Interspeech 2018
2-6 September 2018, Hyderabad, India

All-Conv Net for Bird Activity Detection: Significance of Learned Pooling

Arjun Pankajakshan, Anshul Thakur, Daksh Thapar, Padmanabhan Rajan & Aditya Nigam

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi
arjunp@projects.iitmandi.ac.in, {anshul_thakur,s16007}@students.iitmandi.ac.in,
{padman,aditya}@iitmandi.ac.in

Abstract

Bird activity detection (BAD) deals with the task of predicting the presence or absence of bird vocalizations in a given audio recording. In this paper, we propose an all-convolutional neural network (all-conv net) for bird activity detection. All the layers of this network including pooling and dense layers are implemented using convolution operations. The pooling operation implemented by convolution is termed as learned pooling. This learned pooling takes into account the inter feature-map correlations which are ignored in traditional max-pooling. This helps in learning a pooling function which aggregates the complementary information in various feature maps, leading to better bird activity detection. Experimental observations confirm this hypothesis. The performance of the proposed all-conv net is evaluated on the BAD Challenge 2017 dataset. The proposed all-conv net achieves state-of-art performance with a simple architecture and does not employ any data pre-processing or data augmentation techniques.

Index Terms: bird activity detection, all-conv net, learned pooling

1. Introduction

Manual monitoring of birds can be a tedious and difficult task due to the wide range of habitats, such as islands, marshes and swamps, occupied by different bird species [1]. Many bird species are nocturnal which makes it less feasible to manually monitor them. Moreover, it requires an experienced bird watcher to accurately identify the bird species. Acoustic monitoring bypasses the need of manual labour and provides a convenient way to monitor birds in their natural habitats without any physical intrusion [2].

Recently, there is a boom in sophisticated automatic recording devices which can be programmed to record audio data for many days. This vast amount of audio data can be used for acoustic monitoring, which can give insight into avian diversity, migration patterns and population trends in different bird species. Bird activity detection (BAD) [3] is generally the first module in any bioacoustic monitoring system. BAD distinguishes the audio recordings having bird vocalizations from those recordings which do not contain any bird call or song. Hence, BAD helps in removing the audio recordings which do not contain any bird vocalizations from further processing. This reduces the amount of data to be processed for other tasks such as bird vocalization segmentation and species identification.

The task of BAD becomes challenging due to the variations present in bird vocalizations across different bird species. These variations can be due to the different frequency profiles or different frequency-temporal modulations of different species. This behaviour is more evident in the vocalizations of *Emerald dove* and *Cassin's vireo*. The sounds produced by *Emerald dove* are characterized by low-frequency and have little variation in frequency along time. On the other hand, the vocalizations of *Cassin's vireo* relatively occupy high frequency bands and exhibit different kinds of frequency and temporal modulations. These variations present in bird sounds make it difficult to model the bird vocalizations class effectively. Apart from this, the background of audio recordings is also unpredictable. Different biotic and abiotic non-bird sounds can form the background in any audio recording. Hence, the task of BAD can be seen as the classification between the universe of bird sounds and the universe of non-bird sounds. The extreme variations present in both the sets make this simple two-class classification problem challenging. An ideal bird activity detector should work well across different bird species in different recording environments.

Many studies in the literature have targeted the task of BAD. In our earlier studies [4, 5], the frameworks based on SVM powered with dynamic kernels are proposed for the task of BAD. In [4], a GMM based variant of the existing probabilistic sequence kernels is proposed while in [5], an archetypal analysis based sequence kernel is successfully utilized for BAD. Both these frameworks require less amount of training data and generalize well on unseen data. However, the classification performance of these frameworks is not up-to the level of state-of-art BAD frameworks. In [6], a masked non-negative matrix factorization (Masked NMF) approach for bird detection is proposed. Masked NMF can handle weak labels by applying a binary mask on the activation matrix during dictionary learning. In [7], a convolutional recurrent neural network (CRNN) for BAD is proposed. Convolutional layers of this network extract local frequency shift invariant features while recurrent layers learn the long-term temporal relations between the features extracted from short-term frames. The classification performance of this network is comparable to the *state-of-art* methods on the evaluation data of the BAD 2017 dataset. This method does not utilize any pre-processing or data augmentation. However, the recurrent layers are computationally expensive to train which increases the overall computational time required to train the CRNN. In [8], two CNNs (*sparrow* and *bulbul*) are proposed. The ensemble of these two networks provide state-of-art bird activity detection on this dataset.

