Exploration of Working Memory Dynamics Through Deep Learning Neural Networks

Christian Wawrzonek - 2016

Advisor: Timothy Buschman

Collaboration: Pavlos Kollias, Matt Panichello

Senior Thesis 2016 Proposal: October 6th, 2015

1 Motivation and Goal

How exactly do populations of neurons encode information over very short time scales? Given extensive training, neurons are able to change the weighted connections between them in order to encode information. However, very short timescales of only a few seconds are far too short to change neural weights. However, humans and other animals are clearly able to maintain small amounts of information presented over very short timescales and reproduce this information later. This is the problem of working memory, and very little is known about the underlying neural dynamics. The goal of this project is to train a neural network on complex working memory tasks and analyze resulting neural patterns.

2 Background

Previous models of working memory, such as bump attractor networks, have been very successful at modeling working memory [5]. However, these models only succeed by manually creating a rigid structure for training and are unlikely to be the strategy which occurs in natural learning. A major aim of this model is to train a network with very few constraints on encoding paradigms and analyze the strategies developed by the network. Last semester, my spring independent work consisted of engineering and studying the dynamics of a deep learning neural network modeling a delayed saccade task, a relatively simple spatial working memory task coding the locations of stimuli in 2-D space. This semester, the aim is to advance this model significantly, adding new dimensions to the task and providing more thorough analysis of the neural coding paradigms that arise.

A major new dimension that we wish to add to the model and explore is the concept of retrocuing. This phenomenon occurs during working memory tasks in monkeys that are trained to store multiple objects in working memory and cued to report only a subset of those objects. When cued to report an object, there is some baseline success rate with which they perform. In this case, it is assumed that any errors in reporting are simply a degradation of the stored information. However, if a monkey is re-presented with the cue (not the stimulus) after a delay period, reporting accuracy increases. Where once it was assumed that this information was simply lost, it is now hypothesized that there is simply a competing representation of multiple objects in working memory. Once re-presented with the cue, the representation of the undesired object collapses, allowing the correct object representation to reacquire attentional resources.

Effective training of this phenomenon in a neural network would potentially offer significant insight into the dynamics of working memory.

3 Approach

This new model will follow a similar development strategy to the model last semester. It will be a deep learning neural network, utilizing Hessian Free Optimization to bypass previous problems with training in deep learning models [1, 2]. We will attempt to employ new training strategies in order to train more complex tasks, such as training the network with progressively training with increasing difficulty in an attempt to guide learning.

There are many suggestions I wrote about last semester to improve the training and performance of our simple model, including exploring different local connectivity heuristics and further optimization of hyper parameters.

4 Plan of Action

Given that I already have a working framework from which to build on, the first steps will be to transform the previous network to handle more complex decision and working memory tasks. Once the model can reliably and robustly model these new aspects of working memory, I can begin perturbing the network with task variations such as noise, randomness, and of course, retrocuing. These and other variations, coupled with rigorous mathematical analysis, will hopefully offer novel insight into working memory encoding paradigms.

Given enough time and steady progress, there should also be enough time to explore new architectures for the network beyond our simplistic schema.

5 Evaluation

Obviously, the most fundamental requirement for this project is that it reliably and robustly model more complex spatial working memory tasks. Once this goal is achieved, perturbations and variations in working memory tasks should produce results comparable to behavior results in observed in monkey trials. Another major metric of success will be novel insights made through mathematical analysis of neural network connection and activation patterns. Last semester, analysis of network paradigms was shallow and underwhelming given time constraints, consisting primarily of simple principle component analysis, and I expect significantly more rigorous network analysis with this project.

6 References

References

[1] S. Hachreiler, ‘The Vanishing Gradient Problem During Leaning Recurrent Neural Nets and Problem Solutions,” Inr. J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 6, no. 2, pp. 107-1 16. 1998.

[2] Martens, J. (2010). Deep learning via Hessian-free optimization. In Proceedings of the 27th International Conference on Machine Learning.

[3] Miller P, Brody CD, Romo R, Wang XJ. 2003. A recurrent network model of somatosensory parametric working memory in the prefrontal cortex. Cereb Cortex. 13:1208–1218.

[4] Song, S., Sjöström, P. J., Reigl, M., Nelson, S. and Chklovskii, D. B. Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol. 3, e68 (2005).

[5] Wimmer, K., Nykamp, D.Q., Constantinidis, C. Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci. 17, 431–439 (2014).