Data Augmentation for Acoustic Biodiversity Monitoring — Capstone Final Report

Li Sun, Ziye Tao, Yang Xiang, Yanqi Luo, Meichen Dong

December 2022

Abstract

Rare and inconspicuous species characterize the tropical ecosystem, but the insufficient data from them is a common barrier for the application of advanced data mining tools. To deal with the issue of imbalanced data, data augmentation methods could play a fundamental role. So several audio and image augmentation methods, including noise injection, pitch changing, frequency and time masking are carried out and evaluated based on the model's performances of detecting the sound of rare species. It is found out random combination of audio augmentations achieves the best overall performance, and extra spectrogram augmentations do not significantly further improve the performance. The validity and effectiveness of balancing data of rare species through some augmentation methods promote the process of detecting future rare species and help better protect biodiversity.

1 Introduction

1.1 Background

Tropical ecosystems are particularly characterized by the high number of rare and inconspicuous species, which are often the main focus of conservation and wildlife management programs. However, data collection of rare species is challenging in rainforests, which would lead to data imbalance when collecting the sounds. Leaving the imbalance problem unsolved will cause inaccuracy in later model training and implementation. Data augmentation could be critical in order to balance the data. In this paper, we compare different data augmentation approaches for training. Regarding audio data augmentation, we examined four methods with the trained data: Noise Injection, Time Shifting, Pitch Changing, and Speed Changing. For image data augmentation, we tried Loudness Augmenter, Frequency Masking, and Time Masking. We aim to improve the data augmentation process to increase the training data size for rare species, better understand the sounds of rare species, and promote the model performance of rare species.

1.2 Contribution

(1) **Exploration**: Exploratory analysis was carried out on the audio data to better understand the frequency feature and variation of all the rare species across different time and locations.

(2) **Audio & Image Augmentation**: Four audio (Noise Injection, Time Shifting, Pitch Changing, and Speed Changing) and three image data augmentation (Loudness Augmenter, Frequency Masking, and Time Masking) methods were compared and evaluated on the model performance of detecting rare species.

(3) **Data Balancing**: Combination of augmentation methods are done and same-level model performance is achieved with much larger size of data, which proves the validity of augmentation methods on the rare species, and contributes to future detection of sounds from rare species.

(4) **Pipeline Construction**: A user-friendly code pipeline was built in Github for future easy reference.

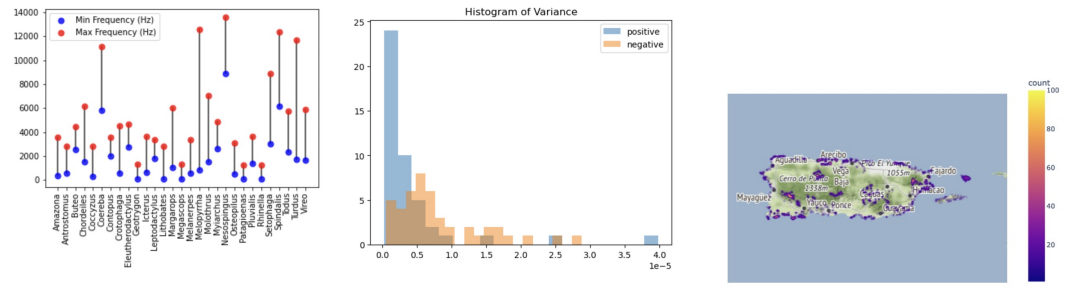
2 Related Work

Biodiversity monitoring helps campanies and organizations to manage threats of species loss and climate change[1]. Accurate sound recognition systems can play an important role in biodiversity conservation and ecological research [7]. Data augmentation helps organizations train systems with limited data. There are multiple data augmentation methods for audio classification including traditional methods on the raw audio signal, as well as the current approach of linear interpolation on the spectrogram [4]. The traditional methods extract features from audio traces and perform preprocessing feature extraction, and classification on them [2]. Some basic techniques of audio augmentation include noise injection, shifting time, and changing pitch and speed [3]. Based on the basic techniques, [9] proposed a low-implementation cost audio augmentation technique that could be used to process the raw signal. The techniques contained VTLP-based augmentation, tempo perturbation-based augmentation, and speed perturbation-based augmentation.

