DOI 10.5162/SMSI2020/D2.2

Visualizing Neural Network Decisions for Industrial Sound Analysis

Sascha Grollmisch[1,2], David Johnson[1], Judith Liebetrau[1]
[1] Fraunhofer IDMT, Ilmenau, Germany
[2] TU Ilmenau, Ilmenau, Germany
jsn@idmt.fraunhofer.de

Summary:
Recent research has shown acoustic quality control using audio signal processing and neural networks to be a viable solution for detecting product faults in noisy factory environments. For industrial partners, it is important to be able to explain the network's decision making, however, there is limited research on this area in the field of industrial sound analysis (ISA). In this work, we visualize learned patterns of an existing network to gain insights about the decision making process. We show that unwanted biases can be discovered, and thus avoided, using this technique when validating acoustic quality control systems.

Keywords: Industrial sound analysis, acoustic quality control, machine learning, neural networks, visualization, layer-wise relevance propagation

Background, Motivation and Objective

Neural networks have improved classification systems in audio research fields such as Acoustic Event Detection and Music Information Retrieval. Similar approaches have also been shown to be useful for a acoustic quality control systems [1]. Instead of differentiating between sound events or music genres, the task is to detect machinery and product faults using sounds containing only subtle changes, which are often audible to experienced machine operators. The aim of Industrial Sound Analysis (ISA) is to automatically detect these differences in audio signals within the human auditory range. In [1], the surfaces of metal balls were able to be classified with high accuracy using a deep feed forward neural network (DNN) even with noisy conditions. Even though high classification performance was reported, the decision making process of the DNN was not investigated. Understanding networks' decisions is important for creating explainable classifiers and avoiding potential biases. In this work, we visualize information about a DNN's decision using layer-wise relevance propagation technique (LRP) [2]. Additionally, an artificial bias is added to the dataset to show how such a problem could be discovered and possibly avoided before the quality control system is implemented in real-world production lines.

Visualization Techniques

Several methods have been developed to visualize the decision making process of non-linear classifiers such as neural networks, attempting to make the so called "black box" more understandable. One approach is to propagate the gradient of the resulting class back through the neural network to the input feature space. The state-of-the-art LRP method modifies the backpropagation rules such that the back-propagated signal is weighted with each layers activations to produce less noisy heatmaps [2]. LRP has been shown to be effective in fields such as image recognition for visualizing information about a network's decisions and uncovering unwanted biases in the dataset as well as in the classifier, e.g. identifying that a neural network makes decisions using visible watermarks present only in some images [3].

Dataset and Experiment

For the use case of metal ball surface detection the IDMT_ISA_METAL_BALLS dataset was published together with results from a DNN baseline system [1]. The dataset contains several metal balls with three different surfaces (*eloxed*, *coated*, and *broken*) which pass by a microphone on a metal slide. This sound is automatically recorded and cut to 400 ms for further analysis. The reported baseline system uses a magnitude spectrogram obtained from a STFT as input, and achieved 98.8% file-wise accuracy on a separate test dataset. To visualize the decision making process of the DNN, we concatenated spectral time frames of the test data and overlayed them with corresponding heatmaps obtained using LRP. The heatmaps were extracted using the iNNvestigate

framework [4] with alpha 1 and beta 0. Furthermore, we use the IDMT_ISA_METAL_BALLS dataset to explore the potential for uncovering biases using visualization techniques. For this experiment, we created a modified dataset in which a 10kHz sine wave is added to the *broken* class samples to determine if a classifier uses this bias to make its decisions.

Results

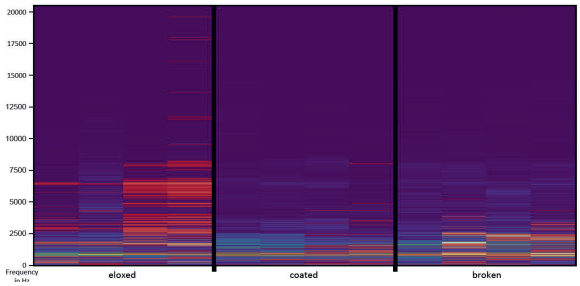

Fig. 1. Decision heatmaps for each class (eloxed, coated, and broken) without bias.

