4.3 Towards Classification of Technical Sound Events with Deep Learning Models

Constantin Rieder[35], Markus Germann[36] and Klaus Peter Scherer[37]

Abstract: Sounds of machines and mechanical systems contain a lot of information about the observed object and its state. Experienced engineers and technical service staff can often identify or classify a certain technical object with state via its sound. An equivalent automated system with such capabilities is difficult to realise because of noisy unknown surroundings. In this paper, we show an approach to implement the mentioned characteristics with deep learning methods and enhance the power of a technical assistance system.

Keywords: Deep Learning, Sound Analysis, Information Systems

Introduction

Information and mobile assistance systems are becoming increasingly important in the context of the digitisation and the development of technical services, such as technical customer service. Generally, these systems are intended to provide the user the appropriate information when carrying out his work, to provide suitable assistance, to support decisions and if necessary, to guide the user. This paper presents one of the emphasized topics from the project MARS (Multimodal Information Systems with Robust Semantics). The main project idea is the development of an information system, which recognizes the desired information need based on different input modes and generates a corresponding reaction. These input modes include natural language queries, visual information and sound signals. With regard to the latter point, the basic idea is to extract features from incoming sound signals or certain sound events from technical equipment and machines to classify the mechanical object or rather to detect the condition of the inspected object such as a machine failure. The assistance system needs the identification of the sound event in order to be able to initiate appropriate support measures regarding the identified technical object.

The realization of the presented idea will use machine learning methods in order to train neural networks to detect and classify a certain technical object or its condition by analysing its sound patterns from an audio recording. The implementation of such a classification procedure needs a large amount of data. Unfortunately, it is very difficult to find

[35] Karlsruhe Institute of Technology, Institute for Automation and Applied Informatics, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany, Constantin.Rieder@kit.edu
[36] Ibidem, Markus.Germann@kit.edu
[37] Ibidem, Klaus-Peter.Scherer@kit.edu
cbe doi:18.18420/provided-by-editor-02

recordings or datasets that cover the technical and industrial sound schemes extensively. For this purpose, we use the AudioSet from Google, a suitable dataset, which provides a relatively large quantity of the desired technical and industrial sound events for training and evaluation [Ge17].

Used Dataset

The AudioSet from Google Research was used for the experimental implementation of the classifications. One of the main advantages of the AudioSet is that the audio material is well prepared for machine learning. The current release consists of over 2 million hand-labeled 10-second clips. The individual clips are from YouTube videos and the labels are taken from the AudioSet ontology, a hierarchical set of over 600 audio event classes. It includes a broad spectrum of sounds ranging from human voice to music and machine sounds to general ambient sounds [Go18].

Data Representation

The AudioSet offers a compact representation of the audio sources in a CSV format and a feature set of extracted 128-dimensional audio features (per second sound recording). These audio features are stored in 12.228 TensorFlow Record Files and are approx. 2.4 GB in size. The features are stored as *tensorflow SequenceExample* protocol buffers. The context part contains meta information such as the video ID, start and end time as well as the labels contained in the sequence in coded form. Furthermore, the Protocol Buffer contains the audio features themselves. These are stored in the form of byte lists as 128-bit quantified features. For each second in the sequence such a byte list is created [Go18].

Data Preprocessing

The focus will be on the application in a technical and industrial environment, therefore irrelevant sounds such as “Human Sounds”, “Animal Sounds” and “Music” were removed in a rough cut in the first step. In the next step of preprocessing, the intersection of relevant and non-relevant entries was included. The relevant entries were separated from the intersection by loading every single TFRecord, comparing the concrete labels with the desired target set, removing the non-relevant units and rewriting the TFRecord File, so that mainly mechanical components were entered as features. In this way, an acceptable subset with a size of approx. 1.2 GB was created from the entire feature set. The feature set coming from the preprocessing will be served as input and processed in the next phase with deep learning procedures to train the artificial neural networks. Several different models are used for this purpose.

Concept

In the first realization steps, a prototypical framework was conceptualized with corresponding modules for different tasks. The audio features module provides the respective labelled audio segments. The acoustic model training module performs the deep learning steps to learn the acoustic model (see Fig. 1).

