Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks

Elly C. Knight[a,d]

Sergio Poo Hernandez[b]

Erin M. Bayne[a]

Vadim Bulitko[b]

Benjamin V. Tucker[c]

[a] *Department of Biological Sciences, University of Alberta, Edmonton, Canada*

[b]*Department of Computing Science, University of Alberta, Edmonton, Canada*

[c]*Department of Linguistics, University of Alberta, Edmonton, Canada*

[d]Corresponding author. Address: Department of Biological Sciences, CW 405 Biological Sciences Building, University of Alberta, Edmonton, AB, T6G 2E9, Canada. Email: ecknight@ualberta.ca

Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks

A variety of automated classification approaches have been developed to extract species detection information from large bioacoustic datasets. Convolutional neural networks (CNNs) are an image classification technique that can be operated on the spectrogram of an audio recording. Using CNNs for bioacoustic classification negates the need for sophisticated feature extraction techniques; however, CNNs may be sensitive to the parameters used to create spectrograms. We used AlexNet to classify spectrograms of audio clips from 19 species of birdsong. We trained and tested AlexNet with the spectrograms and observed that mean classification accuracy ranged from 88.9% to 96.9% depending on the parameters used to create the spectrogram. Classification accuracy was highest when we used a composite of four spectrograms with different combinations of scales for frequency and amplitude. Classification accuracy also varied depending on the FFT window size of the spectrogram. Overall, our results suggest that optimal spectrogram parameters for CNN classification may differ from those used for other human visualization or other classification approaches. We suggest that if spectrogram parameters are appropriately selected, classification accuracy similar to current state-of-the-art methods can be achieved using off-the-shelf software and without the need to extract domain-specific features.

Keywords: autonomous recording unit; birdsong; classification; signal processing; machine learning; spectrogram

Introduction

The need for rapid and accurate classification of acoustic sounds is increasing as ecologists use new recording technology to collect acoustic data over extended periods of time. Bioacoustic sampling with autonomous recording units (ARUs) is increasingly popular because it facilitates collection of a permanent time-series record of acoustic diversity and can be easily scaled over large areas. Recordings collected by ARUs can be used for a variety of ecological purposes, including biodiversity assessment (Sueur and Farina 2015), monitoring ecosystem disturbance (Deichmann et al. 2017), monitoring population trends (Furnas and Callas 2015; Jeliazkov et al. 2016), behavioural studies (Ehnes and Foote 2014), modelling habitat associations (Campos-Cerqueira and Aide 2016), studying phenology (Willacy et al. 2015), detecting rare or inconspicuous species (Sidie-Slettedahl et al. 2015), and monitoring range shifts (Potamitis et al. 2014). However, audio recordings must be processed to extract species data, which can be time-consuming if large volumes of acoustic data are collected (Shonfield and Bayne 2017).

A multitude of automated species classification approaches have been developed to improve the efficiency of processing audio recordings. Researchers have built classifiers for bird (reviewed by Priyardarshani et al. 2018), bat (Armitage and Ober 2010; Walters et al. 2012), frog (reviewed by Xie et al. 2016), fish (Noda et al. 2016), cricket (Brandes et al. 2006; Jaiswara et al. 2013), bee (Gradišek et al. 2016), and monkey species (Heinicke et al. 2015; Turesson et al. 2016) from audio recordings. Extracting species detection information from full length audio recordings has four general steps: signal detection, signal pre-processing, feature extraction, and classification (Xie et al. 2016; Priyadarshani et al. 2018). We focus here on the latter three steps that occur after acoustic signals have been separated from the background noise of a recording. Signal pre-processing can involve transformation to a visual representation (e.g., spectrogram), filtering of unwanted background noise, and segmentation of the acoustic syllables for further processing. The next step, feature extraction, computes feature values for each of the segmented syllables. Commonly used features in species classification are Mel Frequency Cepstral Coefficients (Cheng et al. 2010), time and frequency domain features such as signal bandwidth or spectral centroid (Huang et al. 2009), time-frequency features such as minimum and maximum frequency (Acevedo et al. 2009), or specialized approaches such as ridge-based features (Dong et al. 2015; Xie et al. 2015). The final step, classification, compares those feature values to the feature values of labelled training data. Machine learning techniques are most commonly used for comparison, including Hidden Markov models (Chu and Blumstein 2011), k-nearest neighbour (Huang et al. 2009; Bang and Rege 2014), support vector machines (Acevedo et al. 2009; Armitage and Ober 2010), and random forests (Armitage and Ober, 2010; Noda et al. 2016). Non-machine learning approaches including band-limited energy detection (Charif et al. 2010) and cross-correlation (Katz et al. 2016) have also been employed.

Recent state-of-the-art approaches to automated species classification include image classification techniques such as convolutional neural networks (CNNs) that operate directly on the spectrogram (Stowell et al. 2016; Stowell et al. 2018). CNNs are artificial neural networks that combine convolutional, max pooling, and fully connected layers within a deep (i.e., multilayer) artificial neural network. A convolutional layer consists of a collection of kernels, which allow the network to detect arbitrary patterns in the spectrographic image that correspond to arbitrary patterns of sound in the audio recording. When such a pattern is detected, its location within the image is recorded by activating corresponding elements in the output of the convolutional layer. Next, the max pooling layers aggregate the activations in a convolutional layer into a smaller image, which helps make pattern detection less location sensitive. Eventually a visual pattern detected by a kernel anywhere in the spectrographic image can carry its activation all the way to the appropriate output of the CNN and predict the classification of the input species.

CNNs were first applied to audio recordings for automatic speech recognition (Deng et al. 2013) and have since been used to classify bird and whale species from acoustic recordings with high accuracy (Cakir et al. 2017; Salamon and Bello 2017; Stowell et al. 2018). One of the reasons CNNs have been proposed for bioacoustic classification is that the convolutional layers can render the network location invariant (Cakir et al. 2017; Salamon and Bello 2017). In other words, a CNN should be able to accurately classify a pattern regardless of where it is located within the spectrographic image (Bunne et al. 2018), which means the algorithm can be robust to variation in the timing and frequency of the target acoustic signal within a recording.

