Self-adaptive Artificial Intelligence

Rogério de Lemos
School of Computing
University of Kent, UK
r.delemos@kent.ac.uk

Marek Grześ
School of Computing
University of Kent, UK
m.grzes@kent.ac.uk

***Abstract*—Machine learning tools, like deep neural networks, are perceived to be black boxes. That is, the only way of changing their internal data models is to retrain these models using different inputs. This is ineffective in dynamic systems that are prone to changes, like concept drift. A new promising solution is transparent artificial intelligence, based on the notions of interpretation and explanation, whose objective is to correlate the internal data models with predictions. The research question being addressed is whether we can have a self-adaptive machine learning system that is able to interpret and explain its data model in order for it to be controlled. In this position paper, we present our initial thoughts whether this can be achieved.**

***Index Terms*—machine learning, artificial intelligence, self-adaptation, AI transparency**

I. Introduction

This position paper is not about how to apply artificial intelligence (AI) to self-adaptation, but how to apply self-adaptation to AI. The former has been a recurrent and viable approach for supporting the self-adaptation of software systems [2]. The latter is based on recent trends in AI, whose goal is to make AI more transparent. That is, the application of techniques for interpreting and explaining what the model has learned [5].

AI techniques, through data models (i.e., models learned from data), are useful to deal with uncertainties when process models are difficult to obtain [7]. However, considering that either the system or its environment may evolve, data models cease to be accurate, thus leading to *concept drift*, i.e., the data model is not updated according to the distribution of the changing input data [8]. When this happens there is the need for the AI technique to adapt its data model, and the challenge is how to maintain an accurate nonlinear data model under concept drift. This can be achieved either by directly manipulating the data model or recomputing it by using new data. In the context of concept drift, the focus of this paper is with the direct manipulation of the data model, hence self-adaptive artificial intelligence (AI) [1].

As noted above, one way for maintaining an accurate data model when facing concept drift is to repeatably update the machine learning models, which requires repeated cycles of training, testing and deployment. Such an approach may not be effective, responsive or robust to dynamic aspects. From the perspective of control systems, this is essentially an open loop control system. It is known that the design of open control systems is only able to cope with a narrow type of uncertainty, which is usually application dependent.

An alternative solution for dealing with concept drift is to employ a self-training ensemble of models (e.g. classifiers [1]). Without the aid of external supervision to update the ML model of a classifier, this solution relies on a feedback loop to control iterative replacement of old classifiers with new ones. Such a solution broadens the type of uncertainties that the ensemble as a whole is able to cope, leading to better performance but with a higher price in resource consumption. Self-adaptive AI can be achieved by using a MAPE-K loop like framework for controlling the structure of the ensemble, i.e., connecting and disconnecting classifiers, based on the performance of the individual classifiers. However, instead of manipulating the parameters of nonlinear ML models, the structure of the ensemble is being manipulated. The focus of this paper is restricted to single model classifiers since transparent AI is essentially related to the manipulation of parameters.

The incorporation of a feedback control loop into most classes of individual ML-based classifiers is challenging because their mappings from inputs to predictions are complex, which is difficult to control at the parametric level. The claim made in this position paper is that a novel promising solution for manipulating ML models is transparent artificial intelligence (AI), which is a technique that allows humans to interpret and explain predictions of ML models. This is achieved by providing evidence on how a hierarchy of model parameters responds to data for the purpose of prediction. From the viewpoint of self-adaptation, instead of relying just on monitoring the inputs and outputs of a machine learning model, such as deep neural network (DNN), the motivation for using transparency is to promote interpretation of these models for allowing their explanation for the purpose of controllability. That is, to identify factors in the nonlinear machine learning models that impact prediction, and how these factors can be adapted for improving resilience against change.

Self-adaptive AI could be applied in a wide range of software engineering contexts, essentially whenever process models are either impossible to be obtain or too costly to be implemented. This would range from specific activities associated with the feedback control loop or the whole control loop. However, a major challenge is how to integrate machine learning techniques with techniques that rely on process mod-

[1] Since artificial intelligence (AI) is broad, we focus on machine learning (ML) in order to explain our initial thoughts regarding self-adaptive AI.

els designed by human experts, and how to obtain the same level of assurances. Regarding the latter, this is one of the aims of transparent AI, that is, to increase the trust on AI techniques in order to make them more resilient regarding their performance, safety, security and accountability [3].

