Team 42

Anomalous Sound Detection

Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring

AVALLE Dario
EURECOM
avalled@eurecom.fr

FONTANA Umberto
EURECOM
fontana@eurecom.fr

SORBI Marco
EURECOM
sorbi@eurecom.fr

SPINA Gabriele
EURECOM
spina@eurecom.fr

***Abstract*—The unsupervised anomalous sound detection (ASD) task consists in detecting unknown anomalous sounds of a machine knowing the normal behavior of the same. In this report, we propose the SepSTgramNet, an architecture that derives its structure from the STgramNet but with a separate elaboration of the channels. In the result section, we show that this architecture outperforms the original model in this novelty detection task thanks to the different processing dedicated to spatially non-correlated channels.**

I. Introduction

The anomaly detection task consists in identifying abnormal sounds emitted from a machine using its audio samples. The collection of anomalies is usually a hard job, such that the anomalies are often available in a small number or not available at all. For this reason, the training data contains only audio samples that record the normal behavior of the target machines. A standard approach to this challenge is the usage of Autoencoders (AE) [1], which learn the underlying distribution of non-anomalous data and use the reconstruction error of an audio sample as an anomaly score (a higher reconstruction error is an index of the presence of an anomaly). We approached the challenge using visual-features classifiers, trained to recognize a single machine among all the others, solving a classification task in practice. Once the model is trained for this task, we can use the score it assigns to a given machine as an anomaly score. Our main contribution to this challenge, which is inspired by the work of Liu et al. [2], is the extraction of temporal features together with the mel-spectrogram of the audio as input to a final classification model. In particular, we made use of the STgramNet proposed in [2] and our own modified architecture SepSTgramNet, which differentiates in the way the features are combined.

II. Data Exploration

The original DCASE challenge [3] included 6 type of machines, while this challenge made use only of one machine type, which is the slide rail type.

Two datasets were available, both containing a train and test partition:

- Development dataset: contains a total of more than 2000 normal samples from the three machines with ids 0, 2, and 4, while the test set contains approximately 800 anomalous and 300 non-anomalous audios with the same machine ids of the train.
- Evaluation dataset: contains approximately 2000 additional non-anomalous training samples with ids 1, 3, and 5. The same ids are present in the test set which is used for the submission on the Kaggle platform.

Each recording is a single-channel audio of 10 seconds duration that includes both a target machine's operating sound and environmental noise. A standard way to visualize this kind of data is to use the mel-spectrogram, which is a representation of the short-term power spectrum of a sound on a mel scale of frequency. Two examples, of a normal and of an anomalous record, are illustrated in Fig. 1, where a logarithmic transformation was applied for visualization purposes. It is possible to notice a clear difference in the spectral representation between the two.

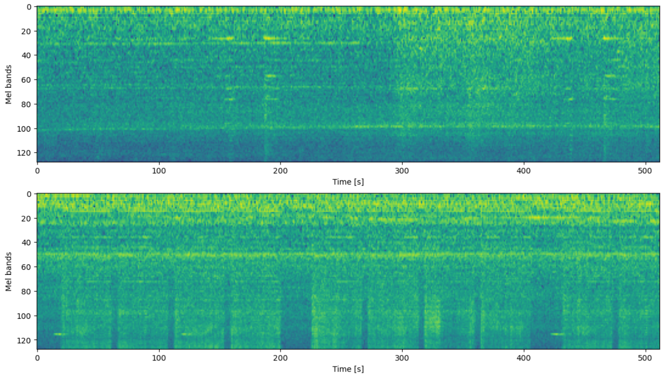

Fig. 1: Log-mel spectrogram of a normal (figure above) and of an anomaly (figure below) record

We tried in all the experiments to use the log-mel-spectrogram, and the normal spectrogram, by replacing the mel-spectrogram or by adding them as a feature, but they were dropped in the end as they consistently lowered the performance of our models.

III. Models

The novelty detection task can have different approaches other than ours, such as the aforementioned Autoencoder architectures, the clustering approach, and the self-supervised approach. In particular, we started investigating a contrastive learning self-supervised approach [4], which consists in training a model to learn the general features of the data by minimizing the distance between similar samples and maximizing the distance between different ones. The initial results were though not promising and for this reason, we didn't investigate further and this technique is not the subject of this report.

The approach we chose, instead, is to transform the outlier detection into a multiclass classification. The main idea behind this solution is to train a classifier to recognize the machine id of the sample in a supervised learning framework. The anomaly score will be the probability of the sample being anomalous. In our setting, it was computed as follows:

$$s_i = 1 - Softmax(z_i)|_{id_i} \tag{1}$$

where z_i is the output of the model for the ith sample and id_i is the machine id of the ith sample. The anomaly score ranges from 0 (not anomalous) to 1 (anomalous).

The baseline we used to compare the improvements is a pre-trained ResNet50, in which the fully-connected layer has been fine-tuned for our classification task. As an improvement, we used also the STgramNet architecture proposed in [2], which combines the spectral features with a learned temporal embedding.

A. STgramNet and TgramNet

The TgramNet architecture [2] is a model that permits the extraction of the temporal features from an audio signal. It is a neural network composed of a 1-dimension Convolutional layer followed by three blocks containing, in sequence, a Layer Normalization operation, a Leaky ReLU activation function, and another 1-dimension convolutional layer. The final output is an embedding having the same shape as the mel representation of the signal, 128×512. An example of such embedding is provided in Fig. 2.