An additional advantage of CNNs is that they remove the need to develop sophisticated feature extraction techniques because they operate on the spectrogram; however, this dependency on the spectrographic image suggests the accuracy of bioacoustic classification with CNNs could be particularly sensitive to the pre-processing parameters used to compute the spectrogram. In this paper, we assess the effect of spectrogram parameters on the accuracy of bioacoustic classification using a CNN. To make our methods accessible to ecologists, we did not build our own custom CNN. Instead, we used a pre-trained out-of-the-box network available in Matlab, AlexNet (Krizhevsky et al. 2012), to classify the primary vocalizations of 19 species of boreal bird. We tested the effect of five spectrogram parameters on the classification accuracy of AlexNet: frequency scale, amplitude scale, number of spectrograms, FFT window length (ms), and frequency resolution. We chose our spectrogram parameters and their settings independent of prior assumptions because the classification mechanism of CNNs remains poorly understood and we did not want to assume it would be optimized by existing conventions for human visualization or other classification approaches. Finally, we compared the species-level classification accuracy of AlexNet to a simple, frequency-based classifier.

Methods

Acoustic Data Collection

We were interested in understanding whether AlexNet could differentiate between species of differing vocalization frequency because the spatial invariance of CNNs could potentially apply to the frequency domain as well as the temporal domain (Bunne et al. 2018). We thus selected 19 species of bird for our acoustic dataset that had vocalizations of varying frequency (Table 1). Audio clips of the song or primary vocalization for each species were selected from the Bioacoustic Unit database (http://bioacoustic.abmi.ca). The Bioacoustic Unit database is a collection of 16-bit audio recordings collected in the boreal forest of Canada with Song Meter SM2+ recorders (Wildlife Acoustics, Concord, MA, USA), which are dual-channel recorders with omnidirectional microphones spaced 10” apart. Recordings were made with a sampling rate of 44.1 kHz, no high-pass filter, and a bit depth of 16. Vocalizations of the focal study species were identified from recordings during aural interpretation by expert human listeners. Each vocalization was then manually clipped from the full-length recording with approximately 0.1 seconds of additional recording at the beginning and end of the vocalization using the program Audacity (Audacity Team, 2018). For species with song repertoires, we used only the dominant song type. In total, we used 3048 WAV format audio clips (Table 1) of varying amplitudes, signal-to-noise ratios, and partial masking by other species’ vocalizations.

Frequency Characterization

We first characterized the dominant frequency of each of our 19 bird species. We used the warbleR package (Araya-Salas and Smith-Vidaurre 2016) in R (R Core Team, 2017) to extract 100 evenly spaced measurements of dominant frequency from the left channel of each test clip. To remove the influence of background noise produced by the audio recorders, we set a bandpass filter for each species following the maximum and minimum frequencies of the vocalization of each species as described in the Birds of North America online (Rodewald 2015; Table 1) and only extracted the dominant frequency of samples with amplitude above 30% mean frequency. We calculated the dominant frequency of each clip as the mean of the 100 measures of dominant frequency. We characterized each species by calculating the mean and standard deviation in dominant frequency across all clips for that species.

Spectrogram Creation

We converted each audio clip to a spectrogram in preparation for training and testing our classifier. Audio files were transformed into spectrograms by applying the Fast Fourier Transform (FFT), as introduced by Koenig et al. (1946). The FFT is a class of algorithms designed to maximize the speed of calculating the discrete Fourier transform, with the most common FFT algorithm requiring the number of discrete frequency points be a power of 2 (summarized in Fulop 2011). The discrete Fourier Transform, when applied to an audio sample of a predetermined length, calculates a spectrum of the sounds analyzed, which are expressed as amplitude (generally decibels) as a function of frequency (Hertz or cycles/second). A spectrogram is then created by plotting a series of spectra from the short-time (i.e., brief segments of the signal) discrete Fourier Transforms (discrete short-time Fourier transform; STFT). The spectrographic image resolution depends on window size and is a tradeoff between the time and frequency domains. For example, longer time windows provide increased spectral resolution (narrowband spectrogram) while shorter time windows provide increased temporal resolution (wideband spectrogram).

We selected five spectrogram parameters to manipulate with no prior assumptions about discipline-specific spectrogram creation methods. 1. Frequency scale: we built spectrograms with a linear and a log frequency scale because we predicted that a log transformation of the frequency scale would improve classification accuracy by allowing for better discrimination of species that vocalize at low frequencies (Figure 1). 2. Amplitude scale: we built spectrograms with a dB and a log dB scale (i.e., log of the log) because we predicted that a log dB scale would improve the classification accuracy of AlexNet by allowing for better discrimination of the lower amplitude details in each clip (Figure 1). 3. Number of spectrograms: we built a composite image of multiple spectrograms with different combinations of the scales for amplitude and frequency because we predicted it would improve classification accuracy by allowing the network to pick the most useful information in the combination for each species (Figure 1). 4. FFT window length: we built spectrograms of varying window length (0.5, 1, 5, 10, 50, 100 ms; Figure 2) because window length affects both the temporal and frequency resolution of the spectrogram and we were interested in how this tradeoff affects classification accuracy. We predicted that classification accuracy would be highest for intermediate values of time window length, similar to those used to visualize birdsong (e.g., 10 ms). 5. Number of frequency segments: we built spectrograms of varying frequency resolution because CNNs are limited by pixel resolution input and we were interested in the effects of limiting this resolution. We independently manipulated the frequency resolution by dividing each discrete Fourier Transform into varying numbers of segments (10, 25, 50, 75, 113; Figure 2) and summing the amplitude within each segment. We predicted that more segments (i.e., higher frequency resolution) would improve the classification accuracy of AlexNet.

