Xception Based Method for Bird Sound Recognition of BirdCLEF 2020

Jisheng Bai[1], Chen Chen[1], and Jianfeng Chen[1]

Northwestern Polytechnical University, Xi'an, China
{baijs, cc_chen524}@mail.nwpu.edu.cn
chenjf@nwpu.edu.cn

Abstract. In this paper, we present an Xception based method for bird sound recognition of BirdCLEF2020. The goal of BirdCLEF2020 is to detect and classify 960 bird species within the provided soundscape recordings, it is more complex to differentiate such a large number of birds than BirdCLEF2019. In our approach, logmel or loglinear spectrograms are extracted as features, and some data augmentation techniques are utilized to improve the performance of detecting the bird sounds. Finally, we evaluate our system on BirdCLEF2020 test dataset and achieve a classification mean average precision (c-mAP) score of 0.0421.

Keywords: Bird sound recognition · Xception · Data augmentation.

1 Introduction

It is difficult to take clear photos of birds, which requires an open environment, professional equipment, high level of photography, and it takes a lot of time to actively look for the objects to be photographed. Instead, we can identify a bird's specie through a segment of its voice. Hearing the sound and distinguishing birds is important for many environmental and scientific purposes. Some successful techniques will be used in monitoring of ecological environment in the future. Birds are highly sensitive to the environment. If some areas are polluted, some birds will gradually fly away. Therefore, the changes of bird habits and population can reflect the changes of the environment. If some microphones are installed in the forests, the information of bird population and activities can be collected, which can significantly improve the efficiency and accuracy compared with manual observation. The sounds of different birds are indeed specific. A large number of sound data can be collected by installing specific recording equipment. This can be done automatically in an unattended environment, without the need for people to invest extra time and money.

With the development of deep learning, a large body of research in sound classification is proven to outperform traditional methods in bird sound classification [5]. Convolutional neural networks(CNNs) show great feature extraction

ability and results in many computer vision tasks. The image-based architectures, Inception-v3 for example, can obtain the best performance in sound classification or what ever the targeted domain [6].

2 Dataset

The training data consists of audio recordings for bird species from South and North America and Europe. The Xeno-canto community contributes this data and provides more than 70,000 high-quality recordings across 960 species to this year's challenge. Each recording is accompanied by metadata containing information on recording location, date and other high-level descriptions provided by the recordists.

The test data consists of 153 soundscapes recorded in Peru, the USA, and Germany. Each soundscape is of ten-minute duration and contains high quantities of (overlapping) bird vocalizations [1].

The number of recordings for each specie are calculated, although the training dataset contains over 70,000 recordings, there are less than 90 recordings in some species, the rest species are between 90 and 100 recordings. For example, the bird specie with a ebird name of "banfru1" only contains 1 recording, but there are 100 recordings in another specie "whcspa". The number distribution of the dataset is shown in Fig.1. The imbalance of the dataset can have effect on recognizing bird species.

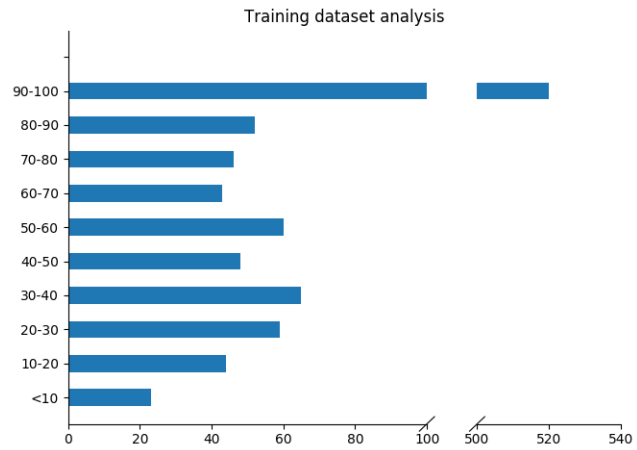

Fig. 1. The number distribution of training dataset recordings

3 Data preparation

3.1 Bird sing separation

To separate bird sound and background noise from a original recording, we apply dilation, erosion and smooth masking. Similar techniques are presented in [11] and used in [9] and [2]. We then separate all recordings into 960 bird sing species and noise classes. Details are described as following:

