

Nervous mind

Bert G. J. Frederiks

juli, 23 2001

In order to explain consciousness, I want to explain in lay-man's words the main principles of some of the most important neural networks, being "error backpropagation" (backprop), which, because of publication problems, was reinvented 4 times, and "Adaptive Resonance Theory" (ART), which is more than 30 years old, but only now getting to be valued for what it is worth. So that's old news, but it's getting time for this news to reach the newspapers. The step from nervous tissue to the mind and consciousness, and the other way around, is full of pitfalls, but neural networks are quite boring without looking in that direction.

What is a neural network?

A neural network is a network of nerve-cells, also named "neurons", like we have in our brains. Neurons have many incoming signals, and (usually) one outgoing signal which goes to many other neurons. The cells in our cerebrum each have on average 10,000 entries. Artificial models usually only have a hand full. Cells may be connected to each other, to sensors (eyes, ears, skin, etcetera), or to muscles and glands.

Trough its inputs a neuron may be stimulated. When a neuron is stimulated above a certain threshold, it will become activated, by which it will itself start to stimulate other neurons. Important is, mathematically spoken, that this is process is non-linear. The stimulation has to take a certain threshold. In general a large part of its inputs will have to be stimulated in order for a neuron to become activated.

A neural network learns by itself

Special about a neuron is further more that the sensitivity of each of its inputs can be adjusted by the neuron itself, by which a neuron can teach itself something. The general principle here is, that if a neuron should have been activated and is indeed activated, then all inputs that have attributed positively to this will become a little more sensitive, while the others will become a little less sensitive, etc. OK, you might say, but how does a neuron "know" whether it should have been activated? And if it knows this, then why should it learn this? The simple answer is that a neuron learns this for those moments at which it isn't told whether it should have been activated. In other words, we have to make a distinction between learning and showing what has been learned. While learning we put a so-called "teaching-input" at the output of a neuron. From

this teaching-input, and its ordinary inputs, the neuron knows how to adapt itself.

You can imagine that, when a single neuron has the tendency to produce a certain output, given a certain input, then a network of neurons as a whole will, when given a certain pattern as input, tend to produce a certain pattern on its output. In other words, such a neural network will associate patterns with each other. Pattern-association, that is automatically coupling patterns which often occur together at the same time, is indeed one of the things for which a neural network can be constructed.

My favorite example of a possible use of this is that of an oil-company searching for oil. This is done by detonating explosives in the ground, and then listening to the reflected sounds. These sounds are analyzed by experts, and from this it is decided whether the chance of there being oil is large enough to start drilling. This is a complex analysis, for which intuition, or at least a clever guess, cannot be missed. But a neural network is also be particularly well suited for this analysis.

For this we create a neural network consisting of layers. At the bottom we situate an input-layer, consisting of neurons which are connected tot signals registered by microphones. Next we have some so-called “hidden” layers, and finally, at the top, an output-layer. Activation-signals run from input layer, through the hidden layers, to the output layer.

In our example the output-layer only exists of two neurons. Activation of the first means: “Yes, there is oil in the ground”, and activation of the second means: “No, there is no oil in the ground”. We train the neural network using old, recorded sounds from the archive, whereby we know through experts or drilling-results whether there was oil in the ground or not. As such we will make the neural network associate certain patterns with “Yes, there is oil in the ground”, and others with the opposite. Neural networks can in principle learn this such that they also become very good at deciding from totally knew sound-patterns whether there is oil in the ground or not—I’ll explain this further on.

The adaptation or learning of such a neural network as a whole is of course more complex than that of a single neuron, but the principles remain the same. At the output-side nothing changes; there we simply tell the neuron whether it should have been activated or not. But if this neuron knows this, then it can also tell all neurons in the hidden layer below whether they should have been activated or not, etcetera. This mechanism is called “error-backpropagation”, or “backprop” for short. It has been reinvented several times. Paul Werbos probably was the first, but they idea became famous through the PDP-group.

Remark that for almost any association the whole network participates because all neurons influence each other. Because each neuron contributes only a very little bit to any association, a neural network is able to remember many things even when some of its neurons die. In other words, knowledge is distributed over all neurons. More precisely it is distributed over the specific sensitivities between neurons.