Apart from performing data augmentation on raw audio signals, augmenting the visual representation of audio traces, in particular, is a popular approach as well [11]. A spectrogram is a bi-dimensional graph with two dimensions of frequency and time, adding a third dimension indicating the signal intensity [4]. With the affiliation of volume, the effective size of training data increased which improved the performances of deep networks in image classification[5]. Recently, spectrogram augmentation has been applied to animal sound classification. [8] trained a six-layer CNN on audio spectrograms to classify audio recordings targeting bird species. Same-class audio samples and samples of noise are combined to perform data augmentation techniques. [4] proved the effectiveness of spectrogram augmentation compared with audio augmentation. They tested different data augmentation techniques such as image augmentation, signal augmentation, and spectrogram augmentation for training CNN for birds/cats sound classification problems. They stated that a combination of spectrogram Augmentation and Signal Augmentation is the most beneficial to classify animal sounds. Similarly, [10] stated that mixed frequency Masking data augmentation has the greatest efficiency compared to traditional methods on the raw audio signal and the current popular augmentation of linear interpolation and nonlinear mixing on the spectrum.

3 Data and Exploration

In total, there are around 31k 5-sec audio data and 2k 1-min audio data from Puerto Rico. The 5-sec audio data is catergorizd into 45 species, and for each species, there are 200 positive samples and 500 negative samples. 80% of 5-sec data was used to do training, and the rest of them was used to do validation. 1-min data was used to do testing which each audio was split into 55 5-sec recordings, with stride 1. The predicted probability for each sample was the maximum over all 55 recording pieces.



(a) Frequency Range for Each Species (b) Histogram of Variance of Each Class (c) Data Sample Location Distribution

Figure 1: Exploratory Data Analysis

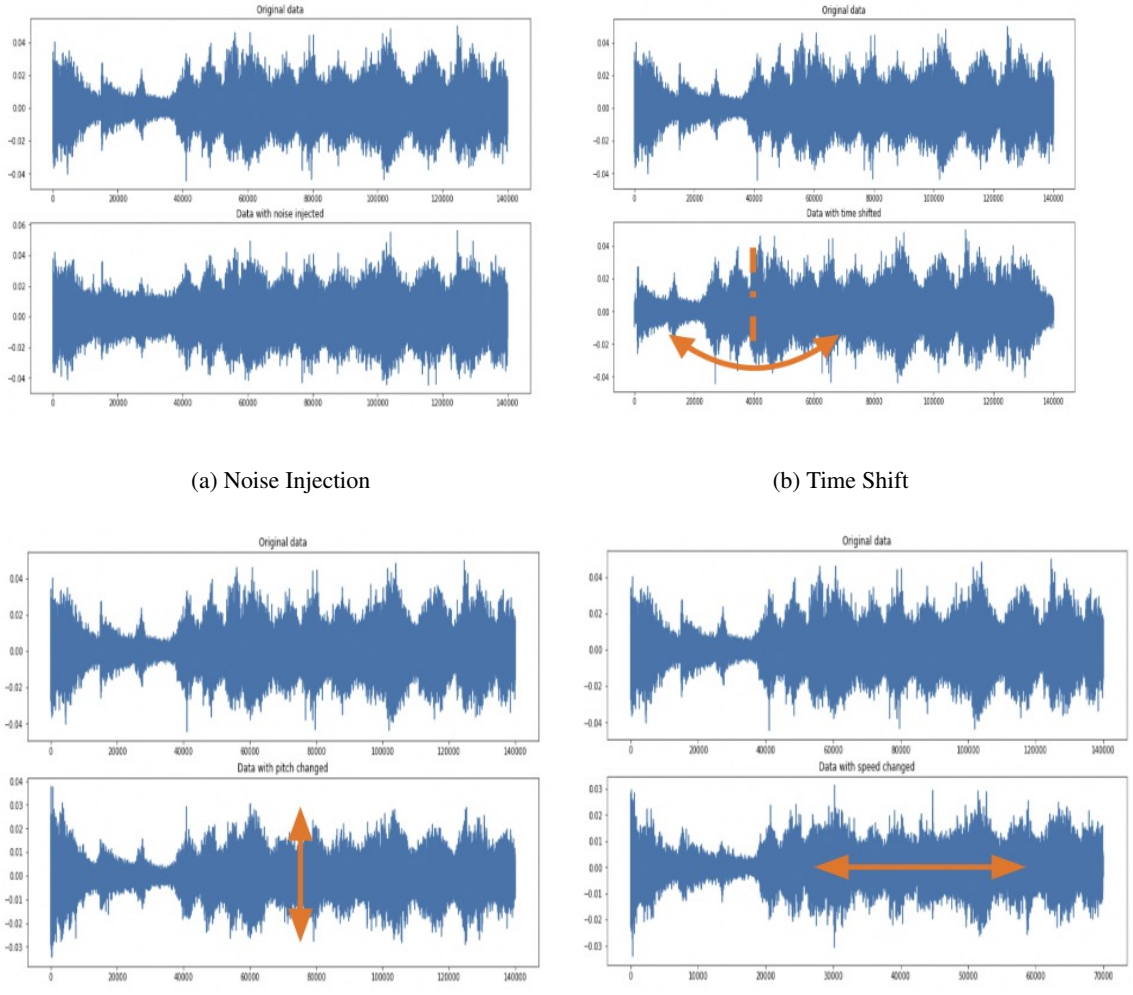
Figure 1(a) shows the frequency range of different species. Some species, such as Geotrygon, Rhinella, and

