UNIVERSITY OF CHICAGO / MEDICINE & BIOLOGICAL SCIENCES DIVISION

Overcoming Catastrophic Forgetting Using Recurrent Neural Networks

Gregory Grant, Chaihyun (Catherine) Lee, Varun Iyer
Freedman Laboratory, 5812 S. Ellis Ave. MC0912, P-419 – Chicago, IL, 60637

The Problem

Artificial neural networks (ANNs), artificial intelligence computing systems inspired by *in vivo* neural pathways, imitate human learning by analyzing datasets consisting of well-defined inputs and expected outputs. While training on each dataset, an ANN optimizes its internal parameters to maximize task performance, but generally cannot build upon past experience. This phenomenon, referred to as "catastrophic forgetting," identifies that when a standard ANN trains on a new task, it rapidly forgets any previously learned tasks as its weights adjust to this newest task.

By implementing a Recurrent Neural Network (RNN), we can model both short-term and long-term neural processing through complex processing circuits and synaptic memory stabilization.

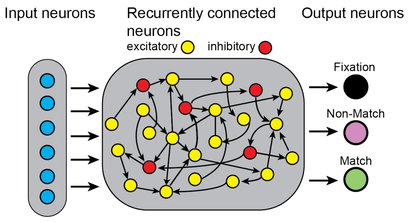


Figure 1. The Recurrent Network
In recurrent neural networks, like the one shown here, the ability of neurons to connect to other neurons in the hidden layer allows for feedback, short-term memory, decision-making, and other complex processes.

I. Dendrites

A standard *in vivo* neuron is composed of a soma (processor), dendrites (receivers), and an axon (sender). Electrical signals are transmitted from one neuron's axon to another's dendrites, and networks of such connections facilitate task solving and decision making.

Current neural networks usually only simulate the soma. We have added dendrites to each neuron in the network to allow a single set of neurons can solve multiple tasks by utilizing different dendrites for each task, a phenomenon we refer to as "activation trajectories." However, this concept, based on recent research (Cichon and Gan, Nature, 2015), fails when dendrites are naively applied to our network without any other changes. Deliberately matching dendrites to tasks produces the expected high performance for all tasks, but to allow the network to develop such activation trajectories organically, we turn to implementations of neural circuits and disinhibitory pathways.

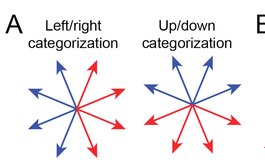


Figure 2. Categorization Task (top)
A. Our RNN model was trained on tasks based on two different categorization rules. Arrows shown in red belong to one category; those in blue belong to the other. B. Dendrites are intended to be used for task-specific gating of inputs to the neuron. Based on the current rule, dendrites are expected to either be inhibited or allow activation.

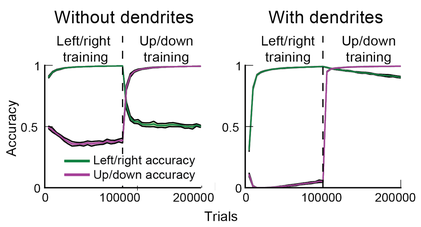
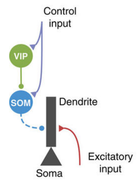


Figure 3. Dendrite Performance (bottom)
As mentioned in the text, the use of dendrites allows the network to perform at high accuracy for both tasks. For these tests, we used deliberately-matched dendrites to simulate a fully-optimized set of task pathways through the network. Without dendrites, we can see the expected effect of catastrophic forgetting.

II. Disinhibitory Pathways

In our network, we developed initial conditions to create the structure for disinhibition pathways (DPs), a type of information-gating neural circuit. We chose to mimic a type of *in vivo* DP known to establish task-specific dendritic behavior in ambient task conditions (Yang et al., Nature, 2015).

Our initial conditions, chosen through a grid-search in a statistical model, were able to demonstrate that the formulation of these DPs can result in single-task activation trajectories. By initializing our networks with the DPs in place, we were able to slightly increase the rate at which our networks learned new tasks. However, the network did not assign inherent importance to these pathways, which caused them to be ignored in favor of simpler solutions. To solve this further problem, we considered encouraging the network to form such pathways on its own.



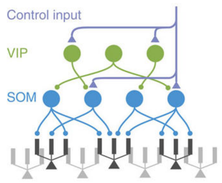


Figure 4. Disinhibitory Pathway (left)
A common disinhibitory pathway (DP) in our brain is composed of VIP, SOM, and PV neurons. Control inputs target both VIP and SOM inhibitory neurons, while VIP mainly projects to SOM neurons. SOM neurons, then, project to PV neurons. This effectively "gates" information through the DP.

Figure 5. DP Network Trajectories (right)
With such DPs wired into the network, the model is able to use multiple activation trajectories that are task-specific. This diagram shows one such layout.

(Yang et. al, Nature, 2015) Figures 4 and 5 are used with permission.

III. Motif Circuits

Motifs are small neural circuits, or groups of neurons, that fulfill a single, simple purpose. One such circuit is a simple memory circuit, where a pair of reciprocally connected neurons feed information into one another. In RNNs where we encourage these memory motifs (by penalizing the network when motifs are absent), the networks are able to retain information over long delay times more easily than their no-motif counterparts.