- Every recording is loaded with a sample rate of 22050Hz.
- Short-time Fourier transform(STFT) is utilized to calculate spectrograms with a window length of 1024 and hop length of 512.
- We calculate median value for each row and column, then set every element in the spectrogram to 1 if it is 1.5 times bigger than the median of its related row and column, otherwise it's set to 0.
- Binary erosion and dilation filters are implemented to distinguish noise and signal parts. The filter size is 4 by 4 square.
- Here we create a one-dimension vector named indicator vector, its i_{th} element is set to 1 if its related column has at least one 1, or it is 0.
- Finally, we smooth the indicator vector twice by a dilation filter of size 8 by 1 then use it as a mask to separete original bird recordings. Each recording can be divided into lots of signal and noise parts, all signal parts are concatenated as one and the same as noise. We cut all recordings of every species into 5 seconds parts. After these, we can get 960 folders of each specie of 5-second bird sound and noise recordings.

3.2 Data augmentation

In recent years, some data augmentation techniques are successfully applied in sound recognition tasks. Two data augmentation metheds are utilized in our approach.

Some time and frequency augmentation methods have been used in [9] and [11]. In our proposed method, we simply use some of the techniques to augment the dataset and they can be described as follows:

- Load a bird sound file from random position (it starts from the beginning if it reach the end).

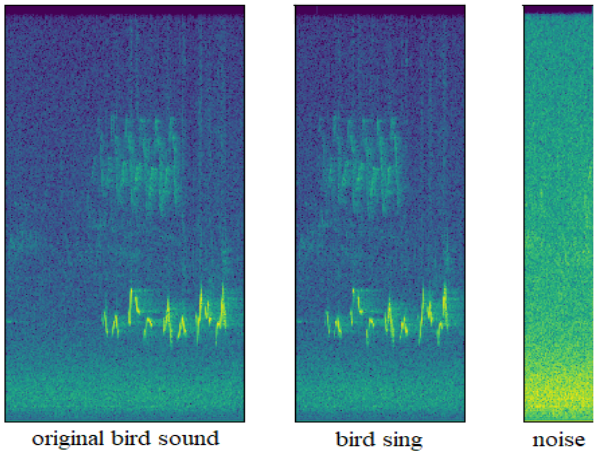

Fig. 2. An example of bird sound separation

- Add most three files on the top of a 5-second bird sound fragment with independent chance. These three files are bird sing from the same bird specie, bird noise from the same specie and noise from another specie with chance of 0.3,0.3 and 0.5 respectively. A amplitude factor between 0.3 and 0.5 is applied during the process.
- STFT is used to generate spectrogram from the added file with a window size of 1024 and hop length of 512.
- Normalization and logarithm is applied to calculate logmel or loglinear spectrogram with 128 Mel-bands, frequencies beyond 11025Hz and lower than 50Hz are removed.
- Due to the input size of Xception, different interpolation filters are applied to resize the spectrograms into 299*299*1.

To handle the imbalanced dataset and prevent overfitting, we apply mix-up during training stage [13]. It can be expressed as:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{2}$$

where x_i, x_j are input features, y_i, y_j are target labels, $\lambda \in [0, 1]$ is a random number drawn from the Beta (c, a) distribution. We can get more training samples without extra computing resource, new samples are linear interpolated of real samples.

3.3 Denoising

All the bird sounds come from diverse area around the world with special environmental background noise. In bird sound recordings, loud raining sound,

quiet background, sound of wind and sound of other animals may exist and even cover some bird sounds. To eliminate this, we propose a denoising method to do spectral subtraction on a spectrogram:

- Load a 5-second sound file and generate an amplitude spectrogram.
- Calculate mean amplitude values over frames, mark the minimum 20 frames and calculate the mean amplitude values of these frames, we can get a primary subtraction vector. Then subtract the primary subtraction vector over frequencies for each frame, we can get the final denoised spectrogram.

