Interspeech 2018
2-6 September 2018, Hyderabad

Deep learning techniques for koala activity detection

Ivan Himawan[⋆] *Michael Towsey*[⋆] *Bradley Law*[†] *Paul Roe*[⋆]

[⋆] Queensland University of Technology, Brisbane, Australia
[†]Forest Science Unit, NSW Department of Industry-Lands, Australia
{i.himawan, m.towsey}@qut.edu.au, brad.law@dpi.nsw.gov.au, p.roe@qut.edu.au

Abstract

Automatically detecting koalas in the real-life environment from audio recordings will immensely help ecologists, conservation groups, and government departments interested in their preservation and the protection of their habitat. Inspired by the success of deep learning approaches in various audio classification tasks, in this paper, the feasibility of recognizing koalas' calls using a convolutional recurrent neural network architecture (CNN+RNN) is studied. The benefit of this architecture is twofold: firstly, convolutional layers learn local time-frequency patterns from the audio spectrogram and secondly, recurrent layers model longer temporal dependencies of the extracted features. In our datasets, the performance of CNN+RNN is evaluated and compared with standard convolutional neural networks (CNNs). The experimental results show that hybrid CNN+RNN architecture is beneficial for learning long-term patterns in spectrogram exhibited by koalas' calls in unseen conditions. The proposed method is also applicable for detecting other animal calls such as bird sound where it achieves 87.46% area under curve score on the bird audio detection challenge evaluation data.

Index Terms: koalas' call, deep learning, convolutional recurrent neural network, sound classification, constant Q transform

1. Introduction

The implications of expanding urbanization in many parts of the world are affecting biodiversity. In South East Queensland, for example, increasing human population density was identified as the main factor in the decline of populations of koalas (*Phascolarctos cinereus*) [1, 2]. A recent Australian Senate inquiry (Commonwealth of Australia, 2011) recommended the implementation of habitat mapping to assist in the management of koalas, highlighting the need for reliable distribution models for this species. To monitor wildlife, ecologists often use acoustic sensors as an effective approach to collect data at large spatiotemporal scales. The recorded acoustic data provide the means for ecologists to identify particular species and to conduct species-richness surveys based on the animals' mating calls in their habitat. In order to perform classification tasks from recordings, trained professionals need to listen to a large batch of recordings over many hours. Therefore, tools which help to automatically identify koalas' calls can improve the efficiency of monitoring.

Automatic detection of animal vocalizations, such as mating calls and birdsong, has been the subject of intensive research over decades. Recently, bioacoustics or ecoacoustics [3], has become one of the "big data" research areas. With the proliferation of high-quality acoustic sensors (i.e., microphones) that can be mounted on smartphones and robots, an ever-growing quantity of recordings is being generated, far more than can feasibly be analyzed manually. Detecting bird sounds in audio recordings is one example of such remote monitoring projects. Bird species are good indicators of environmental health and easily detectable by sound rather than by vision [4]. The use of detection and classification of other animals, such as frogs [5, 6], bats [7], and marine mammals [8], using acoustic recordings has also been studied extensively for the conservation of natural ecosystems.

A variety of machine learning techniques have been explored for automatic detection of animal vocalizations from acoustic recordings. Typically, acoustic features are extracted from waveform envelopes or spectrograms into representations that best characterize the signal. These features can be broadly categorized into temporal and spectral features that summarize the content of ecological interest. For example, Towsey et al. [9] extract fourteen acoustic indices, with varying degrees of correlation with bioacoustic activity, which are relevant for determining bird species richness in audio recordings. These acoustic indices, along with features derived from spectrogram images, are also used to detect and classify frog calls [10]. Other techniques for detection are based on energy thresholding, spectrogram cross-correlation, and Hidden Markov Models (HMMs), which are widely-employed in bioacoustic software (SongScope, XBAT, Raven) [4].

Recently, deep learning techniques have revolutionized the applicability of machine learning in speech, vision, and text processing. A large volume of data is usually required for training high-dimensionality models using deep neural networks, which have many hidden layers and millions of parameters, in order to achieve state-of-the-art classification accuracy. Indeed, ecoacoustics research may benefit substantially from "big data" analysis by applying deep learning models for detection and classification tasks. For example, the winning entry in the BirdCLEF [11], a yearly contest on methods for bird-species identification, was based on deep convolutional neural networks (CNNs). Also, several top-performing systems employed CNNs as solutions in the bird audio detection challenge 2017 [4] which deal with the estimation of the presence/absence of bird sounds in short-duration recordings [12, 13, 14].

