CS230

Modeling Visually Guided Behavior of Mice in Naturalistic Virtual Reality

Linnie Jiang
Department of Neurobiology
Stanford University
linniej@stanford.edu

Javier C. Weddington
Department of Neurobiology
Stanford University
javiercw@stanford.edu

Abstract

Animals efficiently make decisions from high-level sensory observations. However, it is unknown how the visual system represents the sensory environment to coordinate observations and actions. Our study of the brain has not included an understanding of how visual cortex responds to naturalistic stimuli. Here, we present a novel virtual reality system to understand the mechanisms of cortical representations in mammals using a deep neural network. We find that using a deep neural network yields better performance than using a simpler, traditional linear nonlinear (LN) model. Our results indicate that the mechanism by which the deep neural net performs well is biologically relevant.

1 Introduction

Why use mice to study vision? In short, there are far more similarities found in visual processing than differences. Orientation selective cells, receptive fields, higher visual processing, and visual discrimination among many others properties are conserved in lower and higher mammals[47]. Despite mice being non-foveated and having severely diminished visual acuity in comparison to primates, the computational properties of vision are largely conserved. The relative size of the mouse and its cortex confer the benefit of studying the mouse visual system in naturalistic contexts. Full immersion in virtual reality and access to more areas of the visual cortex (with fewer, less intensive surgeries) makes the mouse a great model organism for studying vision. Encoding models of the mouse visual cortex

How do brains represent real-world, high dimensional visual inputs? Visual neuroscience typically breaks down stimulus features into components such as frequency, wavelength, orientation, and position. However, it is unknown how much this decomposed version of a visual stimulus can explain responses to natural visual scenes. This thesis proposal seeks to contribute to our understanding of the neural underpinnings of sensory processing by studying autonomous behavior of mice in fully immersed virtual reality. In this context, higher level properties of object features may be better descriptions of visual responses than lower level ones. An obvious example for this possibility are the behaviors elicited in foraging, predator and prey behaviors. Object features like size, color, self-generated motion vs externally-generated motion, category, and distance from the animal are likely more useful for ethological behaviors. In this thesis, I propose to investigate how natural videos drive responses in the mouse primary visual cortex.

How does spontaneous behavior affect visual representations? Another visual operation likely to receive increased attention in freely moving animals is the computation of object structure and scene layout through the parallax afforded by self-directed motion. Behavioral components like self-driven

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

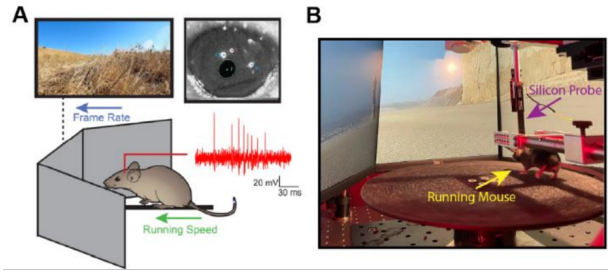

Figure 1: A. Schematic of nvr rig. A head-mounted mouse runs on a wheel. The velocity of the wheel dynamically changes the frame-rate of a naturalistic stimulus (upper left) surrounding the animal on three screens, providing closed-loop locomotor-dependent optic flow. During behavioral sessions, high-speed videography tracks the pupil, allowing us to reconstruct the animal's center-of-gaze on the monitors. Additionally, the brain is available for optical or electrophysiological activity monitoring (red trace, raw voltage from silicon probe reveals high-amplitude action potentials). B. Image of a mouse during a recording session running in the "beach" environment.

