A neuronal model for the shaping of feature selectivity in IT by visual categorization

M. Szabo [a] R. Almeida [a] G. Deco [b] M. Stetter [a,*]

[a] *Siemens Corporate Technology, Information & Communications, München, Germany*

[b] *ICREA and University Pompeu Fabra, Dept of Technology, Barcelona, Spain*

Abstract

Neurophysiology results have shown that learning a visual categorization task shapes the selectivity of inferiotemporal cortex neurons to task-relevant features of the stimuli. In this work we propose a biologically realistic mean-field neuronal model of a two layer network to explain these experimental results. We show that the enhancement of feature selectivity in one layer of the model can emerge due to input coming from another layer, corresponding to a region encoding stimulus category, possibly in prefrontal cortex. Further, we explore the behavior of the network in function of the weights of the connections between its two layers.

Key words: IT, PFC, perceptual learning

1 Introduction

In a recently performed neurophysiology experiment in monkeys, Sigala and Logothetis have studied how the selectivity to stimuli features of inferotemporal cortex (ITC) neurons is affected by performance in a visual categorization task [1]. The stimuli used were characterized by several features, some of which (diagnostic features) were relevant for the categorization task and some of which were irrelevant. For a given stimulus, each feature took one of several different values. The experimental results showed that, after training, selectivity for the different levels of the diagnostic features was enhanced in comparison to selectivity for the levels of other features. In our work we propose a model to explain these experimental results. ITC and prefrontal

* corresponding author
Email address: stetter@siemens.com (M. Stetter).

cortex (PFC) are two interconnected cortical areas thought to be involved in the performance of visual tasks, such as visual recognition, categorization and memory. In this context, recent studies have suggested that PFC is mainly associated with cognitive dependent processing (such as categorization), while ITC is more associated with feature processing [2]. Further, top-down signals from PFC to ITC are thought to partially determine ITC neuronal responses [2; 3]. Taking such results into account, we hypothesize that the enhancement of selectivity for the levels of the diagnostic features reported in [1] for ITC neurons might emerge from the interaction between ITC and another cortical region, possibly in PFC, in which the previously learned categories are encoded. In order to test our hypothesis and account for the experimental results, we propose a two layer model in the framework of biased competition and cooperation [4]. One layer corresponds to a part of ITC and is organized into pools of neurons which receive feature specific inputs. The other layer corresponds possibly to a region in PFC and contains neuronal pools which are connected to the pools in ITC in a way which allows them to encode for the categories of the stimuli. Using mean-field simulations we show that the enhancement of selectivity for level of the diagnostic features can result from top-down information to ITC, signaling category.

2 Methods

This work is based on a biologically realistic network of spiking neurons. This model is a modification of the network introduced by Brunel and Wang [5], which has been extended and successfully applied to explain several experimental paradigms [4; 6; 7]. The simulations presented are done using a mean-field approximation consistent with the model used [5], which allows an exhaustive analysis of the regimes as a function of the parameter space.

In the present case the model is organized into two connected layers (see Figure 1). Each layer consists of $N_E = 800$ excitatory pyramidal cells and $N_I = 200$ inhibitory interneurons, which are fully connected, unless stated otherwise. The excitatory neurons of each layer are organized into several pools, characterized by shared inputs and weights of connections. The layer corresponding to ITC is organized into five excitatory pools: four stimulus selective pools with 80 neurons each and one pool, called non-selective, constituted by all other excitatory neurons. The neurons in the four selective pools receive external inputs encoding stimulus specific information. For simplicity, in our work we consider that the stimuli presented are characterized by two features, with two levels each (high and low). We also consider just two possible categories and that these categories are determined exclusively by one of the features, the diagnostic feature. The four selective pools in the ITC layer are denoted according to the specific inputs they receive: one pool receives input when

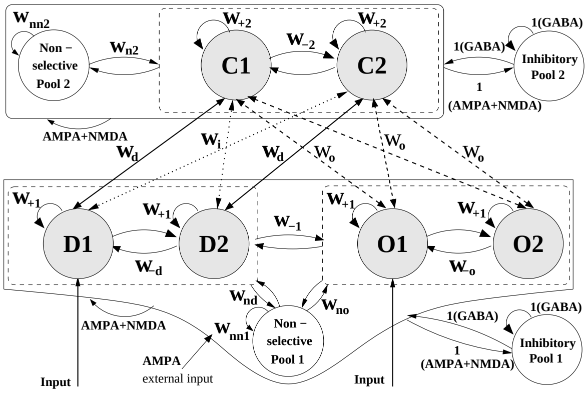

Fig. 1. Schematic representation of the architecture of the two layer model used.

the stimuli is characterized by the diagnostic feature being high (pool D1), one when the diagnostic feature is low (D2), one when the other feature is high (O1) and one when it is low (O2). The layer corresponding to PFC is organized into one non-selective pool and two pools selective for category 1 (C1) or category 2 (C2). The stimuli with diagnostic feature high belong to category 1 and the stimuli with diagnostic feature low belong to category 2. The selectivity of the category selective pools emerges from their connections with the selective pools of the ITC layer.

