To appear in The Handbook of Brain Theory and Neural Networks, Second edition,
(M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002.
http://mitpress.mit.edu

SPARSE CODING IN THE PRIMATE CORTEX

Peter Földiák

School of Psychology

University of St Andrews

St Andrews KY16 9JU

U.K.

Peter.Foldiak@st-andrews.ac.uk

INTRODUCTION

Brain function can be seen as computation, i.e. the manipulation of information necessary for survival. Computation itself is an abstract process but it must be performed or implemented in a physical system. Any physical computing system, be it an electronic computer or a biological system consisting of neurons, must use some form of physical representation for the pieces of information that it processes. Computations are implemented by the transformations of these physical representations of information. The brain receives information via the sensory channels and must eventually generate an appropriate motor output. But before we can even study the transformations that are involved, we need at least some fundamental understanding of the internal representation that these transformations operate on. Neurons represent and communicate information mainly by generating (or 'firing') a sequence of electrical impulses. Electrophysiological techniques exist for the recording of these impulses from isolated single neurons in the living brain. Single cell recording has revealed a remarkably close and highly specific relationship between sensory stimuli, neural activity, and behavioral reports of perceptual states. However, the encoding of events and states into a sequence of neural impulses in a large number of neurons can be highly complex, especially in the cerebral cortex, which contains an elaborate working model of the world. One of the basic questions about this neural code is whether an information item (e.g. a specific sensory stimulus caused by an object) is represented by the activity of a single, individually meaningful cell, or is it only the global activity pattern across a whole cell population that corresponds to interpretable states? There are now strong theoretical reasons and experimental evidence suggesting that the brain adopts a compromise between these extremes, using a relatively small (though in absolute number still substantial) subset of neurons to represent each item. This is often referred to as sparse coding (Dayan and Abbott, 2001).

Sparse Coding

The brain must encode the state of the environment and its own internal states by the firing pattern of a large but fixed set of neurons. For simplicity, consider coding by units that are either "active" or "passive" where the code assigns states to the subsets of active units. An important characteristic of such a code is the activity ratio, the fraction of active neurons at any one time. At its lowest value is local representation, where each state is represented by a single active unit from a pool in which all other units are silent. For instance, letters on a computer keyboard (ignoring the Shift and Control keys) are locally encoded. In dense distributed representations, each state is represented on average by about half of the units being active. Examples of this are the binary (ASCII) encoding of characters used in computers or the coding of visual images by the retinal photoreceptor array. Codes with low activity ratios are called sparse codes.
The activity ratio affects several aspects of information processing such as the architecture and robustness of networks, the number of distinct states that can be represented and stored, generalization properties, and the speed and rules of learning (Table 1).
The representational capacity of local codes is small: they can represent only as many states as the number of units in the population, which is insufficient for any but the most trivial tasks. Even when the number of units is as high as that in the primate cortex, the number of discriminable states of the environment well exceeds this number. Making associations between a locally encoded item and an output, however, is easy

	representational capacity	memory capacity	speed of learning	generali-zation	interference	fault tolerance	simultaneous items
local	very low	limited	very fast	none	none	none	unlimited
sparse	high	high	fast	good	controlled	high	several
dense	very high	low	slow	good	strong	very high	one

Table 1.Properties of coding schemes

and fast. Single-layer networks can learn any output association in a single trial by local, Hebbian strengthening of connections between active representation units and output units, and the linear separability problem does not even arise. In such a lookup table, there is no interference between associations to other disriminable states, and learning information about new states does not interfere with old associations. This, however, also means that there will be no generalization to other discriminable states, which presents a serious problem. If the discrimination is poor, that is a problem in itself. Alternatively, if the discrimination is fine enough to make the necessary distinctions we can expect a system never to experience precisely the same pattern of stimulation twice, and no past experience will ever be used in a new situation.

