Proc. of SoftComputing'96, Iizuka, Japan, ed. T. Yamakawa, G. Matsumoto, pp. 482-485.

Categorization, prototype theory and neural dynamics.

Włodzisław Duch
Department of Computer Methods, Nicolauss Copernicus University,
ul. Grudziadzka 5, 87-100 Toruń, Poland.
Tel/Fax +48-56-21543, e-mail: duch@phys.uni.torun.pl
WWW: http://www.phys.uni.torun.pl/kmk

Keywords: cognitive science, neurodynamics, feature spaces, categorization, neural networks

ABSTRACT

Model of categorization plausible from the neurobiological point of view is outlined. A link between neural systems and the theory of psychological spaces is presented, leading to a model of mind in which physics of mental events is determined by neural dynamics. This model is used to discuss psychological category learning experiments.

1 INTRODUCTION.

Categorization, or creation of mental categories, is one of the most important cognitive processes. It is also one of the most difficult processes to understand if one tries to see it from the point of view of both psychology and neuroscience. Current research on category learning and concept formation frequently ignores constraints resulting from the neural plausibility of postulated mechanisms. Connectionist models are at best loosely inspired by the idea that neural processes are at the basis of cognition. An explanation given by a formal theory, even if it fits psychological data, may allow for predictions, but it does not give us more understanding of human cognition than a few-parameter fits allowing for prediction of sun eclipses gave the ancient astronomers. Correlation does not imply causation.

Several models of categorization of perceptual tasks have been compared by Cohen and Massaro [1], including Fuzzy Logical Model of Perception (FLMP), Gaussian Multidimensional Scaling Model (GMM), Theory of Signal Detection (TSD), Feedforward Connectionist Model (FCM) and Interactive Activation and Competition Model (IAC). All these models predict probabilities of responses in a prototypical two and four-response situations in an almost equivalent way. The main purpose of this contribution is to outline a path from neuroscience to psychology and base the *ad hoc* categorization models on more solid foundations.

2 BRAIN AND INFORMATION PROCESSING.

There is growing theoretical and experimental evidence [2] that the original idea of local reverberations in groups of cortical neurons coding the internal representations of categories, put forth by psychologist Donald Hebb already in 1949, is essentially correct. Local circuits seem to be involved in perception and in memory processes. Analysis of integration of information from the visual receptive fields in terms of modules composed of dense local cortical circuitry [3] allows for explanation of a broad range of experimental data on orientation, direction selectivity and supersaturation. It would be most surprising if the brain mechanisms operating at the perceptual level were not used at higher levels of information processing. Neocortex has highly modular organization. Neurons are arranged in six layers and grouped in macrocolumns containing in turn microcolumns (110 neuron in association cortex). Successful models of memory, such as the tracelink model of Murre [4], make good use of this modular structure. Probably each episodic memory is coded in a number of memory traces that are simultaneously activated and their activity dominates the global dynamics of the brain, reinstating similar neural state as during the actual episode.

A fruitful hypothesis relating psychological concepts to brain activity is based on the following reasoning. There is good experimental evidence, coming from the recordings of the single-neuron activity in the inferotemporal cortex of monkeys performing delayed match-to-sample tasks (cf. [2]), showing that the activity of a neural cell assembly (NCA - presumably a microcolumn within a macrocolumn) has attractor dynamics. Several stable patterns of local reverberations may form, each coding a specific perceptual or cognitive representation. Via axon collaterals of pyramidal cells extending at distances of several millimeters, each NCAs excites other NCAs coding related representations. From the mathematical point of view the structure of local activations is determined by attractors in the dynamics of neural cell assemblies. Such networks should be properly described as

a collection of mode-locking spiking neurons. Simple models of competitive networks with spiking neurons have been created to explain such psychological processes as attention (cf. [5]). Realistic simulations of the dynamics of microcolumns, giving results comparable with experiment, should soon be possible, although have not been done yet. Recently Amit and Brunel [6] have solved the problem of spontaneous activity and stability of the background dynamics of networks of spiking neurons. Solution of this basic problem requires modular structure of the network, including inhibitory interneurons within NCAs. Learning creates local attractors without destabilizing the background dynamics. Predictions from such models are directly compared with neurophysiological experiments.

