BIOACOUSTIC FEW SHOT LEARNING WITH CLASS AUGMENTATION

Technical Report

Aaquila Mariajohn
Aaquilaa.arzela@gmail.com

ABSTRACT

This document details the results and techniques used for the submission for the DCASE 2022 Task 5 challenge. The goal is to identify positive shots of the required sample throughout the audio clip using few-shot learning. Prototypical networks are used for the few-shot learning training and inference models. The lack of data was compensated with augmentations.

1. INTRODUCTION

Task5 of the DCASE 2022 challenge is to use a few-shot learning approach to identify instances of the required sample from the bioacoustics audio recording. The need for this challenge arises from the lack of necessary training data in the bioacoustic field and the presence of unknown classes at inference time which is not the case for a traditional supervised learning approach.

2. FEW-SHOT LEARNING & PROTOTYPICAL NETWORKS

Few-shot learning is a type of meta-learning which is well suited for a small subset of the actual test class dataset, with fewer labels per class [1]. where the model learns to distinguish between different classes. Prototypical networks incorporate different approaches during the training and validation phases [1]. During the training stage, prototypical networks are trained with a model which outputs an embedding vector. The loss function used is prototypical loss which is the softmax of the distance of the sample embedding from the class embedding as given in [1]. During the inference stage, any distance measure is used to compare the current sample and a positive label of the class required and a negative label.

3. DATASET, FEATURES & MODEL

The dataset used for this challenge is the DCASE 2022 Task 5 Development Set [2]. The training set consists of a total of 47 classes with around 21 hours of total duration. A common sampling rate of 24kHz was used with a maximum frequency of 12kHz. For feature extraction, a frame size of 0.025 seconds with a hop of 0.0125 seconds was used. Log Mel spectrogram was used as the preferable feature. These features were grouped into segments with a segment length of 8 frames and a hop of 4 frames. The data were normalized using min-max normalization before training. Besides this data augmentation was used to increase the number of classes and also the labels. For increasing the number of labels, time-domain noise addition method was used. For increasing the number of classes, segment flipping and mirroring techniques were adopted as the performance of prototypical networks increases with more classes and since the CNN-based model treats each segment as an image.
The model used is a simple CNN-based model with 4 layers of dimensions 256, 256, 512, 64, and each with batch normalization, ReLU, and max-pooling layers.

4. RESULTS

The training was done with around 50000 episodes and a total of 184 classes with 5 samples per class. The model was trained with 5 shots and 5 classes, 500 episodes per iteration. An accuracy of 96% was achieved on both the training and validation sets. The evaluation was done with a threshold probability of 0.5, 4 iterations per file. The probability of threshold was chosen by iterating over multiple values around 0.5. An F1 score of 43.892 overall was obtained for this model. Other approaches with different numbers of shots and ways were also attempted. 5 way 5 shots gave the best performance even though theoretically increasing the number of ways should improve the performance. Given below is the comparison of 3 different approaches adopted.

Model	F1 (%)	Precision	Recall
5shot- 5way	43.892	0.44	0.43
5shot - 10way	41.28	0.34	0.51
10shot - 5way	33.5	0.25	0.49

5. CONCLUSION

Prototypical networks have been proven as a good approach for few-shot learning for image-based classification [1]. Even though prototypical networks work well with other data, it is not effective for audio data due to the complex nature of the data. A typical shot in image classification would have only 5 samples for a 5-shot classification. But training with just 5 shots will not be effective as 5 shots of the data may have only a small part of the label. For audio data, each occurrence of the positive label is further divided into multiple segments depending on the segment size used. Hence, it is not straightforward for audio as compared to image classification.

6. REFERENCES

[1] https://doi.org/10.48550/arZiv.1703.05175
[2] https://dcase.community/challenge2022/task-few-shot-bioacoustic-event-detection