We used a fully-crossed design to create spectrograms for every combination of our five parameters (150 sets of spectrograms in total). Each set contained a spectrogram for each of the 3048 audio clips in our dataset. We created spectrograms using a modified discrete STFT so that we could manipulate the selected parameters. First, we divided the WAV file up by the selected FFT window length and calculated the discrete Fourier Transform. For the log dB amplitude scale, we took the log of the amplitude values from the discrete Fourier Transform. We then allocated the amplitude values into the selected number of frequency segments and summed the values in each segment. For the log frequency scale, we spaced the frequencies contained in each frequency segment logarithmically. We converted each spectrogram to an RGB image with the jet colormap from MATLAB and 100 distinct colors. The amplitude scale was thus specific to each spectrogram, with blue indicating the minimum amplitude within that clip and red indicating the maximum amplitude within that clip. Finally, we bicubically resized each spectrogram to 227 pixels to 227 pixels to fit AlexNet's image input requirements. All spectrograms were resized to this image size regardless of audio clip length. We created all spectrograms from the raw audio; we did not bandpass filter or implement pre-processing noise reduction to the signal. We averaged the two channels of the clip because the 10" distance between the microphones did not cause substantial differences in time of arrival. We used rectangular windowing and limited the frequency range between 0.1 to 10 kHz for all FFT transformations. All spectrograms were created in MATLAB using audioread, fft, colormapping, and imresize functions (The Mathworks Inc. 2017).

Network Setup

We used an out-of-the-box network called AlexNet (Krizhevsky et al. 2012) from the MATLAB neural network toolbox. AlexNet is a CNN that is comprised of 5 convolutional layers, 3 max pooling layers, 3 fully connected layers, and 1 softmax layer (Appendix 1). AlexNet is pretrained on Imagenet (Deng et al. 2009), an open-source image database, to classify a color image into one of 1000 categories. We then changed the last fully connected layer of the network from 1000 class outputs to 19 class outputs; one for each of the selected bird species. We used a grid search to determine the best hyperparameters for the number of epochs (10, 50, 100, 200) and batch size (3, 5, 10, 50, 100).We fixed the learning rate of 0.0001 because we have previously found it yields high test accuracy with reasonable training time for a variety of image classification tasks (*unpublished data*). We selected 50 epochs and a batch size of 5 as our hyperparameters for all subsequent experiments because the combination yielded a high average test accuracy (96.4%) with reasonable training time (40 minutes on Nvidia GTX Titan, Maxwell architecture).

Network Training and Testing

We trained and tested AlexNet separately on each of the 150 spectrogram sets using the stochastic gradient descent algorithm implemented by MATLAB's trainNetwork function. We ran 4 cross-validation trials for each spectrogram set with 75% of the dataset selected randomly for training, and 25% withheld for testing. We used the class with the highest predicted probability in the softmax layer as the predicted species. To estimate class accuracy of each species, we ran ten trials using our top-performing spectrogram settings (composite spectrogram, 50 ms FFT time window, 113 frequency segments; see Results for details) and calculated the confusion matrix for each trial. We ran these as additional trials to increase the sample size for our estimate of best class accuracy per species and to analyze the relationship between dominant frequency and class accuracy.

Statistical Analysis

We calculated the classification accuracy for each trial (4 trials x 150 spectrogram sets = 600 samples) as the number of test spectrograms classified correctly, divided by the total number of test spectrograms. We used a general linear model of the single spectrogram trials (4 trials x 120 spectrogram sets) to test for an effect of spectrogram frequency scale, amplitude scale, FFT window length, and number of frequency segments on the classification accuracy of AlexNet. We included FFT window length as a second-order polynomial because visualization of the raw data showed a nonlinear relationship. We then tested our prediction that the log frequency scale improved the network's ability to classify low frequency species by regressing the mean dominant frequency of each species and the interaction with frequency scale type against the classification accuracy of each species from the ten confusion matrix trials.

Next, we used all spectrogram sets (5 trials x 150 spectrograms) to test whether the composite spectrogram improved classification accuracy. We used a general linear model with the composite spectrogram set as the reference category to compare the composite spectrogram to the four frequency and amplitude scale combinations for single spectrograms (linear - dB, linear - log dB, log - dB, log - log dB). We included FFT window length as a second-order polynomial and number of frequency segments to account for any covariation.

Finally, we investigated the importance of dominant frequency for species classification because the spectrogram resolution results suggested frequency was particularly important for classification by AlexNet. We used linear discriminant analysis in the MASS package in R (R Core Team, 2017) to create a classification matrix based on the dominant frequency of each audio clip, as measured during frequency characterization. We chose linear discriminant analysis because it can use univariate or multivariate input to classify data into a predetermined number of categories. We ran 10 cross-validation trials with 75% of the dataset selected randomly for training, and 25% withheld for testing. We then compared the confusion matrix of the linear discriminant analysis classifier to the confusion matrix of the AlexNet classifier trained and test composite spectrograms. All statistical analyses were completed in R version 3.4 (R Core Team 2017) using the MASS package and base functions.

Results

Spectrogram Resolution

FFT window length and number of frequency segments both had a strong effect on the classification accuracy of AlexNet ($t_{474} = 14.09$, $P < 0.001$, $t_{474} = 6.47$, $P < 0.001$). Classification accuracy increased quickly with decreasing FFT window length and peaked near 50 ms windows (Figure 3). Classification accuracy increased with increasing number of frequency segments and was highest at the maximum number of frequency segments (113; Figure 3).

