Talk: The Explanatory Problems of Deep Learning in Artificial Intelligence and Computational Cognitive Science: Two Possible Research Agendas

AISC 2017, 14th Conference of the Italian Association of Cognitive Sciences, Bologna.

Antonio Lieto
Università di Torino, Dipartimento di Informatica and ICAR-CNR, Palermo, Italy
lieto@di.unito.it

Endowing artificial systems with explanatory capacities about the reasons guiding their decisions, represents a crucial challenge and research objective in the current fields of Artificial Intelligence (AI) and Computational Cognitive Science [Langley et al., 2017].
Current mainstream AI systems, in fact, despite the enormous progresses reached in specific tasks, mostly fail to provide a transparent account of the reasons determining their behavior (both in cases of a successful or unsuccessful output). This is due to the fact that the classical problem of opacity in artificial neural networks (ANNs) explodes with the adoption of current Deep Learning techniques [LeCun, Bengio, Hinton, 2015].
In this paper we argue that the explanatory deficit of such techniques represents an important problem, that limits their adoption in the cognitive modelling and computational cognitive science arena.
In particular we will show how the current attempts of providing explanations of the deep nets behaviour (see e.g. [Ritter et al. 2017] are not satisfactory. As a possibile way out to this problem, we present two different research strategies.
The first strategy aims at dealing with the opacity problem by providing a more abstract interpretation of neural mechanisms and representations. This approach is adopted, for example, by the biologically inspired SPAUN architecture [Eliasmith et al., 2012] and by other proposals suggesting, for example, the interpretation of neural networks in terms of the Conceptual Spaces framework [Gärdenfors 2000, Lieto, Chella and Frixione, 2017]. All such proposals presuppose that the neural level of representation can be considered somehow irrelevant for attacking the problem of explanation [Lieto, Lebiere and Oltramari, 2017]. In our opinion, pursuing this research direction can still preserve the use of deep learning techniques in artificial cognitive models provided that novel and additional results in terms of “transparency” are obtained.
The second strategy is somehow at odds with respect to the previous one and tries to address the explanatory issue by avoiding to directly solve the “opacity” problem. In this case, the idea is that one of resorting to pre-compiled plausible explanatory models of the word used in combination with deep-nets (see e.g. [Augello et al. 2017]). We argue that this research agenda, even if does not directly fits the explanatory needs of Computational Cognitive Science, can still be useful to provide results in the area of applied AI aiming at shedding light on the models of interaction between low level and high level tasks (e.g. between perceptual categorization and explanantion) in artificial systems.

References

Agnese Augello, Ignazio Infantino, Antonio Lieto, Umbarto Maniscalco, Giovanni Pilato and Filippo Vella, 2017. Towards A Dual Process Approach to Computational Explanation in Human-Robot Social Interaction. IJCAI 2017 Workshop Proceedings on "Cognition and Artificial Intelligence for Human-Centred Design", Melbourne, 19-25 August 2017.

Peter Gärdenfors, P., 2000. Conceptual spaces: The geometry of thought. MIT press.

Chris Eliasmith, Terry Stewart, X. Choo, X., Bekolay, T., DeWolf, T., Tang, Y., Rasmussen, D., 2012. A large-scale model of the functioning brain. *Science*, 338 (6111), 1202–1205.

Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. *AAAI 2017*.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Antonio Lieto, Christian Lebiere, and Alessandro Oltramari. The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, 2017.

Antonio Lieto, Antonio Chella, and Marcello Frixione. Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures*, 19:1–9, 2017.

Samuel Ritter, Barrett, D. G., A. Santoro, & Botvinick, M. M. (2017). Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. *ICML 2017*, arXiv preprint arXiv:1706.08606.