Dense distributed, or "holographic," codes, on the other hand, can represent a very high number (e.g. in binary codes, 2^N) of different states by combinatorial use of (N) units. In fact, this power is largely superfluous, as the number of patterns ever experienced by the system will never approach this capacity, and therefore dense codes usually have high statistical redundancy. The price to pay for the potential (but unused) high information content of each pattern is that the number of such patterns that an associative memory can store is unnecessarily low. The mapping between a dense representation and an output can be complex (a linearly nonseparable function), therefore requiring multilayer networks and learning algorithms that are hard or impossible to implement biologically. Even efficient supervised algorithms are prohibitively slow, requiring many training trials and large amounts of the kind of training data that is labeled with either an appropriate output or reinforcement. Such data is often too risky, time consuming, or expensive to obtain in a real system. Distributed representations in intermediate layers of such networks ensure a kind of automatic generalization, however, this often manifests itself as unwanted interference between patterns. A further serious problem is that new associations cannot usually be added without retraining the network with the complete training set.

Sparse codes combine advantages of local and dense codes while avoiding most of their drawbacks. Codes with small activity ratios can still have sufficiently high representational capacity, while the number of input-output pairs that can be stored in an associative memory is far greater for sparse than for dense patterns. This is achieved by decreasing the amount of information in the representation of any individual stored pattern. As a much larger fraction of all input-output functions are linearly separable using sparse coding, a single supervised layer trained by simple supervised or reinforcement learning methods, following perhaps several unsupervised layers, is more likely to be sufficient for learning target outputs, avoiding problems associated with supervised training in multilayer networks. As generalization takes place only between overlapping patterns, new associations will not interfere with previous associations to nonoverlapping patterns.

Distributed representations are tolerant to damage to the units or noise. However, redundancy far smaller than that in dense codes is sufficient to produce robust behavior. For instance, by simply duplicating units with 99% reliability (assuming independent failures), reliability increases to 99.99%. Sparse representations can be even more tolerant to damage than dense ones if high accuracy is required for a representation to be recognized or if the units are highly unreliable.

Sparseness can also be defined with respect to components. A busy scene may be encoded in a distributed representation while, at the same time, object features may be represented locally. The number of simultaneously presented items decreases as activity ratio increases because the addition of active units eventually results in activation of so many units, that the representation of unrelated items ("ghost" items) will be apparent in the set of active units.

To utilize the favorable properties of sparse representations, densely coded inputs must be transformed into sparse form. As the representational capacity of sparse codes is smaller, this cannot be achieved perfectly for all possible patterns on the same number of units. Information loss can be minimized by increasing the number of representation units or by losing resolution - but only in parts of pattern space that are usually not used. Both measures seem to be taken in the cortex. First, the number of neurons in the primary visual cortex is about two orders of magnitude higher than the number of optic nerve fibers that indirectly provides its input. Second, the natural sensory environment consists of patterns that occupy only a small fraction of pattern space; that is, it has large statistical redundancy. Barlow (1972) suggested that it is the nonaccidental conjunctions, "suspicious coincidences" of features, or "sensory clichés" that must be extracted that give good discrimination in populated regions of pattern space. By making explicit, local representations for commonly occurring features of the natural environment, such as facial features, our visual system is much better at discriminating natural images than, for instance, random dot patterns. As events are linked to the causes of sensory stimuli in the environment, such as objects, rather than arbitrary

combinations of receptor signals, associations can be made more efficiently, based on such direct representations (Barlow, 1991). There is substantial evidence that the visual system is well adapted to the special statistical properties of the natural visual environment by utilizing sparse coding (Baddeley et al., 1997; Vinje and Gallant, 2000, Simoncelli and Olshausen, 2001).