Spectrogram Scale

Frequency scale and amplitude scale both had an effect on the classification accuracy of AlexNet, although the effect of frequency scale ($t_{474} = -3.20$, $P = 0.001$) was greater than the effect of amplitude scale ($t_{474} = 2.24$, $P = 0.03$; Figure 4). Classification accuracy was 0.6% (± 0.2% SE) higher for the linear frequency scale than the log scale. Classification accuracy was 0.4% (± 0.2% SE) higher for the log dB amplitude scale than the dB scale. There was a weak interaction between frequency scale and the mean dominant frequency of a species ($t_{756} = -1.92$, $P = 0.06$); classification accuracy for the log scale was 0.3% (± 0.3% SE) lower than the linear scale for every 1 kHz increase in dominant frequency (Figure 5). Contrary to our predictions, there was no difference in classification accuracy between the linear and log frequency scale for species that vocalize at lower frequency.

Composite Image

The composite spectrogram had higher classification accuracy than the four frequency and amplitude scale combinations for single spectrograms (linear - dB, linear - log dB, log - dB, log - log dB), but it was not significantly higher than the linear frequency and log dB amplitude single spectrogram (t_{592} = -0.74, *P* = 0.45; all other *P* < 0.001; Figure 4). The mean accuracy of the composite image was 94.9% (± 0.2% SE).

Classifier Comparison

The classification accuracy of the linear discriminant analysis classifier based on sound frequency attributes was 51.7% (± 0.8% SE; Figure 6). Species with similar frequencies were misclassified most often, particularly the passerine species with similar intermediate frequency ranges (REVI, SWTH, WTSP, ALFL, BCCH, HETH). In contrast, the AlexNet classifier most frequently misclassified species that have similar song structure (SWTH and HETH, ALFL and LEFL, BBWA and AMRE).

Summary of Classifier Performance

Classification accuracy across all trials of all spectrograms ranged from 85.5% to 98.6%, with a mean of 94.3%. The spectrogram settings that our analyses indicated would maximize classification accuracy (composite spectrogram, 113 frequency segments, 50 ms FFT window length) yielded a mean classification accuracy of 96.9% (± 0.2% SE). The spectrogram settings that our analyses indicated would minimize classification accuracy (log frequency scale, dB amplitude scale, 10 filters, 0.5 ms window length) yielded a mean classification accuracy of 88.9% (± 0.2% SE).

Discussion

High classification accuracy is an important goal for automated species classification from audio recordings because it affects how well the resulting data represents the ecological conditions of the recording. We showed that the accuracy of a CNN bioacoustic classifier built with AlexNet is affected by choice of the parameters used to convert each audio clip to a spectrogram. We found that mean classification accuracy ranged from 88.9% to 97.3% depending on the parameters used to create the spectrogram. Classification accuracy was affected by the scales used for frequency and amplitude, by the FFT window length, and by the number of frequency segments of the spectrogram. Current state-of-the-art approaches to species classification from audio clips typically achieve classification accuracy above 95% on similarly sized datasets (Nicholson 2016; Salamon and Bello 2017; Zhao et al. 2017), and our results for 19 species are comparable. We achieved this high classification rate with off-the-shelf software (AlexNet) and relatively short training time (< 40 minutes), suggesting that state-of-the-art bioacoustic classification is becoming increasingly accessible to ecologists and can be achieved if appropriate spectrographic parameters are selected. For ecologists wishing to use AlexNet for birdsong classification of audio clips, we recommend using an online tutorial (e.g., MathWorks 2018) and a computer with a graphics processing unit (GPU). AlexNet can be run without a GPU, but the processing time will be approximately an order of magnitude longer. Ecologists should note, however, that using AlexNet as a recognizer similar to other out-of-the-box software programs (e.g., MonitoR, Katz et al. 2016; Kaleidoscope, Wildlife Acoustics Inc.) will require signal detection prior to application of the AlexNet classifier.

The results of our study and others suggest that frequency is likely an important characteristic for birdsong classification. Our linear discriminant analysis classifier based on mean dominant frequency achieved a classification accuracy of 52.9%. Band-limited energy detectors that use only the frequency range of an acoustic signal for classification of audio clips (Charif et al. 2010) are reported to achieve similar accuracy (43%; reviewed by Priyadarshani et al. 2018). Even though CNNs can be spatially invariant (Bunne et al. 2018), frequency is likely also important for CNN classification of birdsong. We observed an increase in classification accuracy with increasing number of frequency segments, which improved the frequency resolution of the spectrogram. Unlike the linear discriminant classifier, however, the AlexNet classifier did not commonly misclassify species of similar mean dominant frequency. Instead, AlexNet commonly misclassified species of similar song structure like Swainson's Thrush (*Catharus ustulatus*) and Hermit Thrush (*C. guttatus*), which suggests AlexNet's classification is based, at least in part, on other frequency characteristics such as frequency range and harmonics. Future work should use spectrogram manipulation to determine which specific frequency characteristics contribute to CNN classification accuracy.

Classification accuracy was highest for intermediate values of FFT window length however, the maximum accuracy was achieved with a longer window length (50 ms) than is generally used for birdsong visualization (10 ms) which was contrary to our predictions. This longer time window further emphasizes the importance of frequency resolution in birdsong classification with AlexNet. Some authors have predicted that temporal resolution should be particularly important in birdsong classification (Priyardarshani et al. 2018) because auditory experiments indicate the temporal information in birdsong is important for communication by birds (Dooling et al. 2002).

Graciarena et al. (2010) found no improvement in classification accuracy by Gaussian mixture models when they increased time window resolution for Mel frequency cepstral coefficient extraction. We suggest, however, that the impact of FFT window length on classification accuracy is likely to vary across datasets. Graciarena et al. (2010) showed that the importance of temporal resolution for Mel frequency cepstral coefficient extraction was dependent on the bird species.

We found that using a log dB scale for amplitude increased the classification accuracy of AlexNet. Previous authors have shown that CNNs are capable of classifying energy modulated patterns (Salamon and Bello 2015), which are particularly common in birdsong, and the log dB scale for amplitude may further improve this capability by allowing for better discrimination of the lower amplitude details in each clip. We suggest using a log dB scale for amplitude may also improve the ability of AlexNet to classify clips that were recorded at greater distances by emphasizing the lower amplitudes of those clips via the hue of the spectrographic image. Audio signals recorded far from the sound source are particularly difficult to classify due to the attenuation and spherical spreading of sound as it travels (Knight and Bayne 2018). Future work should compare dB and log dB spectrogram scales for classification of clips recorded at known distances.

