Adaption of Layerwise Relevance Propagation for Audio Applications

Roman Kiyan[1], Nils Poschadel[1], Stephan Preihs[1], Jürgen Peissig[1]
[1] *Institut für Kommunikationstechnik, Leibniz Universität Hannover,*
E-Mail: name.surname@ikt.uni-hannover.de

Introduction

Machine learning (ML) models, especially deep neural networks (DNNs), excel at particular tasks in signal detection and parameter estimation under conditions that cause traditional algorithms to struggle. For instance, convolutional recurrent neural networks (CRNNs) can surpass manually designed algorithms in performance for direction of arrival (DOA) estimation of sound sources among noise and reverberation [1]. Due to a DNN model's high complexity, the high-level problem solving *strategy* embodied by a trained network is obscure, however – the model is a *black box* [2], which is generally problematic. It is not clear, for example, whether a model has learned a valid strategy for tackling the task at hand or if its results may be based on training data artifacts altogether [3, 4].

The field of explainable artificial intelligence (XAI) is concerned with means to gain insight into such black box models. This work is concerned with applying the Layerwise Relevance Propagation (LRP) method [5] – typically used with image processing models – to sound source DOA estimation CRNNs [6, 7] in order to explore the specifics of interpreting the method's results in the context of audio applications.

Explaining Neural Network Predictions

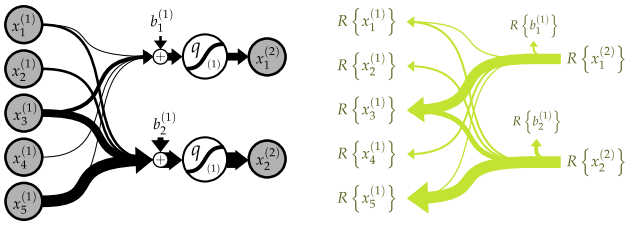
LRP is an XAI technique that aims to attribute the output produced by a model for a particular input example to the individual input features in terms of how much each feature *contributes* to the result. For a DNN this is achieved by re-tracing signal propagation through the model in reverse direction. Starting at the output layer, a conservative quantity termed *relevance* is propagated to the next lower layer according to one of several relevance redistribution rules. The method's result is obtained by repeating the procedure layer by layer until the input layer is reached.

For a fully connected (or convolutional) DNN layer which implements a mapping from nodes ${x_i}^{(l)}$ in layer l to nodes ${x_j}^{(l+1)}$ of the next layer $l+1$ according to

$$ {x_j}^{(l+1)} = q\left(b_j + \sum_i {x_i}^{(l)} \cdot w_{ij}\right), \tag{1} $$

with weights w_{ij}, biases b_j and a nonlinear activation function q, the basic relevance redistribution rule known as LRP-z [2, 5] is given by

$$ R\left\{{x_i}^{(l)}\right\} = \sum_j \frac{{x_i}^{(l)} \cdot w_{ij}}{b_j + \sum_{i'} w_{i'j} \cdot {x_{i'}}^{(l)}} \cdot R\left\{{x_j}^{(l+1)}\right\}, \tag{2} $$



(a) Forward pass (b) LRP backward pass

Figure 1: Illustration of LRP relevance redistribution for a fully connected neural network layer.

where the notation $R\{x\}$ indicates relevance assigned to variable x. LRP thus redistributes each upper layer node's relevance $R\left\{{x_j}^{(l+1)}\right\}$ to the lower layer nodes ${x_i}^{(l)}$ according to their contributions ${x_i}^{(l)} \cdot w_{ij}$ in relation to the total value of ${x_j}^{(l+1)}$ prior to the activation function. This principle is illustrated by Fig. 1.

Modifications of LRP including numerical stabilization (LRP-ϵ) or improved suitability for convolutional layers (LRP-$\alpha\beta$) exist [2, 5], with best practices regarding the selection of rules and hyperparameters having been surveyed [8]. LRP has also been modified for recurrent structures such as Long Short Term Memory (LSTM) [9].

While the typical domain of LRP is image processing, the method has been applied to neural networks for audio analysis including estimation of sound source DOA [1] or elevation [10] as well as speech classification [11].

