Shallow Neural Networks, Transfer Learning, and Template Matching Applied to Bird Audio Classification

Nathan Kueterman

University of Dayton

Dayton, OH

nkueterman1@udayton.edu

*Abstract*—Deep learning is an emerging field that has shown promising results for image classification. The focus of this study was to analyze the performance of modern machine learning algorithms in the application of classifying bird species from recorded calls. Bird songs are specific to each species and have a distinct signature, which was the aspect of the data being leveraged. The data set was collected by various bird enthusiasts and uploaded to xeno-canto.org where the files are open to the public. Statistical analysis was performed on these calls and predictions were made on which species produced the audio which were limited to American Robins and Mourning Doves. An audio spectrogram served as the input image to the neural networks while the raw audio signal was used for template matching. The technique that had the highest accuracy was the transfer learning approach, which utilized the pre-existing neural network known as AlexNet. The shallow neural net had a slightly lower accuracy, while the rudimentary Spectral Angle Mapper (SAM) classifier performed at the lowest accuracy. All the classification techniques utilized have associated trade-offs which are explored in the conclusion of this study.

Keywords—machine learning, deep learning, spectrogram, neural network, birds, classification, bioacoustics, epoch, batch normalization, Rectified Linear Units (ReLU)

# Introduction

The classification of a bird call strictly based on audio is something that currently only a trained human ear can accomplish. That trained ear is also limited to the amount of calls that it has encountered and with up to 10,000 species existing on Earth it is likely beyond human comprehension to reliably classify all the species on these grounds. This study develops a network that could potentially classify most of these species, or at least the relevant species of a particular region. The main limitation for this comprehensive solution is the limited amount of data available for some species in remote locations or the species that use calls more scarcely. There are also other considerations such as the noise of the recording environment and the ability to scale the data up to a level of hundreds or even thousands of classes.

If implemented in a reliable, scalable manner, the techniques utilized here could benefit conservation, ecology, and archival efforts across the globe. There really is not a need to build a tool that can classify all species at the same time but rather about 100 species that the algorithm might encounter being in a somewhat limited geographical region. There are networks that can classify this quantity of classes with high level of accuracy. [1] These algorithms have not been tested fully in a real-time, in-the-field environment which would require the most robust algorithm available to deal with the challenges of this classification problem. This study provides a rudimentary evaluation of the possible approaches to creating this ideal classifier.

The data used in [1] was gathered via preexisting data sets recorded in a multitude of different environments. In this study, the data was collected by an organization known as Xeno-Canto which provides an online database of bird calls recorded by bird enthusiasts all around the world. The database is open to download the files as well as upload in the hopes of building a comprehensive database for archival and other bioacoustics research such as classification via deep learning. In this study the data was limited to two classes: American Robins and Mourning Doves. This was done to test the fundamental aspects of the classification methods rather than to test each’s scalability.

Three methods were utilized to classify the bird calls. These methods are tested, analyzed, and compared in the following sections.

# Approach

## Shallow Neural Network

The first step in generating a shallow network is to organize and pre-process the data. The organization process consisted of segmenting the training/validation versus the testing data. The training/validation consisted of about 450, 10 second audio clips for each of the two classes. This was an arbitrary number of samples that was decided upon in order to get a relatively well-trained network. The training and validation were split up in a 7:3 ratio by using a built-in function in MATLAB that divides them into this ratio randomly. In order to filter out some the noise and unwanted, inaudible frequencies the audio files were converted into images via an auditory spectrogram. This is a spectral analyzation tool that outputs a visual representation of the signals frequencies as they vary with time. Neural networks have a history of high-accuracy performance with image classification so filtering out the noise and representing the audio as images was necessary for training the network for the significant features.

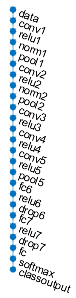
Designing the network had few requirements: minimal number of layers, input size compatible with 3-band images, and ability to extract features that relate to spatial characteristics. This was done by modeling a network designed in [source] for the purpose of optical character recognition. The only modifications needed were to deal with the different dimensionality of the data under test. The network consisted of three convolutional layers with batch normalization, ReLU, and max pooling for down-sampling and stability. At the end of the network was a fully connected layer to interpret the global features, and finally, a class output to define the class of the sample.

