Bioacoustic classification framework using transfer learning

Pedro Bonito Baptista, Cláudia Antunes

Instituto Superior Técnico
{pedrobonitobaptista,claudia.antunes}@tecnico.ulisboa.pt

Abstract. The field of bioacoustics plays an important role on preventing and reducing human impact on environment by enabling the development of tools capable of performing automated analysis of environmental data. Deep learning methods have been successful on automating the process of species identification in environmental recordings, requiring nonetheless a large number of training samples per species. In this paper, we address the problem of automating species detection in noisy environments with limited training data, proposing a framework for training a convolutional neural network (CNN) on Mel spectrograms to predict a set of species present in the Rainforest Connection's acoustic recordings. We leverage transfer learning by using a pre-trained model as a way to reduce training requirements, both the amounts of data and time. Finally we explore several window sizes, data augmentation techniques and predictive thresholds to improve the model's performance.

Keywords: Bioacoustic classification · Deep learning · Convolutional Neural Networks (CNN) · Data augmentation · Transfer learning

1 Introduction

Bioacoustics focuses on the analysis of the sounds produced by or affecting living organisms, especially the ones related to communication. Prior bioacoustics research was heavily dependent on manual labor to segment, detect and label animal activity, present in hours of field recordings. Consequently, recent research overlaps the work developed by Rainforest Connection (NGO) which focuses on developing bioacoustic monitoring systems to ensure the rainforest's conservation, being also a prominent source of environmental audio data.

Deep learning methods have been successful on automatic acoustic identification, through image analysis dedicated architectures, such as convolutional networks. However they require a large number of training samples per species. This limits applicability to rarer species, which are central to conservation efforts. Thus, the Kaggle competition "Rainforest Connection Species Audio Detection" [1] encouraged contenders to develop solutions capable of automate high-accuracy species detection in noisy soundscapes with limited training data.

[1] https://www.kaggle.com/c/rfcx-species-audio-detection

This work explores an end-to-end approach that encompasses the extraction of the Mel spectrograms from the raw audios to the training of a *Convolutional Neural Network*, exploring transfer learning and data augmentation as a way of improving the models' performance. This study is the starting point of the definition of a bioacoustic classification framework. The final approach is the result of several experiments on different window sizes and data augmentations techniques. It includes 5-second-long Mel spectrograms and relies on the SpecAugment method to increase the training set size, achieving an accuracy of 91%, a mean precision of 77% and a mean recall of 78%.

This paper is organized into six sections. Section 2 describes the concepts addressed by this work and section 3 introduces the work related with automatic bioacoustic analysis and classification. Section 4 presents the proposed methodology, section 5 details the obtained results and section 6 discusses them briefly.

2 Basic Concepts and Notation

The research direction that this work will concern is *sound event detection and classification*. The goal of *sound event classification* is to determine which acoustic event appears in an audio sample, not taking into account its corresponding time and its number of occurrences. Nevertheless, as acoustic events can overlap temporally, acoustic event classification sometimes is not a practical problem. On the other hand, *sound event detection* labels temporal regions within an audio recording, with their start and end time, as well as with the event's type. In both research directions, a *frame* (or sound clip) indicates the unit of analysis and may contain several events that may overlap in time.

The referred classifiers, in sound event detection, ideally, have each one of the acoustic events instances in the training data, labeled with their start and end time. This type of labels is referred as *strong labels*, nevertheless, acquiring them is a costly process that also requires careful attention to detail by the annotator. On the other hand, the labels that do not contain any information about the temporal location of each event or the number of occurrences in the recording are called *weak labels*.

A *sound spectrogram* is an image of the time-varying spectral representation, produced by applying the *short-time Fourier transform (STFT)* to successive overlapping frames of an audio sequence. The horizontal dimension corresponds to time and the vertical dimension corresponds to frequency. The relative spectral intensity of a sound at any specific time and frequency is indicated by the color/grayscale intensity of the image.

Model performance and capability to capture the natural variability of data can be increased with the use of data augmentation techniques. Such signal transformations may include time shifting, volume control or adding additive noise to the acoustic data. The first concept consists of shifting a sound event in time and the second controls the volume of the acoustic signal. Additive noise consists of summing noise to the original signal, whether that represents Gaussian noise, uniform random noise, or a background recording, process further detailed in [1]. The developed approach adds *Gaussian or Pink noise*, with respect to the *Signal-to-Noise Ratio (SNR)*, that adaptively sets an appropriate noise level based on the amplitude of the original sound signal. Furthermore, Gaussian noise, referenced as white noise, is a noise over the whole frequency range, oppositely to Pink which has a gradual decrease in noise intensity from low frequency to low frequency bands, approximating the characteristics of noise of the natural world.

