Perceptual Decision Making Investigated via Sparse Decoding of a Spiking Neuron Model of V1

Jianing Shi  
Biomedical Engineering  
Columbia University  
js2615@columbia.edu

Jim Wieland  
Biomedical Engineering  
Columbia University  
djw21@columbia.edu

R. Theodore Smith  
Ophthalmology  
Columbia University  
rts1@columbia.edu

Paul Sajda  
Biomedical Engineering  
Columbia University  
ps629@columbia.edu

Abstract—Recent empirical evidence supports the hypothesis that invariant visual object recognition might result from non-linear encoding of the visual input followed by linear decoding [1]. This hypothesis has received theoretical support through the development of neural network architectures which are based on a non-linear encoding of the input via recurrent network dynamics followed by a linear decoder [2], [3]. In this paper we consider such an architecture in which the visual input is non-linearly encoded by a biologically realistic spiking model of V1, and mapped to a perceptual decision via a sparse linear decoder. Novel is that we 1) utilize a large-scale conductance based spiking neuron model of V1 which has been well-characterized in terms of classical and extra-classical response properties, and 2) use the model to investigate decoding over a large population of neurons. We compare decoding performance of the model system to human performance by comparing neurometric and psychometric curves.

I. INTRODUCTION

An open question for the visual neuroscience, computer vision and machine learning communities is “how does the human visual system achieve robust object recognition in the face of clutter, noise and uncertainty”? At the core of this question is the way in which the visual system might identify invariances in the visual world which it can exploit. One hypothesis is that the visual system remaps the visual input to a high dimensional space in which recognition/classification can be done via hyperplanes. Several models developed based on such a hypothesis have shown impressive results in terms of recognition performance. Feedforward architectures, such as HMAX [4], propose that a combination of linear and non-linear filtering operations, arranged in a hierarchical structure, can map the visual world into a high dimensional space where objects can be linearly decoded with high recognition accuracy. The relationship between such models and the visual system is that 1) the linear and non-linear operations can be loosely related to the response properties of simple and complex cells seen in primary visual cortex, and 2) the hierarchical cascade of the two operations can be seen as the cascade of activity from primary visual cortex up through extrastriate areas. Noticeably absent in such feedforward architecture is an recurrent connectivity and thus dynamics which is clearly prevalent in visual cortex [5], [6]. The machine learning community has recognized that the combination of a simple linear decoder and a highly recurrent, non-linear network with fixed connections can provide universal computational capabilities. Termed liquid state machines (LSM) [2] and echo state networks (ESN) [3], these algorithms have been related to the brain’s neural microcircuitry, and potentially link the anatomy and physiology of the vast recurrent circuitry in cortex with general computation capabilities.

Perceptual decision making is a potential paradigm for investigating some of the invariances of visual recognition and tying these to the level of performance degradation that is a function of the decision difficulty. There has been substantial work investigating two alternative forced choice (2-AFC) tasks in which the goal is to discriminate one type of visual stimulus from another (e.g. motion direction, face vs car/house) in the presence of noise, clutter and/or varying uncertainty. Neural correlates of perceptual decision making were found in monkey experiments for a motion detection task [7], and recently in a reaction-time visual detection task [8]–[10]. In these studies, the authors constructed neurometric functions and compared them with psychometric curves, constructed from behavioral responses, in order to identify those neurons which may contain information directly related to the monkey’s decision. Recently, similar approaches have been used for human decision making, with neurometric functions constructed from electroencephalography (EEG) [11]. In this case, the authors found that EEG could be decoded using a linear decoder, specifically logistic regression, such that a subset of the resultant neurometric curves were indistinguishable from...
psychometric curves. This suggests that macroscopic neural signals, such as scalp potentials measured via EEG, contain substantial information about the stimulus and/or information pertaining to human decision making. Given that activity from individual neurons and macroscopic signals, such as scalp EEG, have been analyzed in this way, it begs the question of whether activity from large populations of neurons could be similarly analyzed and related to perceptual decisions.

In this paper we consider linear decoding of populations of spiking neurons, and investigate the decoding accuracy relative to human performance (see Fig. 1). We use a large-scale spiking neuron model of primary visual cortex (V1) as a way to generate large populations of spike trains and as a non-linear encoder of the visual stimulus. We employ a linear decoder that exploits the spatial-temporal dynamics generated by the V1 model neurons. Specifically we use a $\ell_1$-regularized logistic regression model and study the decoding accuracy with respect to the V1 circuitry.