## Transfer Learning with AlexNet

The pre-trained AlexNet neural network was acquired from MathWorks. To completely train a network of this size and to this level of accuracy would take copious amounts of processing power and time. Luckily, the weights of the network have already been calibrated and must only be shifted to more accurately extract the relevant features from the bird call data set. This is done with the same methodology of training an un-trained network except that the weights do not have a random initiation and some of the weights, in the early layers, do not get changed at all. The first few layers extract more global, low-level features which tend to improve performance regardless of the application. The task-specific, later layers are shifted more to the bird calls high-level features not seen in the ImageNet training data set.

AlexNet is originally trained for image classification which makes it an ideal candidate for transfer learning in this study. The main caveat for AlexNet is it heavily relies on the input images to be of the size 227x227x3 to achieve optimal classification accuracy. This is not ideal, as the spectrogram images are of size 875x656x3. Fortunately, the aspect ratio of the spectrograms is 1.33:1 and scaling to 1:1 will not significantly affect performance in a negative manner. Besides augmenting the image size prior to feeding into the network, the pre-processing procedure is identical to that of the shallow neural network. In this case, the training and validation data are different because a different random seed was used to divide the two sets. This should not have a noticeable effect on the training process as the validation data is used strictly as a tool to determine when to stop the training process.

The architecture of AlexNet contains approximately twice as many as the shallow network used in the previous section. AlexNet has 5 convolutional layers and 3 fully connected layers. After each of the convolutions, the ReLU activation function is applied to produce a relative output. Dropout is utilized before the first two fully connected layers, which prevents the network from overfitting to the test data. The output layer of the network also had to edited to output two classes to define the separation requirement of the overall neural network. In comparison, the shallow neural network has a very similar architecture in terms of the convolutional techniques used to extract features, while AlexNet has a more rigorous filtering of the convolution outputs with the three fully connected layers at the end. Figure 1 compares the schematics of the two networks under consideration. Both networks follow a very linear scheme.



1. Side-by-side comparison of AlexNet (left) versus the shallow neural network (right) architectures.

## Template Matching with SAM

This method was included in analysis to benchmark the deep learning approaches from the previous section. Unprocessed audio data was supplied to the SAM detection algorithm. The single call that visually had the least amount of noise and highest magnitude in relevant frequencies was manually selected as the truth signature for each class. This was a discretionary analysis performed on the spectrograms. The data was again segmented into 10 second clips except they were left in vector form. The sampling frequency of the clips is 44.1 kHz so each of the clips had 441,000 samples. All test clips were concatenated so that the calculation could be performed in one operation via matrix operations instead of iterating through each test call.

The SAM algorithm is supervised in the sense that the ground truth is provided to be projected onto the test data. SAM utilizes the raw, unwhitened data for classification. Formulated below is the mathematical representation of the algorithm.

(1)

The ground truth signatures are represented in this equation by the variable and is the matrix of all the test calls. This equation utilizes the quotient of the product of the signatures with the product of their respective magnitudes. The known spectra of the ground truth are compared to the unknown test samples in the context of their spectral angle, hence the spectral angle mapper.

# Experimental Methodology

All algorithms need a consistent, rigorous methodology to evaluate their performance. This is especially true when performing a comparative analysis where comparison must be done according the same metric. The outputs of the neural networks are categorical, either American Robin or Mourning Dove. The SAM algorithm outputs a value between -1 and 1 so the evaluation cannot be done without a decision system.

Neural networks can be tailored to output exactly what answer is desired. In this case, the networks final output layers were edited to output the predicted categorical class. Following the predicted outputs, the true classes are known so the accuracy can be determined via a simple binary equals operation. This can then be broken down into what is known as a confusion matrix which displays the ratio of the correct predicted class versus the misclassifications of the class. This method was used to rank the performance of the two neural network approaches directly.