To make a step towards psychology possible attractor states of neurodynamics should be identified, basins of attractors outlined and transition probabilities between different attractors found. In the olfactory system it was experimentally found [7] that the dynamics is chaotic and reaches attractor only when an external input is given as a cue. The same may be expected for the dynamics of NCAs. Specific external input provides a proper combination of features that activates a category coded by the NCA. From the neurodynamical point of view external input puts the system in a basin of one of the local attractors. Such neural networks map input vectors **X** (cues) into fuzzy prototypes. Although the exemplar theory of categorization is usually presented as an alternative to the prototype theory [8] neurodynamics lies at the basis of both theories. Since neural dynamics in biological networks is noisy (spontaneous background cortex activity and other sources) several similar exemplars become so fuzzy that a single prototype is formed. To see it clearly a complementary description via feature spaces is introduced.

3 ENCODING CATEGORIES IN FEATURE SPACES.

A classic category learning task experiment has been performed by Shepard *et.al.* in 1961 and replicated by Nosofsky *et.al.* [9]. Subject were tested on six types of classification problems for which results were determined by logical rules. For example, categories of Type II problems had the XOR structure (i.e. XOR combination of two features determines which category to select) that may be described by the following dynamical system:

$$V(x,y,z) = 3xyz + \frac{1}{2}\left(x^2+y^2+z^2\right)^2$$

$$\dot{x} = -\frac{\partial V}{\partial x} = -3yz - \left(x^2+y^2+z^2\right)$$
$$\dot{y} = -\frac{\partial V}{\partial y} = -3xz - \left(x^2+y^2+z^2\right) \quad (1)$$
$$\dot{z} = -\frac{\partial V}{\partial z} = -3xy - \left(x^2+y^2+z^2\right)$$

This system has 5 attractors (0,0,0), (-1,-1,-1), (1,1,-1); (-1,1,1), (1,-1,1); the first attractor is of the saddle point type and defines a separatrix for the basins of the other four. Such dynamical system may be realized by different neural networks. In this example, as well as in the remaining five types of classification problems [9], it is easy to follow the path from neural dynamics to the behavior of experimental subjects during classification task. Starting from examples of patterns serving as point attractors it is always possible to construct a formal dynamics and realize it in the form of a set of frequency locking nonlinear oscillators [10].

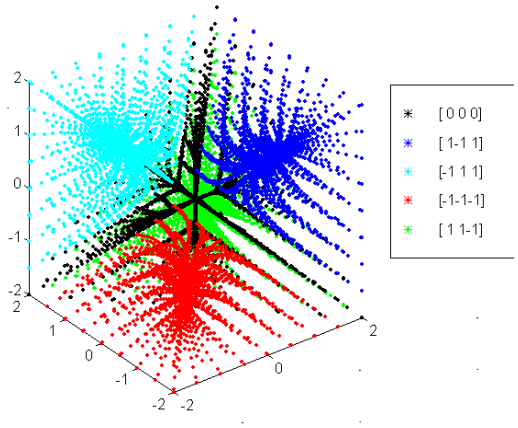

Fig. 1. Trajectories showing the basins of five attractors for the Type II classification problem of Shepard *et.al.*

It is convenient to describe such classification problems in a feature space. In case of Shepard experiments it contains axis for shape, color and size. Feature spaces, called also psychological spaces, are quite popular among psychologists. Our goal is to show how neural dynamics is connected to processes in the feature spaces. Neural dynamics models physical processes at the level of brain events while feature spaces model mental processes providing precise language to speak about the mind events. Psychological models of categorization should be justified as approximations to real neural dynamics. Attractors activated by specific inputs $\mathbf{X}_{inp}$ divide the input space into areas corresponding to basins of different attractors. For example, a cortical microcolumn may learn to solve the A.XOR.B problem establishing attractors

presented in Fig. 1. In the input space (feature space) the four vertices of the cube will represent the shortest transients of the phase space trajectories and the basins of attractors will belong to the neighborhood of these vertices. Introducing the density of feature space objects $M(\mathbf{S})$ proportional to the length of transients of the neural dynamics (the time it takes to reach an attractor from a given initial conditions $\mathbf{S_0}$) neural dynamics defined by activity of a large number of neurons may be approximated by simple gradient dynamics in the feature space [11].

