Technical Sound Event Classification Applying Recurrent and Convolutional Neural Networks

Constantin Rieder, Markus Germann, Samuel Mezger and Klaus Peter Scherer
Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen, Germany

Keywords: Deep Learning, Sound Analysis, Information Systems, Intelligent Assistance.

Abstract: In many intelligent technical assistance systems (especially diagnostics), the sound classification is a significant and useful input for intelligent diagnostics. A high performance classification of the heterogeneous sounds of any mechanical components can support the diagnostic experts with a lot of information. Classical pattern recognition methods fail because of the complex features and the heterogeneous state noise. Because of no explicit human knowledge about the characteristic representation of the classes, classical feature generation is impossible. A new approach by generation of a concept for neural networks and realization by especially convolutional networks shows the power of technical sound classification methods. After the concept finding a parametrized network model is devised and realized. First results show the power of the RNNs and CNNs. Dependent on the parametrized configuration of the net architecture and the training sets an enhancement of the sound event classification is possible.

1 INTRODUCTION

In the context of intelligent diagnostics, information and intelligent support systems are developed and in use for technical services concerning surveillance of machine components. In addition to 2-dimensional pattern recognition (optical digital image analysis), also 1-dimensional sound analysis is becoming increasingly important because of the information it provides about the interior of a component. Experienced engineers are often able to detect and identify the technical condition (fault or normal) of the components due to the their emission of complex sounds. The main idea is to support the human experts and diagnosticians by a condensed, evaluated information about processes and behaviour of the components. Based on this, an intelligent decision making and fault detection is possible. Once the technical sound events are detected in a very reliable way, intelligent reaction and regulation processes can follow. However, the problem is the audio event classification itself, including the audio event detection. Classical pattern recognition has no chance to recognize the different sound classes, since no explicit features can be described. In this proposal a new method is presented. Based on different training sets neural networks are examined, the different architectures are applied and finally a parametrized convolutional network is conceptualized and realized. The results are evaluated in different steps. Section II describes the general approach used to tackle the problem with Recurrent Neural Networks, in particular the Long short-term memory network. It contains an overview over the corpus used as a basic training set and furthermore an evaluation of the artificial neural networks. In section III a solution approach based on Convolutional Neural Networks is designed and demonstrated. This requires certain transformations of the training data.

2 SOUND EVENT CLASSIFYING USING RECURRENT NNs

The LSTM network (Long-Short-Term-Memory Network) is one of the most popular variant of a Recurrent Neural Network (RNN). It adopts the gated architecture and extends it with the ability to handle the vanishing gradients and is able to learn dependencies which are more extensive. It implements a kind of long-lasting short-term memory (Hochreiter and Schmidhuber, 1997). This fact is responsible for the power of this type of RNNs and achieves good results in the sound event classification. The neurons of the hidden layer are generated from LSTM cells that aggregate the output from four components. These

Rieder, C., Germann, M., Mezger, S. and Scherer, K.
Technical Sound Event Classification Applying Recurrent and Convolutional Neural Networks.
DOI: 10.5220/0009874400840088
In *Proceedings of the 1st International Conference on Deep Learning Theory and Applications (DeLTA 2020)*, pages 84-88
ISBN: 978-989-758-441-1

components of a common LSTM-cell consist of three gates that regulate how much of the data is forgotten, updated and output, and one cell core with linking logic. The input gate determines which values enter the cell, the Forget Gate determines whether information remains or is forgotten and the output gate determines how the remaining values are output. Inside the core of the cell, the information flows are controlled by linking logic to derive the state of the cell.