II. MATERIALS AND METHODS

A. Perceptual Decision Making Paradigm

We used a well-studied 2-AFC face vs. car discrimination, where the stimuli consisted of a set of 12 face (Max Plank Institute face database) and 12 car grey-scale images [11]. All images were $512 \times 512$ pixels, with 8 bits/pixel, and were equated in their magnitude spectrum, which was taken as the average magnitude spectrum across all images. In order to control the difficulty of the perceptual decision, the phase spectra of the images were manipulated using the weighted mean phase method [12]. This resulted in a set of images graded by phase coherence.

The sequence of images were presented to the model in a block design, where an image was flashed for 50 ms, followed by an inter-stimulus interval (ISI) of 200 ms. Ten phase coherences were used (0%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%). The image type and coherence level were selected at random during the presentation. All images were repeated 30 times in the simulation resulting in a total of 7200 trials. Fig. 2 summarizes the stimulus for the experiment.

For comparison we collected data from 10 human subjects for the same task (coherence 20%, 25%, 30%, 35%, 40%, 45%) in order to compare the discrimination from the model with human discrimination performance. For human psychophysics experiments, images were identical, as was the image duration, though the ISI differed relative to the model simulations. The ISI for human psychophysics experiments was longer and randomized between 1500-2000 ms.

B. V1 Model Summary

The recurrent V1 model has been described in [6], [13] and is well-characterized in terms of its classical and extra-classical response properties. It consists of a feedforward lateral geniculate nucleus (LGN) input and a recurrent spiking neuron network of primary visual cortex (V1).

In summary, the model consists of a layer of $N$ ($64 \times 64$) conductance based integrate-and-fire point neurons (one compartment), representing about a $2 \times 2$ mm$^2$ cortical sheet. 75% of the neurons are excitatory and 25% are inhibitory. Dynamic variables of each neuron are the membrane potential $v_i(t)$ and spike train $\mathcal{S}_i(t) = \sum_k \delta(t-t_{i,k})$, where $t$ is time and $t_{i,k}$ is the $k$th spike of the $i$th neuron ($i = 1, \ldots, N$),

$$C_i \frac{dv_i}{dt} = -g_{L,i}(v_i - v_L) - g_{E,i}(\mathcal{S}_E)(v_i - v_E) - g_{I,i}(\mathcal{S}_I)(v_i - v_I)$$

with LGN cells modeled as rectified space-time separable linear filters.

C. Sparse Decoding of V1 Model Activity

We use a linear decoder to map the spatio-temporal activity in the recurrent V1 model to a decision on whether the input stimulus is a face or a car. We employ a sparsity constraint on the decoder to control the dimension of the effective feature space. Sparse decoding is also consistent with previous research investigating decoding performance using multi-unit recording in monkeys [9].

Given the spike trains for a population of neurons, we construct a spatio-temporal word which is subsequently used as a feature vector for the linear decoder, see Fig. 3. Specifically, feature vector $x_i$ ($i = 1, \ldots, m$, where $m$ is the number of samples) is constructed by computing the mean in a time bin. This results in a feature vector whose dimension $n$ takes “the number of neurons in the model cortex” by “the number of time bins”. The class labels of each sample (either face or car) are given as $b \in \mathbb{R}^m$, taking values of $\{-1, +1\}$. We seek the solution of the following constrained minimization problem,

$$\arg \min_{w} \sum_{i=1}^{m} f((w^T x_i + v)b_i)/m + \lambda \|w\|_{\ell_1},$$

(1)

where $\lambda > 0$ is the regularization parameter, $w \in \mathbb{R}^n$ is the weights for the linear decoder, while $v \in \mathbb{R}$ is the offset. $f$ is the logistic loss function, defined by $f(z) = \log(1 + \exp(-z))$. Such a formulation essentially minimizes the average logistic loss, with a Lagrange multiplier for the $\ell_1$ norm of the weights.
The resulting linear decoder can be geometrically interpreted as a hyper-plane defined by $w^T x + v = 0$, which separates the classes of face and car.

The sparsity constraint is an important form of regularization. $\ell_1$-regularized logistic regression leads to sparse solution in the weights $w$, thus termed sparse logistic regression, and consequently serves as an avenue for feature selection. This approach allows us to use a rigorous form of optimization to identify neurons in the network which convey the most information about the discrimination. We apply an interior-point method to solve this optimization problem [14].