3.4 Features

During the training, separated bird sing, augmented bird sound and spectral subtraction sound with logmel and loglinear spectrograms of 128 bands are generated as input. All the spectrograms are calculated by logarithm.

4 Network architecture

4.1 Xception

Inception-v3 is one of the state of art architectures in image classification challenge [12]. And it is confirmed that Inception-based CNNs on Mel spectrograms provide the best performance [4]. The best network for bird song detection seems to be the Inception-v3 architecture and it preforms better than even the more recent architectures [10].

Xception is a improvement of Inception-v3 proposed by Google. In [3], the correlation between channels and spatial correlation should be dealt with separately. The convolution operation in the original Inception-v3 is replaced by a separate concept (Extreme Inception), which is the basic module of Xception. The Xception network is composed of a series of separable convolution, residual connection similar to ResNet and some other conventional operations.

4.2 Training strategy

To recognize 960 species and handle such a large amount of recordings, we used Xception architecture instead of other CNN architectures. As for features, we selected logmel and loglinear spectrogram as input. We applied a denoising method and a data augmentation method during the data preprocessing. Pytorch was implemented to train model, and python librosa library was applied to process recordings and generate features.

During the training, categorical cross entropy was used as loss function and stochastic gradient descent was used as optimizer with weight decay of 1e-4 and a constant learning rate of 0.001.

5 Results

5.1 Evaluation metrics

The evaluation metric is the classification mean Average Precision (c-mAP), considering each class c of the ground truth as a query. This means that for each class c, all predictions are extracted from the run file with ClassId(c), rank them by decreasing probability and compute the average precision for that class, which can be expressed as

$$c - mAP = \frac{\sum_{c-1}^{C} AveP(c)}{C} \tag{3}$$

where C is the number of species in the ground truth and $AveP(c)$ is the average precision for a given species c computed as:

$$AveP(c) = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{n_{rel}(c)} \tag{4}$$

where k is the rank of an item in the list of the predicted segments containing c, n is the total number of predicted segments containing c, $P(k)$ is the precision at cut-off k in the list, $rel(k)$ is an indicator function equaling 1 if the segment at rank k is a relevant one (i.e. is labeled as containing c in the ground truth) and n_{rel} is the total number of relevant segments for c.

5.2 Performance on validation split

For each validation or test file, we smoothed the prediction vector by applying a moving window size of 5 seconds. Then we only selected first 3 maximum probabilities as the result of a 5-second test fragment.

To achieve the performance of our methods, we only trained 76 bird species included in the validation dataset, the best c-mAP scores of the experiments are shown in Table 1, data augmentation with adding sounds randomly is denoted as AR, spectral subtraction is denoted as Spec sub. Experiment 1 and 2 only used the extracted bird sing to generate features, the rest of experiments extracted features Continuously from original recordings.

From Table 1 we can see, logmel with mix-up method achieves the best c-mAP score, loglinear performs well with separated bird sings and data augmentaion with randomly adding.

5.3 Performance on test split

Four teams submitted 29 submissions in LifeCLEF 2020 Bird Monophone. Finally we got the 2^{th} rank among the teams and achieved a best c-mAP score of 0.0421 and r-mAP of 0.0671. Details are shown in Table 2 [7] [8]. Three different methods were tested and results are shown in Table 3:

Table 1. Best c-mAP scores of the experiments on validation split

ID	Features	Data augmentation	Dnoised	C-map
1	Logmel	/	/	0.074
2	Loglinear	/	/	0.096
3	Logmel	AR	/	0.083
4	Loglinear	AR	/	0.092
5	Logmel	Mix-up	/	**0.154**
6	STFT	/	Spec sub	0.083
7	Logmel	/	Spec sub	0.084
8	Loglinear	/	Spec sub	0.065

