When Computers Learn Frog Calls: Deep Neural Network Classification of Frog Species

Ami Kano, Kathryn Meldrum, Tyler Valentine
University of Virginia
ak7ra, kmm4ap, xje4cy@virginia.edu

May 12, 2023

Abstract

Amphibian species are facing increasing rates of extinction. In an effort to better monitor the changing population levels and understand how to help at-risk species, researchers have relied on automated audio recording devices placed throughout their habitats. These devices produce thousands of audio files and are time-consuming for researchers to parse through manually. Automated software for audio classification has been applied to problems such as speech or music recognition for over a decade; however, applications for conservation efforts are more recent. Previous studies have shown success with applying convolutional neural networks (CNNs) for classifying audio of bird species. This work focuses on classification of six frog species. Our results indicate high prediction accuracy with our test set and an ability of the model to generalize to new, long-form audio files.

1 Introduction

According to the International Union for the Conservation of Nature, 41 percent of amphibians globally are endangered [1]. In Ecuador, one of the most biodiverse countries in the world, 57 percent of its amphibian species are endangered. The decline of these species is caused by an array of factors, including habitat loss and emergent diseases [2].

Conservation biologists in Ecuador have been taking action to reverse these trends. With a focus on protecting frog species, they have been taking frogs from the wild, breeding them in their laboratory, and then releasing the offspring into their natural habitats [3]. One challenge with this approach is the difficulty of searching for frog species which are continuing to decrease in number. Although some conservation efforts for mammals involve physical or motion-triggered camera traps, these methods would leave frog species too vulnerable to predators. Further, cameras are not suitable for the small, ectothermic frogs, many of which blend into their habitat [4].

Although it is challenging to find these frogs, it becomes even more difficult to reliably monitor changes in their locations and populations over time. Traditional sampling methods, where biologists are sent to the field to collect and catalog samples, are expensive and can result in incomplete and limited datasets. Biologists cannot remain in the field for 24 hours a day and are unable monitor various sites simultaneously.

To address these challenges, scientists have employed audio detection of species throughout the past decade. In 2013, a group of scientists in Puerto Rico developed hardware and software that could collect, process, and analyze audio recordings in real-time. By placing automated recording stations in remote locations, the group was able to receive audio recordings sent via a radio antenna up to 40 km away to their base station [5].

Thousands of audio recordings from various wetlands and forests must be searched for animal detection. Some scientists aim to solve this problem through the use of citizen science. Frog Find, a project developed by the Conservation Science Research Group at the University of Newcastle, has uploaded thousands of 30-second audio files to the Zooniverse, the world's largest and most popular platform for people-powered research. Their goal is for users to listen to each recording and identify the presence or absence of select endangered frog species in Australia to help with their conservation efforts.

Automated software is powerful for its ability to perform this task more efficiently. As a research field, computational bioacoustics has accelerated due to the growth of affordable digital sound recording devices, progress in big data, signal processing, and machine learning. However, as recent as 2017, reviews and textbooks did not give much emphasis to deep learning as a tool [6].

Although deep learning has been applied to audio classification tasks (such as speech and music) for over a decade, applications in wildlife acoustics are much more recent. One example is the Life-CLEF bird identification challenge in 2016; [7]. participants were asked to identify all of the active singing bird species in each audio file with machine learning models. Out of a total of 83 research groups worldwide who registered for the competition, six of the registrants were noted for implementing highly successful methods. Half of those teams applied convolutional neural networks (CNN) after converting the audio files into spectrograms and applied a variety of different models, such as AlexNet.

The ability to use CNNs for classifying audio is made possible by Mel-spectrograms, an image that represents the frequency compositions of an audio signal over time. The x-axis of a Mel-spectrogram shows time, while the y-axis shows frequency and the amplitude is shown by color. This allows for the viewer to essentially 'see' the variables stored in a audio tensor.

While bird species classification models have become more sophisticated and accurate in recent years, less attention has been placed on audio classification of other species targeted by conservation efforts. Our work aimed to close this gap by applying deep learning for classification of frogs. Since several successful teams chose to convert raw audio files to Mel-spectrograms for use in training convolutional neural networks, we chose to take the same approach.