In this paper, we investigate the feasibilty of recognizing koala calls using deep learning architectures. Male koala calls consist of a single repeating or oscillating element with a series of inhalations and exhalations at lower frequencies lasting for 30 seconds or more [15]. Structurally similar calls are observed in female koala calls [16]. Specifically, we propose a combined convolutional and recurrent neural networks (RNNs) architecture for this task. The main idea of this integration is to use CNNs as a feature extractor and recurrent layers to model the long-term dependencies. Similar combined CNN+RNN architectures have also been proposed in automatic speech recognition [17], speaker verification antispoofing [18], as well as music classification [19]. Our research led us to extracting features using the constant Q transform (CQT), a perceptually-

10.21437/Interspeech.2018-1143

inspired approach to time-frequency analysis, which captures low frequency at higher frequency resolution.

This paper is organized as follows. Section 2 presents related works. Section 3 presents datasets, feature extraction method, and deep learning architecture proposed in this study. Section 4 and 5 describe the experimental setup and results. This is followed by conclusions in Section 6.

2. Relation to prior works

Compared to the traditional machine learning (ML) techniques, the use of CNNs eliminates the need for feature extraction which is the most important and time-consuming part of detection and classification tasks. However, in practice, the automatic feature learning approach (i.e., train using raw audio as input in order to autonomously discover feature representations) does not outperform the system trained on mid-level representation of audio (e.g., spectrogram) [20].

CNNs differ from traditional networks by learning filters that are shifted in both frequency and time by making the explicit assumption that the input data is an image. However, it lacks longer temporal context information which is beneficial for processing patterns with temporal characteristics. This shortcoming is addressed by integrating CNNs for local feature extraction and RNNs into a single network to learn the temporal information of the extracted features [21].

3. Data and Methods

3.1. Datasets

Koala males emit loud bellows during the breeding season, and this behavior can be used for estimating occupancy [22]. The raw acoustic data were collected from 63 sites from *Willi Willi* National Park in New South Wales (NSW) during night-time. At each site, one SongMeter (SM2 Wildlife Acoustics) is deployed to record koala bellows. The koala calls are considered to be detectable by SongMeters at up to at least 100 m. The raw acoustic data (recorded at a 22,050 Hz sampling rate) are then annotated with the presence or absence of koalas. In total, we used about 3.6 hours of annotated koala calls as positive examples and more than seven hours of other audio clips as negative examples (e.g., noise, crickets, frog and bird calls, vehicles) in datasets. We created a uniform six-second-long audio clip from the dataset in order to generate input features. This produces 2181 and 4337 clips for the positive and the negative classes, respectively. The datasets are split into a training and a test set in the proportion of 80% and 20%, respectively.

Time and frequency shifts were applied as data augmentation techniques to artificially enlarge the training dataset to reduce overfitting. For the time shift, the spectrogram is cut into two parts by a small random amount in time for the second part and the second part is placed in front of the first. Concerning the frequency shift, a small shift in spectra (4 bins) is applied.

3.2. Feature extraction

The constant Q transform employs geometrically spaced frequency bins with Q-factors (ratios of the centre frequencies to bandwidth) are all equal across the entire spectrum [23]. This results in the time-frequency signal representation with higher frequency resolution at lower frequencies while providing a higher temporal resolution at higher frequencies. The Fourier-based approaches on the other hand, lack frequency resolution at lower frequencies and lack temporal resolution at higher frequencies due to the increasing Q-factor when moving from low to high frequencies. Note that the CQT is essentially a wavelet transform with high Q factors [24].

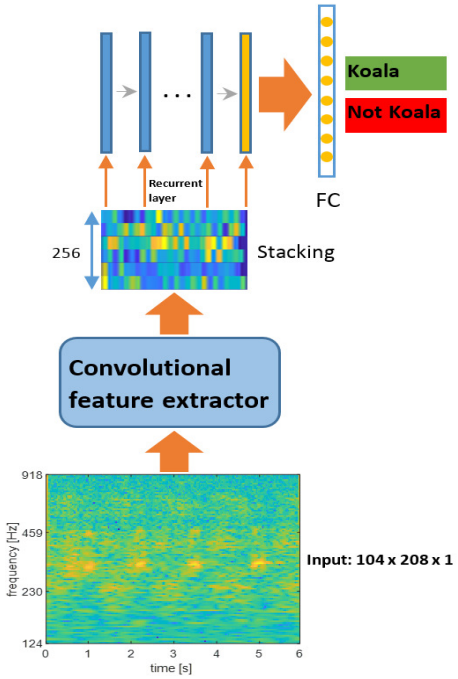


Figure 1: *Diagram of CNN+RNN architecture for koala detection. The output of recurrent layer is followed by a fully connected (FC) layer and a final classification softmax layer.*

