SIMULATED SLEEP HELPS TO GENERALIZE KNOWLEDGE IN A SPIKING NETWORK TRAINED WITH SPIKE-TIMING DEPENDENT PLASTICITY

Timothy Tadros,
Neurosciences Graduate Program
UCSD

Giri P Krishnan,
Department of Medicine
UCSD

Maxim Bazhenov
Department of Medicine
UCSD

ABSTRACT

Artificial neural networks are known to generalize poorly to new examples; although they excel at representing data observed in the training set, they are unable to represent data drawn from different distributions. In the mammalian brain, evidence suggests that sleep promotes generalization of learned examples. To address the validity of this hypothesis, we utilized a previously developed spiking neural network trained with spike-timing dependent plasticity (STDP) to perform digit classification on the MNIST dataset. We demonstrate that incorporating an offline, sleep-like period after training leads to generalization and robustness to novel inputs.

1 INTRODUCTION

Although artificial neural networks (ANNs) can rival human performance on various tasks, ranging from complex games (Silver et al. (2016)) to image classification (Krizhevsky et al. (2012)), they have been shown to underperform when the testing data differs in specific ways even by a small amount from the training data (Geirhos et al. (2018)). This lack of generalization presents several issues when ANNs are utilized in the real world. Primarily, ANNs are often trained on refined datasets of images designed to best capture the image content, whereas in real-world scenarios, they may be tested on disturbed or noisy inputs, not observed during training. Creating more robust neural networks will pave the way forward for using these promising neuro-inspired architectures in the real-world.

It has been hypothesized that in the mammalian brain sleep helps to create generalized representations of the information learned during the awake state (Stickgold & Walker (2013)). Sleep has been identified as being critical for memory consolidation - a process of converting recent memories into long-tern storage (Rasch & Born (2013)). During sleep, there is reactivation of neurons involved in previously learned activity (Stickgold (2005)) and this reactivation is likely to invoke the same spatio-temporal pattern as the pattern observed during training in the awake state (Wilson & McNaughton, 1994). Sleep reactivation, or replay, serves to strengthen synapses involved in a learned task through spike-timing dependent plasticity rules (STDP). Sleep, through STDP, can increase a subject's ability to form logical connections between memories and to generalize knowledge learned during the awake state (Payne et al. (2009)).

Similarly, research suggests that sleep can help extract the gist of a task by strengthening connections pertinent to all memories while weakening connections, through synaptic downscaling, relevant to a single, spurious memory (Lewis & Durrant (2011)). This body of neuroscience work suggests that a sleep-like phase applied in training neural networks may allow for gist extraction of the training data, leading to increased generalization and robustness to the underlying distribution of the training data. Our hypothesis is that sleep could aid in increasing a neural network's generalization performance by reducing the impact that small additions of noise can have on the network's classification accuracy.

2 NETWORK ARCHITECTURE AND SIMULATED SLEEP

To address the validity of this hypothesis, we utilized a spiking neural network trained with STDP previously proposed to perform digit cassification on the MNIST dataset (see Diehl & Cook (2015) for details). The MNIST dataset represents a simple task for artificial intelligence whereby the network must learn to classify grayscale images of handwritten digits (LeCun et al. (1998)). The spiking network consists of 3 layers: an input layer, an excitatory middle layer and an inhibitory layer. Neurons in the input layer receive input proportional to the intensity of each pixel in the MNIST images. The input layer projects to a layer of excitatory neurons with an all-to-all connectivity matrix and the weights of these connections are updated by an STDP rule. In addition, the excitatory layer projects to and receives lateral inhibition (which promotes competition amongst neurons) from the inhibitory layer. The neurons within each layer are governed by leaky-integrate-and-fire dynamics. Additionally, each neuron in the excitatory layer has a threshold parameter which is governed by a homeostatic rule to ensure balanced activity (see Figure 1 for a summary of the architecture).