2 Related Work

The creators of the dataset - Terneux, Nicolalde, Nicolalde, and Merino-Viteri - produced a paper called *Presence-absence estimation in audio recordings of tropical frog communities* [8]. This paper shows a Gaussian Mixture Model trained to predict the species of a frog call. Although the specific kind of model is different, their paper is quite relevant to this project as we have chosen to use their data. The segmentation of audio files described in the paper is also loosely mimicked in this project.

In addition, there are many past works that attempt to classify animal calls. For example, classifying bird calls with neural networks has been attempted countless times, with a notable example being *Bird sound recognition using a convolutional neural network* [9] by Incze, et al. The authors of this paper convert audio data into spectrograms and create a convolutional neural network. The specific CNN described in this paper is a fine-tuned version of a pre-trained MobileNet model, but other papers have attempted to design their own CNN.

3 Data

The data for this project were collected in the Yasuni Rainforest. The audio files ranged from 2-20 minutes and the species in each audio file was identified and annotated by researchers from the Museo de Zoología QCAZ. The original dataset contained ten species, each with at least 39 call examples. We limited this project to the six species that had at least 100 frog calls represented in the audio files in order to ensure that we would have enough training data.

3.1 Preprocessing

In order to prepare the data for a convolutional neural net model, we converted the frog calls into images. To do this we used the python library Librosa, which has functionality to identify points of higher amplitude in audio files, such as frog calls, bird calls, and talking. We segmented half a second before and after each onset point, resulting in a one-second audio clip with a centered onset point. The one-second audio files were then converted to Mel-spectrograms, also using the librosa library, to allow us to employ an deep learning image classification model. Then, 100 images of each of the six frog species' calls were manually selected to use for model training and testing. Additionally, 100 background classification pictures were selected which consisted of a variety of non-frog-call spectrograms including silence, insect calls, bird calls, and microphone movement.

4 Model

The choice of model was based on a need to maximize classification accuracy while minimizing model complexity. The AlexNet architecture, which won the ImageNet competition in 2012 with a high accuracy and only five convolutional and three fully-connected layers [10], satisfied both of these requirements. To use architecture in our study, we modified several of the feature map dimensions, as shown in Figure 1. For the first input to the first convolutional layer, the input images were all of the same size with a height of 640 pixels and a width of 480 pixels. The final layer has a size of seven for the seven possible classes.

The model was trained with 100 images from each class for a total of 700 images with the PyTorch library implemented in python. After splitting the dataset randomly into 70 percent for training and 15 percent each for validation and testing, the model was trained with 75 epochs with the training set. It was found that the best model was after epoch 46, where the validation accuracy was 99 percent and the model did not appear to be overfitting to the training data. This final model was applied to the testing data set to achieve an accuracy of 97.1 percent.

5 Evaluations

5.1 tSNE

T-distributed stochastic neighbor embedding (tSNE) is a visualization technique that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map [11]. Using tSNE, visualizations that show how each data point is interpreted by the CNN were created.

Figure 2 is a tSNE visualization of the datapoints in the validation set as interpreted by the first epoch version of the CNN model. It is clear that the classes are not effectively clustered. This is due to the low quality of the underlying model; the validation loss and accuracy for the model at epoch were approximately 1.935 and 8.6% respectively.

However, after several epochs, the models begin to distinguish the classes. As shown in figure 3, at the 19th epoch the tSNE visualization depicts the datapoints of each class being plotted closely together. This illustrates how the CNN model is improving its classification prediction performance by updating the weight parameters towards the optimal values every epoch. The validation loss and accuracy of the model at epoch 19 are approximately 0.092 and 97.1% respectively, which is a further reflection on how well the model distinguishes the different classes.

5.2 Saliency Maps

Saliency maps in deep learning mark the importance of each of each pixel in an image that is being classified. [12]. In the case of our Mel-spectrograms, we employed saliency maps to investigate whether the most important pixels towards classification were the highest amplitude parts of the heat map. In other words, if the most important pixels were where humans would see the shape of the frog call. This was especially important for this particular project because many of our training and testing images came from only a few unique files which may have had constant, distinctive background noise. This setup could make our model susceptible to picking up on background indicators, such as a particular bird call, as an identifier for a particular species rather than the species' call itself. A Saliency map for each species is shown in Figures 4-9, these images were taken from the test set of each species.

