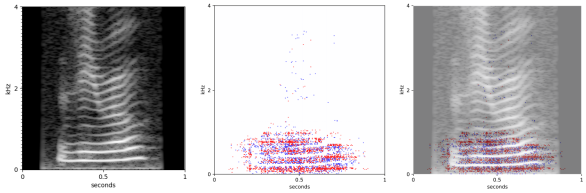
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2021
http://excel.fit.vutbr.cz

Interpretation of Deep Neural Networks in Speech Classification

Marek Sarvaš*



Abstract

The growing problem of the popularity of using deep neural networks is their black box representation. The lack of transparency is raising questions about their reliability, credibility, or vulnerability to adversarial attacks. This caused rising demand for neural network explainability. The goal of this paper is to replicate existing experiments on a gender classification model and extend these experiments to analyze and uncover vulnerabilities of a network trained for gender classification on audio signal spectrograms. The easiest way to explain something is through visualization. For this, a layer-wise relevance propagation technique was chosen in this work because it produces easy-to-understand heatmaps of features relevant to a neural network. The heatmaps are produced by back-propagating relevances through a network from the output to the input layer. Two neural network models with AlexNet and ResNet architecture were used. Experiments with AlexNet model show that the network's predictions are highly dependent on a small number of time-frequency (TF) bins. By augmenting the training data using obtained relevance maps, I managed to lower the dependency on these bins. As a result, the prediction accuracy, when these bins were not present, was increased by 15%. The proposed approach can potentially lead to increased robustness of models, preventing or reducing the impact of adversarial attacks. Interpretation of ResNet model showed dependencies on lower frequencies and time. Producing interpretable heatmaps of the ResNet model required the implementation of more robust LRP rules.

Keywords: neural network interpretation — Layer-wise Relevance Propagation — deep neural networks — speech classification

Supplementary Material: N/A

*xsarva00@fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Deep neural networks are, nowadays, heavily used as state-of-the-art solutions to problems like image, audio processing, or natural language understanding. Yet, they still represent a black box where input comes into the neural network and prediction comes out, but inner decision-making remains hidden. By analyzing some high performing models trained for image classification, discoveries showed that predictions were dependant on artifacts such as image watermark[1] or background[2]. Even though these models have high accuracy of predicting ground truth on train or test datasets, the reasons for these predictions are considered wrong. Such problems of the models are hard to uncover on limited datasets and end up revealed after a while, if at all.

As demand for explainable neural networks is rising, more discoveries and experiments are made. Because the easiest way to explain and understand something is through visualization, interpretation of image classification models can be easily understood. This paper aims to bring more insight into how deep neural network models decide when predicting a person's gender from speech recording processed as a spectrogram, following previous work on this topic [3].

Convolution neural network with AlexNet architecture was trained on spectrograms of audio signals from AudioMNIST dataset[3], achieving 97% accuracy of predicting correct gender. The Layer-wise relevance propagation was chosen as an explanation method for its reliability, implementation efficiency. Produced heatmaps of relevances revealed that the model's predictions were heavily dependent on few time-frequency (TF) bins. I manage to lower dependency on these few TF bins and boost the model prediction accuracy, when values of these bins are set to zero, by 15%. The same technique could be used on other speech processing models, uncovering decisions and potentially making them more robust and safer to adversarial attacks.

2. Interpretation of neural networks

Convolutional neural networks (CNN), as a type of the Deep neural network (DNN), are very popular for image and even audio classification problems. These architectures are composed of a chain of multiple layers, where every layer has a significant number of learnable parameters. CNNs architectures have great performance in pattern recognition and their advantage comes mainly from reduced number of learnable parameters due to the utilization of convolution and pooling layers [4].