3 Related Work

Research on automatic bioacoustic analysis lead to the application of sound recognition techniques such as *Gaussian mixture model (GMM)* [2] or *hidden Markov model (HMM)*[3] to conduct species detection. The *Mel Frequency Cepstral coefficients (MFCC)* and spectrograms were the most common input features used by these algorithms, successful in the identification of individuals species but that failed to perform with recordings with high diversity. Consequently, subsequent research focused on developing algorithms capable of classifying multiple species in audio data.

In recent years, Convolutional Neural Networks (CNNs) have outperformed the former models in visual recognition tasks, namely in large-scale image and video recognition, mostly due to the late availability of large public datasets of images such as the ImageNet [4]. Several CNN architectures were applied to this dataset to perform image classification, in particular fully connected *Deep Neural Networks (DNNs)* such as AlexNet [5], VGG [6], resNet[7], among others.

Transfer learning is used to avoid the large amount of training data and time that deep neural networks with initially randomized weights require to achieve reasonable performance. In particular, given the context of environmental data in which labels are costly, one can take advantage of this technique by retraining with new data a model already optimized for a similar dataset to improve performance. The ResNet50 [7] model is a classic neural network used as backbone for many computer vision tasks and it was trained on the ImageNet dataset, which contains over one million images across 1000 classes. Despite not containing spectrograms, models pre-trained on this dataset learn a variety of image features and have been successfully tuned to spectrogram classification [8][9]. Consequently, CNNs are able to exploit the adjacency properties of audio signals and recognize patterns in the spectrum image, achieving state-of-the-art performance in sound event detection and classification.

Furthermore, the competition "Rainforest Connection Species Audio Detection", previously referenced, encouraged contenders to develop models that aimed to automate the detection of several species in the RFCx audio files. Thus, this competition provides multiple audio processing methodologies and models architectures which concern species detection in noisy soundscapes with limited training data, such as the ones presented in [8][9].

Supervised learning algorithms need a considerable amount of labelled data to achieve good performance, requiring a time-consuming annotation process dependent on expertise annotators, a limiting factor in the context of environmental data. In this sense, *data augmentation* emerges as a strategy to diminish this problem, as it synthetically generates new labeled samples from the existing ones, expanding the effectiveness of the training set. A more detailed overview of the techniques is given in [10], including methods that apply various audio signal transformations, such as time stretching, pitch shifting, dynamic range compression, adding random noise, etc. Additionally, SpecAugment [11] is a simple data augmentation method for speech recognition, that contrasts with the most common ones, as it is directly applied on Log-Mel spectrograms instead of raw audios.

4 Methods

This study proposes a bioacoustic classification framework using transfer learning of deep neural networks. Thus, this sections focuses on detailing each step of the suggested end-to-end pipeline that results in a **classification model**, process represented in Fig.1.

The starting point consists of converting the sequence of sounds (**raw audio)**, that is, the time series, into audio features that can capture the distinctive properties of each event. Given the results obtained by deep neural networks in image classification problems our **feature extraction** step focuses on the extraction and processing of the Mel spectrograms, that are image representations of sound. So, we explore the learning ability of deep neural networks, namely Convolutional Neural Networks, describing the **training** process of the model with the mentioned spectral shape features.

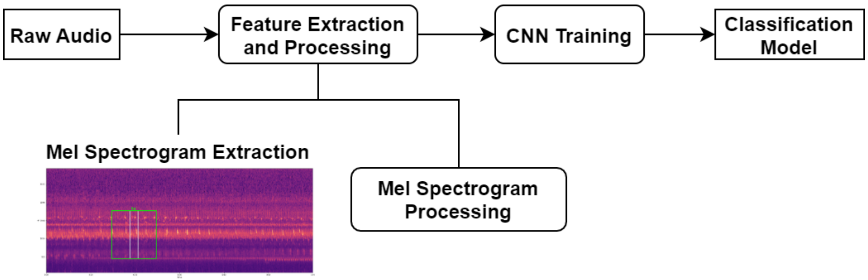
Raw Audio
Feature Extraction and Processing
CNN Training
Classification Model
Mel Spectrogram Extraction
Mel Spectrogram Processing

