Technische Universität Berlin
Fakultät I - Geistes- und Bildungswissenschaften
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation

Master Thesis Exposé
Visualization of feature maps from convolutional neural nets for an audio-based patient monitoring system

Abgabedatum: 24.10.2021

Betreut von Prof. Dr. Stefan Weinzierl
Dr. Athanasios Lykartsis
Markus Hadrich

Autoren:
Yuchen Wang, 414497

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Abstract

Acoumon is an audio-based acoustic monitoring system, using special devices and deep learning models to detect emergency situations of artificially ventilated patients in intensive care, developed at the TU Berlin. The microphone-array device in the a patient's room detects and emits sounds to indicate the emergency situation to the nursing staff and the model through feedback provides predictions based on continual learning. An initial evaluation of the system shows an accuracy of 80% for a binary, balanced dataset. However, it is still uncertain whether the model can be generalized, because a common problem in deep learning is that an accurate interpretation of the model is very difficult. In this master thesis, visualization techniques will be used to explain a model used in the task of a prototype deep learning system for the acoustic monitoring of intensive care patients and analyze whether it can help us better perform classification. Visualization techniques for neural networks provide a visual explanation of the decisions of a machine learning model. In other words, they highlight the areas in the feature maps which have the greatest contribution to the model decisions. These four visualization methods of Vanilla Saliency, Smooth-Grad Grad-CAM are the three most commonly used visualization methods in deep learning at the moment. Therefore, they will be compared and evaluated for the task, in order to analyze whether they can provide information about how this specific task takes place. Specifically, each layer of a convolutional network will be visualized in order to better understand its function and how it contributes to the final decision of the network. The purpose of this thesis is to find the most suitable visualization method for Acoumon project as well as to find the most useful, most important time-frequency information or patterns for this task.

2 Introduction

2.1 Motivation

In recent years, acoustic monitoring has become more and more important, such as the monitoring and evaluation of audio events, home auxiliary equipment and other applications. But to date, acoustic monitoring for medical applications has not made great progress, mainly due to the instability of audio monitoring for complex medical situations and there is not too much relevant open source data available for training. [4]. Acoumon is an audio-based acoustic monitoring system, The training data set comes from real-time recordings of patients and the spatial and sonic environment, then short-length Mel spectrograms are extracted from the recordings to train the convolutional neural network.This project is based on acoustic monitoring and it an attempt to be used in the medical field. As a tool to make the neural network model transparent, visualization is more important in the case which lack sufficient open source data. There are two more classic methods of visualization, deconvolution and guided propagation (Guided-Backpropagation) Through them, we can ünderstandßome of the features learned by the deeper knowledge layer in the CNN model.

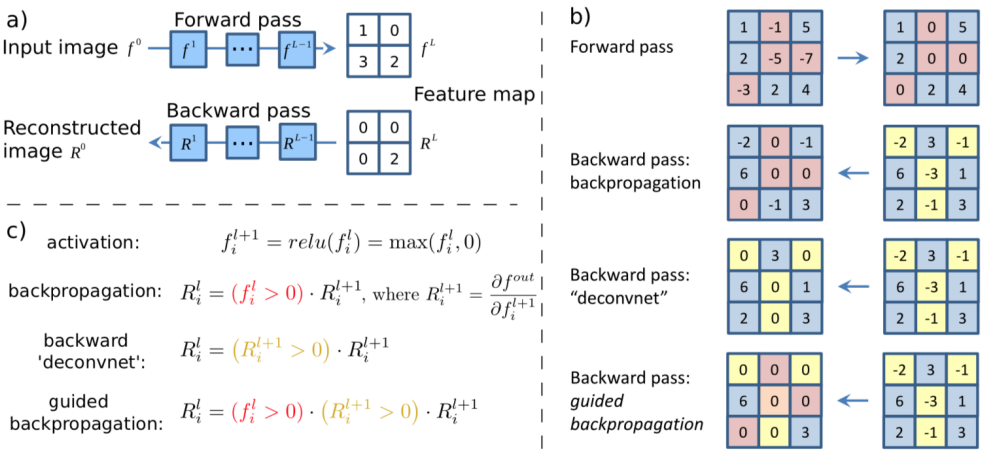

Abbildung 2.1 classic methods of visualization [10]