Patagioenas, have relatively low and small range; while species such as Melopyrrha, Turdus have relatively wide range. Figure 2(b) indicates that though the distributions of variance for positive and negative class are both skewed to the right, the distribution of negative class variance is more spread. It indicates the negative audio is quite diverse. We manually listened to some negative audio and found the those data could be beriefly divided into other species (such as bird, rooster) noise, and human-made noise (such a noise from equipment placing and transportation). That makes sense since by visualizing the data location (Figure 1(c)), we can see that most of the audio were collected from coastal areas and areas along the road.

4 Methods

From data exploration, it shows that the negative set has already been diverse enough, with all types of sounds included, such as human voice and noise of airplane. So augmentation would be only applied to positive training set to enrich the data and enlarge sample size without sacrificing the model performance.

4.1 Audio Augmentation



(a) Noise Injection

(b) Time Shift

(c) Pitch Change

(d) Speed Change

Figure 2: Audio Data Augmentation

4.1.1 Noise Injection

```
data + noise_factor * noise
```

Since the background noise in the positive sample would affect the classification results, noise injection is applied to the audio data. Noise is first generated from a standard normal distribution using code `noise = np.random.randn(len(data))`. Then `noise factor`, a number uniformly chosen from range 0.001 to 0.005, is multiplied to the noise in order to control the noise level.

4.1.2 Time Shifting

```
np.roll(data, shift)
```

Since each test audio would be splitted into 55 5-sec audio samples, the species should might not be in the center of the slices. Thus, a time shifting augmention method could imitate this condition. Time shifting method randomly shifts time to left or right to avoid centralization of feature sounds. The audio will be rolled at most 1 second and at least 0.1 second which is controlled by the parameter `shift`.

4.1.3 Pitch Changing

```
librosa.effects.pitch_shift(y=data, sr=sample_rate, n_steps=n_step)
```

Recall the exploratory data analysis that some species have wide frequency range, 200 positive samples might not be large enough for the model to learn various variations. Besides, the sound pitch might change with the mood or the current situations, it is hard to include all cases in 200 samples. Thus, a pitch changing augmentation is used. The Effect function from librosa library is used to adjust pitch to be higher or lower. The amount of change is controlled by `n_steps`, a random number is chosen from the range -2 to 2.

4.1.4 Speed Changing

```
librosa.effects.time_stretch(data, speed_factor)
```

Similar with pitch variation, animals may also variate the speed to express different emotion or meaning. Thus, a speed changing augementation method is applied on the audio data to imitate these situations as well. This method is also done by implementing function from librosa library. The audio would be randomly speed up by at most 1.25 times faster or speed down by at most 0.75 times slower which is controlled by the parameter `speed_factor`.

4.2 Spectrogram Augmentation

For spectrogram augmentation methods, we implemented functions from a python library called nlaug.

4.2.1 Loudness Augmenter

```
nas.LoudnessAug(zone=(zone_l, zone_r),coverage=coverage, factor=(0.75, 1.25))
```

The loudness of the audio data collected from the same species can be affected for many reasons such as the distance to the equipment. So we consider changing the loudness to augment the spectrogram dataset. We only applied the loudness augmentation on the middle part spectrogram by setting the parameters `zone_l = random(0,0.15)` and `zone_r = random(0.85,1)`. In the chosen part of the spectrogram, the coverage portion of augmentation is determined by the parameter `coverage` which is uniformly chosen between 0.8 and 1. Each input data volume would be adjusted by a different factor which was randomly chosen between 0.75 and 1.25. The volume would be reduced if the factor is between 0 and 1.