It is apparent by our saliency maps that the model does in fact place the highest importance on the frog call pixels when making a classification. We also chose to examine an image from the background class test group. With this saliency map it can be seen that there is a are more than just a few image pixels of relatively high importance, and that these important pixels are found towards the top and bottom of the image, where a frog call would not be located in our image (Figure 10).

5.3 Long Audio Files

A possible application of this model is being able to classify second-length segments from raw field audio

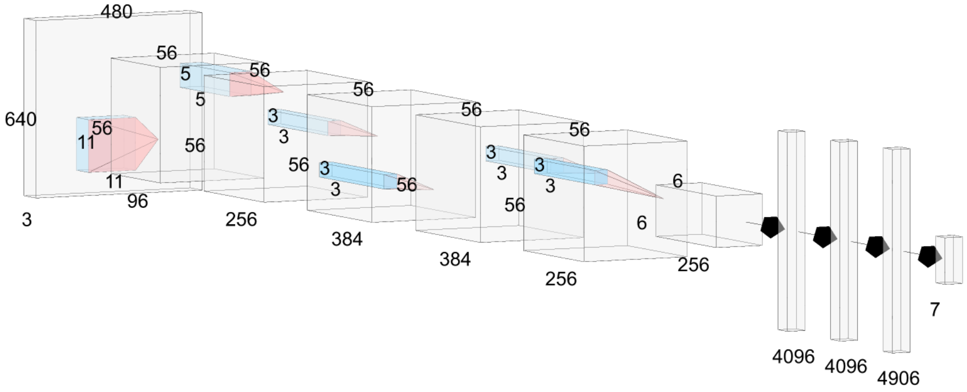

Figure 1: The custom architecture modified from AlexNet. This diagram was created with software developed by Alexander Lenail.

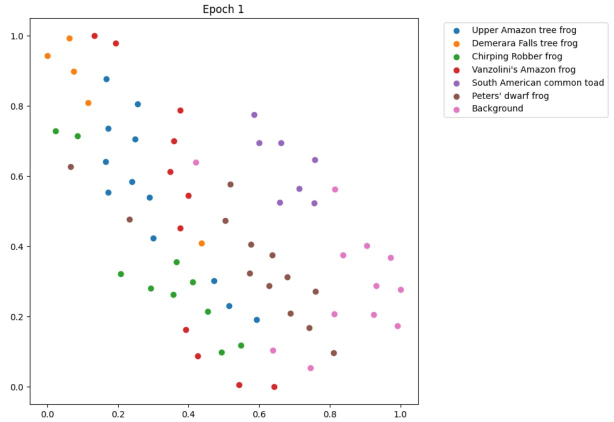

Figure 2: tSNE visualization from the first epoch of the model.

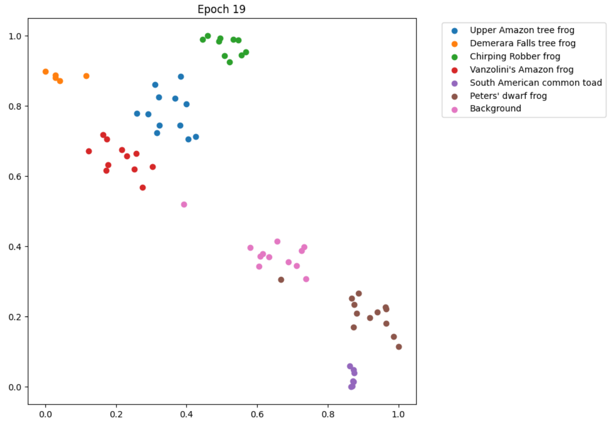

Figure 3: tSNE visualization from the 19th epoch of the model.

