Learning Higher Representations from Bioacoustics: A Sequence-to-Sequence Deep Learning Approach for Bird Sound Classification

Yu Qiao[1], Kun Qian[2(✉)], and Ziping Zhao[1(✉)]

[1] College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
jaderqiao@126.com, ztianjin@126.com
[2] Educational Physiology Laboratory, The University of Tokyo, Tokyo, Japan
qian@p.u-tokyo.ac.jp

Abstract. In the past two decades, a plethora of efforts have been given to the field of automatic classification of bird sounds, which can facilitate a long-term, non-human, and low-energy consumption ubiquitous computing system for monitoring the nature reserve. Nevertheless, human hand-crafted features need numerous domain knowledge, and inevitably make the designing progress time-consuming and expensive. To this line, we propose a sequence-to-sequence deep learning approach for extracting the higher representations automatically from bird sounds without any human expert knowledge. First, we transform the birds sound audio into spectrograms. Subsequently, higher representations were learnt by an autoencoder-based encoder-decoder paradigm combined with the deep recurrent neural networks. Finally, two typical machine learning models are selected to predict the classes, i.e., support vector machines and multi-layer perceptrons. Experimental results demonstrate the effectiveness of the method proposed, which can reach an unweighted average recall (UAR) at 66.8% in recognising 86 species of birds.

Keywords: Sequence-to-sequence learning · Bird sound classification · Bioacoustics · Deep learning · Internet of Things

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61702370), P. R. China, the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300), P. R. China, the Open Projects Program of the National Laboratory of Pattern Recognition, P. R. China, the Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), P. R. China, the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan, and the Grants-in-Aid for Scientific Research (No. 19F19081) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1333, pp. 130–138, 2020.
https://doi.org/10.1007/978-3-030-63823-8_16

1 Introduction

Bird sound recognition refers to the identification of bird species by a given audio. In recent years, the global climate has changed rapidly, and this drastic climate change will lead to a large number of species decrease, which will seriously affect the biological diversity. For this reason, people have come up with many ways to track endangered species. Nevertheless, most of them are expensive for human resources. For instance, observing birds through traditional telescopes can be easily influenced by the weather, which makes the observation of birds less accurate and inconvenient. To overcome the aforementioned challenges, the wireless acoustic sensor networks (WASN) can not only cover the unattended field and/or other places with harsh environment, but also alleviate the influence of weather on bird observation.

In the past decade, numerous efforts have been given to the field of bird sound classification. Many scholars began to use the information implied in bird sound to classify bird species, so as to determine the distribution of birds in a certain area. Large scale acoustic features feeding to an extreme learning machine was introduced in [1,2], which demonstrated an efficient and fast way for recognising bird species by using human hand-crafted features. For machine learning models, SVM was found efficient in previous work [3–5]. There are also applications of convolution neural networks to bird sound recognition. Piczak et al. used convolutional neural networks to do pure audio bird recognition [6]. Three different CNN models were trained according to the difference of time-frequency representations (TFRs): Mel-CNN, Harm-CNN, Perc-CNN. Also, trained a different deep learning framework SubSpectralNet (Subnet-CNN), which is employed to classify bird Sounds. Finally, experiments proved that the performance of classification can be improved by selectively combining the four models separately [7].

In this work, motivated by the success achieved in the field of *natural language processing*, we propose a sequence-to-sequence deep learning model based on recurrent neural network (RNN) for extracting higher representations from the bird sounds without any human domain knowledge. Originally, the sequential to-sequence model is used to deal with speech-to-speech or text-to-speech translation in [8,9]. Similar to the voice of the human, bird sound belongs to a kind of time sequence data, and contains a plenty of semantic information.

For above considerations, sequence-to-sequence structure is introduced to the higher representations learning of bird sounds. As show in Fig. 1, the specific steps include (a) Preprocessing: transform the raw bird sound audio data to spectrograms; (b) Autoencoder Training: the autoencoder-based RNN models is trained by continuously reducing the loss between the prediction sequence and the input sequence; (c) Higher Representations Extraction: the higher representations were learnt by autoencoder-based RNN models; (d) Classifiers Training: a classifier is selected for making the final prediction by the learnt representations. We select two typical machine learning models, i.e., support vector machine (SVM) and multi-layer perceptron (MLP) as the classifiers in this study. (e) Classification Predictions: outputting the results using different classifier.