2.1 Datasets and Preprocessing Steps

The AudioSet (Gemmeke et al., 2017) from Google Research provides the data set for the experimental implementation of the classifications with RNNs and CNNs. In its current version, it consists of over 2 million hand-labelled 10-second clips. The individual clips come from YouTube videos. The labels are taken from the AudioSet ontology developed for this purpose, a hierarchical set of over 600 audio event classes. The ontology covers a wide range of sounds, from the human voice to music, machine sounds, and general environmental sounds. A big advantage of using Google AudioSet (GAS) is that the audio material is commonly available and well prepared for machine learning. The AudioSet offers a compact representation of the audio sources in a CSV format and a set of extracted 128-dimensional audio features (per second of audio recording). These audio features are stored in over 12 thousand TensorFlow Record Files and have a size of about 2.4 GB. The features are stored as Tensorflow sequence example protocol buffers. The context part contains meta information like the video ID, start and end time and the labels contained in the sequence in encoded form. The audio features themselves are also contained in the protocol buffer and are stored as byte lists as 128-bit quantified features. One such byte list is created for every second in the sequence (AudioSet, 2020).

The full range of sounds contained in the GAS is not needed because the focus is on application in a technical and industrial environment. Therefore, in the first step irrelevant sounds like *Human Sounds*, *Animal Sounds* and *Music* were removed in a rough cut. After this class elimination three superclasses are considered (see figure 1), namely *Source Ambigouus Sounds* with 6 subclasses, *Sound of Things* with 13 subclasses and *Channel, Environment and Background* with 3 subclasses. Thus, an n-classification problem must be solved with n=3 at the top level and n=22 at the subclass level.

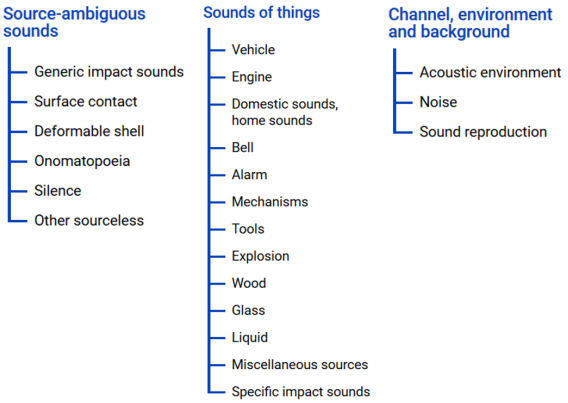

Figure 1: Selected classes from the GAS ontology.

2.2 First Experiments with Used Application Models

In the first experimental setup, different neural networks are used, examined and evaluated for the audio event classification of the above mentioned sound sources. The sound events themselves are represented as time series events over a certain time window. Therefore, the sound events are considered as sample sequences, and the classification task consists in predicting a category for the sequence. The audio features are provided by the AudioSet in frame-level format. They map 10 second blocks at 1 HZ. So the following approaches to frame-level classification with the TensorFlow framework were used for training and classification:

- Deep bag of frames model (Dbof)
- LSTM (Long Short Term Memory) model
- Bidirectional LSTM

These models were preferred because they are suitable for the intended application area according to (Abu-El-Haija et al., 2016) and provide interesting results. For more detailed information and details and advantages of LSTM networks, see (Hochreiter and Schmidhuber, 1997) and for Dbof networks see (Araujo et al., 2018). The relatively strict Top-1 accuracy (Hit@1-Score) was used for the evaluation. This means that the model response (i.e. the one with the highest probability rating) must be the expected response. From the pre-processed data set, all classes were first selected, trained and evaluated. The initial results were moderate and only provided the basis for further adaptations and improvements of the methods used. Problematic aspects were the weak-labeling and parameterization as well as the large number of classes.

2.3 Reduction and Selection of Relevant Corpora

One successful solution to the problem was to reduce the number of classes in a special manner. The number of classes is reduced by categories such as *engine sounds*, *vehicle sounds* and similar others. The restriction of classes showed a significant improvement and the classifier achieved a high top-1 accuracy after a certain amount of training steps. Figure 2 summarizes the evaluation results in a representative manner in relation to the partial corpus for the category engine. Comparable training runs in other categories such as vehicle, mechanisms, tools and other mechanical objects delivered similar results.

