Deep Learning and Embodiment

Pietro Perconti[1] and Alessio Plebe[1]

Department of Cognitive Science,
University of Messina, Italy
{perconti,aplebe}@unime.it

Abstract. Embodiment has become mainstream cognitive science, and has brought several important theoretical and empirical advances. Some embodied cognitive scientists have argued that cognition based on action and bodily states is incommensurate with the standard representationalist and computational approach. Often the case of visual perception is used as the best example for this thesis. In the recent years, the family of algorithms collected under the name "deep learning" has revolutionized artificial intelligence, enabling machines to reach human–like performances in many complex cognitive tasks, especially in vision. Such results are achieved by learning on collections of static images, therefore neglecting actions, dependency on time, and any form of interaction with the environment. Deep learning models were developed with engineering goals in mind, and advancing cognitive science is not in the agenda of this research community. Still, the achievements of deep learning in the case of vision seem to challenge widespread assumptions about visual perception in embodied and enactive cognition.

1 Introduction

Since more than two decades embodiment has taken center stage in cognitive science. Broadly speaking, embodied cognition emphasizes the role played by the body for cognition, in a variety of possible ways. One way could be that the mechanisms for concept manipulation and reasoning cannot be detached from the mechanisms by which our body acts and perceives, as argued by Lakoff and Johnson [44]. Since the body is the locus of actions, embodiment naturally implies *enacted* cognition, as done by Noë [56, 53]. Indeed, the body interacts in its environment, therefore embodiment reconciles cognition with Gibson's ecological psychology [20, 21, 27]. A further important account of embodiment lies in artificial intelligence, as the need of implementing a mechanical body and active controls in order to achieve simulated cognitive capacity [2, 57]. Embodiment has certainly contributed to fundamental advances in cognitive science; but, a controversial aspect has been the rejection of the computational and representational theory of mind [19, 5, 31]. More details about the various faces of embodiment, and its challenge to the computational theory of mind, will be given in §2.

We argue that today a new and unexpected actor may step into this debate: deep learning. This term refers to a family of artificial neural network techniques

that are collecting a number of exciting results. In 2012, the group at the University of Toronto lead by Geoffrey Hinton, the inventor of deep learning, won the most challenging large–scale image classification competition. Soon Hinton was invited by Google, which adopted deep learning for its image search engine. In 2016, the company DeepMind, founded by Demis Hassabis and soon acquired by Google, defeated the world champion of *Go*, the Chinese chessboard game much more complex than chess [73]. The leading Internet companies were among the first to deploy deep learning on a large scale [25], and are also the largest investors in research, way beyond their internal needs.

Deep neural models are the only artificial perception systems that routinely outperform humans in object recognition tasks [80]. This performance is disconcerting for the perspective of embodied (enacted, embedded) cognition, because it is achieved with computations that disregard any action, any dynamics, any interaction with the environment. It is certainly necessary to be cautious in drawing conclusions, because deep neural models are not intended as tools for studying cognition, and are not biologically plausible models (see §4). However, we deem these results may, at least, suggest that the embodied components in vision are important, but not indispensable, for the recognition task.

2 Embodiment and Computationalism

The story usually goes as follows. Cognitive science comes into two steps, the first being a mere computational account for modeling human intelligence and the second a more ecological and biological approach aimed at understanding how the human brain actually works in its environment. The core business in early computational psychology was to design cognitive architectures, that is, models of how a number of information–processing devices interact with each other to perform a given cognitive task. And this is the key goal for the cognitive science even now. But, while early computational psychology does not care how all this processing is actually realized on its matter, the more recent phase of cognitive science would be embodied, meaning that it would be grounded on how the human brain actually works and on how the human body on the whole encodes the information that comes from the world. Milestones in this change in perspective includes the amazing achievements in the field of neuroscience over the last decades, the ecological vision approach by James Gibson [21], and Parallel Distributed Processing (PDP) [68]. All these research programs share the idea that understanding intelligence is not matter of imaging an abstract and disembodied subject in front of a static world, but matter of taking into consideration the whole scene, in which perceptual scenarios endowed with many affordances suggest human bodies how to achieve their goals.

