ANOMALOUS SOUND DETECTION BASED ON ATTENTION MECHANISM

Hayato Mori
Gifu University
Gifu City, Japan
email: mori@asr.info.gifu-u.ac.jp

Satoshi Tamura
Gifu University
Gifu City, Japan
email: tamura@info.gifu-u.ac.jp

Satoru Hayamizu
Gifu University
Gifu City, Japan
email: hayamizu@gifu-u.ac.jp

***Abstract*—For the automation of maintenance of mechanical facilities and devices, anomalous sound detection from machines has been explored. For these years, methods by machine learning and deep learning have been proposed for anomaly detection in various fields. Some deep-learning-based works calculate an anomaly score based on reconstruction errors obtained from an autoencoder model. However, the performance may not be sufficient, depending on characteristics of machines. In this study, we propose a method for detecting anomalous sounds using an autoencoder model with an attention-based mechanism. Given multiple frames of the log-scale mel spectrogram with a missing frame, our model computes the reconstruction error between an predicted frame and the removed frame as an abnormal score. We conducted experiments to compare our scheme to conventional ones, with visualizing attention weights. Our method achieved better performance, and it is found the missing frame can be well predicted using surrounds frames emphasized by the attention model. It is also found our approach can perform well independent on kind of machines and the number of input frames.**

I. Introduction

Mechanical products are indispensable in manufacturing sites and applications, which produce a lot of benefits in our daily lives. Once a machine malfunctions, it causes major accidents and economic damages. To prevent machine breakdowns, technicians with expert knowledge perform regular maintenance. However, keeping mechanical facilities well activated by skilled maintenance technicians requires a lot of costs. Managers want to know whether any machine failure is predicted shortly or not, to conduct the maintenance with the minimum cost in the most suitable timing. Unfortunately, it is still challenging to determine and deal with anomalous condition in the machine in advance.

In recent years, technology has been developed to automatically detect failures from machine drive sound and vibration using machine learning technology [1]. Several machine-learning-based schemes have been proposed, and many successful approaches have employed deep-learning techniques. One of common methods for anomaly detection using deep learning is to use an AutoEncoder (AE) [2]. The AE model consists of two parts: an encoder part and a decoder part. An encoder part converts an input vector into a compact representation, while a decoder part tries to reconstruct the same signal as input data from the representation. In those deep-learning-based schemes, an AE model is firstly created using a training data set only including normal signal data. Because the model can correctly reconstruct normal data, the reconstruction error between input normal data and generated data from the model should be small. On the other hand, the model fails to reconstruct any abnormal data, resulting a higher reconstruction score. Thus, abnormal data can be detected by measuring the reconstruction error between input and output data, as the anomaly score for the created model [3].

In this paper, we propose an attention-based encoder-decoder model for anomalous sound detection. In our model, an input feature consisting of consecutive frames of a log-scale mel spectrogram is put into the encoder, in which a particular one frame data is excluded. A context vector is then obtained from the attention mechanism. The excluded frame is subsequently used as an input to the decoder, utilizing the context vector. The reconstruction error between the predicted frame and the original frame is calculated as the anomaly score. By employing the attention mechanism, important frames for prediction can be selected with larger weights, so that we can make effective prediction for various types of machines in noisy environments. By visualizing the attention weights of the input frames for prediction, we can examine the relationship with the surrounding frames and the excluded one. Experimental results show that by adopting the attention mechanism we achieved improvements for several types of machines in terms of anomaly detection performance. From visualization results of attention weights, we also found that our models could adapt the characteristics of the machine.

The rest of this paper is organized as follows. Section II introduces related works to our research. Our method is described in Section III. Experimental setup, result and discussion appear in Section IV. Finally Section V concludes this paper.