Interpreting complex deep neural networks models remains a challenge. In field of MIR(music information retrieval), the most cases are image classification systems, because the audio is usually converted to the frequency domain through the fast Fourier transform, and then the obtained spectrogram is used as input for the learning of the neural network. Because of the lack of open source data, this visualization method is even more important. Looking for an explanation of classification decisions may reveal the underlying mechanisms of such systems and help enhance them [9]. For example, it could be helpful to use gradient calculation to mark the value of importance of each pixel value which is to reflect the impact of the pixel on the final classification, this mean, it could help us to understand which frequency band is used as an important basis for judgment and played a role in the classification process of the model. In this master thesis, three of these visualization technologies will be introduced, analyzed and compared in detail.

2.2 State of the Art

Deep learning models have recently emerged as powerful alternatives to traditional methods. Notable examples include [8], where a deep feed-forward network learns to estimate an ideal binary spectrogram mask that represents the spectrogram bins in which the vocal is more prominent than the accompaniment. In [2], the authors employ a deep recurrent architecture to predict soft masks that are multiplied with the original signal to obtain the desired isolated source [3].

However, for the task of acoustic detection, one potential weakness is the lack of large training datasets. It takes a lot of time and manpower to collect data and manually label, so there is a problem of bad generalization ability and Although the data of the prediction results is very impressive, it still cannot 100% prove that the model is very reliable. In this aspect, visualization technology can be very useful, as it is robust to adversarial perturbations[1] and therefore more faithful to the underlying model and help achieve model generalization by identifying dataset bias.

A number of previous works, such as [6] have visualized CNN predictions by highlighting 'important' pixels (i.e., change in intensities of those pixels which have the most impact on the prediction score). Specifically, Simonyan et al. [7] visualize partial derivatives of predicted class scores w.r.t. pixel intensities, while Guided Backpropagation and Deconvolution make modifications to 'raw' gradients that result in qualitative improvements. Despite producing fine-grained visualizations, these methods are not class-discriminative. (visualizations with respect to different classes are nearly identical). Other visualization methods synthesize images to maximally activate a network unit or invert a latent representation. Although these can be high-resolution and class-discriminative, they are

[1] An adversarial example is an example created by performing worst-case perturbation on the input of a machine learning model.

not specific to a single input image and visualize a model overall.

As the literature review here shows, there are studies researching CNNs used in the field of audio signal processing [2] [3], but there is almost no research on the use of CNN for acoustic monitoring. As well as many articles comparing and summarizing visualization technology [6], but it is obvious that there is no research that applies visualization technology to the field of acoustic monitoring which is obviously lacking in data sets and has not been explored much. To the best of our knowledge, in this Thesis, we attempt to address this research gap.

2.3 Thesis aims

In this thesis, the goal is to use visualization methods in order to analyze the behaviour of a convolutional neural network for the classification of patients' sounds as an emergency or not, in order to understand how the network makes this decision and on what characteristics of the sound this decision is based. [4]. In order to make the model more reliable and easier to generalize, it is also important that we build more transparent models that have the ability to explain why they predict what they predict. Furthermore, we attempt to examine whether the visualization technologies that is widely used in the field of computer vision and has obtained very good results are suitable for the field of acoustic monitoring. This will address specific questions, e.g., which parts of a spectrogram the algorithm is looking at to make a successful classification, and which parts of a spectrogram make the greatest contribution for the acoustic detection, whether the frequency domain or the time domain is more important, whether different visualization technologies have different effects and if yes, what, and if the precision of the model will be improved through visualization technology, in other words, a qualitative visual inspection of the feature maps might give clues as to which parts of the spectrum or general sound properties are important for deciding whether there is an emergency or not. The above mentioned questions will be investigated in this master thesis. Through the comparison the most suitable visualization technology in the audio field will be found out and which visualization technology gives us the most information. In order to compare, the faithfulness, trust and class discrimination of these visualization methods will be used as metrics.

3 Methods

3.1 Visualization and Auralization of features

In this work there are three methods will be used for visualization, which will be described below.

