Vol 4 | Issue 1 | Pages 80-86

Trends in Artificial Intelligence
ISSN: 2643-6000

Review Article

DOI: 10.36959/643/306

The Hold-Up of the Century: Neural Networks are Coming from Cognitive Science and not Machine Learning. Perspectives to Avoid a New Dark Age of Artificial Intelligence

*Martial Mermillod**

Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000, Grenoble, France

Abstract

Since the new revolution in Artificial Intelligence (AI) and the advent of Deep Learning, the massive trend in AI is mainly driven by the computer science community. However, this statement omits the sometimes complicated relationship between the neural network modeling community and the machine learning community prior to 2012, when "artificial intelligence" or "artificial neural networks" were considered as "has been" at that time. Today, this widespread revolution of AI is also forgetting on the way the multidisciplinary origins of artificial neural networks. The purpose of this article is to (1) recall the origins of artificial neural networks, (2) highlight the fact that deep neural networks simulate a very small area of the human brain in a (very) simplistic way, and (3) provide insights into the importance of continuing interdisciplinary scientific research on artificial intelligence in order to go beyond the current (and important) limits of artificial intelligence.

Introduction

A Venn diagram can lie. For instance, the Venn diagram shown in Figure 1, commonly used to present Deep Learning compared to other machine learning methods, gives the false idea that Deep Learning comes from research mainly conducted by the Machine Learning community. This article is not intended to diminish the importance of computer science in the development of AI. However, the objective is to show that an arrogant view of Machine Learning, which would not need the other disciplines which were at the origin of the current revolution of artificial intelligence (psychology, neurosciences, cognitive sciences), will most likely lead to a new Dark Age of AI.

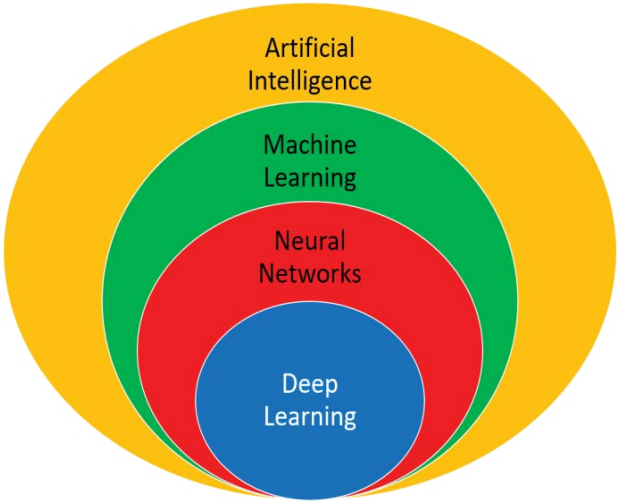

Figure 1: Venn diagram often illustrating the inclusion of Deep Learning among other scientific communities.

The problem with this type of simplistic two-dimensional view of the different scientific communities is that it hides the roots of neural network modeling and the conflicting, or sometimes virtuous, relationships between these communities, many of them not included in this Venn diagrams. An accurate representation should require a three- dimensional Venn diagram (across time) and the massive involvement of other scientific fields in order to understand the true origin of artificial neural networks. Ten years ago, artificial neural networks were generally considered by the mainstream in

***Corresponding author:** Martial Mermillod, Laboratoire de Psychologie et Neuro Cognition, Université Grenoble Alpes, BP 4738040 Grenoble Cedex 9, France

Accepted: November 11, 2020

Published online: November 13, 2020

Citation: Mermillod M (2020) The Hold-Up of the Century: Neural Networks are Coming from Cognitive Science and not Machine Learning. Perspectives to Avoid a New Dark Age of Artificial Intelligence. Trends Artif Intell 4(1):80-86

computer science and mathematics as an inefficient black box, impossible to control and predict, a disgrace for artificial intelligence. This distrust has gone to the point of replacing the term "deep neural networks" with the term "deep learning" for marketing purposes.