II. Related work

Since it is not possible to collect large amounts of abnormal data for anomalous sound detection in the real world, basically we choose a strategy to create a detection model by unsupervised learning using only normal data. In recent years, AE and variational autoencoder have been used as methods based on deep neural networks. Employing acoustic features such as spectrograms or log-scale mel spectrograms as an input, model training is done so that a reconstruction error between the input and a model output could be small. As the reconstruction error, the mean square error is often adopted. The reconstruction error of normal data, which are similar to the training data, should be smaller, while the reconstruction error of abnormal data would be expected to become larger. The reconstruction error between the input and the predicted output is thus calculated as the anomaly score. This strategy can detect anomalous sounds with a high accuracy, on the other hand, in the case of non-stationary sounds that contain pulse waves, the prediction of the target sound becomes more

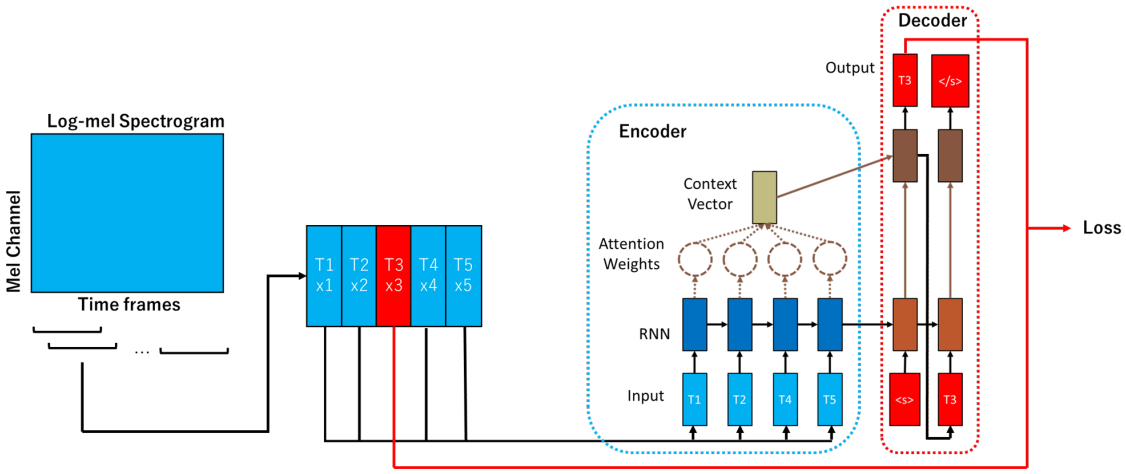


Fig. 1. A flow of our propose method.

difficult and the reconstruction error tends to be high even for normal data.

An et al. proposed an anomaly detection method using the reconstruction probability from the variational autoencoder [4]. Anomaly detection by the reconstruction probability outperforms autoencoders and principal component-based methods. Experimental results show that the proposed method outperforms autoencoder based and principal components based methods. Utilizing the generative characteristics of the variational autoencoder enables deriving the reconstruction of the data to analyze the underlying cause of the anomaly. The structure of the encoder-decoder model was used to detect abnormal images.

Interpolation Deep Neural Network (IDNN) [5] is a model that takes multiple frames of a log-scale mel spectrogram, in which the central frame is removed, as an input and predicts the interpolation of the removed frames as an output. The model consists of a neural network with fully connected layers, and can make predictions that take into account information from surrounding frames. Anomalous sounds are detected based on the reconstruction error, which is the difference between the predicted frame and the true frame. It is hypothesized that the prediction of the central frame rather than the edge frames would make the reconstruction error consistent with the anomaly. Their experiments show significant improvements compared to AE in detecting abnormalities in non-stationary machine sounds.

III. Method

In this study, we employ an attention-based model to detect anomaly for machine sounds. Similar to the above related work, the model tries to reproduce the excluded frame of consecutive acoustic frames, and calculate the anomaly score. In this section, we describe an attention mechanism and explain the architecture of our proposed method for detecting anomalous sounds.