The main contributions in this work are: Firstly, we introduce the unsupervised sequence-to-sequence deep learning approach to the field of learning higher representations from bird sound. Secondly, we investigate the effect by using different topologies of the deep neural networks. Finally, we analyze and discuss the deep learnt features' performances on recognising bird sounds. We hope this study can facilitate the relevant work in finding more robust and efficient acoustical features from the bioacoustics in future.

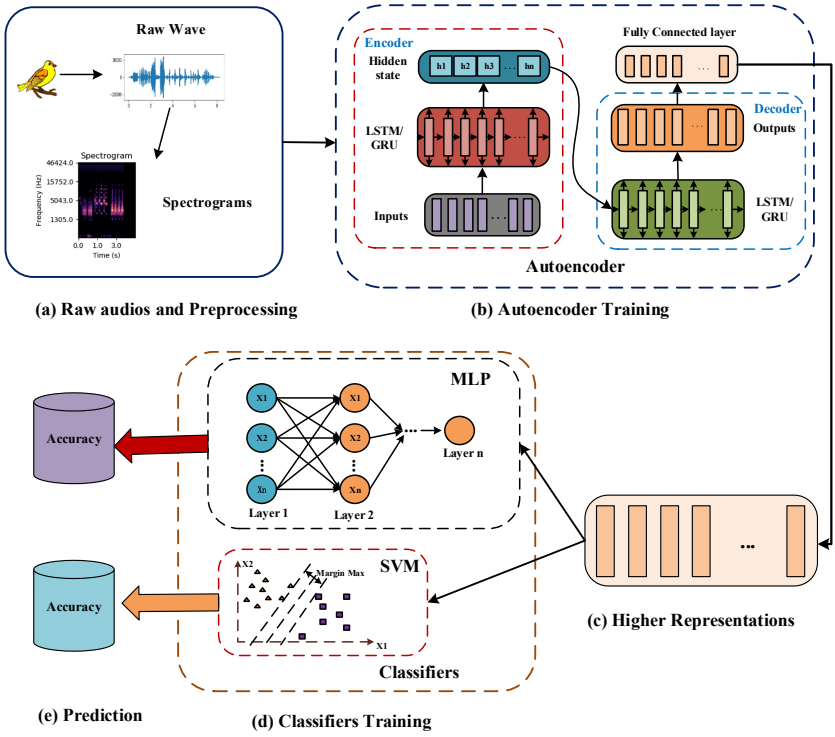

Fig. 1. The framework of proposed Seq2Seq based higher representation learning system for bird sound classification.

This paper is organized as follows: Firstly, we introduce the methods used in Sect. 2. Section 3 introduces experimental design, including description of the database, data preprocessing, experimental setting and results. And the discussion will be given Sect. 4. Finally, we conclude this study in Sect. 5.

2 Methods

2.1 Sequence-to-Sequence Deep Learning Approach

Sequence to Sequence (Seq2Seq) learning was firstly proposed by Kyunghyun Cho et al. [10], which has been demonstrated to be efficient in the field of machine translation and speech recognition [11].

Here, we will describe the underlying framework of RNN Encoder-Decoder briefly, which proposed by Sutskever et al. [12]. In the Encoder-Decoder framework, encoder readers input sequence and transform it into a vector v. Here, we assume $X = (x_1, x_2, ..., x_t)$ as input sequence, and $Y = (y_1, y_2, ..., y_t)$ as output sequence, then

$$\mathrm{h}_t = f(x_t, h_{t-1}) \tag{1}$$

Where f is the nonlinear function of RNN hidden layer, h_t is the hidden state at time t, which is calculated by the input x_t at time t and the hidden state of the previous layer h_{t-1}.

$$\mathrm{v} = q(h_1, h_2, ..., h_t) \tag{2}$$

Where in Eq. (2), encoder converts the hidden state at all moment into a vector v through a nonlinear function q, vector v contains the key information extracted from the input sequence.