The first victim in this way of reasoning was, of course, the notion of *mental representation*. It suggests, in fact, the idea of a sharp dualism between the subject and the world in which mental representations are a sort of mediation device consisting essentially in information processing. The supporters of the 4E cognition (embodied, embedded, enactive, and extended), on the contrary,

prefer a more dynamic view in which individuals interact with other things in the environment in a more direct way. This is, however, highly misleading. The notion of mental representation itself, in fact, does not entail any denial of bodily possibilities and constraints, being simply an abstract rule able to link kinds of environmental events and bodily encoding in a systematic way. How an abstract rule can have a causal role on the world is the most celebrated achievement in the representational computational theory of mind. The point is, however, that computational psychology is neutral about ecology of cognition. It is simply, so to say, the way to solve the mind-body problem, not the way to discriminate ecological and dynamic accounts in understanding how knowledge works. Mental representations could be dynamic and ecological constructs, as many 4E cognition supporters desire, if we are able to sketch out its functioning in the right way, as nowadays it is possible to do on the basis of neuroscience and neural networks [58, 32, 59, 60]. It is matter of conceiving the right idea of what mental representations are, not to rule out them from the scene. And, perhaps, with these considerations in our mind it becomes possible to reconcile embodied cognition with classical computational psychology [52]. The real problem in embodiment and enactivism is the expectation that only by adopting a 4E cognition account we can deal with cognition in the right way. In other words, if a cognitive architecture is not modeled in a human–like manner, we should not be able to understand the cognitive process we are interested in. This is, however, precisely the point deep learning models put into discussion. While, in fact, their basis is grounded in PDP and, in the end, in the neural networks account, deep learning models do not follow any further biologically inspired constraint and achieve their results in a mere mathematical way and even without any particular cognitive concern in mind. This is like a scandal for embodied cognition: what about biological constraints, if we can get human–like cognitive performances in another way?

A similar scandal occurred some years ago in the field of cognitive ethology. Since Gordon Gallup [17] devised the *mark test*, or mirror test, many other ethologists were engaged in testing the capacity to recognize own image reflected in a mirror. It results that, besides humans, the other species which show this ability are those closer to us. The best results have been obtained with great apes. One can think that there is a phylogenetic reason for this. You need a human–like brain to have the self-recognition capacity. In this framework, a little scandal arises when it emerges that also magpies, a songbird from the crow family, is endowed with self-recognition ability [61]. Magpies, in fact, have a brain very different from that of primates and other mammals, insofar it does not include any neocortex, that is, the large structure on the outer surface of the brain where self-recognition, like most higher-order processing, takes place. As in the case of the deep learning performances in objects recognition, we have to be cautious. Recent findings suggest a homology between certain neuronal cell types of the avian dorsal telencephalon and the cell types of mammalian neocortical circuits [33, 3]. But, in any case, we have to moderate the neocortex enthusiasm, as we have to moderate the embodiment fanaticism.

3 Deep Neural Models

Deep neural models, also named deep learning models, are responsible for the current resurgence of Artificial Intelligence after several decades of slow and unsatisfactory advances [71]. Deep learning, in all its variations, has achieved unprecedented success in a vast range of applications, often approaching human performance [49]. Deep learning evolved from artificial neural networks, introduced in the '80s with the PDP (*Parallel Distributed Processing*) project [68]. The basic structure of the "parallel distributed" is made of simple units organized into distinct layers, with unidirectional connections between each layer and the next one. This structure, known as *feedforward network*, is preserved in most deep learning models. PDP reestablished a strong empiricist account, with models that learned from scratch any possible meaningful function just by experience. The success of PDP was largely due to an efficient mathematical rule, known as *backpropagation*, which adjust the connections between neurons according to a number of input/output examples shown during trainig. The mathematics of learning in deep networks is an evolution and a refinement of the same mathematical rule for learning in PDP models, and in fact Geoffrey Hinton [28] was one of the main contributors to the PDP project.

