[47] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

[48] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Details are in main paper and more in the supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] More details are in the supplementary materials.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]