2.1 Uncovering Clever Hans predictors

In image classification, several experiments were proposed to discover if it is possible to explain what features are the predictions based on. An interesting fact is that these experiments uncovered the so-called Clever-Hans predictors [5]. The clever Hans predictor is a term for a deep neural network model whose decisions and performance may perform extremely well on train and validation datasets but may fail in real-life applications because it learned to use undesirable features in images. For example experiments in the article uncovered that high performing model based its prediction of a horse in the image on the occurring watermark [1]. In other experiments [2], the model learned to distinguish between wolf and husky based on snow in the background. These experiments show the importance of interpretation methods and neural network interpretation in general to prevent previously mentioned problems. However, a lot of these discoveries were made by accident rather than targeted research, but they can be the right step towards a better understanding and improving neural networks.

2.2 Difficulties in explaining Neural Networks

Explaining deep neural networks comes, according to *Toward Interpretable Machine Learning*[6] research, with three main difficulties caused by the complexity of the models. The complexity comes from the number of layers that perform linear and non-linear transformations of the input. The first difficulty comes from a combination of neurons that are activated locally, by the small fraction of data points, and neurons activated more globally. Thus the output of the networks is affected by global as well as local effects in the input. The second difficulty comes from the presence of a *shattered gradient*[7] effect in ReLU neural networks with higher depth, where the gradient becomes more fast-changing. This can cause problems in explanation methods that depend on the usage of the model's gradient, such as sensitivity analysis or simple Taylor decomposition [8]. The last difficulty is finding a reference point as the base of the explanation. This problem comes from local explanation methods where the explanation methods are based on a comparison of predicted output and the reference point. The output can change rapidly based on the reference point, even when the reference point does not carry any significant information for further interpretation.

2.3 Comparison of different methods

Different methods were proposed to attempt to explain various neural network models. The article by W. Samek and G. Montavon[6] summarized three explanation methods belonging into distinct groups of explanation, each with different advantages and disadvantages.

First method is Occlusion analysis [9], which is a specific type of perturbation analysis, where input features of neural network or whole patches are being occluded. For example, when explaining models trained for image classification, square regions of the input image are replaced with grey or zero values. The relevance heatmap is obtained by measuring the effect of occluded regions on the prediction and accuracy of the explained model. This method is the easiest to implement, does not require access to the source code of a model, but is the worst of the three mentioned methods in terms of runtime efficiency.

Another method is Integrated Gradients belonging to a group of methods for explaining deep neural networks based on their gradients. Other variant is, for example, SmoothGrad [6]. The Integrated Gradient method utilizes sensitivity of backpropagation methods and implementation invariance of gradients. On the other hand, it suffers from the shattered gradient problem [7]. In addition, this method is almost as slow as Occlusion analysis.

The last proposed method is Layer-wise relevance propagation (LRP) which belongs to a group of backward propagation techniques. The goal of LRP is to produce heatmaps of positive and negative relevances. These techniques utilize deep neural networks' layered structure. They scale better when used on complex deep neural networks than gradient-based methods, but can be used also on different machine learning models. The heatmap is obtained by backpropagating relevances from the output layer through the model to the input layer. This method does not use a model's gradient, therefore is resistant to a shattered gradient effect. In [6], LRP was placed in first and second place in terms of runtime efficiency and human interpretability. On the other hand, LRP depends on access to the neural network's source code as its implementation depends on a model's structure.

Every method produces slightly distinct relevance scores and heatmaps. In computer vision, for example, LRP tends to highlight features mostly in favor of positive relevances. The occlusion method highlights important regions in the image. And the integrated gradient highlights relevant time-frequency bins but shows more negative relevance in heatmap than LRP.

2.4 Speech interpretation

Because one of the best ways to interpret neural networks is through visualization, the interpretation of audio signals can be more challenging than image interpretation. It seems that interpreting neural network models for audio classification is not as popular or is in progress due to the higher difficulty. To gain new insight into audio signal classification, a few experiments were proposed in articles by W. Samek and G. Montavon[3] and S. Becker with others [6]. The experiments were done on raw waveforms and audio spectrograms using the AudioMNIST[3] dataset. Layer-wise relevance propagation was chosen as an explanation method for used CNN models. In both cases, raw waveforms and audio spectrograms, LRP highlighted features based on their contribution to the prediction. Results showed that raw audio signals are not the best way to explain neural networks. LRP highlighted relevant features in a raw waveform, but they are hard to interpret to obtain new information about neural network decisions. On the other hand, spectrograms brought insight into how lower frequencies affect gender predictions. The approach in this work is inspired by the article created by W. Samek and G. Montavon [3].