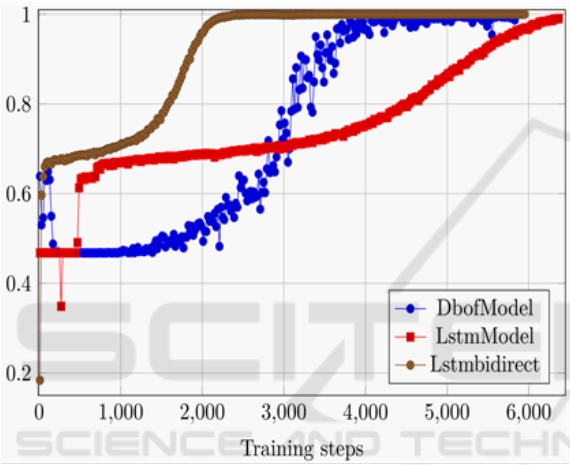

Figure 2: Improvement of results on reduced corpus Engine Sounds.

Considering the results of the runs using the reduced corpora, bidirectional LSTM delivered the best performance in the evaluation, followed by the Dbof model and the LSTM model. Overall, the reduction of the entire corpus to certain categories showed significantly better results than the application to all classes. In the next step of the project, CNNs will be used for classification.

3 SOUND EVENT CLASSIFYING USING CONVOLUTIONAL NNs

CNNs are the method of choice for the analysis of images with deep learning methods. There is extensive research work and well developed networks dedicated to this task. We will use these methods and advantages to implement the use case of technical sound event classification. Considering that CNNs can be very powerful in classifying image data, at first it seems to be inapplicable to use them for classifying sound events, but it show promise for audio classification (Hershey et al., 2016).

This requires a huge amount of training data consisting of images labeled with their corresponding classes. Transfer learning, however, can overcome this barrier by using pre-trained Neural Networks, in our use case CNNs. This can significantly reduce the effort required to train a powerful and complex neural network. This becomes possible by using components of proven and powerful image recognition networks, which have already learned basic and crucial differentiating features and capabilities for image recognition. A variety of powerful CNNs have been developed for image classification, such as Inception, VGG, DenseNet and MobileNet, to mention some of them.

To do this, the previously applied concept for training the model must be changed. The following figure 3 shows an overview of the conceptual scheme.

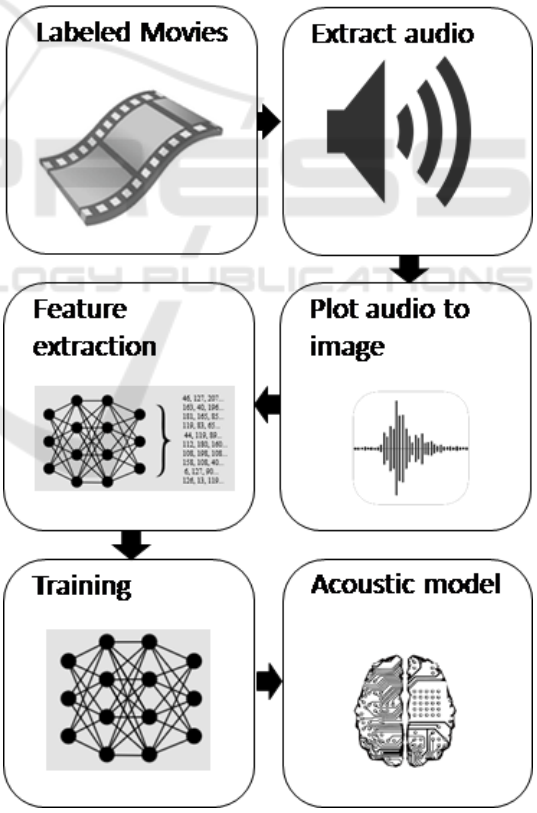

Figure 3: Acoustic CNN Model Training scheme.