The "deep" addition to PDP style of feedforward network is just in the number of layers between the input and output layers, usually called "hidden" layers. Neural models can learn increasingly complex function by augmenting the number of units. This way, however, the number of parameters to optimize increases as well, and learning becomes more difficult. In particular, it was observed that increasing the number of units by adding layers was much less efficient than increasing the width of a single hidden layer [10].

A novel learning strategy, again invented by Hinton, succeeded in breaking the limit of no more than three layers [29], paving the road for deep models. Currently, the most successful learning method is stochastic gradient descent [37, 72], not much different from the good old backpropagation.

There is a fundamental difference in aims between the first generation of artificial neural networks and deep neural models. The former was motivated primarily by "Explorations in the Microstructure of Cognition", as the title of the book by Rumelhart and Mcclelland indicated [68]. Conversely, deep neural models are developed with engineering goals in mind, without any ambition or interest in exploring cognition, even if most of the protagonists are the same of earlier artificial neural networks, like Hinton. A striking example is the recent invention of a deep model known as *variational autoencoder* [38, 64], which mathematical formulation is very close to the free–energy principle for predictive brains of Friston [12, 13]. Despite the large resonance of Frinston's theory within cognitive science, all the proposers of variational autoencoder are either unaware or fully disinterested of this coincidence. Most of the components of deeep learning – for example reinforcement learning or recurrent networks – owe indeed a dept to neuroscience and cognitive science, as PDP far legacy, but this connection is now neglected, all that matters is the pragmatic success in applications. The success is so resounding to stimulate some reflections on the relevance of deep

models for cognitive science as a whole [50, 7], or for the so-called "general" artificial intelligence [43, 45]. These are important considerations for cognitive science, but they just happen not to be the issues to be examined here. Our focus is only on the results achieved by deep learning in artificial vision, and their relevance for embodiment in cognitive science.

4 Disembodied Vision

There are several reasons for regarding vision as the case where results achieved by deep learning are challenging for embodied cognition. First, vision is a paradigmatic case used in support of embodied cognition, as seen in §2. Vision is also the most successful field of application for deep learning, as recognized by the scientific community of vision science [80]:

> For decades, perception was considered a unique ability of biological systems, little understood in its inner workings, and virtually impossible to match in artificial systems. But this status quo was upturned in recent years, with dramatic improvements in computer models of perception brought about by 'deep learning' approaches [...] For as long as I can remember, we perception scientists have exploited in our papers and grant proposals the lack of human-level artificial perception systems [...] But now neural networks [...] routinely outperform humans in object recognition tasks [...] Our excuse is gone

An additional reason of interest for vision is that deep models for this application have a peculiar architecture, often regarded closer to the brain than ordinary layered neural networks. This type of architecture is called *Deep Convolutional Neural Network* (DCNN), because it integrates the convolution operation [67] within a layered learned structure. This strategy was first proposed by [15] in the architecture called *Neocognitron*, where "neo" is with reference to his earlier *Cognitron* [14]. The Neocognitron alternates layers of *S–cell* type units with *C–cell* type units, which naming are evocative of the classification in simple and complex cells by Hubel and Wiesel [30]. The S–units act as convolution kernels, while the C–units downsample the images resulting from the convolution, by spatial averaging. The crucial difference from conventional convolution in image processing is that the kernels in Neocognitron are learned. The first version of the Neocognitron learned by unsupervised self–organization [82], with a winner–take–all strategy: only the weights of the maximum responding S–units, within a certain area, are modified, together with those of neighboring cells. A later version [16] used a weak form of supervision: at the beginning of the training the units to be modified in the S–layer are selected manually rather than by winner–take–all, after this first sort of seeding, training proceed in unsupervised way.