However, today Deep Learning (or more precisely, Deep Neural Networks), and specifically Convolutional Neural Network (CNN) have been considered to be the new benchmark of Artificial Intelligence (AI), revealing groundbreaking possibilities in scientific, industrial, economic as well as societal and ethical domains. In particular, these algorithms have involved massive and widespread investments in the engineering and industrial communities. Concomitantly with this sudden trend, the interdisciplinary roots of neural network modeling have been forgotten and, consequently, so too has the understanding of what these powerful algorithms are able to do and what they are not or not yet able to do. Unfortunately, many researchers in the scientific community, including many actual users of CNN, are often unaware of the interdisciplinary roots of this type of AI and use it as a magicwand without any awareness of its limitations. The aim of this article is to recall the brain- inspired origins of deep-learning. We will then consider the perspectives of neural network modeling in order to illustrate the dangers inherent in using CNN as a magic wand, without any awareness of the part of the brain for which this type of neural network is the formal analog and of what types of cognitive functions such neural architectures are not, in their current forms, suitable for.

Where does the Current Revolution in AI Come from and are Psychology and Neuroscience Distracting for Artificial Intelligence?

From the late 1990s to 2012, research on artificial neural networks was seen as “has been”. Geoffrey Hinton himself, psychologist and computer scientist, recently proclaimed "godfather of AI", stated that during his studies in the 1970s: "in psychology they had very, very simple theories, and it seemed to me it was sort of hopelessly inadequate to explaining what the brain was doing. So then I took some time off and became a carpenter."Geoffrey Hinton probably forgot that his own work on neural network modeling was the direct offspring of other psychologists and neuroscientists such as McCulloch & Pitts, Donald Hebb and Frank Rosenblatt. His own career was launched by the seminal work of another psychologist, David Rumel hart, on back-propagation algorithm.

This disconnection between psychology and neuroscience in one side and computer science on the other is widespread in today's AI community. For example, in 2019, the French parliamentarian and mathematician Cédric Villani launched a large national project on the basis of his earlier report “Making sense of artificial intelligence”. Following the original call, an international board consisting of the leading international experts in AI evaluated the link between AI and psychology/ neuroscience, in the following terms: “Reviewers were unanimous in finding the emphasis on neuroscience to represent a distraction from the goals of the program” [namely, to develop AI]. This very important comment stated by an international board that brings together many prominent AI researchers is extremely thought- provoking and suggested to us several important questions concerning the past, present and future of AI and, more particularly, the relevance (or otherwise) of studying human intelligence in order to develop AI.

Interestingly, most of the current international AI funding goes mainly to people who were most opposed to neural network modeling ten years ago. It's very easy to find out by following their track-records. So, what suddenly changed in their minds? The current revolution in AI has mainly been driven by researchers such as Geoffrey Hinton, Yoshua Bengio and Yann LeCun. A key event occurred in December 2012 at the Neural Information Processing Systems conference [1] when Geoffrey Hinton and his team revealed outstanding image classification performance by a CNN. This groundbreaking capability was made possible by the conjunction of three factors: i) efficient algorithms in neural network architectures; it should be noted, however, that CNN had already been discussed in the scientific literature for some time [2]; ii) the possibility of accessing a vast amount of data, thus permitting the neural network to learn the statistical regularities present in these data, and iii) the use of powerful Graphics Processing Units (GPU) and more recently Neural Processor Units (NPU) that made it possible to speed-up the processing times of these algorithms. In the light of these impressive performances, massive investments were made by the largest internet companies, among others, in order to develop the new potential of AI in vast fields of industrial application. For instance, autonomous vehicles, drones, robotics, the health industry, security and defense systems now make widespread use of CNN (or deep learning) in addition to symbolic AI and other machine-learning methods that have traditionally been used by internet companies.

To what Extent is Deep Learning Inspired by the Human Brain?

The origin of artificial neural networks can be found in princeps discoveries in neuroscience and cognitive psychology that led to the first model of formal neurons [3] and described how information is transferred by means of synaptic weights through the soma of a neuron to its axon (i.e. a transfer function). Complementing this work on formal neurons, the Canadian neuropsychologist Donald Hebb specified how biological neural networks are able to learn by modifying the synaptic weights between pre- and post-synaptic neurons [4]. This seminal work led to the first artificial neural network, which was developed by the cognitive psychologist Frank Rosenblatt [5] in 1958: the Perceptron. Naturally, these princeps mathematical attempts to understand how neurons generate behavior were followed by important developments, for instance at the level of the temporal dynamic of the transfer functions in the neurosciences [6] or the use of more efficient gradient descent algorithms in computer science in order to optimize the modifications of synaptic weights in deep layers of formal neurons [7]. The backpropagation algorithm, which makes it possible to optimize the modifications of the synaptic weights, and the advent of the Multi-Layer Perceptron (MLP) developed by the psychologist David E. Rumelhart and

the Parallel & Distributed Processing Group were responsible for the widespread dissemination of these computational models in the cognitive sciences during the late 80's and early 90's. Deep learning is the direct offspring of these studies in biological and artificial neural systems [7]. However, despite the clear interdisciplinary origin of deep learning, the extent to which such artificial neural networks still share principles in common with the human brain is a matter of debate.

