Construction and Improvements of Bird Songs' Classification System

Haiwei Wu[1] and Ming Li[2]

[1]Sun Yat-sen University, China, [2]Duke Kunshan University, China
wuhaiweideyouxiang@gmail.com, ming.li369@dukekunshan.edu.cn

Abstract. Detection of bird species with bird songs is a challenging and meaningful task. Two scenarios are presented in BirdCLEF challenge this year, which are monophone and soundscape. We trained convolutional neural network with both spectrograms extracted from recordings and additionally provided metadata. Focusing on the soundscape situation, we applied bird event detection to reduce false alarm. Besides, we rescored the retrievals using masks which are designed for all species being identified. In addition, context information was also taken into consideration in our system. Our system was evaluated in BirdCLEF 2018 and we achieved an official mean average precision (MAP) score of 0.6548 for monophone classification without background bird songs and 0.5882 for identification with background bird songs. For soundscape, we achieved 0.1196 in classification mean average precision (C-MAP).

Keywords: sound detection, bird song, convolutional neural network

1 Introduction

BirdCLEF challenge is hosted by the LifeCLEF lab [1, 2, 3]. The aim of the competition is to train models which can classify different bird species by bird songs. Data of bird songs in this challenge are collected and displayed on www.xeno-canto.org. This year, a training set of 36,496 bird songs' audios covering 1500 species is provided. As for evaluation, two scenarios are focused on [3]. The first scenario is the identification of bird species with given monophone recordings. Each of these recordings includes mainly one bird's song. For this scenario, 12,347 unlabeled bird songs' audios are provided for evaluation. The second scenario is the detection of species of soundscape recordings. Participants are required to find out the most likely species for each segment of 5 seconds. In the contest this year, a well-labeled soundscape's evaluation set of 20 minutes including 240 segments of 5 seconds and a test set of 6 hours including 4382 segments of 5 seconds are provided. In this note, construction of our basic system for the first scenario and improvements focusing on soundscape scenario will be introduced.

The training features of our model mainly consist of two parts. The original part is the frequency information of each recording and the additional part is the metadata [4] including latitude, longitude, elevation and time information.

For the original part, audios are converted into features on the frequency domain. Every 5 seconds' segment of recordings is turned into a time-frequency image with the resolution of 512 × 256 pixels [4, 5, 6]. The Problem of audio classification is transformed into the problem of image classification where convolutional neural network performs very well [4, 5, 6, 7]. In our system, the original spectrograms are fed into a multi-layer convolutional neural network. Additional metadata are provided in the given XML files. Before the last fully connected layer, the additional features are concatenated to the flattened convolutional neural network layer. Together, the concatenated features are then used to compute the remaining layers. Besides a regular multi-layers' convolutional neural network, we also tried out ResNet [8].

Above is the method of our model training. Based on our model, we made some improvements focusing on the problem of soundscape in the test period. Firstly, a simple bird event detection [4, 9, 10] was applied before spectrograms being classified by our trained neural network. Secondly, we designed a mask for each kind of birds. Every time after getting the list of bird species from neural networks, we sorted it and rescored the top 3 or 5 species by our model after applying our masks. Thirdly, we considered the previous and next 5 seconds' information for current evaluation using a simple mechanism.

Pytorch was used for our model training and evaluation.

2 Feature preparation

We transform the problem of bird songs' classification to image classification. Each 5 seconds' segment of given audios is turned into a spectrogram with the resolution of 512 × 256 pixels [4, 5, 6]. A sliding window is used to segment the audios with an overlap of 4 seconds. For the reason that some spectrograms contain mostly noises, a simple approach introduced by [4, 5, 6] is used to separate the spectrograms into training samples and noise samples. The noise samples here are also used for data augmentation latter. Data imbalance is a severe problem in the data. For bird species whose spectrograms are less than a given number, over-sampling [11] using augmented data is applied.

