Retaining Knowledge for Learning with Dynamic Definition

Zichang Liu  
Rice University  
Houston, TX 77025  
zichangliu@rice.edu

Benjamin Coleman  
Rice University  
Houston, TX 77025  
brc7@rice.edu

Tianyi Zhang  
Rice University  
Houston, TX 77025  
tz21@rice.edu

Anshumali Shrivastava  
Rice University/ThirdAI Corp.  
Houston, TX 77025  
anshumali@rice.edu

Abstract

Machine learning models are often deployed in settings where they must be constantly updated in response to the changes in class definitions while retaining high accuracy on previously learned definitions. A classical use case is fraud detection, where new fraud schemes come one after another. While such an update can be accomplished by re-training on the complete data, the process is inefficient and prevents real-time and on-device learning. On the other hand, efficient methods that incrementally learn from new data often result in the forgetting of previously-learned knowledge. We define this problem as Learning with Dynamic Definition (LDD) and demonstrate that popular models, such as the Vision Transformer and Roberta, exhibit substantial forgetting of past definitions. We present a first practical and provable solution to LDD. Our proposal is a hash-based sparsity model RIDDLE that solves evolving definitions by associating samples only to relevant parameters. We prove that our model is a universal function approximator and theoretically bounds the knowledge lost during the update process. On practical tasks with evolving class definition in vision and natural language processing, RIDDLE outperforms baselines by up to 30% on the original dataset while providing competitive accuracy on the update dataset.

1 Introduction

Motivation: A common machine learning pipeline is to first define a task objective based on historical experience, then deploy a model trained to optimize the objective over collected data. However, machine learning models rarely operate in such a static environment. Instead, models must be dynamically updated to accommodate changes in class definitions without forgetting previous knowledge. Consider the case of fraud detection, where models must adapt to emerging scam patterns while also providing protection against previously-known patterns. Machine learning models are also used to identify harmful microbes given gene sequences, where it is important to perform well on new variants or newly discovered microorganisms. A similar situation arises when models are used to filter inappropriate content such as hate speech and violence. New phrases or previously-innocuous language may become inappropriate due to world events or attempts by users to circumvent the filter. While the classification task remains the same (fraud or safe, harmful or not, inappropriate or appropriate, etc.), class definitions evolve over time and likely differ significantly from the point of the initial training. We refer to this setting as Learning under Dynamic Definition (LDD). It should be
noted that this notion is very different for several other notions of distribution shift in the literature, which we review in the next section.

A standard solution is to retrain the model on all available data (original dataset and data for the new definitions). This methodology is inefficient due to data storage and extra computation, and even prohibitive when dealing with large models and streaming data. Furthermore, it prevents real-time and on-device learning. An ideal and efficient solution is to update the trained model incrementally only with data for the new definitions (e.g., using online gradient descent). However, this is a challenging process due to the tendency of machine learning models, in particular neural networks, to forget previous knowledge. This situation, known as collapse/catastrophic forgetting [25], causes performance decay on the old definitions.

Our Proposal: Catastrophic forgetting occurs because gradient updates from training examples in the update process are able to change parameters that are tuned for optimal performance on the original dataset [30]. Inspired by recent connections between random partitions and distribution models [19], we propose RIDDLE (Retaining Information in a Dynamically Defined Learning Environment), a novel architecture with meaningful parameter groupings. RIDDLE works by partitioning the input space and assigning a set of parameters to each partition. Gradient updates are restricted only to locally relevant parameters, identified by cheap hashing computations. RIDDLE can be viewed as a Mixture-of-Expert-Layer where the gating function is Locality Sensitive Function (LSH) and each expert is just one number. At the same time, RIDDLE can be viewed as a kernel method where the prediction is an efficient weighted kernel density estimation. RIDDLE can be plugged into different neural network architectures, just like fully connect layers, to ease the memory loss while utilizing the learned features, as shown in Figure 2.

We summarize our technical contributions below.

- We introduce the RIDDLE (in Section 3), consisting of a set of parameters indexed by locality-sensitive hash functions. RIDDLE is computationally efficient and requires less memory than popular machine learning models with similar accuracy.
- We prove that RIDDLE is universal function approximator and can therefore represent any function (Section 4). We prove a formal connection between parameter usage and distribution similarity in Theorem 3.2.
- In Section 5, we apply RIDDLE in the dynamic definition setting for four real-world tasks. Our method outperforms baselines by up to 30% on the original dataset while achieving competitive accuracy on the new dataset.

