XAI in time series and multimodal learning

Valerio Guarrasi†*, Giulio Iannello†, Rosa Sicilia†, Paolo Soda†, Lorenzo Tronchin†

†Unit of Computer Systems and Bioinformatics, Universita' Campus Bio-Medico di Roma,
*Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Italy

{valerio.guarrasi}@uniroma1.it
{g.iannello, r.sicilia,p.soda,l.tronchin}@unicampus.it

Abstract

The pervasiveness of artificial intelligence in our daily lives has raised the need to understand and trust the outputs of the learning models, especially when involved in decision processes. As a result, eXplainable Artificial Intelligence has captured more and more interest in the scientific community, providing insights into the behaviour of these systems, ensuring algorithms fairness, transparency and trustworthiness. In this contribution we overview our work on the explainability of deep learning models applied to time series and multimodal data.

1 Introduction

Artificial Intelligence (AI) has proven to effectively support the decision process [Fu, 2011] and in particular deep learning techniques have achieved state-of-the-art performance [Fawaz *et al.*, 2019]. Despite the impressive prediction accuracy attained in several applications, there is still the need to *explain* the decisions of the learning models proposed. As a result, eXplainable Artificial Intelligence (XAI) has captured more and more interest in the scientific community [Guidotti *et al.*, 2018; Du *et al.*, 2019; Gilpin *et al.*, 2018] since the complex nature of the models, such as deep neural networks, makes it impossible for the user to understand and validate the decision process. XAI aims to provide an insight into the behaviour and processes of these systems, ensuring algorithms fairness, identifying any potential bias in the training data and allowing complex AI models to be more transparent and understandable to humans [Gilpin *et al.*, 2018].

Among a growing body of literature about XAI, in our laboratory we are directing our efforts toward two issues. The first concerns the explainability of Deep Learning (DL) models working on Time Series (TS) data. Indeed, the rising capabilities of storing and registering data have increased the number of temporal datasets, boosting the attention on TS classification models, raising the need to explain their decision. In this context, we present the application and evaluation of three XAI methods in a real-world multimodal task of anomaly detection on telematics data. We dealt with the challenge of explaining multivariate TS (MTS) and showing how to adapt different methodologies, originally designed for images, to this domain.

The second concerns the explainability of Multimodal Deep Learning (MDL) models, a topic at its infancy in the current literature. With the recent availability of a larger data repository, we expect to have the possibility to explore more complex deep architectures, studying how to learn shared representation and how to combine the unimodal networks. Nevertheless, more complex models exacerbate the problem of understanding what the predictions rely on, and also which modalities and features hold an important role [Joshi *et al.*, 2021].

2 XAI in time series

We take into account the challenge of explaining a real-world multimodal task of anomaly detection on telematics data from vehicles' black-box, where the available modalities are acceleration MTS and velocity univariate TS (UTS). Moreover, in this application there is also a supervised classifier trained to recognise if a crash event occurred or not. The peculiarity of MTS is that they are characterised by complex non-linear temporal dependencies between their attributes, i.e. the points of each UTS are connected with the other sequences via the time dimension. This key issue makes the development of an XAI approach for MTS-based anomaly detection rather challenging. The analysis of the literature shows that the study on XAI for MTS is limited: indeed, more efforts have been directed towards data diverse from TS, i.e. images and tabular data, not taking into account the complex relationship retained in multivariate time series. Thus as first contribution, we studied how to employ three XAI methods suited to explain models working on images and how to extend their application to a multimodal architecture working on telematics MTS acquired by car's black-boxes.

A further issue that we tackled in this work is evaluating the provided explanation for the MTS. Indeed, as highlighted in [Doshi-Velez e Kim, 2017], different kinds of explanations produced after interpreting machine learning models may not be equally explainable, so a pressing need is emerging for quantifying the quality of the explanations produced, a topic that is only in its infancy in the current literature [Doshi-Velez e Kim, 2017; Schlegel *et al.*, 2019].