Data augmentation is necessary for building robust models and handling data imbalance. Adding noises is a commonly used data augmentation method. We try to add two kinds of noises to spectrograms. For each epoch of training, 10 percent of data are added Gaussian noises and 10 percent are added noise samples.

Gaussian noises: Gaussian noises [4] are commonly used for augmentation. Adding Gaussian noises is a regular method for building robust classifying networks. Models are able to ignore this kind of noises after training. We add these noises with randomly chosen weights to our spectrograms and re-normalize the results.

Noise samples: Besides Gaussian noises, noise samples are also considered and added to our spectrograms. Noises of audios recorded by similar equipment under similar environments often share some common patterns. Adding similar noises will help improve the performance. During data processing, we have obtained many spectrograms which are thought to be noise samples [4, 5, 6]. We randomly choose some of them and add them to current features with random weights. Re-normalization is also used after addition.

Researchers [4] noted that considering metadata will do good to the performance of the model. As for our metadata, we consider latitude, longitude, elevation, and the time of a recording. We simplify the method of metadata processing in [4]. From these provided metadata, we are able to obtain a vector of 7 elements [4]. Values of elements [4] are shown below:

1. Latitude and Longitude provided, 1 if available, 0 if not;

2-3. Latitude and Longitude, normalized between 0 and 1;

4. Elevation provided, 1 if available, 0 if not;

5. Elevation, normalized between 0 and 1;

6. Time of recording provided, 1 if available, 0 if not;

7. Time information directly normalized between 0 and 1;

3 Model construction

We use a relatively shallow architecture of convolutional neural network as our basic model [4, 6]. Finding the best architecture of network is very time-consuming and we tend to find out some new methods to improve the performance in the test period. Our basic network consists of 6 convolutional layers and 3 fully connected layers. Max Pooling layers are added after each convolutional layer. Each convolutional and fully connected layer is followed by a batch normalization [12] layer to avoid parameters getting too extreme and fasten the process of convergence as well. Dropout [13] is also used after each fully connected layer to reduce the problem of overfitting. As for activation function, we select exponential linear units (ELU) [14], which is thought to be a proper choice [4, 6].

As the problem can be viewed as a multi-class identification problem, cross entropy loss is used here to be minimized. We use Adam [15] as our optimizer. Adam optimizer can be regarded as RMSprops [16] with momentum, which makes the best use of the first moment and the second moment of the gradient. Parameters can be updated more stably using it.

Learning rate decay technique [17] is used in our training process. At the very beginning, learning rate is set to 0.0001. After nearly 15 epochs of training, it is lowered to 0.00001 in order to optimize the updating. We stop the process when the accuracy converges.

Above we mention that metadata is also used for training in our system. Spectrograms are flattened to a vector of 512 elements by our convolutional neural network. We construct an additional fully connected layer for metadata [4]. Vectors of 7 elements are transformed to vectors of 100 elements through this layer. For the limitation of time, the output dimension of this layer is not further explored here. Later, we concatenate the 512 and 100 elements and feed them into the next fully connected layers. Finally, a softmax layer [18] of 1500 elements outputs the predicted probability for each bird species.

4 Improvements

In the competition of last year showed that the performances on soundscape still had a large room for improvement. The performance of model has a great impact on the final result. While for the limitation of hardware resources and time, we did not lay stress on the model training. Instead, we tried to find out methods that make the best use of our current models. Several methods we applied will be introduced below.

Bird event detection: False alarm of target species will influence the metric of C-MAP. Introduction of bird event detection [19] is able to reduce false alarm and improve the final performance. At the very beginning, we planned to use the soundscape evaluation set to train a neural network. While for the limitation of labeled data, performance was not good enough for use. At last, we directly used the method mentioned above [4, 5, 6] to separate bird songs and noises. If a spectrogram is regarded as noise, classification will not be done on it.

