Iconic Training and Effective Information: Evaluating Meaning in Discrete Neural Networks

Igor Aleksander and David Gamez

Department of Electrical Engineering, Imperial College, London SW7 2BT, UK
i.aleksander@imperial.ac.uk, dgamez@imperial.ac.uk

Abstract

In discussions about the physical support of conscious experience, a recent trend has been introduced (by Tononi and various colleagues) that measures the capacity of a network to discriminate among different states and integrate the information generated by this discrimination. This capacity to generate and integrate information can be used to understand the information processing in a network and Tononi has claimed that it is also linked to conscious experience. This paper describes experiments in which networks of weightless neurons were used to explore how different connection patterns and architectures affected the effective information generated by a network. The training of these networks using easily recognizable images made it easy to monitor their internal states, and this supports the interpretation of the system using the mental stance, which is described in a companion paper. By applying the same training to different architectures we were also able to study how the informational relationships depended on a combination of training and other dynamic effects.

Introduction

When a system enters a particular state, the amount of information associated with that state depends on the number of other states that are available to the system. For example, when a person looks at a landscape, they distinguish between that particular landscape and all of the other landscapes that they have seen, and they also distinguish the landscape from the enormous number of non-landscape perceptions that they are capable of. However, when a simple photodiode is exposed to light from a landscape it switches on in exactly the same way that it does when it is exposed to a light bulb or any other source of light. One difference between the person and the photodiode is that the person generates a lot of information when they rule out the large number of non-landscape states, whereas the photodiode generates very little information because it only ever rules out one state – the state of no light being present. A second difference is that a person experiences a landscape as a single integrated scene, whereas an array of photodiodes in a camera employ no such integration – leaving this to the observer of the image.

The measurement of the information that is generated and integrated by a system can help us to understand the relationships between its mental states (Gamez and Aleksander, 2009). For example, one partial state might be representing the presence of red in the environment and another partial state might be representing the presence of a cube. If these states are integrated together, then the system can be said to 'be aware of' the presence of a red cube, but if the states are not integrated together, then the system can only be said to be independently aware of redness and cubeness. The link between information integration and consciousness (Tononi, 2008) also makes it possible to generate predictions about the phenomenal states of a system by calculating its areas of maximum information integration (Gamez, 2008), and Tononi (2008) has made some proposals about how this approach could be used to understand the qualitative character of phenomenal states.

Over the last 15 years a number of algorithms have been put forward to measure the information generated by a system and the extent to which this information is integrated. One of the first of these measures was neural complexity (Tononi, Sporns and Edelman, 1994), which calculates the difference between the sum of the entropies of the individual components of a system considered independently and the entropy of the system considered as a whole. Tononi and Sporns (2003) put forward an algorithm that measures the mutual information between two halves of a subset of a system when one half is in a state of maximum entropy. This procedure is repeated on all possible bipartitions and subsets to find the most integrated parts of the system. More recently Balduzzi and Tononi (2008) have developed a way of calculating the generation and integration of information on the basis of the *a priori* and *a posteriori* repertoires of a system's current state. This approach is summarized in detail in the next section.

In this paper we focus on the first aspect of the information integration problem: the amount of

information that is generated when a system enters a particular state – what Balduzzi and Tononi (2008) call the *effective information*. This has some overlap with information integration, both because the effective information is the result of causal interactions within the system and because we consider the effective information that is generated over several time steps through interactions between different neurons. To study the generation of effective information we simulated a number of different networks of weightless neurons and trained them using images of famous faces. By monitoring the evolution of these networks towards a stable state it was possible to evaluate the effect of connection patterns and architecture on the effective information generated by the networks, and the discussion at the end of this paper makes some speculative suggestions about their information integration. A theme that permeates this paper is that there is a distinction between the information generated by learned states and that generated by the spurious effects due to the connection patterns of the network.

The first section of this paper covers Balduzzi and Tononi's (2008) algorithm for measuring effective information and explains how this can be extended to calculate the information integrated by a particular state. The next part sets out the experimental set up that was used, covering both the weightless neurons and the training with a set of images. The subsequent sections describe experiments that explored how different connection patterns and architectures affect the effective information generated by a network. The paper then concludes with a discussion and suggestions for future work.