Using a log scale for frequency resulted in lower overall classification accuracy because it decreased the classification accuracy of species that vocalize at higher frequencies. We predicted the opposite; that the log scale would increase classification accuracy by improving AlexNet's ability to differentiate between species that vocalize at lower frequencies. We noted, however, that summing the amplitude in each frequency segment reduced the relative amplitude of the lower frequencies for the log frequency scale, which may have confounded any potential improvement in classification accuracy of low frequency species that the log scale provided. Many top-performing classifiers use a mel scale, which is a log-based scale designed to approximate the frequency-band sensitivity of human hearing (Stowell et al. 2018). We suggest that a log scale may improve classification accuracy if a traditional discrete STFT is used to construct spectrograms without frequency segments.

Overall, we showed the choice of spectrogram parameters is important for bioacoustic classification using CNNs. Other authors have also shown that spectrogram parameters can affect the classification accuracy of CNNs and other approaches. Xie et al. (2018) showed that the algorithm used to create the spectrogram can affect CNN classification accuracy, with chirplet spectrograms outperforming a short-time Fourier Transform spectrogram. Ulloa et al. (2016) showed that using spectrograms with intermediate resolution and zero window overlap resulted in the highest classification accuracy when using normalized cross-correlation for classification. Colour scale may also affect CNN classification accuracy and the perceptually-uniform viridis colour scale has been shown to improve classification accuracy over the jet colour scale used here (Amiriparian et al. 2017). Overall, our results suggest that optimal spectrogram parameters may differ from those used for human visualization or other classification approaches, and thus should be selected independent of prior assumptions. Optimal spectrogram parameters will also likely also vary depending on the target bird species or community. We therefore encourage practitioners to use our results as a starting point for optimizing spectrogram settings before using a multi-species CNN classifier to process audio recordings.

Acknowledgements

Our sincere thanks to the editor and several anonymous reviewers for their thoughtful comments, which greatly improved the clarity and utility of our manuscript. We thank

Nicole Boucher, Alex MacPhail, and Easwaramurthyvasi Chidambara-vasi for their help in clipping audio files. This work was supported by the Alberta Conservation Association and the Alberta Biodiversity Monitoring Institute. Thanks to NVIDIA for donation of graphical processing units used in data processing. Funding for ECK and SPH was provided by Natural Sciences and Engineering Research Council of Canada.

References

Acevedo MA, Corrada-Bravo CJ, Corrada-Bravo H, Villanueva-Rivera LJ, Aide TM. 2009. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecol Inform. 4:206–214.

Amiriparian S, Gerczuk M, Ottl S, Cummins N, Freitag M, Pugachevskiy S, Baird A, Schuller B. 2017. Snore sound classification using image-based deep spectrum features. Interspeech, August 20-24; Stockholm (Sweden). International Speech Communication Association. p. 3512-3516.

Araya-Salas M, Smith-Vidaurre G. 2016. warbleR: An Rpackage to streamline analysis of animal acoustic signals. Methods Ecol Evol. 8:184-191.

Armitage DW, Ober HK. 2010. A comparison of supervised learning techniques in the classification of bat echolocation calls. Ecol Inform. 5:465–473.

Audacity Team. 2018. Audacity(R): Free Audio Editor and Recorder Version 2.2.2.

Bang AV, Rege PP. 2014. Classification of bird species based on bioacoustics. J Adv Comput Sci Appl. 1:6-10.

Brandes TS, Naskrecki P, Figueroa HK. 2006. Using image processing to detect and classify narrow-band cricket and frog calls. J Acoust Soc Am. 120:2950–2957.

Bunne C, Rahmann L, Wolf T. 2018. Studying invariances of trained convolutional neural networks. arXiv:1803.05963.

Cakir E, Adavanne S, Parascandolo G, Drossos K, Virtanen T. 2017. Convolutional recurrent neural networks for bird audio detection. 25th European Signal Processing Conference, August 29-September 2; Kos Island (Greece). IEEE Signal Processing Society. p. 1744-1748.

Campos-Cerqueira M, Aide TM. 2016. Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. Methods Ecol Evol. 7:1340–1348.

Charif RA, Waack AM, Strickman LM. 2010. Raven Pro 1.4 User's Manual. pp. 1–379.

Cheng J, Sun Y, Ji L. 2010. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. Pattern Recognition. 43:3846–3852.

Chu W, Blumstein DT. 2011. Noise robust bird song detection using syllable pattern-based hidden Markov models. 2011 IEEE International Conference on Acoustics Speech and Signal Processing, May 22-27; Prague (Czech Republic). IEEE Signal Processing Society. p. 345–348.

Deichmann JL, Hernández-Serna A, Amanda Delgado CJ, Campos-Cerqueira M, Aide TM. 2017. Soundscape analysis and acoustic monitoring document impacts of natural gas exploration on biodiversity in a tropical forest. Ecol Indic. 74:39–48.

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. ImageNet: A large-scale hierarchical image database. http://image-net.org/

Deng L, Hinton G, Kingsbury B. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. 2013 IEEE International Conference on Acoustics Speech and Signal Processing, May 26-31; Vancouver (BC). IEEE Signal Processing Society. p. 8599–8603.

Dong X, Towsey M, Truskinger A, Cottman-Fields M, Zhang J, Roe P. 2015. Similarity-based birdcall retrieval from environmental audio. Ecol Info. 29:66–76.

Dooling RJ, Leek MR, Gleich O, Dent ML. 2002. Auditory temporal resolution in birds: Discrimination of harmonic complexes. J. Acoust. Soc. Am. 112:748–759.