displacement in a fully immersed virtual environment are understudied and underappreciated among visual neuroscientists. The spatiotemporal continuity between frames from animals moving through space is a fundamental and evolutionarily component of visual perception. Even rodents, with their relatively poor visual acuity, use self-directed displacement glean information about stable, moving, and harmful objects in their environment. There are many more examples across species that suggest there are suppressive mechanisms for self vs externally generated motion. Monkey head movements (vestibulo-ocular movements) Sharks, skates, and rays all have mechanisms of selectively suppressing the self-generated sensory effects of behaviors. Some of these mechanisms of suppression are through negative prediction of the upcoming signal. Another mechanism for differentiating self from external movement is by integrating information from an efference copy of motor commands and vestibular motion all of which signal self-motion to the organism. Lastly saccadic suppression occurs right before and during an eye movement. The mechanism of self-guided visual behavior has potential to reveal new phenomenology about visual processing in animals. Despite there being clear incentive to incorporate locomotion, most visual experimentalists have investigated head-fixed animals, which mostly eliminates this important behavioral component. Although motor input to the mouse visual cortex has been identified, it is largely unclear whether the effects of motor input on the neural code conform to one of these concepts. To investigate this question, I will study the effect of locomotion on responses in the visual cortex by looking at responses in closed and open-loop natural environments. These two conditions will allow us to dissociate visual responses to parallax and other locomotor behaviors.

Eye movements are often overlooked in experiments with rodents. Mice, unlike primates, do not have a fovea, and have not been trained to fixate on a target, adding a significant challenge when modeling visual computations in the brain of awake, behaving mice. The NVR system bridges the gap by employing high-speed videography of the animal's pupils, which can be used post-hoc with infrared fiducials placed on the stimulus monitors, as a means of determining the center of gaze. This will allow us to determine much more accurately the retinal image in order to create more accurate encoding models. In addition, our novel eye tracking system will allow us to determine the independent contributions of eye movements to visual encoding in mice. The Baccus Lab has created a three layer neural network model that accurately captures retinal responses, nearing the fundamental limit imposed by neural variability[11]. These models capture a wide range of retinal computations, and have internal units that correspond to retina interneuron recordings. As such, these models can potentially be used to discover new computations, and generate specific new hypotheses as to the neural basis of those computations[12]. We have constructed and validated a Naturalistic Virtual Reality (NVR) system that will be used to analyze cortical responses to natural stimuli during behavior. In our system, a head-fixed mouse is allowed to run on a treadmill which controls the presentation of ethologically natural stimuli, representing forward motion along a linear track. Stimuli were captured as videos taken from the height of a mouse. To sample different natural settings, we recorded videos in five local environments such as wooded hike trails, city streets, Stanford streets, the beach, and open green meadows. Stimuli are presented during neural recordings In two conditions 1) an open loop (OL) condition in which visual stimuli were repeated without feedback from the treadmill. 2)

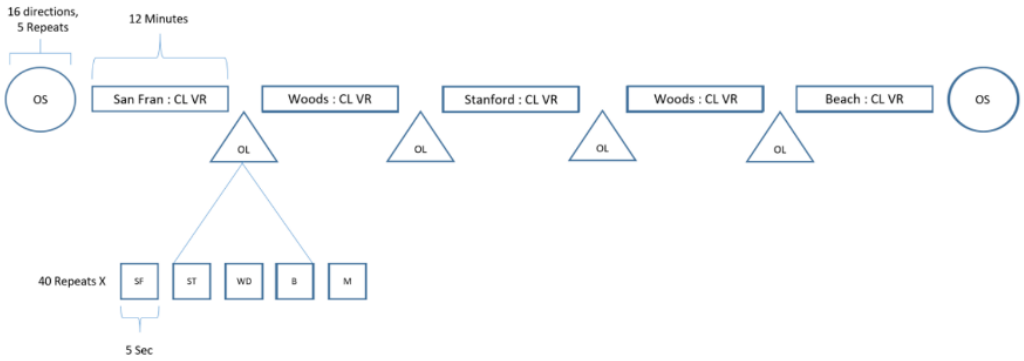

Figure 2: Stimulus presentation schedule. CL: Closed Loop; VR: Virtual Reality; OS: Orientation Selective stimulus; OL: Open Loop repeats; San Fran/SF: streets of San Francisco; ST: Stanford; WD: Woods; B: Beach; M: Meadow

a closed loop VR (CL VR) condition in which motion of the treadmill advanced the frames of the movie. Although the animal was head fixed, it was free to run and move its eyes. In addition, drifting grating stimuli as shown in Figure 1 spaced by 5 seconds of gray screen. Simultaneously, we recorded running, eye movements, and the stimulus frames shown.

