Explainable AI for Audio via Virtual Inspection Layers

Johanna Vielhaben
Fraunhofer Heinrich Hertz Institute
johanna.vielhaben@hhi.fraunhofer.de

Sebastian Lapuschkin
Fraunhofer Heinrich Hertz Institute
sebastian.lapuschkin@hhi.fraunhofer.de

Grégoire Montavon
Technical University Berlin
Free University of Berlin
gregoire.montavon@fu-berlin.de

Wojciech Samek
Fraunhofer Heinrich Hertz Institute
Technical University Berlin
BIFOLD – Berlin Institute for the Foundations of Learning and Data
wojciech.samek@hhi.fraunhofer.de

Abstract

The field of eXplainable Artificial Intelligence (XAI) has made significant advancements in recent years. However, most progress has focused on computer vision and natural language processing. There has been limited research on XAI specifically for audio or other time series data, where the input itself is often hard to interpret. In this study, we introduce a *virtual inspection layer* that transforms time series data into an interpretable representation and enables the use of local XAI methods to attribute relevance to this representation.

1 Introduction

Research in Explainable Artificial Intelligence (XAI) has predominantly concentrated on Computer Vision (CV) and Natural Language Processing (NLP), leaving other data domains like time series, e.g. audio samples, under-explored. Often, local XAI methods are used to explain model output by providing feature-wise attribution scores. These scores are presented as heatmaps overlaying the sample to highlight relevant features. In this way, the explanation relies on the interpretability of the input features, which is challenged for time series like the waveform of an audio sample. To address this problem, we propose to incorporate *Virtual Inspection Layers*, leveraging an invertible transformation to an interpretable domain, to enhance the interpretability of explanations [13]. Specializing to time series, we consider the Discrete Fourier Transform (DFT) which maps the data to a representation, where meaningful features related to frequencies and their magnitudes can be extracted. These are more easily interpretable, so can use the DFT as a Virtual Inspection Layer and apply existing local attribution methods to provide interpretable explanations in the frequency domain.

2 Related work

The field of XAI, primarily focused on CV and NLP, has seen a surge in research efforts dedicated to time series analysis, see [12, 7] for a systematic review. In principle, XAI methods from CV and NLP can be easily applied to time series classifiers using similar architectures such as CNNs or RNNs

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

[5]. Notably, Layerwise Relevance Propagation (LRP) [2] has been used for explaining classifiers for audio [3], and other time series data, such as ECG analysis [9], and EEG analysis [10]. Further, Gradient × Input (G×I) has been applied to ECG classifers [8].However, these methods provide attribution scores for individual time points, limiting their interpretability for time series [12]. Our proposed virtual inspection layer allows for assessing relevance scores in an interpretable domain, such as the frequency or time-frequency domain, enabling explanations using feature-wise post-hoc XAI methods for classifiers operating in the time domain.

3 LRP propagates relevance to interpretable representation via Virtual Inspection layer

Virtual inspection layer We view a neural network as a composition of layer- or block-wise functions,

$$f(x) = f_L \circ \cdots \circ f_1(x)\,.$$

We can use a a local XAI method to quantify the relevance $R_f(x_i)$ of each feature i in x towards $y = f(x)$. While the datapoint x representation is not interpretable for humans, we assume there is an invertible transformation $\mathcal{T}(x) = \tilde{x}$, that renders x interpretable. Now, we can now quantify the relevance of $\tilde{x}_i$ by by attaching the inverse transform to the network,

$$f(x) = f_L \circ \cdots \circ f_1 \circ \mathcal{T}^{-1} \circ \underbrace{\mathcal{T}(x)}_{\tilde{x}}, \tag{1}$$

and compute the relevance scores $R'_f(\tilde{x}_i)$ for the interpretable representation. This does not necessiate re-training the model on $\{\tilde{x}\}$. In general, an interpretable-representation-inducing bottleneck can be inserted at any layer of the network.