Effective and Integrated Information

According to Balduzzi and Tononi (2008), when a system, X, enters a state, x_1, the information generated by x_1 is a function of the size of the system's repertoire of possible states and how much the uncertainty about the repertoire is reduced by entering state x_1. Balduzzi and Tononi suggest that the information generated by x_1 can be measured by comparing the *a priori* and *a posteriori* repertoires of the system. The *a priori* repertoire, $p^{max}(X_0)$, is the probability distribution of the states of the elements when they are in maximum entropy and each state is equally likely. Under these conditions a system with n binary elements will have 2^n possible states and the probability of each state is $1/2^n$. The *a posteriori* repertoire, $p(X_0 \rightarrow x_1)$, is the probability distribution of the states of the elements that could have led to the system entering x_1. The effective information, *ei*, of state x_1 is defined by Balduzzi and Tononi as the relative entropy of the *a posteriori* and *a priori* repertoires, as expressed in Equation 1:

$$ei(X_0 \rightarrow x_1) = H[p(X_0 \rightarrow x_1) \| p^{\max}(X_0)] \quad (1)$$

Since the maximum entropy is constant, Equation 1 can be rewritten as Equation 2:

$$ei(X_0 \rightarrow x_1) = H(p^{\max}(X_0)) - H(p(X_0 \rightarrow x_1)) \quad (2)$$

The effective information in equations 1 and 2 could be the result of interactions between all of the elements in the system, or it could be the sum of the effective information generated by independent groups of elements. To measure a system's *information integration* it is necessary to consider the relationship between the information generated by the system as a whole and the information generated by groups within the system. If the information generated by the system as a whole is the same as the sum of the information generated by its parts, then the system as a whole is not generating or integrating any information.

To measure this relationship, Balduzzi and Tononi put forward a measure of integrated information, which is the relative entropy between the *a posteriori* repertoire of the system as a whole and the combined *a posteriori* repertoires of the parts of the system, as expressed in Equation 3:

$$ei(X_0 \rightarrow x_1 / P) = H\left[p(X_0 \rightarrow x_1) \middle\| \prod_{M^k \in P} p(M_0^k \rightarrow \mu_1^k) \right], \quad (3)$$

where $ei(X_0 \rightarrow x_1/P)$ is the effective information of a particular partition, P, of the system into two or more parts, M_k is a part of the system, and μ_k is a state of M_k. To calculate $ei(X_0 \rightarrow x_1/P)$, each part is considered as a system in its own right and the *a posteriori* repertoires are calculated by treating inputs from the other parts as noise.

From the point of view of information integration, the most important partition of a system is the one that decomposes it into maximally independent parts. This partition is called the minimum information partition (*MIP*) of the network, and it can be understood to be a 'natural' way of dividing the system 'along its joints'. For example, the minimum information partition of the elements in Figure 1 is shown by the dotted line. With this partition $ei(X_0 \rightarrow x_1/P)$ is zero because the information integrated by the parts is the same as the information integrated by the system as a whole.

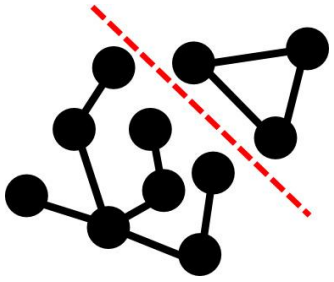

Figure 1: Minimum information partition of a network.

The minimum information partition is found by dividing up the network in all possible ways and identifying the partition in which the normalized value of $ei(X_0 \rightarrow x_1/P)$ reaches a minimum. When the partition is equal to the

entire system Equation 3 always yields zero because the *a posteriori* repertoire of the parts equals the *a posteriori* repertoire of the entire system. To fix this problem the effective information of the whole system is calculated from the *a priori* and *a posteriori* repertoires using Equation 1. Once the minimum information partition has been found, the information integration of a state of the network, $\Phi(x_1)$, can be calculated from the effective information of the minimum information partition, as shown in Equation 4:

$$\phi(x_1) = ei(X_0 \rightarrow x_1 / P^{MIP}) \quad (4)$$

While Balduzzi and Tononi (2008) calculate the effective information on the basis of states that are the direct cause of the current state, in this paper the effective information is calculated on the basis of an indirect causal relationship between an initial state of the network and its final stable state. This is a convenient way of dealing with recurrent networks that settle into a small set of stable states and a similar approach could be used to investigate the relationship between events at the brain's sensors and the appearance of these events in consciousness 500 ms later (Libet, 1982, 1993), after the brain has moved through a number of different states.