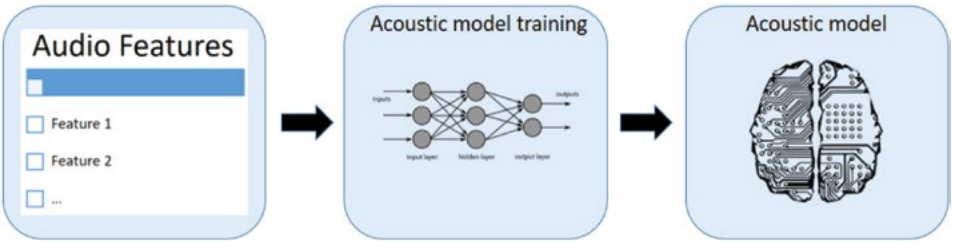

Figure 2: Training the model

From the application point of view, the process begins with a module that delivers a sound segment from an audio recording. The next module extracts the audio features and transfers them to the acoustic model. Now the acoustic model is used to classify the unknown set of features and generates a label accordingly. In the best case, the label matches with the corresponding information from the information system which is then made available to the user (see Fig. 2).

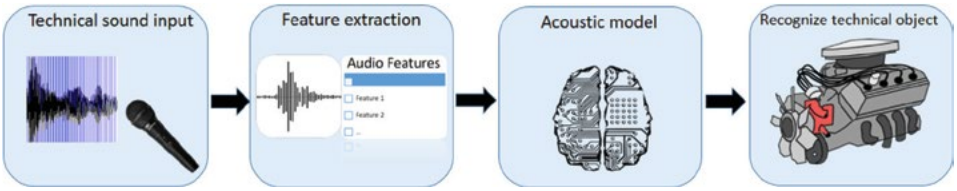

Figure 3: Using the acoustic model to recognize and classify technical objects

Together, the modules presented above form the targeted overall framework for the technical sound recognition. Fig. 3 shows the structure and the general idea of the planned technical sound recognition procedure. The incoming technical sounds are processed by the feature extraction unit, which extract the input features for the sound recognition engine unit from audio waveforms of technical sounds. The sound recognition unit is responsible for the matching of the incoming extracted features. It uses the acoustic model, which has been trained for specific sounds of technical objects before. In this unit, further processing of the result can take place or the assistance services responsible for the recognized object can be initiated.

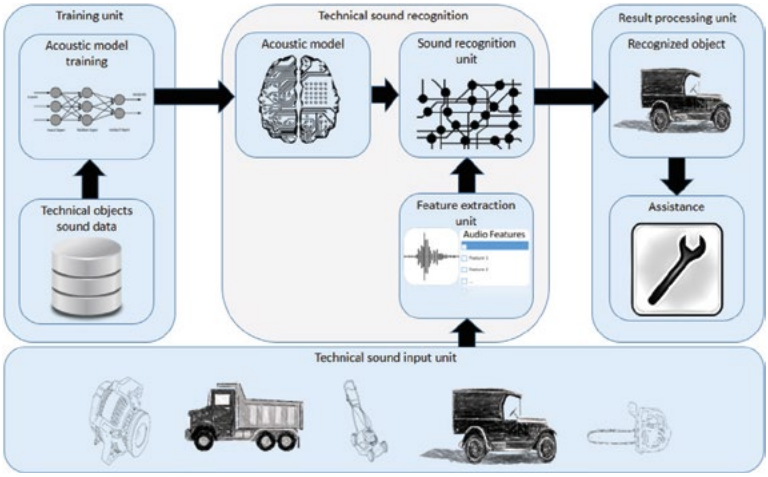

Figure 4: Structure of the mechanical sound recognition system

First experiments and evaluation

In an experimental setup, different neural networks were trained and evaluated for technical sound recognition. Sound events are ongoing events over time. Hence, sound events can be seen as sequences and the task in sequence classification is predicting a category for the sequence. AudioSet provides the audio features in frame-level format representing 10 second chunks at 1 Hz. Therefore following frame-level classification approaches were used with the TensorFlow [Te18] framework:

• Deep bag of frames model (Dbof)

• LSTM (Long Short Term Memory)

• Bidirectional LSTM

The models were selected because they are suitable for the intended field of application and, according to [Ab16], provide promising results. More technical details and advantages on LSTM can be found in [HS97] and on Dbof in [Ar18]. 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) must be the expected answer. Initially, all classes from the pre-processed data set were selected, trained and evaluated. Based on the first results, the used methods should be furthermore adapted and improved. In addition to the parameterization, the main problem was the large number of classes. Another reason could be the weak labeling of the data set. One successful solution to the problem was to reduce the number of classes in a special manner. The number of classes was restricted by categories such as engine sounds, vehicle sounds and 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. Fig. 4 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.