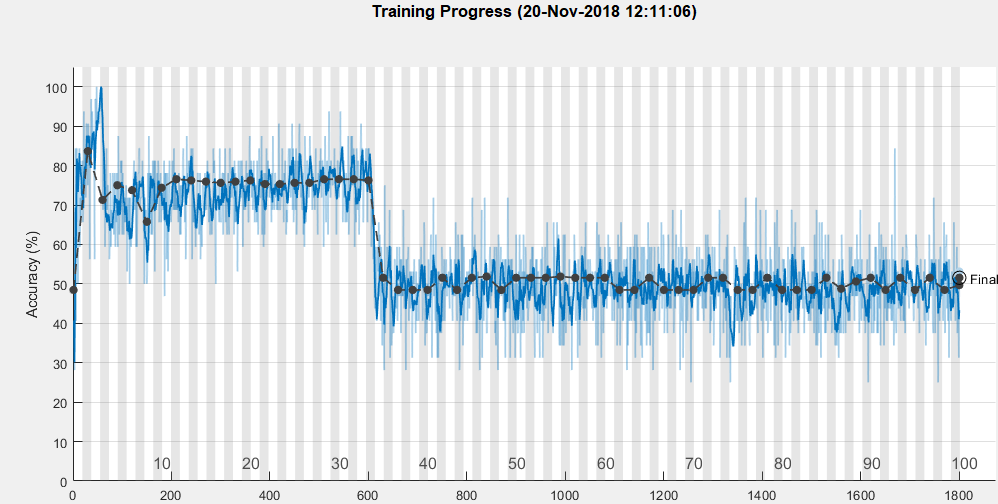
As previously stated, the SAM classification approach required a decision threshold for the normalized, relative scores produced by the algorithm. There is no defined decision threshold to achieve the most accurate results. Relatively, the best decision system can be determined experimentally which was done in this study. The first attempted method was to use the median value of the SAM score of the respective class being predicted as a threshold and if it is greater than this parameter the class is true, otherwise output is the other class. The thresholding parameter needed some tuning, so another attempt was using the mean value. The chosen decision methodology was to directly compare the SAM score of the test sample on a case by case basis and whichever vector had the higher absolute magnitude was defined as that respective class. This comparative test achieved the highest level of accuracy of the different decision systems tested.

Now that all three approaches have a consistent performance metric, a threshold for acceptable performance needed to be established. To an ear that has not heard either of these calls, the average classification would be approximately 50%. This classification percentage can simply be calculated via the overall ratio of correct classifications to the amount of test samples.