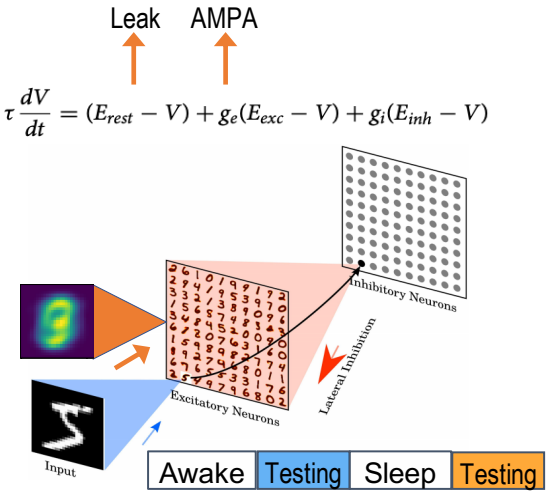

Figure 1: Network architecture and sleep changes (adapted from Diehl & Cook (2015)). Changes to network during sleep include presenting the average image and increasing leak and AMPA currents.

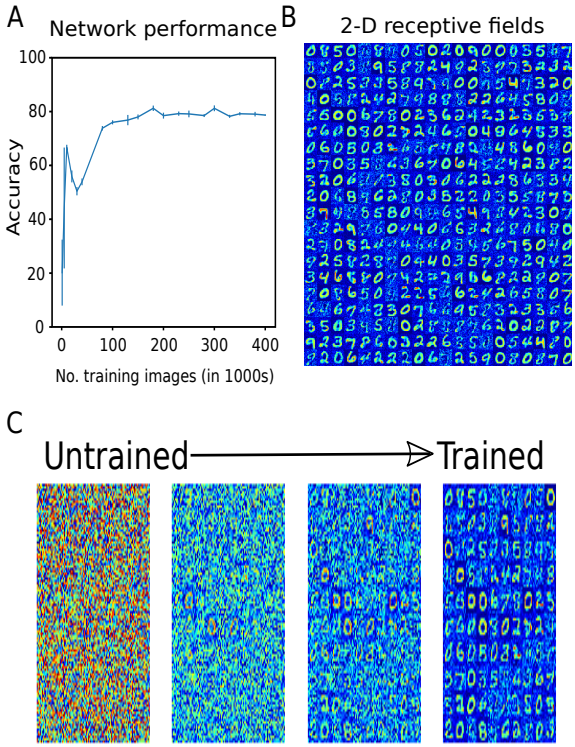

Figure 2: Network is able to learn the digit classification task. A. Test accuracy as a function of number of training images seen. B. Receptive fields of neurons in the excitatory layer form into 2-D spatial filters. C. Network receptive fields at different stages in awake (before sleep) training.

As the network is presented with more images, the network is able to classify a greater percentage of images correctly, by modifying weights from the input to the excitatory layer to compute 2-dimensional spatial filters of the MNIST digits (see Figure 2A-C). 2-D receptive fields are computed by reshaping the weights connecting to a single neuron in the excitatory layer into the same dimension as the input images. Then, these 2-D receptive fields are aligned in order to visualize all receptive fields learned by the network.

While this network can learn the task, it only reaches high levels of performance ($>$ 80%) after training on more than 100,000 images. We took a partially trained network (between 20 and 80% of the full training image set) and applied a sleep-like phase after the learning phase. During simulated sleep, we modified the intrinsic and synaptic currents to mimic changes in neuromodulator levels, while presenting noisy Poisson input based on the statistics of the MNIST input (see Figure 1 for dynamical equation updates). These changes capture cellular and synaptic changes which occur during stage 3 sleep, and result in an increase in activity, mirroring the "up-state" of slow-wave sleep (Wei et al. (2016)). During sleep, the same STDP and threshold updating rules are used. We compared performance before (awake) and after sleep by computing the

classification of the network on different testing images. Classification is done by assigning each neuron in the excitatory layer a label (0-9) based on which set of digits produce the maximum mean firing activity in that neuron. Networks are tested from various random initializations ($n = 5$) to measure variability of the training and sleep phases.

3 RESULTS

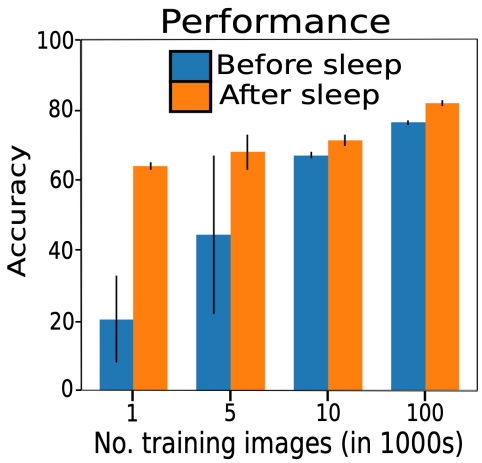

Figure 3: Sleep improves classification performance. Accuracy before (blue) and after sleep (orange) for different training levels, tested on 5 different network initializations.