Networks of Weightless Neurons

Weightless Neurons

The experiments in this paper were carried out using weightless neurons, which compare stored patterns in their lookup tables with the *N* inputs that they receive from other neurons (see Figure 2). When the input pattern matches a known pattern (possibly with some degree of approximation), then 1 or 0 is output (depending on the training). When the input pattern does not match a known pattern (possibly with some degree of approximation), then the neuron outputs a random sequence of 1s and 0s.

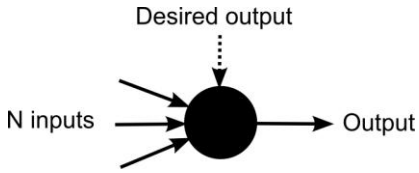

Figure 2: Weightless neuron.

A weightless neuron is trained by presenting it with an N-input pattern and setting the desired output to 1 or 0, which causes an output of 1 or 0 to become associated with the N-input pattern in the neuron's lookup table. If the stored output value is contradicted during a training sequence, then the stored lookup state is deleted for the contradicted input pattern. An approximate match between the input and the patterns in the lookup table is carried out by evaluating the Hamming distance between the input pattern and the stored patterns: if the Hamming distance is over a threshold set by the generalization parameter, then the input pattern is said to match. If multiple patterns match the input at a given level of approximation, then a random sequence of 1s and 0s is output.

Dynamic Neural Systems

In the first set of experiments a 98x98 neuron layer was used that did not have external inputs. The *n* inputs of each neuron were either connected to another neuron selected at random from the network (called a *distributed* connection) or connected to their near neighbors in a way that was specific to individual experiments (called a *local* connection). In the second set of experiments two 98x98 neuron layers were used, with one acting as the input to the second.

In the majority of the experiments the networks were trained on the set of famous faces shown in Figure 3 using what is known as an 'iconic' method. This created stable states for the chosen training patterns by forcing the input image to be both the output of the network and the 'desired' output of the neurons.[1] Once a network has been trained on these images, its stable states often match one of the images exactly, and when this occurs the state is referred to as an *experience state*. The advantage of working with two dimensional layers trained with recognizable images is that visual inspection can easily be used to monitor the evolution of the patterns over time.[2]

Figure 3: Images used to train the networks: Left to right, top to bottom, Einstein, Mandela, Obama, Zarkozy.

In all of these experiments the simulation and training of the networks was carried out using the NRM software.[3]

Effective Information and Connectivity

This series of experiments was designed to explore how the connectivity of the network affected the information that was generated by the stable states. The network was

[1] Aleksander (1996, Chapter 4) explains iconic training in detail.

[2] This fits in with the monitoring of the internal states of self-organizing systems that is discussed in Gamez and Aleksander (2009).

[3] See Barry Dunmall's site for more information about NRM: http://www.iis.ee.ic.ac.uk/eagle/barry_dunmall.htm.

trained so that it could move (with varying degrees of success) from a large number of initial starting states to a small number of stable finishing states. The training was carried out by setting the appropriate pixel of one of the four face patterns in Figure 3 as the desired output of each neuron in the network and also as the actual output of each neuron with the learning mechanism switched on. This created a potentially stable experienced state in the network and the system was tested by putting it into a non-experienced state and checking that it settled into an experienced state. The experienced states became attractors in the state dynamics of the system, with the dynamics depending on the connectivity of the neurons.

Experiment 1: Connectivity and the Retention of Experienced States

In the first experiment the 98x98 neuron layer was randomly interconnected with 32 connections per neuron. Figure 4, top row, shows that with this level of connectivity the network consistently settled into one of the four experienced states.

Figure 4: Experiment 1A, connectivity: 32 random connections. The upper row shows the individual stable states as they appeared in the 98x98 weightless neuron network. The lower row shows the time development (left to right, t=0, t=1, t=2, t=4) of the system starting in a random state and ending in one of the four experienced states.

The effective information that was generated when the network entered one of its four final states can be calculated as follows. We know that the network as a whole has 2^{9604} possible states. In the *a priori* repertoire each of these states has an equal probability of $1/2^{9604}$, making the entropy of the *a priori* repertoire 9604. If a quarter of these states, or 2^{9602}, lead to each final state, then the *a posteriori* repertoire will consist of 2^{9602} states with probability $1/2^{9602}$ and 2^{9604}-2^{9602} states with probability 0, giving an entropy of the *a posteriori* repertoire of 9602. By Equation 2, the effective information generated by each stable state is thus 9604–9602 = 2 bits of information. This is low considering the size of the network, but it makes sense if one considers that one out of 2^{9602} states lead to each stable state, and so little information is gained by entering one of the stable states.

