

Magnification in Neural Information Channel Maps

M. D. Plumbley

Division of Engineering, King's College London

Strand, London WC2R 2LS, UK

Email: Mark.Plumbley@kcl.ac.uk

ABSTRACT: Many cortical maps are found in biological organisms. In these maps, related units tend to be collected close to each other, and important inputs typically have a magnified representation on the map. We consider an interpretation of these structures as maps of parallel information-bearing channels, rather than winner-take-all maps such as the Kohonen self-organizing map (SOM). By considering the information transmitted across the map in the presence of output noise and a total power limit, magnification arises naturally as a requirement for most efficient representation of information by the map.

1 INTRODUCTION

The Kohonen (Kohonen, 1982) self-organizing map (SOM) is an attractive and useful neural network which, in its two-dimensional form, allows complex data to be visualized on a two-dimensional surface. Part of its attractiveness is undoubtedly the link it has with the cortical maps that are found in biological organisms. These biological maps include the map of frequencies in the bat auditory cortex, the map of the visual field in the human visual cortex, and maps of skin sensors in a large number of animals.

A common feature of these cortical maps is that the representation scale is non-uniform: some regions seem to be magnified when compared with others. For example, the area of the visual cortex which corresponds to visual input to the fovea is magnified much more than that for peripheral vision; the cortical area given over to touch sensors in the lips and fingers is proportionally much greater than that given over to the back; and the bat has a much magnified representation for frequencies around its echo-location frequency than elsewhere (Anderson and Hinton, 1981).

The areas which are magnified often seem to correspond to the more important sensory input regions, which are more critical to the survival or general performance of the animal. In our example above, information from the fovea is required for discernment of detail in an object; information from the fingertips or lips is needed for fine manipulation of tools or food; while the detailed frequency differences around the bat's echo-location frequency are important if it is to find and catch its prey.

In this paper we will attempt to quantify this idea with the help of information theory and the concept of an *information channel map*. This differs from the usual Kohonen SOM in that the output units have graded activities that are independent, rather than taking the form of a winner-take-all arrangement. We will see that attempting to optimize transmission of information across such a map, which is subject to noise and power constraints, can naturally lead to a requirement for the 'important' parts of the input to be magnified.

2 INFORMATION THEORY

If we are to quantify what we consider to be 'important' about a sensory input, one measure we could use would be the *information* it gives us. Information has been used by a number of authors in the neural networks field in recent years.

Linsker (1988) suggested that a neural network should try to maximize the information transmitted from its inputs to its outputs: this is his *Infomax* principle. A number of authors have found this information-theoretic approach to be fruitful. For example, Atick and Redlich (1992) found that optimization of information gave a good prediction for the shape of retinal filters, and Plumbley (1993) found it useful for the analysis of learning in decorrelating networks.

The central concepts from information theory are *entropy* $H(X)$ of a random variable (RV) X and the *mutual information* $I(X, Y)$ between two RVs X and Y (Shannon, 1949).

Entropy is a measure of uncertainty in a RV. For a discrete RV which taking values $X = x_i$ with probabilities p_i it is defined as

$$H(X) \triangleq - \sum_i p_i \log p_i \quad (1)$$

using $p \log p = 0$ when $p = 0$. This entropy $H(X)$ is always non-negative, is maximized when the probabilities p_i are all equal (maximum uncertainty), and zero (i.e. minimized) when $p_i = 1$ for some i and zero for all others. In the latter case, the value of the RV is known with probability 1, and there is no uncertainty.

We can think of the *mutual information* $I(X, Y)$ between two RVs X and Y as the entropy (uncertainty) of X before we know Y , less the remaining entropy of X after we know Y . In other words, how much uncertainty in X do we remove by knowing Y ? It can be shown that mutual information can be directly defined as

$$I(X, Y) \triangleq \sum_{ij} p_{ij} \log \frac{p_{ij}}{p_i p_j}. \quad (2)$$

See e.g. Haykin (1994) for an introduction to Information Theory in a neural network context.

3 SELF-ORGANIZING FEATURE MAPS

The Kohonen (1982) self-organizing map (SOM) is a 2-layer neural network, with one layer of input neurons, and one layer of output neurons. The input neurons are set to component values of the current input pattern vector. The output layer of neurons is a winner-take-all (WTA) network, so that one and only one output neuron is ‘on’ for each input pattern: the remainder are ‘off’. Each output neuron has a *target* input, represented by a weight vector, to which it will respond most strongly: the neuron whose target is closest to the current input pattern gets to turn on.

The output neurons are arranged on a spatial grid (often 2-dimensional and rectangular), which is used during the network’s learning process. The learning algorithm forces spatially neighbouring neurons on the grid to respond to similar input vectors. In addition, areas of the map which correspond to input patterns \mathbf{x} which occur with high probability density $p(\mathbf{x})$ are given a larger representation (although not necessarily directly proportional).