2 Related Work

Continual Learning: Continual learning refers to learning and remembering multiple tasks from a stream of information. Current approaches can be split into three main categories (1) memory replay to complement neural networks [45, 29], (2) constraints on network updates [48, 28, 52], and (3) dynamic architectural adjustments to accommodate new information, such as additional layers [40, 61]. Recently, few-shot learning methods are also applied into continue learning to solve sequential learning with limited data [8, 58, 53]. Continual learning methods encounter similar challenges to the ones presented by learning under dynamic distribution, most notably catastrophic forgetting. However, continual learning attempts to solve a sequence of related tasks, often expressed as task embeddings, over data that does not necessarily reflect changes in class definitions.

Train-Test Distribution Shift: Transfer learning and test-time adaptation attempt to remedy mis-alignments between the training and test distribution. Although such methods may seem relevant in learning with dynamic definition, they have dramatically different goals. Transfer learning attempts to
adapt a model from a data-rich source domain to a data-scarce target domain, without regard for the performance on the source domain \[55\]. While several works on train-test distribution shift focus on robustness, it is from the perspective of out-of-distribution detection \[24\] or uncertainty estimation \[43\] \[33\] \[57\]. Our objective is to provide models that perform well on both the original dataset and new dataset, where class definitions are different, rather than simply detect or adapt to the distribution change.

**Online Learning:** Online learning regards optimization as a process where a model is trained by interacting with a stream of data. Online learning algorithms attempt to minimize the regret, or difference between the loss of the predictions by the incrementally-trained model and the loss of the globally optimal model \[43\] \[46\]. Online learning methods attain minimal regret against adaptive adversaries and could be expected to perform well in the dynamic distribution setting. However, the best performing methods for non-convex online learning require retraining over all previous data at every time step \[55\], making them practically unsuitable for the tasks we consider.

**Locality-Sensitive Hashing in Machine Learning:** Locality-sensitive hash (LSH) functions are a central component of our proposal. An LSH family is a set of functions that maps similar inputs to the same hash value (with high probability). LSH is an incredibly powerful and well-studied framework, with the following formal definition \[27\] \[21\] \[20\].

**Definition 2.1 ((S_0, cS_0, p_1, p_2)-sensitive hash family).** A family \( \mathcal{H} \) is called \((S_0, cS_0, p_1, p_2)\)-sensitive with respect to a similarity function \(\text{sim}(\cdot, \cdot)\) if for any two points \(x, y \in \mathbb{R}^D\) and \(h\) chosen uniformly from \( \mathcal{H} \) satisfies:

- If \(\text{sim}(x, y) \geq S_0\) then \(Pr[h(x) = h(y)] \geq p_1\)
- If \(\text{sim}(x, y) \leq cS_0\) then \(Pr[h(x) = h(y)] \leq p_2\)

The probability \(Pr[h(x) = h(y)]\) is known as the collision probability of \(x\) and \(y\), which we use the term “collision” to mean that two points map to the same hash value. For the purpose of our arguments, we use a stronger notion of LSH, in which the collision probability is a monotonic increasing function of the similarity between \(x\) and \(y\). That is \(Pr[h(x) = h(y)] \propto f(\text{sim}(x, y))\).

Under the above conditions, the collision probability \(Pr[h(x) = h(y)]\) forms a positive semi-definite radial kernel \[19\]. A number of useful kernels can be obtained via LSH, including MinHash for the Jaccard kernel \[12\], signed random projections for the angular kernel \[13\], and p-stable LSH for a kernel in Euclidean space \[20\].

Recent works have shown that LSH can be used for efficient statistical estimation \[34\] \[41\] \[19\] \[33\] \[35\] \[37\] \[62\], compressive machine learning \[44\], nearest neighbor search \[34\] \[41\] \[33\] and to accelerate neural network training and inference \[17\] \[16\] \[32\] \[40\] \[43\]. A particularly relevant result is the RACE sketch \[19\], which solves kernel density estimation using the number of examples that fall into the random partitions induced by an LSH function.