A. Attention mechanism

An attention mechanism is an architecture proposed for machine translation and has been used for various tasks in image and speech processing fields [6] [7]. The mechanism is often employed with Recurrent Neural Networks (RNNs). Given each input vector, an RNN model calculates inner-state vectors, which are used to obtain attention weights. Let us denote an output frame index by t, and an input frame index by u. An attention weight α_{tu} can be then expressed as:

$$\alpha_{tu} = \frac{\exp(\text{score}(\boldsymbol{h}_{t-1}, \boldsymbol{s}_u))}{\sum_{u'=1}^{U} \exp(\text{score}(\boldsymbol{h}_{t-1}, \boldsymbol{s}_{u'}))} \tag{1}$$

$$\text{score}(\boldsymbol{h}_{t-1}, \boldsymbol{s}_u) = \boldsymbol{a}^\top \tanh(\boldsymbol{W}_1 \boldsymbol{h}_{t-1} + \boldsymbol{W}_2 \boldsymbol{s}_u) \tag{2}$$

where $\boldsymbol{h}_t$ indicates a hidden state vector in a decoder, $\boldsymbol{s}_u$ represents a hidden state vector in an encoder, and $\boldsymbol{a}$ is the weight vector. The attention weights indicate which vector is important or should be focused on for prediction among the input vectors. A context vector is subsequently obtained, by summing up hidden state vectors in the encoder with corresponding attention weights. The context vector is finally used for prediction.

B. Proposed method

Similar to conventional schemes, in our approach, only normal sounds are chosen as training data, in order to create a model with a distribution of the data set. At first, among acoustic features extracted as log-scale mel spectrograms, multiple consecutive frames $\{\boldsymbol{x}_{d+1}, \boldsymbol{x}_{d+2}, \ldots, \boldsymbol{x}_{d+i}, \ldots, \boldsymbol{x}_{d+n}\}$ are focused on, where d is an arbitrary starting frame number. One particular frame $\boldsymbol{x}_{d+i}$ is excluded from those frames, while the others are utilized as input frames to the attention encoder. Consequently, the number of input frames is $n-1$. The excluded frame $\boldsymbol{x}_{d+i}$ is used as an input for the decoder when training it and estimating an anomaly score. Figure 1 shows an overview of the proposed method when $n = 5$ and $i = 3$. The first frame (<s>) and the last one (</s>) in the decoder section are zero vectors with the same dimension as one frame. In the decoder, we employ Gated Recurrent Units (GRUs) [8] as an RNN layer. The decoder is trained by feeding the correct token at each input. The whole loss function is designed to measure a mean squared error between predicted frames and true frames. The loss function is given as follows:

$$\text{Loss} = \frac{1}{N-1} \sum_{k=1}^{N-1} (\boldsymbol{x}_k - \boldsymbol{y}_k)^2 \tag{3}$$

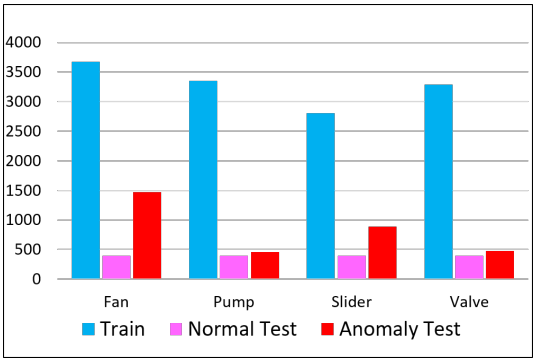


Fig. 2. The numbers of files in the data set.

TABLE I
CONDITION OF LOG-SCALE MEL POWER CALCULATION.

Condition	Value
Data length	10 sec
Sampling Frequency	16 kHz
Frame length	64 msec (=1024 points)
Frame shift	32 msec (=512 points)
Window function	hanning
Mel filter bank	128

where $\boldsymbol{y}_k$ is a k-th predicted vector, and N represents the number of decoder time steps.

