Anomalous Sounds Detection Using A New Type of Autoencoder based on Residual Connection

Technical Report

Yunqi Chen [1], *Yuheng Song* [2], *Ting Cheng*[3]

University of Electronic Science and Technology of China, Chengdu, China
[1] chenyunqi@std.uestc.edu.cn
[2] 2018011211019@std.uestc.edu.cn
[3]citrus@uestc.edu.cn

ABSTRACT

This report describes our submission for task2 (Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring) of the DCASE 2020. In this report we propose networks of fully connected autoencoder based on residual connections, which can increase the accuracy of anomaly sound detection. As for data preprocessing, we use data augmentation methods to generate more data from existing data. Our feature extraction is still carried out with log mel spectrogram. Finally, our method has achieved average AUC of 0.7912 and average pAUC of 0.6105 on the development dataset.

***Index Terms*—**DCASE 2020, autoencoder, fully connected, residual connection, data augmentation.

1. INTRODUCTION

Anomaly sound detection plays an important role in our production and life. It can detect anomaly sounds that people cannot detect in time and remind people in time. Because of its characteristics, it can be widely used in many places, especially when it is used to detect whether the machine is faulty, it can timely detect the faults that occur during the operation of the machine, and remind workers to promptly check, which can avoid many dangers caused by machine failure in time.

The use of unsupervised learning for anomaly sound detection has many advantages. We can easily collect a large amount of normal sound data and input it to the neural network for training. Although we can also collect abnormal sound data, the rarity, randomness and diversity of abnormal sound data make it difficult for us to collect all abnormal sound data. Therefore, in this report we will propose a method based on fully connected residual autoencoder for abnormal sound detection

2. DATA PREPROCESSING

2.1. Sound Data Augmentation

Data augmentation helps to generate synthetic data from existing data set such that generalization capability of model can be improved. Firstly, we add white noise with the standard deviation of 0.005 to the origin sound data to generate the noisy sound data. Secondly, we shift the sampled signal to the right by 1600 points to generate the time shifted sound data. Thirdly, we increase the speed of our sound signal by 1.2 times to get the frequency shifted sound data. Finally, we combine the origin sound data, noisy sound data, time shifted and frequency shifted sound data as our data inputs.

2.2. Acoustic Feature Extraction

The sampling rate of each normal sound data is 16 kHz. The log mel spectrum is extracted as sound features. For detailed parameter settings, the number of frames is 5; the number of mel filter banks is 128; the FFT length is 1024; the FFT shift length is 512.

3. NETWORK STRUCTURE

There are some researches [1], [2], in which residual autoencoders has been proposed. Based on their research, we design 4 residual connection autoencoder as our network to finish anomaly sound detection inspired by the Resnet.

For the reason why we design these structures, it is because that as the depth of the network rises, the effect of the network gradually reaches saturation. The residual connection can solve this problem. Furthermore, the case of the AE, one is trained to minimize the reconstruction error of the normal training data, and the anomaly score is calculated as the reconstruction error of the observed sound. Thus, the AE provides small anomaly scores for normal sounds. However, it gives no guarantee to increase anomaly scores for anomalous sounds. Indeed, if the AE is generalized, the anomalous sounds will also be reconstructed and the anomaly score of anomalous sound will be small [3]. So we also choose to uses residual connection in the encoder part to try to increase anomaly score. Then I will describe these three models in detail and give the corresponding structure

Firstly, we depict a two-iteration architecture, with the goal of the first iteration being to encode the original input and the goal of the second iteration being to encode the residual from the first level's reconstruction. And the output of the autoencoder is the sum of the first level's reconstruction and the second level's reconstruction. The structure of this model is shown in figure 1. We named it iteration-residual model.

Secondly, we use residual connection before the bottleneck layer. The first Dense layer is connected to the fourth Dense layer, the second Dense layer is connected to the fifth layer, the third layer is connected to the sixth layer. After the bottleneck layer, we use 6 Dense layers to constitute an autoregressive model as our

decoder. The structure of this model is shown in figure 2. We named it encoder-residual model.

Thirdly, we tried to fuse the first model with the second model. Differently we only use residual connection before each level's bottleneck layer to reduce network complexity, which will be benefited to our final accuracy. The two structures of this model is shown in figure 3 and figure 4. We named it mix-residual-13 model and mix-residual-1324 model.

Batch normalization (BN) is included in both the models to accelerate the learning process and improve the baseline level by regularization term. BN has been applied in most of the recent network architectures and been explained to be incompatible with dropout. Therefore, we have decided not to adopt the dropout.