The convergence between Neocognitron and the PDP project was done by [46], applying backpropagation to an architecture composed by two layers of Fukushima's S-cell type, followed by ordinary PDP neural layers. It was an early step towards DCNN. Like the artificial neural networks of the PDP project, this

mixture of Neocognitron and backpropagation met with a relative good success, especially in the field of character recognition [47], but it was not the main choice within mainstream computer vision. A major shift came about when DCNNs, like ordinary layered networks, became "deep", once again thanks to the work of Hinton together with his PhD student Krizhevsky [42]. This model dominated the ImageNet Large-Scale Visual Recognition Challenge, the major competition in computer vision. The model dropped the previous error rate for the ImageNet Challenge from 26.0% down to 16.4%. This first success steered computer vision towards DCNNs, and several refinements continued to improve performances, even surpassing those of human subjects [63].

It is crucial for our purposes to see in details how this result has been achieved. ImageNet is an image database organized according to the hierarchy of nouns in the lexical dictionary WordNet, in which each lexical entry is associated with hundreds of images [69]. The Visual Recognition Challenge uses a subset of ImageNet made of 1000 different categories, corresponding to synsets in WordNet, with roughly 1000 images in each category, About 1,2 million images are used for training the models, and 150,000 are used for testing. The DCNN models are simply exposed, several times, to all training images together with their known category. The images are all of the same size, 256×256. The model is unaware of any further information: nothing about the context of each image, nothing about the relations between categories, nothing about the poses each object can assume in space, nothing about affordances exposed by objects, nothing about how objects can change their aspect in time. In summary, the model learns to recognize objects in a fully disembodied way.

The main lesson learned by the computer vision community from embodied cognition [70, 56, 57] should be that trying to understand an image as a static task is hopeless. However, vision turns out to be much easier when the agent interacts with the environment, when vision is treated as an interactive process. Yet artificial vision in embodied and enactive systems have never achieved performance anywhere near to that of DCNN models. Most often active vision models have been developed for very simplified and easy tasks, for example using just the two categories of circles and diamonds [1], or the four categories cat, dog, giraffe and horse [81]. In a comparison of the best active vision models [9], errors on a set of 100 different individual objects were around 40% or worst.

Let us reinstate here that it is out of the scope of this paper the discussion about the pros and cons of deep learning, we provided references for that in §3. Our reflections derive from the empirical observation of the impressive advantage of deep learning in vision, over all other existing methods, including methods that have attempted to adhere to embodied and enactive cognition.

Still, one may object that DCNN models are engineered software far from the way natural vision works, therefore cannot be used to evaluate cognitive theses. Indeed, "neurons" in deep learning models bear little resemblance to their biological cousins. However, recent studies revealed surprising similarities between patterns of activation in layers of convolutional neural models, and patterns of voxels in subjects seeing the same images. One of the first attempt to

relate results of DCNN with the visual system was based on the idea of adding at a given level of an artificial network model a layer predicting in the space of voxel response, and to train this layer on sets of images and corresponding fMRI responses [23]. Using this method, a model very similar to AlexNet [4] was compared with fMRI data [24], training the mapping to voxels on 1750 images. The model responses were predictive of the voxels in the visual cortex above chance, with prediction accuracy slightly below 0.5 for area V1, and of slightly below 0.3 for area LO. The same technique has been further exploited, by generating artificial fMRI data, using stimuli of classical vision experiments, such as simple retinotopy or face/places contrast, for which good agreement between synthetic fMRI responses and DCNN was found [11].