This result can be generalized into a formula expressing the maximum effective information that is generated when a network enters a trained stable state. If the network has a repertoire of n_{tot} possible states, then the entropy of the *a priori* repertoire is $\log_2(n_{tot})$ and the entropy of the *a posteriori* repertoire of one of the stable states is $\log_2(1/(n_{tot}/t))$. Putting these expressions into Equation 2 gives Equation 5:

$$ei^{\max}(X_0 \to x_1) = \log_2(t), \quad (5)$$

which is the maximum possible effective information that can be generated when the network enters a trained stable state, assuming that n_{tot}/t of the states in the *a priori* repertoire lead to one of the t stable states. The following experiment shows how the connectivity of the network may prevent it from reaching this maximum.

In the next experiment the connectivity was reduced to six connections per neuron. As can be seen from the final stable states shown in the top row of Figure 5, the Obama and Zarkozy images continued to be stable final states, but the Mandela and Einstein states merged into a single stable end state.

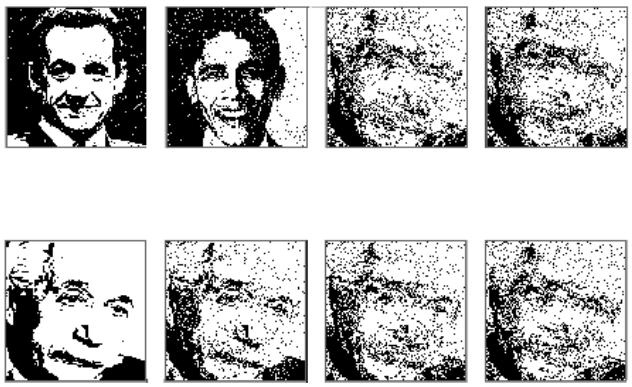

Figure 5: Experiment 1B, connectivity: 6 random connections. The upper row shows that two of the experience states were reasonably sustained, while the remaining two fused into one state. The lower row shows the deterioration over time into one of the confused experience states.

In the case of the fused Mandela/Einstein stable state, $2^{9604}/2$ states led to this state and the effective information generated by this state was 1 bit, which is less than the maximum of $\log_2(4) = 2$ bits. This shows that the reduced connectivity prevented the network from generating the maximum possible effective information that could be expected from the training.

Analytic comment. The above results may be linked to integration as follows. A particular neuron with n inputs can receive 2^n distinct binary patterns (n-tuples). If the same n-tuple is present for two experience states that

require opposite outputs for that neuron, it is said that this neuron is *contradicted* for that *n*-tuple. That is, the neuron transmits no information in this instance and cannot take part in the process of integration. An exact prediction of the number of contradictions depends on similarity relationships between the experience states. While this is beyond the scope of this paper, it is noted that the similarity between the Mandela and Einstein images is greater (due to large areas of white) than between others. This illustrates well that for these two patterns a sufficient number of neurons has been excluded from the integration process for the integration to fail and for the experience states to be lost.

Experiment 2: Effect of Connection Localization on Effective Information

In the previous experiments the neurons had no geometrical metric associated with their interconnection. The next set of experiments examined whether localized connections affected the effective information generated by the system. Again, the 98x98 network was used and each neuron was connected to 25 other neurons in a 5x5 square and trained using the image set on the centre bit of the 5x5 square to create a stable state. The system was tested by starting it in one of the experience states, and it was clear that these were integrated and stable. Noisy versions of the images were then tested and finally the initial state was set completely to noise to see if the experience states emerged.