Masking and rescoring: For birds belonging to a specific species, the frequency of their songs always falls in a certain range. Outside this range of frequency, any other information including environmental sound or songs of other kinds of birds can be considered as noises. Inspired by this idea, we designed a mask for each kind of birds. We accumulated spectrograms of a species on the frequency axis and normalized it. The range of values under 0.6 would be masked. Here, we consider 0.6 a relatively proper threshold. The masks for all birds being classified can be viewed as band-pass filters [7]. According to the output for each 5 seconds' segment of the neural network, bird species will be sorted by their probabilities. Top 3 or 5 species will be selected and spectrogram will be applied the band-pass filters of these chosen species separately. After being masked, these 3 or 5 new spectrograms will be rescored by the neural network. Using this method, we can reduce the interference and obtain a more accurate result with our current model. In our experiment, we rescored top 3 retrievals. Illustration Fig.1 describes the whole process in detail.

Considering context: We found that, at most of the time, a bird song often lasts for a period of time more than 5 seconds. For a 5 seconds' segment in soundscape, the final result is strongly relevant to the result of previous and next 5 seconds' segments. This context information is considered in monophone scenario by overlapping while seldom considered in soundscape. Here, we simply added the outputs of the previous and next 5 seconds' segments to current output with a given weight which can be 0.2 or 0.3 and so on. Here, we set this value to 0.3 which we found that it resulted in a relatively better result in validation set. By this method, we took the context into consideration of classification.

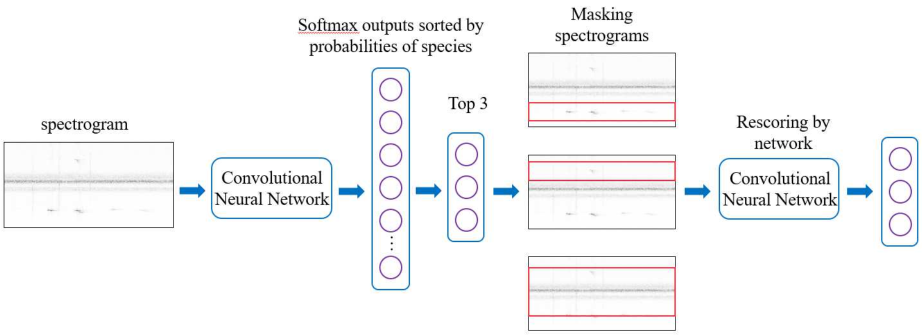
spectrogram
Convolutional Neural Network
Softmax outputs sorted by probabilities of species
Top 3
Masking spectrograms
Rescoring by network
Convolutional Neural Network

Fig. 1. Masking and rescoring method

5 Results

We totally trained 4 models for our classification task. Methods of data augmentation and addition of metadata are introduced above. Besides the basic convolutional neural network, we also trained a Resnet for further improving the final fused results.

1. ConvNet with Data augmentation without metadata addition;
2. ConvNet with Metadata addition without data augmentation;
3. ConvNet with Data augmentation and metadata addition;
4. Resnet with Data augmentation without metadata addition.

This year, a labeled soundscape's evaluation set is given. We are able to test our improvemnts with it. Model 3 is used to test the effect of our methods. From table 1, we can see that masking and rescoring method as well as context considering can improve C-MAP.

Table 1. Evaluation of systems using different methods

Basic	Masking and rescoring	Considering Context	Both methods
0.16942	0.17346	0.23918	0.24508

5.1 Submissions

To fuse different systems, we added the outputs of different models and normalized the final result. Our submissions' details are described below:

Monophone scenario:

DKU_SMIIP run2: The final output of model 1;

DKU_SMIIP run3: Fusion of model 2 and 3;

DKU_SMIIP run4: Fusion of model 1, 2 and 3;
DKU_SMIIP run5: Fusion of model 1, 2, 3, 4.

Table 2. Official scores for monophone

runs	MAP (without background species)	MAP (with background species)
run2	0.5896	0.5278
run3	0.6476	0.5814
run4	0.6541	0.5883
run5	0.6548	0.5882

From table 2, we can find that with increasing of the fused systems, the performance is getting better. As expected, system of run5 has the highest scores on MAP without background species among our submissions.