### 3 RIDDLE for Learning with Dynamic Definition

**Learning with Dynamic Definition:** We formally define the problem as follows.

**Definition 3.1.** Given a model \(f(x; \theta)\) and a loss \(\ell(f(x; \theta), y)\), consider an original dataset \(D^o = \{(x^o_i, y^o_i)\}_{i=1}^m\) drawn i.i.d. from a distribution \(P^o\) and an update dataset \(D^u = \{(x^u_i, y^u_i)\}_{i=0}^m\) drawn i.i.d from distribution \(P^u\). Let \(\theta^o\) be the parameters that optimize \(\mathcal{L}_o(\theta) = \sum_{D^o} \ell(f(x; \theta), y)\). The dynamic definition learning problem is to find a solution minimizing \(\mathcal{L}_{o+u}(\theta)\) when we are only given access to \(\theta^o\) and \(D^u\).

\[
\mathcal{L}_{o+u}(\theta) = \sum_{(x,y) \in D^o} \ell(f(x; \theta), y) + \sum_{(x,y) \in D^u} \ell(f(x; \theta), y)
\]

In this paper, we consider the model performance in response to the changes from \(P^o\) to \(P^u\), specifically by allowing \(x\) to be drawn from different input spaces. Note that Definition 3.1 generalizes in a straightforward way to sequences of dataset updates by considering \(D^o\) to be the union of all prior data. The difficulty of the problem may be expressed in terms of the Kullback-Leibler divergence \(D_{KL}(P^o || P^u)\). Intuitively, small changes in distribution are easier to handle because \(\theta\) need not change substantially to perform well on \(D^u\).
3.1 Architecture

RIDDLE is a 2D parameter array $S$ that contains $L$ rows and $R$ columns $S \in \mathbb{R}^{L \times R}$. The model is indexed by a set of $L$ randomly generated LSH functions $h_l : \{1, \ldots, R\} \rightarrow \mathbb{Z}$ for $l = 1, \ldots, L$. The prediction evaluated at a query $x$ is defined as following:

$$f(x; S, h) = \frac{1}{L} \sum_{l=1}^{L} S[l, h_l(x)]$$

Algorithmically, the process consists of three steps: (1) hash code computation, (2) partition look-ups, and (3) output averaging. Specifically, the model output for a query $x$ is an average over the $L$ parameters retrieved from $S$ using the hash codes $h(x)$. Figure 1 illustrates a toy example.

For a classification problem with $C$ classes, the 2D array is repeated $C$ times, resulting in a 3D tensor of parameters $S \in \mathbb{R}^{C \times L \times R}$. The same LSH functions are used for all $C$ repetitions, and the classification prediction for query $x$ is defined as follows:

$$f(x; S, h) = \arg\max_{c \in C} \left( \frac{1}{L} \sum_{l=1}^{L} S[c, l, h_l(x)] \right)$$

RIDDLE can be used as a stand alone machine learning model on simple datasets but can also be combined with existing neural networks. Specifically, RIDDLE can use the learned representation from neural network (Figure 2). RIDDLE can also be implemented in standard deep learning frameworks and trained using SGD (Algorithm 1). It should be noted that the parameters of LSH functions $h(x)$ are not trainable. Only the values in $S \in \mathbb{R}^{L \times R}$ will be updated by SGD. The inference process consists only of inexpensive hash evaluations and memory look-ups. We use LSH functions based on sparse projections [2], which compute $h(x)$ via an inner product with parameter vectors randomly drawn from $\{-1, 0, 1\}$ with probability $\{1/6, 2/3, 1/6\}$. With such a hash function, computation can be reduced to addition, subtraction and memory lookup.

3.2 Intuition

RIDDLE may seem conceptually different from neural networks at first gasp, however, RIDDLE can be viewed as a combination of Mixture-of-Expert-Layers [23] with Kernel Methods [26]. To understand RIDDLE as a mixture of experts, observe that the gating function for every row in RIDDLE is a randomly initialized hashing function $h_l$, and the expert is one number stored at $S[i, h_i(x)]$. RIDDLE can be viewed as an ensemble of $L$ such Mixture-of-Expert-Layers. On the other side, RIDDLE is also a kernel density estimation process where the LSH collision probability is the kernel function. The number stored at $S[i, h_i(x)]$ is an efficient estimation of the weighted kernel values between the test sample and training samples.