Fig. 1 shows heatmaps for the unchanged dataset, plotted on top of the magnitude spectrograms, indicating the origin of the neural network's decisions in red (slightly important) to yellow (very important). While heatmaps are a valuable tool for image classifiers, they are harder to interpret or validate for spectrograms of industrial sound sources, especially when the correct solution is unknown beforehand. The important frequencies vary slightly for all examples of one class. This makes it hard to identify an easy to understand decision pattern for each class.

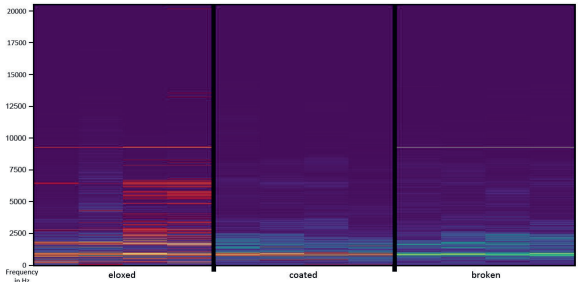

Fig. 2. Decision heatmaps for each class with 10kHz sinusoidal bias in the broken class.

While validating the classifier for complex sound scenarios can be difficult, uncovering unwanted behavior may be feasible using visualization methods. We demonstrate this using a sine wave as an unwanted bias. Adding the sine wave to the *broken* class during training and test improved the baseline accuracy from 98.8% to 99.4%, since that class could be classified perfectly with the added bias. The plots in Fig. 2 show that the model is trained to make decisions for the *broken* class using the added sine wave. Additionally, the absence of energy at 10kHz for decisions regarding the *eloxed* class is also noticeable.

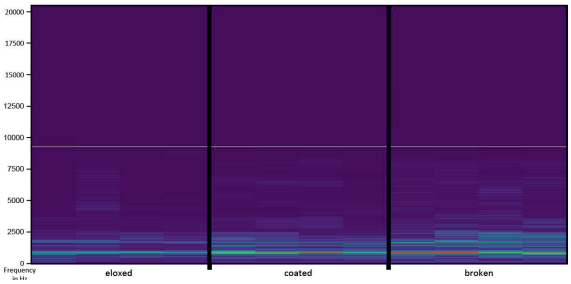

Fig. 3. Decision heatmaps for each class with a 10kHz sinusoidal bias in each class.

By adding the sine wave to all classes in the test set, the accuracy dropped to 33.3%. All files were classified as *broken* showing that our artificial bias worked. Furthermore, it can be seen in Fig. 3 that a bias instead of the actual characteristics of the original audio was picked up by the classifier, providing an explainable reason for the misclassifications.

Conclusion

While neural networks are a promising direction for building acoustic quality control systems, the reason for the classifiers decision are hard to explain due to the non-linearity of the model. State-of-the-art visualization techniques from image recognition research, such as LRP, have the potential to provide insights on the decision making process of the neural network. Compared to natural images it may be difficult to validate the complex frequency patterns which were found. However, our experiments showed that it is a potential method for discovering unwanted biases in datasets. Future work could be to transfer the resulting heatmaps to the audio domain, in addition to the visual domain, to make the decisions audible and possibly easier to interpret.

References

[1] S. Grollmisch, et al., Sounding Industry: Challenges and Datasets for Industrial Sound Analysis, *EUSIPCO*, A Coruna, Spain, 2019

[2] S. Bach, et al., On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLoS ONE 10(7): e0130140* (2015), doi: 10.1371/journal.pone.0130140

[3] S. Lapuschkin, et al., Unmasking Clever Hans Predictors and Assessing What Machines Really Learn, *Nature Communications* (2019), doi: 10.1038/s41467-019-08987-4

[4] M. Alber, et al., iNNvestigate neural networks!, arxiv: 1808.04260

This work has been partially supported by the German Research Foundation (BR 1333/20-1, CA 2096/1-1).