Online Task-Free Continual Learning with Dynamic Sparse Distributed Memory (DSDM)

1. Introduction

Dynamic Sparse Distributed Memory (DSDM) is an associative, content-addressable memory model designed to address the challenges of **online, task-free continual learning**. Unlike conventional learning paradigms that rely on task boundaries or repeated access to previous data, DSDM learns from **non-stationary data streams** in a single pass, dynamically adapting its memory structure as new patterns emerge.

The core motivation behind DSDM is to overcome **catastrophic forgetting**, a fundamental limitation of gradient-based neural networks in continual learning. By replacing global weight updates with a **dynamic memory of sparse, distributed representations**, DSDM enables stable learning across evolving data distributions without explicit task information.

2. Dynamic Sparse Distributed Memory (DSDM)

2.1 Memory Structure

The memory (M) in DSDM is represented by two matrices:

- Address matrix $A \in \mathbb{R}^{K \times m}$
- Content matrix $C \in \mathbb{R}^{K \times d}$

Each row ((a_i, c_i)) corresponds to a memory neuron, where:

- (a_i) is an **address vector** (input feature representation)
- (c_i) is a **content vector** (typically a one-hot encoded class label)

The memory is initialized using the first observed sample ((x_1, y_1)). Unlike classical Sparse Distributed Memory (SDM), which assumes a fixed number of randomly initialized memory nodes, **DSDM starts with an empty memory and grows dynamically**, adding new neurons as needed based on the structure of the incoming data. This adaptive mechanism provides higher storage capacity and more faithful modeling of evolving data distributions.

2.2 Online and Task-Free Learning Mechanism

DSDM operates in a **fully online and task-free** manner, processing each incoming data sample ((x_t, y_t)) only once.

For every new input (x_t), the algorithm identifies the **Best Matching Unit (BMU)**:

$$\mathrm{BMU} = \arg\min_i \|x_t - a_i\|_2$$

The resulting distance, referred to as the *to-BMU distance* (d_{BMU}), determines how the memory is updated.

To dynamically regulate memory growth, DSDM introduces a **recursive temperature (RT)**—an exponential moving average of recent to-BMU distances. RT increases when the input stream shifts (e.g., new classes or tasks), and decreases once the memory adapts. This adaptive threshold governs neuron creation:

- **If** (d{BMU} > RT):
  A **new memory neuron** is created by appending (x_t) to (A) and (y_t) to (C).

- **If** (d{BMU} <= RT):
  The information is **distributed among existing neurons**. Both addresses and content vectors are updated using a **softmin weighting function**, where closer neurons receive higher influence.

This learning strategy enables **local generalization without overwriting previously learned knowledge**, significantly reducing catastrophic forgetting.

2.3 Memory Pruning

To prevent unbounded growth, DSDM enforces a maximum memory capacity (Q). When the number of neurons (K) exceeds this capacity, a **density-based pruning strategy** based on the **Local Outlier Factor (LOF)** algorithm is applied. Highly dense neurons, assumed to represent redundant information, are removed while preserving coverage of the feature space.

2.4 Inference Process

During inference, for a query input (x_q), DSDM computes the Euclidean distance between (x_q) and all stored addresses. These distances are then converted into activation weights using a **softmin function** parameterized by temperature (beta):

- Large (beta): broader, more distributed activation
- Small (beta): sharper, nearest-neighbor-like behavior

The final prediction is obtained as a weighted sum of content vectors:

$$y_q = \sum_{i=1}^{K} w_i \cdot c_i$$

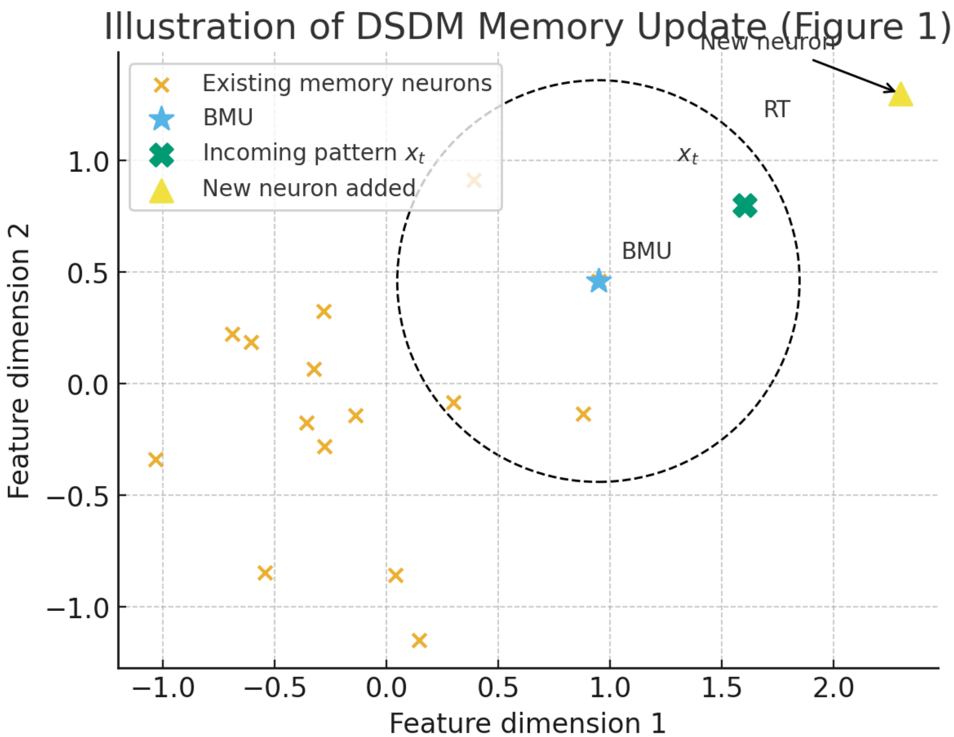

Figure 1: *Illustration of DSDM memory operation.*

Address space showing existing memory neurons, Best Matching Unit (BMU), recursive temperature threshold (RT), and the creation of a new neuron when an incoming pattern lies outside the RT boundary.