Encouraging further motifs proved to be too complex to reduce into a simple penalization function for our network, but this proof-of-concept memory cell motif shows that dynamic growth of complex pathways (including DPs) and circuits is possible within the realm of RNNs.

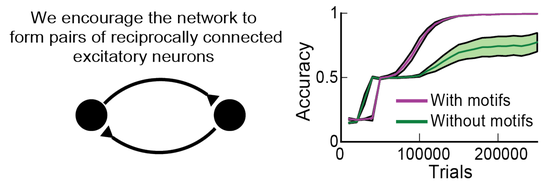


Figure 6. Motif Circuit and Performance
The simplest possible motif is the two-neuron memory cell, shown here. When the network is encouraged to use these motifs, it learns new tasks more quickly than without. One tasks's motifs are lost in the next task, however, due to catastrophic forgetting.

Acknowledgements / Contact

Many thanks to Dave Freedman and Nicolas Masse of the Freedman Lab, with whom this research was conducted.
Correspondence should be addressed to Gregory Grant (gdgrant@uchicago.edu), Catherine Chaihyun Lee (clee0505@uchicago.edu), and Varun Iyer (varuni@uchicago.edu).

IV. Weight Stabilization

ANNs mimic memory creation through parameter optimization, but have no impetus to preserve important memories. This is the root cause of catastrophic forgetting, and must be solved for the benefits of dendrites, DPs, and motif circuits to be fully engaged.

Long-term memory structure can be preserved through synaptic plasticity, for which there are multiple approaches. One approach tracks the relative importance of each network parameter to each solved task. If a parameter is vital to a task solution, it is fixed in place (Zenke, et al., arXiv, 2017). This synaptic regularization allows for new tasks to be learned without interference to previous memories, and was successful in our trials. We also added a parameter-altering cascade model to each synaptic parameter to simulate the varying timescales of molecular interactions *in vivo* and isolate important parameters (Benna and Fusi, Nature, 2016). This latter addition has been unsuccessful, and requires further testing.

$$\tilde{L}_\mu = L_\mu + \underbrace{c\sum_k \Omega_k^\mu \left(\tilde{\theta}_k - \theta_k\right)^2}_{\text{surrogate loss}}$$

Figure 7. Synaptic Regularization Model (top)
This equation shows the addition of the synaptic regularization to the standard loss function used when optimizing neural networks. The current parameter values (θ) are compared with "ideal" parameters ($\tilde{\theta}$) for each task k, and then scaled by their importance (Ω). This determines the strength of the synaptic regularization (Zenke et. al, arXiv, 2017).

$$C_k \frac{\partial U_k}{\partial t} = g_{k-1,k}(U_{k-1} - U_k) + g_{k,k+1}(U_{k+1} - U_k)$$

Figure 8. Cascade Model (bottom)
This equation models the transfer of synaptic value (U) between elements in the cascade network. The first element in the cascade, the synaptic parameter itself, feeds value into the network, and then receives it in turn based on the connections (g) and element sizes (C). This allows the network to maintain the strength of important parameters (Benna and Fusi, Nature, 2016).

The Future

We have explored several methods to make our network more efficacious in maintaining short-term and long-term memories, for solving the issue of catastrophic forgetting. Instead of using dendrites naively, an architecture should be implemented so that the network itself engages dendrites appropriately for different tasks. Using disinhibition pathways and motif circuits encourage this type of structure, and allow the network to choose effective activation trajectories. These structures can then be held fast through weight stabilization, allowing networks to successively learn new tasks.

For our network, we have limited the scope of synaptic plasticity implementation to long-term memory consolidation models, but these models lack computational efficiency. We hope to move away from such models to make the process more closely resemble neurobiological processes. We are also considering teaching the network using reinforcement learning (RL). RL is a machine learning framework highly analogous to human learning that places ANNs in interactive environments instead of simply showing them datasets. Using RL alongside the continual learning characteristics of our model will allow us to more intuitively develop possible solutions to catastrophic forgetting, as a result of the similarity to human learning.

References

Benna, Marcus K, and Stefano Fusi. "Computational principles of synaptic memory consolidation." *Nature Neuroscience*, vol. 19, no. 12, 3 Dec. 2016, pp. 1697-1706., doi:10.1038/nn.4401.
Cichon, Joseph, and Wen-Biao Gan. "Branch-Specific dendritic Ca2 spikes cause persistent synaptic plasticity." *Nature Communications*, vol. 520, 30 Mar. 2015, pp. 180-185., doi:10.1038/nature14251.
Yang, Guangyu Robert, et al. "A dendritic disinhibitory circuit mechanism for pathway-Specific gating." *Nature Communications*, vol. 7, 20 Sept. 2016, doi:10.1038/ncomms12815.
Zenke, Friedemann, et al. "Continual learning through synaptic intelligence." 12 June 2017, arxiv.org/abs/1703.04200. arXiv:1703.04200v3

THE UNIVERSITY OF CHICAGO