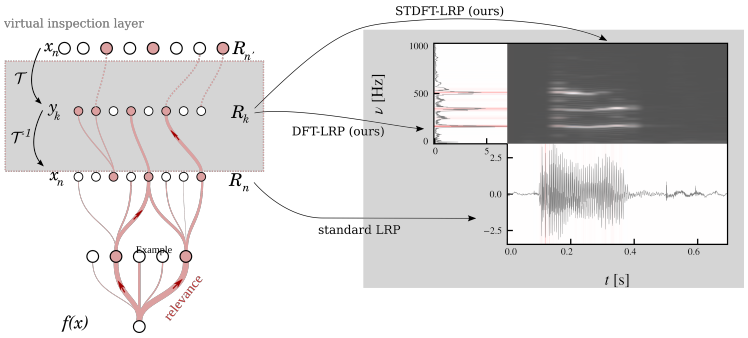

Figure 1: Schematic overview of virtual inspection layers and (ST)DFT-LRP.

Relevance Propagation for the Discrete Fourier Transformation For a neural network trained in time domain, we can use LRP to quantify the relevance R_n of each time step x_n towards the prediction. Here, we lay out how to propagate relevance one step further into the frequency domain. A signal in time domain x_n, $n = 0, ..., N-1$ is connected to its representation in frequency domain $y_k \in \mathcal{C}$, $k = 0, ..., N-1$, via the DFT. The DFT and its inverse are linear transformations with complex weights,

$$y_k = \mathrm{DFT}(\{x_n\}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n \left[\cos(\frac{2\pi kn}{N}) - i\sin(\frac{2\pi kn}{N})\right] \tag{2}$$

$$x_n = \mathrm{DFT}^{-1}(\{y_k\}) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} y_k \left[\cos(\frac{2\pi kn}{N}) + i\sin(\frac{2\pi kn}{N})\right]\,. \tag{3}$$

Now, we attach a virtual inspection layer performing the inverse DFT in Eq. (3) to the model, before the first layer f_1 that operates on the signal in time domain, see Figure 1. For real valued signals $x_n \in \mathcal{R}$ we can express the inverse DFT as,

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \mathrm{Re}(y_k)\cos\Big(\frac{2\pi kn}{N}\Big) - \mathrm{Im}(y_k)\sin\Big(\frac{2\pi kn}{N}\Big)\,. \tag{4}$$

We assume that relevance values R_n are available and that they are of form $R_n = x_n c_n$ (a property ensured by most LRP rules, in particular, LRP-$0/\epsilon/\gamma$). Now, we wish to propagate relevance scores R_n for x_n onto y_k. For this we employ the generic LRP-rule for propagating relevance scores R_j at layer j onto neurons of the lower layer i,

$$R_i = \sum_j \frac{z_{i,j}}{\sum_i z_{i,j}} R_j \tag{5}$$

for which we have to quantify the contribution $z_{k,\mathrm{Re},n}$, $z_{k,\mathrm{Im},n}$ of neuron $\mathrm{Re}(y_k)$, $\mathrm{Im}(y_k)$ to x_n. Simply, the inverse DFT in Equation (4) is a homogeneous linear model, i.e. of type $f(x) = wx$, and we can defined the contribution as the value of the neuron itself times the weight,

$$z_{k,\mathrm{Re},n} = \mathrm{Re}(y_k)\cos(\frac{2\pi kn}{N}), \;\; z_{k,\mathrm{Im},n} = -\,\mathrm{Im}(y_k)\sin(\frac{2\pi kn}{N})\,. \tag{6}$$

Now, we apply Eq. (5) to aggregate the contributions of each neuron $R_{k,\mathrm{Re}}$, $R_{k,\mathrm{Im}}$ towards the model output and find,

$$R_{k,\mathrm{Re}} = \mathrm{Re}(y_k)\sum_n \cos(\frac{2\pi kn}{N})\frac{R_n}{x_n}, \;\; R_{k,\mathrm{Im}} = -\,\mathrm{Im}(y_k)\sum_n \sin(\frac{2\pi kn}{N})\frac{R_n}{x_n}\,. \tag{7}$$

Here, we assume $R_k = 0$ if $x_k = 0$ and define $0/0 = 0$. For numerical stability, we add a small-valued constant ϵ to the denominator. We leverage the addivity of LRP attributions, and define $R_k = R_{k,\mathrm{Re}} + R_{k,\mathrm{Im}}$, which can be abbreviated by

$$R_k = r_k \sum_n \cos(\frac{2\pi kn}{N} + \varphi_k)\frac{R_n}{x_n}\,. \tag{8}$$

where r_k, φ_k denote amplitude and phase of y_k.