3.3 RIDDLE for Learning with Dynamic Definition

Before we determine the class of functions that are representable using our model, we first demonstrate that this architecture addresses the problem from Definition 3.1. We consider a RIDDLE model $f(x; S, h)$ following the notation from Section 3.1. That is, we are given a model with parameters $\theta = (S, h)$ trained on $D^n$. We wish to adapt the model on $D^n$ using gradient descent.

Parameter Separation: The RIDDLE partitions parameters in such a way that inputs from different distributions are likely to use different parameters. The parameter update may be written as:

$$S_{l+1} = S_l[l, h_l(x)] + \eta g[l] \text{ for } l = 1 \ldots L,$$

where $g = \nabla f \left( \frac{1}{L} \sum_{l=1}^{L} S[l, h_l(x)], y \right)$. From this expression, we can see that $L$ parameters are involved in the computation of $f(x; S, h)$ - specifically, the parameters selected by the hash functions $\{h_1(x), \ldots, h_L(x)\}$. Based on the LSH property (Definition 2.1), we expect unrelated inputs to have different hash values and therefore use different parameters. When $P^n$ is significantly different from $P^o$, $x^o \in D^n$ and $x^o \in D^n$ are likely to activate different partition and thus use different parameter sets.
We formalize this intuition below by showing that the difference between \( f(x; S, h) \) and its initialization is bounded by an estimate of the distribution \( P^o \).

**Theorem 3.2.** Given a dataset \( D \), train a RIDDLE model \( f(x; S, h) \) by running Algorithm 1 for \( e \) epochs with learning rate \( \eta \) and initialization \( S_0 \). Suppose the gradient norms are bounded by \( G \). Then

\[
|E_h[f(x; S, h) - f(x; S_0, h)]| \leq \epsilon \eta G \text{KDE}(x, D)
\]

where \( \text{KDE}(x, D) \) is the kernel density of \( x \) over \( D \) using the kernel of the LSH function \( h \).

Theorem 3.2 allows us to prove the following guarantee:

**Theorem 3.3.** Using the notation from Definition 3.2, let \( S_o \) be the sketch that optimizes the loss \( L_o(S) \) over \( P^o \) and \( S_{o+u} \) be the sketch obtained by training over \( D^u \) for \( e \) epochs with learning rate \( \eta \) and initialization \( S_o \). If the loss is \( \text{L-Lipschitz} \) with \( G \)-bounded gradients, then:

\[
E_h[L_o(S_{o+u}) - L_o(S_o)] \leq \sum_{x \in P^o} G L \eta \text{KDE}(x, D^u)
\]

The loss difference in Theorem 3.3 quantifies the amount of information about \( D^o \) that is forgotten by the model when trained on \( D^u \). The excess loss on example \( x \in D^o \) is bounded by the kernel density estimate \( \text{KDE}(x, D^u) \). When \( D^o \) and \( D^u \) are drawn from different distributions, which is the characteristic of learning under dynamic distribution, the density becomes small (i.e. as the divergence \( D_{KL}(P^o || P^u) \rightarrow \infty \), \( \text{KDE}(x, D^u) \rightarrow 0 \)).

**Parameter Importance:** The RIDDLE allows us to determine the importance level of each parameter for encoding information from the original distribution with simple counting. We may do this by counting the number of inputs from \( D^o \) that fall into each partition, using a set of counters \( C \in \mathbb{R}^{L \times \bar{R}} \). Parameters with a large counter \( c \) can be regarded as important for the original distribution, as they participate in prediction for a large fraction of examples from \( D^o \). Our intuition is that parameters corresponding to partitions with large count values are critical for performance on \( D^o \) and should not be adjusted. We use this idea to improve performance via partition-dependent learning rates. Specifically, we allow the learning rate \( \eta(c) \) depend on the partition count \( c \). We find that the following update rule provides good empirical results:

\[
\eta(c) \propto \frac{\bar{c} + \alpha}{c \log c + 1}, \text{ where } \alpha \text{ is a constant}
\]

4 RIDDLE is a Universal Function Approximator

In this section, we analyze the class of functions that can be represented using the RIDDLE. Using techniques from the sketching literature, we prove a probabilistic guarantee.
We are interested in the representation capabilities of weighted kernel sums over an 

Applied to our problem, a universal kernel is one which can approximate any well-behaved function where the expectation is taken over the choice of LSH functions. We begin with a proof that any

Theorem 4.3.