The use of synthetic fMRI data is pursued also with a different strategy [36], constructing a statistical model of the activity in higher visual cortex, by combining a wide range of information from previous studies. This model allows the interpolation of novel responses as needed for experimental purposes. Using this method Bryan Tripp [77] was able to test similarities with cortical responses and DCNN models, on various different properties: population sparseness; orientation, size and position tuning; occlusion; clutter; and so on. The DCNNs tested were AlexNet [42] and VGG-16 [74]

An alternative method for comparing DCNN models and fMRI responses was offered by the representational similarity analysis, introduced by Nikolaus Kriegeskorte [41, 40]. This method can be applied to any sort of distributed responses to stimuli, computing one minus the correlation between all pairs of stimuli. The resulting matrix is especially informative when the stimuli are grouped by their known categorial similarities. The whole idea is that the responses across the set of stimuli reflect an underlying space in which reciprocal relations correspond to relations between the stimuli. This is exactly the idea of *structural representations*, one of the fundamental concepts in cognitive science [76, 54, 60]. The representational similarity analysis is applied by [35] in comparing responses in the higher visual cortex, measured with fMRI in humans, and with cell recording in monkeys, with several artificial models. This study is very interesting because it includes, in addition to AlexNet, few models with more biological plausibility.

The most biologically plausible model is VisNet [83, 66], organized into five layers, which connectivity approximates the sizes of receptive fields in V1, V2, V4, posterior inferior temporal cortex, and inferior temporal cortex. The network learns by unsupervised self-organization [84] with synaptic modifications derived from Hebbian rule[26]. VisNet does not fully adhere to embodied and enactive cognition, however, it attempts to include in a biologically plausible model the perception of an object when acting on it or when the object is moving. For this purpose, learning includes a specific mechanism called *trace memory*, since learning of a single cell is affected by a decaying trace of previous cell activity. This rule is an attempt to reproduce an embodied and enactive component of vision, where invariant recognition of objects is learned by seeing them when moving under various different prospective.

Kriegeskorte and co–workers [35] constructed several representational similarity matrices on a set of natural images spanning multiple animate and inanimate categories, comparing voxels in the inferior temporal cortex (IT) with models (the study actually compared 37 different models, of which only AlexNet and VisNet are of interest here). The analysis revealed that AlexNet was significantly more similar to the IT structural representation of the categorical distinction animate/inanimate than VisNet.

This impetus of studies on the analogies between DCNN and the visual system has led to a broad discussion in the visual neuroscience community on the relevance of deep learning models for their scientific objective. Positions span from a mostly positive acceptance [18, 80], to a cautious interest [48, 22], down to more skeptical stances [55, 65, 8]. There is obviously a large number of structural features of the visual system that drastically departs from a DCNN model. Just to mention few: visual maps in the cortex have many strong interconnections and a very large number of weaker connections [79, 78, 51]; receptive field sizes change within a cortical map, and the degree of changes is larger in higher cortical areas [34]; receptive field are also modulated by tasks [39]; scene dynamics affects recognition areas, in addition to motion areas [75]. And, most of all, it is certainly true that the visual system in the brain is embodied and enactive [6, 62]. Still, the point is that DCNN is at the same time the only type of model that achieves human performances in vision, and the only model that displays similarities with brain activation of subjects seeing the same pictures. This fact may suggest that there is a core processing, necessary for discriminating the content of the scene, which works essentially as a computation on local patterns of information. This local process seems to be relatively independent from environmental and bodily cues.

5 Conclusions

The performance achieved by deep learning models in visual pattern recognition is a highly unexpected circumstance in contemporary cognitive science, so deeply influenced by the 4E cognition account and, in particular, by embodiment concerns. It should be simply a thing not occurring. Despite mainstream expectations, the fact that it actually happened shows how much the opposition between classical cognitive science and embodiment concerns is highly misleading. As above mentioned, computational psychology and classical mental representations are ecologically neutral. In addition, we are faced with a sort of micro–singularity in the progress of cognitive science, i.e., a case in which artificial intelligence surpass human intelligence: a new challenge arising from the field of artificial intelligence for both the cognitive science and common sense. This is perhaps the moral of the story: we have to revise our current theoretical expectations in order to accept that biologically–inspired and embodied cognitive architectures are not a warranty for a successful processing.

