IAF SYMPOSIUM ON INTEGRATED APPLICATIONS (B5)
Tools and Technology in Support of Integrated Applications (1)

Author: Dr. Guglielmo Faggioli
AIKO S.r.l., Italy, guglielmo@aikospace.com

Mr. Mattia Varile
AIKO S.r.l., Italy, mattia@aikospace.com

ONBOARD EXPLAINABLE ARTIFICIAL INTELLIGENCE

Abstract

Deep learning (DL) is perhaps one of the most promising technology in the development of forthcoming autonomous space systems. The improvements made in sectors like image-recognition, natural language processing, or in areas like biology, suggest that the space sector should seriously consider the implementation of DL systems in future projects. The main challenge will concern achieving reliable and efficient onboard DL-based space systems.

DL models are seen by many as black boxes, in the sense that a complete view of information learned during the training phase is missing. This lack of knowledge represents a big issue in order to trust their predictions.

Due to this nature, it is difficult to understand which aspects and features of the input data drive the decisions of the network. Moreover, decisions are driven by combinations of data features, and understanding the decision-making process of neural networks is challenging. These are big issues to deal with in order to achieve high confidence in AI predictions. One step closer to the solution of these issues is using new validation methods, which do not solve directly our lack of knowledge but which provide tools capable of giving useful interpretations about how a model is working. Explainable Artificial Intelligence (XAI) framework provides interesting tools in this direction. This work focuses on exploiting the theory of Shapley's value to contribute in the direction of XAI. The proposed approach is a concept of game theory, that tries to understand what is the contribution of each agent in a collaborative system. The Machine Learning community focused on validation is already applying widely this method to DL techniques, for example in the biomedical sector, and they are obtaining reliable interpretations about what their models are actually learning and how they do so. In this framework, Neural Networks (NN) are seen as cooperative systems while their input features are seen as agents that cooperate in order to make the best prediction. Understanding the contribution of each feature to the overall network output provides a better view of the information learned by the network. Shapley's value quantifies this contribution by averaging among all the possible ordered subset of the features and of the related contribution.

This work shows the potential of using explainability tools on NN, highlighting how this approach makes the results human-readable and understandable. This work aims to extend the background on XAI for space applications, promoting their use in space missions.