Connectionist Models

James L. McClelland & Axel Cleeremans

In: T. Byrne, A. Cleeremans, & P. Wilken (Eds.),
Oxford Companion to Consciousness.
New York: Oxford University Press, 2009.

Connectionist models, also known as Parallel Distributed Processing (PDP) models, are a class of computational models often used to model aspects of human perception, cognition, and behaviour, the learning processes underlying such behaviour, and the storage and retrieval of information from memory. The approach embodies a particular perspective in cognitive science, one that is based on the idea that our understanding of behaviour and of mental states should be informed and constrained by our knowledge of the neural processes that underpin cognition. While neural network modelling has a history dating back to the 1950s, it was only at the beginning of the 1980s that the approach gained widespread recognition, with the publication of two books edited by D.E. Rumelhart & J.L. McClelland (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986), in which the basic principles of the approach were laid out, and its application to a number of psychological topics were developed. Connectionist models of cognitive processes have now been proposed in many different domains, ranging from different aspects of language processing to cognitive control, from perception to memory. Whereas the specific architecture of such models often differs substantially from one application to another, all models share a number of central assumptions that collectively characterize the "connectionist" approach in cognitive science. One of the central features of the approach is the emphasis it has placed on mechanisms of change. In contrast to traditional computational modelling methods in cognitive science, connectionism takes it that understanding the mechanisms involved in some cognitive process should be informed by the manner in which the system changed over time as it developed and learned. Understanding such mechanisms constitutes a significant part of current research in the domain (Elman et al., 1996; Mareschal, Johnson et al., 2007; Mareschal, Sirois, Westermann, & Johnson, 2007; Rogers & McClelland, 2004).

Connectionist models take inspiration from the manner in which information processing occurs in the brain. Processing involves the propagation of activation among simple units (artificial neurons) organized in networks, that is, linked to each other through weighted connections representing synapses or groups thereof. Each unit then transmits its activation level to other units in the network by means of its connections to those units. The activation function, that is, the function that describes how each unit computes its activation based on its inputs, may be a simple linear function, but is more typically non-linear (for instance, a sigmoid function).

Representation, processing, and learning in connectionist networks

Representation can take two very different forms in connectionist networks, neither of which corresponds to "classical" propositional representations. One form of representation is the pattern of activation over the units in the network. Such patterns of activation are generally thought to constitute the representation one has of something while perceiving,

processing or remembering it. Units in some connectionist networks are specifically designated in advance by the modeller to represent specific items such as identifiable visual features, letters, words, objects, etc. Networks that employ such units for all cognizable entities of interest are called localist networks – the representation of an item is in a sense localized to the single unit that stands for it. Most networks, however, rely on distributed representation – the idea that an item, such as a word, object, or memory, is instantiated by a pattern of activation over a large ensemble of units. In such systems, each representation depends on the activation of many units, and each unit is involved in representing many different objects. Distributed representations in early models were assigned by the modeller, using for instance a set of primitive features as the components of the representation. In later work using multilayer networks with hidden units, the learning methods allow the patterns to be determined from an initial random starting place through a learning process.

The other form of representation in connectionist networks consists of the values of the connection weights linking the processing units. The connection weights are generally viewed as the repository of prior experience that survives the patterns of activation produced during the experience itself. Such connection weights may be positive or negative real numbers. In some models they are set by hand by the modeller but in most cases they are set through a learning process, as discussed below.

Importantly for the relevance of connectionist models to the study of consciousness, neither the patterns of activation in a network nor the connection weights linking them are subject to direct inspection or manipulation by some other part of a connectionist system. Instead, activations in one part of a system simply directly influence the activation of connected units elsewhere. Thus, while it is possible to build connectionist networks that have weights that allow, for example, patterns of activation on two sets of units to be compared, with another set of units then reporting the degree of sameness or difference, this is not necessary for processing to occur and indeed most connectionist networks do not embody such mechanisms. Connection weights are generally assumed to be completely inaccessible to any form of inspection. They can and do, however, encode associative relationships between arbitrary patterns (such as the association between a name and a face) and structured or lawful mappings between patterns (such as the relationship between the present tense form of a word and its past tense, or the features shared by an ensemble of patterns in an ***implicit learning** experiment).

Another important feature of connectionist systems is the fact that the patterns of activation that are formed during processing are not subsequently stored in the system as memories. Instead, they are thought of as leaving a trace in the network through the adjustments they produce in the connection weights. These adjustments can then allow the pattern (or something like it) to be reconstructed at a later time, as a form of memory. Thus, long-term knowledge in connectionist networks is always encoded by the connection weights, whereas the temporary results of processing occur through activation patterns over units.