A simple unsupervised algorithm for learning such representations in a nonlinear network using local learning rules has been proposed (Foldiak, 1990) which uses Hebbian forward connections to detect nonaccidental features, an adaptive threshold to keep the activity ratio low, and anti-Hebbian, decorrelating lateral connections to keep redundancy low. Simulations suggest that these three constraints force the network to implement a sparse code with only little information loss. Another interesting effect can be observed: high probability (i.e. known or expected) patterns are represented on fewer active units while new or low probability patterns get encoded by combinations of larger numbers of features. This algorithm is not limited to detecting only second-order correlations, so it seems suitable for multilayer applications. Other algorithms adjust connection weight by explicitly maximizing measures of sparseness, successfully producing V1 simple cell-like receptive fields (e.g. Olshausen and Field, 1997), though the biological implementation here is less direct.

Sparse Coding in the Cortex

It is easy to measure sparseness in network models, where the responses of all units can be observed. An idealized "wavelet" filter model of simple cell responses in primary visual cortex has shown that wavelet coefficients of natural images show high kurtosis; that is, for natural images, most wavelet units have outputs near zero and only a small subset of units gives large responses, there being a different subset for each image (Field, 1999). Evaluating the sparseness of coding in brains, however, is difficult: it is hard to record a set of neurons simultaneously across which sparseness could be measured. Techniques, such as optical recording and multiple electrode recording, may eventually yield data on the density of coding, but there are presently formidable technical difficulties to overcome. We have more information about neurons' breadth of tuning across various stimulus sets than about sparseness per se. Coding across stimuli and across cells are, however, closely related (Table 2). For instance, the sparseness averaged across stimuli and narrowness of tuning averaged across units must be equal.
What evidence is there for sparse coding from single unit recordings in sensory cortex? The most immediate observation during physiological experiments is the difficulty of finding effective stimuli for neurons in most cortical areas. Each neuron appears to have specific response properties, typically being tuned to several stimulus parameters. In primary visual cortex, many neurons only respond strongly when an elongated stimulus, such as a line, edge, or grating, is presented within a small part of the visual field, and then only if other parameters, including orientation, spatial frequency (width), stereoscopic disparity, and perhaps color or length fall within a fairly narrow range. This suggests that at any moment during the animal's life, only a small fraction of these neurons will be strongly activated by natural stimuli. The problem of finding the preferences of cells becomes severe in higher visual areas, such as area V4, and

		cell								
		1	2	3	4	5	6	7	8	9
	1			0						
	2			0						
	3	1	0	1	0	0	1	0	1	1
stimulus	4			0					sparseness	
	5			1						
	6			0						
	7			0	breadth of tuning					

Table 2. Sparseness vs. breadth of tuning

especially in infero-temporal cortex (IT). Cells' preferences in IT are often difficult to account for by reference to simple stimulus features, such as orientation, motion, position, or color, and they appear to lie in the domain of shape (Gross, Rocha-Miranda, and Bender, 1972; Perrett et al., 1982, Tanaka, 1996). Cells here show selectivity for complex visual patterns and objects, such as faces, hands, complex geometrical shapes, and fractal patterns, and the responses are usually not predictable from responses to simple stimuli. Cells responding to faces but not to a large collection of control stimuli could be considered, on the one hand, to be very tightly tuned cells in the space of all possible stimuli. On the other hand, they may have quite broad tuning and show graded responses to stimuli within the specific categories for which they show selectivity. To estimate, therefore, how often these cells are activated in behaving animals would require much more accurate knowledge of the animals' visual environment and their behavior, or access to the cell's response during natural behavior.