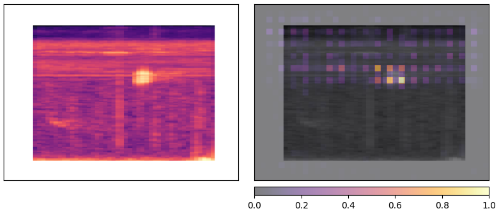

Figure 4: Upper Amazon Tree Frog Saliency Map

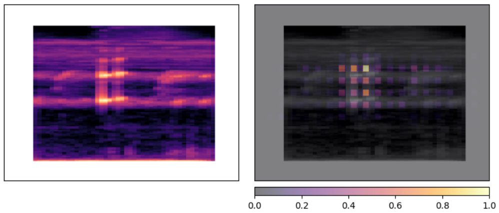

Figure 5: Demerara Falls Tree Frog Saliency Map

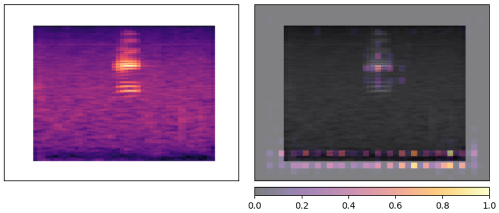

Figure 6: Chirping Robber Frog

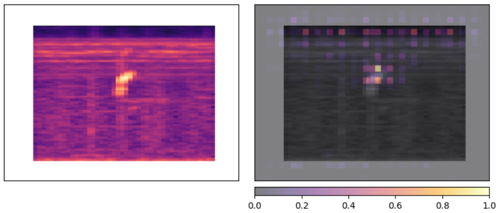

Figure 7: Vanzolini's Amazon Frog Saliency Map

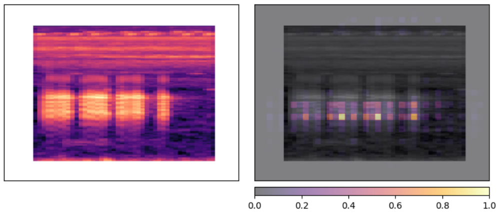

Figure 8: South American Common Toad Saliency Map

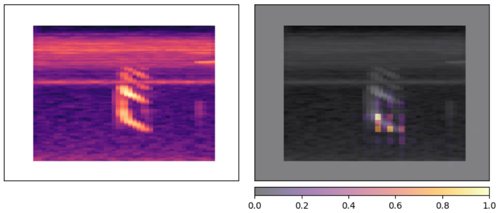

Figure 9: Peter's Dwarf Frog Saliency Map

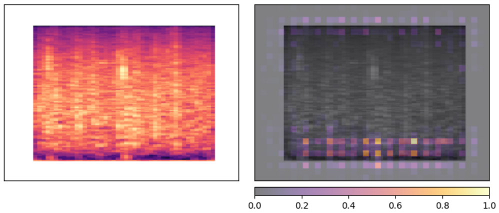

Figure 10: Background Class Saliency

for species monitoring. By testing the model on audio that has not been segmented perfectly about a frog call as we did with the previous testing data, we can demonstrate the model's utility towards largely unprocessed data.

5.3.1 City Noise Audio File

As a control, we tested a two-minute audio file that was recorded in Central Park, New York City. This data was not processed further than segmentation into consecutive, 1-second clips and conversion into Mel-spectrograms. We classified the spectrograms sequentially, to identify presence of South American Frogs in each second of the file as shown in Figure 11. As expected, the file was largely classified as background. Only one second was classified as a frog, with low confidence.

5.3.2 Yasuni Rainforest Audio File

After demonstrating that the model would not classify non-frog sounds as frogs, we applied the same segmenting method to an audio file from the Yasuni Rainforest. This audio file came from the same source as the data we trained our model on but was unlabeled and contained multiple frog species. We used

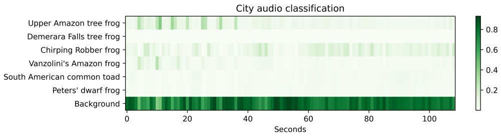

Figure 11: City Audio Classification.