The data set for GAS is divided into a balanced training set, a balanced eval set and an unbalanced training set. This split has also been adopted unchanged for the experiments with the CNNs. The unbalanced set was used as the training set and the eval set as the test

set. The structuring of the classes followed the audio set ontology. The biggest changes with respect to the RNNs were in the class selection (subset from the domain) and the raw data. The pre-extracted features of GAS as raw data were not applied and were implemented and generated by a separate feature extraction process (sound to spectrogram to features). For this procedure, the corresponding source files of the sounds are downloaded and converted into the audio format WAV by using the ontology as an orientation. The audio files are split into 10-second chunks in the next step. For the image analysis images are of course needed. For this purpose the created chunks are transformed and plotted into spectrograms. It is expected that audio examples of an individual class will represented by color and shape similarities in different regions of the spectrograms. The remaining part of the neural network must then assign these characteristics to the classes to be learned. The following models are considered for image classification:

- InceptionV3
- MobileNet

The application of the technical sound classification is also extended by the transformation of the sound files to images, as the following scheme shows.

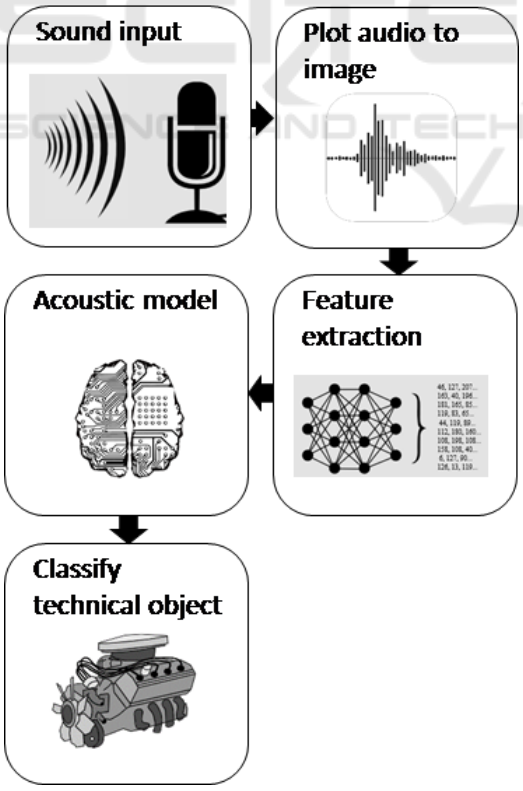

Figure 4: Classification scheme with applying the acoustic model.

3.1 Transformation and Visual Representations of Audio Sets

As the above schemes show, the audio signals are transformed into a visual 2-dimensional representation. These are then used as input for the feature extraction. The extracted features are used to train the CNNs. The feature extraction from the visual representations is also used as input generator for the application of the trained models. The visual representation is generated by a Fast Fourier Transform from the time series into the frequency domain. Spectrograms are generated from the WAV audio files. There are different types of spectrograms that can be generated from audio signals. The following figure 5 shows an extract of four audio chunks from the used technical sounds as MEL-spectrograms.

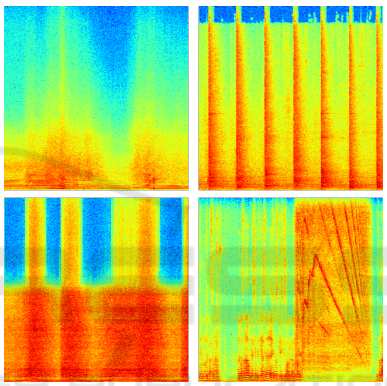

Figure 5: Generated spectrograms.

For the 2-dimensional pattern recognition, i.e. picture analysis, very powerful methods are known and used for classification of 2 dimensional patterns. Also for deep learning algorithms, experiences and effective results can be produced. Because of that the sound patterns are transformed into images with special features. One method consists of a fast Fourier transformation and cepstrum analysys from the time domain into the frequency domain and return. The models must be selected by the learning methods. A parametrized model is developed to guarantee different influences from net parameters to the results. The consequence is a generated information concerning robustness, accuracy and performance. Following parameter set can have an impact on the model generation of the training corpus.