The next step of our proof is to construct a RIDDLE capable of approximating

The crucial property is that LSH kernels are

Lemma 4.2.

Therefore, we consider the easier task of learning the model parameters values directly in practice

Constructing a RIDDLE over a carefully chosen \( N \) point dataset can approximate any continuous and bounded function. However, learning both the data and data weights with \( N \) being arbitrarily large poses difficulty in optimization. We simplify the optimization by observing that the output of the learning process is not a dataset but a model. Therefore, we consider the easier task of learning the model parameters values directly in practice as shown in Algorithm 1. This is a slight relaxation of the problem, however, we argue that with a row-sum constraint and a reasonable uniqueness assumption about the hash partitions, the two processes are equivalent in practice. Thus, the RIDDLE is a universal function approximator.

4.1 Weighted LSH Kernel Sums are Universal Function Approximators

We are interested in the representation capabilities of weighted kernel sums over an \( N \)-point dataset \( D = \{x_1, \ldots, x_N\} \). We consider the specific case where the kernel is the collision probability of an LSH function. Such kernels are referred to in the literature as LSH kernels [19].

\[
    f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)
\]

The crucial property is that LSH kernels are universal. Informally, a kernel is universal if a weighted kernel sum over a carefully-chosen point set can approximate any function with arbitrary accuracy [42]. Some restrictions apply; the size of the point set is allowed to be arbitrarily large and the accuracy is computed over any compact subset of \( \mathbb{R}^d \) rather than the entire space.

Definition 4.1. Universal Kernel [42]: Let \( S(\mathcal{X}) \) denote the space of bounded and continuous functions on a compact domain \( \mathcal{X} \). Then a kernel \( K(x, y) \) is universal if it is continuous and induces a reproducing kernel Hilbert space that is dense in \( S(\mathcal{X}) \).

Applied to our problem, a universal kernel is one which can approximate any well-behaved function over compact subsets of \( \mathbb{R}^d \) using a linear combination of kernels. The theory of universal kernels allows for a convergence-style proof for weighted LSH kernel sums. We begin by showing that the \( p \)-stable LSH functions are universal kernels. We defer proofs to the appendix.


Theorem 4.3. Given a continuous and bounded function \( g(q) \) and any \( \epsilon > 0 \), there exists a set of coefficients \( \{\alpha_n\} \), set of points \( \{x_n\} \) and an integer \( N \) such that

\[
    f_N(q) = \sum_{n=1}^{N} \alpha_n K(x_n, q) \quad \|f_N(q) - g(q)\|_{\mathcal{X}} \leq \epsilon
\]

where \( K(x_n, q) \) is the \( L^2 \) LSH kernel and \( \mathcal{X} \) is any compact subset of \( \mathbb{R}^d \).

4.2 RIDDLE for Weighted LSH Kernel Sums

The next step of our proof is to construct a RIDDLE capable of approximating \( f_N(q) \) from Theorem 4.3. We consider a RIDDLE model with \( S \in \mathbb{R}^{\mathbb{R}^d \times \mathcal{X}} \) constructed from a set of points \( \{x_n\} \) with
We consider four datasets that reflect dynamic distributions. Every dataset is organized into two parts: a training dataset $D$ and an update dataset $D^u$. The training dataset $D$ consists of images of 4, 1, 7, 9, 0, 8. The update dataset $D^u$ consists of images of 5, 2, 3, 6.

Although we often use the average to do the aggregation in practice, we analyze the median-of-means estimator $\tilde{S}$.