Categorization based on prototypes is characterized by large basins of attractors, corresponding to large and fuzzy objects in feature spaces. A prototype is not simply a point with average features for a given set of examples, but a complex fuzzy object in the feature space. If categorization is based on exemplars basins of attractors corresponding to these exemplars should be small and the feature space objects well localized. Noise in neural system will destroy weak local attractors, changing a set of localized objects representing exemplars to a fuzzy prototype with some internal structure. A reasonable approximation (called further the FSM approximation) to the neural dynamics represented in the feature space is:

$$\mathbf{S}(0) = \mathbf{X}_{inp}$$
$$\dot{\mathbf{S}}(t) = \beta \nabla_{\mathbf{S}} M(\mathbf{S};t) / \left(1 + g\left(M(\mathbf{S};t)\right)\right) + \eta(t) \quad (2)$$

where β is a step size constant, memory function $M(\mathbf{S};t)$ represents time-dependent (due to learning) object density in the feature space and η is a noise term representing spontaneous spiking activity. The denominator contains a function *g(x)* equal to zero for small *x* and taking large values around local maxima of the memory function, slowing down the state vector dynamics near memorized categories. In effect the time it takes to go from object A to B may be different than the time it takes to go from B to A. If the cue $\mathbf{X}_{inp}$ is sufficiently similar to a memorized object (corresponding to an attractor in neural dynamics) the state vector following the gradient of $M(\mathbf{S};t)$ will stay within this object. If not, the noise term will bring the state vector $\mathbf{S}(t)$ close to one of the memorized objects and the probability of different responses will depend on the local topography of the feature space. This dynamics should model probability and the timing of different answers when specific cues $\mathbf{X}_{inp}$ are given. It would be ideal to fix the form of the *g(x)* function comparing the gradient dynamics to the neurodynamics being modeled.

People learn relative frequencies (base rates) of categories and use this knowledge for classification. This is known as the base rate effect. Frequently repeated stimuli create deep basins of attractors (large densities of feature space objects). The size of these basins depends on the inherent noise and variability of the stimuli. Such effects are relatively simple to model. The **inverse base rate effect** [12] shows that in some cases predictions contrary to the base rates are made. Names of two diseases, *C* (for Common) and *R* (for Rare), are presented to participants, the first linked to symptoms *I* and *PC*, and the second *I* and *PR*. Thus *PC* and *PR* are perfect predictors of the disease *C* and *R*. Associations $(I,PC) \rightarrow C$ are presented 3 times more often than $(I,PR) \rightarrow R$. After a period of learning participants are asked to predict which disease corresponds to a novel combination of symptoms. For a single symptom *I* most (about 80%) predict *C*, in agreement with the base rates. For combination of symptoms *PC+I+PR* most (60%) choose *C*, again with agreement with the base rates (cf. Fig.2). However, 60% participants associate the combination *PR+PC* with the disease *R*, contrary to the base rate expectations.

For many years this effect has eluded explanation until Kruschke and Erickson [13] have introduced a model integrating six psychological principles of human category learning: error-driven association learning, rapid shift of attention, base rate learning, short term memory effects, strategic guessing and representations based on exemplars and their fragments. While strategic guessing in novel situations (assigning novel stimuli to still-to-be-learned categories) is certainly a higher order cognitive process all other principles may be absorbed in construction of representations of categories rather than in processes acting on these representations.

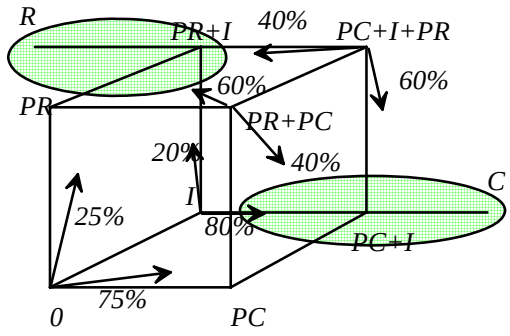

Fig. 2. Feature space for symptoms *I, PC, PR.* Combination *PC+PR* leads in about 60% responses to prediction of *R* and in 40% to prediction of C disease.

