Explainable-by-design Deep Learning

Plamen Angelov, *Fellow, IEEE,*

I. Extended Abstract

MACHINE and AI justifiably attract the attention and interest not only of the wider scientific community and industry, but also society and policy makers. However, even the most powerful (in terms of accuracy) algorithms such as deep learning (DL) can give a wrong output, which may be fatal. Due to the opaque and cumbersome model structure used by DL, some authors started to talk about a dystopian "black box" society. Despite the success in this area, the way computers learn is still principally different from the way people acquire new knowledge, recognise objects and make decisions. People do not need a huge amount of annotated data. They learn by example, using similarities to previously acquired prototypes, not by using parametric analytical models. Current ML approaches are focused primarily on accuracy and overlook explainability, the semantic meaning of the internal model representation, reasoning and its link with the problem domain. They also overlook the efforts to collect and label training data and rely on assumptions about the data distribution that are often not satisfied. The ability to detect the unseen and unexpected and start learning this new class/es in real time with no or very little supervision is critically important and is something that no currently existing classifier can offer. The challenge is to fill this gap between high level of accuracy and the semantically meaningful solutions. The most efficient algorithms that have fuelled interest towards ML and AI recently are also computationally very hungry – they require specific hardware accelerators such as GPU, huge amounts of labeled data and time. They produce parametrised models with hundreds of millions of coefficients, which are also impossible to interpret or be manipulated by a human. Once trained, such models are inflexible to new knowledge. They cannot dynamically evolve their internal structure to start recognising new classes. They are good only for what they were originally trained for. They also lack robustness, formal guarantees about their behaviour and explanatory and normative transparency. This makes problematic use of such algorithms in high stake complex problems such as aviation, health, bailing from jail, etc. where the clear rationale for a particular decision is very important and the errors are very costly. All these challenges and identified gaps require a dramatic paradigm shift and a radical new approach. In this talk the speaker will present such a new approach towards the next generation of computationally lean ML and AI algorithms that can learn in real-time using normal CPUs on computers, laptops, smartphones or even be implemented on chip that will change dramatically the way these new technologies are being applied. It is explainable-by-design. It focuses on addressing the open research challenge of developing highly efficient, accurate ML algorithms and AI models that are transparent, interpretable, explainable and fair by design. Such systems are able to self-learn lifelong, and continuously improve without the need for complete re-training, can start learning from few training data samples, explore the data space, detect and learn from unseen data patterns, collaborate with humans or other such algorithms seamlessly.

P. Angelov is with LIRA, School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA, UK, p.angelov@lancaster.ac.uk.

References

[1] P. Angelov and E. Soares, "Towards explainable deep neural networks (xdnn)," *Neural Networks*, vol. 130, pp. 185–194, 2020.

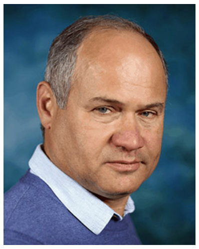
[2] E. A. Soares, P. P. Angelov, B. Costa, M. Castro, S. Nageshrao, and D. Filev, "Explaining deep learning models through rule-based approximation and visualization," *IEEE Transactions on Fuzzy Systems*, 2020.

[3] P. P. Angelov and X. Gu, "Deep rule-based classifier with human-level performance and characteristics," *Information Sciences*, vol. 463–464, pp. 196–213, 2018.

[4] ——, "Toward anthropomorphic machine learning," *Computer*, vol. 51, no. 9, pp. 18–27, 2018.

[5] ——, *Empirical approach to machine learning*. Springer, 2019.

[6] P. P. Angelov, X. Gu, and J. C. Príncipe, "A generalized methodology for data analysis," *IEEE transactions on cybernetics*, vol. 48, no. 10, pp. 2981–2993, 2017.



Plamen Angelov Prof. Angelov (MEng 1989, PhD 1993, DSc 2015) is a Fellow of the IEEE, of the IET and of the HEA. He is Governor of the International Neural Networks Society (INNS) being his Vice President for two tterms till end of 2020. He holds a Personal Chair in Intelligent Systems at Lancaster University, UK. He has authored or co-authored 350+ peer-reviewed publications in leading journals, peer-reviewed conference proceedings, 6 patents, three research monographs (by Wiley, 2012 and Springer, 2002 and 2019) cited over 10800+ times with an h-index of 54. He is the founding Director of LIRA (Lancaster Intelligent, Robotic and Autonomous systems) Research Centre (www.lancaster.ac.uk/lira) which includes over 50 academics across 15 Departments from all Faculties of the University. He has an active research portfolio in the area of computational intelligence and machine learning and internationally recognised results into online and evolving learning and explainable AI. Prof. Angelov leads numerous projects (including several multimillion ones) funded by UK research councils, EU, industry, UK MoD. His research was recognised by 'The Engineer Innovation and Technology 2008 Special Award' and 'For outstanding Services' (2013) by IEEE and INNS. He is also the founding co-Editor-in-Chief of Springer's journal on Evolving Systems and Associate Editor of several leading international scientific journals, including IEEE Transactions on Fuzzy Systems, of the IEEE Transactions on Cybernetics (the IEEE Transactions with the highest impact factor, 11.47), IEEE Transactions on AI as well as of several other journals such as Fuzzy Sets and Systems, Soft Computing, etc. He gave over a dozen plenary and key note talks at high profile conferences. Prof. Angelov was General co-Chair of a number of high profile conferences. He was also a member of International Program Committee of 100+ international conferences (primarily IEEE). More details can be found at www.lancs.ac.uk/staff/angelov.