Associative learning should go deep

Esther Mondragón,[1,2,*] Eduardo Alonso,[1,2,*] and Niklas Kokkola[1,2,*]

[1]City, University of London, London EC1V 0HB, UK

[2]Computational and Animal Learning Research Centre, St Albans AL1 1RQ, UK

*Correspondence: e.mondragon@cal-r.org (E. Mondragón), E.Alonso@city.ac.uk (E. Alonso), Niklas.Kokkola.1@city.ac.uk (N. Kokkola).

Keywords: Associative learning; deep neural networks

Abstract: Conditioning, how animals learn to associate two or more events, is one of the most influential paradigms in learning theory. It is nevertheless unclear how current models of associative learning can accommodate complex phenomena without ad hoc representational assumptions. We propose to embrace deep neural networks to negotiate this problem.

Associative learning describes how two or more events (be they stimuli or responses) become associated (Box 1). This deceptively simple idea is one of the fundamental pillars in the study of learning and cognition. It has been proven to operate at both behavioural and neural levels, with a wide range of procedures and organisms, and to underlie higher-order cognitive processes (rule learning, concept formation). The rules of association formation may be simple but the world upon which they operate is not necessarily so. We argue that whereas models of associative learning often assume an arbitrary connectionist architecture, using deep networks to learn stimulus representations would allow for biologically plausible, hierarchical representations, better model comparison, and ultimately more accurate predictive models of learning. Although there is an on-going debate on the explanatory power of associative learning theory (see e.g., [1]), recent studies on the neural bases of trial and error learning [2], and the role of associative learning in evolutionary biology [3] and social interaction [4] seem to bolster the status of associative learning as one of the cardinal paradigms in behavioural neurosciences. The crux of the controversy nonetheless does not question experimental evidence, of which plenty exists, but whether such evidence is supported by current models within the terms of reference of traditional associative learning theory.

The last decade has seen a surge of increasingly sophisticated computational models of association formation, stemming from both neuroscience and artificial intelligence (see e.g., [5,6]). For instance, reinforcement learning algorithms have been remarkably successful in modelling the role of dopamine in reward learning [7] and are at the heart of cutting-edge studies in model-free and model-based associative learning [8]. Typically, such models are embedded in neural networks that correct a prediction error iteratively. Indeed, neural network architectures seem to be a logical way of representing connections between events, and the update rule they implement intuitively corresponds with the way predictions are adjusted as a result of learning. These update rules are also justified with respect to probability theory, i.e. Bayes' rule. Notwithstanding their merits, there are still critical phenomena whose interpretation poses formidable challenges for such models –and this has been taken as evidence of the limited scope of associative learning theory itself: Importantly, but not exclusively, evidence of learning between motivationally neutral stimuli questions whether reward is an essential component of learning (the role of reward); learning about absent stimuli may suggest the involvement of stored information (the role of memory); solving complex stimulus and temporal structural discriminations seems to require postulating non-linear relationships in stimulus pattern integration (the nature, configural or elemental, of the stimulus representation); and the notion of goal-directed behaviour raises incertitude on what the elements of an association might be (the content of learning). These are paradigmatic examples of topics that existing models of associative learning fail to explain in a systematic, consistent corpus.

It is our claim that this inadequacy can mainly be ascribed to a representational problem deriving from such models being instantiated in connectionist networks, which even with numerous hidden layers rely on hand-crafted inputs and suffer in terms of robustness and generalization. Advances in deep neural networks, aka Deep Learning, may provide us with powerful tools for modelling how representations of events are formed, connected and learned about. The underlying idea is to exploit large (deep) neural networks consisting of multiple levels of abstractions [9], facilitated amid breakthrough progress in Big Data, GPU computational power, and the development of “smart” training heuristics and architectures. Various techniques have been formulated to allow the formation of long-range dependencies along either the depth of a network (feed-forward nets) or temporally (recurrent nets). An example in the former case is the Rectified Linear Unit (ReLU) activation function, which has been contextualized in the stability analysis of synaptic connections. ReLU preserves error gradients due to its binary derivative and when combined with the dropout technique wherein units are randomly removed from the network during training, produces distributed, robust representations. This ensures similar network inputs activate similar high-level abstractions and removes detrimental feature co-dependencies, thereby improving generalization (and discrimination). It is the unique synthesis of these techniques that makes Deep Learning suitable. Although it is arguable whether deep neural networks learn or act as human beings [10], they have been extraordinarily efficient in recognising complex images and audio signals [11], and in solving intricate control tasks [12].