First, the audio signal is downsampled with an anti-aliasing filter to one twelfth of the original sampling rate (1,837 Hz). Hence, f_{max}, the highest frequency analyzed, set to half of the new sampling rate (918 Hz). The transform is computed with f_{min}, the lowest frequency analyzed, set to 124 Hz, and central frequencies given by $f_k = f_{min} \cdot (2^{\frac{1}{b}})^k$, using the technique described in [25]. In our approach, we use a total number of 104 bins, with the setting of $b = 36$ bins per octave. Following the CQT, the spectrogram is converted into a log scale,

$$Spectrogram(t,\omega)|_{dB} = 20\ \log_{10}(|X^{CQ}(t,\omega)|) \quad (1)$$

with the six second duration audio clip resulting in 208 CQT frames. The noise-reduction procedure can be optionally applied, for example, by subtracting the median value computed for each spectral band in a spectrogram from every frame. The CNNs input data should have a unified form, hence the input feature shape for spectrogram is 104×208.

3.3. Models

The CNN architecture consists of 3 convolutional layers. We use a receptive field of 3×3 followed by a max pooling operation for every convolutional layer. Rectified linear unit (ReLU) is used as an activation function. Dropout is employed in convolutional layers with rate 0.5 to address overfitting. Xavier initialization is used for convolutional layers [26]. The activations from the filters of the last convolutional layers are stacked over the frequency axis and fed to the LSTM (Long short-term memory) layer, a special RNN structure to avoid exploding gradients problem. The feature maps at the output of the CNN are formulated as a 3D tensor, where 26 is the number of time steps mapped from the 208 time steps in the original spectrogram. Thus, 26 recurrent layers should be constructed in the

Type	Filter/Stride	Output	#Params
Conv1	3 x 3 / 1 x 1	104 x 208 x 32	320
MaxPool1	4 x 2 / 4 x 2	26 x 104 x 32	-
Conv2	3 x 3 / 1 x 1	26 x 104 x 64	18K
MaxPool2	4 x 2 / 4 x 2	7 x 52 x 64	-
Conv3	3 x 3 / 1 x 1	7 x 52 x 128	73K
MaxPool3	4 x 2 / 4 x 2	2 x 26 x 128	-
LSTM	-	64	82K
FC1	-	64	4K
FC2	-	2	130
Total			177K

Table 1: *Proposed CNN+RNN architecture. The data shape indicates frequency × time × number of filters.*

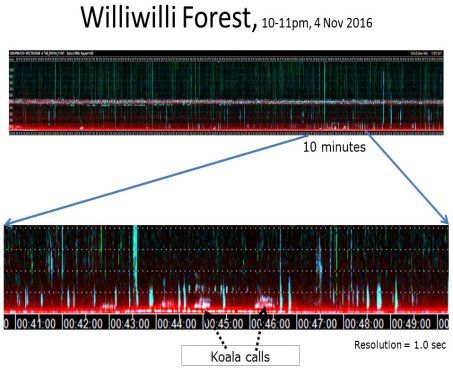


Figure 2: *Detecting koalas from false-color spectrogram.*

RNNs. We apply recurrent networks with 64 LSTM cells. The RNN (many-to-one configuration) output is followed by fully connected layers. The combined CNN and RNN are optimized jointly by employing backpropagation algorithm. Figure 1 depicts CNN+RNN architecture used in this study.

A softmax layer with two nodes is used (one for the koalas and one for non-koala). The network is trained using Adam optimizer [27] with momentum of 0.9, learning rate of 10^{-3}, and a batch size of 32. The binary cross-entropy is used as a loss function. Tensorflow [28, 29] is used to implement the models. In this paper, we compare the CNN+RNN implementation with a standard CNN baseline. Table 1 and 2 show CNN+RNN and CNN architectures in details, respectively.