3. Layer-wise relevance propagation

Layer-wise relevance propagation (LRP)[10] computes activation scores in forward pass and subsequently propagates the output of the network as relevance scores in a backward direction towards the input layer using propagation rules. Information about different rules and their effect is derived from [11] and [12]. The propagation process is conservative analogous to Kirchhoff's current law in electrical circuits. In neural networks, this means that all activation energy or relevance (in backward propagation) flowing into the neuron has to flow out of the neuron, i.e., being redistributed into the lower layer. The product of back-propagation is a heatmap showing relevant features that have a positive (red) and negative (blue) impact on the model's prediction.

The basic equation, also referred to as the LRP-0 rule, to propagate relevances through the model is defined as:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_j a_j w_{jk}} R_k \,, \tag{1}$$

where j and k are neurons of two consecutive layers, R_j is relevance of neuron j, R_k is relevance of neuron k computed in previous layer, or in case of the output layer, R_k is output of the model, a_j is activation energy of neuron j, i.e. output value of the activation function, and w_{jk} is value of weight between neuron j in given layer and neuron k in previous layer (in the back-propagation direction from the output layer to the input layer). The numerator $a_j w_{jk}$ represent contribution of neuron j to neuron k. The basic rule can be improved to produce a more robust explanation of a model.

The first improvement is denoted as the LRP-ε rule consisting of a constant value ε added to the denominator (2). The addition of ε causes small or contradictory relevances of neuron k to be absorbed, producing a less noisy heatmap with fewer input features presented.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_j a_j w_{jk} + \varepsilon \cdot sign(\sum_j a_j w_{jk})} R_k \,. \tag{2}$$

Another possible improvement from LRP-0 is a rule denoted as LRP-γ (3) achieved by disproportionately favoring the positive contribution of relevances. The

value of γ determines how much are positive relevances favored over negative ones producing more stable, smooth, and less noisy heatmaps.

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\sum_j a_j \cdot (w_{jk} + \gamma w_{jk}^+)} R_k \,. \tag{3}$$

In equation 3, w_{jk}^+ is defined as positive part of weight of neuron j, i.e., $w_{jk}^+ = \max(0, w_{jk})$.

4. Experimenting with neural network interpretation

The following experiments aim to determine LRP faithfulness, analyze the accuracy of trained models with different validation datasets and attempt to reduce high model sensitivity to a small number of features.

4.1 Gender classification

In this work, I implemented a model and Layer-wise relevance propagation using PyTorch library [13]. The model is a convolutional neural network with AlexNet architecture [14] consisting of five convolution layers with ReLU activation functions and max-pooling layer, and three fully connected layers for classification. The properties of layers and size of kernels were adjusted as described in articles by W. Samek and G. Montavon [3]. Layer-wise relevance propagation was implemented in two steps. Firstly, the activation energy and parameters of each layer are stored using forward-hooks in the forward pass of input through the model. Secondly, the heatmap is produced by propagating the model prediction backward using stored parameters of each layer and the LRP-0 rule. In this case, this rule is sufficient enough to provide heatmaps interpretable to humans and correctly finding the most relevant time-frequency bins, as shown in Figure 3.

Heatmap produced by the trained model is shown in Figure 1, where red highlights positive relevances and blue negative ones. After adding the heatmap on the top of the original spectrogram Figure 2 we can see that the model's gender prediction is truly based on the lower frequencies of audio recording, which corresponds with the study of female and male fundamental frequencies[15].