### III. RESULTS

We first investigated whether and how the activity of V1 model was selective to phase coherence levels in the stimulus. Fig. 4 shows the mean firing rate averaged over all 4096 cortical neurons and all 30 trials for different phase coherence levels. Clearly the average firing rate in the network, for face trials (Fig. 4a) or car trials (Fig. 4b), decreases as the coherence level decreases. This is consistent with the notion of V1 neurons being selective to oriented edge energies, which is reduced as the phase coherence decreases.

We then considered the difference in activity between face and car trials, averaged over all neurons in the network. Fig. 4c shows a graded response in the average difference, with the trend being toward smaller differences for lower coherence levels. To investigate how discriminating this difference was on a trial to trial basis we constructed histograms of the network-average for both face and car trials. Fig. 4d shows the resulting neurometric functions constructed by computing the area under the ROC curve (i.e. $A_z$) at each coherence level and fitting a Weibull function to these points. Note that we show neurometric curves constructed from the time windows, within which we compute the average firing rates, that result in the best discrimination. For the Magno system this window is 40-60 ms post-stimulus. The resulting neurometric curve is inferior compared to the psychometric curve, averaged over behavioral data of 10 human subjects. This suggests that the network-average is a suboptimal decoding strategy and that a more optimal decoding strategy requires exploiting both the spatial and temporal dynamics.

We next evaluated discrimination performance of the sparse linear decoder based on the spatio-temporal activity. A high-dimensional word, as in Fig. 3b, was constructed such that both rate based and timing based discriminating features could be exploited. Fig. 5 compares the resulting neurometric function with the psychometric curves. The corresponding neurometric function is comparable, if not perfect, to the psychometric behavior. This suggests that even at an early stage of V1, information of the complex stimuli, can already be reliably decoded by fully utilizing the rich dynamics. The difference between the neurometric and psychometric curves suggest that it is possible 1) there exists loss of information downstream in the decision making process, or 2) modulation such as attention is absent in the V1 model.

The sparse decoder not only makes decoding of high-dimensional features feasible, but also facilitates an avenue for feature selection. The cardinality of the decoder weights, namely the number of non-zero components in $w$, determines the effective word size and the number of neurons needed to optimally decode at each coherence level. In Fig. 6, we observed that the number of neurons for optimal decoding decreases as the coherence level increases, i.e. as the task gets easier. The difference between the word size and the number of neurons suggests that activity at multiple times for a subset of neurons is informative for discrimination and in some sense...
reflects the relevance of intrinsic dynamics to decoding.

IV. CONCLUSION

We have used an anatomically and physiologically detailed spiking neuron model of V1 to investigate the decoding accuracy for a relatively complex decision making task. A sparse decoder, based on $\ell_1$-regularized logistic regression, identifies the most informative dimensions distributed across the spatio-temporal cortical “word”. The resultant neurometric curve is comparable to the psychometric behavior. We find that a relatively small number of neurons participate in the decoding process for easy decisions, while a larger number is required for more difficult decisions.

We do not mean to suggest that such decisions regarding complex objects such as faces and cars are necessarily made at an early stage of V1, but instead that the information for decision making is preserved at V1 and can be extracted using a sparse decoding strategy. In addition, the supra-optimal results suggest that information loss is likely to occur downstream in the decision making process, especially for decision with high uncertainty. Recent monkey experiments have shown that optimal decoding of V1 activity, measured optically, outperforms the monkey in both speed and accuracy [8], [10]. Work based on multi-unit recording in monkey also suggests that activity of a large population of V1 neurons is combined suboptimally to mediate behavioral performance in a visual detection task [9].

Our model probes the capacity for the accuracy of perceptual decision making in V1, yet does not yield predictions about reaction time. However interesting to note is, as there is a dichotomy between strictly feedforward and recurrent network models for early stages of the perceptual decision making process, a similar dichotomy exists for models of the later stages of the process, where the decision variable is formed and the reaction time is reached [15], [16]. Recent work has argued for recurrence and subsequent network dynamics playing an important role in accumulating evidence, forming the decision and generating the response [16]. A challenge for future work is to consider how models of the decision making processes can be integrated with more realistic models of the stimulus transformation, as via this V1 model, in order to provide a more comprehensive picture of the perceptual decision making process.

V. ACKNOWLEDGEMENT

This work was supported by NGA NURI Grant HM1582-07-1-2002 and NIH Grant R01EY015520-01A2.

REFERENCES