The decoder is often trained to predict the output of next time y_t, which is obtain by vector v and all of the previous predictions $y_1, y_2, ..., y_{t-1}$, such as Eq. (3):

$$\begin{aligned} & p(y_1, y_2, ..., y_t | x_1, x_2, ..., x_t) \\ = & \prod_{t=1}^{t} p(y_t | x_1, x_2, ..., x_{t-1}, y_1, y_2, ..., y_{t-1}) \\ = & \prod_{t=1}^{t} p(y_t | v, y_1, y_2, ..., y_{t-1}) \end{aligned} \tag{3}$$

The decoder probability distribution at a given time can be expressed as

$$\prod_{t=1}^{t} p(y_t | v, y_1, y_2, ..., y_{t-1}) = g(h_t, y_{t-1}, v) \tag{4}$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_t, and h_t is the hidden state of the RNN.

Motivated by the success of Seq2Seq, we introduce and propose an autoencoder based RNN model in bird sound classification task.

The Mel spectrum is a time-dependent sequence of frequency vectors, which represents the amplitude of the MEL frequency band of a piece of audio. In

the recurrent autoencoder, the Mel spectrum is fed to the multi-layer encoder RNN firstly, and then updates the hidden state of the encoder according to the input frequency vector. The final hidden state is reconstructed by a full connection layer that contains information about the entire input sequence. Finally, a multi-layer decoder RNN reconstructs the original input sequence utilizing the reconstructed features.

Here, we'll mainly train a Seq2Seq model for the extraction of the higher representations. Our aim is to extract the features of the full connection layer from the trained Seq2Seq model, which is the key for the later retraining of the classification model.

2.2 Evaluation Metrics Method

Considering the imbalanced distribution of the MNB database, we use the unweighted average recall (UAR) as the evaluation metrics for this study. UAR is defined as the averaged recall achieved by the model in recognising different classes. Compared to the conventionally used accuracy, UAR is more rigorous in the case of imbalanced data. For details of UAR, it can be referred to [13].

3 Experimental Design

3.1 Database

In this study, we use the database provided by the Museum für Naturkunde Berlin (MNB)[1], Berlin, Germany. To make an applicable training process, we eliminated the species which contain less than 20 audio recordings, which resulted in a database having 86 species in total (5 060 audio recordings with a whole length of approximately 4.0 h). We split the whole database into three sets, i.e., train (60%), development (20%), and test (20%), respectively. All the hyper-parameters of the classifiers will be tuned and optimised by the dev set, and applied to the final test set.

3.2 Preprocessing

Since the sample time in the database is different, before extracting the spectrograms, we found that the high frequency part of bird song could be included by converting the original audio into 4 s. Therefore, we adopted the following processing: if the time is less than 4 s, fill it according to the silence; instead, it only intercepts to 4 s.

In addition, because the sampling frequencies of bird sounds are not consistent, so according to Nyquist's sampling law: when the sampling frequency f_s is greater than 2 times of the highest frequency f_{max} in the signal, that is $f_s > 2f_{max}$, the sampled digital signal can completely retain the information in the original signal. Based on this law, when extracting the spectrograms, the

[1] http://www.animalsoundarchive.org/RefSys/Statistics.php.

highest frequency of all the audios are controlled to about half of the sampling frequency. In this way, the extracted spectrograms contain the information of the high-frequency part of the bird sound, so that the extracted features can contain more effective information, thus ensuring the accuracy of subsequent training.

In order to reduce the influence of noise on the classification results, when extracting the features of the spectrograms, we found that it was better to control the amplitude below -50 db.

To conclude, the raw audio data of bird sound will be transformed to spectrograms with the window width w $= 0.08$ s, the window overlap $0.5\text{w} = 0.04$ s, and $\text{N}_{mel} = 128$ Mel frequency bands, with amplitude clipping below -50 db.

3.3 Experimental Setting

In the phase of Seq2Seq learning, we used the open source toolkit, i.e., auDeep [14,15]. When investigating the topologies of the deep learning models, we firstly study the long short-term memory (LSTM) [12] and the gated recurrent unit (GRU) [16] based RNNs. Then, we compare the different Encoder-Decoder structures with the combinations of the unidirectional RNN and the bidirectional RNN (BiRNN). Additionally, we change the hidden layer numbers with 2, 3, or 4 to find the differences in capacity of learning higher representations. Our experiment is going to be performed for 64 batch size, learning rate 0.001 and 20% dropout.