4.2 AlexNet model performance

The AlexNet model was trained on the AudioMNIST dataset[3] on 200 epochs with the learning rate set to $1e-4$. The training set consists of 6000 spectrograms with a size of 227×227 time-frequency bins. Half of the training set was male recordings and half female. The achieved accuracy of this model was 97.8% on

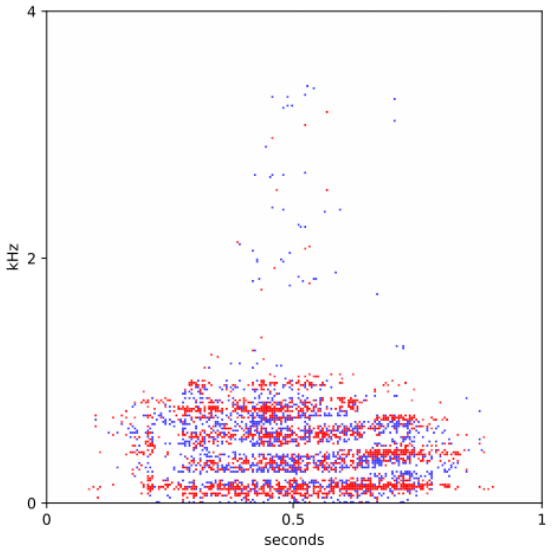


Figure 1. Produced heatmap of correctly classified female audio recording.

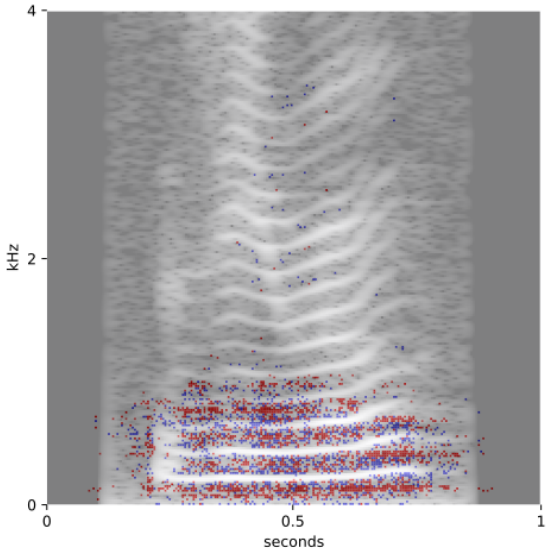


Figure 2. Heatmap on top of spectrogram showing that time-frequency bins representing lower frequencies are most important for correct prediction.

a test set consisting of 1500 male and 1500 female recordings. The faithfulness of implemented LRP on AlexNet models was tested using the pixel-flipping method. The Pixel-flipping method sets the value of a spectrogram's bins to 0 w.r.t. heatmap generated by LRP. Time-frequency bins are set to zero from the most relevant to least by sorting all bins in a heatmap based on their value in descending order. Then, the bins with the highest positive relevance values are chosen. Assuming that LRP works correctly and highlights the most relevant bins, the model accuracy should decrease rapidly when these bins are not present. The blue curve in Figures 3, 4 and 6 represents AlexNet's accuracy on input data modified by pixel-flipping using LRP. The orange curve represents the accuracy of the

same model, but the input data are modified by pixel-flipping randomly. Figure 4 provides a closer look at the accuracy decrease by modifying data from 0 to 1% using a heatmap produced by LRP. The results in Figure 3 show that the model is highly dependant on a small number of features. The accuracy dropped from 97.8% to 12.3% when only 0.5% of bins were set to zero. This result differs from the original paper. I do not know how was the pixel-flipping implemented in the original paper. However, the drop in the model's accuracy in my experiments was caused by the model's high dependency on a small number of time-frequency bins. Therefore, setting these bins to zero created a big predominance of bins with negative relevance scores towards the correct prediction. The presence of bins with negative relevance values and the absence of bins with positive relevance values towards correct prediction caused the model to predict the opposite gender nearly every time.