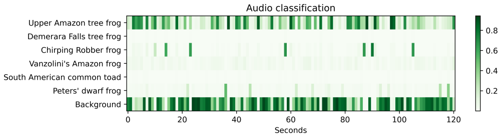

Figure 12: Rainforest Audio Classification

a two-minute section from 7:50-9:50 minutes in the 'cc149A.wav' unidentified long recording. While we were unqualified to explicitly annotate seconds of this file for different frog species, we can see that the resulting classifications of the file, in Figure 12, contain several types of frogs as well as some background, which seems to accurate to what can be heard in the file.

6 Conclusion

6.0.1 Impact

Neural network models that classify calls of non-bird species have not been explored deeply. The CNN model created for this project proves that calls of non-bird species, such as those of frogs, can successfully be classified by deep learning models. The success of this project suggests the viability of expanding research in computational bioacoustics to frog species.

6.0.2 Future Work

The primary limitation of the model presented here is that it thus far can only identify six of the many species of frogs present in the area it is designed for. In order to train the model to identify more species, greater sample sizes of audio data from other frog species must be collected. Additionally, this type of model could be used to track other types of endangered animals who have distinctive calls, such as whales, manatees, and bats.

7 Metadata

The presentation of the project can be found at:

`https://github.com/ak7ra/frog_classification/blob/main/presentation_recording.mp4`

The code/data of the project can be found at:

`https://github.com/ak7ra/frog_classification/`

References

[1] Y. Jiang, W. B. Liao, and A. Kotrschal, “Small-scale dams deplete frogs and toads,” *Conservation Science and Practice*, vol. 4, May 2022.

[2] H. M. Ortega-Andrade, M. R. Blanco, D. F. Cisneros-Heredia, N. G. Arévalo, K. G. L. d. Vargas-Machuca, J. C. Sánchez-Nivicela, D. Armijos-Ojeda, J. F. C. Andrade, C. Reyes-Puig, A. B. Q. Riera, P. Székely, O. R. R. Soto, D. Székely, J. M. Guayasamin, F. R. S. Pesántez, L. Amador, R. Betancourt, S. M. Ramírez-Jaramillo, B. Timbe-Borja, M. G. Laporta, J. F. W. Bernal, L. A. O. Cachimuel, D. C. Jácome, V. Posse, C. Valle-Piñuela, D. P. Jiménez, J. P. Reyes-Puig, A. Terán-Valdez, L. A. Coloma, M. B. P. Lara, S. Carvajal-Endara, M. Urgilés, and M. H. Y. Muñoz, “Red list assessment of amphibian species of ecuador: A multidimensional approach for their conservation,” *PLOS ONE*, vol. 16, p. e0251027, May 2021.

[3] N. Forrester, “Guardian of ecuador's diverse — and vanishing — frog species,” *Nature*, vol. 615, p. 960–960, Mar 2023.

[4] T. Hammond, “Finding frogs in the field using new technology,” Aug 2019.

[5] T. M. Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez, “Real-time bioacoustics monitoring and automated species identification,” *PeerJ*, vol. 1, p. e103, Jul 2013.

[6] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, Mar 2022.

[7] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, and A. Joly, "LifeCLEF Bird Identification Task 2016: The arrival of Deep learning," in *CLEF: Conference and Labs of the Evaluation Forum*, vol. CEUR Workshop Proceedings, (Évora, Portugal), pp. 440–449, Sept. 2016.

[8] A. E. Terneux, D. Nicolalde, D. Nicolalde, and A. Merino-Viteri, "Presence-absence estimation in audio recordings of tropical frog communities," Jan 2019. arXiv:1901.02495 [cs, eess, stat].

[9] Incze, H.-B. Jancsó, Z. Szilágyi, A. Farkas, and C. Sulyok, "Bird sound recognition using a convolutional neural network," in *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, p. 000295–000300, Sep 2018.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, (Red Hook, NY, USA), p. 1097–1105, Curran Associates Inc., Dec 2012.

[11] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, p. 2579–2605, 2008.

[12] M. Brahimi, M. Arsenovic, S. Laraba, S. Sladojevic, K. Boukhalfa, and A. Moussaoui, *Deep Learning for Plant Diseases: Detection and Saliency Map Visualisation*, p. 93–117. Human–Computer Interaction Series, Cham: Springer International Publishing, 2018.