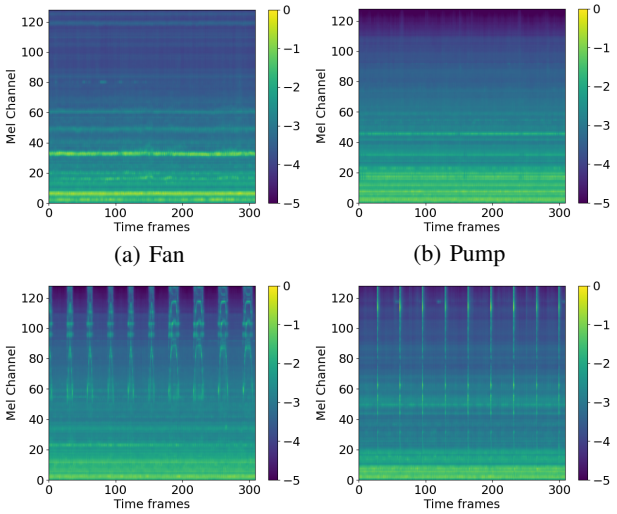
The calculation of an anomaly score for test data is done as follows. First, consecutive frames in a test set is put into the trained attention model encoder, to predict a missing frame. Second, a mean square error between the predicted frame and the excluded true frame is calculated as a reconstruction error. This step is predicted for all frames, and the average of the reconstruction errors is calculated as the anomaly score of the data set.

IV. EXPERIMENT

We conducted an experiment on anomalous sound detection using the method described above. The purpose of this experiment is to investigate the relationship between the surrounding frame and the anomalous sound detection. We extended the interpolation of the IDNN method using explicit attention, and investigated the relationship between the attention weights and the characteristics of machine sounds by visualizing them, while maintaining the accuracy.

A. Data

To evaluate the proposed method, we conducted experiments using the MIMII data set [9]. This data set is the same as the one used in the DCASE2020 Challenge Task2 [10]. This data set is designed to assist machine-learning and signal-processing communities for automated facility maintenance. The corpus contains normal and anomalous sounds of machines. Normal sounds were recorded for different types of industrial machines (valve, pump, fan, and slider), and to resemble a real-life scenario, various anomalous sounds were recorded (contamination, leakage, rotating unbalance, and rail damage). The sound data is characterized by the addition of factory noise, which is close to the real environment. Figure 2 shows the number of files for each machine. Each data was recorded at 16 kHz sampling rate, having 10-second length.



(a) Fan (b) Pump (c) Slider (d) Valve

Fig. 3. Examples of spectrograms.

TABLE II
DEEP NEURAL NETWORK ARCHITECTURE OF AUTOENCODER.

Layer	Units
Input	640
Fully Connection	128
Fully Connection	128
Fully Connection	128
Fully Connection	128
Fully Connection	8
Fully Connection	128
Fully Connection	128
Fully Connection	128
Fully Connection	128
Fully Connection	640

TABLE III
DEEP NEURAL NETWORK ARCHITECTURE OF IDNN.

Layer	Units
Input	512
Fully Connection	64
Fully Connection	32
Fully Connection	16
Fully Connection	32
Fully Connection	64
Fully Connection	128

TABLE IV
AUC OF AE, IDNN AND PROPOSED METHOD.

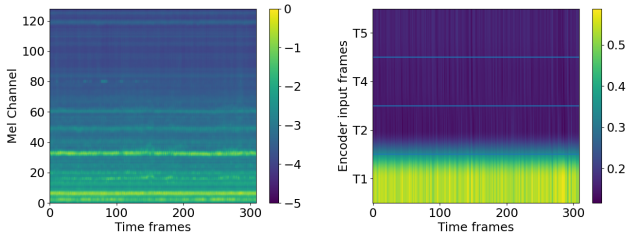
Algorithm	Fan	Pump	Slider	Valve
AE	65.83	72.89	84.76	66.28
IDNN	66.52	72.89	86.04	87.36
Proposed method	66.59	73.02	93.34	94.68

B. Feature extraction

We used log-scale mel power coefficients as an input feature. Table I shows details of feature extraction, and Figure 3 depicts spectrograms. The horizontal axis indicates time (frame index) and the vertical axis is mel channel (frequency). The fan and pump almost consist of stationary mechanical sounds, while the slider and valve have mechanical sounds of non-stationary movement with periodic pulse waves.