Theorem 4.5 (RIDDLE Estimator). Given a dataset $D$ of weighted samples $\{(x_i, f_i)\}$, let $h(x)$ be an LSH function drawn from an LSH family with collision probability $\mathbb{P}_{\text{LSH}}$. Let $S$ be a 2d parameter array constructed using $h(x)$. For any query $q$, we have:

$$\mathbb{E}(S[h(q)]) = \sum_{i} \alpha_i, K(x_i, q), \text{var}(S[h(q)]) \leq \left( \sum_{i} \alpha_i, \sqrt{K(x_i, q)} \right)^2$$

The key insight is that each row of $S$ is an unbiased estimator for weighted sums of LSH kernels. By finding the central tendency of the $L$ rows, we obtain a sharp concentration around $f_N(q)$. Although we often use the average to do the aggregation in practice, we analyze the median-of-means estimator $\tilde{S}$, as it allows us to prove an exponential concentration of the estimate around the weighted KDE. Note that while this simplifies the proof, it does not affect the convergence to $f_N(q)$ as $L \to \infty$. We combine the variance bound from Theorem 4.4 with the median-of-means guarantee to obtain a relationship between the RIDDLE parameters $S$, weighted kernel values, and estimation error.

Theorem 4.5 (Weighted-KDE Estimation Error). Let $Z(q)$ be the median-of-means estimate constructed using the $L$ unbiased estimators. Then with probability $1 - \delta$,

$$|Z(q) - f_K(q)| \leq \epsilon = 6 \frac{\tilde{f}_K(q)}{\sqrt{L}} \sqrt{\log 1/\delta}$$

where $f_K(q)$ and $\tilde{f}_K(q)$ are the weighted KDE with kernels $K(x, q)$ and $\sqrt{K(x, q)}$, respectively.

The universal approximation property of our model follows by observing that the error $\epsilon$ in Theorem 4.5 goes to zero as $N \to \infty$ and $\epsilon \to 0$ in Theorem 4.3 as $L \to \infty$.

5 Experiments

In this section, we first evaluate RIDDLE on practical dynamic distribution learning task in Section 5.1. Further, since we prove RIDDLE as a universal function approximator, we provide an extensive study on the expressiveness and efficiency of RIDDLE in Section 5.2. Code is available at https://github.com/lzcemma/RIDDLE.

5.1 RIDDLE for Learning under Dynamic Definition

Here, we investigate whether RIDDLE can adapt to new distributions while retaining good performance on the original distributions.

5.1.1 Datasets

We consider four datasets that reflect dynamic distributions. Every dataset is organized into two parts: the original dataset $D^o$ and the update dataset $D^u$. $D^o$ and $D^u$ have significant semantic differences and do not share any overlapping training data.

**MNIST Binary**: We randomly partitioned digits into two groups (e.g., 4, 1, 7, 5, 3 and 9, 0, 8, 6, 2) and train the model to predict the group assignment of an image. For the original distribution, we randomly choose 3 digits from the first group and 3 digits from the second group such as $D^o$ only consists images of 4, 1, 7, 9, 0, 8. The update dataset consists of images of the remaining digits.
Figure 2: RIDDLE can be used as a stand-alone model, as shown in (a). We use this setting for experiments on MNIST-binary and all experiments studying expressiveness and efficiency. RIDDLE can also be combined with existing neural networks, as shown in (b). We replace the classification layer with RIDDLE for experiments on CIFAR10, ImageNet, and News.

**CIFAR10:** The task is to predict whether the object in the image is capable of carrying things (e.g., car and horse can carry things, while a frog cannot). The original dataset \( D^o \) contains images of 6 out of 10 objects that can and cannot carry things, and the update dataset \( D^u \) contains the remaining 4 objects.

**ImageNet:** The task is to predict whether a person is allergic to the object in the image. We select the image classes overlapping with national allergen directory \( D^o \), and assign images to the allergen and non-allergen classes based on the common allergen group in national allergen directory \( D^o \). The original dataset \( D^o \) contains a subset of allergens and non-allergens, while the update dataset \( D^u \) contains the remaining objects. This task was built to mimic the seasonal addition and sensitization of new allergens.

**News:** The original task for the News dataset is to predict the news topics given a caption and short description. We re-task the dataset to predict whether a person would be interested in a new article by assigning some topics as interesting and the rest as not-interesing. The original dataset \( D^o \) contains news from a subset of interesting and non-interesting topics, while the update dataset \( D^u \) contains the rest. We organize this dataset to reflect changes in a user’s interests.