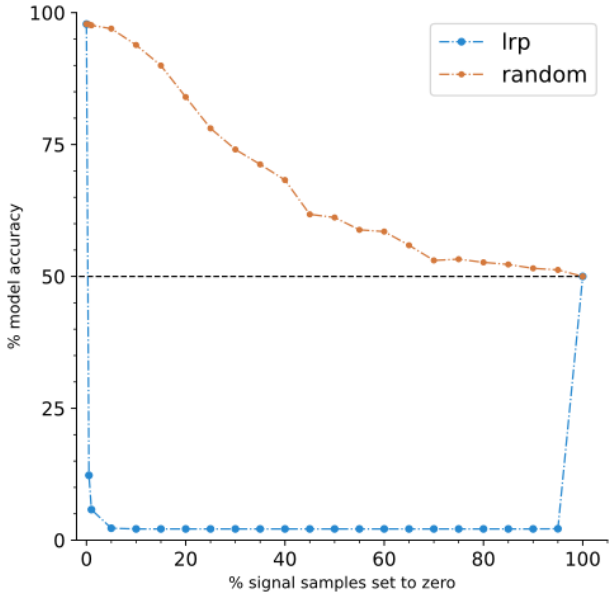


Figure 3. AlexNet model trained on AudioMNIST dataset accuracy w.r.t. percentage of each spectrogram's bins set to zero in validation set.

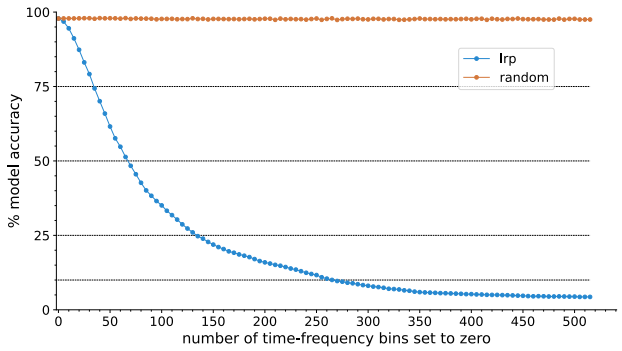


Figure 4. Same principle as Figure 3 but closer look of accuracy drop from 0% to 1% of bins set to zero.

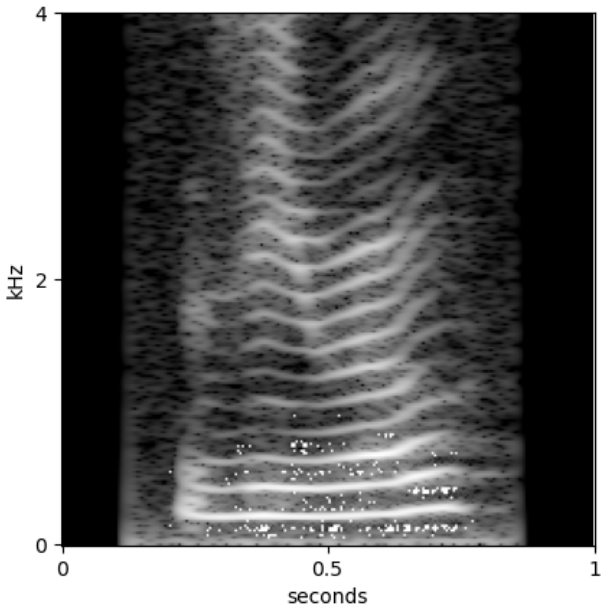


Figure 5. Spectrogram with 0.5% most relevant bins, i.e., bins with the highest positive relevance values, are set to zero.

4.3 Improvement with LRP augmented dataset

High dependency on such a small number of features as showed in the previous experiment is not ideal and can make the model vulnerable to noisy input or adversarial attacks. Because of this, I propose to use LRP as an improvement tool for a previously trained model. Data of the AudioMNIST training set was modified w.r.t. the most relevant bins according to LRP as follows. The value of 0.5% of the most relevant bins Figure 5, in each spectrogram, was set to the mean of their Moore neighborhood. The pre-trained model from 4.2 was then again trained on this augmented dataset. Evaluation using the original AudioMNIST validation set and pixel-flipping method showed that this newly trained model has slightly higher accuracy on the original dataset. In addition, the performance of the model when the most relevant bins are not present has been more than doubled as shown in the Table 1 and Figure 6.