C. Experimental setup

In this study, training models were created for each of the four types of machine sounds, and their performance was tested using the test data. We used the architecture illustrated in Figure 1. Four frames except the central frame were composed as an input data to the encoder. The decoder used the sequence of three frames, in which the first and last frames were zero vectors while the second one corresponds to the central vector. We used the area under the receiver operating characteristic (ROC) curve, that is AUC, as an evaluation metric.



(a) Reconstructed spectrogram (b) Attention weights

Fig. 4. Results of normal fan data.

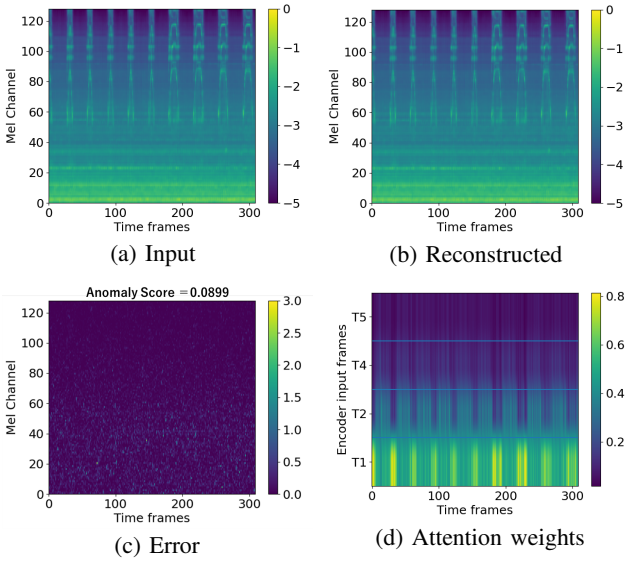
As a comparison to the proposed method, experiments were also conducted for AE and IDNN. Table II and III show the network structures of AE and IDNN, respectively. AE is the same system as the baseline of DCASE2020 task2. For the autoencoder, the block having consecutive five frames was used. Given the same block as our proposed scheme, IDNN predicted the central frame. In both AE and IDNN, the activation function was ReLU, and on each layer batch normalization was applied. Both loss functions for model training were based on mean squared errors. In all the experiments, Adam [11] was used as the optimizer.

D. Result and discussion

Table IV shows the AUC of AE, IDNN and our proposed method for calculating the anomaly score in the test data. Compared to AE, IDNN and our proposed method improved the AUC for all the machine sounds. In particular, the AUC of slider and valve, which have machine sounds with non-stationary movement, have been greatly improved. Furthermore, the proposed method further improved the AUC of slider and valve by about 7% compared to IDNN. The system with the top score for Challenge Task2 had a higher AUC than the proposed method, but used ensemble models with elaborate tuning. The proposed method had a high AUC for a single model and was not complex, so our model can be applied to various machine sound data.

Figure 4 illustrates results for normal fan data with stationary rotation; (a) is a reconstructed spectrogram, and (b) depicts attention weights in which the horizontal axis indicates the predicted frame index and the vertical one is input frame index ($i = 1, 2, 4, 5$). As seen in the figure, only the first frame had a concentrated weight, while the other frames had smaller weights. The reason for this was that the machine is stationarily moving, and there was no significant difference between frames. Therefore the central frame ($i = 3$) can be predicted almost only using the first frame. Another stationary machine, i.e. pump, showed the same trend in attention weights as fan.