4. Experiments

4.1. Evaluation metric

The koala call detection system is evaluated from the receiver operating characteristic (ROC) using area under the curve (AUC) measurement. One important advantage of CNNs is the ability to learn local filters from input patches. When it comes to audio spectrogram data, filter dimensions correspond to time and frequency. Thus, by varying filter shape, in other words, by making it wider or higher, the network can be adjusted to learn the time dependencies and the frequency context separately. While this is not in the scope of this work, the hyperparameters for the final network configuration were obtained empirically, as shown in Table 1 and 2.

4.2. Baseline CNN

The benefits of using a recurrent layer after the convolutional layers and training a CNN for use as a baseline was investigated. Instead of a recurrent layer, after the last convolutional layer, the feature maps are flattened and then connected to the fully connected layers in a standard CNN.

Type	Filter/Stride	Output	#Params
Conv1	3 x 3 / 1 x 1	104 x 208 x 32	320
MaxPool1	4 x 2 / 4 x 2	26 x 104 x 32	-
Conv2	3 x 3 / 1 x 1	26 x 104 x 64	18K
MaxPool2	4 x 2 / 4 x 2	7 x 52 x 64	-
Conv3	3 x 3 / 1 x 1	7 x 52 x 128	73K
MaxPool3	4 x 2 / 4 x 2	2 x 26 x 128	-
FC1	-	1024	6.8M
FC2	-	2	2K
Total			6.9M

Table 2: *CNN architecture. The data shape indicates frequency × time × number of filters.*

Models	AUC	AP
CNN+RNN	0.9909	0.988
CNN	0.9908	0.988

Table 3: *AUC scores and average precision (AP) scores on testing dataset.*

5. Results

All models (CNN+RNN and baseline CNN) are evaluated using five-fold cross validation, with a single fold held out as a test set for each round of validation, while the remaining folds are used as training data. The reported results in Table 3 are the average values of the test score across three separate cross-validation runs. The network weights are initialized using a different seed for each run. Overall, the results from the CNN+RNN and CNN models were comparable in terms of performance. While over 99% AUC was obtained, the trained models are useless if they perform poorly in recordings come from different locations.

To evaluate the performance of the classifiers when detecting koalas in different acoustic soundscape, one hour nighttime recordings from a specific site were obtained. The audio signal was first segmented into overlapping 6-second clips (50% overlap). The posterior probabilities of the target class (koalas) from the network output are used as the prediction/confidence scores of the classifiers (after smoothing using a median filter). Typically, these posteriors are used to make the decision (koalas or non-koalas) by comparing them with a fixed threshold. Figure 3 and 4 show the scores over the one-hour recording produced by the CNN+RNN and standard CNN models, respectively. The ground-truth predictions, which were ten koalas in total (plotted in red line), were obtained by a trained expert listener. Overall, both the CNN+RNN and CNN models can predict a koalas call with a high level of confidence (e.g., > 0.5, in practice the threshold should be determined by ROC curve analysis). However, the standard CNN model struggled at the start of the audio clip, due to the noise created by an airplane flying overhead, and both the CNN+RNN and CNN models missed the third koala counted from the right (at 44:20), as its faint calls were masked by the noise made by a low flying airplane.

Visual inspection using a false-color spectrogram [30] is used in Figure 2 to show koala-based activity at around minutes 45:00 and 46:30. These koalas were detected by the classifiers with a high level of confidence, as shown in Figure 3. Note that most non-koala segments were detected with very low confidence values (almost zero) by CNN+RNN. Similar findings were also observed for the CNN+RNN model when detecting

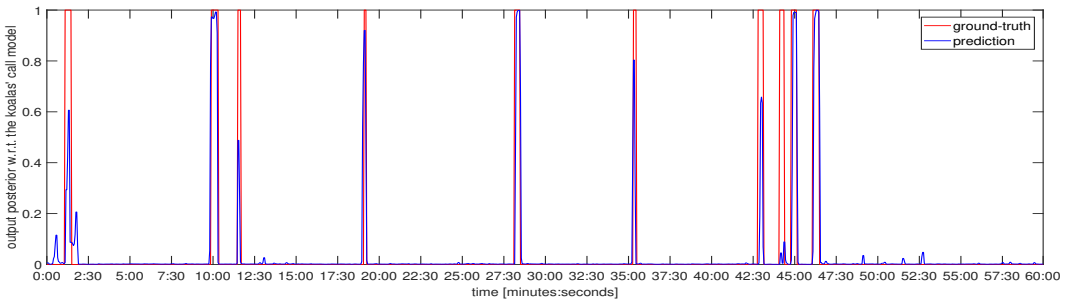


Figure 3: *Output posterior w.r.t. the koalas' call model over the one-hour recording produced by CNN+RNN.*

