The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)

Reducing Catastrophic Forgetting in Self Organizing Maps with Internally-Induced Generative Replay (Student Abstract)

Hitesh Vaidya, Travis Desell*, **Alexander Ororbia***

Rochester Institute of Technology
20 Lomb Memorial Dr., Rochester, NY 14623
{hv8322, tjdvse}@rit.edu, ago@cs.rit.edu

Abstract

A lifelong learning agent is able to continually learn from potentially infinite streams of pattern sensory data. One major historic difficulty in building agents that adapt in this way is that neural systems struggle to retain previously-acquired knowledge when learning from new samples. This problem is known as catastrophic forgetting (interference) and remains an unsolved problem in the domain of machine learning to this day. While forgetting in the context of feedforward networks has been examined extensively over the decades, far less has been done in the context of alternative architectures such as the venerable self-organizing map (SOM), an unsupervised neural model that is often used in tasks such as clustering and dimensionality reduction. Although the competition among its internal neurons might carry the potential to improve memory retention, we observe that a fixed-sized SOM trained on task incremental data, i.e., it receives data points related to specific classes at certain temporal increments, it experiences severe interference. In this study, we propose the c-SOM, a model that is capable of reducing its own forgetting when processing information.

Introduction

Lifelong machine learning, otherwise known as continual and never-ending learning (Chen and Liu 2016), stands as one of the greatest challenges facing artificial intelligence research, especially with respect to models parameterized by deep neural networks (DNNs). In this problem context, an agent must attempt to learn not just one single prediction task using one single dataset, but rather, it must learn *across* several task datasets, much as human agents do, aggregating and transferring its knowledge as new pattern vectors are encountered. DNNs particularly struggle to learn continually due to the well-known classical fact that they tend to *catastrophically forget* (French 1999), or rather, they completely erase the knowledge acquired during the learning of earlier tasks when processing samples from new tasks.

While catastrophic forgetting has been explored extensively in DNNs, very little work has focused on the occurrence and reduction of forgetting in less popular architectures, such as self-organizing maps (SOMs) (Kohonen 1982). SOMs are a type of competitive neural system where internal neurons compete for the right to activate in the presence of a data pattern and synaptic parameters are adjusted through a Hebbian update. Neuronal units can be arranged in a topological fashion, i.e., a neighbourhood based on Euclidean distance between the activation values of units, as well as a spatial fashion, i.e., units arranged in Cartesian plane. Given that units in a competitive learning system like the SOM tend to specialize for certain types of patterns, i.e., data prototypes, such a system appears to be suited for combating catastrophic forgetting (since, in theory, competing units would lead to reduced neural cross-talk, a source of forgetting (French 1999)). Furthermore, approaches, such as (Gepperth et al. 2015), have advocated the use of SOMs for continual learning. However, the SOM in its purest form, as we observe in our experiments, is itself prone to forgetting (Mermillod, Bugaiska, and Bonin 2013).

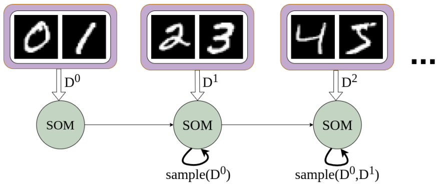

Figure 1: Our proposed c-SOM architecture.

In this work (Figure 1), we propose the c-SOM (continual SOM), an SOM that actively reduces the amount of forgetting it experiences. Specifically, we modify the SOM's decay to be task-dependent and extend its units to self-induce a form of generative rehearsal to improve memory retention.

Methodology

Problem Setup: Consider a sequence of N tasks, denoted by $\mathcal{S} = \bigcup_{t=1}^{T} \mathcal{T}_t$. Each task $\mathcal{T}_t$ has a training dataset (with C classes), $\mathcal{D}^{(t)}_{train} = \bigcup_{i=1}^{N_t} \{(\mathbf{x}^{(t)}_i, \mathbf{y}^{(t)}_i)\}$, where $\mathbf{x}^{(t)}_i \in \mathcal{R}^{D \times 1}$ is a data pattern and $\mathbf{y}^{(t)}_i \in \{0, 1\}^{C \times 1}$ is its label vector (N_t is the number of data points in task T_t). Similarly, $\mathcal{D}^{(t)}_{test}$ is the test dataset for task $\mathcal{T}_t$. When a lifelong learning model is finished training on task $\mathcal{T}_t$ using $\mathcal{D}^{(t)}_{train}$, $\mathcal{D}^{(t)}_{train}$ will be lost as soon as the model proceeds to task $\mathcal{T}_{t+1}$ to $\mathcal{T}_T$ ($\mathcal{D}^{(t)}_{test}$ is

*These authors contributed equally.

Algorithm 1: c-SOM Training

```
Input: Task input, $\mathbf{x}_i^{(t)}$, c-SOM weights $\{\mathbf{w}_v, \mathbf{w}_v^{\sigma}\}$
Parameter: $\sigma$, $\lambda$
for $t = 0$ to $T$ do
    $\sigma_t, \lambda_t \leftarrow$ decay parameters for t // (Eqn 1 & 2)
    Update $\mathbf{w}_v$ on ($\mathbf{x}_i^{(t)}$, $\sigma_t$, $\lambda_t$) // (Hebbian update)
    Update $\mathbf{w}_v^{\sigma}$ (for all units) // (Welford update)
    Generate $K$ samples from $m_r$ randomly chosen, trained c-SOM units & then retrain SOM on them
```

only used for evaluation). The objective is to maximize the agent's performance on task $\mathcal{T}_t$ while minimizing how much its performance on prior tasks $\mathcal{T}_1$ to $\mathcal{T}_{t-1}$ degrades.