5.1.2 Baselines

We compare RIDDLE with two updating strategies on different architectures.

**Update:** We train a neural network to full convergence on the original dataset \( D^o \). Then, for updating, we train models only on the update dataset \( D^u \), where the initialization is the parameters obtained on \( D^o \). This approach does not require data storage nor training computation over the original dataset.

**Elastic Weight Consolidation (EWC):** \( D^o \) was proposed to mitigate catastrophic forgetting by selectively slowing down learning on the weights important for previous tasks using Fisher information.

5.1.3 Experiment Details

For tasks where it is reasonable (CIFAR10, ImageNet and News), we exploit recent advances in large scale pretrained models and replace the final layer with RIDDLE. Pretrained models are usually trained on a diverse corpus of unlabeled data and are likely to provide good featurization across distributions. \( D^o \) also observe that pretrained models are more resistant to forgetting.

5.1.4 Main Results

Table A.1 summarizes the test accuracy of all models after training on the original dataset and after the update. After training on the original dataset, we observe that all models achieve roughly the
Table 1: This table summarizes the accuracy after training on original dataset, and after training on update dataset. The RIDDLE obtains comparable accuracy on original test set and update test set after trained on corresponding train set. The RIDDLE’s accuracy is significantly higher than all baselines on original test set after updating.

We are interested in the accuracy on original dataset after updating the model on update dataset. We observe a 20% to 30% accuracy drop on the original accuracy with naive updating. Models such as Resnet18 and ViT display catastrophic forgetting behavior, eventually performing no better than random. EWC is effective in prevent forgetting to some extend, reducing the degradation to around 10% to 20%. The RIDDLE significantly outperforms both baselines combined with all architectures across datasets up to in terms of original test set accuracy. Indeed, the RIDDLE keeps the accuracy degradation below 7%. For example, the accuracy drop is only 0.9% on Mnist-Binary. On the updated test set, the RIDDLE achieves competitive accuracy (Table A.1), indicating that it is learning the properties of both distributions.

Note that ViT outperforms Resnet18 on both CIFAR10 and ImageNet, which confirms our intuition that pretrained models are inherently more robust to distribution changes. By switching the final classification layer to the RIDDLE, we increase the accuracy by a large margin. The experiment results strongly support our hypothesis that the RIDDLE retains the ability to adapt to new distributions while being less vulnerable than established methods to knowledge loss.

5.1.5 Ablation on Update Rule

There are two main drivers of performance when training RIDDLE over dynamically evolving definitions: the hash function and the parameter-dependently learning rate.

Without a learning rate that depends on parameter importance, the test accuracy on the original MNIST test set after updating is 94.9%, which is 26% higher than Update-MLP, 4% higher than EWC-MLP. Using the dynamic learning rate, RIDDLE achieves 98%. There could be smarter or simpler update functions based on the parameter count, which we leave for future work.