References

1. Beer, R.D.: The dynamics of active categorical perception in an evolved model agent. Adaptive Behavior **11**, 209–243 (2003)
2. Brooks, R.A.: Intelligence without representation. Artificial Intelligence **47**, 139–159 (1991)
3. Calabrese, A., Woolley, S.M.N.: Coding principles of the canonical cortical microcircuit in the avian brain. Proceedings of the Natural Academy of Science USA **112**, 3517–3522 (2015)
4. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. CoRR **abs/1405.3531** (2014)
5. Chemero, A.: Radical embodied cognitive science. MIT Press, Cambridge (MA) (2009)
6. Churchland, P.S., Ramachandran, V., Sejnowski, T.: A critique of pure vision. In: Koch, C., Davis, J. (eds.) Large-Scale Neuronal Theories of the Brain. MIT Press, Cambridge (MA) (1994)
7. Cichy, R.M., Kaiser, D.: Deep neural networks as scientific models. Trends in Cognitive Sciences **23**, 305–317 (2019)
8. Conway, B.R.: The organization and operation of inferior temporal cortex. Annual Review of Vision Science **4**, 19.1–19.22 (2018)
9. De Croon, G.C., Sprinkhuizen-Kuyper, I.G., Postma, E.: Comparing active vision models. Image and Vision Computing **27**, 374–384 (2009)
10. de Villers, J., Barnard, E.: Backpropagation neural nets with one and two hidden layers. IEEE Transactions on Neural Networks **4**, 136–141 (1992)
11. Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B.: Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage **152**, 184–194 (2017)
12. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. Journal of Physiology – Paris **100**, 70–87 (2006)
13. Friston, K., Stephan, K.E.: Free–energy and the brain. Synthese **159**, 417–458 (2007)
14. Fukushima, K.: Cognitron: a self-organizing multilayered neural network. Biological Cybernetics **20**, 121–136 (1975)
15. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36**, 193–202 (1980)
16. Fukushima, K.: Neocognitron: a hierarchical neural network capable of visual pattern recognition. Neural Networks **1**, 119–130 (1988)
17. Gallup, G.: Chimpanzees: Self-recognition. Science **167**, 86–87 (1970)
18. Gauthier, I., Tarr, M.J.: Visual object recognition: Do we (finally) know more now than we did? Annual Review of Vision Science **2**, 16.1–16.20 (2016)
19. Gelder, T.v.: What might cognition be, if not computation? Journal of Phylosophy **91**, 345–381 (1995)
20. Gibson, J.J.: The senses considered as perceptual systems. Houghton Miflin, Boston (MA) (1966)
21. Gibson, J.J.: The Ecological Approach to Perception. Houghton Miflin, Boston (MA) (1979)
22. Grill-Spector, K., Weiner, K.S., Gomez, J., Stigliani, A., Natu, V.S.: The functional neuroanatomy of face perception: from brain measurements to deep neural networks. Interface Focus **8**, 20180013 (2018)