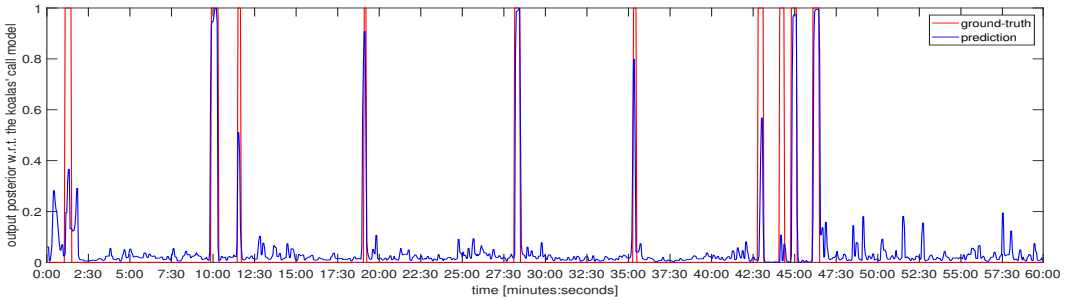
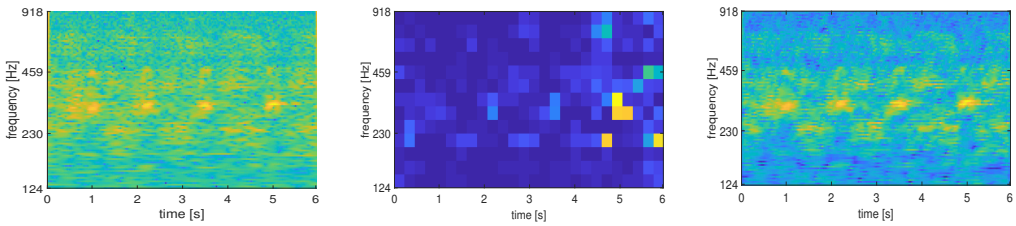


Figure 4: *Output posterior w.r.t. the koalas' call model over the one-hour recording produced by CNN.*



(a) *original spectrogram.* (b) *heat map.* (c) *reconstruction from layer one.*

Figure 5: *Example of the positive class.*

different koala population in South East Queensland (values are not reported in this paper due to restricted-use data). This was not the case for CNN, where the background noise clips produced relatively higher scores compared to CNN+RNN. Since the input to a CNN model is an audio spectrogram, choosing one filter shape over another will potentially impact both learning time and frequency features at the same time. As small squared (3×3) filters were used, the CNN model learnt localized patterns that were constrained by the size of the filters that were represented in the sub-band for a short-time. Thus, it was difficult for the CNN model to capture long temporal dependencies due to the use of these small squared filters [31]. This suggests that CNN+RNN is able to generalize over koala sounds in different environments even though the difference in performance is minimal for the testing data.

To identify the features that the network was learning from, a map of the test samples were computed to highlight the parts of the spectrogram that were important for the prediction of a koalas call. To compute the map, a small part of the image, which formed a 8×8 block, was occluded by setting its intensity to 0. The difference between the probabilities for the whole image and the occluded one indicates the contribution of the occluded part for classification [32, 33]. This procedure was repeated by applying a sliding block with no overlap to occlude a small part of the image. Figure 5.b. shows this map, which can be interpreted as a heat map, with the hot (yellow) pixels contributing more to feature classification than the cold (blue) pixels. We also used the deconvolutional method [32, 34] to project the activations from the feature space to the input space to reveal relative importance of the input pixels for the observed activations (Figure 5.c.). Results show that there is a strong activity in the frequency bands (i.e., 230Hz-459Hz) occupied by the koala's call.

The deep learning architecture developed for detecting koala's call can also be used for other application such as bird activity detection. Using the bird audio detection task [4], the AUC scores for CNN+RNN and CNN methods obtain 87.46% and 83.57% on the unseen evaluation data (scoring server online: http://lsis-argo.lsis.org/scores), respectively. For this task, the input feature is the spectrogram derived from the short-time Fourier transform instead of CQT.