Soundscape scenario:
DKU_SMIIP run1: The output of model 3;
DKU_SMIIP run2: Fusion of model 2 and 3;
DKU_SMIIP run3: Fusion of model 1, 2 and 3;
DKU_SMIIP run4: Fusion of model 1, 2, 3, 4.

Table 3. Official scores for soundscape

runs	C-MAP (classification mean average precision)
run1	0.1071
run2	0.1161
run3	0.1147
run4	0.1196

In submissions of soundscape scenario, result of run3 is worse than run2 out of expectation. The reason is possibly that the weights of fusion are not properly set. Further exploration should be done on a better fusion method.

6 Conclusion and future work

In this competition, there are two scenarios, monophone and soundscape. We trained models using the convolutional neural network with bird songs' spectrograms. Besides the regular model training, we made data augmentation to improve the robustness. We also added metadata to further improve the performance.

Focusing on soundscape scenario, we made some improvements based on our current models in the test period. Firstly, bird event detection was introduced to reduce false alarm. Secondly, masks were designed for each kind of birds. Rescoring is done on the top 3 or 5 of sorted bird species list after being masked. Thirdly, context is considered by adding outputs of previous and next 5 seconds' segments to current output.

Above methods still have many spaces for improvement. Bird event detection [19] can be done using neural network models if enough labeled data provided. Bandpass filters of birds can be more delicate. In our work, context information is considered using a relatively simple method. During the evaluation, we found that this kind of information can obviously improve the performance. Further investigations need to be done in this direction.

In addition, due to the lack of hardware resources and time, performances of our basic models still have room for improvement. Further, more model structures and fusion methods will be explored.

References

1. Joly, Alexis and Goëau, Hervé and Botella, Christophe, Glotin, Hervé and Bonnet, Pierre and Planqué, Robert and Vellinga, Willem-Pier and Müller, Henning. (2018). Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of AI. In: Proceedings of CLEF 2018.
2. Joly, Alexis and Goëau, Hervé and Glotin, Hervé and Spampinato, Concetto and Bonnet, Pierre and Vellinga, Willem-Pier and Lombardo, Jean-Christophe and Planque, Robert and Palazzo, Simone and Muller, Henning. (2017). LifeCLEF 2017 lab overview: multimedia species identification challenges. In: Proceedings of CLEF 2017.
3. Goëau, Hervé and Glotin, Hervé and Planqué, Robert and Vellinga, Willem-Pier, and Stefan, Kahl, Joly, Alexis. (2018). Overview of BirdCLEF 2018: monophone vs. soundscape bird identification. In: Proceedings of CLEF 2018.
4. Fazekas, B., Schindler, A., Lidy, T., & Rauber, A. (2017). A multi-modal deep neural network approach to bird-song identification. In: Proceedings of CLEF 2017.
5. Sevilla, Antoine, L. Bessonne, and H. Glotin. (2017). Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: Proceedings of CLEF 2017.
6. Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., & Eibl, M. (2017). Large-scale bird sound classification using convolutional neural networks. In: Proceedings of CLEF 2017.
7. Fritzler, A., Koitka, S., & Friedrich, C. M. (2017). Recognizing bird species in audio files using transfer learning. In: Proceedings of CLEF 2017.
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In: Proceedings of CVPR 2016.
9. Sprengel, E., Martin Jaggi, Y. K., & Hofmann, T. (2016). Audio based bird species identification using deep learning techniques. In: Proceedings of CLEF 2016.
10. Lasseck, M. (2013). Bird song classification in field recordings: winning solution for NIPS4B 2013 competition. In Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada (pp. 176-181).
11. Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

12. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
13. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.
14. Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.
15. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
16. Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2), 26-31.
17. Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.
18. Hinton, G. E., & Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. In: Advances in neural information processing systems (pp. 1607-1614).
19. Stowell, D., Wood, M., Stylianou, Y., & Glotin, H. (2016). Bird detection in audio: a survey and a challenge. In: Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on (pp. 1-6). IEEE.