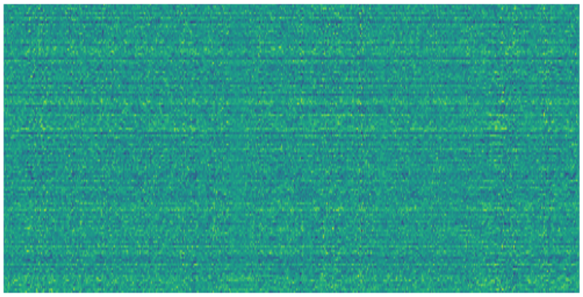

Fig. 2: Temporal features extracted by the TgramNet

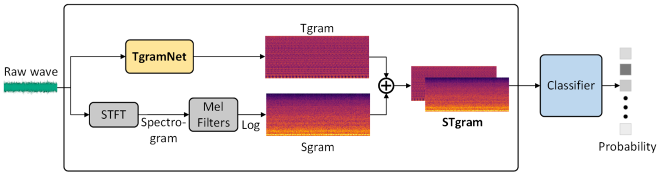

Fig. 3: STgramNet Architecture

Once the temporal features are extracted, they are concatenated channel-wise to the mel-spectrogram of the audio sample. The two-channels image, combination of the temporal and spectral features, is then fed into a visual classifier, in our case a ResNet18. This entire structure constitutes the STgramNet architecture, represented in Fig. 3. Since ResNet18 operates on three-channels inputs, we manually set the third channel of the image to zero.

B. SepSTgramNet

As already mentioned, in the STgramNet the classifier is a CNN having as input an image with the temporal-spectral features. We thought this could be a problem because the convolution operator aims to extract meaningful information from the surroundings of a single point. In the STgramNet case, the information contained in the same area of the two channels are not necessarily correlated. This is the reason why we implemented the SepSTgram, where these two channels are passed separately to two CNNs (ResNet18 backbones), the outputs of which are then combined and passed to a network of two fully connected layers to perform the classification. The architecture is shown in Fig. 4.

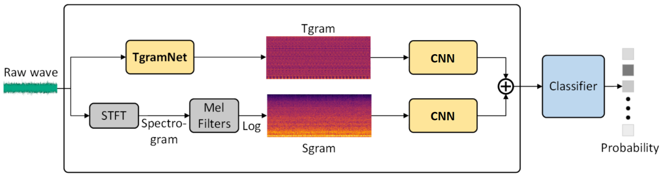

Fig. 4: SepSTgramNet Architecture

IV. Experiments

The models have been trained to distinguish the machine ids using only non-anomalous sounds. During the evaluation, instead, we didn't use the classification results, but the score attributed to the machine id of the sample, computed as in Equation 1.

We used a pre-trained ResNet50 as a baseline for the task. It was trained using the mel-spectrogram of each sample, replicated into three channels, and only the last fully-connected layer was fine-tuned. We also tried to use the log-mel-spectrogram or the frequency spectrogram instead of the mel-spectrogram, but this resulted in poor performance of this model. This is also the reason why we dropped the idea of using them also for the other models.

The TgramNet architectures of STgramNet and SepSTgramNet were trained from scratch, while the ResNet18 backbones of the two architectures were initialized to a set of pre-trained weights and re-trained entirely. Also in this case the one-channel images were replicated to three channels to match the input of ResNet18.

All the models were trained using an Adam optimizer and a cross-entropy loss function. For each model, the hyperparameters were tuned on the development set, before being tested on the evaluation set.

V. Results and Conclusions

The performance of each model in the development and in the evaluation dataset is shown in Table I, which shows also the best hyperparameters found. The metric used to compute the results is the ROC-AUC score.

TABLE I: Table of Results

Model	Hyperparameters	Dev Score	Eval Score
ResNet50 (baseline)	*Epochs: 10,* *batch size: 32,* *lr :* 10^{-3},	0.867	0.753
STgramNet	*Epochs: 10,* *batch size: 4,* *lr :* 10^{-4}	0.943	0.876
SepSTgramNet	*Epochs: 9,* *batch size: 4,* *lr :* 10^{-3}	**0.991**	**0.918**

The SepSTgramNet was the best scoring model in the challenge, showing that elaborating the temporal and the spectral features separately improves the novelty detection performance over the STgramNet. It can be noted that the evaluation score of the models is always lower than the development score, as the two sets include different machines, and the hyperparameters found for the development set probably differ from the optimal ones on the evaluation set.

VI. Future Works

For future works, is possible to use the SepSTgramNet configuration with other architectures besides ResNet, such as DenseNet or ConvNeXt. The work of Liu et al. [2] uses the ArcFace loss [5] as an alternative to the cross-entropy and its integration could determine an improvement in the recognition task. It can also be interesting for future work to evaluate the effect of well-calibrating the probabilities of the softmax output when computing the anomaly score of each sample. Another possible job is to extend this solution to different machine types to verify the flexibility of the model. A final work could be dedicated to the investigation of other possible scales different from the mel-scale, which tries to simulate human hearing and may not be the perfect fit for this task.

References

[1] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," *CoRR*, vol. abs/2102.07820, 2021.

[2] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous Sound Detection Using Spectral-Temporal Information Fusion," 2022.

[3] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, pp. 81–85, November 2020.

[4] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," 2018.

[5] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 5962–5979, oct 2022.