Ehnes M, Foote JR. 2014. Comparison of autonomous and manual recording methods for discrimination of individually distinctive Ovenbird songs. Bioacoustics. 24:111–121.

Fulop SA. 2011. Speech spectrum analysis. Berlin: Springer. Chapter 3, The Fourier power spectrum and spectrogram; p. 69–106.

Furnas BJ, Callas RL. 2015. Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. J Wildl Manage. 79:325–337.

Graciarena M, Delplanche M, Shriberg E, Stolcke A, Ferrer L. 2010. Acoustic front-end optimization for bird species recognition. 2010 IEEE International Conference on Acoustics Speech and Signal Processing, March 14-19; Dallas (TX). IEEE Signal Processing Society. p. 293–296.

Gradišek A, Slapničar G, Šorn J, Luštrek M, Gams M, Grad J. 2016. Predicting species identity of bumblebees through analysis of flight buzzing sounds. Bioacoustics. 26:63-76.

Heinicke S, Kalan AK, Wagner OJJ, Mundry R, Lukashevich H, Kühl HS. 2015. Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods Ecol Evol. 6:753–763.

Huang C-J, Yang Y-J, Yang D-X, Chen Y-J. 2009. Frog classification using machine learning techniques. Expert Syst Appl. 36:3737–3743.

Jaiswara R, Nandi D, Balakrishnan R. 2013. Examining the effectiveness of discriminant function analysis and cluster analysis in species identification of male field crickets based on their calling songs. PLoS One. 8e75930–11.

Jeliazkov A, Bas Y, Kerbiriou C, Julien J-F, Penone C, Le Viol I. 2016. Large-scale semi-automated acoustic monitoring allows to detect temporal decline of bush-crickets. Glob Ecol Conserv. 6:208–218.

Katz J, Hafner SD, Donovan T. 2016. Assessment of error rates in acoustic monitoring with the R package monitoR. Bioacoustics. 25:177–196.

Koenig W, Dunn HK, Lacy LY. 1946. The sound spectrograph. J Acoust Soc Am. 18:19–49.

Knight EC, Bayne EM. 2018. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics. DOI: 10.1080/09524622.2018.1503971

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 1097-1105.

MathWorks Inc. 2018. Get started with transfer learning. Accessed February 12, 2019. https://www.mathworks.com/help/deeplearning/examples/get-started-with-transfer-learning.html.

Nicholson D 2016. Comparison of machine learning methods applied to birdsong element classification. Proceedings of the 15th Python in Science Conference, July 11–17; Austin (TX). SciPy. p. 57–61.

Noda J, Travieso C, Sánchez-Rodríguez D. 2016. Automatic taxonomic classification of fish based on their acoustic signals. Applied Sciences. 6:443–12.

Potamitis I, Ntalampiras S, Jahn O, Riede K. 2014. Automatic bird sound detection in long real-field recordings: applications and tools. Appl Acoust. 80:1–9.

Priyadarshani N Marsland S Castro I. 2018. Automated birdsong recognition in complex acoustic environments: a review. J of Avian Biol. 49: jav-01447.

R Core Team 2017. R: a language and environment for statistical computing. Version 3.4.3 Vienna (Austria):R Foundation for Statistical Computing.

Rodewald P, editor. 2015. The Birds of North America Online. Ithaca (NY): Cornell Laboratory of Ornithology.

Salamon J, Bello JP. 2017. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters. 24:279–283.

Salamon J, Bello JP. 2015. Feature learning with deep scattering for urban sound analysis. 23rd European Signal Processing Conference, August 31-September 4; Nice (France). IEEE. p. 724–728.

Shonfield J, Bayne EM. 2017. Autonomous recording units in avian ecological research: current use and future applications. Avian Conserv Ecol. 12(1):14.

Sidie-Slettedahl AM, Jensen KC, Johnson RR, Arnold TW, Austin JE, Stafford JD. 2015. Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. Wildl Soc Bull. 39:626–634.

Stowell D, Wood M, Stylianou Y, Glotin H 2016. Bird detection in audio: A survey and a challenge. 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing, Sept 13–16; Salerno (Piscataway (NJ)): IEEE Signal Processing Society. p 1–6.

Stowell D, Wood M, Pamula H, Stylianou Y, Glotin H. 2018. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. Methods Ecol Evol. DOI:10.111/2041-201X.13103

Sueur J, Farina A. 2015. Ecoacoustics: The ecological investigation and interpretation of environmental sound. Biosemiotics. 8:493–502.

The Mathworks Inc. 2017. MATLAB and Statistics Toolbox. Massachusetts USA.

Turesson HK, Ribeiro S, Pereira DR, Papa JP, de Albuquerque VHC. 2016. Machine learning algorithms for automatic classification of marmoset vocalizations. PLoS One. 11e0163041–14.

Ulloa JS, Gasc A, Gaucher P, Aubin T, Réjou-Méchain M, Sueur J. 2016. Screening large audio datasets to determine the time and space distribution of Screaming Piha birds in a tropical forest. Ecol Inform. 31:91–99.

Walters CL, Freeman R, Collen A, Dietz C, Fenton MB, Jones G, Obrist MK, Puechmaille SJ, Sattler T, Siemers BM, Parsons S, Jones KE. 2012. A continental-scale tool for acoustic identification of European bats. J Appl Ecol. 49:1064–1074.

Willacy RJ, Mahony M, Newell DA. 2015. If a frog calls in the forest: bioacoustic monitoring reveals the breeding phenology of the endangered Richmond range mountain frog (*Philoria richmondensis*). Austral Ecology. 40:625–633.