Information theory has already been used in SOM learning. Linsker (1989) derived an alternative learning algorithm for the SOM by maximizing the mutual information between the input and the output of the map, and Holthausen and Breidbach derived a number of alternative feature map algorithms from entropy measures (Holthausen and Breidbach, 1997). However, the winner-take-all mechanism means that only one node is on at any time, no matter how large the grid. From the point of view of efficient information transmission, this may be rather inefficient, since it would usually make more sense to use all of the different outputs independently. We shall therefore consider how magnification effects might arise in a map where many nodes in the map are free to have non-zero values at the same time. This leads to us considering a map of information channels – an *information channel map* – rather than a map of features – a *feature map* (Fig. 1).

4 INFORMATION DENSITY ACROSS AN INFORMATION CHANNEL MAP

In an information channel map, we would like to define what proportion of the total information is transmitted by each channel. However, entropy (and information) can be ‘shared’ by a number of variables (or channels), so that the total entropy is less than the sum of the individual entropies (i.e. in general we have $H(X, Y) < H(X) + H(Y)$).

We hypothesize that it is possible to define an information density measure that allows us to do this. To be useful and intuitively reasonable, it is likely to have at least the following properties: summability, symmetry, and accountability.

By *summability*, we mean that we should be able to integrate entropy distribution across a domain to obtain the total entropy represented across that domain. In the case of a map with a finite number of output neurons, we should simply be able to sum the entropy distribution assigned to each of the neurons in the map to get the total amount of information distributed across the map.

Considering a ‘map’ with just two outputs X and Y , since we know that $H(X, Y) < H(X) + H(Y)$ if X and Y are not independent, then the portion of total entropy that is distributed to X and Y will be less than the entropy of each point calculated alone.

This leads to the second property we would like: that of *symmetry*. If some of the total entropy in the output is represented equally at two points (e.g. if their responses are completely correlated then the portion of

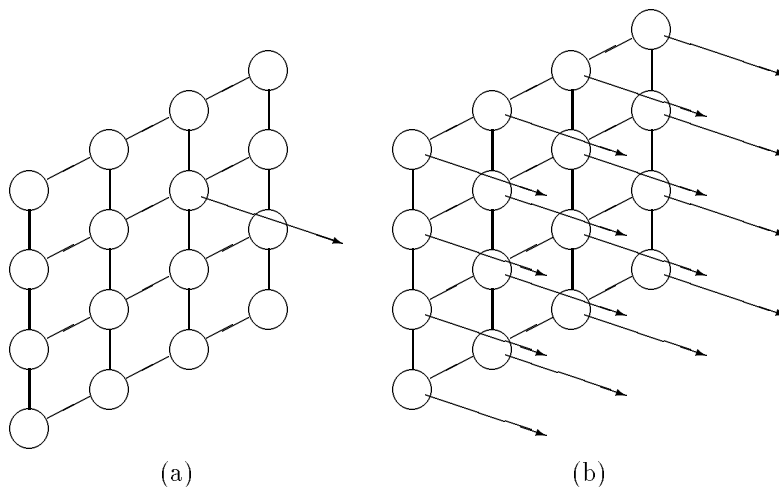


Figure 1: Information is transmitted through a winner-take-all feature map, such as the Kohonen Map (a), by the identity of a single ‘on’ neuron. We are searching for an alternative model for an information channel map, where information is transmitted in parallel (b).

the entropy distribution assigned to these two points should also be equal. For example, if the response of two neurons was always exactly the same, we should expect them to have the same information density. In other words, swapping variable labels should not change the results of the information distribution.

A final property we would like is *accountability*. If one of the variables in the map contains all the entropy (or information), we would like the entropy distribution measure to reflect that. So if $H(X, Y) = H(X)$ while $H(Y) = 0$, then our entropy distribution should show that X is delivering all of the total entropy.

Tackling this problem of defining a suitable information density measure is the subject of current work (Plumbley, 1997; Plumbley, 1998).

5 EXAMPLE: ONE-DIMENSIONAL GAUSSIAN CHANNEL MAP

For the purposes of this example, we shall assume that we can approximate our grid of neurons by a continuous neural field, and that it is possible to define an information density function over this field. In particular, we would like the concept of *information capacity density*, the maximum value of information density at a point in a neural field, to be well-defined.

In this example, the neural field is one-dimensional, with position v on the field. Consider a normally distributed signal $Y(v)$ with mean zero and variance $S(v)$ at v . Suppose that it is subject to additive noise of variance $N(v)$, and total power (variance) limit $P = \int S(v)$.

It is a well-known result from information theory, that the total information capacity of this system is maximized when $S(v) + N(v)$ is a constant (Shannon, 1949). For example, if $N(v) = N$ is constant with v (so-called *white* noise), then we should have $S(v) = S$. In this case the information capacity density, which is $0.5 \log(S(v)/N(v))$, will also be uniform.

If v is in fact a frequency f , then the obvious solution to this is the well-known *whitening filter* (Shannon, 1949). In this case, we simply apply a multiplier to the source of the signal $S(f)$ at each f so that the resulting signal is uniform over f . This is the solution achieved by a simple linear filter.

The disadvantage of this weighting method is that it allocates equal information density to all frequencies. The solution is restricted to this since it is not a simple matter to change the frequencies used to transmit the information. However, if we are not working in the frequency domain, we have more flexibility in the solution we can adopt. In this case, it is possible to allocate more information density to certain areas.