A problem with backprop is that biological systems do not exist in this form. Another disadvantage is that these kind of networks learn quite slowly. You have to show them the patterns to be associated many hundreds of times, preferably in random order. This makes that the network can adapt as ideal as possible to all these patterns with regard to each other. An accompanying disadvantage

then is that these networks may unlearn old associations through learning new ones.

A neural network automatically abstracts features

Besides pattern-association neural networks can also be used to abstract features from patterns. For this we make a neural network such that we do not have to tell it, through a teaching input, what the features to be distinguished are, to the contrary, it will distinguish this itself.

We do this with the aid of neurons which try to switch each other off. In stead of trying to activate each other, these neurons try to inhibit each other. An inhibited neuron cannot inhibit any other neuron any more. In the most extreme case only one winner will remain. Such a winner may then—automatically—provide a teaching input for underlying neural layers. Remark that, with regard to neural activation going from input-layer to output-layer, this network is not different from the layered network described above. But perpendicular to this—that is, within each or certain neural layers—neurons try to inhibit each other.

A consequence of this neural wiring is that, if the number of possible winners is limited and smaller than the number of patterns shown to the neural network, then the network will, in the end, group patterns on distinguishing features. This means that the winning neurons each will come to stand for certain features of the patterns shown to the network.

How does this work? Take, to keep it simple, a network which only has inhibiting neurons in the top neural level. Suppose we randomly show this network two patterns, A and B . Let us start with pattern A . There will then be a winner neuron in the top neural level. Which neuron this is, is, the first time at least, pure coincidence but we know for sure that only one can win, and only one will win. Call this neuron a . Using the principle of backprop again, this winner neuron a next provides a teaching-input to underlying neural layers. As such the neural network learns to associate A with output-neuron a . If pattern A is different enough from pattern B , then this will, because of this difference, more likely not activate neuron a , but it will activate some other winner neuron, which we may call b . Remark that we create a teaching-input without a teacher. It must be said that in the simple case it could rather easily happen that both pattern A and pattern B activate the same winning output-neuron, but there exist numerous tricks to prevent this.

Now take a situation where there are more patterns than there are inhibiting neurons. It is clear then, that it is impossible for every pattern to have it's own output-neuron. But output-neurons could be coupled to resembling elements of patterns, such as certain forms or colors, in other words (general) features or properties of these patterns. A structure like this will come into existence automatically because, not only can there be only one winning neuron, there also always will be one winning neuron, and if we keep showing all patterns to the network randomly, then it will keep adapting until everything fits best.

If we allow for more than one neuron to win, then certain combinations of output-neurons will become activated, which together form an analysis or deconstruction of the input pattern. Wherever we, human beings, look at, it will immediately decompose in a number of parts. I refer to this as an attention pattern.

The principle of inhibition can be build into any neural network. It automatically makes networks a little more “intelligent”. It makes that a neural network structures itself, by finding the most important features, organizing itself around these, and by forgetting the less important details more easily. This is abstraction in a very abstract sense—I call it immanication, with immanence being the opposite of transcendence. It will, for instance, also help with regard to the example of the oil-company mentioned earlier. In our own brains there are many neurons which seem to have functions dealing with this.

A biologically more plausible model: attention and mirroring

A problem with backprop is that inside our brains error-backpropagation cannot work that way. Over 30 years ago Stephen Grossberg invented a mechanical-mathematical principle which does not have this problem. He calls his theory “Adaptive Resonance Theory”, or “ART” for short, and he explicitly connects the principle of resonance in his theory with consciousness. In completely different words than he uses I would like to try to explain his theory. In my explanation I, further more, will refer to the human brain and mind. ART as Grossberg describes it is logically and mathematically sound, but it is too detailed and complex for me to show you some of the more general issues I want to address. The fact that Grossberg brings mathematics and consciousness together in one scientific theory remains magnificent, whatever one may think of the contents of it. The fact that he wrote down his theory, and built his neural machine, so long ago, and only receives recognition know is, I think, sad, and in my opinion it is shows how science works in daily practice when we are dealing with truly knew developments.