Sleep improves performance on networks trained with small training dataset. After training in the awake state, the network is able to accurately classify the MNIST digits. However, at different levels of awake training (measured by how many images in the training set the network has observed), incorporating an off-line sleep period after awake training notably increases classification accuracy on a novel test set (Figure 3). Most notably, at very small levels of training (1000 images), the trained network classifies the test set with 20% accuracy. However, after a sleep-like period where noisy input is presented to the network, classification accuracy reaches 60%. This effect is pronounced even at higher levels of awake training, suggesting that a sleep-like period can promote one-shot learning and greater generalization of the task structure.

Sleep promotes increased generalization. As noted above, neural network-based classifiers often suffer from poor robustness. If a network is trained on intact, undistorted images, then the network will fail to classify distorted images, even if the distortions are not significant enough to affect human-level perception. To test the effect of sleep on a network's robustness, we added noise to the MNIST images, either by adding random Gaussian noise or applying a blur filter to the images (Figure 4A). We found that the network after undergoing a sleep period is able to classify more images correctly even as the images are further distorted (Figure 4B). These results mirror the results from biology which suggest that sleep can help a subject extract the gist of a task and generalize knowledge learned during a waking period.

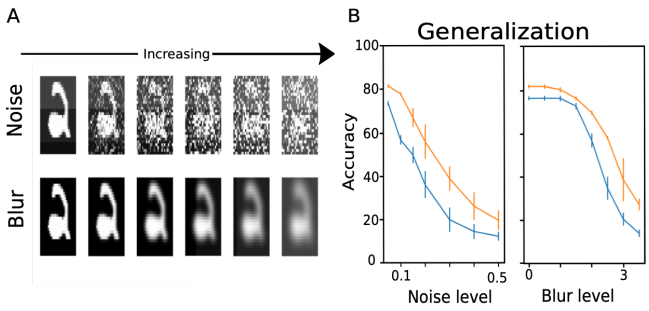

Figure 4: Sleep improves generalization to noisy images. A) Examples of noisy and blurred images. B) Accuracy as a function of noise added before and after sleep.

Sleep prunes task-irrelevant neurons from the network. We next analyzed which component of the network, changing neuronal thresholds or synaptic plasticity, contributed the most to the accuracy increase after sleep. We observed that most neurons experienced an increase in their thresholds due to the constant activity presented during sleep and the homeostatic rule used to change thresholds (Figure 5A). However, neurons with well-formed 2-D receptive fields were qualitatively more likely to have decreasing thresholds after sleep (Figure 5C). Oppositely, neurons with noisy 2-D receptive fields were more likely to experience an increase in their firing thresholds following sleep (Figure 5D). We quantified this phenomenon by looking at the average neighborhood pixel variance using $3x3$ pixel squares. Receptive fields with low neighborhood pixel variance are likely to be more refined since there is little variability between neighboring pixels. In contrast, noisy receptive fields should have high neighborhood pixel variance. There was a significant correlation between neighborhood pixel

variance and threshold change (Figure 5B), suggesting the hypothesis that sleep improves performance and robustness by pruning task-irrelevant neurons from the network by increasing their firing thresholds.

We confirmed this as the main source of improvement after sleep by testing the network in four conditions: using either before- or after-sleep weights and before- or after-sleep thresholds. The largest performance increase was observed when after-sleep thresholds were used (no significant difference between normal sleep and only using after-sleep thresholds, $p = 0.22$). However, when pre-sleep thresholds were used along with the STDP changes that resulted from sleep, performance did not improve significantly. This suggests that in the default network architecture, sleep improves performance by altering the thresholds in a manner in which task-specific neurons can respond more acutely (because of reduced thresholds) to the images presented during testing.

Finally, we analyzed the effect of reducing inhibition and fixing the thresholds during sleep in order to determine the role of synaptic plasticity changes that occur during sleep on generalization. We were able to see the same performance increase after sleep by reducing inhibition in the network, as competition between neurons was reduced (normal sleep vs. only STDP changes, $p = 0.85$). We explored the synaptic weight changes during sleep and uncovered two main principles (not shown here). First, in neurons with well-formed receptive fields, there is very little synaptic weight change after sleep. Second, in neurons with task-irrelevant receptive fields, there is an overall synaptic downscaling of connections, mirroring the results from the threshold analysis above. Overall, these results support the role of sleep in memory consolidation and generalization of knowledge learned during the waking state. Moreover, this line of work supports the synaptic homeostasis hypothesis of sleep which suggests that slow-wave sleep improves performance by downscaling synaptic weights (Tononi & Cirelli (2006)).