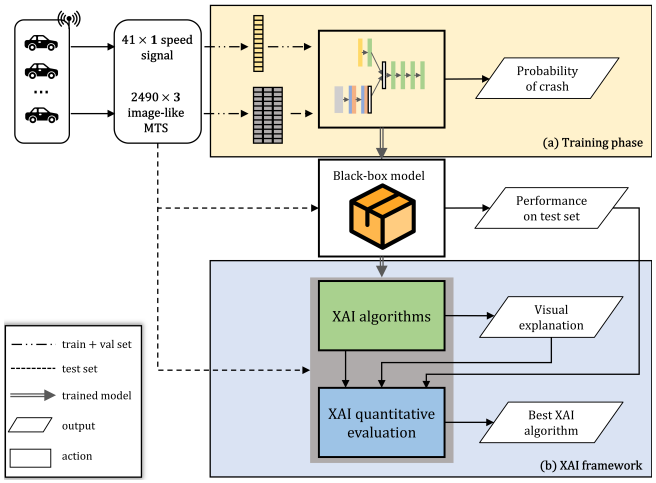

Figure 1: An overview of our approach.

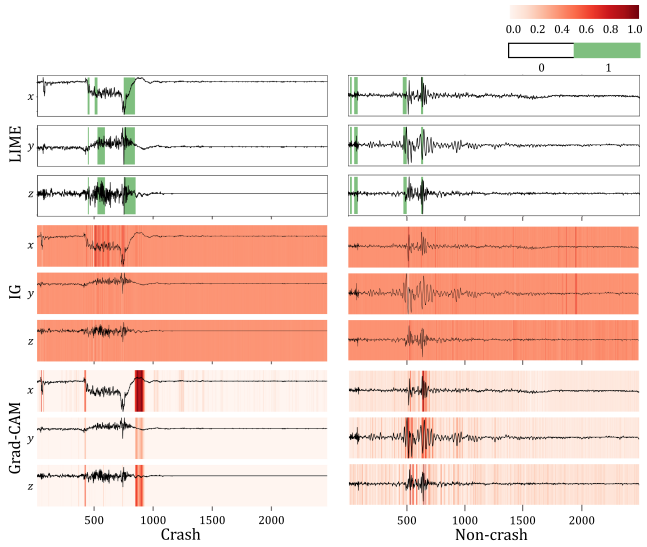

Figure 2: Explanation over the acceleration signal for a crash and non-crash event using LIME, IG and Grad-CAM.

2.1 Methods

Figure 1 shows an overview of the proposed approach that is able to explain a multimodal anomaly detection architecture identifying car crash events from telematics data of vehicles. Hence, we first train the black-box multimodal model for classification (panel (a) in Figure 1), which consists of a Convolutional Neural Network (CNN) that can learn local temporal-spatial patterns that are intuitive to visualise for the end-user. Then, the trained CNN, the test samples and the performance on the test set are used as inputs of the XAI framework to generate visual explanations of the decision provided by the model and to evaluate the quality of such explanation (panel (b) in Figure 1).

To consider both temporal and spatial relationships between each dimension of a MTS, we represent each sample as a 2D image where each pixel does not retain only visual features, such as the shape, the intensity and the texture, but also temporal features across each UTS included in the MTS. In this way we gain advantage of analysing the MTS as a whole, leveraging the rich literature about the explainability on images, and of maintaining the flexibility to search for the best architecture and performance without the constrain of designing a model for explicitly achieving explainability.

Therefore, given the mulimodal CNN initially designed, we then investigated its explainability by employing three well-known XAI algorithms originally designed for images, providing a *Saliency Map (SM)*, which is an efficient way of pointing out what causes a certain outcome. The three approaches are: (i) an *agnostic solution* that is generalisable by definition to any model and returns a comprehensible local predictor, namely LIME [Ribeiro *et al.*, 2016]; (ii) a *model-specific solution*, namely Grad-CAM [Selvaraju *et al.*, 2017], designed explicitly for convolutional architectures that consider the CNN activation on the specific input sample. (iii) a *model-inspection approach*, namely IG [Sundararajan *et al.*, 2017], a method that derives explanation by examining the internal model behaviour when modifying the input sample. Indeed, as LIME aims to approximate the decision surface of a complex model using an interpretable one, the explanations from the surrogate models cannot be perfectly faithful with respect to the original model [Rudin, 2019]. For this reason we also customized an XAI method suited for CNNs and another more general suited for deep architectures, namely Grad-CAM and IG, respectively. We therefore present how to customise and employ them to deal with a multimodal architecture working on multivariate telematic data.

