DATA AUGMENTATION METHODS EXPLORATION FOR SOUND EVENT DETECTION

Technical Report

Marco Bertola

Universitat Pompeu Fabra

ABSTRACT

In this technical report is describe the submission of a system for DCASE2022 Task4: *Sound Event Detection in Domestic Environments 2022* [1]. Sound Event Detection (SED) systems have gained great attention in the past few years, motivated by emerging applications in several different fields such as smart homes, autonomous cars, and healthcare. Their performances can heavily depend on the availability of a large amount of strongly labeled data. Generating or retrieving this data is often difficult and costly. The aim of this work is to explore, combine and compare different data augmentation techniques to balance out the lack of strongly labeled data. As conclusion, the best result is submitted to DCASE 2022 Task4 challenge.

Index Terms— Sound Event Detection, Data Augmentation, DCASE

1. INTRODUCTION

The main purpose of a SED task is to detect and classify sound events within a sound clip identifying the relative boundaries, mostly called onset and offset. In the last years, neural networks have been heavily applied to SED tasks achieving good accuracy results and making progress in the current state-of-the-art [2]. However, training a neural network to perform this kind of task requires a lot of annotated audio data with relative annotations to provide class labels, onset, and offset of the events. We refer to this kind of data as strongly labeled. Generated strongly labeled data can be very difficult and costly and it is mostly human-based. Synthesizing the data can partially balance out the problem by taking advantage of automation, but it also presents some trade-offs. In fact, training SED models with a strongly labeled synthesized dataset could bring the model to overfit on a few sound examples [3]. Therefore this scenario requires utilizing some weakly labeled data, which provides only the class information, and some unlabeled data, which provides no information at all [3]. However, this previous type of data results much simpler to obtain. In this kind of context, data augmentation techniques can also provide a solution for improving model generalization and balancing out a lack of strongly labeled data.
Following this consideration, we firstly decided to enlarge and extend the range of experiments already proposed by Nam et al.[3], and secondly, we experimented the outcome generated by combining together some of the previous techniques

2. AUDIO DATA AUGMENTATION

Data augmentation describes a set of techniques where a subset of training instances are manipulated and used as additional training instances. This kind of approach has been widely applied in image processing task in the early recent. A canonical example of this is the presentation of rotated, scaled, and translated images in image classification.
First applications of audio data augmentation have mostly involved speech recognition tasks and, in particular, in applying audio effects to spoken training examples. This typically took the form of adding noise, approximating reverberation or other channel effects. In 2017, for instance, a study conducted by Ko et al.[4] showed that combining clean and reverberated training data generates a considerable improvement in the close-talking scenario. Moreover, another study of 2018 by Liang et al.[5] on the ASR task, demonstrated that enforcing noise-invariant representations by penalizing differences between pairs of clean and noisy data can increase model accuracy, produce models that are robust to out-of-domain noise, and improve convergence speed. In 2017, Park et al. presented SpecAugment [6], a pretty straightforward approach that is still one of the most commonly used data augmentation methods right now. The suggested augmentation policy consisted of warping the features, masking blocks of frequency channels, and masking blocks of time step. This approach has also recently inspired data augmentation experiments applied to SEE tasks [3]
In summary, the intuition behind data augmentation is to enrich the sample distribution represented in training data to better approximate the population that will be sampled from during evaluation.

3. EXPERIMENTS

The aim of the experiment is to try to balance out the presence of synthetic data by enhancing the model's generalization capacity via data augmentation. The audio data augmentation techniques we experimented with have been reverberation, time and frequency masking, noise, and frame-shift. Time masking, frequency masking and mixup, in particular, has been applied with different settings. As a baseline, we decide to pick as a reference the PSD1 and PSD2 student results without any data augmentation technique. Data processing in the experiments from *01* to *07* has been randomly applied in real-time during the training step with a probability *P(A)=0.5*. As a final step, we also tried to combine together to different of the previous techniques with different data processing policies.

4. RESULTS

The outcome of the experiments has been tracked and reported in Table 1.

Model ID	Methods	P(A)	PDS1	PSD2
00	Baseline w/o d.a.	0.5	0.324	0.500
01	Reverb	0.5	0.310	0.475
02	Mixup 0.2	0.5	0.353	0.531
03	Mixup 0.5	0.5	0.340	0.520
04	Frequency and Time Masking 80	0.5	0.322	0.513
05	Frequency and Time Masking 120	0.5	0.330	0.524
06	Frame-shift	0.5	0.325	0.513
07	Mixup 0.2 + Frequency and Time Masking 120	0.5	0.300	0.439
08	Mixup 0.2 + Frequency and Time Masking 120	0.3	**0.356**	**0.554**

Table 1: Results

5. CONCLUSION

As a part of DCASE2022 *Sound Event Detection in Domestic Environments 2022* challenge, different data augmentation techniques have been explored showing slight improvements or degradation from the baseline. Model *08* has been submitted for the final challenge evaluation.

6. REFERENCES

[1] http://dcase.community/challenge2022/.

[2] https://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments-results/.

[3] Nam, Ko, Lee, Kim, Jung, Choi, and Park, "Heavily augmented sound event detection utilizing weak predictions," *IEEE Transactions on Multimedia*, vol. 17, no. 10, Oct. 2021. [Online]. Available: http://ieeexplore.ieee.org/document/7100934/

[4] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5220–5224, iSSN: 2379-190X.

[5] D. Liang, Z. Huang, and Z. C. Lipton, "Learning Noise-Invariant Representations for Robust Speech Recognition," *arXiv:1807.06610 [cs, eess]*, July 2018, arXiv: 1807.06610. [Online]. Available: http://arxiv.org/abs/1807.06610

[6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Interspeech 2019*, pp. 2613–2617, Sept. 2019, arXiv: 1904.08779. [Online]. Available: http://arxiv.org/abs/1904.08779