2 Related work

We have read neuroscience literature to understand traditional approaches to recording neural activity in the mouse cortex differ from our data. We have studied cortical anatomy to draw inspiration for our model. We will also build upon work done in a previous paper in which a 3-layer CNN was used to model the retina. Because our model will be a more extended one of visual cortex as well (downstream processing centers), we will envelop that model and add to its implementation.

3 Dataset and Features

Our neural data was collected from the mouse primary visual cortex using high density 64 channel probes spanning 1000 micrometers to record from all 6 cortical layers. The sampling rate of our amplifier was 30 KHz. We sorted the dataset to select cells from the recording using software created by the Baccus Lab. The output of the sorted dataset are spike times for which the neuron was active during the experiment. We binned those spike times at 10ms parameterizing a firing rate for the cells in our experiment. We use these binned spike times to train from the neural data collected in experiments to train our model.

The original movie shown to the retina of the mouse was a $508 \times 638 \times 3 \times n_{frames}$ (RGB). For our model inputs, we grayscaled and downsampled the movies to make them $40 \times 80 \times 1 \times n_{frames}$. We trained our model on closed loop virtual reality where the mice's spontaneous activity advances the frames of the movie. We tested our model on the open loop repeats where mouse locomotion did not drive stimulus changes. The videos for the closed and open loop stimuli were taken from identical distributions. However, the frames in the training dataset are not present in the test set. The justification for downsampling and grayscaling the inputs to the model are 1) to have model inputs that will allow the model to train faster and 2) the visual acuity of the mouse retina is poor and a deprecated image could be a fair estimation of image statistics the mouse observed.

4 Methods

Since our goal is to learn how complex a model is needed to capture responses of the mouse primary visual cortex without behavioral variables, we will compare simpler models (see Figure 3 for the Linear Nonlinear model) to the CNN models shown to capture retinal responses to natural images. We randomly initialized a hyperparameter sweep through L1 and L2 regularization, dropout, and learning

rate for the a Linear Nonlinear (LN) model and a Batch Normalized Convolutional Neural Network (BNCNN) as shown in Figure 4. We gave input the size of $(HISTORY, batch\_size, pix\_x, pix\_y)$ and got a vector of size 22 with the activations of the final layer of the network in both cases.

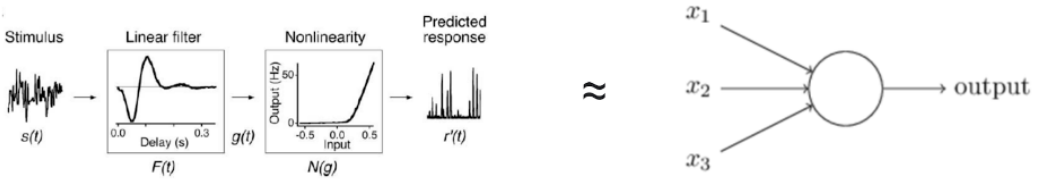

Figure 3: Model architecture of the Linear Nonlinear (LN) model: stimulus is convolved with a linear filter, and the output is passed through a nonlinearity that removes negative activity. This mimics a one-layer neural network.