Regarding the evaluation procedure, there is no consensus in the literature on methods to assess explainability [Molnar, 2020]. In this respect, we customised the two strategies presented in [Schlegel *et al.*, 2019] for UTS, by adapting them to the MTS domain. We exploited a perturbation-based XAI evaluation, measuring the performance drop D of the anomaly detection system respectively disrupting the time points identified by the XAI method as relevant (D_{XAI}), and random regions of the signal (D_{random}). The assessment is based on the assumption that if relevant/random feature (time points) gets changed, the model's performance should decrease/stagnate. Finally, as we aim to conduct an exhaustive evaluation, we performed the evaluation procedure considering all test set samples in an hold-out setup.

2.2 Main results

For the anomaly detection task, we ran a hold-out obtaining on the test set recall and precision values equal to 70% and 63% respectively.

In Figure 2, we report the saliency maps attained by the three XAI approaches on a crash sample (left column in Figure 2) and on a non-crash sample (right column in Figure 2). Regarding the crash sample explainability, both IG and LIME show as significant for the prediction the x-axis time instants of the peak and immediately before it. Instead, Grad-CAM highlights as important the signal trend following the event: after the shift in x acceleration values, there is a transition period that leads the car's cinematic to a new stable state. IG and Grad-CAM saliency maps on y and z axes do not show

Table 1: Each box contains the performance drop per XAI method and perturbation type. The colour code is based on the difference between D_{XAI} and D_{random}: green if $D_{XAI} > D_{random}$, red otherwise.

Perturbation type	Drop LIME	Drop IG	Drop Grad-CAM	Drop Random
Zero	54.3%	58.2%	20.1%	0.7%
Swap	13.8%	15.2%	6.5%	14.3%
Mean	10.0%	5.5%	2.7%	3.6%

activation, outlining two main findings: (i) x component is the most significant UTS for predicting the event as a crash; (ii) y and z are de-correlated with respect to x-axis. This finding is feasible as, in general, the strongest fluctuations of the acceleration for a crash event occurs along the direction of travel of the vehicle, i.e. the x-axis, proving the reliability of the explanations. As a result, IG and Grad-CAM are able to exploit the cross-correlation in UTS learned from the CNN, namely defining which UTS is most valuable for the prediction. LIME explanation does not present this evidence since it uses a surrogate model to approximate the CNN, being an agnostic method, so it does not directly inspect the convolutional architecture's inner workings.

For the non-crash instance, we can observe a less clear focus of the black-box model on a precise pattern. IG and Grad-CAM heatmaps exhibit lower and more distributed intensity signals than the crash samples saliency maps: long-term time relationships are detected as important for the non-crash prediction. Indeed a bumpy road is characterised by less intense and longer disturbances with respect to those that emerge in a crash event.

Overall, the explanations reasonably find the crucial features used from the model to perform the anomaly detection task, casting light on the black-box's decision process.

To provide an exhaustive comparison between the three methods, Table1 shows the results emerging from XAI quantitative evaluation. Note that each table box is highlighted in green if $D_{XAI} > D_{random}$ is satisfied, in red otherwise. From the results, IG-based perturbations account for the most significant drop in performance and it is the only method that always exceeds the random drop. Hence, the quantitative analysis suggests that IG is the more informative explainability method for the anomaly detection task as it is able to detect the time points valuable for the CNN to perform the prediction. In contrast, Grad-CAM results in the least reliable algorithm from the quantitative evaluation as two out of three times it is overtaken by random drop. Moreover, we find the temporal trend to have a minor influence for the CNN to perform the classification task as we observe from the drops in the values of *swap* and *mean* metrics compared to the *zero* one.