Deep Learning Share Principles in Common with Very Specific Parts of the Human Brain

The resurgence of neural network modeling was based on image recognition and classification and this is the field in which these neural systems have found their most widespread application. In the human brain, these cognitive functions are achieved via the occipito-temporal pathway from perceptual areas (e.g. the primary visual cortex, or V1) to low-level cognitive areas (e.g. the infero-temporal cortex). To this end, CNN use different layers of convolution and pooling before formal classification by a MLP [8]. A naïve question (but one very often raised by experts of AI not familiar with cognitive neuroscience) is: what is the formal equivalence in the brain with these mathematical techniques?

At the root of this neural pathway, it has been known for some time that receptive fields representing the probability of a neuron responding after a physical stimulation are able to simulate V1 simple cells [9], and this fact was mentioned by Yann LeCun in the very early days of Deep Learning [2], before neural networks had started to be used as a magic wand in the field of AI. Subsequent studies have shown that a Gabor function tuned to different orientations and spatial frequencies nicely approximate the receptive fields of V1 simple cells [10] and that an artificial neural network is able to efficiently reduce the dimensionality of the physical world by means of a convolution of the kernel of these receptive fields in the spatial domain [11] or, as demonstrated by our own work, by means of a multiplication in the spectral domain [12,13]. In other words, in contradiction of the claim made by many parts of the community that the convolutional layers of CNN is not bio-inspired, this mathematical operation in the first layers of neurons is actually a good approximation to the receptive fields of biological neurons (albeit simplified and optimized compared to biological processes). But what about the pooling process that follows each convolution layer? Is this process inspired by the human brain?

In their seminal work, Hubel & Wiesel not only revealed the properties of the receptive fields of V1 neurons but also pointed to a pooling process that takes place between subsequent layers of perceptual neurons [9]. For instance, the V1 complex cells achieve a larger receptive field than V1 simple cells by pooling together different simple cells attuned to the same orientation and spatial frequency. This pooling process continues along the infero-temporal pathway of the primate cortex to reach a level of abstraction to general shapes [14,15] and also engenders category-specific neurons that are sensitive to entire categories irrespective of the basic perceptual properties [16]. Similar neurons have been found in the human medial temporal lobe [17] but, even more surprisingly, intra-cellular recordings in the human hippocampus have shown that specific neurons can process a level of abstraction that makes them responsive to high-level abstract concepts such as a specific individual identity [18]. For instance, individual neurons in the human hippocampus not only exhibit a selective response to Jennifer Aniston, the Tower of Pisa or Pamela Anderson, irrespective of surface properties (pose, orientation, scaling, picture or drawn stimulus) but also achieve direct access to the concept through the written name "Pamela Anderson" [18]. With regard to the potential future development of AI, it should be noted that this activity of concept-specific neurons is strongly correlated with the

