Proceedings of the DataNinja sAIOnARA 2024 Conference 78-80 DOI: 10.11576/dataninja-1178

Concept extraction for time series with ECLAD

Antonia Holzapfel ANTONIA.HOLZAPFEL@DSME.RWTH-AACHEN.DE
Andres Felipe Posada-Moreno ANDRES.POSADA@DSME.RWTH-AACHEN.DE
Sebastian Trimpe TRIMPE@DSME.RWTH-AACHEN.DE
Institute for Data Science in Mechanical Engineering (DSME), RWTH Aachen University, Aachen, Germany

Abstract

Concept Extraction (CE) methods are being increasingly used in the image domain for explaining deep learning models, which are not inherently interpretable. However, there have not been transfer studies yet for their usage in the time series domain. The purpose of this work is to explore the use of CE methods in time series. We propose to modify the ECLAD algorithm for this domain by changing the latent space representation used to extract concepts. This method is then tested on an InceptionTime model trained on the Gunpoint dataset. Preliminary results show that we can successfully extract concepts from time series models on datasets with local features and provide conceptual explanations that effectively explain how the model works.

Keywords: Explainable Artificial Intelligence, Concept Extraction

1. Introduction

Over the last years, the field of explainability (XAI) has emerged to explain what happens under the lid of black box models. Particularly, Concept Extraction (CE) methods have been developed to produce global explanations of models in the form of human-understandable "concepts". CE methods are particularly interesting because they provide a direct link between the inner representations of the model and human understandable visualizations.

While several CE methods are available for image data [1–3], no CE methods have been developed or transferred to time series. On time series, global methods like (global) Grad-CAM [4] have been used. However, they fail to provide global explanations beyond the aggregation of local ones.

In this work, we explore the use of CE methods in time series. This is, if in the time series domain, CE methods can extract patterns that are distinguishable from each other and the model is sensitive to. These patterns may be related to meaningful features of the dataset. Specifically, we modify the CE method ECLAD [1] by adapting the latent representations it uses, in coherence to time series classification models. We test the method using an InceptionTime model trained on the Gunpoint dataset. Preliminary results show that the adapted ECLAD provides meaningful model explanations.

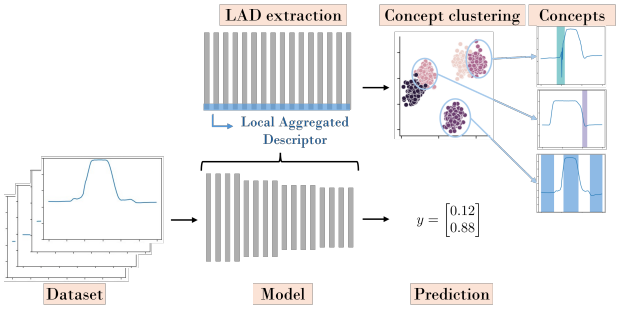

Figure 1: Overview of ECLAD on time series

2. CE for Time Series

In this work, we use ECLAD as a basis for CE in time series classification models. ECLAD is an algorithm that provides explanations based on three key steps: The encoding of the latent space of neural networks, the mining of patterns and the assessment of how relevant they are for the predictions of the model. We modify the first and second steps to be compatible with time series data, and explore which latent space representations make sense in this domain.

In the first step, ECLAD uses the notion of **Local Aggregated Descriptors (LADs)**, which are descriptors of how models encode a region at different levels of abstraction. These are obtained by aggregating the activation maps of multiple layers L of a CNN model. In the case of time series data, the inputs of the model have the shape $x_i \in \mathbb{R}^{w \times d}$, where

w, and d are the length and channels of an input. Since the latent representations of CNN based models inherit the dimensionality of the input vectors, our proposed LADs are timestep-wise descriptors denoted as $d_{x_{ts}} \in \mathbb{R}^{w \times 1}$.

After extracting the LADS of several inputs, a set $\Gamma = \{\gamma_{c_1}, \ldots, \gamma_{c_{n_c}}\}$ of centroids $\gamma_{c_j} \in \mathbb{R}^{1 \times c^*}$ defining the **concepts** can be obtained by applying a **mini-batch k-means algorithm**, where c^* is the sum of the number of units in the layers in L. The centroids represent similarly encoded time subsequences. For a human-understandable visualization, the concepts c_j can then be located in an input x_i by creating a **mask** $m^{c_j}_{x_{ts}} \in \mathbb{R}^{w \times 1}$ that analyzes the LAD $d_{x_i,(b)}$ at timestep b of the input, and assessing whether it belongs to a cluster γ_{c_j}.