5. Speaker ID classification

To extend previous experiments with AlexNet, I applied the LRP method to a more complex model for speaker classification. This model is based on ResNet34 architecture[16] with some changes to perform well on a designated task. This architecture is using mainly a combination of 2D convolutional layers (Conv2d) and 2D batch-normalization (BatchNorm2d) layers with dense layers at the end. Pretrained model was provided by VUT FIT with a classification accuracy of 96%. Input data for this network are 64-dimensional

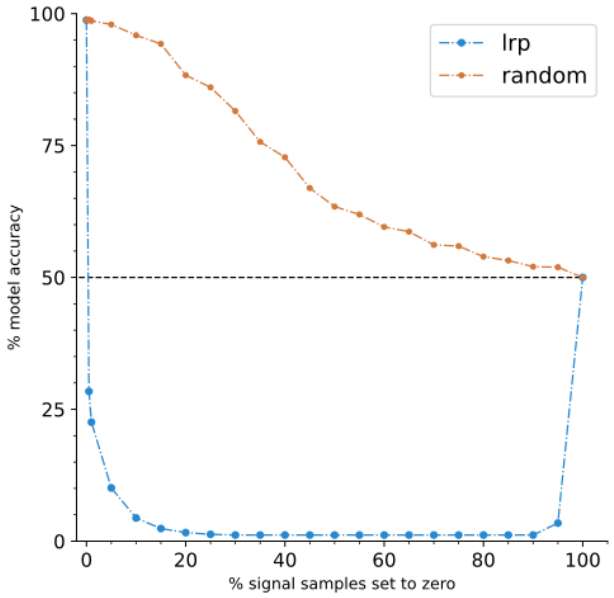


Figure 6. Previously trained AlexNet model trained again on augmented AudioMNIST dataset. Accuracy is showed w.r.t. percentage of each spectrogram's bins set to zero in validation set.

Table 1. Each model with its prediction accuracy is presented in separate row. *AlexNet* represents basic CNN model trained on AudioMNIST dataset, *AlexNet_augmented* is additionally trained on augmented AudioMNIST dataset, as described in subsection 4.3. Other columns show model's accuracy on validation datasets where 0% and 0.5% of the most relevant spectrogram bins were set to zero, respectively.

Model	**0%**	**0.5%**
AlexNet	97.8%	12.3%
AlexNet_augmented	98.7%	28.4%

filter banks from the VoxCeleb dataset[17][18][19] augmented with noise and music. The prediction of the model is tensor of size $(N, 5994)$, where N represents batch size, and classified speaker ID is obtained as $y = argmax(logits)$, where logits represent the value of each speaker. A spectrogram of the input data is shown in the Figure 7. Heatmap produced only by using the LRP-0 rule is shown in Figure 8. In the case of this ResNet model, produced heatmaps are noisy and hard to interpret. The LRP-0 method creates clusters of mixed time-frequency bins with positive and negative relevance values. Because the LRP-0 rule tends to create noisy heatmaps, more robust rules, such as LRP-ε and LRP-γ, need to be used.

I updated the relevance propagation using LRP-ε with $\varepsilon = 0.5$. In this experiment, the LRP-0 is

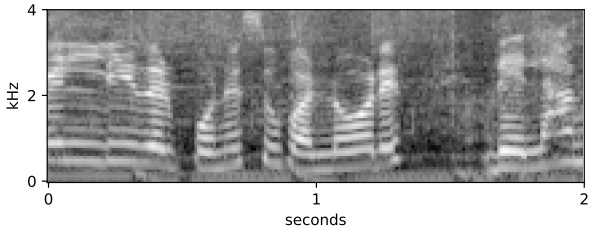


Figure 7. Spectrograms of VoxCeleb features of the 2s segment of recording.