Our contention is based on the evidence that many learning phenomena do involve the formation of complex associations both in the interaction of structured sequences of paired events and, critically, in the formation of the stimulus representation per se; which deep learning naturally accommodates. In particular, we hypothesize that Convolutional Neural Networks (CNNs) show the necessary algorithmic and computational characteristics, namely, sparse connectivity and shift invariance, whilst keeping the error correction of many associative learning models, to account for phenomena that have thus far escaped a cohesive associative learning analysis. Crucially, in contrast to standard multi-layer networks, CNNs do not use ad hoc features, rather they define hierarchies of layers which automatically learn representations at different levels of abstraction –such representations emerge from the aggregation of lower level features through convolutional and pooling layers.

Specifically, the capability to distinguish between elements which are common or unique to different stimuli is essential in solving non-linear discriminations and determinant in the formation of within-compound associations and in mediated phenomena. Current models of associative learning do not establish a mechanism to extract, bond, and compute common and unique features, but rather conceptualise them ex nihilo. CNNs, which hierarchically filter information using different receptive fields (kernels) might offer a solution. In these systems, similar inputs result in similar activation patterns within the network, offering a plausible substratum for producing commonalities in representation. For instance, to solve a non-linear discrimination as the one described in Box 2, both elemental (a) and configural (b) theories rely on hand-crafted internal units. Contrarily, a deep neural network (c) could produce the effect without hypothesizing arbitrary constructs: The network would be trained on raw sensory input by application of kernels along its depth, producing abstractions ranging from low-level to compressed high-level representations. The ReLU activation function would preserve the error gradient when backpropagating through the network and, in combination with the dropout technique, would promote sparse, decorrelated, and noise invariant representations; ensuring the same abstractions would be active in trials of the same type. The correlation of activation of common and unique features would lead to the associative formation of respective unitized nodes, which would link to the outcome. Through backpropagation, the network would learn to associate irrelevant features, negatively correlating them to the outcome – thus solving the discrimination. This procedure would naturally extend to complex discriminations, automatically extracting patterns pertinent to the task.

Summarizing, current models of associative learning rely on bespoke stimulus representations and on the addition of multiple layers to connectionist networks. We contend that this approach may have exhausted its explanatory scope: (a) representations need to be generated by the learners and (b) multi-layer networks must be accompanied with computational techniques that make them efficient. Deep Learning does precisely that. With this paper, we would like to spur the interest in this new technology in the associative learning community.

References

1. Abrahamse, E. *et al.* (2016) Grounding cognitive control in associative learning. *Psychol. Bull.* 142, 693-728
2. Eshel, N. (2016) Trial and error: Optogenetic techniques offer insight into the dopamine circuit underlying learning. *Science* 354 (6316), 1108-1109
3. Enquist, M. *et al.* (2016) The power of associative learning and the ontogeny of optimal behaviour. *R. Soc. Open Sci.* 160734 DOI: 10.1098/rsos.160734
4. Heyes, C. (2016) Homo imitans? Seven reasons why imitation couldn't possibly be associative. *Phil. Trans. R. Soc. B* 371, 20150069; DOI: 10.1098/rstb.2015.0069.
5. Schmajuk, N. and Alonso, E. (2012) Computational Models of Classical Conditioning. *Learn. Behav.* 40, 231-240

6. Mondragón, E. *et al.* (2014). SSCC Temporal Difference: A Serial and Simultaneous Configural Compound-Stimuli representation for TD learning. *PLoS ONE* 9 (7): e102469.
7. Nasser H. *et al.* (2017) The dopamine prediction error: contributions to associative models of reward learning. *Front. Psychol.* 8 DOI:10.3389/fpsyg.2017.00244
8. Dayan, P. and Berridge, K.C. (2014) Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cogn., Affect. Behav. Neurosci.* 14, 473-492
9. Goodfellow, Y. *et al.* (2016) *Deep Learning*. Cambridge, MA: The MIT Press
10. Lake, B.M. *et al.* (2016) Building Machines That Learn and Think Like People. *Behav. Brain Sci.* 24, 1-101
11. Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR* (arXiv:1409.1556).
12. Silver D. *et al.* (2016) Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529, 484-489

Box 1: A classical conditioning example. When a stimulus is perceived a central representational node becomes active. Pairings of two stimuli engender concurrent activation of their internal nodes. In a typical procedure, a stimulus A is paired with an outcome (O, aka unconditioned stimulus or reinforcer), a stimulus able to elicit an unconditioned response (UR). A, on the other hand, is said to be neutral to that response. With pairings, a link is progressively formed between the stimulus' nodes, as a result of which A becomes a conditioned stimulus for the outcome. Thereupon, presentations of A alone will activate O's central representation, eliciting a conditioned response (CR) (top panel). The strength of the association (w) between A and O increases with the number of pairing trials as the error, that is the difference between the value of the prediction of O by A and the actual value of the occurrence of O, is reduced (bottom panel).