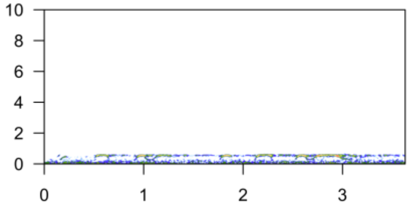
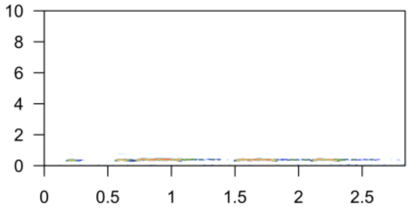
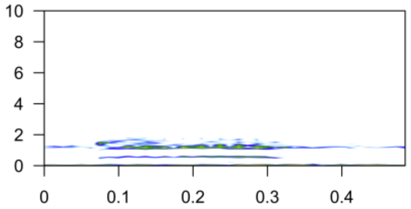
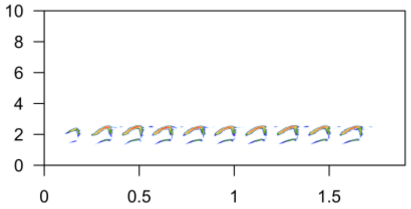
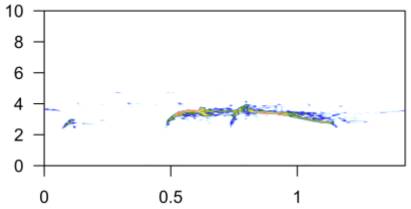
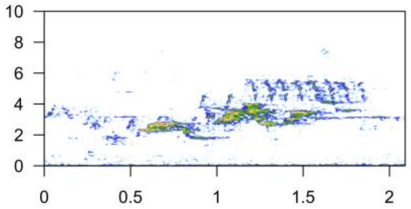
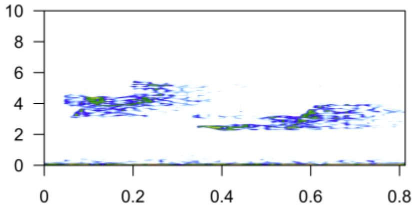
Xie J, Ding C, Li W. 2018. Audio-only bird species automated identification method with limited training data based on multi-channel deep convolutional neural networks. arXiv:1803.01107

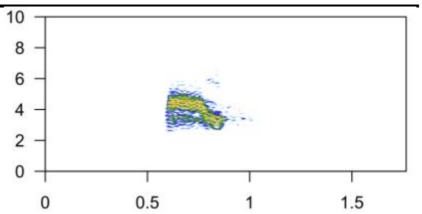
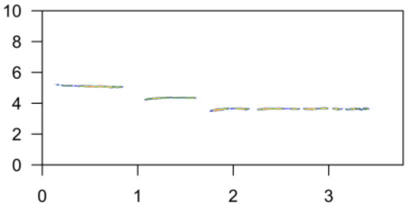
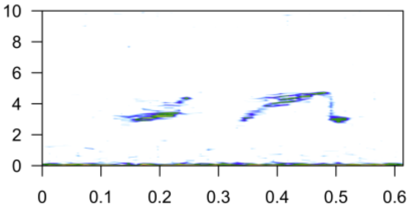
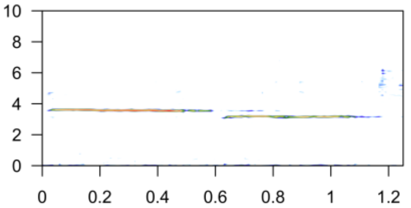
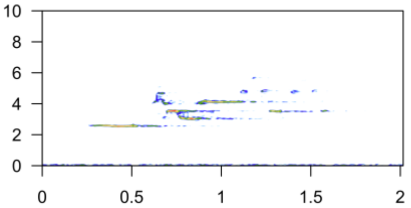
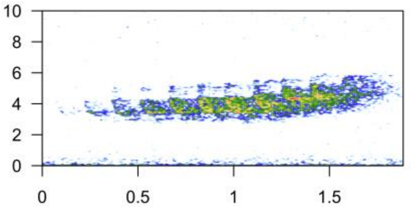
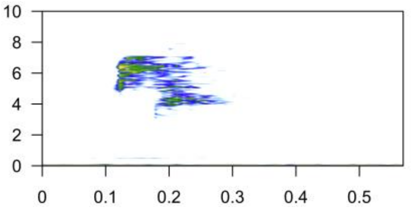
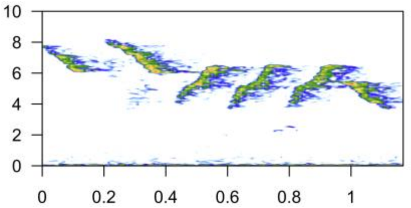
Xie J, Towsey M, Zhang J, Dong X. 2015. Application of image processing techniques for frog call classification. 2015 IEEE International Conference on Image Processing, September 27-30, Québec City (QC). IEEE. p. 4190–4194.

Xie J, Towsey M, Zhang J, Roe P. 2016. Frog call classification: a survey. Artif. Intell. Rev. 49:375-391.

Zhao Z, Zhang S-H, Xu Z-Y, Bellisario K, Dai N-H, Omrani H, Pijanowski BC. 2017. Automated bird acoustic event detection and robust species classification. Ecol Inform. 39:99–108.

Table 1. Bird species included in a bioacoustic classifier built in AlexNet. Typical spectrograms were constructed with a 1024 FFT size and are plotted with time (s) on the x-axis and frequency (Hz) on the y-axis. Species are sorted in order of ascending mean dominant frequency.