3.1.1 Visualization

Vanilla saliency: This is a very popular visualization tool that is used to show why the deep learning model made its final decision and clearly show the input features that led to the decision. In the visualization method, the saliency map is based on gradient calculation, which keeps the complexity low and can run quickly, making visualization simple. In summary, Vanilla saliency tries to detect how much confidence each pixel in the input image has on the classification model.

SmoothGrad: In order to understand the decision of the classification model, a common method is to find the area of the input feature that most contributes to the final classification and highlight it. At the same time, sensitivity maps display a lot of noise and some highlighting pixels seem to be randomly selected. SmoothGrad can help reduce visual noise effects in this case, and it can be combined with other sensitivity map algorithms. The core idea is to take an image of interest, sample similar images by adding noise to the image, then take the average of the resulting sensitivity maps for each sampled image. In [1] it is also found that the common regularization technique of adding noise at training time has an additional 'de-noising' effect on sensitivity maps. [9]

Grad-CAM: Gradient-weighted Class Activation Mapping. As its name suggests, this method uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept [5]. Grad-CAM is a class-discriminative localization technique that generates visual explanations for any CNN-based network without requiring architectural changes or re-training. For Image classification, Grad-CAM lends insight into failures of current CNNs, showing that seemingly unreasonable predictions have reasonable explanations. It can also help in diagnosing failure modes by uncovering biases in datasets. It can also help untrained users successfully discern a 'stronger' network from a 'weaker' one, even when both make identical predictions [5].

Comparison and Evaluation: In order to compare the above mentioned methods and their evaluate their suitability for analyzing convolutional neural network, the evaluating

class discrimination, trust, faithfulness and interpretability [5] of each method will be investigated. Class discrimination means whether the model could still make the same classification judgment between many different classes and select the necessary audio features, in other words, whether this visualization method has a good discriminating power for the class of interest. The trust of a visualization to a model is to compare different prediction explanations and evaluate which seems more trustworthy. For example, whether the extracted visualized feature areas are consistent with the actual audio events of interest (e.g., the voice in a voice separation task) instead of the algorithm focusing on other areas of the spectrogram which are irrelevant for the task. The faithfulness of a visualization to a model is its ability to accurately explain the features learned by the model. This can be done as in [11], where input images from the test set are be point-wise multiplied with the class-conditional saliency maps from the visualization methods to generate explanation maps. The explanation maps will then be used as an input for the trained model and it will be measured how often and how much the confidence of the prediction of the model decreases or increases compared to the prediction of the original image. Additionally, those pixels that are detected to have an impact on the decision will be deleted or added accordingly, so as to detect the corresponding features leading to the class prediction.

4 Preparatory Works

In the preliminary preparations, I participated in projects and courses related to machine learning and became familiar with several common models, such as Naive Bayes Classifier, Decision trees, etc. Through these models I achieved a better understanding of the role and meaning of different layers in the convolutional neural network, and through my involvement as a student assistant in the Acoumon project (code review and debugging parameters) I gained a deeper understanding of how a neural network can be coded with Tensorflow and Keras. In addition, in order to better understand the visualization technology, I have read several related application papers. In the literature research, the principles of each visualization technology and the advantages and disadvantages found after combining examples have been understood.

5 Time Schedule

The following is the specific time schedule of the work, the total length is 6 months.

Preliminary research and planning	01.07.2021-01.08.2021
Adjustment of existing models for visualization purposes	01.08.2021-01.09.2021
Development and evaluation of the visualization methods	01.09.2021-01.10.2021
Writing the thesis	01.10.2021-01.12.2021

Literaturverzeichnis

[1] Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[2] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *ISMIR*, pages 477–482, 2014.

[3] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. 2017.

[4] Hadrich Lykartsis and Weinzierln. A prototype deep learning system for the acoustic monitoring of intensive care patients. 2021.

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[6] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 429–436. Springer, 2015.

[9] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[10] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[11] Haofan Wang, Mengnan Du, Fan Yang, and Zijian Zhang. Score-cam: Improved visual explanations via score-weighted class activation mapping. *arXiv preprint arXiv:1910.01279*, 2019.