Inspired by the success of CNN-based architectures for BAD and simplicity of all-conv neural network proposed in [9], we propose an all-convolutional neural network (all-conv net) for BAD. This network is characterized by alternating convolution and pooling layers followed by dense layers, where each of these layers are implemented by the convolution operation itself. Instead of max-pooling, the local features obtained using a convolution layer are pooled using a learned aggregation, implemented using convolution operations. This aggregation function is referred as learned pooling, which aggregates the contemporary information present in different feature-maps. On the contrary, max-pool operations aggregates the informa-

10.21437/Interspeech.2018-1522

tion present in each feature-map individually (see section 2). This behaviour of learned pooling helps in obtaining better discriminating features, leading to a better classification performance in comparison to the normal max-pooling. Moreover, the proposed all-conv net has lesser number of trainable parameters in comparison to the existing state-of-art neural networks for the task in hand and does not utilize any pre-processing and data augmentation.

The rest of this paper is organized as follows. In Section 2, the proposed deep all-conv framework is described in detail. Performance analysis and conclusion are in sections 3 and 4, respectively.

2. Proposed Framework

In this section, we describe the proposed all-conv net for BAD. First, we describe Mel-spectrogram, a spatio-temporal feature representation obtained from the raw audio recordings. Then, we describe the proposed all-conv net which maps the input Mel-spectrogram to a two dimensional probabilistic score vector containing probabilities for the presence and absence of any bird activity.

2.1. Feature representation

Mel-spectrogram, a spatio-temporal feature representation, obtained from the audio recordings are used as input to the proposed all-conv net. Short-term Fourier transform (STFT) is employed to obtain the spectrogram from the input audio recordings. The audio signal is windowed using a frame size of 20 ms with no overlap and a Hamming window. 882 FFT points are used to obtain the Fourier transform of an audio frame. This process converts a 10 second (*duration of recordings in the BAD 2017 dataset*) audio recording into a 441×500 dimensional spectrogram representation. Each frame of this spectrogram is converted into a 40-dimensional vector of log filter bank energies using Mel filter-bank [7]. Hence, each 10-second audio recording is represented by a 40×500 dimensional Mel-spectrogram.

2.2. All-conv net for BAD

The proposed architecture consists of four pairs of convolution and learned pooling layers followed by two $(1, 1)$ convolution layers. The input to the network is a 40×500 Mel-spectrogram. A kernel size of 5×5 with a stride of 1×1 is used in the convolutional layers. To avoid over-fitting, as data is not abundant, the proposed architecture has only 16 filters in each convolutional layers. In order to pool the feature maps, the convolution with kernel size of 5×1 and 2×1 with a stride of 5×1 and 2×1 respectively is used at the subsequent layers. In the later part of the network, we have two 1×1 convolutional layers with 196 and 2 filters respectively. The output of these layers is a two-dimensional feature map, which is converted into a probabilistic score vector using soft-max function. The elements of this vector signify the probability of the presence and absence of the bird activity in any input Mel-spectrogram.

For regularization, batch normalization [10] has been used after convolutional layers and dropout [11] with the probability of 0.25 and 0.5 has been used after convolutional layers and learned pooling layers respectively. The weights of the network has been initialized using random normal distribution. The network is optimized using Adam optimizer [12] with a learning rate of 0.001 and a decay of 10^{-6} and binary cross-entropy as the loss function. The network is shown in Fig. 1. An exhaustive grid search is applied to establish the various parameters such as the number of convolutional layers, the number of filters, the size of filter kernel and drop-out rates. Following are the main highlights of the proposed network:

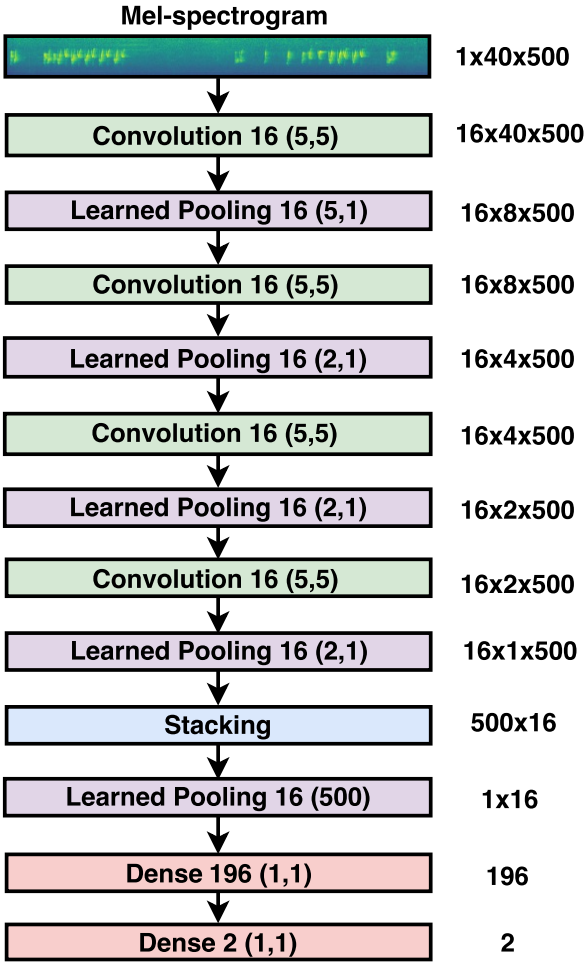

Figure 1: *Proposed all-conv architecture for bird activity detection*

Input: The input to the network is Mel-spectrogram of dimensions 40×500, extracted from audio wave files (see Section 2.1).

Learned pooling: Replacing max-pooling with convolution pooling layers incorporates the inter-feature dependencies. To pool the feature maps, we utilise strided convolutions instead of max-pooling. Throughout the learning process, the temporal dimension is kept constant. Although, for BAD, preserving temporal dimension may not make any difference. But for other applications which are extension of BAD such as bird activity/event segmentation, it is essential to preserve the time information. By the series of convolution and learned pooling operations, the input data has been transformed into a vector with 16 unique spatial feature representations of size $16 \times 1 \times 500$. Intuitively, each of these 1×500 feature vectors can be interpreted as a position vector which points to certain regions in the input Mel-spectrogram. The elements of these position vectors are coordinates of the peak responses learned by corresponding

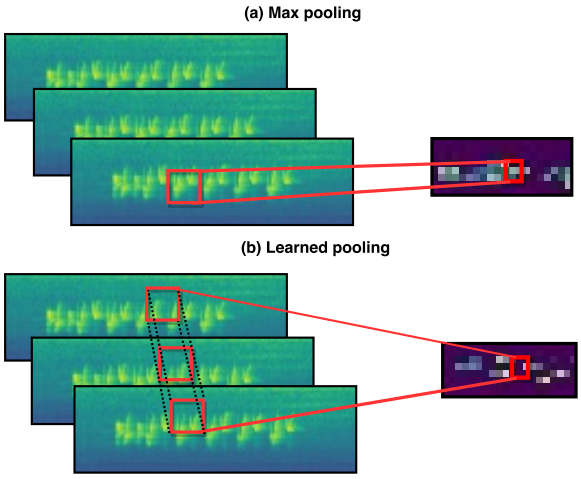

Figure 2: *Illustration of the difference between (a) max pooling and (b) learned pooling.*

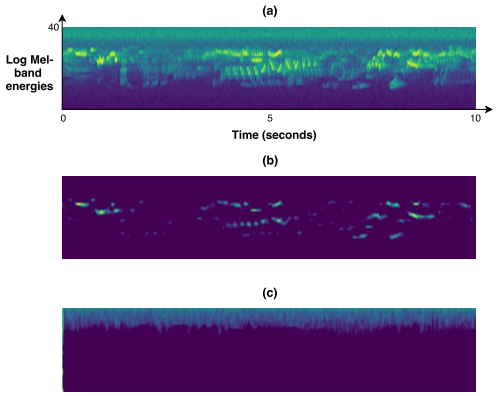

Figure 3: ***(a)*** *Mel-spectrogram of an audio recording containing bird activity* ***(b)*** *Response of the* 8*th filter, learned in the first convolution layer, for the input Mel-spectrogram shown in (a)* ***(c)*** *Response of the* 11*th filter, learned in the first convolution layer, for the input Mel-spectrogram shown in (a).*

filters.

After this stage, these 16 feature vectors are stacked together to form a new $2D$ feature representation of size 500×16 corresponding to each input Mel-spectrogram. Temporal pooling is applied on this $2D$ representation to obtain a 16-dimensional feature vector. This temporal pooling is implemented using $1D$ strided convolution. This helps in compressing the $2D$ matrix representation into a low dimensional feature representation.

Convolutional fully connected layers: The fully connected layers at the end of the network are implemented by using 1×1 convolutions in which the number of filters in each of the convolution layers are equivalent to the number of neurons in a fully connected layer. The first layer consists of 196 filters and finally, we reduce our feature map to 2 dimensions to get the required probability of the presence/absence of the bird activity.

Activations: ReLU (Rectified Linear Unit) activation has been applied over all the convolutional and the learned pooling layers. While on the fully connected layer of 196 filters (which has been implemented using convolutional layer), we have applied sigmoid activation. Softmax is applied over the final 2-dimensional output of the network to get the probabilities of presence/absence of bird activity.