(PLEASE INSERT FIGURE 1 HERE)

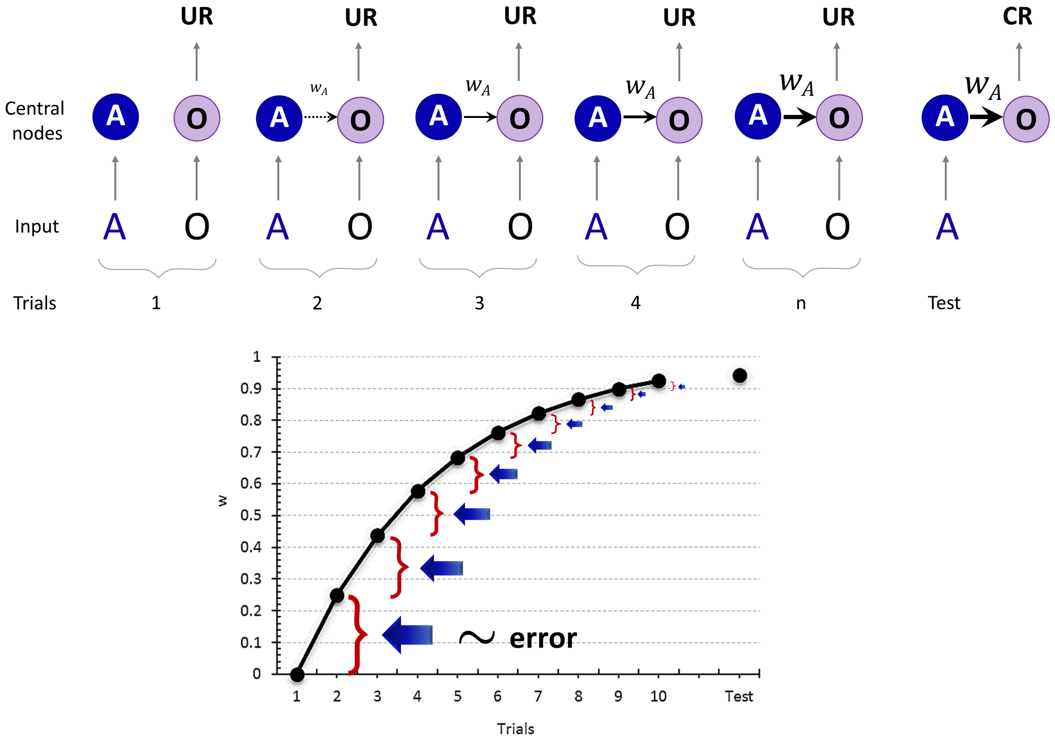
Box 2: Elemental, Configural and Deep Learning network architectures. In a negative patterning discrimination, single A and B presentations are followed by an outcome (A → O and B → O), and combined presentations are not (AB → no O). To explain discriminative performance, elemental learning models (a) posit that the combination of stimuli conveys an extra feature X (in the form of explicitly added cues (Z), subtracted cues (Y), or a mixture of both) distinctive from those in A and B alone. Thus, during learning, A and B will become linked to the outcome (excitatory link, black) whereas X will develop an inhibitory link (red) to O which will prevent the response from occurring. Configural models (b), on the other hand, simply consider that, partially activated by A and B, the compound AB is distinctively represented in a hidden layer together with the stimuli themselves. During learning, direct excitatory and inhibitory associations will be formed from each node in the hidden layer to the outcome. In contrast, in one schematic Deep Learning solution of the discrimination combining CNN and associative learning rules (c), receptive fields would produce hierarchies of abstractions from an input of raw features. Common elements activation would be highly correlated, fostering associations between them and creating unitized nodes of different common categories. In parallel, unique features would be extracted to nodes corresponding to the input patterns, which would link forward to the outcome and backward to the input. Next, nodes activated by the same input would associate (e.g., presentations of A would result in the association between the common elements nodes and the unique A features nodes) and a new abstraction (A in such case) formed in subsequent layers. The proximity of the newly formed A, B and AB abstractions to the outcome would link them preferentially to it. At this point the network cannot solve the discrimination because A and B nodes separately predict the outcome but AB – which combines the elements from both – disconfirms it, and thus the error is high. Backpropagation of the outcome error would trigger the formation of associations between the unique elements nodes and promote the emergence of a node of unique A and B features that would be extracted in a new layer and linked to the outcome. Through backpropagation of the outcome error, features in fully connected layers within the network could learn to inhibit other features. Thus, the common elements, better outcome predictors, would strengthen their link to the outcome promoting inhibitory links between the unique elements and the outcome. Note that this a schematic description of a network that would in practice comprise a large number of layers and for which ReLU and dropout techniques are needed, and that the elements in the unitized nodes are not conceptualized features but explicit and automatically extracted by the filters.