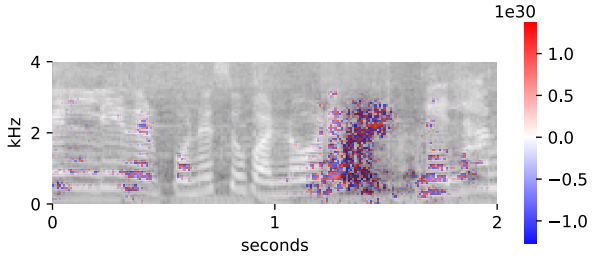


Figure 8. Heatmap produced by LRP-0 overlayed on top of features of a 2s segment of recording from VoxCeleb dataset.

only used for linear and 1D batch normalization layers near the output layer. For hidden layers consisting of Conv2d and BatchNorm2d layers and input layer, the LRP-ε rule was used. Heatmap produced by this updated method, with positive (red) and negative (blue) time-frequency bins, is shown in Figure 9. By introducing the LRP-ε rule, the heatmap is less noisy, bins with positive and negative relevance values are no longer mixed, and they are more distributed along the time axis. In Figure 10 is shown evaluation of the model with the pixel-flipping method. The pixel-flipping evaluation for speaker ID interpretation is based on the same principle as in previous experiments with AlexNet. Based on these results, I assume that the Resnet model for speaker ID classification is based on the lower frequencies (similar to the AlexNet model) in time.

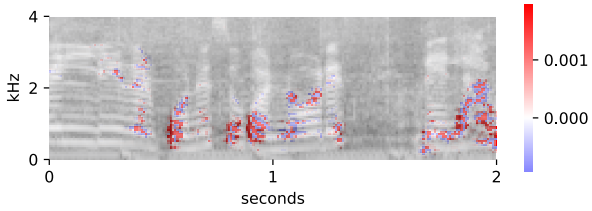


Figure 9. Heatmap produced by the combination of LRP-0 and LRP-ε overlayed on top of features of a 2s segment of recording from VoxCeleb dataset.

Even though using LRP-0 and LRP-ε showed promising results, I tried adding the LRP-γ rule with value $\gamma = 5$ for $1/3$ of the hidden layers in the upper part of the model, i.e., layers closer to the input layer. Figures

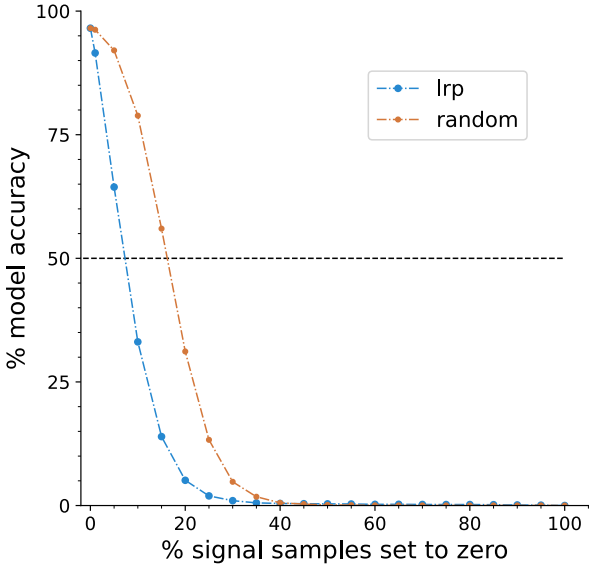


Figure 10. ResNet model's accuracy w.r.t. percentage of each spectrogram's bins set to zero. Blue curve represents setting time-frequency bins to zero in descending order from the ones with the highest positive relevance value. Random curve represents setting time-frequency bins to zero at random.