2.3. Learned pooling vs. max-pooling

As discussed earlier, learned pooling learns an aggregation function that utilizes the correlations among different feature maps. On the other hand, max-pool does not take into consideration these inter-feature map relations. In max-pool, each feature map is processed individually and any aggregated feature map is a function of only one input feature map. However, in learned pooling, an aggregated feature map is a function of multiple input feature maps. This behaviour is illustrated in Fig. 2. The analysis of the filters learned at the first convolution layer of the proposed all-conv net confirm that the information learned by each filter can be complementary to other filters. This is illustrated in Fig. 3. The analysis of Fig. 3(b) illustrates that the 8th filter of the first convolution layer of the proposed all-conv net is only concerned with learning the bird vocalizations. On the contrary, 11th filter is learning the background information (Fig. 3(c)). Thus, it can be deduced that each filter is learning a different behaviour or event. The utilization of this complementary information in the aggregation function can lead to more discriminative features. The learned pooling learns this sort of aggregation, leading to better bird activity detection (see Section 3).

2.4. Discriminative nature of the learned features

The first 10 layers of the proposed all-conv net can be seen as a feature learner while the last 2-dense layers act as a classifier. In this subsection, we analyse the discriminative ability of the 16-dimensional features generated by the network at the 10th layer (*before the dense layers*). Fig. 4 shows the feature representations for five different audio files present in the BAD 2017 dataset. These files cover five different cases i.e. first one contains a birdcall mimicked by a human (*fake birdcall*), second contains the background without any birdcall (*pure noise*), third one contains birdcalls (*pure birdcall*), fourth one has both speech and birdcall (*speech+birdcall*) and fifth one contains music only (*music*). The analysis of Fig. 4 highlights the ability of the proposed all-conv net to differentiate between bird activity and various other acoustic events. The feature coefficients corresponding to the first few filters (i.e. $1, 3, 4$ and 6) exhibit high magnitudes for birdcalls only. As a result, a clear discrimination between birdcalls and other sounds, such as speech and music, is evident in these features. This highlights the effectiveness of the proposed all-conv net for the task in hand.

3. Performance Analysis

3.1. Dataset and evaluation metric

The performance of the proposed all conv-net is evaluated on the BAD challenge 2017 dataset [13]. The dataset is divided into two parts: development and evaluation. The development dataset has 15690 audio recordings, out of which 7710 has bird activity. The evaluation dataset has 8620 recordings. All of these recordings are 10 seconds long and are sampled at 44.1 kHz. More information about the source and recording conditions of BAD 2017 dataset can be found at [13]. The 80% of the development data is used for training the network while rest

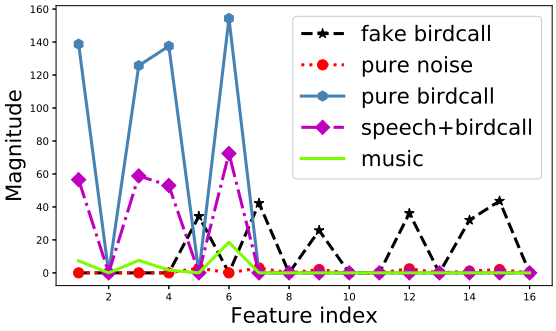

Figure 4: *16-dimensional feature vectors obtained before the dense layers*

of the 20% is used for validation purposes. The number of examples in both bird and non-bird classes are forced to be the same.

The area under the Receiver Operating Characteristic curve (AUC) [14] is used as a metric to evaluate the performance of the proposed network.

3.2. Comparative methods

The performance of the proposed all-conv net is compared with various other methods proposed for BAD on the evaluation dataset of BAD challenge 2017. We have used rapid probabilistic sequence kernels (RPSK) [4], masked NMF (MNMF) [6], RCNN [7], *bulbul* and *sparrow* networks [8] and DenseNets [15] as comparative methods. Apart from the methods listed here, the performance of the proposed all-conv net can be compared with all the entries of the BAD 2017 challenge. The performance of all the entries on the evaluation dataset is listed at [16].

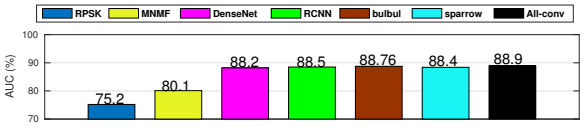

Figure 5: *Classification performance of the proposed all-conv net along with various other methods.*

3.3. Classification performance

The classification performance of the proposed all-conv net along with various comparative methods is depicted in Fig. 5. The analysis of this figure makes it clear that non-deep learning methods (RPSK and MNMF) are not able to perform up-to the level of deep learning based frameworks. The performance of all-conv net is comparable to the state-of-art deep learning frameworks including *bulbul*-net which is the winning entry of BAD challenge 2017. Also, it must be noted that the proposed all-conv net does not utilize any pre-processing and data augmentation. Incorporating these steps in the proposed architecture may lead to an improvement in the classification performance.