To perform a short–time DFT (STDFT) one computes the DFT of windowed signals,

$$v_{m,k} = \mathrm{DFT}(\underbrace{x_n \cdot w_m(n)}_{sm,n})\,, \tag{9}$$

to map the signal in time to time-frequency domain. Analogously to the DFT, we can propagate relevance through the inverse short-time STDFT

$$\tilde{x}_n = \frac{\sum_m \mathrm{DFT}^{-1}(\{v_{m,k}\})}{\sum_m w_m(n)}\,. \tag{10}$$

by

$$R_{m,k} = r_{m,k}\sum_n \cos(\frac{2\pi kn}{N} + \varphi_{m,k}) \cdot W_n^{-1}\frac{R_n}{x_n}\,, \tag{11}$$

where $W_n = \sum_m w_m(n)$.

4 Results

For real-world audio data, we 1) test which feature domain – time, frequency or time-frequency – is the most *informative* to the model across different XAI methods, and 2) compare the *faithfulness* of different XAI methods in each feature domain. We base our evaluation on a one-dimensional CNN digit classification model trained on the raw waveforms from the AudioMNIST dataset [3] , which achieves a test accuracy of 96%. As XAI methods, besides LRP, we consider Integrated Gradient (IG) [11], Gradient times Input (G×I) (e.g. [1]), and Sensitivity [6]. We compute LRP relevances in time domain and apply ST(DFT)-LRP via Equation (4) and Equation (10) to propagate relevances R_n from time domain x_n to frequency y_k and time-frequency $v_{m,k}$ domain. Relevance scores for the remaining XAI methods are evaluated for a model with a (ST)DFT virtual inspection layer attached to the original input layer. We choose a rectangular window of size $H = N/10$ and hop length $D = H$ for the STDFT.

Informativeness First, we compute the Shannon entropy of the heatmaps to measure their complexity [4]. In the most *informative* domain, relevance will be concentrated on only a few features that are sufficient for the prediction, which results in heatmaps with low complexity. We list the mean complexity over all heatmaps in Table 1. For each method except Sensitivity, the frequency domain shows the lowest complexity, i.e. is most informative for the model, followed by time and time-frequency domain. Notably, the visual impression of the heatmaps in Figure 2 suggests that complexity is lower for time-frequency than for the time domain, as relevance shows distinct peaks at certain frequencies in frequency *and* time-frequency domain, but is distributed rather uniformly in time domain. We suspect that the higher complexity of time-frequency heatmaps compared to time domain is due to artefact fringes in the spectrum, produced by the sharp edges of the rectangular window. The complexity of Sensitivity heatmaps is the same for frequency and time-frequency features because the method only takes into account the gradient, i.e. the weights of the DFT.

Faithfulness Second, we perform feature flipping in time, frequency, and time-frequency domain to benchmark *faithful*. Here, we either flip features to a zero baseline in order of their relevance scores (smallest destroying feature, SDF) or start with an empty signal and add the most relevant features first (smallest constructing feature, SCF). After each feature modification, i.e. addition or deletion, we measure the model's output probability for the true class. For flipping in frequency or time-frequency domain, we set the amplitude of $y_k, z_{k,m}$ to zero. In time domain, we set the time point x_n to zero. For comparability of the feature flipping curve across domains, we scale them to the ratio of modified features. A relevance attribution method that is *faithful* to the model reflects in a steep descent or ascent in true class probabilities after flipping or adding the truthfully as most important annotated features, respectively, and consequently in low or high AUC scores, that we list in Table 1. For all domains, ((ST)DFT)-LRP delivers the most *faithful* relevance heatmaps, followed by IG, G×I, and Sensitivity, according to both, SCF and SCF AUC scores.