Processing in connectionist networks occurs through the propagation of activation signals among the processing units, via the weighted connections. The process is generally regarded as a continuous-time process subject to random variability. Thus, it is widely assumed that the state of activations of units evolves continuously over time, and is not completely

deterministic so that the same input can give rise to different real-time trajectories. While in reality much of our experience consists of a fairly continuous flow e.g. of visual or auditory experience, many connectionist networks settle over time after the presentation of an input into what is called an attractor state (see ***attractor networks**) – a stable pattern of activity that tends to remain in place until some sort of reset occurs so that the next input can be presented. Such attractor states may be seen, perhaps, as representing the sequence of states a perceiver might enter into while making fixations on different points in a static scene, or experiencing a set of discrete items presented one after another in a sequence. In this respect, Rumelhart et al. (1986), reflecting on the implications of a connectionist approach to cognition and consciousness, noted for instance that "consciousness consists of a sequence of interpretations—each represented by a stable state of the system" (p. 39). This process is often further simplified as a simple, one-pass, feed-forward process that allows the pattern of activation over a set of input units, together perhaps with an internal pattern left over from the last round of settling, to influence the next state of the network. Networks of this type – known as Simple Recurrent Networks – have many of the essential properties of PDP models, and so have been a subject of fairly intensive study, particularly in the domains of language processing (e.g., Elman, 1990) and implicit learning (e.g., Cleeremans & McClelland, 1991)

Learning in connectionist models is the process of connection weight adjustment. In contrast to traditional models in cognitive science, most connectionist models learn through experience, that is, through repeated exposure to stimuli from the environment. Two broad classes of learning mechanisms can be distinguished based on whether adjustments to the connection weights are dependent on an error signal or not. In the former case, learning is said to be supervised for it is driven by the difference between the current response of the network and a target response specified by the environment. Such supervised learning (e.g., back-propagation) instantiates the computational objective of mastering specific input-output mappings (i.e., achieving specific goals) in the context of performing specific tasks. By contrast, unsupervised learning (e.g., Hebbian learning) instantiates the different computational objective of capturing the correlational structure of the stimulus environment, so enabling the cognitive system to develop useful, informative models of the world. Unsupervised learning procedures do not depend on the availability of a "teacher" signal, but instead determine adjustments to the connection weights based on the simultaneous activation of connected units, so instantiating Hebb's (1949) notion that "neurons that fire together wire together" — the same principle that is also observed in the neural process of long-term potentiation.

In simple models, the environment specifies the states of all units in the system, in which connection weights can be seen as establishing associative links between these units. In localist networks, individual connections mediate meaningful associative relationships. In distributed models, however, the situation is more complex. In such systems, if one wishes to associate, let us say, the sight of a rose with the smell of a rose, and if the sight and smell are each represented as a pattern of activation over a set of units, then the connection weight changes needed to store the association may in many cases impact all of the connections. It remains surprising to many people that many different associations can be stored in the same set of connection weights, especially if, as is often done, one or more layers of intermediate

units are interposed between input and output. The consequence of this is to then introduce intermediate representations that can re-represent input patterns in new ways.

Connection adjustment schemes in connectionist networks are used both for processes usually thought of as 'learning' and also for processes usually thought of as 'memory'. In the former case, connectionist models generally rely on very gradual connection adjustment procedures to give rise to a set of connections that implement a learned skill such as reading (mapping patterns representing the sounds of words to other patterns representing their sound and meaning). One such learned skill is the ability to anticipate the successive elements of a sequence from preceding elements, as a result of prior experience with many such sequences.

In the case of memory for a particular episode or event, the event is by definition experienced only once. To store it effectively in memory it is generally assumed that relatively large changes to connection weights must be made at or very near the time of the actual experience. In many models, the elements of the experience are thought of as forming a widely distributed pattern of activation (perhaps spanning many brain areas) representing all different aspects of the experience. An architecture that may involve many layers of intermediate units with very easily modifiable connections in its deepest layers then essentially allows all elements of the pattern to become inter-associated with all other elements. As a result, later presentation of a unique subset of the pattern can then allow the rest of the pattern to be reinstated over the units of the network.

Connectionism and consciousness

A starting place for consideration of the relevance of connectionist models and our concept of consciousness lies in the distinction, made above, between patterns of activation and the knowledge stored in connections. It is likely that the patterns of activation over some brain area are associated with states of conscious experience. Thus, one may experience a friendly beagle through the pattern of white and brown blotches of its fur, the yipping sounds it makes, the excited frisking about that it does when it meets you, and this may be crucially dependent upon active patterns of activation in a relevant ensemble of brain areas. One may imagine or remember an experience with the beagle by activating (pale, incomplete, and likely distorted) versions of these patterns over the same ensemble of brain areas when the beagle is not physically present. Within the connectionist framework it is often imagined that not all aspects of these patterns are likely to be consciously experienced; those that are especially emphasized by control processes and persist in a relatively stable form as attractor states may, however, be more likely to be consciously experienced, both during the initial event itself, and during a subsequent episode of imagination or memory. Stability of representation has been proposed as a ***computational correlate of consciousness** by Mathis and Mozer (1995, see also ***attractor networks**) and also by philosophers O'Brien & Opie (O'Brien & Opie, 1999) in their "Vehicle theory of phenomenal experience".