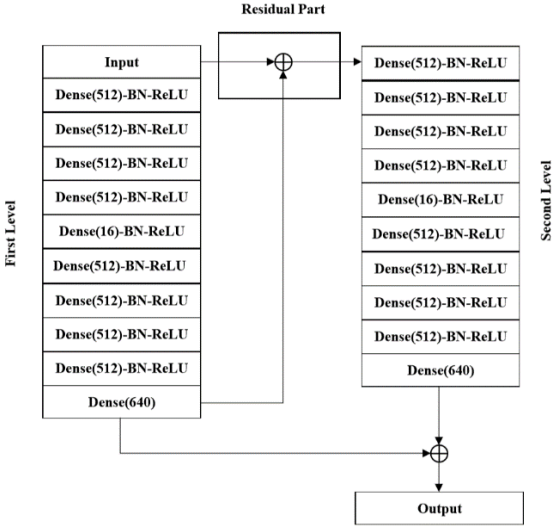

Figure 1: Iteration-Residual Model

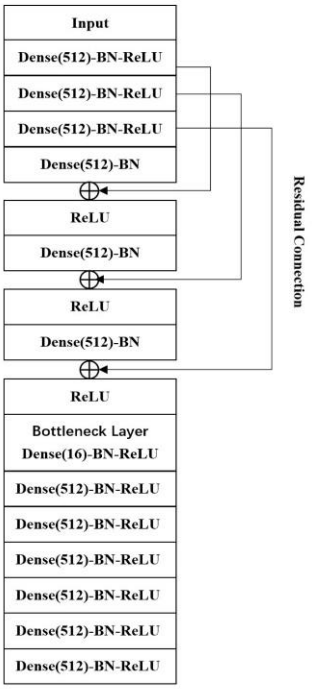

Figure 2: Encoder-Residual Model

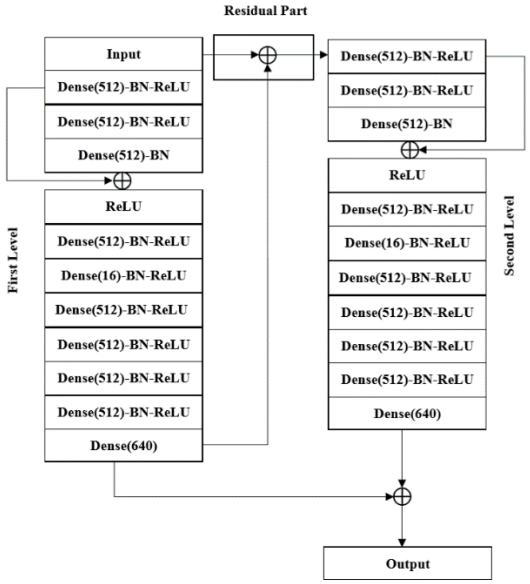

Figure 3: Mix-Residual-13 model

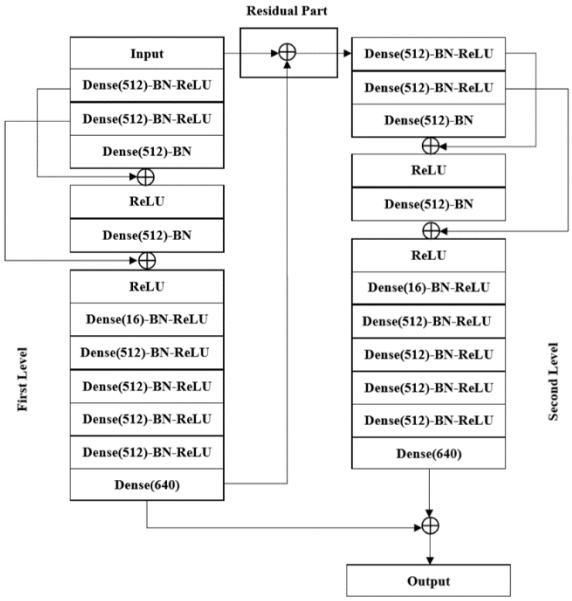

Figure 4: Mix-Residual-1324 model

4. EXPERIMENTS

4.1. Datasets

The datasets used in the training model is the development training dataset, which contains 6 machine types. Each machine type has three or four machine IDs. Each machine ID's dataset consists of around 1,000 samples of normal sounds. Each sample is a single-channel 10-sec length audio that includes both a target machine's operating sound and environmental noise.