6 MAGNIFICATION

Suppose now that we have a system with input random function $X(u)$ defined over some continuous domain u , and an output random function $Y(v)$ similarly defined over some continuous v . Suppose also that we have

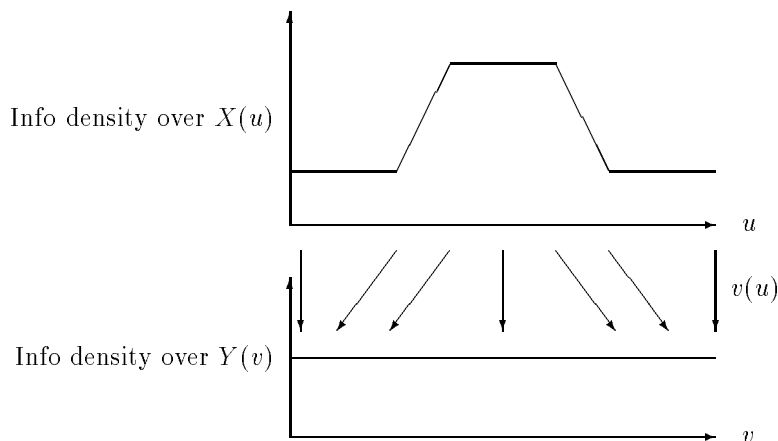


Figure 2: Magnification of input range over some regions of u , allowing a non-uniform information density over u , while we have a uniform information density over v .

a continuous map from an input range u to an output range $v(u)$. The output $Y(v)$ is determined from $X(u)$ ‘locally’ to $v(u)$, i.e. changes to $X(u)$ will cause the most significant effects in $Y(v)$ ‘close’ to $v(u)$.

In this case, we can get a non-uniform information capacity density by modifying the mapping $v(u)$. The information capacity density will be scaled in the forward direction by $1/(dv/du)$. Thus if we arrange to have uniform density at the output $Y(v)$, the input information density local to an input point u will be proportional to dv/du (fig. 2). The factor dv/du is the *magnification factor*.

If the information capacity density of the output field is uniform, this forward magnification factor dv/du is then simply proportional to the information density we wish to have in the input field at u .

7 DISCUSSION

We have not covered the locality aspect of map formation here, only the magnification aspect. This information density measure as used here does not tell us why maps should form with local responses close to each other. It is expected that the cost of long-distance computation may provide the answer here.

However, this approach does give us an information-theoretic reason why ‘important’ parts of sensory inputs might be magnified in cortical maps. If the important areas correspond to those where high information density is required, then a magnified representation in a cortex which is power-limited and subject to noise will give that higher information density.

For the human visual system, the magnification factor around the fovea from the retina to the visual cortex allows the information capacity density to be concentrated around the centre of the visual field. For the bat, this view suggests that its audio cortex should have the magnified representation around its radar frequency, since the magnified representation gives the bat a much larger information density around that frequency than it does at other frequencies.

8 CONCLUSIONS

Cortical feature maps are found in many biological organisms. We have suggested that the magnification of important regions which is found can be explained by considering the maps to be formed of parallel information-bearing channels. This is a slightly different approach to the more usual winner-take-all maps used in, for example, the Kohonen SOM.

We hypothesize that we can define a measure of information capacity density across an information channel map. If we wish to optimize information capacity across such a map which has uniform output noise and a limited total power (variance) available, we should spread the information capacity density uniformly over the map.

This uniform information capacity density could be achieved by scaling inputs values at each position, but this would result in the input information capacity density also being uniform. If a non-uniform information

capacity density is desired at the input (e.g. the fovea needs a higher information capacity density) then the representation of this region should be magnified in proportion to the information capacity density required.

REFERENCES

- Anderson, J. A. and Hinton, G. E. (1981). Models of information processing in the brain. In Anderson, J. A. and Hinton, G. E., editors, *Parallel Models of Associative Memory*, pages 9–48. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Atick, J. J. and Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation*, 4:196–210.
- Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Macmillan, New York.
- Holthausen, K. and Breidbach, O. (1997). Self-organized feature maps and information theory. *Network: Computation in Neural Systems*, 8:215–227.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.
- Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer*, 21(3):105–117.
- Linsker, R. (1989). How to generate ordered maps by maximising the mutual information between input and output signals. *Neural Computing*, 1:402–411.
- Plumbley, M. D. (1993). Efficient information transfer and anti-Hebbian neural networks. *Neural Networks*, 6:823–833.
- Plumbley, M. D. (1997). Towards an information density measure for neural feature maps. In *Proceedings of the 1996 NEuroNet Workshop on Independent and Principal Component Analysis Methods in Image Analysis and Video Processing*, Thessaloniki, Greece.
- Plumbley, M. D. (1997). Information density and cortical magnification factors. In Baddeley, R., editor, *Proceedings of the 2nd Workshop on Information Theory and the Brain, Newquay, Sept 1996*. In press.
- Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37:10–21.