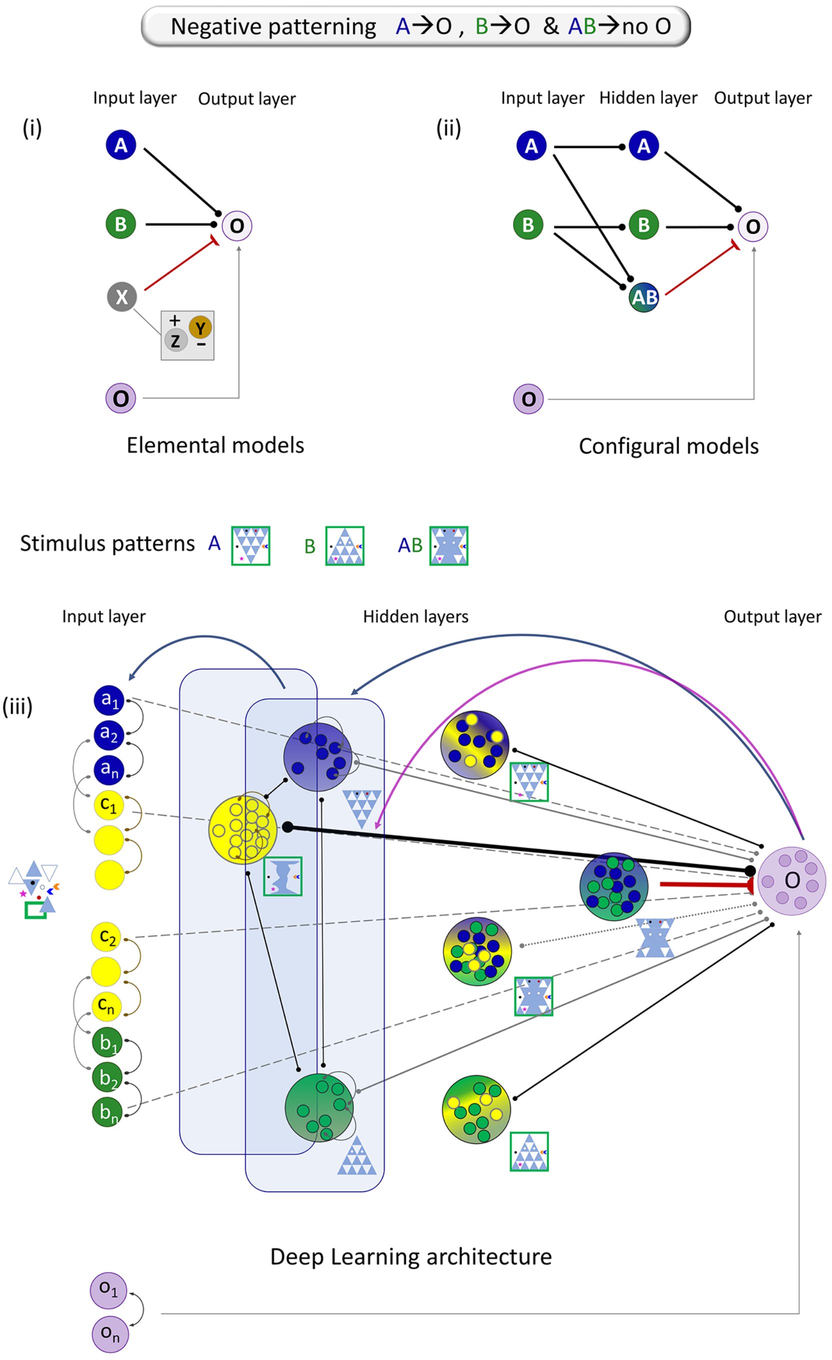
(PLEASE INSERT FIGURE 2 HERE)

FIGURE CAPTIONS

Figure 1: A classical conditioning example, where an organism is exposed to the repeated occurrence (trials) of a pair of stimuli, A and O.

Figure 2: Learning architectures in elemental, configural and Deep Learning models of a negative patterning discrimination.




Negative patterning A→O , B→O & AB→no O
(i)
Input layer
Output layer
A
B
X
O
+
Y
Z
–
O
Elemental models
(ii)
Input layer
Hidden layer
Output layer
A
B
A
B
AB
O
O
Configural models
Stimulus patterns A
B
AB
(iii)
Input layer
Hidden layers
Output layer
a1
a2
an
c1
c2
cn
b1
b2
bn
O
Deep Learning architecture
o1
on