When tuning the hyper-parameters of the models, we use a grid searching strategy in development set and apply the optimised values to the test set. For SVM, the kernels are selected from *linear*, *radical basis function (RBF)*, *poly*, and *sigmoid*. The *Gama* and C values are all tunned from 10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}. For MLP, the *Alpha* value is tuned as the same grid as *Gama* and C values. The hidden layer structures are optimised from [(500, 500, 500), (600, 600, 600), (650, 650, 650), (700, 700, 700), (750, 750, 750), (800, 800, 800), (850, 850, 850), (900, 900, 900), (950, 950, 950), (1000, 1000, 1000), (1200, 1200, 1200)]. Both of the SVM and the MLP models are implemented in Python script based on the scikit-learn library [17]. To eliminate the effects of outliers, all of the features are standardised before fed into the classifiers.

3.4 Experimental Results

By adjusting the topologies of the autoencoder, network depth and various parameters of the classifiers, the best parameters of the final experiment are shown in the Table 1.

The results using LSTM and GRU based RNN models (two hidden layers) are shown in Table 2 and Table 3, respectively. In this study, a two hidden layer GRU (BiRNN-BiRNN as the Encoder-Decoder) based model can reach the best performance. In particular, when fed into a MLP classifier, the UAR can be reaching at 66.8% for recognising totally 86 species of birds.

Table 1. The parameters of final model.

Hyperparameter	Value
RNN cell	GRU
Encoder depth	2
Decoder depth	2
Encoder	Bidirectional
Decoder	Bidirectional
Kernel	*rbf*
C	100
Alpha	0.1
MLP hidden layers	(850, 850, 850)

Table 2. The results (UAR: %) achieved by two-layer LSTM RNN models.

Encoder	Decoder	SVM		MLP	
		Dev	Test	Dev	Test
RNN	RNN	33.2	33.7	34.6	31.9
BiRNN	RNN	27.8	30.3	25.7	26.9
BiRNN	BiRNN	55.1	**52.8**	49.0	**46.8**
RNN	BiRNN	22.8	22.6	29.3	25.3

Figure 2 illustrates the comparison between different topologies of the models. It is demonstrated that, a two-layer BiRNN-BiRNN structure can be the best option in this work.

4 Discussion

As a pilot study on using Seq2Seq deep learning approach to extract higher representations from the bird sounds, we can find that, it is feasible to build an efficient framework for recognising bird sounds without any human hand-crafted features. In addition, we may find that, the selection of the deep learning topologies can effect the final model's performances (see Table 2 and Table 3). Among the experimental results in this study, GRU based RNN can be superior to LSTM based RNN in learning higher representations from the bird sounds (a significance level at $p < 0.001$ by one-tailed z-test). When adding the hidden layers of the RNN models, we may find a decrease in final performance (see Fig. 2). It is reasonable to think that due to the current limited size of the database, the model seems to be vulnerable to be over-fitting. In future work, we will implement our approach in larger size bird sound databases. An interesting finding is that, when introducing the BiRNN structure, the performance can be improved (see Table 2 and Table 3). Similar to human speech, bird sound may also have the strong contextual information, which can be extracted not only

Table 3. The results (UAR: %) achieved by two-layer GRU RNN models.

Encoder	Decoder	SVM		MLP	
		Dev	Test	Dev	Test
RNN	RNN	62.0	60.3	60.9	58.9
BiRNN	RNN	64.4	62.4	60.6	58.9
BiRNN	BiRNN	68.0	**65.7**	63.3	**66.8**
RNN	BiRNN	62.3	62.2	62.4	62.1

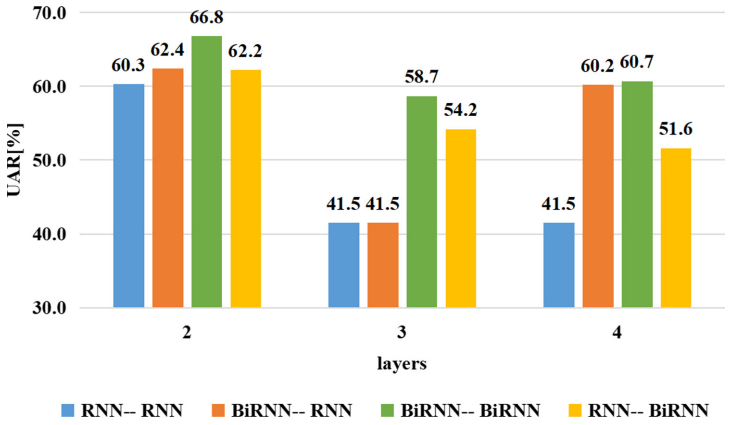

Fig. 2. The results (UARs: %) achieved by different topologies of the proposed model (GRU RNN) evaluated by test set.

