The Engine of Thought — a Bio-Insipred Mechanism for Distributed Selection of Useful Information

Harri Valpola

Department of Biomedical Engineering and Computational Science
Helsinki University of Technology, Finland

Abstract

In humans, consciousness is a process taking place on the neocortex. I argue that from an information-processing perspective this process is, among other things, distributed selection of useful information. I review simulation results from a computer model of the process. The model is able to learn abstract categories and invariant features in an unsupervised manner. It is also able to segment out individual objects and switch between different objects. Object representations emerge from subsymbolic representations.

1. Introduction

My research group is developing a brain-based cognitive architecture. The goal is to understand the information-processing principles of the brain and to apply them for building intelligent machines. We are roughly following the path taken by the evolution, using the vertebrate, mammalian and eventually human brains as our guide along the way.

Our guiding principle is that the brain has evolved for intelligent *behaviour*. In order to understand the brain, we need to understand the problems for which the brain is the solution. For the most part of the evolutionary history of the brain, the problem was very concrete and related to interaction with the environment. This is why we use robotics as our target application area and focus first on sensorimotor coordination rather than higher cognition and symbolic reasoning. Neuroscience has taught us that higher cognition relies on the same brain structures which originally evolved for concrete motor tasks. We therefore think that once we understand the computational principles of various parts of the brain, we can apply the same principles to higher cognition, too.

Currently we have models of the cerebellum, basal ganglia and neocortex up and running, and we are later going to include a model of the hippocampus. In my talk, I will focus on the neocortex because it is the part which is most relevant for consciousness. I will take an information-processing perspective: consciousness is the process which happens on the neocortex. My aim is to explain what this process is doing, for what purpose and what kinds of emergent properties it has.

In a nutshell, the neocortex is, among other things, a distributed information selection system. The cortex consists of a large number of processing elements, cortical areas, organized into a hierarchy. Each area processes bottom-up information and selects the bits which it deems important in light of the information which it receives from the others (top-down and lateral input).

2. Bayesian decision theory

Before going into further details, let me first give a bit of theoretical background: Bayesian decision theory is the golden standard of intelligent behaviour. The recipe is tremendously simple:

1) First gather *all* information about the world, combine this into a distribution which quantifies how much the agent believes in different propositions (states of the world, facts, outcomes of actions, etc.). This information is represented in terms of the joint probability of all the relevant quantities and can be computed following a few simple rules (Bayes' rule and the marginalization principle).

2) Then choose the action which maximises the expected utility—in other words, the action which the agent believes will, on average, lead to the best outcome.

In Bayesian decision making, one is first supposed to lay down the facts (all the facts) and *then* make an informed decision. Decision making is neatly separated into two stages: analysis and action selection. Computational issues aside, the Bayesian recipe has been shown to be optimal. It is simply the clever thing to do. It is also the optimal way to learn since any parameters of the internal model can be treated just as any other unknown variables of the external world.

Recently there has been a lot of interest in "the Bayesian brain" since many aspects of cortical processing can be explained from a Bayesian point of view. However, beautyful and powerful as the theory is, one runs into serious difficulties in the real world where computational issues cannot be neglected. The problem with this "optimal" approach is that in practice there is so much information to be absorbed that the processing becomes completely overwhelmed with nitty-gritty details already in the first stage. In practice, it is never possible to consider all the possibilities (past, present and future paths of the whole universe).

3. Distributed selection of useful information

The solution which the neocortex seems to have adopted is to divide and conquer: distribute the process of analysing and selecting relevant information. Selection is a type of decision and in theory it is suboptimal to "jump into conclusions" before all the available information is integrated and all the possibilities considered. In practice, however, selection is necessary in order to avoid choking the system with an overwhelming amount of irrelevant information.

Cortical areas can be considered as agents whose goal is to maximise the amount of useful information and, importantly, minimise the amount of irrelevant information: to find *something new and interesting*. Let me elaborate a bit:

> *Something*. Each area has its own "receptive field" from which it receives information. Other things being equal, it is sensible to try to maximise the amount of information that the area passes forward. This can be accomplished by minimising the reconstruction error of the inputs. Often the inputs are noisy and it pays to use "prior information" to make sense of them. In the cortex, there are numerous top-down and lateral projections that "modulate" the processing. The Bayes rule tells how predictions from lateral and top-down information can be used for improving the estimate of what the bottom-up information means.
>
> *New*. It is useless for a cortical area to represent some piece of information if the other areas already represent the same thing. While top-down and lateral predictions about the bottom-up information may increase the probability of a given feature actually being present, a predicted feature also loses value. It no longer pays to represent it. It has been shown that in noisy, low-contrast situations, predictability increases the cortical responses while the opposite is true in high-contrast situations.