Fig. 1: Classification model training flowchart

4.1 Feature Extraction and Processing

The proposed window function defines the frame's center (window center) as the sum of half of the maximum time interval (maximum delta) of a given dataset, to the interval's beginning (interval start) of the concerned sample. The start (window start) and end (window end) of the frame is the result of subtracting and adding to the center, respectively, the selected frame length (window duration) divided by two.

$$\text{window center} = \text{interval start} + (\text{maximum delta} \ / \ 2)$$
$$\text{window start} = \text{window center} - (\text{window duration} \ / \ 2)$$
$$\text{window end} = \text{window center} + (\text{window duration} \ / \ 2)$$

Additionally, it is important to note that the sampling rate by which the audio is extracted must be taken into consideration when building the mentioned window. All in all, the window function allows for a training set composed only by frames that are linked to a given event.

The extracted Mel spectrogram (**Mel spectrogram extraction**) is represented in Fig.1 by the green box, being the white box the representation of the event's time interval. Each Mel spectrogram is computed using the *librosa* Python package with the default settings (sampling rate = 48 kHz, NFFT = 2048, hop length = 512, window length = 2048, Hann window) specifying however the number of mel bands (n_mels = 224) and if available the minimum and maximum frequency. The frequency interval corresponds to the minimum and maximum value registered in the dataset, with a 10% margin to increase the considered interval.

The resulting Mel spectrograms, as a part of the **Mel spectrogram processing** step, are converted to units of decibel (dB), resized to the dimensions supported by the model, that for the ResNet50 correspond to 224x244 images, and normalized with the min-max scaling. Finally, the spectral features are converted to color images, that is, images with RGB channels and given the transfer learning setting, a final step is necessary in which the spectrogram is processed to the adequate image format of the selected backbone model. For the ResNet50 model, for instances, the images are converted from RGB to BGR, then each color channel is zero-centered with respect to the ImageNet dataset, without scaling.

4.2 Model Training

The proposed model uses the pre-trained ResNet50 weights used for ImageNet classification, and includes only the feature extraction layers of this model, excluding the remaining layers, referred as the network "top". Hence, the knowledge obtained in image classification, namely the detection of basic image features, can be transferred (**transfer Learning**) to the task at hand by using the weights of the optimized model. In this sense, by freezing some layers of the pre-trained model and only training the last several layers the model can be fine-tuned to our problem.

Our **model architecture** comprises two *fully connected (FC)* layers, the first consists of 512 nodes and uses the "Relu" activation function that converts negative inputs to 0. This layer is followed by a batch normalization and drop-out layer with a drop-out rate of 50%, in which each node is ignored with a 50% probability. The final layer, given the binary classification setting, consists of one node that passes through the sigmoid function.

The **training** step consists of training each model on the spectral features to obtain a classification model for each binary classification setting. Also, we use the Adam optimization method with a learning rate of $1 * 10^{-4}$ and decay of $1 * 10^{-7}$. Moreover, the binary cross entropy loss function is utilized and 30 epochs are applied. These parameters result from the fine tuning process in which we analysed the values who favored the model's performance.

Model performance and capability to capture the natural variability of data can be increased with the use of **data augmentation** techniques. Thus, two approaches are followed as a way of increasing the training set, process further evaluated in Section 5.2, in which the first randomly adds one of the two additive noises, Gaussian or Pink, to the audio signal, time shifting and controlling its volume afterwards, and the second applies the SpecAugment technique to the Mel spectrogram.

5 Results

5.1 Multispecies bioacoustic classification

The proposed bioacoustic classification framework is evaluated on the audio data from the Kaggle competition that was collected from about 700 sampling sites across the mountains of Puerto Rico at a sampling rate of 48 kHz with 24 kHz bandwidth, following a schedule of 1-minute audio recording every 10 min, as described in [8]. It concerns the classification of 24 bird and amphibian species which inhabit the tropical mountains and it provides two datasets. The first has information about the true positive events, having a labelled interval which refers to the specie call, and the second has information about the false positive events, detailing by opposition the interval where a certain specie does not appear. Each sample on both datasets provides information about the specie present in the audio sample, the sound´s song type as well as the frequency and time interval of the event.