Interpreting Relevance Heatmaps

The models examined in this work are CRNNs for DOA estimation of one-second segments of speech signals [12] sampled at 16 kHz and convolved with simulated [13] Higher Order Ambisonics (HOA) [14] room impulse responses and with added diffuse babble noise (created by averaging speech signals).

Fig. 2(a) depicts the magnitude spectrogram of an example input signal, with relevance values for each time-frequency bin (sum of magnitude and phase channel relevance) obtained from an analysis of a second order Ambisonics model being displayed as a heatmap in Fig. 2(b). When LRP is applied to audio processing DNNs, which oftentimes use some form of spectrograms as their inputs, the results are typically displayed graphically and interpreted through visual inspection of the time-frequency domain data [1, 11]. While this is a valid approach, one must keep in mind that spectrograms are fundamentally distinct from images in several ways. On the one hand, audio signals are inherently *additive*, which im-

plicates that each time-frequency bin in a spectrogram does not unambiguously belong to one of the signal components present in a scene. This behavior is unlike images, where each pixel is generally occupied by a single object only [18]. Moreover, spectrograms differ from images with respect to the meaning that *channels* carry. In images, channels are used to represent color, whereas a multichannel spectrogram may bear a variety of meanings based on the specific configuration – such as HOA signals or intensity vector components.

These specifics complicate the process of gaining knowledge from LRP analysis in audio applications. The main contribution of this work lies in the proposition of an approach based on further analysis of the results produced by LRP enabling observations beyond simple visual comparison of input data and relevance heatmaps.

Experimental Setup

Training and test data generation for the experiments in this work adheres to the procedures established by Poschadel et al. [6, 7]. The model architecture, which is summarized in Tab. 1, is identical to that presented by Poschadel et al. [6, 7] except for the models performing single-source DOA classification, thus using a softmax output activation and assigning scores to 425 direction classes. Among the analyzed models, four models use HOA signals of up to fourth order whose Short Time Fourier Transforms (STFTs) serve as model inputs. Magnitude and phase of the complex-valued spectrograms are treated as individual channels. Results presented here are for a second order Ambisonics model operating on normalized input data (normalization of magnitude and phase channels through subtraction of the mean and division by the standard deviation of training set magnitude and phase data, respectively). Another model uses spectrograms of the x, y and z components of the active and reactive intensity vectors computed from first order Ambisonics signals [1, 6].

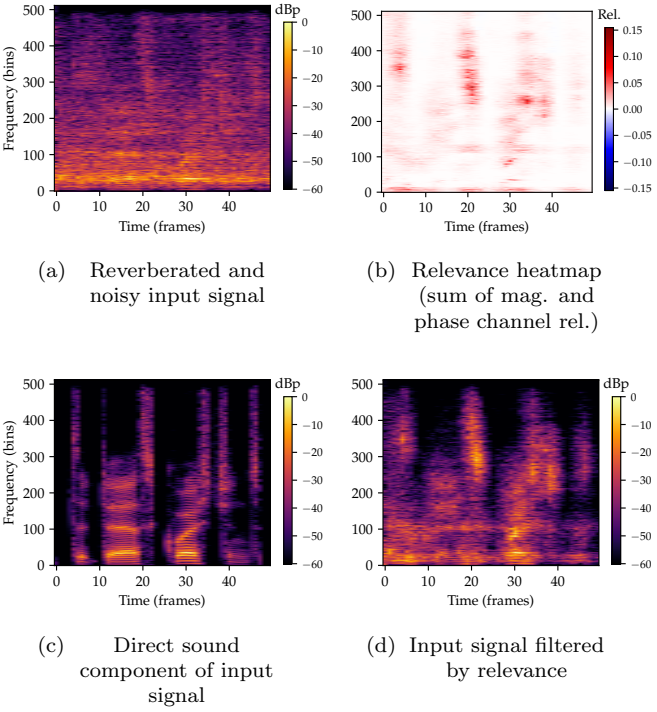
A custom LRP framework for use with the TensorFlow 2 Python library [15] has been created, inspired by and partly based on existing implementations [9, 16, 17]. LRP-$\alpha\beta$ with $\alpha = 1$, $\beta = 0$ has been used for the convolutional layers, LRP-ϵ with $\epsilon = 0.01$ has been used for fully connected layers. The existing variant of LRP for LSTM [9] is designed for LSTM layers which pass only their block outputs at the final time step to the following neural network layer. Since the bidirectional LSTM (BiLSTM) layers in the models examined here yield their block outputs at each time step, the LRP algorithm for LSTM has been modified to also inject relevance from the next upper layer at each time step during the LRP backward pass. For the LSTM layers, $\epsilon = 0.01$ has been used. Relevance attributed to biases in the fully connected and BiLSTM layers has been redistributed among lower layer nodes in order to ensure relevance conservation from layer to layer [2].