Take two neural networks, each consisting of about seven neural layers, and lay them on top of each other, but in opposite direction with regard to each other’s input and output side. So, layer 7 lays on top of layer 1, layer 6 on layer 2, etcetera. Neurons are, in between each of these networks, connected one to one such that they tend to sensitize each other when activated themselves but without them being able to activate each other. A such two such neurons, each from the other neural network, in fact together form one single neuron—in the explanation of Stephen Grossberg it is indeed exactly this. The result should be that the two networks tend to mirror each other, but with the neural activation flowing in the opposite direction with regard to each other.

The idea is that we can perceive, abstract, and deconstruct things through the first neural network, while we can imagine and construct things through the second network. For this we connect the input-layer of the first to our senses. This network I call the deconstruction neural network. The output of this network I connect to a winner-take-almost-all neural layer consisting of mutually inhibiting neurons. This winner-take-almost-all layer actually functions as an attention mechanism. This neural attention mechanism does no more than to allow for a small number of winners to stimulate the second neural network.

So the winning neurons of the attention mechanism deliver their output on the input of the second network which I shall name the construction neural network. The attention mechanism activates a small number of neurons in the

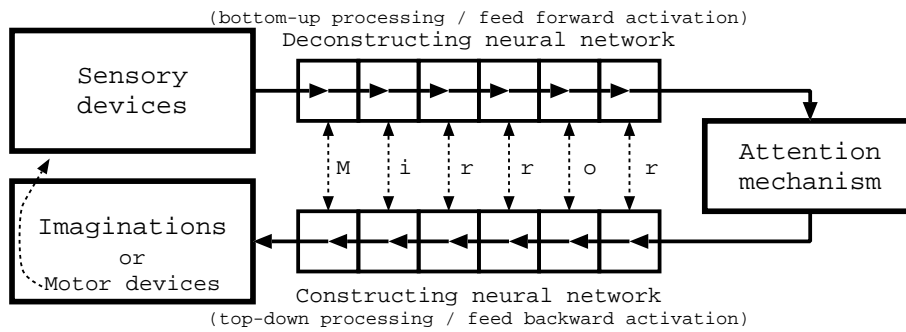


Figure 1: A hierarchical neural network split in two halves

construction neural network. From this the activation flows back in the direction of the senses, something which we experience as imagination when we do not actually see the thing imagined.

Attention-like mechanisms exist in every neural layer, but in lower neural layers there is not a single attention mechanism, instead there are hundreds to hundreds of thousands, each of which cover a small area in the neural network. Their function is abstraction. Imagine: You look at a painting. That is a complex pattern. It enters your neural network through your eyes. Inhibiting neurons bring about abstraction. In the first layer lines, colors, and simple movement is detected. In higher neural layers ever more complex forms are detected. In the end the attention mechanism only allows for five winners. From these five winners neural activation flows back through the construction neural network, in the direction of the eyes.

Because the deconstruction and the construction neural network tend to mirror each other, the back-flowing activation will tend to be a (re)construction of what is perceived. When this succeeds the network enters into a state which Grossberg calls “resonance”. Resonance is a relatively stable situation of the neural network which, given the properties of neurons, will automatically lead to strong learning for the simple reason that neurons, in this state of resonance, will stimulate, or not stimulate, each other for a relatively long period of time. In a human embryo the directions of activation will be rather random, and the accompanying resonance will therefore be little, but there too the activation will be directed such that the two networks tend to mirror each other. As such we form ourselves an imagination along with what we see. In other words, in this way something may attract our attention. The other way around, because the constructive neural network influences the constructive network in exactly the same way, our imagination of something will also make us attend to something. In this case we first, as it were, have an attention pattern, and from this choices in what we see are made. Along both routes, but especially along the latter route, a certain point of attention will lead to another point of attention, again and again. This is further more helped by the fact that specific attention naturally is quickly exhausted. As such we have one basis for ‘thinking’ which humans and other mammals have in common.

Remark that the rather transcendental information which enters my eyes—being just a bunch of colors—is transformed into the activation of only a small

number of neurons, which together form a pattern of attention. So attention in fact is the summit of abstraction. The mirroring between the constructive and deconstructive neural network thereby is a biologically possible way of error-backpropagation, because mirroring in fact entails that every difference in activation between the two neural networks is an ‘error’. This error is detected and corrected in both networks at neural level, and all this solely on the base of mechanical principles.