All neurons receive a background external input modeled as a Poisson train with rate equal to 2.4 kHz. The neurons in the ITC layer can receive extra external inputs coding for the stimulus presented. In this case, the rate of the Poisson train is increased by $\lambda_{stim} = 50$ Hz. The excitatory recurrent connections are considered to be mediated by AMPA and NMDA receptors, the external connections are assumed to be mediated by AMPA receptors, and the inhibitory connections are GABAergic [5]. We modulate the conductance values for the synapses between pairs of neurons by connection weights, which can deviate from their default value of 1. The structure and function of the network is achieved by differentially modulating these weights within and between pools of neurons. The structure is set so that the sum of weights of the connections to each neuron is 1, to assure stability. The weights are considered fixed, after some tuning process, compatible with what would result from Hebbian learning. The labeling of the weights is defined in Figure 1. Since we want to model the enhancement of selectivity in the ITC layer due to categorization, we chose to leave no structure in this layer, thus implementing cooperation between all pools [4]. That is, all weights are taken equal to one, except for the weights from and to the non-selective neurons which were computed to assure stability (the same was done for the non-selective pool of the PFC layer). In the PFC layer, within the same category pool, the neurons are strongly coactivated, and are therefore connected with a stronger than average weight, in this case considered to be $w_{+2} = 2$. Neurons from different category selective pools are likely to have anti-correlated activity resulting in weaker than aver-

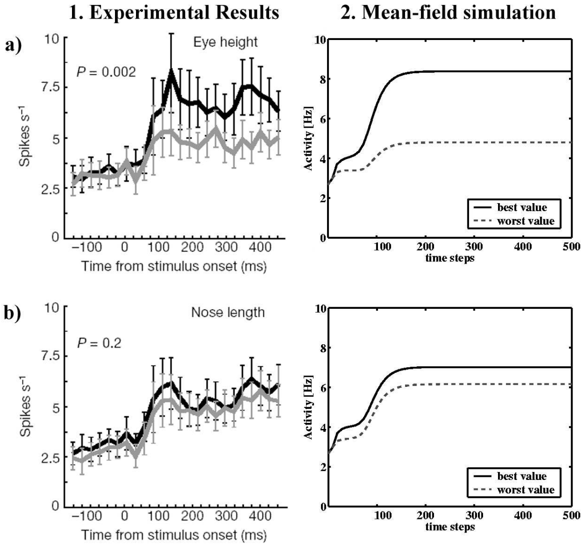

Fig. 2. Experimental results (1) (adapted from [1]) and mean-field simulations (2) showing average spiking rates of feature selective neurons according to their best (black solid lines) and worst (gray and dashed lines) responses to the levels of diagnostic (a) and non-diagnostic (b) features.

age connections, in this case set to $w_{-2} = 0$. Finally, the connections between the two layers are set as follows: the neurons in pools O1 and O2 and the category neurons are assumed to have uncorrelated activities, and hence we chose the same weight w_o, for all the connections between them. The activities of the diagnostic selective neurons and the corresponding category neurons are likely to be correlated, so we hypothesize that the connection strengths between D1 and C1 (and D2 and C2, respectively) increase so that $w_d > w_o$. Likewise the neurons in D1 and C2 (D2 and C1, respectively) probably have anti correlated activities, resulting in weaker weights $w_i < w_o$. The absolute strengths of the weight parameters between layers are explored to analyze the different operational modes of the model. **The mean-field formulation we use ([5]) can be directly related to the cellular properties, but describes the population averaged firing rate of each individual pool by only one equation.**

3 Results

Figure 2 shows results experimentally measured (2.1) and results from one mean-field simulation (2.2). **The experimental results show the population average for all recorded visual responsive neurons, when different combinations of features were presented.** Four responses were selected: the highest (black line) and lowest (grey line) responses sorted according to one diagnostic feature (Figure 2.1.a) and the highest (black line) and lowest (grey line) responses sorted according to one non-diagnostic feature (Figure 2.1.b). **The simulation results were obtained by doing the corresponding calculations in our model. In order to reproduce the experimental data, we also took into account all the neurons responding to the presented stimuli, so the average firing rate over**

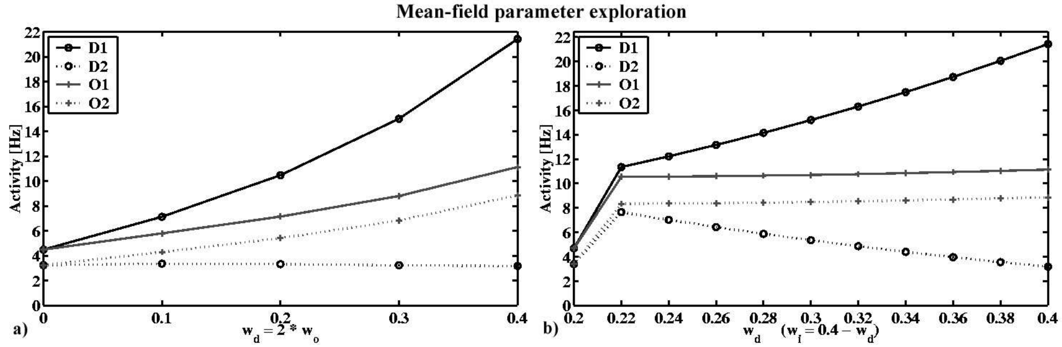

Fig. 3. Average spiking rates of the feature selective neuronal pools as function of w_d, when $w_d = 2w_o$ and $w_i = 0$ (a) and when $w_o = 0.2$ and $w_i = 0.4 - w_d$ (b).