Number of trainable parameters: Table 1 shows the number of trainable parameters in the deep frameworks considered in this study. Although the comparative deep learning frameworks considered in this study show comparable performances to the proposed all-conv net, the number of trainable parameters in the all-conv net is significantly lesser than the other methods.

Table 1: *Number of training parameters in various deep learning frameworks considered in this study*

Framework	Trainable parameters
DenseNet [15]	328004
RCNN [7]	806000
Bulbul [8]	373169
Sparrow [8]	309843
All-conv (Proposed)	154414

Input shape (1x40x500)

All-conv net	Max-pool variant
5x5 Conv2D 16 ReLU	5x5 Conv2D 16 ReLU
5x1 Conv2D 16 ReLU	5x1 MaxPool2D
5x5 Conv2D 16 ReLU	5x5 Conv2D 16 ReLU
2x1 Conv2D 16 ReLU	2x1 MaxPool2D
5x5 Conv2D 16 ReLU	5x5 Conv2D 16 ReLU
2x1 Conv2D 16 ReLU	2x1 MaxPool2D
5x5 Conv2D 16 ReLU	5x5 Conv2D 16 ReLU
2x1 Conv2D 16 ReLU	2x1 MaxPool2D
500 Conv1D 16 ReLU	500 MaxPool1D
1x1 Conv2D 196 sigmoid	Dense 196 sigmoid
1x1 Conv2D 2 softmax	Dense 2 softmax
AUC 88.91	85.01

Figure 6: *Architecture of the all-conv net and its max-pool variant.*

Effect of max-pooling and learned pooling on the classification performance: To analyse the effect of learned pooling on the classification performance, we replaced each learned pooling layer of the proposed all-conv net with max-pooling layers. This max-pool variant of the proposed network is shown in Fig. 6. This replacement of the learned pooling with max-pooling showed a relative drop of 4.39% (from 88.9% to 85.01%) in AUC. This justifies the claim that the utilization of the complimentary feature-maps information during pooling leads to the discriminative feature representations which provides better classification performance.

4. Conclusion

In this paper, we proposed an all-conv net for bird activity detection. This network is characterized by the use of convolution operations for implementing all the layers including pooling and the fully connected layers. The performance of this proposed network is comparable to the state-of-art bird activity detection methods. This paper also highlights the use of learned pooling over max-pooling in the proposed network. The experimental observations verifies the superiority of the learned pooling over traditional max-pooling for the task of bird activity detection. Future work may include exploring the learned pooling on other types of the acoustic classification tasks to see if the trends observed in bird activity detection replicate themselves on the other tasks.

5. References

[1] T. S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, no. S1, pp. S163–S173, 2008.

[2] C.-H. Lee, C.-C. Han, and C.-C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1541–1550, 2008.

[3] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *IEEE Int. Workshop Mach. Learn. Sig. Process.*, 2016, pp. 1–6.

[4] A. Thakur, R. Jyothi, P. Rajan, and A. D. Dileep, "Archetypal analysis based sparse convex sequence kernel for bird activity detection," in *Proceedings of Eusipco*, Aug., 2017, pp. 1754–1758.

[5] V. Abrol, P. Sharma, A. Thakur, P. Rajan, A. D. Dileep, and A. K. Sao, "Archetypal analysis based sparse convex sequence kernel for bird activity detection," in *Proceedings of Eusipco*, Aug., 2017, pp. 4436–4440.

[6] I. Sobieraj, Q. Kong, and M. Plumbley, "Masked non-negative matrix factorization for bird detection using weakly labelled data," in *Proc. Eusipco*, 2017, pp. 1819–1823.

[7] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *Proc. Eusipco*, 2017, pp. 1744–1748.

[8] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *Proc. Eusipco*, 2017, pp. 1764–1768.

[9] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. ICLR*, 2015.

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[12] D. P. Kingma and L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[13] "BAD challenge," http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/, accessed: 2017-2-1.

[14] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[15] T. Pellegrini, "Densely connected cnns for bird audio detection," in *Proceedings of Eusipco*, 2017, pp. 1734–1738.

[16] "BAD Challenge 2017 Results," http://c4dm.eecs.qmul.ac.uk/events/badchallenge_results/, accessed: 2018-02-20.