6 The SOM as a Model of Brain Maps

Teuvo Kohonen

The original motivation for the SOM algorithm was an attempt to explain various spatially organized neural "maps" in the central nervous system. In the light of the present knowledge, however, it seems necessary to distinguish three categories of such "brain maps": A1. Feature-sensitive cells that respond, e.g., to specific sensory stimuli. A2. Anatomically organized representations of the body or some receptor surface of it, e.g., in the visual, somatosensory, motor, and auditory cortices. A3. Abstract feature maps that constitute a topographical or topological representation of a specific feature space of the sensory experiences. Examples of such abstract maps are the color map in the visual area V4, and various maps of the auditory space.

Unlike the anatomical maps, the ordered mappings of abstract features cannot be produced by genetically controlled ordered growth of axons, because no ordered receptor system, from which such ordered axons could originate, exists for abstract features. The order that has ensued in the mapping must have emerged by self-organization.

The following three conditions seem to be necessary for the production of biological maps of abstract features: B1. All the cells of such a brain area must receive essentially similar information. B2. There must exist a mechanism for the activation of that particular cell (called the "winner") which, in some sense, is "best fit" to the input information. Its activity shall further be enhanced, for instance by lateral excitation and inhibition, while the activity in the rest of cells is suppressed. B3. There must exist a learning mechanism by which the "winner" and a subset of its spatial neighbors in the area become "tuned" to the prevailing stimulus, while no learning outside this subset occurs. In the long run, when different subsets of cells are activated by different stimuli, a global order along with some dominant features in the stimuli then ensues.

It will be necessary to notice that the above conditions B1 and B2 entail that an area over which the input can be regarded similar, and in which lateral interactions over the area may occur, cannot be very large. In view of the extent to which the afferent axons and intracortical collaterals can spread, a feature map in the cortex may only have a diameter of a few millimeters.

In the biological modeling it is customary to approximate the activation of a neuron by the dot product of its input signal vector $\mathbf{x}$ and its synaptic weight vector $\mathbf{m}_i$, but if this law is used for the definition of the winner in the SOM, then the updating law must be made *self-normalizing*:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}[\mathbf{x}(t) - (\mathbf{m}_i^{\mathrm{T}}(t)\mathbf{x}(t))\mathbf{m}_i(t)] \;, \tag{56}$$

where h_{ci} is the neighborhood function, and the winner is defined by

$$c = \arg\max_i \{\mathbf{m}_i^{\mathrm{T}}(t)\mathbf{x}(t)\} \;. \tag{57}$$

In the next report we shall show how the maximum selection in (57), in principle at least, could be implemented by physiologically plausible networks.

It can be shown that starting with arbitrary initial values $\mathbf{m}_i(0)$, with $\|\mathbf{x}(t)\| = 1$, and h_{ci} sufficiently small, the $\|\mathbf{m}_i(t)\|$ tend to the value 1. Nonetheless (56) preserves the self-organizing property, which can be seen, e.g., from Fig. 3, where the phoneme map has been computed using this equation.

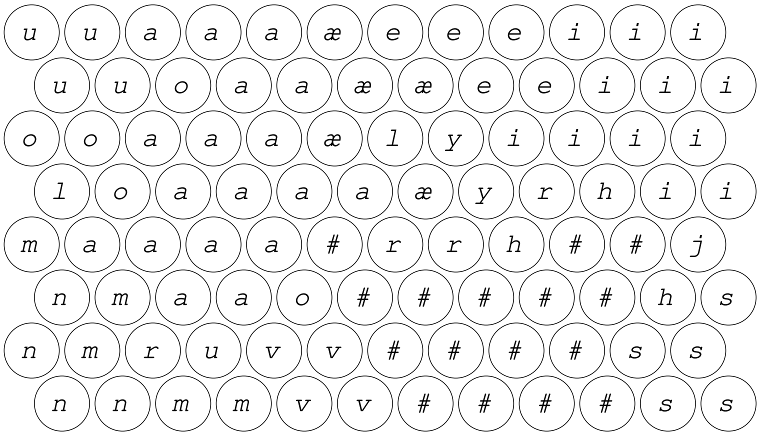

Figure 3: Self-organizing map of Finnish phonemes when 15-channel short-time spectra of natural speech, evaluated at every 20 ms, were used as the $\mathbf{x}(t)$ in (56).

The biological motivation for the bracketed expression in (56) may come from the following argumentation. First, the adaptive changes of a synapse must be made *reversible* in order to keep its weight on the dynamic range during its whole lifetime. Such a law, in the discrete-time formalism, could have the general form

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha[P\mathbf{x}(t) - Q\mathbf{m}_i(t)] \; , \tag{58}$$

where α is a small factor. The term $\alpha P\mathbf{x}(t)$ describes the memory traces due to all presynaptic excitations, and the strengths of the memory traces are assumed proportional to $\mathbf{x}(t)$, while $-\alpha Q\mathbf{m}_i(t)$ represents the *forgetting effect*, which is assumed proportional to $\mathbf{m}_i(t)$. It may be stipulated that the biological forgetting is mostly "active," e.g., Q depends on the degree of activation of the cell, being proportional to $\mathbf{m}_i^{\mathrm{T}}(t)\mathbf{x}(t)$. Furthermore, in order to hold the "memory traces" steady for indefinite periods of time, while being able to change them fast upon demand, it must be assumed that both learning and forgetting (i.e., in general the synaptic changes) not only depend on the presynaptic signal activity, but are also conditioned by some *plasticity control factor* or *"learning factor"* produced by strong activities either at the cell itself or at its *neighboring cells*. If then the activity of the neural network is strongly clustered, i.e., if some kind of competitive process is at work selecting the winner, enhancing its activity, and suppressing the activity in the rest of the cells, then spatial spreading of the "learning factor" to neighboring cells means that P and Q must involve some *interaction kernel* h_{ci} as a factor, relating to the active cell c and cell i. This argumentation then directly results in the adaptation law (56).

Reference

T. Kohonen and R. Hari. Where the abstract feature maps of the brain might come from. *Trends in Neurosciences*, 22:135-139, March 1999.