Channel Relevance

The intensity-based CRNN shall serve as subject for demonstrating an evaluation of channel relevance. With the intensity vector being clearly related to sound source

Table 1: Architecture of the examined CRNN models. N_{in} is the number of input channels (twice the number of Ambisonics channels for HOA models, six channels for the intensity-based model), $N_{\text{filt.}} = 64$ is the number of filters in convolutional layers, $N_{\text{block}} = 50$ is the number of LSTM blocks, $N_{\text{bin}} = 425$ is the number of DOA classification bins. ELU is the exponential linear unit activation function.

Layers	Output dim.
Normalization (HOA only)	$50 \times 512 \times N_{\text{in}}$
Conv. $\rightarrow$ Batch norm. $\rightarrow$ ELU	$50 \times 512 \times N_{\text{filt.}}$
Max. pooling	$50 \times 64 \times N_{\text{filt.}}$
Conv. $\rightarrow$ Batch norm. $\rightarrow$ ELU	$50 \times 64 \times N_{\text{filt.}}$
Max. pooling	$50 \times 8 \times N_{\text{filt.}}$
Conv. $\rightarrow$ Batch norm. $\rightarrow$ ELU	$50 \times 8 \times N_{\text{filt.}}$
Max. pooling	$50 \times 2 \times N_{\text{filt.}}$
Reshape	$50 \times 2 \cdot N_{\text{filt.}}$
BiLSTM	$50 \times 2 \cdot N_{\text{block}}$
BiLSTM	$50 \times 2 \cdot N_{\text{block}}$
Fully conn. $\rightarrow$ ELU	$50 \times 2 \cdot N_{\text{block}}$
Fully conn. $\rightarrow$ softmax	$50 \times 2 \cdot N_{\text{bin}}$



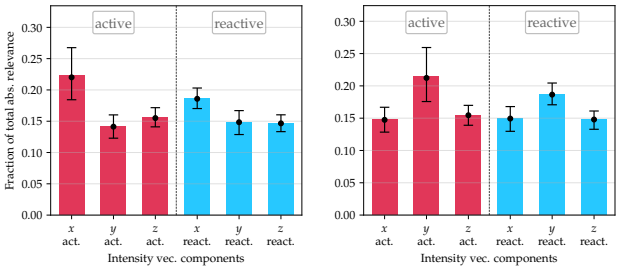
(a) Reverberated and noisy input signal

(b) Relevance heatmap (sum of mag. and phase channel rel.)

(c) Direct sound component of input signal

(d) Input signal filtered by relevance

Figure 2: Magnitude spectrograms and relevance heatmap for one input example for a second order Ambisonics DOA estimation CRNN, all figures for Ambisonics channel 0. For visual clarity, spectrograms are displayed in logarithmic scale w. r. t. their peak value (dBp), but the magnitude spectrogram input to the CRNN is linear.



(a) Examples with $\vec{d} = \vec{e}_{x,\,\mathrm{bin}^+}, \vec{e}_{x,\,\mathrm{bin}^-}$

(b) Examples with $\vec{d} = \vec{e}_{y,\,\mathrm{bin}^+}, \vec{e}_{y,\,\mathrm{bin}^-}$

Figure 3: Channel relevance analysis for an intensity-based DOA estimation CRNN on a set of 1000 examples with DOA $\vec{d}$ corresponding to the direction bins containing the basis vectors $\pm\vec{e}_x, \pm\vec{e}_y$. Bars represent the mean, dots and whiskers the median and 5 %, 95 % percentiles of absolute channel relevance (summed over time and frequency) within each group of examples.