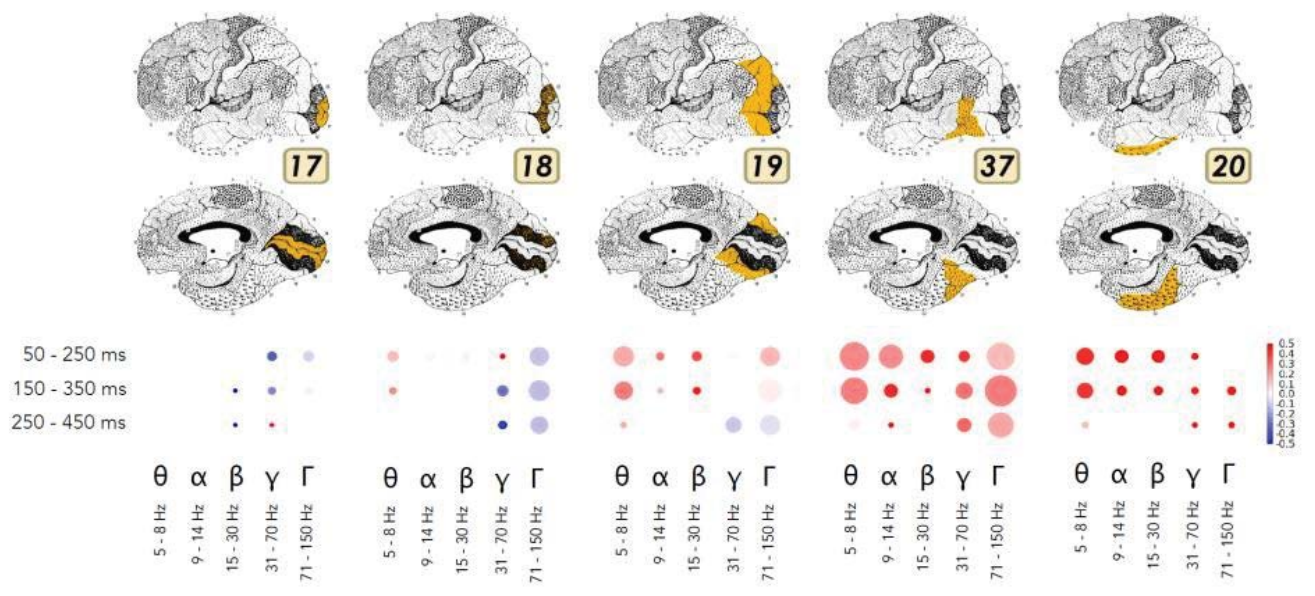

Figure 2: Current research shows that CNN essentially simulates, in a very simplistic way, the ventral stream from the occipito-temporal cortex [22].

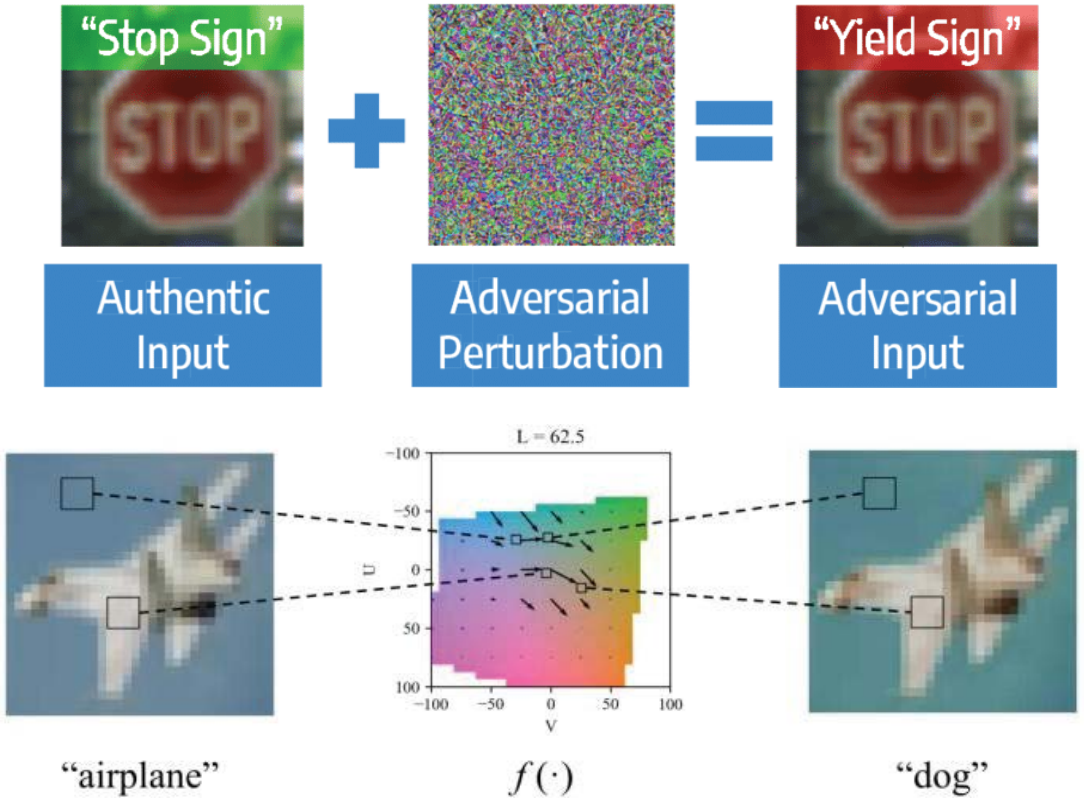

Figure 3: Examples of adversarial attacks.

awareness of the stimulus [19].