all selective pools in the ITC model layer is considered. We run our model for all possible combinations of the input values. After this, we check for each selective pool, which value of the diagnostic feature (nondiagnostic, respectively) produces a higher response and we use this value to compute the average rate representing best value for diagnostic (nondiagnostic, respectively). Similarly the lower response was used to compute the average rate representing the worst value. This average activity over all selective pools based on the sorting regarding the best and worst value of each pool of the diagnostic (nondiagnostic, respectively) is presented in Figure 2.2. The model parameters were set to: $w_d = 0.2$, $w_o = 0.1$ and $w_i = 0$. Four stimuli were used, corresponding to external inputs to the pairs of pools: D1 O1, D2 O1, D1 O2, D2 O2. The simulations results are in good agreement with the experimental results and show that there is an enhancement of the selectivity for the level of the diagnostic feature, as compared to the non-diagnostic feature (the lines in the plots (a) are more separated than those in the plots (b)). Further, since there was no structure in the ITC layer of the model, we show that the enhancement of selectivity emerges due to the feedback from the PFC layer, which signals category.

Figure 3 shows how the selectivity for the diagnostic feature, when compared to that of the other feature, behaves as a function of the weights of the connections between the two layers. Figure 3.a was obtained keeping $w_i = 0$ and changing w_d and w_o, subject to the constraint $w_d = 2w_o$. Figure 3.b was obtained with $w_o = 0.2$ and $w_i = 0.4 - w_d$. In both figures the firing rates of the four selective pools of the ITC layer are plotted as a function of w_d. The results show that when w_d increases in comparison to w_i, both changing w_o (Figure 3.a) or keeping it fixed (Figure 3.b), the difference between the firing rates of the pools D1 and D2 increases, while the difference between the firing rates of O1 and O2 remains approximately constant. The increased difference for the spiking rates associated with the diagnostic feature shows that there is an enhancement of selectivity for the level of this feature in the ITC layer. That there is no corresponding increase in difference for the O pools (due to cooperation) shows that the effect is specific for the diagnostic features and thus that it is mediated by the category specific top-down input from the PFC layer. In both plots of Figure 3, for relatively high values of w_d, the firing activity of the pool D2 approaches the spontaneous level of activity, 3 Hz. In this case the pool D2 looses its general responsiveness to D1 O1 stimuli, which disagrees with the experiment, in which all considered neurons showed responsiveness.

4 Discussion

In this work we present a biologically realistic two layer network model of spiking neurons to account for the enhancement of selectivity in ITC neurons, for stimulus features which are relevant for a learned categorization visual task [1]. Using the biased competition-cooperation framework [4], we show that the enhancement of selectivity can emerge from category specific top-down signals, possible originating in PFC. The effect described can be explained by considering that after a learning period, the connections between the feature encoding layer (ITC) and the category encoding layer (PFC) are such that neurons activated by the level of a feature determinant for categorization are strongly connected to the associated category and weakly connected to the other category. Neurons which receive input specific for a task-irrelevant feature, are connected to the category neurons with an average weight, not necessarily significantly changed during training.

Acknowledgments

This work was supported by the EU grants IST-2001-38099 and QLK6-CT-2002-51439 (to R. Almeida) and by German Ministery for Research grant BMBF 01IBC01A.

References

[1] N. Sigala, N. Logothetis, Visual categorization shapes feature selectivity in the primate temporal cortex, Nature 415 (2002) 318–320.

[2] D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, A comparison of primate prefrontal and inferior temporal cortices during visual categorization, J. Neurosci. 23 (2003) 5235–5246.

[3] H. Tomita, M. Ohbayashi, K. Nakahara, I. Hasegawa, Y. Miyashita, Top-down signal from prefrontal cortex in executive control of memory retrieval, Nature 401 (1999) 699–703.

[4] M. Szabo, R. Almeida, G. Deco, M. Stetter, Cooperation and biased competition model can explain attentional filtering in the prefrontal cortex, Eur. J. Neurosci. In press.

[5] N. Brunel, X. J. Wang, Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition, Comput. Neurosci. 11 (2001) 63–85.

[6] G. Deco, E. T. Rolls, Attention and working memory: A dynamical model of neuronal activity in the prefrontal cortex, Eur. J. Neurosci. In press.

[7] G. Deco, E. T. Rolls, B. Horwitz, ”what” and ”where” in visual working memory: A computational neurodynamical perspective for integrating fMRI and single-neuron data, J. Cogn. Neurosci. In press.