23. Güçlü, U., van Gerven, M.A.J.: Unsupervised feature learning improves prediction of human brain activity in response to natural images. PLoS Computational Biology **10**, 1–16 (2014)
24. Güçlü, U., van Gerven, M.A.J.: Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience **35**, 10005–10014 (2015)
25. Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., Fawzy, M., Jia, B., Jia, Y., Kalro, A., Law, J., Lee, K., Lu, J., Noordhuis, P., Smelyanskiy, M., Xiong, L., Wang, X.: Applied machine learning at Facebook: A datacenter infrastructure perspective. In: IEEE International Symposium on High Performance Computer Architecture (HPCA). pp. 620–629 (2018)
26. Hebb, D.O.: The Organization of Behavior. John Wiley, New York (1949)
27. Heras-Escribano, M.: The Philosophy of Affordances. Palgrave Macmillan, London (2019)
28. Hinton, G.E., McClelland, J.L., Rumelhart, D.E.: Distributed representations. In: Rumelhart and McClelland [68], pp. 77–109
29. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **28**, 504–507 (2006)
30. Hubel, D., Wiesel, T.: Receptive fields and functional architecture of mokey striate cortex. Journal of Physiology **195**, 215–243 (1968)
31. Hutto, D.D., Myin, E.: Radicalizing enactivism: basic minds without content. MIT Press, Cambridge (MA) (2013)
32. Isaac, A.M.: Embodied cognition as analog computation. Italian Journal of Cognitive Science **14**, 239–259 (2018)
33. Karten, H.J.: Vertebrate brains and evolutionary connectomics: on the origins of the mammalian neocortex. Philosophical transactions of the Royal Society B **370**, 20150060 (2015)
34. Kay, K.N., Winawer, J., Mezer, A., Wandell, B.A.: Compressive spatial summation in human visual cortex. Journal of Neurophysiology **110**, 481–494 (2013)
35. Khaligh-Razavi, S.M., Kriegeskorte, N.: Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Computational Biology **10**, e1003915 (2014)
36. Khan, S., Tripp, B.P.: One model to learn them all. CoRR **abs/1706.05137** (2017)
37. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of International Conference on Learning Representations (2014)
38. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proceedings of International Conference on Learning Representations (2014)
39. Klein, B., Harvey, B.M., Dumoulin, S.O.: Attraction of position preference by spatial attention throughout human visual cortex. Neuron **84**, 227–237 (2014)
40. Kriegeskorte, N.: Relating population-code representations between man, monkey, and computational models. Frontiers in Neuroscience **3**, 363–373 (2009)
41. Kriegeskorte, N., Mur, M., Bandettini, P.: Representational similarity analysis – connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience **2**, 4 (2009)
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1090–1098 (2012)
43. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and Brain Science **40**, 1–72 (2017)
44. Lakoff, G., Johnson, M.: Philosophy in the Flesh. The Embodied Mind and its Challenge to Western Thought. Basic Books, New York (1999)

45. Landgrebe, J., Smith, B.: Making AI meaningful again. Synthese **https://doi.org/10.1007/s11229-019-02192-y**, 1–21 (2019)
46. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation **1**, 541–551 (1989)
47. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**, 2278–2324 (1998)
48. Lehky, S.R., Tanaka, K.: Neural representation for object recognition in inferotemporal cortex. Current Opinion in Neurobiology **37**, 23–35 (2016)
49. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017)
50. López-Rubio, E.: Computational functionalism for the deep learning era. Minds and Machines **28**, 667–688 (2018)
51. Markov, N., Ercsey-Ravasz, M.M., Gomes, A.R.R., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M.A., Sallet, J., Gamanut, R., Huissoud, C., Clavagnier, S., Giroud, P., Sappey-Marinier, D., Barone, P., Dehay, C., Toroczkai, Z., Knoblauch, K., Essen, D.C.V., Kennedy, H.: A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cerebral Cortex **24**, 17–36 (2014)
52. Miłkowski, M.: Embodied cognition meets multiple realizability. Italian Journal of Cognitive Science **14**, 349–364 (2018)
53. Noë, A.: Action in Perception. MIT Press, Cambridge (MA) (2004)
54. O'Brien, G., Opie, J.: Notes toward a structuralist theory of mental representation. In: Clapin, H., Staines, P., Slezak, P. (eds.) Representation in Mind – New Approaches to Mental Representation. Elsevier, Amsterdam (2004)
55. Olshausen, B.A.: Perception as an inference problem. In: Gazzaniga, M.S. (ed.) The Cognitive Neurosciences, pp. 295–304. MIT Press, Cambridge (MA) (2014), fifth edition
56. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Science **24**, 939–1031 (2001)
57. Pfeifer, R., Bongard, J.: How the body shapes the way we think: a new view of intelligence. MIT Press, Cambridge (MA) (2007)
58. Piccinini, G.: Computation and representation in cognitive neuroscience. Minds and Machines **28**, 1–6 (2018)
59. Plebe, A.: Cognition and computation. Italian Journal of Cognitive Science **14**, 281–286 (2018)
60. Plebe, A., De La Cruz, V.M.: Neural representations beyond "plus X". Minds and Machines **28**, 93–117 (2018)
61. Prior, H., Schwarz, A., Güntürkün, O.: Mirror-induced behavior in the magpie (pica pica): Evidence of self-recognition. PLoS Biology **6**, 1–9 (2008)
62. Ramachandran, V., Arnel, C., Foster, C., Stoddard, R.: Object recognition can drive motion perception. Nature **395**, 852–853 (1998)
63. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: A comprehensive review. Neural Computation **29**, 2352–2449 (2017)
64. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Xing, E.P., Jebara, T. (eds.) Proceedings of Machine Learning Research. pp. 1278–1286 (2014)
65. Robinson, L., Rolls, E.T.: Invariant visual object recognition: biologically plausible approaches. Biological Cybernetics **109**, 505–535 (2015)