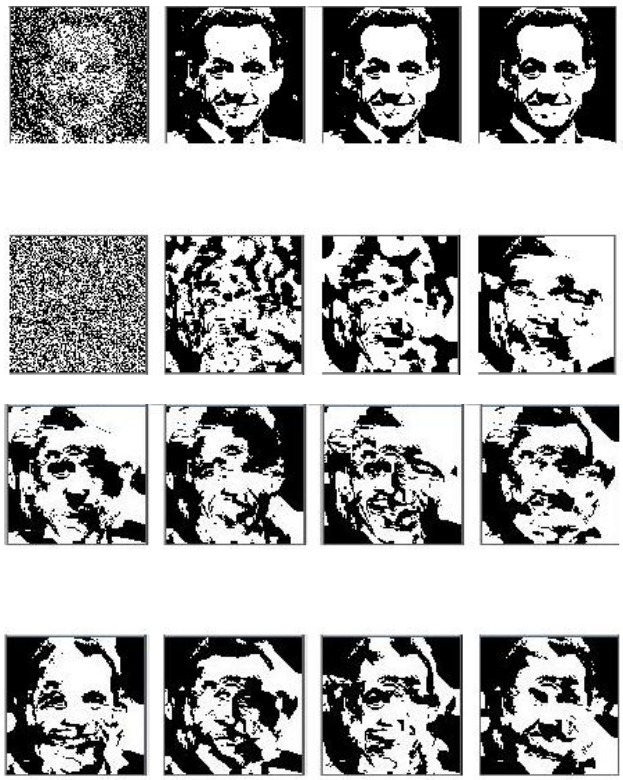

Figure 6: Experiment 2A, localized connectivity, full connection in a 5x5 area.

Top row: Sequence starting with an image state disturbed by 70% noise (70% of the bits of the image arbitrarily set). The last state is stable.
Second row: State sequence starting in a fully random state. The last state is stable.
Last two rows: Final stable states for differing initial random states. No duplicate final states were observed in 30 such tests.

The results in Figure 6, first row, show that localization did not prevent the system from moving into one of the experience states when the network was started with one of the images and 70% noise. However, when the starting state was random the system ended up in a combination of the experience states, which appeared as a nightmarish juxtapositions of images.[4]

The main change introduced by the localized connection pattern was that less initial states led to the experienced states. This had the effect of reducing the entropy of the *a posteriori* repertoire and the net result from Equation 2 is that more effective information was generated by the experience states. The stable mixed up states were also generating more than 2 bits of effective information, although it was not possible to calculate the exact amount with our current experimental setup. Although more effective information was generated by the network, only some states generated information about the experienced world: an issue that must be taken with some care when seeing integration as a measure of consciousness. The confused nature of these final states would also make them much less useful if the system needed to carry out a task in the world.

Analytic comment. In a system of overlapping interconnections trained on experience states (a 5x5 patch overlaps 15 neurons in the next one; a 19x19 patch overlaps 323 neurons in the next one) neighboring patches are likely to have the same response, which creates regions that are sensitive to the same experience state. As time progresses, regions of similarity emerge and grow into clearly defined boundaries.

[4] The authors do not know if such phenomena have been experienced by humans, but if they have been, they could be due to insufficient integration between localized areas in the brain.

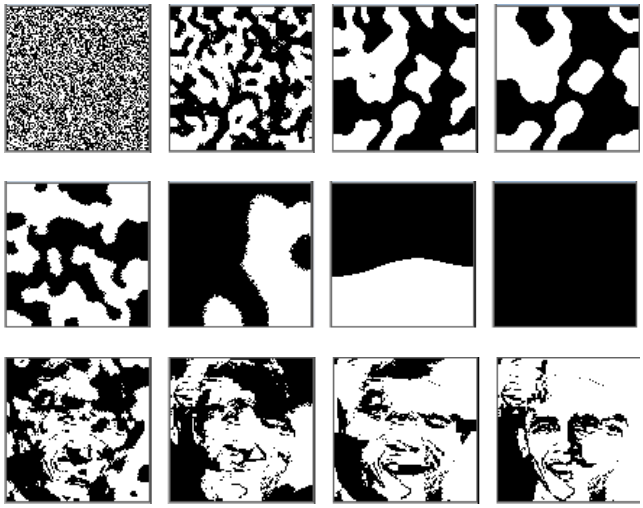

Figure 7: Experiment 2B: false integration mechanisms.
Top row: Patch size variation, left to right, the patch sizes were 5x5, 9x9, 13x13, 19x19.
Second Row: State progression for 25x25 localization patches, trained only on an all-black and an all white pattern, from noise (left) then, left to right, steps 1, 4 and final stable state reached after 22 steps.
Third Row: Final states for the all-black and all-white experiments for the same patch sizes as the top row.