The answers are determined by the sizes of the basins of attractors corresponding to shapes of objects in the feature space. The memory function describing these ob-

jects may be fitted to obtain observed probabilities of answers, as is usually done in psychological modeling [1]. The *C* basin is larger, extends between *I* and *PC+I* vertices, forcing the *R* basin to be flatter and be closer to the *PR+PC* vertex than the *C* basin is, leading to the inverse base rate effect.

Processes acting on representations in feature spaces define physics of mental events, with forces reflecting the underlying neural dynamics. In the absence of cues the state vector **S**(*t*) moves randomly in the feature space. Base rate effects influence the size of the basins of attractors (size of the feature space objects). Specifying value of a feature that frequently appears in combination with other features gives momentum to the state vector in the direction parallel to the axis of this feature, initiating a search for a value of unspecified features (for application of such searches see [14]).

4 SUMMARY

In principle it should be possible to understand categorization and other cognitive processes at the level of dynamics of spiking neural networks, but in practice approximations simplifying description of this dynamics are necessary. The popular feedforward neural networks do not offer a good approximation to real neural dynamics. Their success in modeling psychological data are only due to their ability to approximate arbitrary vector mappings. At best they may capture some correlations, but not proper causation. Psychological models of categorization have been developed in the past 40 years and are already quite sophisticated. To show that these models contain some truth one should try to justify them as approximations to neural dynamics. Therefore it is interesting to note that the FLMP, GMM and TSD categorization models [1] may be derived as static approximations to the dynamic feature space model described here.

Linking neural dynamics with psychological models using feature spaces leads to a complementary description of brain processes and mental events. The laws governing these mental events result from approximations to neural dynamics. Modified feature space models should be useful in analysis of data from many psychological experiments. Learning how to link simplest neural dynamics with feature space representations is just one small step, but many more challenges remain. Hopefully this approach may offer not only good fits to the observations, but also interesting interpretation of mental events.

ACKNOWLEDGMENT

Support by the Polish Committee for Scientific Research, grant 8T11F 00308, is gratefully acknowledged.

REFERENCES

[1] M.M.Cohen, D.W. Massaro, On the similarity of categorization models, chapter 15 in: F.G. Ashby, ed. Multidimensional models of perception and cognition (LEA, Hillsdale, NJ 1992)

[2] D.J. Amit, The Hebbian Paradigm Reintegrated: Local Reverberations as Internal Representations. Brain and Behavioral Science 18(4), 617-626, 2004

[3] D. C. Somers, E.V. Todorov, A.G. Siapas, M. Sur Vector-space integration of local and long-range information in visual cortex (MIT AI memo 1556, 1995)

[4] J. Murre, *TraceLink: a model of amnesia.* Psychological Review 111(1):205-235, 1996.

[5] J.G. Taylor, F.N. Alavi, *Mathematical analysis of a competitive network for attention.* In: J.G. Taylor, ed. Mathematical Approaches to Neural Networks (Elsevier 1993), pp.341-382

[6] D.J. Amit, N. Brunel, Global spontaneous activity and local structured (learned) delay activity in cortex (preprint, Inst. of Physics, Univ. of Rome, 1995)

[7] W.J. Freeman, Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. Biolog. Cybernetics 56 (1987) 139-150

[8] I. Roth and V. Bruce, *Perception and Representation* (Open University Press, 2nd. ed. 1995)

[9] R.M. Nosofsky, M.A. Gluck, T.J.Palmeri, S.C. McKinley and P. Glauthier, Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland and Jenkins (1961). Memory and Cognition 22 (1994) 352-369

[10] H. Haken, *Synergetic Computers and Cognition* (Springer, Berlin 1991)

[11] W. Duch, *On simplifying brain functions.* Proc. of the Second Conference on Neural Networks and their applications, Orle Gniazdo, 1996, pp. 118-124

[12] D.L. Medin and S.M. Edelson, *Problem structure and the use of base-rate information from experience.* J. of Exp. Psych: General 117 (1988) 68-85

[13] J. K. Kruschke, M.A. Erickson, *Five principles for models of category learning*. In: Z. Dienes (ed.), Connectionism and Human Learning (Oxford, England: Oxford University Press 1996)

[14] W. Duch, G.H.F. Diercksen, *Feature Space Mapping as a universal adaptive system*. Comp. Phys. Comm. **87** (1995) 341-371