3. Experimental Evaluation and Results

DSDM was evaluated across a wide range of continual learning benchmarks and consistently demonstrated superior performance in **online and task-free settings**.

3.1 Split MNIST (No Encoder)

On Split MNIST, DSDM achieved a **last accuracy of 94.0 ± 0.2**, significantly outperforming classical SDM (74.2 ± 3.5) and surpassing deep SOTA methods such as CURL and CN-DPM—despite using **no encoder and no backpropagation**. This highlights the effectiveness of DSDM's memory-based learning alone.

3.2 Online Benchmarks: CIFAR-10 and CORE-50

In online comparisons, DSDM outperformed methods like Candidates Voting (CV), which rely on task boundaries:

- **CORE-50 (Q = 5k):** DSDM achieved **57.1%** last accuracy versus CV's 43.1%.
- **Split CIFAR-10 (step=2, Q=1k):** DSDM obtained **67.0%**, exceeding CV's 62.9%.

These results demonstrate DSDM's robustness in realistic, task-free continual learning scenarios.

3.3 Offline Comparisons: CIFAR-100 and CUB-200

Despite being an online single-epoch algorithm, DSDM surpassed several offline continual learning methods:

- **CIFAR-100 (step=2):** 63.3% average accuracy
- **CUB-200 (step=2):** 55.5% average accuracy

Offline competitors required significantly more computational resources, including large batch sizes and multiple training epochs.

3.4 Few-Shot and Gaussian Schedule Experiments

Under the Gaussian schedule—where class boundaries are undefined—DSDM achieved dramatic improvements over Ensemble-based approaches:

- **CIFAR-10 (DSDM + ViT):** 84.9 ± 0.6
- **CIFAR-100 (DSDM + ViT):** 61.4 ± 1.1

This confirms DSDM's capability to dynamically adapt to highly non-stationary data streams.

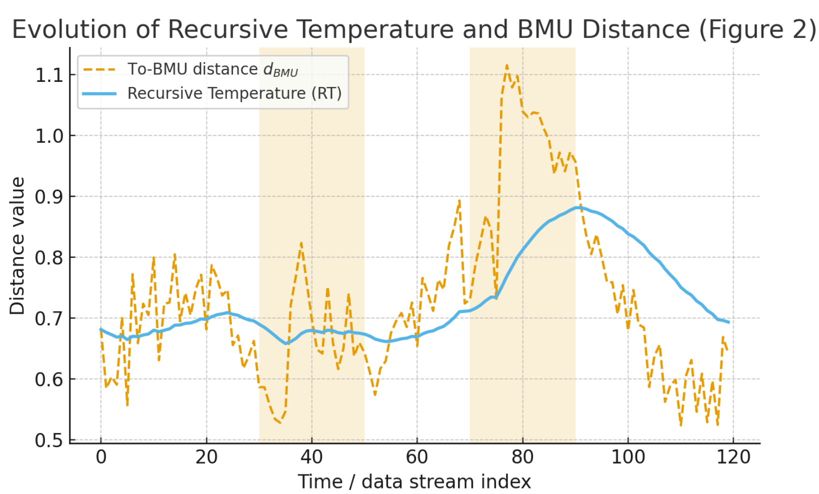

Figure 2: *Evolution of recursive temperature (RT) and to-BMU distance over time.*

RT spikes during distribution shifts and decays as memory adapts, demonstrating dynamic task-free segmentation.

4. Advantages of DSDM

- **True Online and Task-Free Learning:**
  Operates without task boundaries or replay, supporting continuous evaluation.
- **Effective Catastrophic Forgetting Mitigation:**
  Knowledge stored via semi-distributed clusters rather than global weights.
- **High Computational Efficiency:**
  Single-pass learning with batch size 1 and one epoch.
- **Dynamic and Adaptive Memory:**
  Recursive Temperature enables memory growth aligned with data distribution.
- **SOTA Performance Without Gradient Flow:**
  Achieves state-of-the-art accuracy even without backpropagation.

5. Limitations and Dependencies

- **Reliance on Fixed Pretrained Encoders:**
  Performance on complex datasets depends on frozen CNN or ViT encoders.
- **Memory Pruning Requirement:**
  Density-based pruning introduces heuristic decisions.
- **Fixed Hyperparameters:**
  RT, sharpening rate, and inference temperature are not adaptive.
- **Focused on Classification:**
  Extension to regression and anomaly detection remains future work.

6. Conclusion

DSDM represents a powerful alternative to gradient-based continual learning by leveraging a dynamic, sparse, and distributed memory structure. Its ability to achieve strong performance in **online, task-free, and low-data regimes** positions it as a promising framework for real-world continual learning applications.