- Type of signal transformation for the spectograms
- Selection of the color map
- Resolution of the spectrogram images
- Scaling & cropping of the images
- Selection of the audio samples / classes

3.2 Multiclass Classification with CNNs

The spectrograms generated for the respective classes now form the new corpora and serve as input for the training of the selected models. For this purpose, the following six classes were deliberately chosen to represent the technical sounds: *Drill, Hammer, Knock, Sawing, Scrape, Clapping*. The multiclass problem for technical sounds was now addressed with the *InceptionV3* model and the *MobileNet* model.

The runs delivered the following results regarding categorical accuracy, as can be seen in figure 6.

The Inception Model on the evaluation set approached the categorical accuracy of 0.5. The training with MobileNet was convincing with a faster run and delivered accuracy values on the evaluation set also above 0.5 .

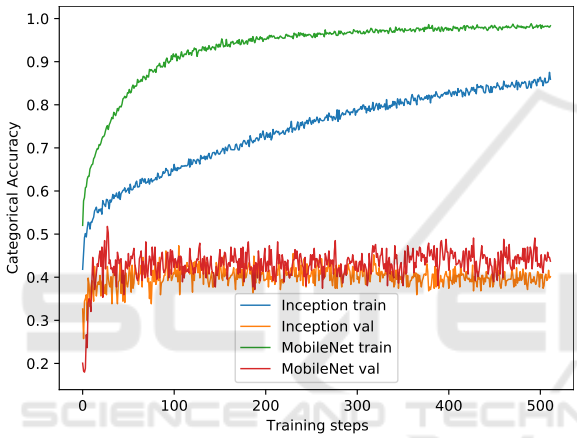

Figure 6: Training of both, the InceptionV3 and the MobileNet model.

4 CONCLUSIONS

The position paper has introduced two technical sound classification approaches that included methods to classify technical sound events by Recurrent Neural Networks and by Convolutional Neural Networks. Experiments on different data sets showed the advantages of the proposed methods over sound data transforming to image data. Most of the sound recordings have been recorded under real conditions. This is a great advantage and corresponds to the use case. However, it is both a difficult situation and a great challenge, because the sound sources were recorded from different environments and distances. They contain other interfering noises and the quality can vary greatly in some cases. One of the most important discoveries so far is the reduction of corpora to selected classes for RNNs, which improved the classification results. In the case of the CNNs the results are still in need of improvement. This will be further pursued in future work. Other models such as *VGG19* and *DenseNet* will be examined and applied. Another possibility could be a combined use of so-called convolutional recurrent neural networks, which is described in (Choi et al., 2017). Other options could be the use of hybrid architectures, which are introduced in (Choi et al., 2017) and (Feng et al., 2017).

ACKNOWLEDGEMENTS

The work presented in this article is supported and financed by Zentrales Innovationsprogramm Mittelstand (ZIM) of the German Federal Ministry of Economic Affairs and Energy. The authors would like to thank the project management organisation AiF in Berlin for their cooperation, organisation and budgeting.

REFERENCES

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675.

Araujo, A., Négrevergne, B., Chevaleyre, Y., and Atif, J. (2018). Training compact deep learning models for video classification using circulant matrices. *CoRR*, abs/1810.01140.

AudioSet, G. (2020). Google AudioSet: A large-scale dataset of manually annotated audio events. Accessed on 01.03.2020.

Choi, K., Fazekas, G., Sandler, M., and Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396.

Feng, L., Liu, S., and Yao, J. (2017). Music genre classification with paralleling recurrent convolutional neural network. *CoRR*, abs/1712.08370.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. W. (2016). CNN architectures for large-scale audio classification. *CoRR*, abs/1609.09430.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.