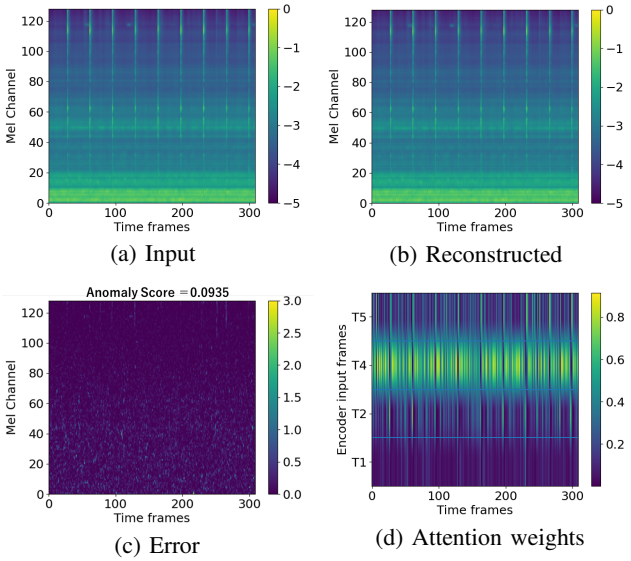
Figures 5 and 6 show results of normal slider and normal valve, having non-stationary movement; an input spectrogram is (a) while an output one is (b), and (c) indicates reconstruction errors in each mel channel; attention weights are illustrated in (d). Obviously, we can see different characteristics from the attention weights for stationary rotation, i.e. fan and pump. In contrast to those machines, the attention weights was not strongly concentrated in any particular frame. This indicates that the prediction was made using incorporating information from surrounding frames. It is observed, in the slider data the weights for the frames before the prediction frame had larger values, and in the valve data weights for the following frames had relatively higher scores. It is also



(a) Input (b) Reconstructed

(c) Error (d) Attention weights

Fig. 5. Results of normal slider data.



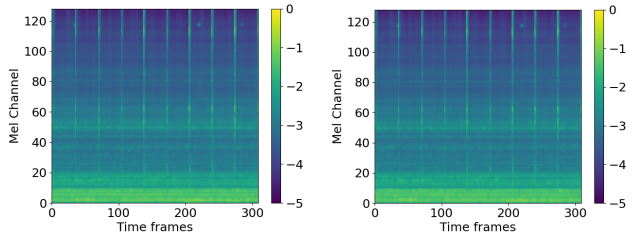
(a) Input (b) Reconstructed

(c) Error (d) Attention weights

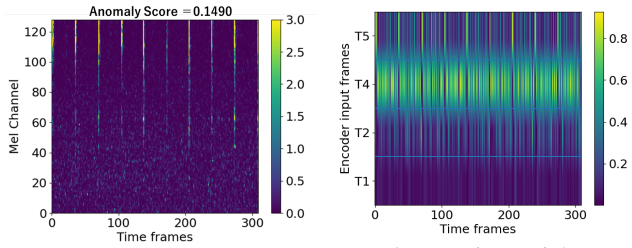
Fig. 6. Results of normal valve data.

confirmed that the attention weights were affected in the area where the pulse wave occurred. It is thus considered that the machine sound of non-stationary movement is able to be well modeled using the features of normal data, by taking into account the information of surrounding frames. In other words, employing the attention mechanism enables us to model normal data much better compared to conventional schemes such as IDNN that used all input frames fairly, and independent on the kind of machines.

Figure 7 indicates results for abnormal valve data. As shown in (c), the abnormal data has a high anomaly score because the part of pulse wave was not reconstructed well. The attention weights were distributed over the entire frame compared to the normal data, which means the difficulty of prediction. Consequently in machines having non-stationary rotation, it is found that the reconstruction error with respect to normal data became small by using the features of the frame that were important based on the attention scheme, while the error

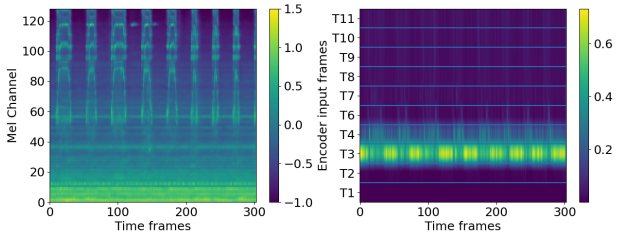


(a) Input (b) Reconstructed



(c) Error (d) Attention weights

Fig. 7. Results of anomalous valve data.