4.2.2 Frequency Masking

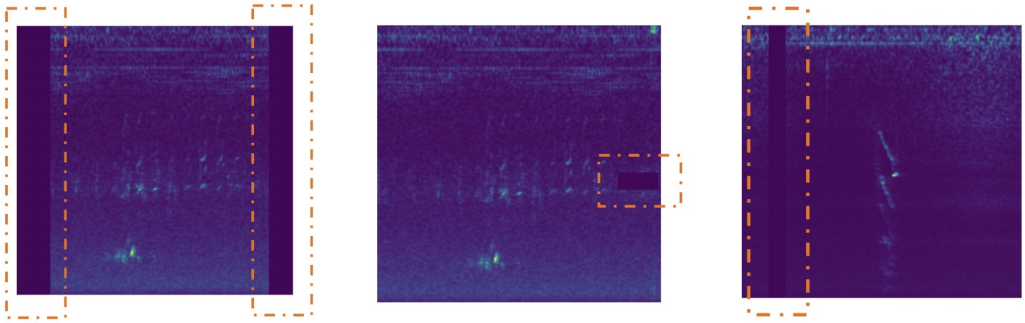
```
nas.FrequencyMaskingAug(zone=(zone_l,zone_r),coverage = coverage, factor=(10,20))
```

Frequency masking is inspired by the image augmentation method "cutout" which is to apply a spatial prior to dropout in the input image and improves the performance of the convolutional neural networks[6]. The spectrogram is masked based on frequency by random values. Different from the loudness augmentation, the augmentation would only be applied at the start(first 0.75 seconds) or the end(last 0.75 seconds) of the spectrogram by setting the parameters `zone`. The portion of augmentation is determined by parameter `coverage` which is randomly chosen between 0.8 and 1. The parameter `factor` is set as $(10, 20)$ which guarantees that at most 20 frequency channels would be masked.

4.2.3 Time Masking

```
nas.TimeMaskingAug(zone=(zone_l, zone_r),coverage = coverage)
```

Similar to frequency masking, time masking is also inspired by the image augmentation method "cutout" but the masking in more vertical in the spectrogram. The spectrogram is masked by random ranges of time using vertical bars. Similarly, with frequency masking, we only applied time masking at the start(first 0.75 seconds) or the end(last 0.75 seconds) of the spectrogram by setting the parameters `zone`. The parameter `coverage` is randomly chosen between 0.1 and 0.2, which will guarantee at most 0.15 seconds would be masked on the spectrogram.



(a) loudness augmenter (b) frequency masking (c) time masking

Figure 3: spectrogram Augmentation

5 Experiment & Result

In total, 35 experiments are carried out, and the work flow is shown in Figure 4. As for loss function, binary cross-entropy loss and masked loss are compared on the model performances, and it shows that among the negative samples of one class, there is no sound of other class of species either. So only the positive samples contain the sound of these 45 classes of rare species. Single and combination of augmentation methods are operated on the original training data, and evaluated on the model performances based on multiple metrics: (1) Recall: the ratio of correctly predicted positive observations to the all observations in actual class. (2) Precision: the ratio of corrected predicted true observations to the total predicted positive observations (3) mAP: the mean of the APs for all classes.

In Figure5, the yellow boxes refer to the four audio augmentation methods, and the blue boxes refer to the three image augmentation methods, where only a single augmentation method is used for each experiment. The purple boxes show the random combination of a group of augmentation methods, where there is 80% chance for each augmentation method to be selected, so the augmentation process for one image vary from another one. The metric scores for each experiment are summarized below in Figure5.

Comparing all the experiments, experiment with only shifting time method achieves worse overall performance than the baseline model, which may indicate that time-related augmentation methods should be used with caution. Random combination of all audio augmentation ("All audio augmentation") achieves the best

overall performance on the test data, with relatively low false negative and false positive rates. After audio augmentation, the extra spectrogram augmentation does not significantly further improve the performances. So it may not be necessary to further apply image augmentation methods after process of proper audio augmentation methods.