In summary, although they are simplified and mathematically optimized compared to a biological brain, CNN reproduce this process of receptive fields/pooling along the occipito- temporal stream. A large set of experimental data in psychology and cognitive neuroscience has recently confirmed this by means of parallel multielectrode intracranial recordings in rhesus macaques [20], fMRI in humans [21] as well as iEEG (Figure 2) in the human visual cortex [22].

Perspectives: Beyond the Current Limits of Deep Learning

The outstanding performance of CNN in classification and recognition tasks was obtained by simulating basic principles of the human brain at the level of the occipito-temporal pathway (in a simplified but mathematically optimized manner). This cognitive efficiency of deep learning has been extended to other perceptual channels (e.g. speech recognition, reading capacities) and/or dynamic processes in order to predict temporal events by including recurrent connections among different layers of artificial neural networks [23-25]. However, despite those success of CNN for pure perceptual categorization tasks, severe limitations and security threat has been discovered by the research community, here is some examples:

The case of adversarial attacks

The introduction of small disturbances on the stimulus to be categorized which are imperceptible to human perception are sufficient to prompt the model to make a bad prediction with high confidence [26,27]. This phenomenon, called adversarial attack, is capable of transforming a stop sign into a priority road sign, quite problematic for a safe autonomous vehicle! It is also possible to transform an airplane into a dog for military drones (Figure 3), which could be even more than problematic. Engineers and computer scientists has been deployed by universities and private companies to find the mathematical roots of this phenomenon and countering these potential adversary attacks, with little success so far.

Interestingly, after a careful review of this literature, few articles addressed the question about: why humans are not sensitive to these adversarial attacks.? It's a pity because, contrary to the statement of Geoffrey Hinton, psychology and cognitive neuroscience "have more than one clue" to address this question. Several models, assessed by experimental data, can explain why and more importantly, how it would be possible to counter adversarial attacks in a simple and efficient way [12,25,28].

The case of catastrophic forgetting

A well-known limitation of artificial neural networks is the inability of a simple neural network to perform lifelong learning. After learning task A, if you stop learning about task A and switch to task B, it will result in an abrupt lost of performance in task A as information relevant to the new task B is incorporated. This phenomenon, called catastrophic forgetting, occurs when the network is sequentially trained on several tasks on static [29] but also dynamic [30] sequences. Despite the different methods proposed to overcome catastrophic forgetting [29], little attention has been given to understand-