An idea related to the above is the proposal by Kirsh (1991, see also this volume) and Koch (2004) that the distinction between implicit and explicit processing is best captured by the extent to which additional computations are required to retrieve content. Thus, an activation pattern that directly encodes some state of affairs would, in this light, constitute a more explicit form of representation than a pattern of connection weights, for instance, because

more computations are required in the latter, but not in the former case, to retrieve the represented content.

Thus, the fact that connectionist models rely on knowledge stored in connections is important for the study of consciousness since it makes it clear how processing can be guided by learned knowledge without that knowledge being accessible to inspection. A central feature of explicit representations is that one is, at least potentially, conscious of having them. However, the knowledge acquired by a trained connectionist network is stored in the form of connection weights between processing units. The information contained in the pattern of connectivity that characterizes a trained network can not be accessed directly by the network itself. Instead, this knowledge can only manifest itself by the influence it exerts on the activation level of the units of the network. In this sense, such weights are not representational: They do not constitute objects of representation in and of themselves. Indeed, Clark and Karmiloff-Smith (1993) have pointed out that in connectionist networks, "knowledge of rules is always emergent. [These models] do not depend on symbolic expressions that stand for the elements of a rule. Instead, they exploit a multitude of subsymbolic representations whose complex interaction produces behaviour which, in central cases, fits the rule" (p. 504). Clark and Karmiloff-Smith continue by noting that such networks have no "… self-generated means of analyzing their own activity so as to form symbolic representations of their own processing. Their knowledge of rules always *remains* implicit unless an external theorist intervenes" (p. 504).

Knowledge is thus always implicit in what Clark and Karmiloff-Smith (1993) dubbed "first-order connectionist networks". In contrast, knowledge in classical, symbolic systems always seems to be at least potentially explicit, to the extent that it is stored in a format (symbolic propositions) that makes it impossible for it to influence behaviour (i.e., to have causal powers) without being accessed or manipulated by an agent (i.e., the processor). In other words, information processing in classical systems always appears to entail access to stored representations in a way that is strikingly different from what happens in connectionist networks.

Finally, an important aspect of connectionist modelling is the use of some parts of a complex connectionist network to control the states of activation in other parts of the network. In this respect, a useful distinction was recently offered by O'Reilly & Munakata (2000) — the distinction between weight-based and activation-based processing. According to O'Reilly & Munakata, "Activation-based processing is based on the activation, maintenance, and updating of active representations to influence processing, whereas weight-based processing is based on the adaptation of weight values to alter input/output mappings" (p. 380). The main advantage of activation-based processing is that it is faster and more flexible than weight-based processing. Speed and flexibility are both salient characteristics of high-level cognition. O'Reilly & Munakata further speculate that activation-based processing is one of the central characteristics of the frontal cortex, and suggest that this region of the brain has evolved specifically to serve a number of important functions related to controlled processing, such as working memory, inhibition, executive control, and monitoring or evaluation of ongoing behaviour. To serve these functions, processing in the frontal cortex is characterized by mechanisms of active maintenance through which representations can remain strongly activated for long periods of time so as it make it possible for these

representations to bias processing elsewhere in the brain. This attentional modulation of activation may have important implications for what aspects of a visual scene become available for overt responding or storage in memory, and indeed, Dehaene et al. (2006) have recently proposed to distinguish between unconscious, preconscious, and conscious processing based precisely on interactions between top-down and bottom-up processing in the brain. Note that such interactions presuppose the existence of recurrent connections—another proposed correlate of consciousness (e.g., Lamme & Roelfsema, 2000). Likewise, Maia and Cleeremans (Maia & Cleeremans, 2005) have proposed that many connectionist networks can be thought as implementing a process of global constraint satisfaction whereby biased competition between neural coalitions result in the network settling onto the most likely interpretation of the current input. Importantly, this suggests a strong link between attention, working memory, cognitive control, and availability to conscious experience, for the mechanisms underlying each of these different aspects of information processing in the brain can be thought of as depending on the operation of the same computational principles.

References

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language, 8*, 487-519.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology : General, 120*, 235-253.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204-211.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Hebb, D. O. (1949). *The organization of behavior*. New York, NY: Wiley.

Kirsh, D. (1991). When is information explicitly represented? In P. P. Hanson (Ed.), *Information, Language, and Cognition*. New York, NY: Oxford University Press.

Koch, C. (2004). *The quest for consciousness. A neurobiological approach*. Englewood, CO: Roberts & Company Publishers.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23*(11), 571-579.

Maia, T. V., & Cleeremans, A. (2005). Consciousness: Converging insights from connectionist modeling and neuroscience. *Trends in Cognitive Sciences, 9*(8), 397-404.

Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism: How the brain constructs cogntion* (Vol. 1). Oxford: Oxford University Press.

Mareschal, D., Sirois, S., Westermann, G., & Johnson, M. H. (2007). *Neuroconstructivism: Perspectives and prospects*. Oxford: Oxford University Press.

Mathis, W. D., & Mozer, M. C. (1995). On the computational utility of consciousness. In G. Tesauro & D. S. Touretzky (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 10-18). Cambridge: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences, 22*, 175-196.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

Rogers, T.T., & McClelland, J.L. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.