> *Interesting*. How do the cortical areas know which pieces of information are relevant? I suggest that top-down and lateral connections are used here, too. If a cortical area finds something new to represent and afterwards the others follow suit, that must have been an important bit of news. Whether others are following can be deduced from the same mapping which is used for predicting the bottom-up features in the previous steps. If bottom-up activity follows the prediction, it was old news, but if the order is reversed, it was something which caught the interest of the others.

The above three points can be implemented with rather simple mechanisms because they all rely on the same mapping with relates the top-down and lateral context with the local bottom-up information. The first point ("something") is about probabilities and answers the question "what is it?" The two other points are about utility: "is it important?"

Ultimately, there are subcortical structures such as cerebellum and basal ganglia which are "by design" interested in certain types of information (e.g. information which predicts motor responses or reward) and which give feedback to the cortex about the relevance of information. This evaluation of relevance trickles down the cortical hierarchy all the way down to primary sensory.

Just as the "optimal" Bayesian approach, this distributed recipe can be used both for selection of relevant information (behavioural timescale, here and now) and learning, which can be considered as selection over longer timescales (select which types of features will be considered at all). The main difference is that the distributed approach relies on specialised experts with a limited scope rather than on a single omniscient decision maker.

It is interesting to note that similar distributed learning, analysis and valuation of information can be identified in the information exhange in human communities such as media or science (cf. citation index).

4. Biased competition and competitive learning give rise to attention and abstractions

The above picture of the cortex as a distributed selection system is by no means totally new. Desimone and Duncan suggested that selective attention results from locally competing neural populations which are reciprocally connected. When long-range connections between the areas bias local competition, a process with the characteristic features of selective attention should emerges. Deco's group has shown with simulations that this so-called biased-competition model of attention agrees very well with neurophysiological and psycophysical experiments.

Our group has extended this work by including learning in the system. It is actually a rather straight-forward idea to couple biased competition with competitive learning which has been used for learning representations for a long time. We have been able to show that this biased competitive learning gives rise to a hierarchy of increasingly abstract representations, much as those found in the neocortex.

Figure 1 gives an example of what a hierarchy of distributed areas might look like.

5. Emergent symbols and the train of thought

The messages used by the brain to communicate between areas (cortical or subcortical) are relatively simple patterns of neural activity. In particular, the code is distributed. Each neuron conveys information about a certain feature of the input and objects thus need to be coded by a population of neurons. This works fine for a single object but leads to potential confusion if multiple objects are present. This is the so-called binding problem.

The binding problem can be solved by representing only one object at a time. The machinery for selecting relevant information can easily be tuned such that it values the *coherence* of the representation. It is enough to favour those pieces of information which agree with top-down and lateral context.

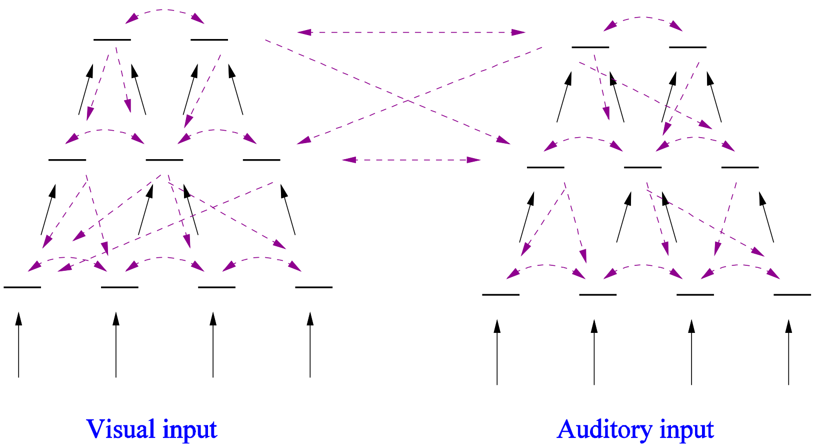

Figure 1. An example of a distributed selection process. The neurons on each area (horizontal bar) receive bottom-up input (solid arrows) and compete locally. Competition is biased by top-down and lateral connections (dashed arrows).

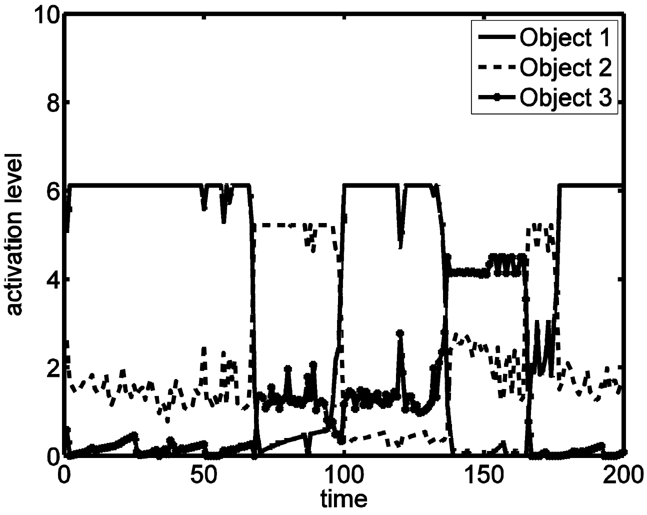

Figure 2. The average activity level of three different coalitions of neurons is shown. There were five cortical areas, each consisting of ten neurons. The neurons of different areas were connected such that there were ten coalitions (each neuron of an area belonged to one of the coalitions). Local competition biased by long-range connections makes one of the populations win the competition. Gradual fatigue starts to erode the active population but it persists because of mutual support from the neurons in the same coalition. When the weakest members of the coalition start failing, the coalition quickly crumbles and is replaced with the next lucky winners.