As previously mentioned, there are 24 annotated species in the provided dataset what would suggest a 24 multi label classification setting. Nevertheless, two species have more than one song type, having both type 1 and 4, revealing the need of two additional labels. As a starting point, the created training set disregards the song type 4 for the mentioned species.

In that sense, our approach transforms the 24 multi label classification setting into 24 distinct classification problems. Thus, this division results in a binary classification setting where we train a model capable of learning the presence or absence for each specie. Moreover, the upcoming sections describe several experiments in which the concerned models followed the architecture described in section 4.2. In particular, the results represent the average of each specie related model's score, taking the example, the accuracy scores displayed in Fig. 2 refer to the average of the accuracy scores of all 24 species. It is also important to note that from these results, the ones presented in section 5.2.1 refer to each specie related model as the analysis discriminates all species.

Also, the training set includes the maximum number of true positive events for each specie, that for the majority of species corresponds to approximate 50 samples. Additionally, it encompasses different quantities of true positive augmented samples, as further described in section 5.2. Lastly, a subset of the available 350 false positive samples is extracted in the same quantity as the true positive subset, that may contain augmented instances. In detail, for the baseline data augmentation approach if one specie has 50 true positive samples, it will have 50 augmented samples and 100 false positive samples, resulting in a balanced training set for each specie.

5.2 Window Size and Data Augmentation

The first approach aims to assess the effect of different window sizes and data augmentation techniques on the performance of each model. Thus, the considered frame lengths were 2, 5 and 10-second-long, as more than 80 percent of the recorded events have intervals smaller or equal to 4 seconds. Also, by including the 2 second window one can verify if smaller frames can capture enough image traits to conduct automate species detection. In addition, for each window size, we trained a model with a training set that did not include augmented recordings (baseline) and compared it to two models whose training set contained samples augmented by the two techniques described in Section 4.2.

As detailed in the previous section, the training set with no augmented recordings includes the maximum of true positive samples, having the same number of negative samples. Conversely, both training sets with augmented instances, from the two aforementioned augmentation techniques, differ from the latter by having augmented samples in the same number as the true positive calls, thus enabling the use of more negative instances. Finally, the evaluation metrics were *accuracy*, which corresponds to the percentage of correct predictions (tp + tn) over the total number of instances evaluated (tp + tn + fp + fn); *precision*, that measures the fraction of an identified event correctly classified (tp ÷ (tp + fp)); and *sensitivy or recall*, which measures the fraction of positive patterns correctly classified (tp ÷ (tp + fn). The test set was classified with a threshold score of 0.6.

As depicted in Fig. 2, the inclusion of the augmented samples in the training set lead to an increase of the accuracy score across all windows sizes. The model trained on the 10-second-long window failed to capture the data's variability leading to the worse results on the precision and recall scores. The 5 second window obtained a significant accuracy increase, registering its best precision and recall score (0.77 and 0.78) with the SpecAugmented spectrograms. Furthermore, the shortest concerned window obtained similar results, in comparison to the 5 second window, in terms of accuracy, achieving nevertheless lower precision and recall results.

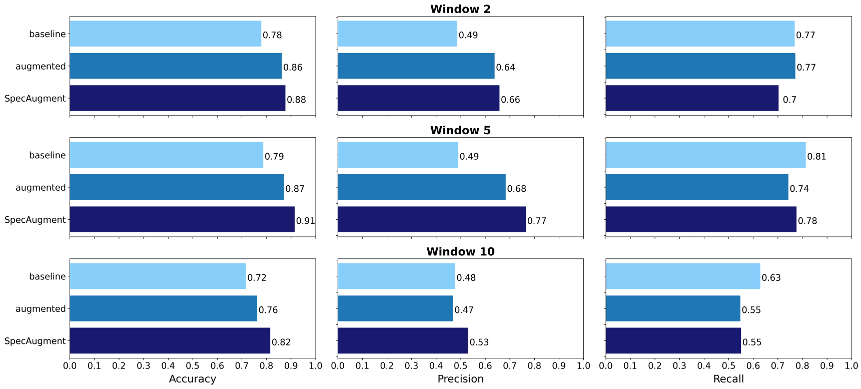

Fig. 2: Window size and data augmentation techniques effect on accuracy, precision and recall.