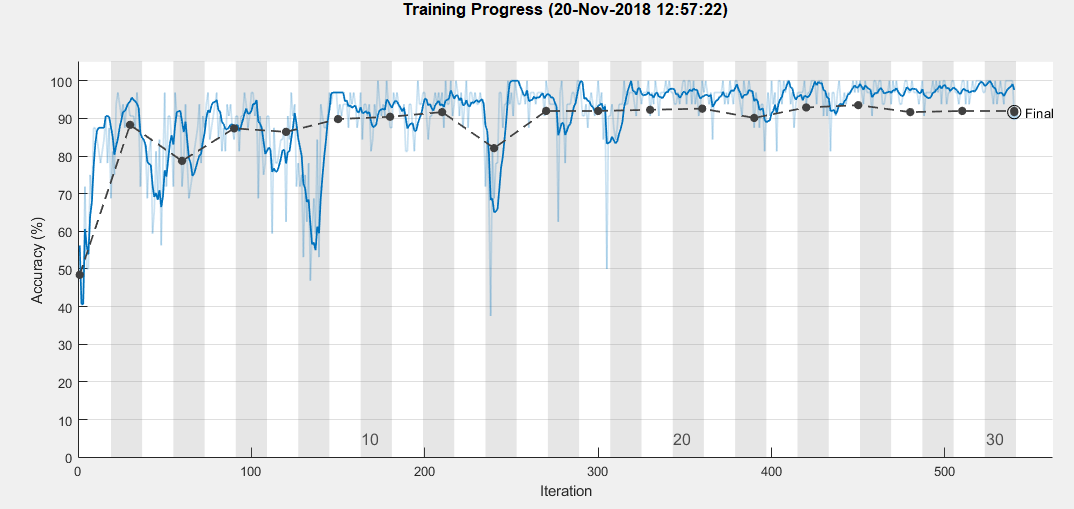
# Results

## Shallow Neural Network

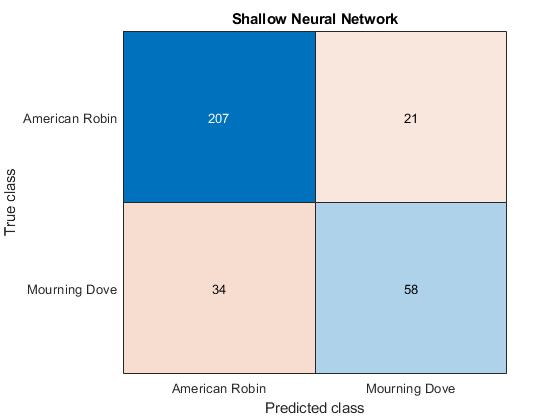
The shallow neural network had an interesting anomaly during the training process that warranted mentioning. At the 30th epoch during training the test and validation accuracy dropped approximately 30% and never returned to the steady 80% accuracy that was being output during the first 30 epochs. When accuracy stops increasing or decreases slightly usually this means that the network is overfitting to the test data. This drastic drop seems even too much to contribute to overfitting, the plot of this training anomaly can be seen in Figure 2.



1. Training progress of the shallow neural network with the 30% accuracy drop near the 30th epoch.



1. Training progress of the shallow neural network limited to 30 epochs.

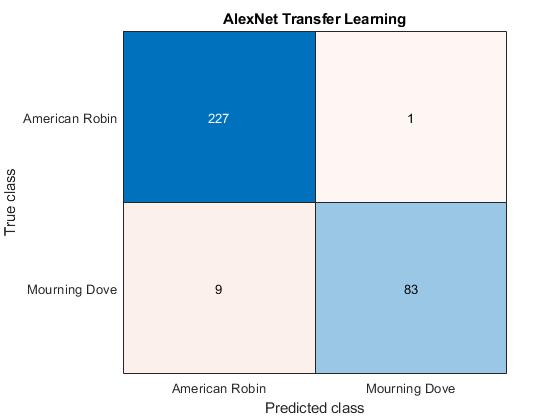


1. Confusion matrix for the shallow neural network bird classification.

The shallow neural network, after being limited to 30 epochs during the training process, as seen in Figure 3, ended with a validation accuracy of 91% while the 100 epoch training process resulted in a validation accuracy of 51%. Using the network limited to 30 epochs and the 320 test samples, the resulting accuracy was 82.8% which can be derived from the confusion matrix in Figure 4.

## Transfer Learning with AlexNet

The re-training process with AlexNet only took 6 epochs to achieve a validation accuracy of 97.8% with the 922 training/validation samples compared to the shallow network that took 30 epochs to reach a validation accuracy of 91.6%. The accuracy after feeding the 322 test spectrograms through the re-trained AlexNet resulted in a 96.9% classification accuracy which can be calculated using the confusion matrix in Figure 5.



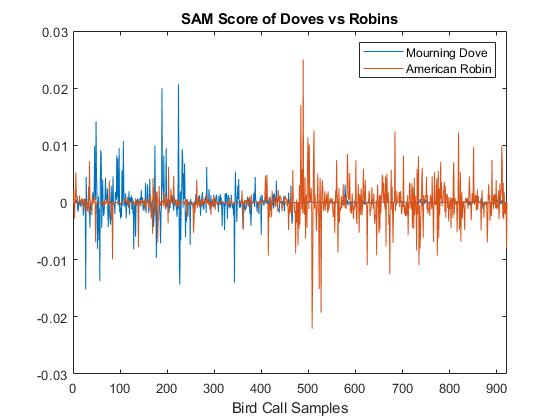
1. Confusion matrix for the AlexNet neural network bird classification.