Without some extra mechanism, the system seeking coherence would quickly get stuck with the first stable coalition which wins out the competition with the other populations. Figure 2 shows the results of a simple simulation which demonstrates what happens when fatigue is added to the system. Fatigue here means that active neurons gradually lose sensitivity. Once some of the neurons of the active population start losing their local competition to other neurons, a domino effect makes the whole coalition collapse and a new coalition takes power.

The emergent behaviour of the system shows relatively discrete switching between individual coalitions. Moreover, learning can give rise to different numbers of coalitions depending on the number of different object categories present in the input data. Figure 3 depicts eight sample images fed into a distributed visual hierarchy. There are six different categories of objects, each with a very large number of potential instantiations. Although there have always been two different objects present in the input stimuli, the system has learned that there are six categories, each represented by a different coalition of interconnected neurons.

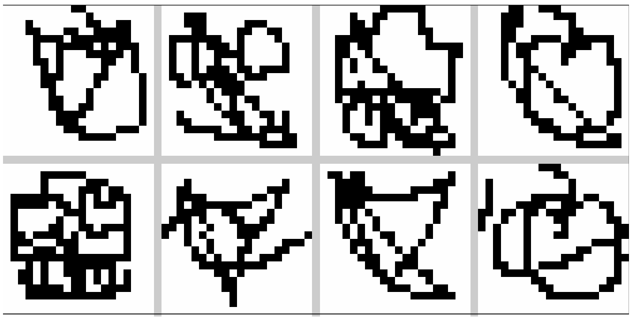

Figure 3. Eight sample images from the training set.

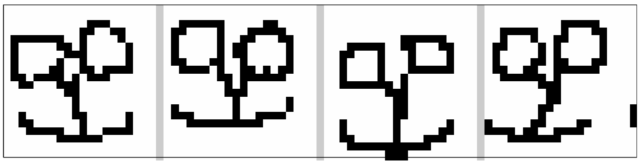

Figure 4. Four samples of test data, each of which have activated the same coalition of neurons.

Figure 4 gives an idea about the variability of inside the categories. Each of the samples has activated the same coalition of neurons. The system has learned to categorise the inputs and recognize in a totally unsupervised manner. Moreover, when shown a picture with two objects as in the training data, the representation keeps switching between two coalitions corresponding to attention switching between the two objects.

The above mentioned system had never seen dynamically changing inputs nor did it have a proper mechanism for representing dynamics. However, it would be relatively easy to add such a mechanism which uses learned dynamics to bias new coalitions. This should give rise to a train of thought, coalitions of neural activations following each other, mirroring the learned dynamics of the external world. In other words, the system would have imagination.

6. Discussion

Machine consciousness is not my goal as such but I believe that consciousness equates with the distributed selection process which is needed for intelligent machines. Rather than ask whether the process is conscious, I would like to know whether the process is useful and can support intelligent behaviour. The system will certainly need further refinements before it can cope with nonlinear dependencies between features and relations between multiple objects. However, the basic principle seems sound and it is promising that the system is able to learn complex invariances and categories without having explicitly been told about them or having seen isolated objects in “laboratory conditions”.

When thinking about how the selection process relates to consciousness, several questions come to mind. Among the easy ones are: can the system make decisions and can it be aware of its own thought processes. The answer seems to be yes to both questions. First, perceptual selection is an important part of decision making: what your visual system picks out affects what you are going to do; mass media affects political decision making by selecting the agenda; and so on. In a distributed system, the border between analysis and decision making becomes obscured and one can argue that the same process which we call attention in sensory modalities corresponds to decision making in cortical motor hierarchy. Second, if the system is able to represent processes and regularities in the outside world, there is no particular reason why it could not do the same for its own internal processes. After all, everything is just neural patterns to the system. Of course there are many other, more difficult questions. Building an intelligent system and studying it seems like a promising way to proceed.

5. Acknowledgements

The simulations and figures are based on the Master's Thesis work of Antti Yli-Krekola (2006).

References

Technical details about the simulations and references to relevant publications can be found in:

Yli-Krekola, A. (2006). *A bio-inspired computational model of covert attention and learning*, Masters Thesis, Helsinki University of Technology.