Inception-v3 Based Method of LifeCLEF 2019 Bird Recognition

Jisheng Bai[1], Bolun Wang[2], Chen Chen[2]
Jianfeng Chen[3], and Zhong-Hua Fu[3]

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

Abstract. In this paper, we present a method of bird recognition based on Inception-v3. The goal of the LifeCLEF2019 Bird Recognition is to detect and classify 659 bird species within the provided soundscape recordings. Log-Mel spectrograms are extracted as features and Inception-v3 is used for bird sound detection. Some data augmentation techniques are applied to improve the robustness and generalization of the model. Finally, we evaluated our system in BirdCLEF test data and achieved 0.055 of classification mean average precision (c-mAP).

Keywords: Bird sound classification · Inception-v3 · Data augmentation.

1 Introduction

Deep learning is proven to outperform traditional methods in bird sound classification [5]. Convolutional neural networks(CNNs) architecture performs well on many computer vision tasks and the convergence of image-based architectures such as Inception-v4 can obtain best performance in sound classification or what ever the targeted domain [6].

The training data of BirdCLEF2019 [7], which is a sub task of LifeCLEF2019 [2] contains about 50,000 recordings taken from xeno-canto.org and covers 659 common species from North and South America [1]. More than 15 and up to 100 recordings are contained per species. And validate split contains 77 recordings. All recordings vary in quality, sampling rate and encoding. Each recording includes metadata providing information of location, latitude, longitude, etc.

To recognize 659 species and train such amount of recordings, we use Inception networks instead of shallow CNN architectures. As for features, we selected log-Mel spectrogram as input. Data augmentation methods are applied during the preprocessing.
We use Ttensorflow to train model and python librosa library to calculate features.

2 Data preparation

2.1 Audio processing

To separate bird sound and background noise, similar method is applied. As it is presented in [12] and used in [9] and [3], we refer to their methods and divide all recordings into 659 different bird song species and one total noise class. Details are described as following:

- Every recording is read in a sample rate of 44100Hz.
- Short-time Fourier transform(STFT) function is use to calculate spectrogram with a window length of 512 and hop length of 256.
- Then we calculate each row and column median, then we set every element in spectrogram to 1 if it is three times bigger than the median of its related row and column, otherwise its set to 0.
- Then we apply binary erosion and dilation to distinguish noise and signal part. 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 4 by 1. And we use it as a mask to divide original bird recordings. Every recording can be divided into many 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, because we would train model, predict validate and test data every 5 seconds. After all the steps, we can get 659 folders contain every species of 5s recordings and one noise folder of all the noise parts.

2.2 Data augmentation

Data augmentation techniques are widely applied in last few years results. All recordings are resampled to 22050 Hz and then filtered by a high pass filter.Then some similar time and frequency augmentation methods used in [9] and [12] are described as following:

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

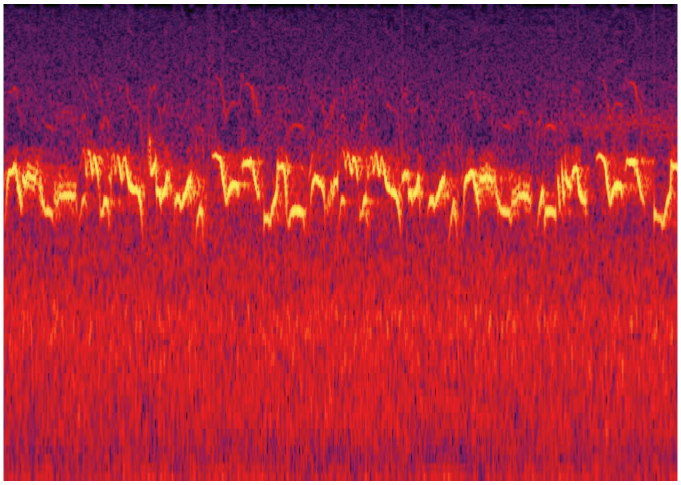

Fig. 1. Example of amerob sound signal part

- Add most four noise files on the top of a bird sound file with independent chance of 0.5. Meanwhile, a dampening factor of 0 to 0.5 are multiplied for each noise file .(In [9], the greatest impact on identification performance is gained by adding background noise. Many systems also use noise overlay as one of the data augmentation methods to improve performance.)
- Using STFT to generate spectrogram from a sound file with a window size of 1024 and hop length of 512.
- Normalization and logarithm is applied to calculate log-Mel spectrogram of 256 Mel-bands, frequencies beyond 10500Hz and lower than 200Hz are removed.
- Due to the size of Inception input, we duplicate the grayscale spectrogram to all three channels. And different interpolation filters are applied to resize the spectrogram.
- Finally the spectrograms are resized into 299*299*3 to fit the input size of Inception.

