Building a hierarchical causal model of natural movies with spiking neurons

Sophie Denève
Group for Neural Theory
ENS Paris

Timm Lochmann
Group for Neural Theory
ENS Paris

Udo Ernst
Institute for Theoretical Neurophysics
University of Bremen

October 24, 2007

Abstract: Several previous models have shown that elementary response properties like simple cell receptive fields with center-surround structure can be explained from specific assumptions about how elementary objects or causes combine to form a visual stimulus. These models state that natural images are caused by hidden objects that are (a) independent and (b) superimpose linearly to create an image. Simple cell V1 responses can be interpreted as the maximum-a-posteriori estimate of the contribution of each object to a given image.

Although such an assumption of linearity yields mathematical tractability, it is a relatively poor model for the description of natural stimuli. The linearity assumption also constrains the corresponding causal model for images (i.e., the generative model) to one layer. This may explain why these approaches did not predict more "high level causes" than gabor-like patterns, and did not predict later processing stages in the visual hierarchy, such as complex cells.

But already for responses of simple cells, these approaches provide a relatively poor description. They do not account for highly non-linear contextual interactions in V1, such as nonclassical receptive fields (ncRFs), divisive inhibition and contrast dependencies. While different mechanistic models have been proposed to account for the observed phenomena related to ncRFs, few studies have investigated their computational role.

On the other hand, approaches not based on generative models but on spatial and temporal invariances (e.g. slow feature analysis, SFA) were able to reproduce many properties of complex cell responses. This suggests that incorporating time (and thus, considering the processing of movies rather than static images) is essential to bring us closer to understanding visual processing in primary visual cortex. However, SFA looses the strength of the generative models approach, which is to directly relate the visual processing with assumptions about the statistics of the image. It is also not entirely clear how this relates to the unfolding of visual responses over time, i.e. to the spike trains of visual cells.

In this work, we attempt to overcome these shortcomings by framing the problem of dynamic vision in time with a simplistic, tractable generative model, where approximate inference and learning can be performed by neurons integrating evidence over time and signalling with spikes. We focus on three aspects of central interest: First, natural images are typically composed of overlapping or occluding objects, i.e. the lighting of a pixel results from one or another object, not a linear combination of both. In this sense, "objects" are binary variables (they are present or absent), not continuous values, and a pixel is lighted according to a non-exclusive "OR". Second, realistic input to perceptual systems has temporal dynamics: The task of perceptual systems can typically be described as online estimation of the changing value of a variable (temporal dynamics are assumed to be Markovian for the sake of tractability). Due to the fluctuating conditions and temporal dependencies in the environment this input is noisy and stems from the recent past. As an illustratory example, the task of a simple cell could be to detect the presence of an oriented edge in a movie from noisy inputs like photons impinging on the retina or spikes from early processing stages. And third, we propose that each visual neuron specializes for one object (one hidden cause) and computes its posterior probability (its probability of presence). As a summary, we assume that movies are composed of objects that appear and disappear randomly, and can conceal each other. The task of visual cells is to learn and report the presence of these "hidden" objects given their past input.

We show that inference in such a model can be efficiently approximated online by a network of spiking neurons. Furthermore, using unsupervised spike-time-dependent plasticity rules we demonstrate that these networks can learn the elementary image features and yield network configurations in which different units code for distinct time-varying causes. Thanks to these learning rule, the spatio-temporal receptive field of a visual neuron can be adjusted online to the spatio-temporal statistics of its preferred stimulus. This model accounts for the center-surround structure of receptive fields found for retinal ganglion cells / cells in the LGN as well as the structure of simple cell receptive fields. Furthermore it may also account for a large set of extra-classical RF effects, in particular changes in the spatio-temporal RF shape of visual neurons as a function of contrast, and orientation and spatial frequency dependent effects of the extra-classical surround. In such a network, the presence of divisive inhibition between neurons with nearby selectivity (i.e. from the ncRF) is essential. While feedforward connections from the input layers represent basic image features, lateral inhibitory connections enforce competition between the causes by means of divisive inhibition. The latter point accounts for the fact that a cause being activated explains the input and should prevent other units from explaining the same input.

This model provides a functional interpretation for the context sensitivity of the classical receptive field of V1 simple cells. Furthermore, it provides a link to empirical data by suggesting an explicit biophysical implementation in terms of shunting inhibition. Since this provides a tractable, non-linear, dynamic generative model of the visual input and incorporates "slow features" in its Markovian statistics, it may help to discover essential higher order statistics in natural movies.