66. Rolls, E.T., Stringer, S.M.: Invariant visual object recognition: A model, with lighting invariance. Journal of Physiology – Paris **100**, 43–62 (2006)
67. Rosenfeld, A.: Picture Processing by Computer. Academic Press, New York (1969)
68. Rumelhart, D.E., McClelland, J.L. (eds.): Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge (MA) (1986)
69. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115**, 211–252 (2015)
70. de Sa, V.R., Ballard, D.H.: Category learning through multi-modality sensing. Neural Computation **10**(5) (1998)
71. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61**, 85–117 (2015)
72. Schmidt, M., Roux, N.L., Bach, F.: Minimizing finite sums with the stochastic average gradient. Mathematical Programming **162**, 83–112 (2017)
73. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature **529**, 484–489 (2016)
74. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2015)
75. Stigliani, A., Jeska, B., Grill-Spector, K.: Encoding model of temporal processing in human visual cortex. Proceedings of the Natural Academy of Science USA **1914**, E11047–E11056 (2017)
76. Swoyer, C.: Structural representation and surrogative reasoning. Synthese **87**, 449–508 (1991)
77. Tripp, B.P.: Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks. In: International Joint Conference on Neural Networks. pp. 3551–3560 (2017)
78. Van Essen, D.C.: Organization of visual areas in macaque and human cerebral cortex. In: Chalupa, L., Werner, J. (eds.) The Visual Neurosciences. MIT Press, Cambridge (MA) (2003)
79. Van Essen, D.C., DeYoe, E.A.: Concurrent processing in the primate visual cortex. In: Gazzaniga, M.S. (ed.) The Cognitive Neurosciences. MIT Press, Cambridge (MA) (1994)
80. VanRullen, R.: Perception science in the age of deep neural networks. Frontiers in Psychology **8**, 142 (2017)
81. Volpi, N.C., Quinton, J.C., Pezzulo, G.: How active perception and attractor dynamics shape perceptual categorization: a computational model. Neural Networks **60**, 1–16 (2014)
82. von der Malsburg, C.: Network self-organization. In: Zornetzer, S.F., Davis, J., Lau, C. (eds.) An Introduction to Neural and Electronic Networks. Academic Press, New York (1990)
83. Wallis, G., Rolls, E.: Invariant face and object recognition in the visual system. Progress in Neurobiology **51**, 167–194 (1997)
84. Willshaw, D.J., von der Malsburg, C.: How patterned neural connections can be set up by self-organization. Proceedings of the Royal Society of London **B194**, 431–445 (1976)