Although my explanation is completely different from that of Stephen Grossberg, it is, I think, largely a summary of his theory, with the addition of ideas of my own. Grossberg has proven that ART neural networks work very well, learn much faster than ordinary backprop networks, and do not suffer from massive loss of memory.

Temporality of imagination and consciousness

With regard to one thing I disagree with Stephen Grossberg. Grossberg couples consciousness to resonance. I think consciousness is impossible without temporality.

With regard to our imagination many things play a role. Especially time is very important. We imagine ourselves things by imagining them in time. Our attention shifts from one aspect to another, and we do not do this without structure. Therefore this temporal structure must, at least in some respect, be laid down in a non-temporal structure. Attention plays an important role in this, because in the attention pattern successive matters can be associated with each other. More specialized planning- and language-mechanisms are equally important, especially for humans. All in all the middle and higher neural layers make for a coherent structure in our thoughts and imagination. They form our “pre-consciousness”. Our actual consciousness only exists in time. That we have the illusion to be conscious of much more is simply caused by the fact that we are not conscious of that of which we are not conscious. We only are conscious of the color of the eyes of a squirrel at the moment that we are conscious of the color of the eyes of a squirrel.

A nice example, and also a metaphor, of temporal images, and ones the reason for me to discover all this, is the working of a movie or television program. Watching a movie, every 2 to 5 seconds a new attention pattern arises—you may have the illusion to see more, but close your eyes and count what you saw. At numerous moments in time such patterns are connected in the here-and-now, for example when a certain person, or a certain house, again attracts our attention. A movie is an image in time, that is a temporal image, but these kinds of association based on recurrent elements also form a more permanent image from which we can remember things. If associations and meanings are woven into each other well enough, then lower neural layers will develop association patterns too. Our consciousness is much like such a movie. Our consciousness is an image (a ‘knowing’) in time.

An important form of mirroring which I, be it in completely different words, have addressed is that which in semiotics, semiology, and linguistics is named “reference”. Here I refer to the mirroring of the world inside our neural network. For instance, the word “cow” refers to a cow. The activation of each neuron in an attention pattern (usually) refers to something in the world. Our neurons

are small mechanisms which are part of the world, so neurons work and are being worked on solely according to ‘physical’ principles. Memory in the form of adaption of sensitivities between neurons play a fundamental role in this mirroring. The mirroring of the world from moment to moment into a neural network—i.e. reference—is in itself of little value. Important is that a structure in time with regard to these mirrorings is remembered, such that it may later lead to comparable structures in time within this mirroring matter, being the neural network. In saying this we go much further than ‘physical principles’. Not that it is in contradiction with it. It just becomes another topic because memory and structures in time come to play a part in understanding it. Not only do we have to think and speak at another aggregation-level because we would otherwise not be able to see the trees through the wood, the system at this level also is a reality in the sense that this cloud of molecules which we call a neural network or brain is a structurally relatively stable situation which as a whole interacts with its environment. The fact that it acts as a whole is due to the structure as described above. As such it is indeed the case that ‘the content’ of the neural network, such as you or I myself, is an actor in causal relations.

Our motor system

If we make a couple of other, big leaps, connecting muscles and bones to our neural network, which are driven by our imagining of our movements (in other words, they are coupled to the output of a special part of the constructive neural network), and further more introducing some clever planning- and energy-mechanisms, plus instincts and an underlying reptile brain, then we have something which qua intellect is already much like an animal. To create a human being we have to explain a little better how language and self-consciousness works.

More words on temporality

Everything exists in time. Nothing special about that. But in what sense plays something which happened earlier a role in a later situation? You might say: I have a memory, and because of that I remember things, but what does that have to do with consciousness? Then I say: OK, but you keep speaking about “I”. Formulate it without “I” and make it into a machine, because that is what I need to explain consciousness. Without “I” one speaks of a more timeless memory, and a more temporal imagination, i.e. a more temporal consciousness. So, an important presupposition for me is that consciousness cannot stand still. I cannot imagine anything with regard to a consciousness that is frozen, but neither can I imagine consciousness without a memory. Of course memory changes too, through learning, but it does so more slowly.