For the last step of ECLAD, the **importance score** of a concept is determined, which quantifies the relevance of its related visual cues towards the prediction of the analyzed model. The metric used in [5] computes how sensitive a model is with respect to the regions containing each concept. For this, said metric computes the pixel-wise sensitivity $r^{c_j}_{x_i}$ of the regions of each image using a concept,

$$r^{c_j}_{x_i} = ||\nabla_x g(f(x_i)) \odot m^{c_j}_{x_i}||_1, \tag{1}$$

where $\odot$ denotes the element-wise product between matrices, $g(y) = ||y \cdot \mathbf{1}^T - \mathbf{1} \cdot y^T||_2$ is a wrapper of the model f, $y \in \mathbb{R}^{n_k}$ is the output of f, n_k is the number of classes and $\mathbf{1}$ is the vector of ones the of the same size as y. The method then proceeds to aggregate $r^{c_j}_{x_i}$ over the images in the dataset and scale the mean relevance of each concept to obtain the final importance scores I_{c_j}.

With the proposed modifications, we can transfer ECLAD to analyse time series classification models. This allows for the extraction of patterns that are meaningful for models and can be represented in human-understandable visualizations. These patterns also respect the equivariance properties of the models.

3. Preliminary Results

To validate ECLAD for time series, we trained an InceptionTime network with 20 Inception blocks on the GunPoint dataset. The dataset contains two classes ("gun" and "no gun"), which can be distinguished by the oscillations before and after the main peak. At convergence, the model obtained an accuracy of 0.975 on the validation dataset.

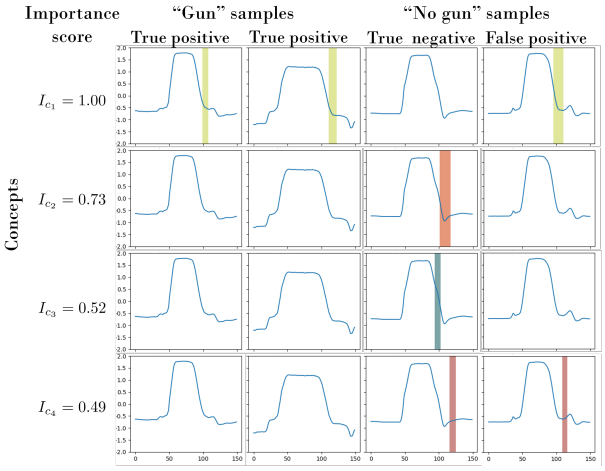

Figure 2: Concept extraction results on InceptionTime trained on the GunPoint dataset.

After training, we extracted ten concepts from the model using our modified ECLAD and visually inspected them to assess how they relate to known features of the dataset and how differentiating they are. The four most important concepts obtained are shown in Figure 2. The first identified concept represents the smooth concave curvature directly after the main peak. The structure is identified consistently, regardless of its exact position, and it strongly contributes to the positive classification. This indicates the presence of concepts within time series classification models.

The second to fourth concepts are also consistently identified and associated to instances of a specific class. These are differentiating features, but the model considers them less important in comparison.

Observing the concepts, it is possible to identify the features that the model is using and those that are causing wrong predictions. For example, for the false positive we can determine that the fourth concept was correctly identified while the lack of the second and third, as well as the presence of the first are confounding factors. These kinds of insight are useful for a user that wants to understand which biases can be present in the model or dataset.

In this extended abstract we focused on assessing the applicability of CE methods to time series. Preliminary results indicate the presence of meaningful features used by the model that our adapted ECLAD could extract.

Acknowledgments

This work is partially funded by the Deutsche Forschungsgemeinschaft (DFG) within the Priority Program SPP 2422 (TR 1433/3-1).

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy -– EXC-2023 Internet of Production -– 390621612.

References

[1] Andrés Felipe Posada-Moreno, Nikita Surya, and Sebastian Trimpe. ECLAD: Extracting Concepts with Local Aggregated Descriptors. *Pattern Recognition*, 147:110146, 2023. URL https://www.sciencedirect.com/science/article/pii/S0031320323008439.

[2] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

[3] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems (NeurIPS)*, 33:20554–20565, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ecb287ff763c169694f682af52c1f309-Abstract.html.

[4] Corne Van Zyl, Xianming Ye, and Raj Naidoo. Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of grad-cam and shap. *Applied Energy*, 353:122079, 2024. URL https://www.sciencedirect.com/science/article/pii/S0306261923014435.

[5] Andrés Felipe Posada-Moreno, Kai Müller, Florian Brillowski, Friedrich Solowjow, Thomas Gries, and Sebastian Trimpe. Scalable Concept Extraction in Industry 4.0. In *Explainable Artificial Intelligence*, pages 512–535. Springer Nature Switzerland, 2023. URL https://link.springer.com/chapter/10.1007/978-3-031-44070-0_26.