## Template Matching with SAM

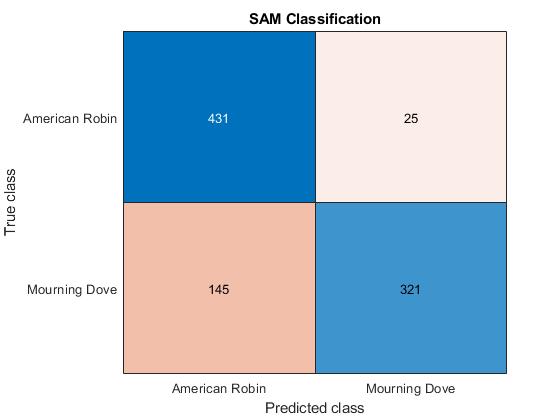
The SAM scores were computed for the 922 test samples using equation (1). They were then classified using the decision methodology discussed in Section III. After computing the SAM scores, they were plotted where a very apparent difference was visible between the two species. The first 466 samples were Mourning Dove calls and the prior 456 were American Robin calls. In Figure 7, the defined line where the scores swap in magnitude can clearly be distinguished at sample 466. Prior to plotting these values, the other metrics, class mean and median, were being utilized for classification. After discovering the maximum absolute magnitude of the true class was this separable, this was now used for the decision methodology. Putting this methodology to use yielded the confusion matrix that can be seen in Figure 8. The SAM template matching approach yielded an accuracy of approximately 81.6%. The fact that there was not any pre-processing done to the data, no iterative training process, and the predictions can be made the fastest of the three approaches, this is an impressive result.



1. Confusion matrix for the SAM template matching bird classification.



1. Plot of SAM scores for the 922 test samples.



1. Confusion matrix for the SAM template matching bird classification.

# Conclusion

Throughout the process of bird call classification, there have been valuable insights into potential solutions as well as problems in utilizing deep learning for bioacoustics classification. The results obtained during this study reveal a promising future for the application of such tools in this field of research. Prior to a few years ago, implementing deep neural networks such as AlexNet could only be performed with application specific devices with hardware acceleration but with the explosion of the field of deep learning, these processes have been optimized enough to justify their application. The main trade-off between deep learning and conventional machine learning approaches is the computational speed decrease that can be observed when implementing neural networks. This is prevalent in the comparison of the three methods time per classification in Figure 6 where the neural networks were both an order of magnitude higher than the conventional template matcher. This does come at a cost because the time per classification has a direct relationship with the accuracy of the algorithm.

The data needed for the neural networks also poses a potential problem as large volumes are needed to train them accurately. In the template matching case, only one high fidelity, low noise signal is needed to classify each specific species. For the neural networks used in this study, the audio file must first be converted into an auditory spectrogram. In this case of AlexNet, it must also be augmented to fit the input size of the network. This is most likely where the bottleneck for speed would occur in the classification process.

In this study only 2 different classes were considered, but in a practical application there would be at least an order of magnitude increase for species to classify between. Neural networks like AlexNet have proven their accuracy in classification problems that consider hundreds of classes. Scalability is where the template matching would not perform as well in. These conventional approaches tend to use singular, more general features whereas the deep learning methods extract a higher quantity of class-specific features.

##### Future Work

The first step in continuing this research would be to do the same procedure and analysis with a larger number of classes and increment this until one algorithm stands out as the most accurate and robust. Another aspect of this study that could be explored is the effect on performance for different pre-processing would have on these algorithms. After optimizing both the pre-processing and classification methods the final, and possibly most exhaustive step, would be integrating this into an application specific device. This would need to be experimented with in the study and take multiple iterations to implement. Finally, there could always be more data so collecting various and more frequent data would only improve performance.

##### Acknowledgments

I would like to give a special thanks to the creators of the organization Xeno-Canto, without them and all the wonderful contributors none of this data would be readily available. My understanding of machine learning concepts would be inferior if it wasn’t for my professor Dr. Jason Kaufman at the University of Dayton. And finally, thank you to birds around the world for producing such elegant acoustics.

##### References

1. Stowell, D. and Plumbley, M. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ, 2, p.e488.
2. Mathworks.com. (2018). Transfer Learning Using AlexNet- MATLAB & Simulink. [online] Available at: https://www.mathworks.com/help/deeplearning/examples/transfer-learning-using-alexnet.html [Accessed 9 Dec. 2018].K. Elissa, “Title of paper if known,” unpublished.
3. [3]"Create Simple Deep Learning Network for Classification- MATLAB & Simulink Example", Mathworks.com, 2018. [Online]. Available: https://www.mathworks.com/help/deeplearning/examples/create-simple-deep-learning-network-for-classification.html. [Accessed: 09- Dec- 2018].
4. J.R. Kaufman, M. Celenk, A.K. White, and A.D. Stocker,“Bobcat 2013: A hyperspectral data collection supporting the development and evaluation of spatial-spectral algorithms,” in Proc. SPIE 9088, June 2014.