DOA, this model lends itself to an analysis of how DOA is reflected in the distribution of LRP relevance across input channels. To this end, a set of 1000 test examples, each featuring one of the four DOAs corresponding to the direction bins that include the orthogonal unit vectors $\pm\vec{e}_x$, $\pm\vec{e}_y$, has been created. After LRP analysis, relevance has been summed across time and frequency for each channel. As highlighted by Fig. 3, the input channel corresponding to the respective example DOA receives a greater share of relevance than the other components among both active and reactive intensity channels. Channels *not* corresponding to the example DOA receive a considerable amount of relevance nonetheless, suggesting that the model is performing a comparison of signals across channels in order to arrive at its DOA estimate.

What is more, Fig. 3 shows that the entirety of active intensity channels receives a greater share of total relevance than reactive intensity. This effect has been further investigated using a technique known as *pixel flipping* [5, 19] – time-frequency bins have been sorted by the relevance attributed to them and successively set to zero while observing the score that the DNN assigns to the true DOA class for the perturbed input. Fig. 4 compares the results of such experiments using a subset of 2583 examples from the original test set with random DOAs where perturbations have been introduced to all input channels or to active and reactive intensity channels only, respectively. These experiments permit to conclude that active intensity serves as a primary source of information to the DNN while reactive intensity provides additional cues – the model is somewhat able to cope with minor perturbations in the reactive intensity channels, but perturbing active intensity causes as sharp a drop in true class score as perturbing all channels, indicating that intact reactive intensity information alone is insufficient for the model's operation. This is consistent with the physical interpretation of active intensity representing net energy transport in the sound field and reactive intensity corresponding to zero-mean energy fluctuations [20].

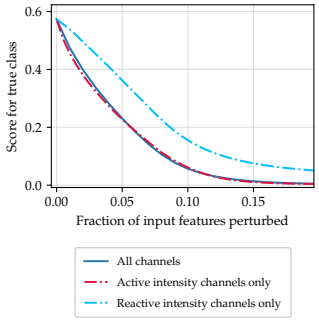


Figure 4: Pixel flipping experiment for an intensity-based DOA estimation CRNN, curves averaged over a set of 2583 examples with random DOAs. Perturbation has been applied by setting time-frequency bins to zero either in all input channels in parallel or in active resp. reactive intensity channels only.

Auralization of Relevance

Due to the additivity of audio signals, interpreting LRP relevance heatmaps such as the one displayed in Fig. 2(b) in terms of sound events can be difficult. As an intuitive approach to this problem, a filtering method that extracts relevant sound features has been developed. When aiming to create a *listenable* representation of LRP relevance, simply computing the inverse STFT of spectrograms weighted by relevance is likely to produce sound artifacts. The developed filtering scheme is essentially a more well-behaved variant of this operation. It is based on frame-wise convolution with impulse responses, extracting relevant audio $a_{i,\mathrm{rel}}(m)$ from the input signal signal $a_i(m)$ according to

$$a_{i,\mathrm{rel}}(m) = \sum_{n=-\infty}^{\infty} \left(a_i(m) \cdot w(m - n \cdot M_{\mathrm{hop}})\right) * g_{i,n}(m) \tag{3}$$

where $w(m)$ and M_{hop} are the STFT window function and hop size (stride) and $g_{i,n}(m)$ is an impulse response for filtering channel i in time frame n. The filters $g_{i,n}(m)$, $g_{i,n+1}(m)$, … are constructed for each frame n from the respective column in the heatmap $H_i(n,k)$, where H_i can be either the magnitude or the phase relevance heatmap (or the sum of both) for the single Ambisonics channel i, for instance. In order to define the $g_{i,n}(m)$, a perfect reconstruction B-band filter bank composed of linear phase FIR filters $\gamma_b(m)$ with $b = 1, \ldots, B$ is used. With K_H being the number of frequency bins in $H_i(n,k)$ and $\Gamma_b(k)$ being the DFT of $\gamma_b(m)$, the filter $g_{i,n}(m)$ is defined as

$$g_{i,n}(m) = \sum_{b=1}^{B} \left(\sum_{k=1}^{K_H} |\Gamma_b(k)| \cdot H_i(n,k) \right) \cdot \gamma_b(m). \tag{4}$$

Fig. 2(d) shows a spectrogram of the reverberated and noisy input signal of Fig. 2(a) filtered using the heatmap of Fig. 2(b) and a filter bank with $B = 20$ bands employing 129-tap filters with equidistant cutoff frequencies designed using the window method. For comparison, Fig. 2(c) shows the noise-free direct sound component of the input signal. It becomes apparent that the DNN attempts to ignore time-frequency regions mostly occupied

by noise and diffuse reverberation, instead focusing on regions where the direct sound speech signal is prominent. This is a non-trivial finding since it demonstrates that the analyzed model has learned *implicitly* to detect human speech based on just the (noisy and reverberated) training examples and corresponding DOA labels. Also, this finding confirms that the model has indeed learned a plausible estimation strategy with respect to the problem that it needs to solve.