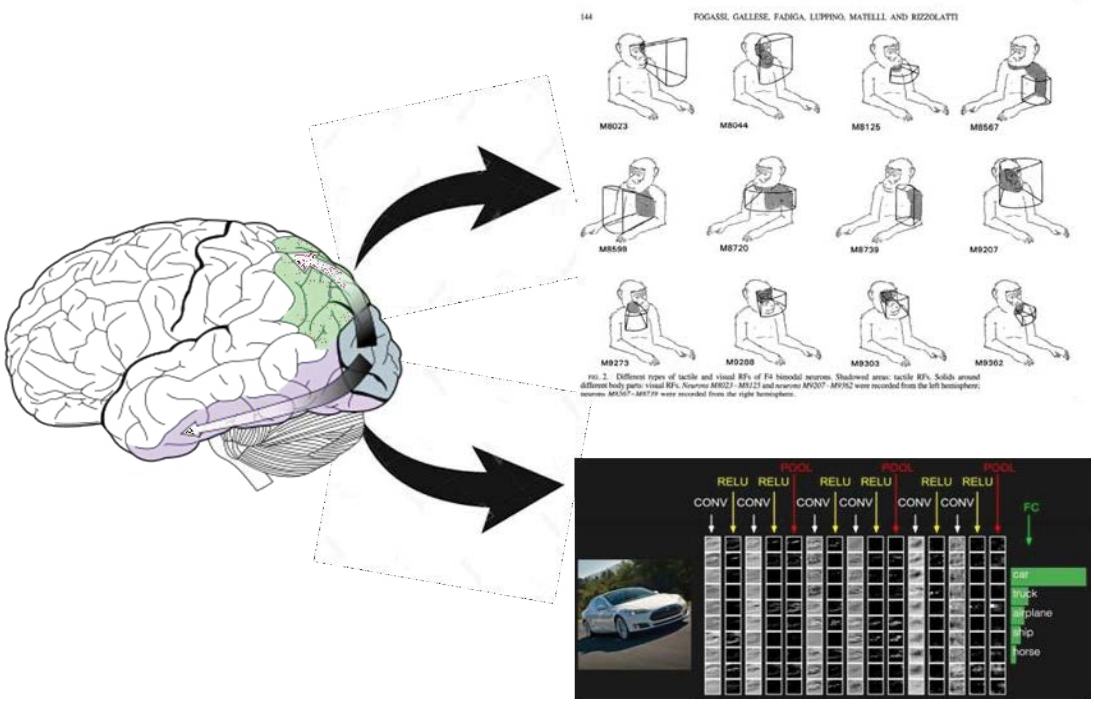

Figure 4: The ventral stream is related to visual recognition, but is in no case suitable for motor control, handle (among a larger neural network) by the dorsal stream, a neural pathway with very different functional properties. Yann LeCun has replicated the ventral stream thirty years ago and it was the root of deep learning. We hope that the scientific community will not wait 30 years more to replicate the dorsal stream for motor control.

Figure 5: A biker killed by an autonomous vehicle.

ing why humans are not sensitive to this phenomenon [30] and especially how to implement a similar method (compared to human brain) for a simpler, agnostic and resource efficient way to overcome catastrophic forgetting [30].

The case of motor control

CNN are currently used for visual recognition in drones or autonomous vehicles. However, as shown on figure 2, CNN in their current forms are simulating in a very simplistic manner cortical areas dedicated to perceptual recognition tasks, but in no case motor control. When it comes to navigating in and representing a complex 3- dimensional world environment, human drivers hopefully use other parts of the brain to efficiently manage complex driving situations (Figure 4). For instance, it has been known for some time in the fields of psychology and cognitive neuroscience that a separate occipito- parietal stream, with specific properties in terms of the spatial and temporal frequencies of the neurons involved, exists in the human brain and permits efficient orientation and action [31].

This anatomical pathway, which interacts strongly with the occipito-temporal pathway, is involved in movement detection [32], action programming and understanding through the mirror neuron system [33,34], as well as in attentional processes [35].

The case of anticipation and meaning of visual scenes

Figure 5 illustrates the last perception of an autonomous vehicle just before killing the biker crossing the road in front

of it. What is the interpretation of the scene for a human driver? Obviously, a human carrying a bicycle. However, the scene understanding for a CNN will likely be very different (e.g. a blue jeans and a wheel crossing the road).

Why this huge difference in interpretation? Human perception is under the massive influence of high-level cognitive processes [25,28,36]. This could be an interesting model (again, "more than a clue") as to why a human driver seeing a human on an advertisement board will not suddenly stop the car in case the people on the board decide to jump off the sign and cross the road! It could be the case for an autonomous vehicle.

Conclusion: Toward a New Dark Age of AI?

Each of the examples provided above is a real shame for artificial intelligence and would be enough to immediately stop the current investments for the development of AI. The huge amount of funding in mathematics and engineering alone, with sometimes little success to efficiently address each of these problems, shows that these disciplines are not effective alone to respond to these questions effectively.

The aim of this article is not to restrain current research in AI, given the tremendous successes of neural networks over the last decade. On the contrary, we want to point out the importance of interdisciplinary research, given that the most recent evidence in psychology and cognitive neuroscience clearly indicates different simple, efficient and reliable solutions, related to other functional properties of the brain (involving other areas than the most famous ventral stream simplistically addressed by CNN). The outstanding results of CNN in pattern recognition were obtained by combining different fields within an interdisciplinary perspective involving neuroscience, psychology, computer science, mathematics or, more generally, cognitive sciences. The same interdisciplinary research will be needed in order to handle other sides of human intelligence adequately. We assert here that the current "blind" application of neural network modeling without regard to their origin and the fundamental properties of these artificial neural systems might not only be useless, but might also be dangerous and, in some cases, even kill people. The need to conduct genuine interdisciplinary research is now more pressing than ever for a more efficient, safer and more reliable AI. With regard to the history of AI, the current issue is just to avoid a new "Dark Age" of AI.

Acknowledgment

This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003) to Martial Mermillod.