from the *forward* direction, but also the *backward* direction. We should make efforts towards finding the contextual information through deeply understanding of the bird vocalisations. Finally, when comparing the classifiers' ability to make the final predictions, we find both of the two machine learning models, i.e., SVM and MLP, can be sufficient to fulfil the task.

5 Conclusion

In this work, we proposed a Seq2Seq deep learning approach for automatically extracting higher representations from bird sounds. The proposed method was demonstrated to be efficient to utilize longer term temporal information and achieved 66.8% of UAR. Moreover, we investigated the effects to the final classification performance by using different deep learning topologies. We found that, a BiRNN-BiRNN structure can reach the highest performance in this study. Future work can be given to the direction of combining the convolutional neural networks and autoencoders to extract more advanced features from the birds' vocalisation. In addition, it is our interest to contribute more to understand in depth about the relationship between the learnt representations and the birds' behaviour activities.

References

1. Qian, K., Zhang, Z., Ringeval, F., Schuller, B.: Bird sounds classification by large scale acoustic features and extreme learning machine. In: Proceedings of GlobalSIP, Orlando, Florida, USA, pp. 1317–1321. IEEE (2015)
2. Qian, K., Guo, J., Ishida, K., Matsuoka, S.: Fast recognition of bird sounds using extreme learning machines. IEEE Trans. Electr. Electron. Eng. **12**(2), 294–296 (2017)
3. Papadopoulos, T., Roberts, S.J., Willis, K.J.: Automated bird sound recognition in realistic settings (2018)
4. Kaewtip, K.: Robust automatic recognition of birdsongs and human speech: a template-based approach. Ph.D. thesis, UCLA (2017)
5. Bang, A.V., Rege, P.P.: Evaluation of various feature sets and feature selection towards automatic recognition of bird species. Int. J. Comput. Appl. Technol. **56**(3), 172–184 (2017)
6. Piczak, K.J.: Recognizing bird species in audio recordings using deep convolutional neural networks. In: Proceedings of International Conference on Genetic & Evolutionary Computing, Fujian, China, pp. 534–543. IEEE (2016)
7. Xie, J., Hu, K., Zhu, M., Yu, J., Zhu, Q.: Investigation of different CNN-based models for improved bird sound classification. IEEE Access **7**(8922774), 175353–175361 (2019)
8. Jia, Y., et al.: Direct speech-to-speech translation with a sequence-to-sequence model. In: Proceedings of Interspeech, Graz, Austria, pp. 1–5. ISCA (2019)
9. Okamoto, T., Toda, T., Shiga, Y., Kawai, H.: Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders. In: Proceedings of Interspeech, Graz, Austria, pp. 1308–1312. ISCA (2019)
10. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of EMNLP, Doha, Qatar, pp. 1724–1734. Association for Computational Linguistics (2014)
11. Cho, K., Van Merrinboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of SSST-8, Doha, Qatar, pp. 103–111 Association for Computational Linguistics (2014)
12. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, pp. 3104–3112. MIT Press (2014)
13. Qian, K.: Automatic general audio signal classification. Ph.D. thesis, Munich, Germany (2018). Doctoral thesis
14. Amiriparian, S., Freitag, M., Cummins, N., Schuller, B.: Sequence to sequence autoencoders for unsupervised representation learning from audio. In: Proceedings of the DCASE 2017 Workshop, Munich, Germany, pp. 17–21. IEEE (2017)
15. Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., Schuller, B.: auDeep: unsupervised learning of representations from audio with deep recurrent neural networks. J. Mach. Learn. Res. **18**(1), 6340–6344 (2017)
16. Deng, Y., Wang, L., Jia, H., Tong, X., Li, F.: A sequence-to-sequence deep learning architecture based on bidirectional GRU for type recognition and time location of combined power quality disturbance. IEEE Trans. Industr. Inf. **15**(8), 4481–4493 (2019)
17. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)