6. Conclusion

This paper investigates deep learning techniques for detecting a koalas call based on combined CNN and RNN architectures by following best practices from literature. It is shown that a CNN+RNN framework is the preferred solution for detecting koala calls in the wild. The proposed method can also be used to detect other animal calls such as bird sound with a superior performance compared to the standard CNN. In future work, we will investigate how to incorporate domain knowledge for designing the network architecture, and to inspect what filters are learning.

7. References

[1] D. of Environment and Q. G. Resource Management, "Decline of the koala coast koala population: Population status in 2008," Retrieved from https://www.cabinet.qld.gov.au/documents/2009/may/koala/Attachments/, 2009.

[2] V. Gonzalez-Astudillo et al., "Decline causes of koalas in south east queensland, australia: a 17-year retrospective study of mortality and morbidity," *Scientific Reports 7*, no. 42587, 2017.

[3] J. Sueur and A. Farina, "Ecoacoustics: the ecological investigation and interpretation of environmental sound," *Biosemiotics*, vol. 8, no. 3, pp. 493–502, 2015.

[4] D. Stowell et al., "Bird detection in audio: a survey and a challenge," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, 2016, pp. 1–6.

[5] J. Xie et al., "An intelligent system for estimating frog community calling activity and species richness," *Ecological Indicators*, vol. 82, pp. 13–22, 2017.

[6] J. Strout et al., "Anuran call classification with deep learning," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2017, pp. 2662–2665.

[7] C. Walters et al., "A continental-scale tool for acoustic identification of european bats," *Journal of Applied Ecology*, vol. 49, no. 50, pp. 1064–1074, 2012.

[8] X. C. Halkias et al., "Classification of mysticete sounds using machine learning techniques," *Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3496–3505, 2013.

[9] M. Towsey et al., "The use of acoustic indices to determine avian species richness in audio-recordings of the environments," *Ecological Informatics*, vol. 21, pp. 110–119, 2014.

[10] J. Xie et al., "Application of image processing techniques for frog call classification," in *Proceedings of IEEE International Conference on Image Processing*, 2015, pp. 4190–4194.

[11] E. Sprengel et al., "Audio based bird species identification using deep learning techniques," *CLEF (Working Notes)*, pp. 547–559, 2016.

[12] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1814–1818.

[13] T. Pellegrini, "Densely connected cnns for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1784–1788.

[14] E. Cakir et al., "Convolutional recurrent neural networks for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2107, pp. 1794–1798.

[15] B. D. Charlton et al., "Perception of male caller identity in koalas (phascolarctos cinereus): acoustic analysis and playback experiments," *PLoS One*, vol. 6, no. 5, pp. 1–8, 2011.

[16] B. D. Charlton, "The acoustic structure and information content of female koala vocal signals," *PloS One*, vol. 10, no. 10, pp. 1–19, 2015.

[17] T. N. Sainath et al., "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2015, pp. 4580–4584.

[18] C. Zhang et al., "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, 2017.

[19] K. Choi et al., "Convolutional recurrent neural networks for music classification," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2017, pp. 2392–2396.

[20] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2014, pp. 6964–6968.

[21] E. Cakir et al., "Convolutional recurrent neural networks for polyphonic sound event detection," *arXiv:1702.06286*, 2017.

[22] W. Ellis et al., "Koala bellows and their association with the spatial dynamics of free-ranging koalas," *Behavioral Ecology*, vol. 22, no. 2, pp. 1–6, 2011.

[23] J. Youngberg and S. Boll, "Constant-q signal analysis and synthesis," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 3, 1978, pp. 375–378.

[24] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, pp. 425–434, 1991.

[25] C. Schörkhuber et al., "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *AES 53rd International Conference on Semantic Audio*, 2014.

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[28] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.

[29] S. Guntupalli, "Bird audio detection challenge," https:https://github.com/swaroopgj/chirps, 2017.

[30] M. Towsey et al., "Visualization of long-duration acoustic recordings of the environment," *Procedia Computer Science*, vol. 29, pp. 703–712, 2014.

[31] J. Pons et al., "Experimenting with musically motivated convolutional neural networks," in *IEEE International Workshop on Content-Based Multimedia Indexing*, 2016, pp. 1–6.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of European conference on computer vision*, 2014, pp. 818–833.

[33] A. Polzounov et al., "Right whale recognition using convolutional neural networks," *arXiv:1604.05605*, 2016.

[34] B. Vikani and F. Shah, "CNN visualization," https://github.com/InFoCusp/tf_cnnvis/, 2017.