3 Network architecture

3.1 Transfer learning from Inception-v3

Inception-v3 is one of the state of art architectures in image classification challenge [13]. And it is confirmed that Inception-based convolutional neural networks 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 [11]. So we selected Inception-v3 as our base model.

Inception models were fine-tuned using neural networks pre-trained on the Large Scale Visual Recognition Challenge (ILSVRC) [10] version of ImageNet, a dataset with almost 1.5 million photographs of 1000 object categories scraped from the web. As it is mentioned in [8], strat training model with pre-trained weights can quickly train and get better performance. But if train model only with last classification layers can lead to worse result, also re-train the whole network cant reach the best performance either.

3.2 Training strategy

During the training, categorical cross entropy was used as loss function and stochastic gradient descent as optimizer with Nesterov momentum 0.9, weight decay of 1e-4 and a constant learning rate of 0.01.
We generated 20 different folders as training data, every folder was augmented with different parameters. We trained these folders with a train batch of 72 and train random order for 50 epochs.

4 Results

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

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

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:

$$AveP(c) = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{n_{rel}(c)} \tag{2}$$

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.

On the validation dataset, we selected max 100 probabilities and it got a c-mAP score of 0.088 and r-mAP (retrieval mean Average Precision) of 0.176. Meanwhile, the max 5 probabilities turned out to be 0.068 and 0.156.

- **result0:** Due to the limited time, it is a pity that we only submitted 1 run. We predicted all the test data and selected max 5 probabilities per 5 seconds as final and the only one submission. Finally we got the 3^{th} rank among the teams and got a c-mAP score of 0.055 and r-mAP of 0.145. Details are showm in Table 1.

– **result1:** We submitted another run after the deadline, and it got c-mAP of 0.065 and r-mAP of 0.164. This run contains max 100 probabilities in a 5-second period in 2.

Participant	c-Map	r-Map
MfN	0.356	0.715
ASAS_1	0.161	0.165
NWPU.jpg	0.055	0.145
PingAn	0.047	0.132
MIHAI ANDREI	0.005	0.006

Table 1. Results of different participants [7]

Item	c-Map	r-Map
Sapsucker	0.082	0.165
Columbia	0.094	0.156
Overall	0.065	0.164

Table 2. Results of additional run

5 Conclusion and future work

We presented a system based on Inception model with some data augmentation techniques for bird recognition and got final c-mAP score of 0.055. And there is a 0.01 c-mAP score improvement of evaluating max 100 probabilities compared to max 5 probabilities in a 5-second sound. To handle more than 50,000 recordings, we selected Inception-v3 which has less parameters and greater feature extracted ability. During training, data augmentation methods were applied to prevent overfitting and improve generalization performance.

Due to the limited time, we could not submit more results and compare the influence of different parameters or architectures. Ensemble of networks could significantly improve results, and it would be apply next year. We will also focus on the performance of CRNN and capsule network for bird recognition. Features can also have great impact on performance sometimes, and some unique data augmentation should be experimented to detect bird species. There is still a lot of room to improve in our future work.

6 Acknowledgement

We would like to thank Stefan Kahl, Herv Goau, Alexis Joly, Herve Glotin and Willem-Pier Vellinga for organizing this task. I especially want to thank Mario Lasseck for sharing his knowledge. I would also like to thank Northwestern Polytechnical University CIAIC for computer resource.

References

1. CrowdAI Homepage (2019), https://www.crowdai.org/challenges/lifeclef-2019-bird-recognition
2. Alexis Joly, Herv Goau, C.B.S.K.M.S.H.G.P.B.W.P.V.R.P.F.R.S.H.M.: Overview of lifeclef 2019: Identification of amazonian plants, south & north american birds, and niche prediction. In: Proceedings of CLEF 2019 (2019)
3. Fazeka, B., Schindler, A., Lidy, T., Rauber, A.: A multi-modal deep neural network approach to bird-song identification. arXiv preprint arXiv:1811.04448 (2018)
4. Herv Goau, 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. Kahl, S., Stter, F.R., Glotin, H., Planque, R., Vellinga, W.P., Joly, A.: Overview of birdclef 2019: Large-scale bird recognition in soundscapes. In: CLEF working notes 2019 (2019)
8. Lasseck, M.: Acoustic bird detection with deep convolutional neural networks. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018). pp. 143–147 (November 2018)
9. Lasseck, M.: Audio-based bird species identification with deep convolutional neural networks. Working Notes of CLEF **2018** (2018)
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
11. 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
12. Sprengel, E., Jaggi, M., Kilcher, Y., Hofmann, T.: Audio based bird species identification using deep learning techniques. Tech. rep. (2016)
13. 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)