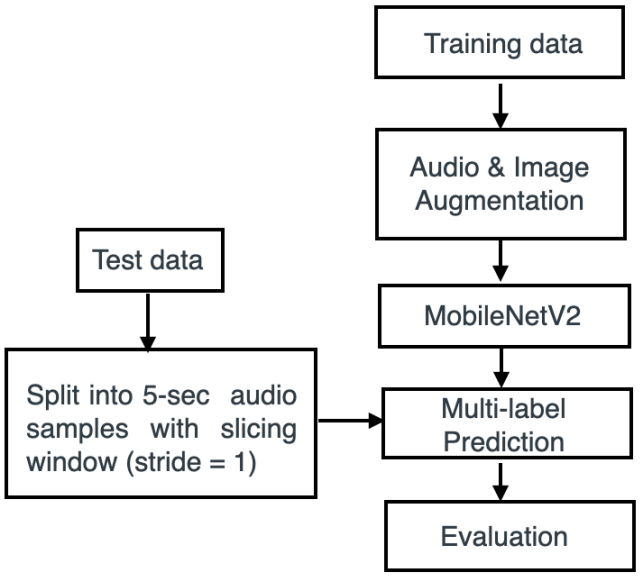


Figure 4: Experiment Work Flow

6 Pipeline

We organize the code and store a working pipeline in GitHub `https://github.com/YangXiang-Sunny/RFCx_data_augmentation` for future use and reproducing the experimental results.

7 Conclusion

In this work, we explored multiple data augmentation approaches in audio classification. It has been experimentally proved that both the audio augmentation methods and the spectrogram augmentation methods are both helpful for improving model performance, especially audio augmentation. In addition, our work shows that a random combination of audio augmentations achieves the best overall performance with a 15.5% mAP increase. For future work, the current methods will be evaluated with data sets from different areas, and other effective combinations of methods and different hyperparameters can be explored.

References

[1] Branko Hilje and T. Mitchell Aide. Calling activity of the common tink frog (diasporus diastema) (eleutherodactylidae) in secondary forests of the caribbean of costa rica. *Tropical Conservation Science*, 5(1):25–37, 2012.

[2] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *International Society for Music Information Retrieval Conference*, 2005.

Validation	Recall	mAP	Precision
Baseline	0.807	0.88	0.89
Noise injection	0.61	0.86	0.879
Shifting time	0.35	0.67	0.85
Changing pitch	0.79	0.9	0.91
Changing speed	0.85	0.84	0.63
All audio augmentation	0.767	0.887	0.894
Loudness	0.745	0.904	0.930
Freq mask	0.492	0.789	0.922
Time mask	0.536	0.840	0.926
All audio & spectrogram augmentation	0.885	0.915	0.886
All audio & spectrogram augmentation w/ mixup	0.360	0.736	0.840

Test	Recall	mAP	Precision
Baseline	0.134	0.27	0.44
Noise injection	0.118	0.235	0.431
Shifting time	0.078	0.158	0.415
Changing pitch	0.39	0.28	0.30
Changing speed	0.187	0.213	0.292
All audio augmentation	0.418	0.312	0.316
Loudness	0.175	0.301	0.419
Freq mask	0.186	0.314	0.417
Time mask	0.071	0.183	0.33
All audio & spectrogram augmentation	0.44	0.28	0.26
All audio & spectrogram augmentation w/ mixup	0.06	0.21	0.38

Figure 5: Final Result

[3] Edward Ma. Data augmentation for audio. *Medium*, 2019.

[4] Loris Nanni. Set of texture descriptors for music genre classification. 2014.

[5] Daniel Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Cubuk, and Quoc Le. Specaugment: A simple data augmentation method for automatic speech recognition. pages 2613–2617, 09 2019.

[6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[7] Ilyas Potamitis. Automatic classification of a taxon-rich community recorded in the wild. *PLoS ONE*, 9(5), 2014.

[8] Sprengel, Jaggi E., Kilcher M, Hofmann Y, and T. Audio based bird species identification using deep learning techniques. *Working Notes of CLEF 2016*, 2016.

[9] Thi-Ly Vu, Zhiping Zeng, Haihua Xu, and Eng-Siong Chng. Audio codec simulation based data augmentation for telephony speech recognition. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019.

[10] Shengyun Wei, Shun Zou, Feifan Liao, and weimin lang. A comparison on data augmentation methods based on deep learning for audio classification. *Journal of Physics: Conference Series*, 1453(1):012085, jan 2020.

[11] Lonce Wyse. Audio spectrogram representations for processing with convolutional neural networks. 06 2017.