Cells with apparent selectivity for faces might be selective for the full configural and textural information present in a preferred face stimulus, or be triggered simply by the presence of two roughly collinear bars (most faces have eyebrows), or a colored ovoid. One of the possible approaches to explore IT cells' preferences have been widely employed (Gross et al., 1972) and has been applied as systematically as possible by Tanaka and his colleagues (Fujita et al., 1992; Tanaka, 1996). Using this approach they try to determine preferred features of cells by simplifying the stimuli that excite them. This method begins by presenting many objects to the monkey while recording from a neuron to find objects that excite it. Component features of effective stimuli, as judged by the experimenters, are then presented singly or in combination. By assessing the cell's firing rate during presentation of each simplified stimulus, the protocol attempts to find the simplest feature combination that maximally excites the cell. This approach suffers from the problem that even simple objects contain a rich combination of color, orientation, depth, curvature, texture, specular reflections, shading, shape, and other features that may not be obvious to the experimenters. As any feature combination may be close enough to the preferences of a cell for it to become excited, the simplified stimuli that are actually presented are only a small subset of all possible combinations, selected according to the experimenter's intuitions. Hence, it is not possible to conclude that the best simplified stimulus found using this method is optimal for the cell, only that it was the best of those presented. A possible improvement in this area may come from automated neurophysiological experimental stimulus optimization procedures, which use more objective on-line search algorithms in a closed loop experimental design to find peaks of neural tuning curves in high dimensional (e.g. image) spaces (Foldiak, 2001). However, it may not always be safe to assume that cells code only one optimal set of features, since it is possible that they could exhibit two or more maxima, corresponding to quite different feature combinations (Young, 1993). Recent results from intermediate stages of visual processing, area V4, however, suggest that cells at these stages encode well localized individual contour components independently of global shape configuration (Pasupathy and Connor, 2001).

Even at higher stages, e.g. in IT, cells can show preferences for patterns that are simpler than real visual objects. One interpretation of these results is that IT might consist of a large number of detectors of pattern "partials," which together might constitute an "alphabet" (Stryker, 1992). The detection of such partials would seem to suggest that these cells will have broader tuning than cells with selectivity for the full configuration. The idea that an IT cell reliably signals the presence of the particular pattern "partial" seems not to be supported by results of Tanaka et al. (1991), who showed that the presence of other visual features can disrupt the cell's response to its "partial," a result which is inconsistent with the visual alphabet concept. Hence, the simplification approach captures neither necessary nor sufficient descriptions of the behavior of IT cells, and does not yet present a clear message on the sparseness of representation.

Finally, we note a difficulty for all attempts to measure sparseness in the cortex. In the extreme case, a cell with tuning so precise that it responds only to a single object will sustain its firing near its background rate when shown anything else. Researchers have only limited time and stimuli available to explore the cell's preferences during an experiment, and invariably go on to the next unit if they cannot determine what it is that the cell prefers, which strongly biases estimates of the specificity distribution. So whether there are any cells with extremely high specificity (approaching the specificity of "grandmother cells"), we cannot expect to find them experimentally using current methods. On the other hand, a cell that appears to respond only to a very limited number of a set of stimuli, as for example some human medial temporal lobe cells shown in Heit, Smith, and Halgren (1988) and the very tightly tuned cell from monkey temporal cortex shown in Young and Yamane 1993), cannot be interpreted as conclusive evidence for extremely narrow tuning because of uncertainty about their responses to untested stimuli.

Beyond sparseness

While breadth of tuning and sparseness are interesting issues both experimentally and theoretically, these issues only cover a minor aspect of neural representation. A much more significant question is what the components of a representation stand for. Imagine a sparse code, where each information item is represented by a randomly selected subset of n units from a pool of N units (e.g. the ASCII code). Alternatively, each unit could represent some meaningful aspect or feature of the item, and each item would be encoded by the combination of meaningful properties or features that are applicable to the item (Barlow, 1972). This scheme could have the same sparseness as the random scheme but the random scheme would only allow us to identify whether the item is present in the representation. The alternative scheme, however, allows much more than that. It allows us to attribute meaning to it. It allows us not only to determine the degree of similarity between items, but the representation of the items also implicitly contain the description of the kind of similarity present between the items. It also allows the system to deal with unknown or new items by generalizing along the relevant dimensions.
Much of the neurophysiological data from high-level visual cortex support Barlow's hypothesis that the neural code is not only sparse, but that the elements of the code stand for meaningful features of the world, such as complex shapes, object-components and faces (Gross, Rocha-Miranda, and Bender, 1972; Perrett et al., 1982, Tanaka, 1996), and even intermediate stages of the ventral visual pathway show selective responses to interpretable shape primitives and contour features, such as angles and curves (Pasupathy and Connor, 2001).