To explore the effect of local connections further, experiments were carried out that varied the size of the local patch. The results in the top row of Figure 7 show that the larger the localized patch, the lower the fragmentation of the state. To reveal the mechanisms at work we also trained the network on all black and all white experiential states. In the second row of Figure 7 one can see how local integration between the neurons led to the development of boundaries in the final stable state. The third row of Figure 7 shows how a region of the 19x19 patch grew into one of the two experience states.

In the experiments shown in Figure 7 the networks were generating significant amounts of effective information because they could enter a large number of stable states and they were more sensitive to their starting points than the networks with non-localized connections. In our current experimental setup it was difficult to measure how much effective information was generated because we did not have an easy way of counting the number of stable states. However it is clear from inspection that locally integrated regions emerged as a function of both training and network structure.

Experiment 3: Effect of Discontinuities on Effective Information

This experiment investigated the effect of a discontinuity that cut direct connections between the top left quadrant and the top right quadrant of the randomly connected 98x98 neuron network.

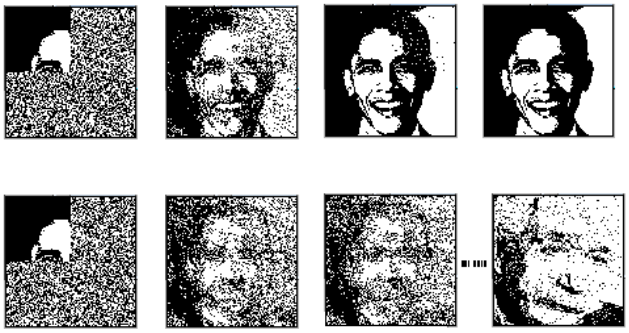

Figure 8: Experiment 3: Effect of a disconnection between the top two quadrants of the network.
Top row: The neuron connectivity was 30 connections drawn from allowed areas of the network. The figure shows the first four time steps, left to right.
Second row: The neuron connectivity was reduced to 8. The first three steps are shown left to right. The last state is for step 32 and did not change, except for noise, thereafter.

The results in Figure 8 show that the discontinuity only interfered with integration for low general connectivity, where integration was expected to fail even without a cut in the network. The importance of the experiment is seen in the first row, where a portion of the Obama image was used as a seed in an otherwise random image. The discontinuity delayed the integration in the top right quadrant which, nevertheless, was eventually achieved to complete the appropriate experience state. The failure in the second row was due to 'false integration', as no sign of the discontinuity was visible in the final state.

It is of some interest to note that this network does not confuse the Einstein and Mandela states, as was the case in the low-connectivity part of Experiment 3. This is due to a fortuitously more amenable set of random connections. In this experiment approximately the same effective information was generated as Experiment 1, although this took place over a longer period of time.

Internalizing the External: Phenomenology

This part of the paper describes experiments that investigated how a network can integrate incoming sensory information with its growing collection of experience states. We were particularly interested in the tradeoff between stable states that enable a network to complete noisy or missing information and stable states that are so strong that they dominate and obscure the sensory information coming from the world.

In these experiments the neural network consisted of a dynamic recurrently connected layer similar to the networks used in previous experiments, which was connected to a second 98x98 layer that acted as an input to the dynamic layer and functioned in a similar way to a retina - see Figure 9.

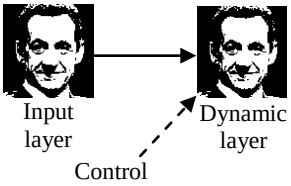

Figure 9: Experimental setup with input layer and dynamic layer

Figure 10 shows the connections of one weightless neuron in the dynamic layer of Figure 9.

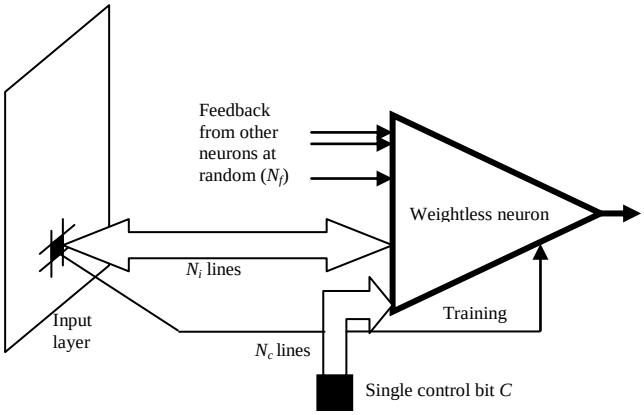

Figure 10: Connections to a neuron in the dynamic layer.