All in all, the results confirm the well known precision-recall relation, in which generally an increase in precision leads to a decrease in recall, and vice-versa. Consequently, a balance is desired if false positives and false negatives are equally significant, which is not the case in our problem's spectrum as recall is slightly more important because false negatives are more costly. From this experiment, both 2 and 5 second long frames appear to be able to capture the distinctive traits of each Mel spectrogram with the 5 second one standing out, being the concerned window from this point forward.

5.2.1 Dynamic Window Sizes

The results obtained in the previous section are strongly marked by the models that fail to differentiate both classes, difficulty amplified with the 10 second window. So, in order to assess if each model would perform better with a tailored window size a different approach was experimented. More concretely, each specie related spectrogram was obtained by taking into consideration the mean time interval of each specie call with a 1 second margin, which implied that, for instances, a specie with an average interval call of 2 seconds would have a 3-second-long Mel spectrogram.

Hence, the average-precision, presented in the definition 1, was the metric used to compare the model trained on the dynamic windows with the one trained on the fixed window size (5 second frame length), that had recordings augmented with the SpecAugment method, as Fig. 3 showcases.

Definition 1. ***Average-precision:*** *Summarizes the precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.* P_n *and* R_n *correspond to the precision and recall at the nth threshold.*

$$AP = \sum_{n} (R_n - R_{n-1}) \, P_n \tag{5.2.1}$$

In particular, Fig. 3 demonstrates the difference in average-precision between each model trained on the dynamic window size (red) and those trained on the 5 second window with SpecAugmented samples (blue). The mean average-precision scores for the dynamic and fixed size approach are 0.52 and 0.63 respectively. Furthermore, species with a mean window size smaller than 5, such as 11 and 18, for instances, are the ones who benefit the most from the dynamic approach. Also, it is possible to understand the impact that models with lower scores have on the metrics depicted in Fig. 2.

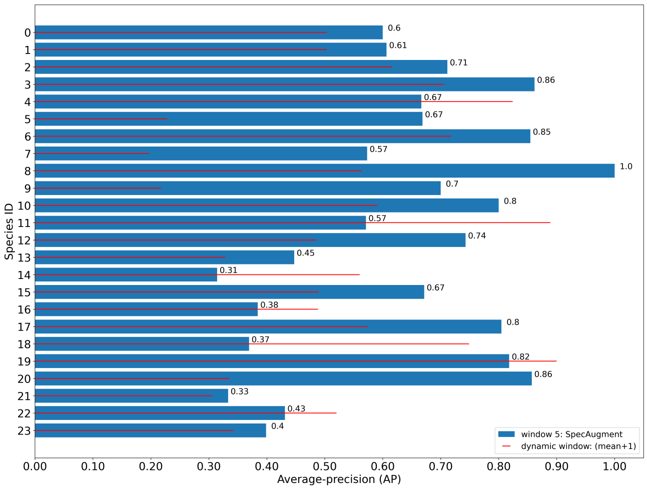

Fig. 3: Average-precision of dynamic and fixed window size approaches.

Lastly, it is important to note that the difference in the training size, that for species 2, 9, 17, 20 and 22 is significantly smaller due to the available TP samples, is not the only cause for a poorer model's performance as the average-precision of specie 17 is higher than the one of specie 23, for example. To sum up, the goal of this approach was to understand if a small combination of window sizes, as a large one would be extremely costly in the prediction step, would favor the model's results. Despite the aforementioned improvement in the species that register smaller calls, the overall performance was not sufficient to justify the cost that a windowed approach would require.

5.2.2 Predictive threshold and Final Remarks

The predictive threshold represents the probability value by which the prediction sample is classified given the binary setting, that is, if the probability returned by the model is superior to the defined threshold the sample will be classified as belonging to the class, and vice-versa. On that account, the previous approaches considered a predictive threshold of 0.6, achieving a precision of 0.77 and a recall of 0.78 with the best model. Nonetheless, one can try to improve the precision score by increasing the predictive threshold.

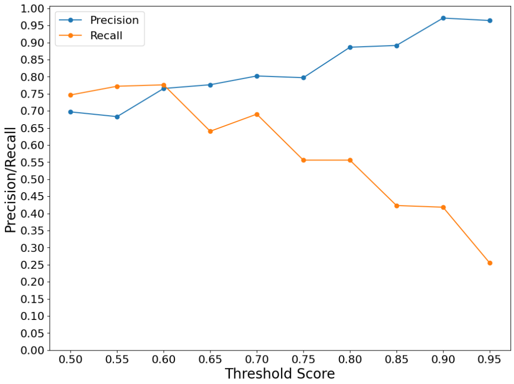

Fig. 4: Precision/Recall threshold curve of a model trained on a 5-second-long window with SpecAugmented samples).