The Model: For our c-SOM (Algorithm 1), σ, λ are, respectively, the initial radius and learning rate values. σ_t, λ_t are their values for task $\mathcal{T}_t$. Both values decay as follows:

$$\lambda_t = \lambda(1 + t * \exp(t/\tau_\lambda))^{-1} \qquad (1)$$

$$\sigma_t = \sigma(1 + t * \exp(t/\tau_\sigma))^{-1}. \qquad (2)$$

$\tau_\lambda = \tau_\sigma = 1000$ are the time constants. u is the best matching unit (BMU) while v is any other unit in the SOM (with m units). $h(u, v, t)$ represents the neighbourhood function between units u and v for task $\mathcal{T}_t$. Each unit v in our SOM is composed of two coupled vectors – its prototype weights $\mathbf{w}_v \in \mathcal{R}^{D \times 1}$ and its running variance weights $\mathbf{w}_v^{\sigma} \in \mathcal{R}^{D \times 1}$, where $\mathbf{w}_v$ is updated via a Hebbian update and $\mathbf{w}_v^{\mu}$ and $\mathbf{w}_v^{\sigma}$ are updated via Welford's online algorithm. $\phi(c)$ returns the number of units trained on class c (out of C), while ρ_v stores the class that unit v maps to. At each simulation step, the c-SOM generates K samples from m_r randomly chosen trained units via: $\mathbf{s}^v = \mathbf{w}_v + \sqrt{\mathbf{w}_v^{\sigma}} \odot \epsilon$ ($\epsilon \sim \mathcal{N}(0, 1)$). The $K \times m_r$ samples are then used to refresh the c-SOM (via the same update rules).

Figure 2 (Left) displays the final output after training a classical SOM on the benchmark Split-MNIST ($C = 1$ per task) with a grid shape of 10×10 (this small grid size was chosen to exacerbate forgetting). As seen among its prototypes, this SOM remembered only classes 6, 8 and 9. Except for 6, classes 8 and 9 are the ones encountered in the last task of Split-MNIST.In contrast, Figure 2 (Right) shows our proposed SOM model with the same shape. Desirably, our model contains units that represent digits in all classes encountered across all tasks of Split-MNIST, exhibiting far less forgetting. Note that all SOMs were trained in a *class incremental* fashion.

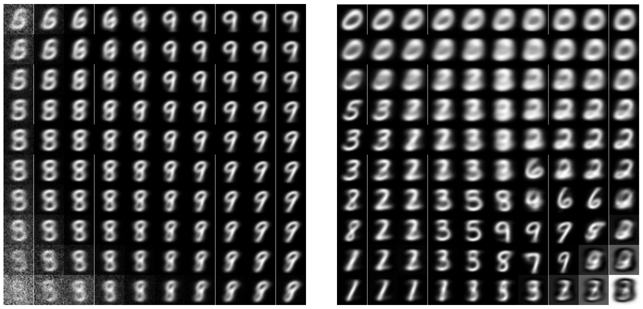

Figure 2: Classical SOM (left) vs. proposed SOM (right).

Algorithm 2: α metric

```
Input: $\mathbf{X}_{test}$, where $\mathbf{X}_{test}^c$ is all $\mathbf{x}_i$ w/ $\arg\max(\mathbf{y}_i) \equiv c$
Parameter: Weights $\{\mathbf{w}_v, \mathbf{w}_v^{\sigma}\}$
$N_c = V/C$, where C = total number of classes
$\rho_v := \arg\min \left( d = \{\|\mathbf{w}_v - \mathbf{X}_{test}^{(c)}\| : \forall\, c \,\epsilon\, C\} \right)$
$\phi = \left\{ \sum_{v=0}^{V} \mathbb{1}(c = \rho_v) : \forall\, c \,\epsilon\, C \right\}$
$\alpha_{\text{mem}} = (\phi_i - N_c)_{RMS}$ // RMS = "root-mean-square"
```

Model	(σ, λ)	α_{mem}	$\mu_{\alpha_{mem}}$	$\sigma_{\alpha_{mem}}$
SOM	(2,0.001)	12.75	17.97	3.02
c-SOM (K=1)	(3,0.01)	11.85	12.49	0.45
c-SOM (K=2)	**(3,0.01)**	**9.9**	**12.15**	**1.25**

Table 1: Parameters with best (minimum) α_{mem} values for each model ($m_r = 1$). $\mu_{\alpha_{mem}}$ and $\sigma_{\alpha_{mem}}$ over 10 trials.

In order to measure the performance of incrementally learned SOMs, we designed a novel metric, α_{mem} (see Algorithm 2). An ideal lifelong learning SOM would have $\alpha_{\text{mem}} = 0$, which would represent an equal representation of classes across units. Therefore, lower α_{mem} values represent better SOM performance. We experimented with different parameter settings of (σ, λ) for three different SOM versions. The σ value was selected out of {2,3,4,5} while the λ was chosen from {0.001, 0.001, 0.007, 0.01}. Table 1 presents the best α_{mem} value along with their corresponding mean and standard deviations across the 10 trials.

Conclusion

In this study, we proposed a variant of the SOM that actively reduces the amount of forgetting it experiences when trained in a class-incremental fashion. Qualitative results on Split-MNIST demonstrate an improvement in memory retention.

References

Chen, Z.; and Liu, B. 2016. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 10(3): 1–145.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135.

Gepperth, A.; Hecht, T.; Lefort, M.; and Korner, U. 2015. Biologically inspired incremental learning for high-dimensional spaces. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 269–275.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1): 59–69.

Mermillod, M.; Bugaiska, A.; and Bonin, P. 2013. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4: 504.