(a) Reconstructed spectrogram (b) Attention weights

Fig. 8. Results of normal slider (10-frame input)

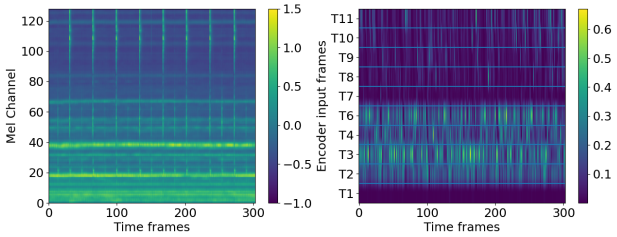
drastically increased for abnormal data.

E. Changing the number of input frames

Based on the above results, the attention mechanism is found to be able to extract important frames for prediction. In order to confirm the effectiveness of the attention, we conducted another investigation by increasing the number of input frames. We set $n = 10$ and carried out the same experiments. Figures 8 and 9 show results for output spectrogram and attention weights of slider and valve, respectively. For non-stationary rotating machine sounds, it is found that the attention weights tended to have larger values near the frame that should be predicted, even when the number of input frames increased. In the stationary rotation (fan and pump), the attention weight was concentrated in the first frame as before. We also found the AUC was almost the same accuracy as the previous experiments. It is thus concluded that the attention structure can be used to properly learn the distribution of sequence of normal data even when the number of frames increases.

V. Conclusion

We proposed an unsupervised deep-learning autoencoder-based method for anomalous sound detection in machines, employing an attention model. Multiple frames from which the central frame was excluded were used as an input for an encoder, and the central frame was used for a decoder, to calculate a reconstruction error as an anomaly score. Experimental results show that the AUC of our approach was



(a) Reconstructed spectrogram (b) Attention weights

Fig. 9. Results of normal valve (10-frame input)

better than those of conventional methods for machine sounds of non-stationary movements. By visualizing the attention weights, we found that the information of the surrounding frames is important when the model predicts the excluded frame for machine sounds of non-stationary movements. Considering the characteristics of the attention mechanism, we tried to increase the number of input frames and still achieved adequate performance.

As our future works, it is necessary to study the effects of the number of frames to be input and the time width of the frames. The use of other models such as transformer [12], which is used in natural language processing, is also our future task.

References

[1] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, and Noboru Harada, “Optimizing acoustic feature extractor for anomalous sound detection based on neymanpearson lemma,” Proc. 25th European Signal Processing Conference (EUSIPCO), pp.698–702, 2017.

[2] Baldi, P., ”Autoencoders, unsupervised learning, and deep architectures, ” Proc. International Conference on Machine Learning (ICML), pp.37–49, 2012.

[3] Chalapathy, R., Chawla, S., “Deep learning for anomaly detection: A survey”, arXiv preprint, arXiv:1901.03407, 2019.

[4] Jinwon An and Sungzoon Cho, “Variational autoencoder based anomaly detection using reconstruction probability,” Special Lecture on IE, vol. 2, no. 1, 2015.

[5] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, Y. Kawaguchi, “Anomalous Sound Detection Based on Interpolation Deep Neural Network,” Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020.

[6] Bahdanau et , ”Neural Machine Translation by Jointly Learning to Align and Translate”, Proc. International Conference on Learning Representations (ICLR), 2015.

[7] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, “Effective Approaches to Attention-based Neural Machine Translation,” arXiv preprint, arXiv:1508.04025, 2015.

[8] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” the Empiricial Methods in Natural Language Processing (EMNLP), 2014.

[9] Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi, “MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection,” arXiv preprint, arXiv:1909.09347, 2019.

[10] DCASE2020 Challenge Task2 “Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring,” http://dcase.community/challenge2020/task-unsupervised-detection-of-anomalous-sounds

[11] Diederk P.Kingma,Jimmy Lei Ba, ”Adam: A Method for Stochastic Optimization”, arXiv preprint, arXiv:1412.6980, 2014

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, ”Attention Is All You Need”, 2017, 31st Conference on Neural Information Processing Systems (NIPS), 2017