Conclusion

XAI techniques aim to provide explanations to the results of ML algorithms such as DNNs, providing knowledge on a model's otherwise hidden decision strategy as well an additional metric – plausibility of the learned strategy – in addition to pure prediction accuracy. In this work, the LRP technique has been applied to CRNN models for sound source DOA estimation and results have been interpreted by leveraging the multichannel nature of the input signals as well as subjecting those signals to a filtering technique in order to extract relevant sound events. Specifically in the context of audio, these additional techniques are proposed as extensions to the LRP technique for further analysis of the attribution results produced by the method. More generally, this work highlights that further insight on the strategy implemented by a DNN model can be gained if results produced by LRP and other XAI methods are viewed in the light of higher-level concepts pertaining to the task at hand than the raw input features used by the model.

References

[1] Perotin, L., Serizel, R., Vincent, E., Guerin, A.: CRNN-Based Multiple DoA Estimation Using Acoustic Intensity Features for Ambisonics Recordings. IEEE Journal of Selected Topics in Signal Processing 13(1) (2019), 22-33.

[2] Lapuschkin S.: Opening the machine learning black box with Layer-wise Relevance Propagation. PhD thesis. Technische Universität Berlin, 2019.

[3] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., Müller, K.-R.: Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE, 109(3) (2021), 247-278.

[4] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R.: Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications, 10, 1096 (2019).

[5] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE 10(7) (2015).

[6] Poschadel, N., Hupke, R., Preihs, S., Peissig, J.: Direction of Arrival Estimation of Noisy Speech using Convolutional Recurrent Neural Networks with Higher-Order Ambisonics Signals. 29th European Signal Processing Conference (EUSIPCO), 2021.

[7] Poschadel, N., Preihs, S., Peissig, J.: Multi-Source Direction of Arrival Estimation of Noisy Speech using Convolutional Recurrent Neural Networks with Higher-Order Ambisonics Signals. 29th European Signal Processing Conference (EUSIPCO), 2021.

[8] Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., Lapuschkin, S.: Towards Best Practice in Explaining Neural Network Decisions with LRP. International Joint Conference on Neural Networks (IJCNN), 2020.

[9] Arras, L., Montavon, G., Müller, K.-R., Samek, W.: Explaining Recurrent Neural Network Predictions in Sentiment Analysis. 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017.

[10] Thuillier, E., Gamper, H., Tashev, I. J.: Spatial Audio Feature Discovery with Convolutional Neural Networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[11] Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., Samek, W.: Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. arXiv preprint 1807.03418, 2018.

[12] Garofolo, J. S. et al.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, 1993.

[13] Wabnitz, A., Epain, N., Jin, C., Van Schaik, A.: Room acoustics simulation for multichannel microphone arrays. International Symposium on Room Acoustics, 2010.

[14] Zotter F., Frank, M.: Ambisonics. Springer International Publishing, 2019.

[15] Abadi, M. et al.: TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[16] Alber, M. et al.: iNNvestigate neural networks! Journal of Machine Learning Research, 20(93) (2019).

[17] Warnecke, A., Arp, D., Wressnegger, C., Rieck, K.: Evaluating Explanation Methods for Deep Learning in Security. IEEE European Symposium on Security and Privacy (EuroS&P), 2020.

[18] Wyse, L.: Audio spectrogram representations for processing with Convolutional Neural Networks. First International Workshop on Deep Learning and Music, 2017.

[19] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R.: Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Transactions on Neural Networks and Learning Systems, 28(11) (2017), 2660-2673.

[20] Jacobsen, F.: A note on instantaneous and time-averaged active and reactive sound intensity. Journal of Sound and Vibration, 147(3) (1991), 489-496.