# DCASE 2021 TASK 2: ANOMALOUS SOUND DETECTION USING CONDITIONAL AUTOENCODER AND CONVOLUTIONAL RECURRENT NEURAL NETWORK

Technical Report

*Wei-Lin Liao[1], Tsung-Han Wu[2], Shu-Yu Chen[2], Yun-Shing Wu[2], Chia-Yin Chen[2], Cai-Yu Yuan[2], Chung-Che Wang[3], Jyh-Shing Roger Jang[2,3]*

[1]Dept. of Mechanical Engineering, National Taiwan Univ., Taiwan
[2]Dept. of Computer Science and Information Engineering, National Taiwan Univ., Taiwan
[3]FinTech Center, National Taiwan Univ., Taiwan

## ABSTRACT

This technical report describes our methods to Task 2 of the DCASE 2021 challenge: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions. We use reconstruction error of a conditional autoencoder and *1 - classification confidence* of a classifier as anomaly scores.

***Index Terms*—**Conditional autoencoder, convolutional recurrent neural network

## 1. INTRODUCTION

The goal of Task 2 of the DCASE 2021 challenge is to detect anomaly sound in a different operation or environmental condition (target domain). All the training examples belong to normal cases, and most of them are in original operation or environmental condition (source domain). More details about the task and datasets can be found in [5], [6], and [7]. In our work, we invoke Conditional AutoEncoder (CAE) and Convolutional Recurrent Neural Network (CRNN) and use reconstruction error of a CAE and *1 - classification confidence* of a CRNN-based classifier as anomaly scores, respectively.

The rest of this report is organized as follows. Section 2 describes details of our approaches. Section 3 shows experimental results. Section 4 shows how we pack our submissions.

## 2. OUR APPROACHES

### 2.1. Conditional Autoencoder

A standard autoencoder consists of an encoder and a decoder. The encoder maps an input vector $X$ to a latent vector $Z$, and the decoder maps $Z$ to $\hat{X}$, which is a reconstructed version of $X$. Mean Squared Error (MSE) is usually used as the loss function. Different types of variation have been used for Anomalous sound detection in [1] and [2].

In our work, CAE is used. A one-hot or two-hot vector corresponds to the class labels of input $X$ is fed into the encoder together with $X$, and fed into the decoder together with $Z$. The input $X$ is composed of 5 consecutive frames of log-mel spectrogram. Four fully-connected layers with 1,024 nodes in each layer are used as encoder and decoder. Mean value of MSE loss for different frames is used as the anomaly score. The n-hot vector for class label(s) can be machine types (one-hot, size 7) or machine types and sections (two-hot, size 13). The size of latent vector $Z$ is variable and the results of different sizes will be shown in the experimental section.

### 2.2. Convolutional Recurrent Neural Network

CRNN is mainly composed of convolutional layers and recurrent layers (and possibly some fully-connected layers). Such a structure was widely used for different tasks like audio classification [3], [4]. For anomaly detection, we can use CRNN as a classifier for different types of machine, and *1 - classification confidence* can be used as the anomaly score.

In our work, the network is composed of several convolutional blocks, two Long Short-Term Memory (LSTM) layers, and a dense layer. Each convolutional block contains a convolution layer, a batch normalization layer, and a maximum pooling layer. Rectified Linear Unit (ReLU) is used as the activation function after the batch normalization layer. The size of the convolution kernel is 3-by-3. The number of output channels of the first block is 32 and increases block by block. Output of the last step of the last LSTM layer is then fed into a dense layer for final output. The classification target is machine types.

## 3. EXPERIMENTAL RESULTS

We test our implementation under different parameter settings and different subset of training data. Results obtained using only the development set are used to select parameters and then train new models using the development set and the additional training set (or additional training set only). Fig. 1. and Fig. 2. shows AUC and pAUC for different machine types and models, respectively. Details of models are described as follows:

- CAE-128-DA: a CAE conditioned on machine types, size of $Z$ is 128, trained using development set and additional training set.
- CAE-S-32-DA: a CAE conditioned on machine types and sections, size of $Z$ is 32, trained using development set and additional training set.
- CAE-S-64-DA: same as CAE-S-32-DA, but the size of $Z$ is 64.
- CAE-S-128-DA: same as CAE-S-32-DA, but the size of $Z$ is 128.

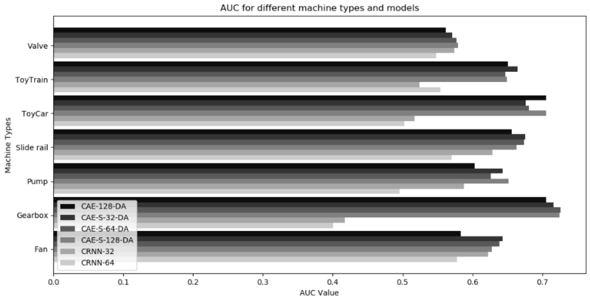

Figure 1: AUC for different machine types and models.

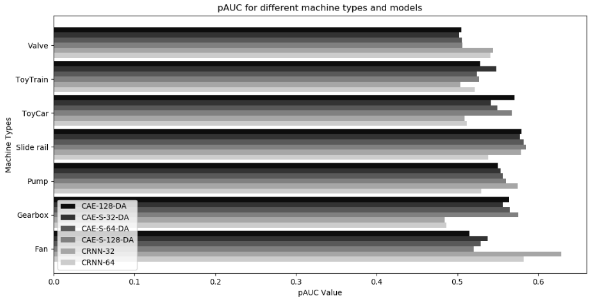

Figure 2: pAUC for different machine types and models.

- CRNN-32: 4 convolutional block, number of output units of LSTM is 32.
- CRNN-64: same as CRNN-32, but the number of output units of LSTM is 64.

## 4. SUBMISSIONS

Based on the above results, we pack our submissions as follows. While comparing different models for one or all machine types, harmonic mean of AUC and pAUC of one or all machine types is used.

- System 1 is CAE-S-128-DA, because it performs the best over all the CAE-based models.
- System 2 is CRNN-32, because it is the better one for the two CRNN-based methods.
- System 3 is CAE-S-32-DA, bedause it performs the best over all the remaining models.
- System 4 is manually ensembled according to individual results. CAE-128-DA is used for ToyCar, CAE-S-32-DA is used for ToyTrain, CAE-S-64-DA is used for SlideRail, CAE-S-128-DA is used for Gearbox and Pump, CRNN-32 is used for Fan and Valve.

## 5. REFERENCES

[1] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, “Unsupervised Anomalous Sound Detection Using Self-Supervised Classification and Group Masked Autoencoder for Density Estimation,” DCASE2020 challenge.

[2] P. Daniluk, M. Goździewski, S. Kapka, and M. Kośmider, “Ensemble of Auto-Encoder Based and WaveNet Like Systems for Unsupervised Anomaly Detection,” DCASE2020 challenge.

[3] F. Grondin, J. Glass, I. Sobieraj, and M. D. Plumbley, “Sound Event Localization and Detection Using CRNN on Pairs of Microphones,” arXiv preprint arXiv:1910.10049 (2019).

[4] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional Recurrent Neural Networks for Music Classification,” in Proc. *IEEE ICASSP*, 2017.

[5] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," arXiv preprint arXiv:2105.02702, 2021.

[6] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," arXiv preprint arXiv:2106.02369, 2021.

[7] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," arXiv preprint arXiv:2106.04492, 2021.