Name	Number of clips	Mean dominant frequency (kHz)	Mean clip length (s)	Mean AlexNet classification accuracy (%)	Typical spectrogram
Barred owl (*Strix varia*)	139	0.28	3.47	97.9	
Great-horned owl (*Bubo virginianus*)	134	0.32	2.58	99.7	
Common raven (*Corvus corax*)	371	1.44	0.49	99.7	
White-breasted nuthatch (*Sitta carolinensis*)	97	2.05	1.75	99.6	
Olive-sided flycatcher (*Contopus cooperi*)	159	3.12	1.21	96.2	
Swainson's thrush (*Catharus ustulatus*)	120	3.46	1.62	90.7	
Red-eyed vireo (*Vireo olivaceus*)	111	3.49	0.76	97.8	

Name	Number of clips	Mean dominant frequency (kHz)	Mean clip length (s)	Mean AlexNet classification accuracy (%)	Typical spectrogram
Common nighthawk (*Chordeiles minor*)	100	3.56	0.85	99.6	
White-throated sparrow (*Zonotrichia albicollis*)	114	3.64	3.88	95.9	
Alder flycatcher (*Empidonax alnorum*)	133	3.67	0.54	94.5	
Black-capped chickadee (*Poecile atricapillus*)	111	3.85	1.11	94.4	
Hermit thrush (*Catharus guttatus*)	293	4.40	1.98	97.8	
Yellow-rumped warbler (*Setophaga coronata*)	141	4.46	1.84	94.9	
Least flycatcher (*Empidonax minimus*)	491	4.84	0.49	98.8	
Yellow warbler (*Setophaga petechia*)	116	5.22	1.25	96.6	

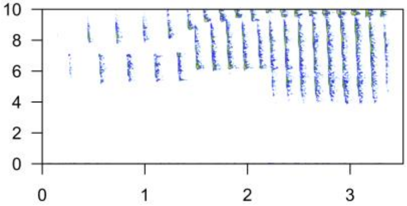
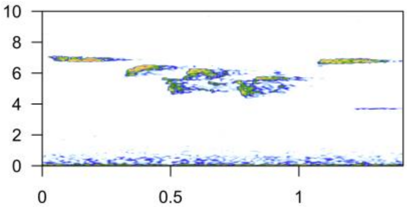
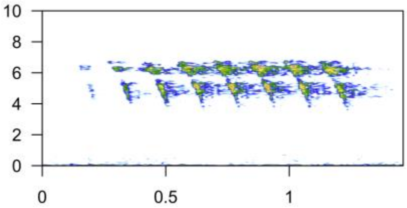
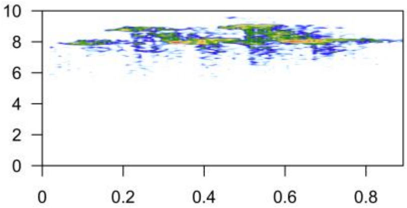
Name	Number of clips	Mean dominant frequency (kHz)	Mean clip length (s)	Mean AlexNet classification accuracy (%)	Typical spectrogram
Tennessee warbler (*Leiothlypis peregrina*)	111	5.25	3.37	100.0	
Brown creeper (*Certhia americana*)	108	5.60	1.36	99.6	
American redstart (*Setophaga ruticilla*)	103	5.82	1.31	96	
Bay-breasted warbler (*Setophaga castanea*)	99	7.20	1.06	92.9	

Figure 1. Spectrogram types used as input for bioacoustic classification with AlexNet. Examples are of a simple sound with increasing frequency through time. From left to right: linear frequency and dB amplitude scales, linear frequency and log dB amplitude scales, log frequency and dB amplitude scales, log frequency and log dB amplitude scales, composite spectrogram: an array of the preceding four spectrograms. Colour scale ranges from minimum (blue) to maximum (red) amplitude for that audio clip.

Figure 2. FFT window length and number of frequency segment values used to create spectrograms for bioacoustic classification with AlexNet. Examples are of an alder flycatcher (*Empidonax alnorum*).

Figure 3. Classification accuracy of birdsong relative to spectrogram FFT window length (ms) and number of frequency segments, as predicted by a bioacoustic classifier built in AlexNet. Spectrograms were constructed with a linear frequency scale and a log dB amplitude scale. Points are raw classification accuracy and have been jittered for visualization. Lines are model predictions with 95% confidence intervals from a linear regression, holding frequency scale as linear, amplitude scale as log dB, and other covariates at their mean.

Figure 4. Classification accuracy of birdsong for different spectrogram inputs for a bioacoustic classifier built with AlexNet. The composite spectrogram was a composite image of the four single spectrograms in a 2x2 array. Error bars represent 95% confidence intervals.

Figure 5. Classification accuracy and mean dominant frequency of 19 species of birdsong, as predicted by a classifier built in AlexNet and trained on spectrograms of the primary vocalizations of each species. Two spectrograms were constructed for each training clip: one with a linear frequency scale and one with a log scale. Best fit lines are model predictions with 95% confidence intervals from a linear regression. Points are raw classification accuracy from confusion matrices and have been jittered for visualization.

Figure 6. Mean classification rate for 19 species of birdsong classified with AlexNet (left) and linear discriminant analysis (right). Each cell represents the mean classification rate for that species combination across ten classification trials. Species are sorted in order of ascending mean dominant frequency. Note the scale is from 0 to

10% classification accuracy on the left and 0 to 100% on the right to highlight differences in misclassification.

Appendix 1. AlexNet is a convolutional neural network originally used for image classification (Krizhevsky et al. 2012). The structure is comprised of 5 convolutional layers, 3 max pooling layers, 3 fully connected layers, and 1 softmax layer, as described below. The network accepts images of size 227 x 227 pixels as input. Each pixel is represented by a triple of color components (RGB).

1. Convolutional layer 1: 96 kernels of size 11x11x3

2. Max pooling layer 1

3. Convolutional layer 2: 256 kernels of size 5x5x48

4. Max pooling layer 2

5. Convolutional layer 3: 384 kernels of size 3x3x256

6. Convolutional layer 4: 384 kernels of size 3x3x192

7. Convolutional layer 5: 256 kernels of size 3x3x192

8. Max pooling layer 3

9. Fully connected layer 1: 4096 neurons x 9216

10. Fully connected layer 2: 4096 neurons x 4096

11. Fully connected layer 3: 11 neurons x 4096

12. Softmax layer