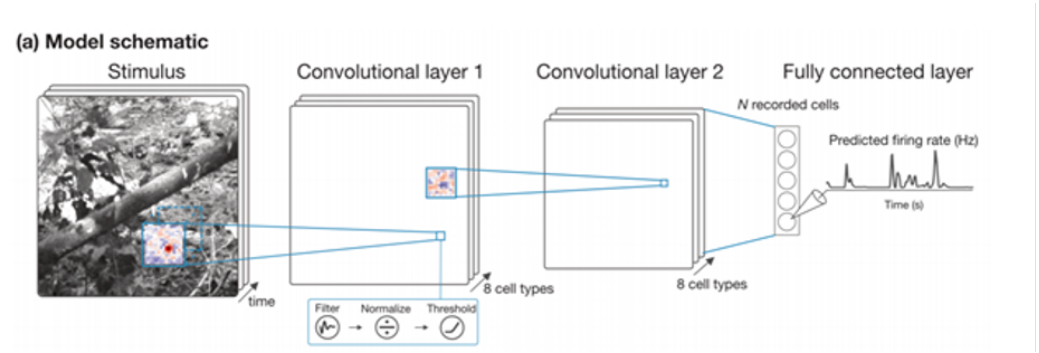

Figure 4: Model architecture of the BNCNN. Two convolutional layers, each with 8 channels, the first with a kernel size of 15, and the second with a kernel size of 11, precede a fully connected layer of 22 units.

5 Experiments/Results/Discussion

Mice running on the wheel might also contribute to some of the variance that we see. We did a grid search as well as random initialization of parameters for the learning rate, L1 regularization of the fully connected layer weights, and L2 regularization of the Conv2D layer weights, choosing values that gave the best performance. The hyperparameters that gave the best performance were 0.001 for the learning rate, 0.01 for the L1 regularization, and 0.001 for the L2 regularization. We also normalized the activations of the final layer with the hyperparameter 0.001. Our mini-batch size was 512 because that was a reasonable size for the data to be between SGD and batch gradient descent, as discussed in lecture. Our primary metric for evaluation was to compare the fully connected layer activations to the ground-truth data for those 22 recorded neurons. We did this comparison by computing the Pearson r correlation coefficient between the vector of activations of the final layer of each model and the vector of neural data.

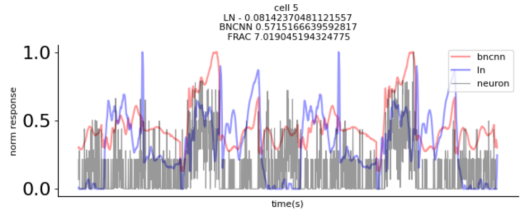

Figure 5: One peri-stimulus time histogram (PSTH), for cell 5 of 22, showing performance via Pearson correlation of the BNCNN and LN model to ground truth data. Fraction describes the ratio of BNCNN Pearson r to LN model Pearson r.

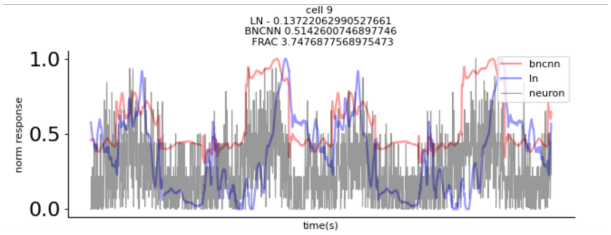

Figure 6: Like Figure 5, PSTH but for cell 9 of 22.

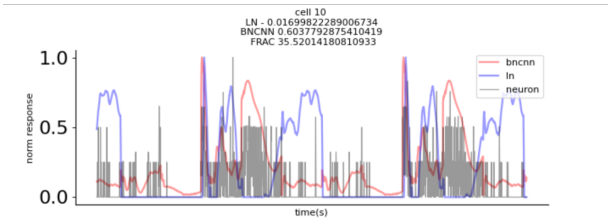

Figure 7: Like Figure 5, PSTH but for cell 10 of 22.

From Figures 5, 6, and 7, we can see that the BNCNN outperforms the LN model, having a 0.4-0.6 Pearson r value rather than a 0.01-0.1 Pearson r value. Upon investigating within the model, we see that layer 1 has Gabor-like filters, which is similar to biological features. Layer 2 (the rest of Figure 8) shows filters like receptive fields.

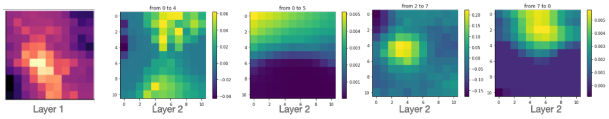

Figure 8: Left: Example of Gabor-like filters in layer 1. Right four images: example of receptive field-like filters in layer 2.