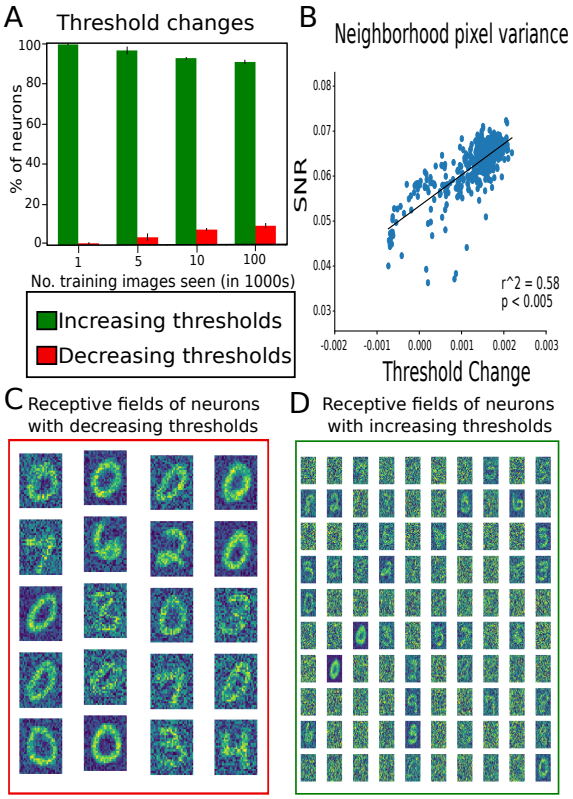

Figure 5: Task-irrelevant neurons fire less after sleep. A. Number of neurons that have increasing or decreasing thresholds after sleep at various stages of training. B. Average neighborhood pixel variance vs. threshold change during sleep. C-D. Example receptive fields for neurons with decreasing (left) or increasing (right) thresholds

4 Conclusions

In this study, we applied an off-like sleep-like phase to the training phase of a spiking network trained to perform the MNIST digit classification task. We found that after any amount of awake training, adding a sleep phase, where noisy Poisson input is passed through the network and activity is elevated, can increase the classification accuracy on a novel test set. Similarly, the network after sleep is able to respond to more diverse representations of the image set, classifying noisy and blurred images more accurately than before sleep. These results mirror work in biology which has shown that sleep can help extract the gist of a task and generalize knowledge learned during the awake state (Stickgold & Walker (2013)). Additionally, these results lend support to the synaptic homeostasis hypothesis which suggests that sleep down-scales synaptic weights to make efficient use of brain space in a energy-conserving manner (Tononi & Cirelli (2006)). Our experiments suggests that down-scaling of synaptic activity is likely constrained to task-irrelevant neurons, thereby containing the representation of the task to a subset of neurons.

REFERENCES

Peter U Diehl and Matthew Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in computational neuroscience*, 9:99, 2015.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 7538–7550, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Penelope A Lewis and Simon J Durrant. Overlapping memory replay during sleep builds cognitive schemata. *Trends in cognitive sciences*, 15(8):343–351, 2011.

Jessica D Payne, Daniel L Schacter, Ruth E Propper, Li-Wen Huang, Erin J Wamsley, Matthew A Tucker, Matthew P Walker, and Robert Stickgold. The role of sleep in false memory formation. *Neurobiology of learning and memory*, 92(3):327–334, 2009.

Björn Rasch and Jan Born. About sleep's role in memory. *Physiological reviews*, 93(2):681–766, 2013.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Robert Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272, 2005.

Robert Stickgold and Matthew P Walker. Sleep-dependent memory triage: evolving generalization through selective processing. *Nature neuroscience*, 16(2):139, 2013.

Giulio Tononi and Chiara Cirelli. Sleep function and synaptic homeostasis. *Sleep medicine reviews*, 10(1):49–62, 2006.

Yina Wei, Giri P Krishnan, and Maxim Bazhenov. Synaptic mechanisms of memory consolidation during sleep slow oscillations. *Journal of Neuroscience*, 36(15):4231–4247, 2016.

Matthew A Wilson and Bruce L McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994.