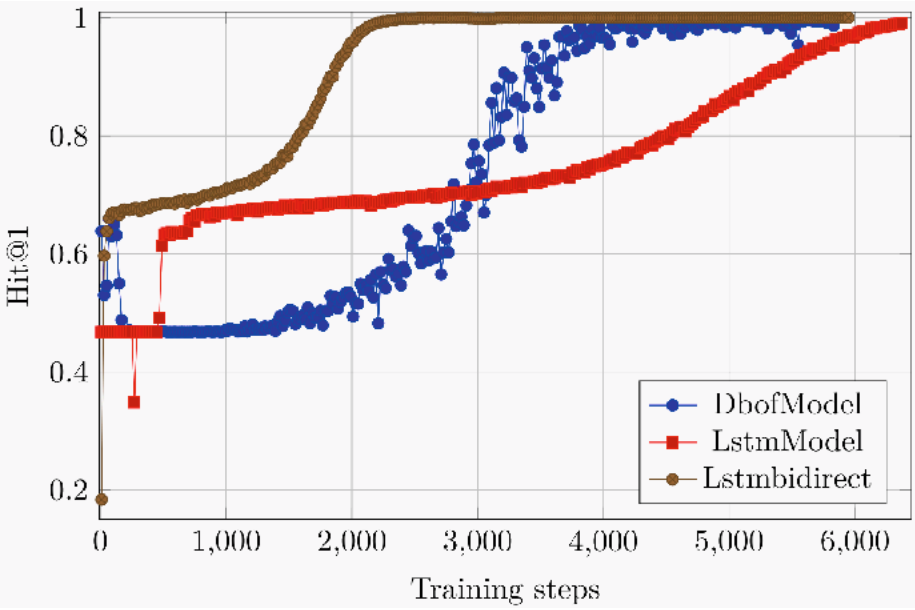

Figure 5: Hit@1 results for category engine

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 steps of the project, the testing of additional models is planned. The use of multi-level attention models could bring more improvements [Yu18]. The use of ResNet with certain adjustments, as shown by Hershey et.al in [He17], is also in consideration.

Conclusions & future work

This paper proposed an approach for audio classification for certain technical sounds, which is a difficult but very interesting problem. With the provision of Audioset by the Sound and Video Understanding teams from Machine Perception research at Google, a good starting position regarding the labeled domain specific data was created. This data set can be used to initiate the investigation of certain sounds such as those of technical objects. The concepts and results presented in this short paper are currently in the initial phase and show the first results and a prototype slice of a subsystem. In the further work steps it is one of the main tasks to further optimize the architecture and to improve the results. Furthermore, additional investigations and experiments are required, on the one

hand on existing methods and on the other hand on other data sets and models.

Acknowledgement

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

References

[Ab16] Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.;Vijayanarasimhan, S.: YouTube-8M: A Large-Scale Video Classification Benchmark. CoRR abs/1609.08675/, 2016, arXiv:1609.08675, url: http://arxiv.org/abs/1609.08675 .

[Ar18] Araujo, A.; Négrevergne, B.; Chevaleyre, Y.; Atif, J.: Training compact deep learning models for video classification using circulant matrices. CoRR abs/1810.01140/, 2018, arXiv: 1810.01140, url: http://arxiv.org/abs/1810.01140 .

[Ge17] Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; Ritter, M.: Audio Set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP 2017. New Orleans, LA, 2017.

[Go18] Google AudioSet Developers: Audioset, accessed: 12.12.2018, Dec. 2018, url: https://research.google.com/audioset .

[He17] 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.; Wilson, K. W.: CNN architectures for large-scale audio classifi cation. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)/, pp. 131–135, 2017.

[HS97] Hochreiter, S.; Schmidhuber, J.: Long Short-term Memory. Neural computation, pp. 1735–80, Dec. 1997.

[Te18] Tensorflow Developers: Tensorflow, accessed: 12.12.2018, Dec. 2018, url: https://www.tensorflow.org/ .

[Yu18] Yu, C.; Barsim, K. S.; Kong, Q.; Yang, B.: Multi-level attention model for weakly supervised audio classification. In: DCASE2018 Workshop on Detection and Classification of Acoustic Scenes and Events. 2018, url: http://epubs.surrey.ac.uk/849626