Table 1: AUC of feature flipping curves for adding (SCF) and deleting (SDF) features, and complexity scores for a digit classifier. The method with the globally highest faithfulness per domain, i.e. highest (↑) AUC for SCF and lowest (↓) AUC for SDF, is marked in bold. Further, the domain with the lowest complexity is marked in bold for each attribution method.

		faithfulness across methods		informativeness across domains
		SCF (↑)	SDF (↓)	complexity (↓)
method	domain			
LRP	frequency	**0.66**	**0.28**	**6.00**
	time	**0.73**	**0.28**	6.69
	time-freq.	**0.69**	**0.31**	7.26
IG	frequency	0.60	0.32	**6.97**
	time	0.34	0.35	7.14
	time-freq.	0.67	0.31	8.52
G×I	frequency	0.51	0.38	**7.03**
	time	0.32	0.37	7.19
	time-freq.	0.58	0.33	8.58
Sensitivity	frequency	0.36	0.59	7.66
	time	0.51	0.41	**6.13**
	time-freq.	0.36	0.59	9.96

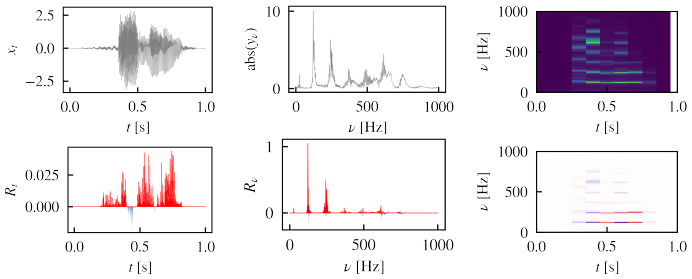

Figure 2: Time, time-frequency, and frequency signal (first row) and relevances (second row) for the digit detection task on the AudioMNIST data. The signal corresponds to a spoken seven.

5 Conclusion

We put forward virtual inspection layers that perform an identity loop via an interpretable representation to facilitate comprehensible explanations. We specialize in DFT for the virtual inspection layer and in LRP for the XAI method. In this way, we demonstrate how to extend LRP to provide interpretable explanations for time series classifiers in both the frequency and time-frequency domain. We envision applications of DFT-LRP in domains where interpreting the time domain representation of the signal is particularly challenging, in particular audio data, but also sensor data or electronic health records.

Acknowledgements

This work was supported by the Federal Ministry of Education and Research (BMBF) as grant BIFOLD (01IS18025A, 01IS180371I); the German Research Foundation (DFG) as research unit DeSBi (KI-FOR 5363); the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant TEMA (101093003); the European Union's Horizon 2020 research and innovation programme (EU Horizon 2020) as grant iToBoS (965221); and the state of Berlin within the innovation support programme ProFIT (IBB) as grant BerDiBa (10174498).

References

[1] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR (Poster)*, 2018.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.

[3] Soeren Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 2023.

[4] Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M. C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanation. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[5] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, Jul 2019.

[6] Niels JS Morch, Ulrik Kjems, Lars Kai Hansen, Claus Svarer, Ian Law, Benny Lautrup, Steve Strother, and Kelly Rehm. Visualization of neural networks using saliency maps. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 2085–2090. IEEE, 1995.

[7] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint 2104.00950*, 2021.

[8] Nils Strodthoff and Claas Strodthoff. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiological Measurement*, 40(1):015001, 11 2019.

[9] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021.

[10] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.

[11] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML, page 3319–3328, 2017.

[12] Andreas Theissler, Francesco Spinnato, Udo Schlegel, and Riccardo Guidotti. Explainable ai for time series classification: A review, taxonomy and research directions. *IEEE Access*, 10:100700–100724, 2022.

[13] Johanna Vielhaben, Sebastian Lapuschkin, Grégoire Montavon, and Wojciech Samek. Explainable ai for time series via virtual inspection layers. *arXiv:2303.06365*, 2023.