Rare Data Augmentation for Audio Event Detection based on Generative Adversarial Network

Zifeng Zhao, Han Lin, Xuanpeng Li
School of Instrument Science and Engineering
Southeast University, China
zifeng_zhao@seu.edu.cn, linhan@seu.edu.cn, li_xuanpeng@seu.edu.cn

Abstract

In current audio event detection, especially in urban sound classification tasks, some categories of audio events still face certain challenges, such as difficulties in obtaining data, small scale and unbalanced distribution, which may probably lead to over-fitting of the detection model or poor generalization. In this paper, we propose a data augmentation solution based on Generative Adversarial Networks (GANs) to implement the data augmentation for rare acoustic audios. This method can improve the precision, recall and F1 score of the model.

Topics: computational intelligence; language and media information

Introduction and background

With the development of deep learning methods in the field of multimedia recognition, data-driven audio event detection becomes a compelling topic. Audio event detection can be used in monitoring, public security, acoustic environment sensing and other fields. Traditional audio data augmentation methods can increase the diversity of data and suppress over-fitting to some extent, but they are far from enough to expand the data set to a much larger scale. Recently, GAN-based methods have achieved huge progress in image generation (for example, DCGAN). Inspired by this, our study focuses on GAN-based audio data augmentation and its impact on data-driven audio event detection models.

Methodology

We employ a Convolutional-Recurrent Neural Network (C-RNN) as our baseline model, which consists of 3 convolution-pooling layers and 2 bidirectional LSTM layers, with a dense layer at the top. We get the confidence of each 10 event categories as the output. For the generative model, the WaveGAN architecture is adopted, which is similar with the famous DCGAN in principal. Different from DCGAN, it turns two-dimensional operation into one-dimensional operation to avoid distortion caused by audio-spectrum transformation.

Data augmentation is carried out only for those relatively rare audios in 10 categories. In the following part we will take gunshot audio as an example. The model with and without data augmentation are taken as the experimental group and the control group respectively, and their performance on gunshot data and on the overall data set was tested. At the experimental group, the gunshot audio generated by GAN was added to the original data set according to the ratio of 1:1 to form a new data set. In order to observe the pure effect of GAN data augmentation on the C-RNN model, no additional filter is used to modify the audio generated by GAN.

Results and analysis

We use the gunshot data in UrbanSound8K to train the GAN model until it converges. Then 374 gunshot audios are randomly generated to augment the data set of the experimental group. For both of the control group (without augmentation) and experimental group (with augmentation), 15 times of model training and testing are conducted.

As can be seen from the experiment data (Figure 1), for all 10 categories of audio classification in UrbanSound8K dataset, after the augmentation for gunshot audios, the precision of the model is 0.5971, the recall is 0.5989 and F1 score is 0.5980, which improves the performance by 0.0222, 0.0246 and 0.0235 respectively.

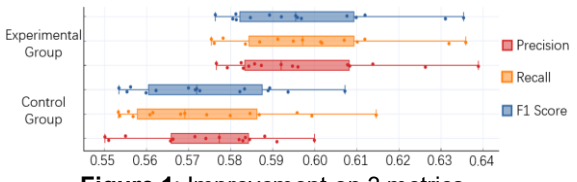

Figure 1: Improvement on 3 metrics

Conclusion

In this paper, we propose an augmentation method for audio data using GAN. The experimental data show that by augmenting rare audios samples (gunshot) in the overall data set using GAN, the model's precision, precision, recall and F1 score can be improved at the same time. We believe that this method probably have similar effect on other types of rare audio signals, and further experiments and analysis will be carried out in the future.