DISCUSSION

The theoretical reasons and experimental evidence discussed here support the hypothesis that sparse coding is used in cortical computations, while the degree of sparseness is still a subject for future research. Even more important may be exploration of the relationship between the responses of single neurons and significant, meaningful features and aspects of the sensory orld. The full description of high-level cells will require far more detailed knowledge of their anatomical connectivity and better understanding of the lower-level sensory neurons out of which their responses are constructed.

REFERENCES

Baddeley, R., Abbott, L. F., Booth, M. J. A., Sengpiel, F., Freeman, T., Wakeman, E. A., Rolls, E. T., 1997, Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proceedings of the Royal Society of London. B264:1775-1783.

Barlow, H. B., 1972, Single units and sensation, Perception, 1:371-394.

Barlow, H. B., 1991, Vision tells you more than "what is where," in Representations of Vision (A. Gorea, Ed.), Cambridge, Eng.: Cambridge University Press, pp. 319-329.

Dayan, P., Abbott, L.F., 2001, Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, MIT Press, Sparse Coding, pp. 378-383.

Field, D.J., 1999, Wavelets, vision and the statistics of natural scenes, Philosophical Transactions of the Royal Sociaty of London A, 357 (1760): 2527-2542.

Foldiak, P., 1990, Forming sparse representations by local anti-Hebbian learning, Biol. Cybern., 64:165-170.

Foldiak, P., 2001, Stimulus optimisation in primary visual cortex, Neurocomputing, 38-40 (1-4) 1217-1222.

Fujita, I., Tanaka, K., Ito, M., and Cheng, K., 1992, Columns for visual features of objects in monkey inferotemporal cortex, Nature, 360:343-346.

Gross, C. G., Rocha-Miranda, C., and Bender, D., 1972, Visual properties of neurons in the inferotemporal cortex of the macaque, J. Neurophysiol., 35:96-111.

Heit, G., Smith, M., and Halgren E., 1988, Neural encoding of individual words and faces by the human hippocampus and amygdala, Nature, 333:773-775.

Olshausen, B. A., Field, D. J., 1997, Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research 37:3311-3325.

Pasupathy A, Connor CE, 2001, Shape representation in area V4: Position-specific tuning for boundary conformation, Journal of Neurophysiology 86 (5): 2505-2519.

Perrett, D.I., Rolls, E.T., Caan, W., 1982, Visual neurons responsive to faces in the monkey temp oral cortex, Exp. Brain Res. 47 (3): 329-342.

Stryker, M., 1992, Elements of visual perception, Nature, 360:301-302.

Simoncelli, E.P., Olshausen, B.A., 2001, Natural image statistics and neural representation, Annual Review of Neuroscience, 24: 1193-1216.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M., 1991, Coding visual images of objects in the inferotemporal cortex of the macaque monkey, J. Neurosci., 6:134-144.

Tanaka K, 1996, Inferotemporal cortex and object vision, Annual Review of Neuroscience, 19: 109-139.

Vinje, W.E., Gallant, J.L., 2000, Sparse coding and decorrelation in primary visual cortex during natural vision, Science 287 (5456): 1273-1276.

Young, M. P., 1993, Visual cortex: Modules for pattern recognition, Curr. Biol., 3:44-46.

Young, M. P., and Yamane, S., 1993, An analysis at the population level of the processing of faces in the inferotemporal cortex, in Brain Mechanisms of Perception and Memory (T. Ono, L. Squire, D. Perrett, and M. Fukuda, Eds.), New York: Oxford University Press, pp. 47-70.