Each neuron in the dynamic layer receives N_f single-bit lines fed back from the outputs of f randomly selected neurons. Neurons also receive N_i parallel lines all carrying one bit of information from the corresponding input layer, as well as N_c lines from a single one-bit control signal, C, that is common to the entire network.

During the first part of the training the control signal was set to 0 and the network was presented with an all white input. The network was taught to output an all white response and this procedure was repeated for an all black input. This training enabled the network to copy the state of the input layer when the control signal was set to 0.

In the second stage of training the network was started with the control signal set to 0. The target image was then presented to the input layer with the dynamic layer in a random state. Next, the control signal was switched to 1 and the training input was activated twice to make the target a re-entrant state (once for the transition from the random state and once to create the re-entrant experience state). We found that it was necessary to repeat this several times with different random starting states to ease the switching from one experience state to another.[5]

[5] The results which follow represent a reasonable outcome after experimenting with several connection and training times to achieve a balance between creating stable states that did not respond to new input and stable states that were responsive to new input.

Experiment 4A: State Switching with an Experience State at the Input.

Once the network had been trained, an Obama image was presented at the input when the dynamic layer was in the Einstein state. The integrative activity of the network between its input and internal state led it to make a transition to the Obama state – see Figure 11.

Figure 11: State transitions from one experience state to another in Experiment 4A. (N_i=2, N_f=16, N_c=16, 20 training steps per image).
Top image: Input at the input layer.
Lower Row: Left to right, state when the top image was presented (t=0), then t=1, t=3 and t=6.

Experiment 4B: State Switching with a Non-experience State at the Input

In this experiment the network was presented with a partial experience state at its input and the dynamic layer was put into a random state. Figure 12 shows how the internal state developed from a random state into a complete Zarkozy image.

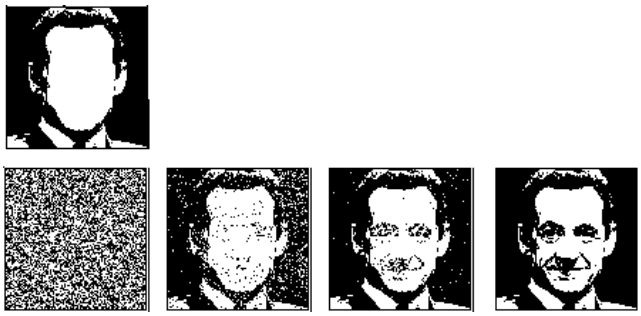

Figure 12: Reaching the experience state from an incomplete input.
Top image: Content of the input layer.
Lower row: Left to right, state of the dynamic layer starting from a random state (t=0) and then 'seeing' the input at t=1, t=3 and t=6.

In both Experiment 4A and Experiment 4B the final state of the network was a combination of the input state, which did not change, and the state of the dynamic layer, which moved into one of the four experience states. The total number of possible states of the network as a whole is

2^{19208}, and so the entropy of *a priori* repertoire is 19,208. In the *a posteriori* repertoire, 1/4 of the input states led to one of the four stable states when they were combined with any state of the dynamic layer, and so the number of states that led to each stable state is $2^{9602}*2^{9604} = 2^{19206}$. The probability of each of these states in the *a posteriori* repertoire is $1/2^{19206}$, and the entropy is 19206. By Equation 2, the amount of information generated when the network entered one of its stable states is 19208–19206 = 2 bits.

Discussion

These experiments demonstrate that it is possible to roughly estimate the effective information that is generated by a stable state of a network of weightless neurons. Instead of calculating the effective information on a state by state basis, we examined the effective information that was generated over time as an initial stimulus, such as noise or a partial face, was processed by the interactions between the neurons to produce a stable state. Although the number of possible states of the network was very high, the effective information generated by the stable states was low because the knowledge that the network was in one of the four experience states did little to reduce uncertainty about the starting state of the network.[6] Localized connectivity made the network more sensitive to its starting state and increased the effective information generated by the network, but this did not improve the network's knowledge about the world because the final stable states were largely mixed up and meaningless.

The human brain has a large number of connections from high level to low level stages of sensory processing (Hupé1 et al., 1998), which suggest that the current state of our brain has a strong effect on our processing of information from the world. Experiment 4B gave a very simple demonstration of this phenomenon when the partial face was filled in by the network's 'expectations' about the input. This experiment showed that the filling in is good at handling noisy data and reducing an uncertain world down to something that the system can respond to, but the cost of such simplification is that much less information is gained about the world.