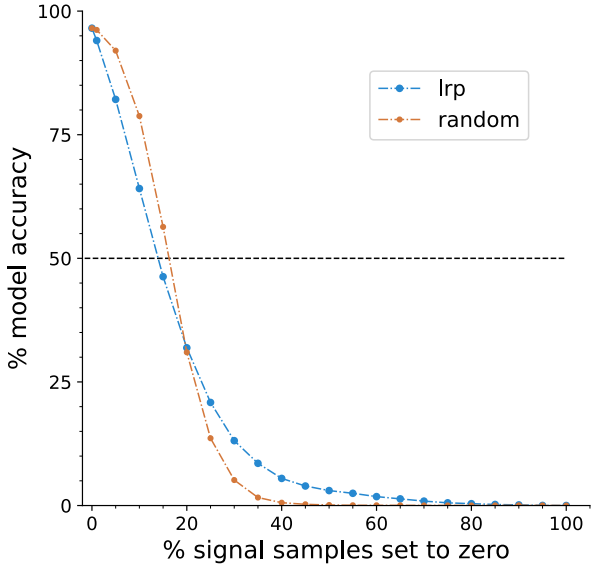


Figure 12. ResNet model's accuracy w.r.t. percentage of each spectrogram's bins set to zero. Blue curve represents setting time-frequency bins to zero in descending order from the ones with the highest positive relevance value. Random curve represents setting time-frequency bins to zero at random.

11 and 12 show produced heatmap and faithfulness evaluation with the pixel-flipping method, respectively. The LRP-γ rule favors the time-frequency bins with positive relevance values (red); therefore, these areas are a bit clearer, and more bins have positive relevance. Because the LRP-γ favors bins with positive relevance values and gamma=5, the positive values are much bigger than in the heatmap produced only with the combination of LRP-0 and LRP-ε. Another effect of using the LRP-γ is that some bins are presented as more relevant; therefore, the drop in accuracy in Figure 12 is not as steep as in Figure 10.

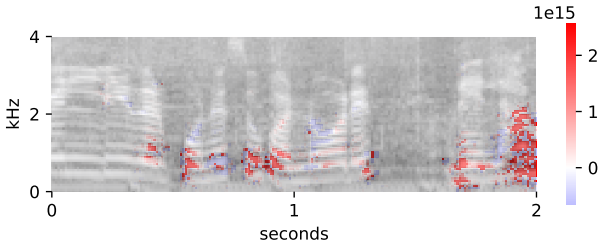


Figure 11. Heatmap produced by the combination of LRP-0 and LRP-ε and LRP-γ overlayed on top of features of a 2s segment of recording from VoxCeleb dataset.

6. Conclusions

This paper presented an analysis and prediction improvement of the deep neural network model trained for speech classification. Model predictions were explained using heatmaps produced by the LRP method, showing the model vulnerability and high dependency on a small number of features presented in lower frequencies. These heatmaps were then used to create a dataset by augmentation of the original AudioMNIST dataset. This new dataset served as a second training dataset for the model, increasing its robustness. The accuracy of the analyzed model was increased from 12% to 28% when the most relevant features were not present in input data. Interpretation of the ResNet model trained for speaker classification showed that this model is more robust than the proposed AlexNet model. The ResNet model classification is based on the lower frequencies, which is similar to the gender classification. The presence of the relevant time-frequency bins throughout the time confirms the expected behavior of this model and shows that the model is well built and trained neural network. Achieved improvement and results can be extended even more in further works. Potentially making speech classification models more resistant to noise or adversarial attacks, or used for gaining more knowledge of specific models and their behavior.

Acknowledgements

I would like to thank my supervisor Ing. Kateřina Žmolíková for her valuable advice, guidance and patience in improving this work. Computational re-

sources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

References

[1] Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. pages 2912–2920, 06 2016.

[2] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. pages 97–101, 02 2016.

[3] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *CoRR*, abs/1807.03418, 2018.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[5] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, Mar 2019.

[6] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher Anders, and Klaus-Robert Müller. Toward interpretable machine learning: Transparent deep neural networks and beyond, 03 2020.

[7] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 342–350, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[8] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[9] Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. *CoRR*, abs/1904.11829, 2019.

[10] Sebastian Lapuschkin, 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:e0130140, 07 2015.

[11] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham, 2019.

[12] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2020.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[15] Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults. 2, 01 1995.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[17] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[18] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019.

[19] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.