The rest of this paper is organised as follows. In the next section, we provide a brief motivational introduction to AI transparency that forms the basis for promoting self-adaptive AI, which is introduced in the following section. Finally, we provide a brief conclusion, and indicate how transparent AI might be relevant to the provision of assurances of machine learning models.

II. AI Transparency

Neural networks have traditionally been seen as 'black box' models because, in contrast to, e.g., decision trees, it is not easy to interpret them or explain their decisions using the values of individual parameters. However, the fact that deep neural networks (DNN) were shown to work surprisingly well in many real-life applications [4], [6], is a natural reason to look for techniques that can increase the transparency of their models. Although the literature offers different interpretations of AI transparency, in the following, we describe it in terms of two key concepts: interpretability and explainability.

In *interpretability*, one is interested in making general observations about the model. The existing methods for interpretability of neural networks are able to detect some hidden features in the data to uncover what kind of knowledge is learned by the machine learning model used to make predictions. One of the goals of interpretability is thus to see how the regularities in the data are captured by the models. Obviously, if a feedback control loop is incorporated into a DNN for improving its performance under uncertainty, interpretability can be seen as a major component, especially when trust has to be considered.

Explainability focuses on particular data examples, and it tries to explain decisions made by the models on those specific examples. For instance, when a particular data example is classified as 'a malicious attack' in cybersecurity, we would like to know why the algorithm decides so. Again, having answers to explainability and incorporating these into a feedback control loop is a clear prerequisite to robust AI models.

III. Self-adaptation and AI Transparency

We believe that techniques for transparency in AI can be used to make the machine learning models respond to changes in the environment or in the input data without recomputing those models from scratch. For example, information about interpretability of a neural network trained for a particular problem will allow us to create effectors or turning knobs for direct intervention into the behaviour of the models. This is essentially the difference between the typical usage of AI, and what we envisage in terms of self-adaptive AI.

Let's consider a scenario in which the main goal is to adjust the machine learning model by minimising the probability of error given specific changes in the data, e.g., concept drift. It should be noted that the solution in this case cannot simply rely in recomputing the model from new data. Instead, we argue that such an optimisation problem can be defined and solved using the results of transparent AI. For example, when the data is drifting in a specific direction, interpretability and explainability can tell us which nodes in a neural network should be inhibited or excited to minimise the probability of error with respect to the regularities in the original data.

Having appropriate definitions of trust and the regularities that should be preserved by the model (e.g. which properties of the input/output mappings should be respected by the model at all times), specific objective functions for mathematical optimisation can be defined to close this human-free feedback control loop. In complex scenarios, continuous optimisation may be required in every iteration of the feedback control loop. This is how the trust-related requirements can be translated into self-adaptability without recomputing the models, bringing potentially more modelling capacity than straightforward, as we could argue, use of new data and purely data-driven and blind adaptation of the models.

IV. Conclusions

This paper has presented our initial thoughts of using transparent artificial intelligence (AI) as a basis to support self-adaptive AI. The claim being made is that the interpretation and explanation of a machine learning (ML) model, based on its inputs and outputs, would allow the model to be more responsive to changes via direct manipulation of the model parameters.

The end goal of our research is related to the resilience of ML techniques. Similar to the identification of code vulnerabilities, the notion of AI transparency is fundamental for identifying potential vulnerabilities since it provides the ability to observe and reason about their decision making, which is fundamental when evaluating their resilience.

References

[1] P. Conca, J. Timmis, R. de Lemos, S. Forrest, and H. McCracken. An adaptive classification framework for unsupervised model updating in nonstationary environments. In P. Pardalos, M. Pavone, G. M. Farinella, and V. Cutello, editors, *Machine Learning, Optimization, and Big Data*, pages 171–184, Cham, 2015. Springer International Publishing.

[2] J. Dowling and V. Cahill. Self-managed decentralised systems using k-components and collaborative reinforcement learning. In *Proceedings of the 1st ACM SIGSOFT Workshop on Self-managed Systems*, WOSS '04, pages 39–43, New York, NY, USA, 2004. ACM.

[3] M. Hind, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, and K. R. Varshney. Increasing trust in AI services through supplier's declarations of conformity. *CoRR*, abs/1808.07261, 2018.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[5] G. Montavon, W. Samek, and K.-R. Mller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1 – 15, 2018.

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[7] H. A. Simon. *The Sciences of the Artificial*. The MIT Press, 1996.

[8] A. Tsymbal. The problem of concept drift: Definitions and related work. Technical report, Trinity College Dublin, Ireland, 2004.