To work out the information integration or Φ of the network it is necessary to calculate the effective information of every possible partition using Equation 4 to identify the minimum information partition. At the time of writing the software for these calculations is still under development and the factorial nature of this analysis heavily constrains the extent to which it can be carried out (Gamez, 2008). Given these constraints we can only make some speculative remarks about the likely information integration of the experimental networks in this paper.

In the randomly connected single layer network it seems likely that the random global connectivity would make whole network into a complex, although there are likely to be smaller higher Φ complexes within the network due to the random nature of the connectivity and the effects of training. The local connections in Experiment 2 would be likely to promote small high Φ complexes, which would probably follow the boundaries shown in Figure 7. In the second set of experiments, the input layer was not as highly integrated as the dynamic layer, and we would predict that the neurons in the input layer would provide sensory data to the dynamic layer without participating in the higher Φ complexes of the dynamic layer.

The work described in this paper is very preliminary and we are currently developing software that will run the Φ calculations on the networks presented in this paper. When we can say more about the information generated and integrated by the networks we will be able to build up a better understanding of the internal states of these experimental networks and make predictions about their consciousness using Tononi's (2008) theory.

Conclusions

This paper has described a number of experiments in which recurrent networks of weightless neurons were used to study the effective information that was generated by stable states. This effective information reached the theoretical maximum for the training set when strong distributed connections caused the network to settle into one of the training states; reducing the connectivity caused the network to generate less than the maximum possible effective information. When the connectivity was more localized the effective information increased, but many of the stable final states were confused and meaningless. The network's ability to reconstruct partial training patterns enabled it to respond effectively to noisy signals, but it also made the network relatively insensitive to its environment.

The illustrations in this paper suggest that effective information measures based on the reconstruction of experience states could be a useful way of evaluating a system's potential for consciousness. Whether this potential is actually exploited is likely to depend on whether the system contains a high Φ complex that is connected to the external world through a process of learning.

The long term aim of this work is to help us to understand cognitive systems as they are being trained and to gain a better understanding of a systems' internal states (Gamez and Aleksander, 2009). We are currently developing software that will enable us to calculate the information integration of a network and build up a more detailed picture of its informational relationships.

[6] If the calculations were carried out on a state by state basis, it is likely that the effective information would be considerably higher.

Acknowledgements

This work is supported by a grant from the *Association for Information Technology Trust.*

References

Aleksander, I. 1996. *Impossible Minds: My Neurons, My Consciousness*. ICP London.

Aleksander, I. 2005. *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines*. Exeter: Imprint Academic.

Aleksander, I., França, F., Lima, P., and Morton, H. 2009. A brief introduction to Weightless Neural Systems. *Proceedings of ESANN 2009*, Bruges.

Aleksander, I. and Morton, H. 2007. Phenomenology and digital neural architectures. *Neural Networks* 20(9): 932-7.

Balduzzi, D. and Tononi, G. 2008. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* 4(6).

Gamez, D. 2008. *The Development and Analysis of Conscious Machines*. Unpublished PhD thesis, University of Essex, UK. Available at: www.davidgamez.eu/mc-thesis.

Gamez, D. and Aleksander, I. 2009. Taking a Mental Stance Towards Artificial Systems. *Proc. AAAI Fall Symp. On Biologically Inspired Cognitive Architectures.*

Hupé1, J.M., James, A.C., Payne, B.R., Lomber, S.G., Girard, P. and Bullier, J. 1998. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* 394: 784-787.

Libet, B. 1982. Brain stimulation in the study of neuronal functions for conscious sensory experiences. *Human Neurobiology* 1: 235-42.

Libet, B. 1993. The neural time factor in conscious and unconscious events. *Ciba Found Symp* 174:123-37.

Tononi, G. 2004. An Information Integration Theory of Consciousness. *BMC Neuroscience* 5:42.

Tononi, G. 2008. Consciousness as Integrated Information: a Provisional Manifesto. *Biological Bulletin* 215: 216-242.

Tononi, G. and Sporns, O. 2003. Measuring information integration. *BMC Neuroscience* 4:31.

Tononi, G., Sporns, O. and Edelman, G.M. 1994. A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* 91: 5033-7.