6 Conclusion/Future Work

These results show that 1) the BNCNN does a remarkable job of correlating to actual neural responses, and 2) it does so by approximating biological mechanisms. Tracking the eye may not be very useful for small changes in pupil position. Because of the poor visual acuity of the mouse, it is possible that if we were to do a Gaussian filter over the inputs of the image, we could perhaps improve our performance in fitting neural data. Alternatively, this model up to the last layer would be identical to one trained without eye movement, as saccades could be modelled as picking different positions in the last-layer activations. This means the receptive fields might be correct, and the linear layer would pool over positions and result in a lower-fidelity response accuracy. Based on intuition and previous work, we initialized our model the way we did and found it to be better than models with different parameters, but a more exhaustive search over a number of different combinations of number-of-layers, number-of-channels, and channels-sizes would be useful in the future.

7 Contributions

JW collected the data and preprocessed the data with the help of Joshua Melander (JM). JM built the LN model. LJ and JM built the BNCNN model. LJ and JW wrote the paper and recorded the video.

References

[1] Volodymyr Mnih et al. Playing Atari with Deep Reinforcement Learning. 2013. *arXiv*: 1312. 5602 [cs.LG].

[2] David Silver et al. "Mastering the game of Go without human knowledge". *Nature* 550.7676 (Oct. 2017), pp. 354–359. DOI: 10.1038/nature24270. URL: https://doi.org/10. 1038/nature24270.

[3] D. L. K. Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". *Proceedings of the National Academy of Sciences* 111.23 (May 2014), pp. 8619–8624. DOI: 10.1073/pnas.1403112111. URL: https://doi.org/10. 1073/pnas.1403112111.

[4] Demis Hassabis et al. "Neuroscience-Inspired Artificial Intelligence". *Neuron* 95.2 (July 2017), pp. 245–258. DOI: 10.1016/j.neuron.2017.06.011. URL: https://doi.org/ 10.1016/j.neuron.2017.06.011.

[5] Ferran Alet et al. Meta-learning curiosity algorithms. 2020. *arXiv*: 2003.05325 [cs.LG].

[6] Malcolm G. Campbell et al. "Principles governing the integration of landmark and self-motion cues in entorhinal cortical codes for navigation". *Nature Neuroscience* 21.8 (July 2018), pp. 1096–1106. DOI: 10.1038/s41593-018-0189-y. URL: https://doi.org/10.1038/ s41593-018-0189-y.

[7] Agrim Gupta et al. Embodied Intelligence via Learning and Evolution. 2021. *arXiv*: 2102. 02202 [cs.LG].

[8] Deepak Pathak et al. Curiosity-driven Exploration by Self-supervised Prediction. 2017. *arXiv*: 1705.05363 [cs.LG].

[9] Deepak Pathak et al. Curiosity-driven Exploration by Self-supervised Prediction. 2017. *arXiv*: 1705.05363 [cs.LG].

[10] Mark H. Plitt and Lisa M. Giocomo. "Experience dependent contextual codes in the hippocampus". *biorXiv* (Dec. 2019). DOI: 10.1101/864090. URL: https://doi.org/10.1101/864090.

[11] Lane McIntosh et al. "Deep Learning Models of the Retinal Response to Natural Scenes". *NIPS 2016*. URL: https://papers.nips.cc/paper/2016/hash/a1d33d0dfec820b41b54430b50e96b5c-Abstract.html

[12] Alexander Schutz et al. "Eye movements and perception: a selective review." *Journal of Vision* (Sept. 2011). URL: https://jov.arvojournals.org/article.aspx?articleid=2191910

[13] Andrew D. Huberman and Cristopher M. Niell. "What Can Mice Tell Us about How Vision Works?" Trends in Neurosciences, vol. 34, no. 9, 2011, pp. 464–473., doi:10.1016/j.tins.2011.07.002.

[14] Niru Maheswaranathan et al. "The dynamic neural code of the retina for natural scenes ". *biorXiv* (Dec 2019). URL: https://doi.org/10.1101/340943

[15] PyTorch library. URL: https://pytorch.org/