Table 2: This table displays the effect of the partition-dependent learning rate. “Global Learning Rate” means that the learning rate is not related to parameter counts.
Table 3: This table summarizes the accuracy comparison. NN1 and NN2 denote two different deep learning models. RIDDLE achieves higher accuracy than at least half of the baselines on all datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>LightGBM</th>
<th>Random Forest</th>
<th>XGBoost</th>
<th>NN1</th>
<th>NN2</th>
<th>RIDDLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Susy</td>
<td>0.7903</td>
<td>0.7880</td>
<td>0.7926</td>
<td>0.8019</td>
<td>0.8024</td>
<td>0.7924</td>
<td></td>
</tr>
<tr>
<td>HAR</td>
<td>0.8532</td>
<td>0.8387</td>
<td>0.8532</td>
<td>0.9027</td>
<td>0.9123</td>
<td>0.9001</td>
<td></td>
</tr>
<tr>
<td>Covtype</td>
<td>0.9565</td>
<td>0.9932</td>
<td>0.9874</td>
<td>0.9836</td>
<td>0.9936</td>
<td>0.9853</td>
<td></td>
</tr>
<tr>
<td>Connect4</td>
<td>0.8171</td>
<td>0.8335</td>
<td>0.8171</td>
<td>0.8558</td>
<td>0.8657</td>
<td>0.8586</td>
<td></td>
</tr>
<tr>
<td>Fashion MNIST</td>
<td>0.8848</td>
<td>0.8777</td>
<td>0.8848</td>
<td>0.8852</td>
<td>0.8889</td>
<td>0.8852</td>
<td></td>
</tr>
<tr>
<td>Boston Housing</td>
<td>1.77</td>
<td>2.05</td>
<td>1.62</td>
<td>2.57</td>
<td>2.16</td>
<td>1.74</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: This table summarizes the efficiency comparison between the RIDDLE and a neural network (NN) with similar accuracy. The RIDDLE reduces memory by up to 3.7x, FLOPs by up to 17x and inference time by up to 3.9x.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Accuracy</th>
<th>Memory(MB)</th>
<th>Flops (M)</th>
<th>Inference Time (µs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>NN</td>
<td>0.9815</td>
<td>3.59</td>
<td>0.90</td>
<td>164.177</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>0.9784</td>
<td>1.36(1.6X)</td>
<td>0.15(6.1X)</td>
<td>42.003 (3.9X)</td>
</tr>
<tr>
<td>Fashion MNIST</td>
<td>NN</td>
<td>0.8852</td>
<td>2.59</td>
<td>0.65</td>
<td>131.385</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>0.8852</td>
<td>1.92(1.6X)</td>
<td>0.13(5X)</td>
<td>39.228 (3.3X)</td>
</tr>
<tr>
<td>HAR</td>
<td>NN</td>
<td>0.9027</td>
<td>2.11</td>
<td>0.53</td>
<td>166.584</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>0.9001</td>
<td>1.18(1.8X)</td>
<td>0.18 (2.9X)</td>
<td>46.567 (3.6X)</td>
</tr>
<tr>
<td>Boston Housing</td>
<td>NN</td>
<td>2.16</td>
<td>1.96</td>
<td>0.51</td>
<td>61.338</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>1.74</td>
<td>0.53(3.7X)</td>
<td>0.03 (17X)</td>
<td>21.862 (2.8X)</td>
</tr>
</tbody>
</table>

5.2 Study on Expressiveness and Efficiency

In this section, we investigate whether RIDDLE is competitive among popular machine learning models in terms of expressiveness and efficiency. Our goal is to show that RIDDLE provides a good model representation for a variety of tasks. We curated a diverse set of datasets from UCI and Kaggle. These datasets cover a wide range of tasks and domains and are widely used as benchmarks in other works [64, 22]. All experiments are conducted on a machine with 96 24-core/2-thread/2-socket processors (Intel Xeon(R) Gold 5220R 2.20GHz) and 8 Nvidia V100 32GB.

In Table 5.2, we observe that RIDDLE outperforms popular linear methods and is competitive compared to deep learning methods. NN1 and NN2 denote two different neural network architectures (Details in Appendix). Both the architecture and evaluation metric is included in Table A.2. We also investigate the efficiency and memory requirement. This is important, as it shows that RIDDLE does not outperform other methods due to an increased parameter size. Table 5.2 shows that RIDDLE attains the same accuracy as baseline methods with a smaller memory footprint, fewer FLOPs and a faster inference speed. Specifically, we use up to 3.7X less memory, use 17X less FLOPs, and achieve 3.9x faster inference speed. The memory improvements likely arise from the meaningful parameter groupings, while the computational improvements are because it only uses addition, subtraction and array lookup operations. This experiment confirms that even though the RIDDLE allocates different parameters to different distributions, we do not require more parameters than existing methods to achieve robustness or accuracy. One limitation now is that memory grows linear with number of class, which can be address with smart hashing mechanism.

6 Conclusion

In this work, we propose an expressive and efficient model RIDDLE for learning under dynamic distribution. Our setup concerns common scenarios where model has to be updated given distribution changes. The model is structured to be invulnerable to forgetting because gradients updates are naturally restricted to only input relevant parameters. Further, RIDDLE can be used for efficient weighted kernel density estimation.

Acknowledgments

This work was supported by National Science Foundation SHF-2211815, BIGDATA-1838177, ONR DURIP Grant, and grants from Adobe, Intel, Total, and VMware.
References


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? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

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]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (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]

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? [Yes]
   (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]