4.2. Training Procedure

We choose the adaptive moment estimation as our optimizer with the learning rate of 0.001. The loss function is the mean square error (MSE) and the contractive loss which is used in contractive autoencoder [4]. The MSE and contractive loss are used adjointly in machine types of "toy-conveyor" and "pump", while the MSE is used solely in the other machine types.

The idea of contractive autoencoder is to make the learned representation to be robust towards small changes around the training examples. It achieves that by using different penalty term imposed to the representation. As for its loss function we need to calculate the representation's Jacobian matrix with regards of the training data. The loss function is as follows:

$$L = \left\| XX - \hat{X} \right\|_2^2 + \lambda \left\| J_h(X) \right\|_F^2, \quad (1)$$

where

$$\left\| J_h(X) \right\|_F^2 = \sum_{ij} \left(\frac{\partial h_j(X)}{\partial X_i} \right)^2. \quad (2)$$

The penalty term is the Frobenius norm of the Jacobian matrix, which is the sum squared over all elements inside the matrix. We could think Frobenius norm as the generalization of Euclidean norm.

We also introduce a callback function during the training process to detect the loss of the model. Moreover, data augmentation is adopted for the machine type of "valve" only.

5. RESULTS

5.1. Results on Development Dataset

The methods of our best models of each machine type are shown in Table 1. The name of model has been mentioned in previous network structure section. we choose these best models for each machine type through lots of experiments and they are proved to be the best. Our best results and the baseline system [5], [6] results on development dataset of each machine type are shown in Table 2 and Table 3.

Table 1: Model choice and loss function choice of each machine type.

Machine Type	Model	Loss
Toy-car	Encode-residual model	MSE
Toy-conveyor	Mix-Residual-13 model	MSE and contractive loss
Fan	Mix-Residual-1324 model	MSE
Pump	Mix-Residual-13 model	MSE and contractive loss
Slider	Mix-Residual-1324 model	MSE
Valve	Iteration-Residual Model	MSE

Table 2: The best results on development dataset of each machine type

Machine Type	AUC	pAUC
Toy-car	0.8162	0.6697
Toy-conveyor	0.7759	0.6250
Fan	0.6956	0.5140
Pump	0.7453	0.5972
Slider	0.9032	0.7209
Valve	0.8117	0.5365

Table 3: Baseline results on development dataset of each machine type

Machine Type	AUC	pAUC
Toy-car	0.7877	0.6758
Toy-conveyor	0.7253	0.6073
Fan	0.6583	0.5245
Pump	0.7289	0.5999
Slider	0.8476	0.6653
Valve	0.6628	0.5098

Compared with the baseline model, it is found that our methods get better results on each machine type. It can do a better job on anomaly sound detection than the baseline system.

5.2. Submissions

Judging by our results on development dataset, we choose two methods to submit at last.

1. Submission 1: This submission is the combination of the best result of each machine type, which the detailed methods have been aforementioned.
2. Submission 2: In this submission, we only use mix-residual-13 model to reduce the number of total parameters.
3. Submission 3: In this submission, we use 3 models for 6 machine types. Types of "toy-car" and "valve" use the iteration-residual model; Types of "toy-conveyor" and "pump" use the mix-residual-13 model; Types of "fan" and "slider" use mix-residual-1324 model.

6. CONCLUSIONS

In this paper, we mainly introduce 4 residual connection autoencoders inspired by the Resnet for anomaly sound detection. And for different machine type, we choose different model structure, which show great results than baseline systems.

7. REFERENCES

[1] Toderici, George & O'Malley, Sean & Hwang, Sung & Vincent, Damien & Minnen, David & Baluja, Shumeet & Covell, Michele & Sukthankar, Rahul. (2015). Variable Rate Image Compression with Recurrent Neural Networks.

[2] Li L. Deep Residual Autoencoder with Multiscaling for Semantic Segmentation of Land-Use Images[J]. Remote Sensing, 2019, 11(18): 2142.

[3] Koizumi Y, Saito S, Uematsu H, et al. Unsupervised detection of anomalous sound based on deep learning and the Neyman–Pearson lemma[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 27(1): 212-224.

[4] Rifai S, Vincent P, Muller X, et al. Contractive auto-encoders: Explicit invariance during feature extraction[J]. 2011.
[5] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. ToyADMOS: a dataset of miniature-machine operating sounds for anomalous sound detection. In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 308–312. November 2019
[6] Harsh Purohit, Ryo Tanabe, Takeshi Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII Dataset: sound dataset for malfunctioning industrial machine investigation and inspection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 209–213. November 2019.