References

1. Krizhevsky A, Sutskever I and Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 1097-1105.
2. LeCun Y, Boser B, Denker JS, et al. (1989) Back propagation applied to handwritten zip code recognition. Neural computation 1: 541-551.
3. McCulloch WS, PittsW (1943) A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5: 115-133.
4. Hebb DO (1949) The organization of behavior: A neuropsychological theory. Wiley 34: 336-337.
5. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65: 386- 408.
6. Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117: 500-544.
7. Rumelhart DE, Hinton GE, McClelland JL, et al. (1986) Parallel distributed processing: Explorations in the Microstructure of Cognition (Vol 1 and 2). Cambridge MA: MIT press.
8. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436.
9. Hubel DH and Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195: 215-243.
10. Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J neurophysiol 58: 1233-1258.
11. Wiskott L, Krüger N, Kuiger N, et al. (1997) Face recognition by elastic bunch graph matching IEEE. Transactions on pattern analysis and machine intelligence 19: 775-779.
12. Mermillod M, Bonin P, Mondillon L, et al. (2010) Coarse scales are sufficient for efficient categorization of emotional facial expressions: Evidence from neural computation. Neurocomputing 73: 2522-2531.
13. Mermillod M, Guyader N, Chauvin A (2005) The coarse-to-fine hypothesis revisited: Evidence from neuro-computational modeling. Brain and Cognition 57: 151- 157.
14. Tanaka K, Saito HA, Fukada Y, et al. (1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey. J neurophysiol 66: 170-189.
15. Tanaka K (1996) Inferotemporal cortex and object vision. Annual review of neuroscience 19: 109-139.
16. Vogels R (1999) Categorization of complex visual images by rhesus monkeys Part 2: single-cell study. The European Journal of Neuroscience 11: 1239-1255.
17. Kreiman G, Koch C, Fried I (2000) Category-specific visual responses of single neurons in the human medial temporal lobe. Nat Neurosci 3: 946-953.
18. Quiroga RQ, Reddy L, Kreiman G, et al. (2005) Invariant visual representation by single neurons in the human brain. Nature 435: 1102-1107.
19. Quiroga RQ, Mukamel R, Isham EA, et al. (2008) Human single-neuron responses at the threshold of conscious recognition. Proceedings of the National Academy of Sciences 105: 3599-3604.
20. Yamins DL, Hong H, Cadieu CF, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111: 8619-8624.
21. Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised but not unsupervised models may explain IT cortical representation PLoS Computat Bio 10: e1003915.
22. Kuzovkin I, Vicente R, Petton M, et al. (2018) Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. Communications biology 1: 107.

23. Elman JL (1990) Finding structure in time. Cognitive science 14: 179-211.

24. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18: 602-610.

25. Mermillod M, Bourrier Y, David E, et al. (2019) The importance of recurrent top-down synaptic connections for the anticipation of dynamic emotions. Neural Netw 109: 19-30.

26. Szegedy C, Zaremba W, Sutskever I, et al. (2013) Intriguing properties of neural networks. arXiv:1312.6199.

27. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. arXiv:1608.04644.

28. Bar M, Kassam KS, Ghuman AS, et al. (2006) Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A 103: 449-454.

29. Kirkpatrick J, Pascanu R, Rabinowitz N, et al. (2017) Overcoming catastrophic forgetting in neural networks Proc Natl Acad Sci U S A 114: 3521-3526.

30. Ans B, Rousset S, French RM, et al. (2010) Self-refreshing memory in artificial neural networks: Learning temporal sequences without catastrophic forgetting. Connection Science 16: 71-99.

31. Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends Neurosci 15: 20-25.

32. Newsome WT, Pare EB (1988) A selective impairment of motion perception following lesions of the middle temporal visual area (MT). J Neurosci 8: 2201-2211.

33. Gallese V, Fadiga L, Fogassi L, et al. (1996) Action recognition in the premotor cortex. Brain 119: 593-609.

34. Rizzolatti G, Fadiga L, Gallese V, et al. (1996) Premotor cortex and the recognition of motor actions. Cognitive brain research 3: 131-141.

35. Buschman TJ, Miller EK (2007) Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science 315: 1860-1862.

36. Kauffmann L, Ramanoël S, Peyrin C (2014) The neural bases of spatial frequency processing during scene perception. Front Integr Neurosci 8: 37.

DOI: 10.36959/643/306