Hence, Fig. 4 displays the variation of mean precision and recall with different thresholds so that the influence of this value on the network performance could be determined. The higher the threshold the higher the precision score, with a significant decrease in recall.The precision-recall balance is achieved somewhere between the 0.50 and 0.65 predictive threshold, with the 0.60 threshold registering the optimal threshold value for the develop approach with a precision of 0.77 and a recall of 0.78.

Following the prior analysis, the last attempt made to improve the precision-recall score had a predictive threshold of 0.6 and a larger training size in comparison to previous approaches. This was accomplished by having 3 subsets derived from the true positive events, the first had the original TP instances, and the other two included augmented samples, using the two data augmentation techniques introduced in section 5.2. Similarly, the false positive subset had the same size as the true positive one. Nevertheless, this approach did not improve the results obtained by the model trained only on the samples augmented with the SpecAugment, so further work will expand on this matter.

6 Discussion

This report is the starting point of an end-to-to pipeline capable of performing bioacoustic classification. In this regard, several experiments were made on multiple levels to validate the proposed approach, from the Mel spectrogram extraction through the use of a window function to the training of a deep neural network model that classifies events in environmental recordings. In detail, we were able to tackle the limited training data challenge with the described data augmentations techniques, improving the baseline approach and contributing to the definition of the mentioned framework.

It is also important to note that it is possible to compare the proposed approach with the ones developed for the Kaggle competition. Nevertheless, in this study we do not make any comparative analysis for two main reasons, the first results of the fact that this research is only the starting point of the bioacoustic classification framework we intend to define, and the second lies on the fact that the competition scoring system in which the performance of the proposed solution is evaluated is not public. For those reasons, future work will concern the mentioned comparison.

The transformation of the 24 multi label classification setting into 24 binary classification problems is a resourceful way to reveal the difficulties that each model faces in the classification setting, that otherwise would be more difficult to disclose. Future work will try to understand why some models fail to perform given the current best performing approach. The lack of results may be because of the confounding signals and noise present in the recordings, which may result in nondistinctive spectrograms. The data augmentation process needs to be optimized to take full advantage of the available true and false positive calls, by combining the SpecAugment method with others which were not yet applied, such as time streching, pitch shifting, among others. Also, alternatives to the ResNet50 model, used as backbone to the current classification model need to be considered, such as EfficientNetB1, VGG16, etc.. Finally, pseudo labeling may have a positive impact on the model's performance by increasing the limited amount of training samples. The semi-supervised learning technique achieves this by training a model on the available set of labeled data and predicting labels on unlabeled data. Later, it combines both datasets, that is, the one with the true labels with the one with the predicted labels and retrains the model. The mentioned techniques will be the focus of future research so that the proposed approach may be improved.

7 Acknowledgements

This work was supported by national funds by Fundação para a Ciência e Tecnologia (FCT) through project VizBig (PTDC/CCI-CIF/28939/2017)

References

1. V.-V. Eklund, “Data augmentation techniques for robust audio analysis,” Master’s thesis, Tampere University, 2019.
2. P. Jančovič and M. Köküer, “Automatic detection and recognition of tonal bird sounds in noisy environments,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–10, 2011.
3. I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, “Automatic bird sound detection in long real-field recordings: Applications and tools,” *Applied Acoustics*, vol. 80, pp. 1–9, 2014.
4. A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” *Neural Information Processing Systems*, vol. 25, 01 2012.
5. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
6. S. Liu and W. Deng, “Very deep convolutional neural network based image classification using small training sample size,” in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734, 2015.
7. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
8. M. Zhong, J. LeBien, M. Campos-Cerqueira, R. Dodhia, J. Lavista Ferres, J. P. Velev, and T. M. Aide, “Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling,” *Applied Acoustics*, vol. 166, p. 107375, 2020.
9. J. LeBien, M. Zhong, M. Campos-Cerqueira, J. P. Velev, R. Dodhia, J. L. Ferres, and T. M. Aide, “A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network,” *Ecological Informatics*, vol. 59, p. 101113, 2020.
10. J. Abeßer, “A review of deep learning based methods for acoustic scene classification,” *Applied Sciences*, vol. 10, no. 6, 2020.
11. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.