– **result0:** We used the randomly adding data augmentation method, the Xception architecture, and loglinear spectrograms in this run. Test output vectors were smoothed to get final prediction. It's the best performance of all runs.
– **result1:** We used the spectral subtraction denoising method, the Xception architecture, and logmel spectrograms in this run. Test output vectors were smoothed as well. Finally we got a c-mAP score of 0.032 and r-mAP score of 0.0592.
– **result2:** We used the Xception architecture and the features are bird sing separated loglinear spectrograms. Test output vectors were smoothed as well. Finally we got a c-mAP score of 0.027 and r-mAP score of 0.0558.

Table 2. Leaderboard of LifeCLEF 2020 Bird [1]

Participant	C-Map	R-Map
mmuehling	0.1282	0.193
NPU_bird	0.0421	0.067
thailsson_clementino	0.0097	0.008
JS_CHUNGNAM	0.072	0.055

Table 3. Results of all runs

Run	Data augmentation	Feature	Denoised	C-Map	R-Map
Run0	Yes	Loglinear	No	0.042	0.067
Run1	No	Logmel	Yes	0.032	0.059
Run2	No	Loglinear	No	0.027	0.059

6 Conclusion and future work

In this paper, we proposed a system for bird recognition based on Xception with some data augmentation and denoising techniques. Finally, we got a c-mAP score of 0.0421 on official test dataset. To handle more than 70,000 recordings, Xception was chosen because of its great feature extraction ability. During training, mix-up and randomly adding data augmentation methods were applied to prevent overfitting and improve generalization performance.
We finally submitted 7 submissions of three main methods. Ensemble of networks is banned this year. We will focus on the performance of convolutional recurrent neural networks and other data augmentation methods without more computing resources for bird recognition. Features can also have great impact on performance sometimes and they would be studied as well. There is still a lot of work to improve in bird sound recognition in the future.

References

1. AIcrowd BirdCLEF2020, https://www.aicrowd.com/challenges/lifeclef-2020-bird-monophone
2. Bai, J., Wang, B., Chen, C., Chen, J., Fu, Z.: Inception-v3 based method of lifeclef 2019 bird recognition. In: CLEF (Working Notes) (2019)
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions (2016)
4. Hervé Goëau, H.G., Planqué, R., Vellinga, W.P., Kahl, S., Joly, A.: Overview of birdclef 2018: monophone vs. soundscape bird identification. CLEF working notes (2018)
5. Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W.P., Planqué, R., Müller, H.: Overview of lifeclef 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 247–266. Springer (2018)
6. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Lombardo, J.C., Planque, R., Palazzo, S., Müller, H.: Lifeclef 2017 lab overview: multimedia species identification challenges. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 255–274. Springer (2017)
7. Joly, A., Goëau, H., Kahl, S., Deneu, B., Servajean, M., Cole, E., Picek, L., Ruiz De Castañeda, R., é, Lorieul, T., Botella, C., Glotin, H., Champ, J., Vellinga, W.P., Stöter, F.R., Dorso, A., Bonnet, P., Eggel, I., Müller, H.: Overview of lifeclef 2020: a system-oriented evaluation of automated species identification and species distribution prediction. In: Proceedings of CLEF 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece. (2020)
8. Kahl, S., Clapp, M., Hopping, A., Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. In: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece. (2020)
9. Lasseck, M.: Audio-based bird species identification with deep convolutional neural networks. Working Notes of CLEF **2018** (2018)
10. Sevilla, A., Glotin, H.: Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: Working Notes of CLEF 2017 -

Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017), http://ceur-ws.org/Vol-1866/paper_177.pdf
11. Sprengel, E., Jaggi, M., Kilcher, Y., Hofmann, T.: Audio based bird species identification using deep learning techniques. Tech. rep. (2016)
12. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
13. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)