Suppose a leaf whirls on the street from point A to point B. Next it whirls to point C. The fact that it first lay on B is important, because otherwise it would probably never have landed exactly on C. This is a kind of memory, but as soon as the leaf lays on C there is no memory of A and B anymore. One peculiarity of consciousness is that it does leave a traceable trail of its own ongoing, and it even does this such that the elements of this trail will return automatically

ever again when necessary. The nucleus of my story is in the relation between this remembered trail trail and the ever changing content of consciousness. This holding on I also name “mirroring”. “Mirroring” in this sense does not refer to the mirroring taking place between the constructive and deconstructive neural network. It refers to the mirroring of the world into the neural network. The result of this mirroring is what one tries do describe in Artificial Intelligence theories (AI), but they do little with the mirroring itself. In contrast for me it is essential to understand this mirroring, that is the relations between the timeless mirroring, the world, and consciousness—the latter two both being temporal.

Why is this an explanation for consciousness? I think you first have to grasp that my idea actually is a much too simple explanation for any sensible individual. Secondly my idea is terribly abstract. There are potentially many possible implementations, even though it is quite difficult to come up with examples. Thirdly, consciousness as mentioned above in fact is only interesting if it is a consciousness *of something*. For this one needs something like a neural network, but my definition of consciousness is in fact much more abstract.

Ant-hill “Aunt Hilly”

An example of a consciousness which is not a consciousness *of something* is given in Hofstadter’s book “Gödel, Escher, Bach” in the form of Aunt Hilly, that is an ant-hill. Ants, in their patterns of movement, form certain repetitive patterns, because they follow each others trails and signals. These patterns are influenced by their environment, and an ant-hill does have a tendency to preserve itself, so so consciousness of something cannot be denied to it, but what does this entail qua content? Almost nothing. As long as we realize what the content of this consciousness entails it is not strange to attribute consciousness to this.

This said it may be a question whether one should call a consciousness which is not a consciousness *of something* “consciousness”? Maybe not, but I like to resolve this by pointing out that systems which do not have consciousness *of something* usually are so very simple qua consciouness that the question is not important at all. The only exception might be AI, but here too it must be said that the topic of AI is just that which happens in the very top level of my system.

Did I answer the question of whether this is an explanation of consciousness? I did if you agree with my definition and description of consciousness. In an abstract sense my description of consciousness is in agreement with my idea of it. Since I have also been able to describe the mechanism of it, I have explained it. The idea is so very abstract that it does not seem simple anymore, but it is.

Mental illnesses

To this simple idea of consciousness one can next put thousands of questions, which all need to be checked. Take the importance of the sensibility of thoughts. Certain mental illnesses make people so confused that their consciousness is clearly limited by it. In a psychoses thoughts can be like a whirling leave, and this can destruct memory and consciousness considerably (desintegration). So sensibility has everything to do with coherence and structure, and this has

everything to do with consciousness because if there is coherence in the timeless “memory”, then the consciousness arising from this may have this coherence too. But if this coherence is too strong, then it may become suffocating. That is exactly the property of neurosis (fixation).

Another important aspect of my consciousness-thesis is that I want to deconstruct the illusion that we are conscious of many things at the same time, which is what most people seem to think. Consciousness at a certain moment is totally senseless to me. This is more an assumption than something a can prove, but I did want to explain how it seems to us that we are conscious of much more on each moment than we in fact are.

Self-consciousness of animals and people

No consciousness is fully without self-consciousness because of its relatively timeless “memory”. The past of one’s consciousness always plays a role in one’s future consciousness. Only the way in which, and the directness with which one’s consciousness participates in this can differ greatly. Human beings can systematically register their own consciousness (the mind’s eye) using mechanisms which form the foundation of language, and language itself. Animals associate events with their own consciousness (through their attention patterns), but they cannot really do this on purpose. They cannot actively associate their contents of consciousness other than through their actions because an earlier content of consciousness is for them largely gone the moment the next is there, with the exception of the timeless memory of this, of course. That would have been the same for humans, where it not that we can use language to manipulate our own and other people’s consciousness. This said, for both humans and animals successive attention patterns will usually overlap, and therefore there will arise associations, either directly or indirectly, in the same way we make them when watching a silent movie. I think animals differ considerably in their talents. Mammals and birds can certainly perceive and remember temporal patterns (in the basal ganglia and/or prefrontal, motor cortex) and through this develop a certain self-consciousness like we, humans, have. I cannot distinguish anything like this in Aunt Hipe.