2.3 Challenges and perspectives

This work represents a first attempt to tailor XAI algorithms to the multimodal nature of the data, suggesting further research in this field. As a first direction, we will investigate XAI methods able to provide a more human-interpretable representation, since the saliency maps provided by all the XAI methods are hard to be interpreted by users. The second direction will focus on developing a multimodal XAI method able to explain both signals available in the telematics data at hand (i.e. acceleration and speed).

2.4 Papers and available resources

This work was published in [Tronchin *et al.*, 2021].

3 XAI in multimodal learning: challenges and perspectives

Multimodal Deep Learning (MDL) studies how deep neural networks can learn shared representations between different modalities that, used together, may offer deeper insights into the data. It investigates when to fuse the different modalities and how to obtain more powerful data representations. Linking information coming from various modalities can leverage the understanding of a field of study. These considerations are of particular relevance for the field of bio-medicine, where MDL has proven to be useful [Ramachandram e Taylor, 2017].

Among the different challenges of MDL, we focus here on supervised multimodal fusion applied to early identify patients at risk of the severe outcome, like intensive care or death, among those affected by SARS-CoV-2, and using chest X-ray (CXR) scans and clinical data. Instead of using manually designed or handcrafted modality-specific features, via DL we can automatically learn and extract an embedded representation for each modality, which is then fused with the others in an end-to-end training process that exploits the loss backpropagation.

It is well-known that the major disadvantage of neural networks is their lack of interpretability. It is hard to understand what the predictions rely on, and which modalities and features hold an important role. For this reason the number of XAI algorithms has grown in the last years. In general, an XAI algorithm is one that produces information to make a model's functioning clear or easy to understand. In spite of the importance of MDL, to the best of our knowledge, no work has studied the XAI methods for fused modalities, particularly in the medical field.

Hence, we are developing an XAI method that can explain model working on the aforementioned biomedical multimodal data. The key idea is to develop an approach working on an embedded representation of the data, created by exploiting a Convolutional Autoencoder (CAE), which is connected in an end-to-end manner also to a multi-layer perceptron which works on the classification task for the prognosis of the severity of the COVID-19 virus. Therefore, the main goal is to construct an optimal embedded space that suits best for the modalities (images and clinical) reconstruction and for the classification task (mild vs severe COVID-19 cases). Exploiting the nature of the model's architecture, we are developing multiple novel XAI algorithms which work on variations on the embedded feature vector to help us in understanding how the classifications are performed by the MLP, via the echoed perturbations on the CAE reconstructions. With clear visual representations of the multimodal feature importance, we would improve trust and transparency by reducing

the model's opacity which makes it difficult for doctors and regulators to trust the MDL models.

Acknowledgements

The work detailed in section 2 was realized in cooperation with Consorzio ELIS (Consel) and Generali Italia S.p.A.Part of this work is also supported by Fondazione TIM with a PhD grant.

References

[Doshi-Velez e Kim, 2017] Finale Doshi-Velez e Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Du *et al.*, 2019] Mengnan Du, Ninghao Liu, e Xia Hu. Techniques for Interpretable Machine Learning. *Communications of the ACM*, 63(1):68–77, 2019.

[Fawaz *et al.*, 2019] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, e Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

[Fu, 2011] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[Gilpin *et al.*, 2018] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, e Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, e Dino Pedreschi. A survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[Joshi *et al.*, 2021] Gargi Joshi, Rahee Walambe, e Ketan Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 2021.

[Molnar, 2020] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[Ramachandram e Taylor, 2017] Dhanesh Ramachandram e Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, e Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[Schlegel *et al.*, 2019] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, e Daniel A Keim. Towards a rigorous evaluation of xai methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4197–4201. IEEE, 2019.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, e Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, e Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[Tronchin *et al.*, 2021] Lorenzo Tronchin, Rosa Sicilia, Ermanno Cordelli, Lorenzo Ricciardi Celsi, Daniele Maccagnola, Massimo Natale, e Paolo Soda. Explainable AI for Car Crash detection using Multivariate Time Series. In *Proceedings of 20th IEEE International Conference on Cognitive Informatics and Cognitive Computing*. In Press, 2021.