Necessity of attention for consciousness

In my final definition of consciousness I left the importance of attention out. The idea is that one does not need this if one asks not whether or not something has consciousness, but what this consciousness looks like. This said, I cannot imagine how there could be a coherent consciousness in which everything comes together, and in which consciousness could gain a respectable structure, without attention. As an admirer of Carl Popper I should therefore add attention as essential to consciousness, since he would say that I should make this into a hypothesis in order to give others the possibility to falsify it. this said, I do know a replacement for the attention mechanism, namely instinct—think of cats—but in fact instinct is a kind of attention too.

Attention makes that even the strongest atheist in the end tends to arrive at a single, most important, ordering principle. In as far as someone does not

succeed in this, his thoughts, or even his personalities, remain like loose sand, for instance with ethics detached from scientific or practical thinking, which is quite wrong in my opinion.

Pain, fear, and superego

Particularly intriguing with regard to consciousness are qualia such as pain, and feelings in general. Understanding these is complex partly because old brain-structures and inborn reflexes play an important role. Instincts are probably connected to our newer brains differently than our senses and muscles. The connections are non-specific, by which I mean that an instinct stimulates many neurons. Through this instincts can globally change our perception. Suppose, for instance, that a newborn baby has the instinct to activate some area of the neural network whenever it sees two darker spots. These darker spots will usually be the eyes of mother or father. As such the human neural network will tend to save everything that father and mother do at a certain location. The other way around the instinct can later “use” this information, through activating this piece of neural network. We could analyze this as being our “superego”.

As such instinct is a kind of inborn, vague attention. A particularity of pain is that it is at the same time a very strong attractor of attention, and also enormously stimulates the neural network as a whole. Although pain itself is not really remembered as such, longterm pain must lead to restructuring of the content of the brains. It is anyway very fatiguing. The temporality of pain is in the fact that it does not stop and that it cannot be ignored. At the same time, through its attraction of attention, is also leads to a narrowing of consciousness. Fear seems to work more or less the same except that there is no other source of pain other than a thought.

A part of the experience of qualia seems to lay in the interaction with older structures of the brain. It is anyway something which we share with many animals. The only, but very distinguishing, particularity of humans is that we can imagine much more things too ourselves, and thereby we can experience much more fears. As such an animal farm is not always a concentration-camp, but vivisection remains silly and scoundrelly—even if for example cats also like to play with the death of their victims, and even if the weighing of interests is often difficult. A human being who panics is qua consciousness not much different from an animal in panic, and hunted and brutally overruled animals do not behave very different from hunted and brutally overruled humans. This is exactly what makes this profession of brain- and neural network studies is so rotten to me; there are so many brutes involved in it.

This said, we do, of course, not live in Paradise. We have to survive too. Just imagine having a manic-depressive daughter and hoping for some medical wonder to cure her—not that I expect scientists to really have other motives than curiosity, fame, and food, but it still remains an argument.

Lack and community

I want to try to give this story a nice ending... I miss something in my exposition of how feelings work. Maybe this is only the missing itself; the fact of knowing that I do not know everything; the hole in my mind, and the desire for something which I do not have, if even a solution to get rid of my pain. But something tells me there is something else. Whether I think of toothache or feelings of beauty, there seems to be more. The curious thing is that I experience this with regard to earthly matters. Consciousness and individuality; that I do understand, but... that feeling of community, and the both peaceful and crude existence; that I do not understand. It seems as if this, which I do not understand, is all that which enters my consciousness, but which lays outside my power, and outside my imagination. It steers and pushes me, and I can go along with it or resist it, or think of a trick to get it within my power anyway. It is fantastic and magnificent.

Literatuur

To contents of this article are addressed more deeply in an Internetboek by the author entitled: "The Time Machine, Prototype of a Conscious Machine", which can be found at <http://tm.bedenkerij.nl/>

Publications of Stephen Grossberg can mostly be found on the Internet too, unfortunately quite often referring to vivisection, but for the nucleus of his theory this is absolutely superfluous. See <http://www.cns.bu.edu/Profiles/Grossberg/>

James Anderson and Edward Rosenfeld wrote a beautiful book in 1998 about many the people involved in the development of neural networks: "Talking Nets. An Oral History of Neural Networks", The MIT Press: Cambridge, London.