Deep Learning for Bird Sound Recognition

Abstract

Precise recognition of bird species through audio relies a lot on careful attention to audio tracks from multiple species which consumes a lot of time and induces human error. To solve this I utilized data from Birdcall Competition data from 2020 originally from xeno-canto. I approached the problem in two stages, first a binary classification to distinguish between the two most frequently occurring species (highest number of samples in the dataset) and then a multi-class classification across all 12 species. Since there was a class imbalnce i downsampled to match the lowest number of samples for bird species. I also used focal loss, data augmentation and trained CNN architecture from scratch which resulted in 80% and 35% accuracies for binary and multi-class respectively.

Intro

Xeno-canto consists of wildlife sounds around the world. For this report, I focused on recognition of 12 Bird Sounds common in the Seattle area. The 12 species are American crow(amecro), American robin(amerob), Bewick's wren(bewwre), Black-chapped Chickadee(bkcchi), Dark-eyed Junco(daejun), House Finch(houfin), House Sparrow(houspa), Northern Flicker(norfli), Red-winged Blackbird(rewbla), Song sparrow(sonspa), Spotted Towhee(soptow) and White-crowned Sparrow(whcspa).

This report uses different Neural Networks architecture for binary and multi-class and shows performance based on hyperaparam tuning, dataset augmentation and use of loss function.

Neural Networks are discussed in detail in the following section.

Theoretical Background

I will explain Neural Networks working through a Example.

In Fig.1, Nodes X1, X2 and X3 are the input features or values (in this spectrogram images).

Each input is connected to every neuron in the hidden layer with a weight.

A1, A2, A3 and A4 make up the hidden layer, Each neuron receives a weighted sum of all input nodes :

$$A_j = f(W_{j1}X_1 + W_{j2}X_2 + W_{j3}X_3 + b_j)$$

Where W is weights, b is bias and f is an activation function(Ex ReLu)
Hidden layers introduce non-linearity to the network.

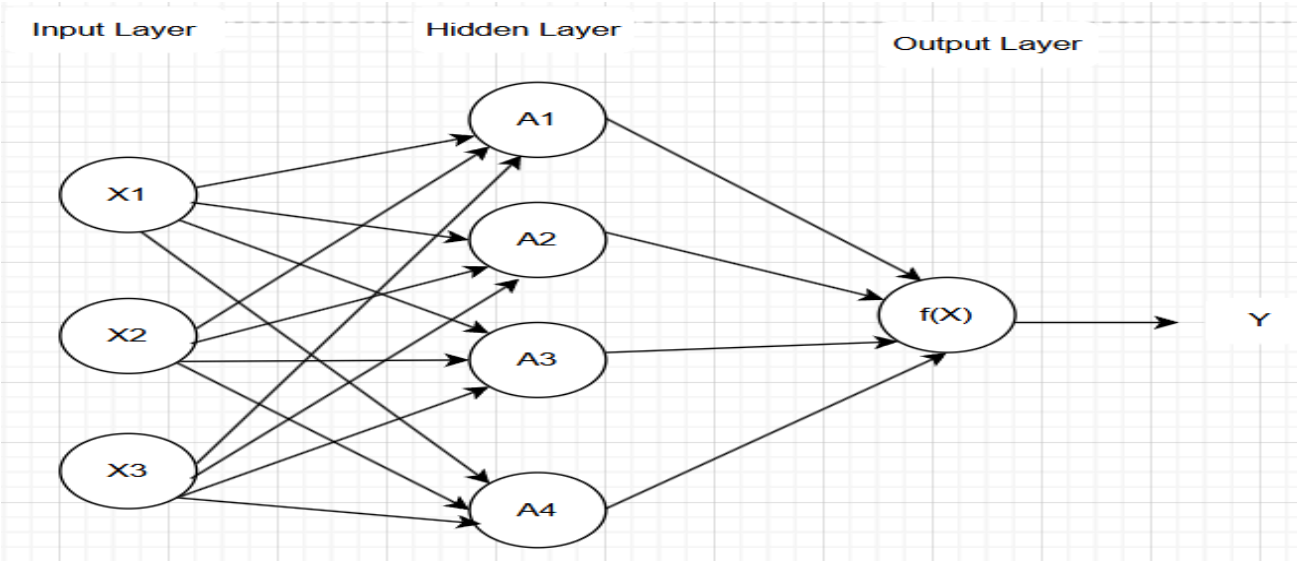

Fig.1 Graphical Example of Feed forward Neural Network

f(X) produces output Y. It takes all input from hidden nodes(A1 to A4) and computes

$$Y = f(W_1A_1 + W_2A_2 + W_3A_3 + W_3A_3 + b)$$

If there is more than 1 hidden layer it is called Multi-layer Neural Network.
One of the most popular uses was image classification and was popularized by CNN.

CNNs is made up of two main types of hidden layers:

- Convolution layers: Detects patterns by scanning images to spot details like edges or textures.
- Polling Layers: Helps simplify things by picking out the most important features.

Convolution Layer

- This layer uses multiple filters to find patterns like edges or textures in an image.
- Each filter slides over the image and looks at small patches one at a time.
- It multiplies its value with the image patch and adds them up.
- If the patch matches what the filter is looking for the result is a high value.

Polling Layer

- This layer retains the important information while shrinking the image into a smaller one
  Ex. max pooling - picks the highest value from each block.

Methodology

The Audio clips were first converted to spectrograms. The data is preprocessed the following way:

- The sound clip is subsampled to 22500Hz.
- The first 3 sec is selected.
- A spectrogram is produced for each 2-second window, resulting in a 128 (frequency) x 517 (time) "image" of the bird call.
- All bird calls for all clips in a given species are saved individually (stored in HDf5 format).

Binary Classification

For Binary, I considered house sparrow and song sparrow and labelled it as 0 and 1 respectively.
I converted data into tensors and split the dataset into 70% train, 15% validation and 15% as test.
I built a custom CNN Architecture with dropout and weighted loss to address class imbalance. I used CrossEntropyLoss and Adam optimizer with early stopping based on validation loss and evaluated based on accuracy.

MultiClass Classification

I trained the Custom Convolution Neural Network(CNN) from scratch. To ensure balanced dataset learning across all 12 classes the dataset was downsampled based on the least(no. of samples) represented species. I also applied on the fly(to save memory) augmentation using time and frequency masking to increase

variations within the dataset. I used focal loss so that model focuses more on classes harder to classify. Training was done using AdamOptimizer with early stopping based on validation loss and model performance was evaluated using top-1 and top-3 accuracy metrics.

Results

Binary Classification

House sparrow (labelled as 0), song sparrow(labelled as 1) had 630 and 263 samples respectively.

No.	Key Features	Loss Function	Accuracy	Comments
1	Basic 2-layer CNN	CrossEntropyLoss	~71%	Overfitting No dropout
2	CNN + Dropout + L2 regularization	CrossEntropyLoss	~66%	Better Regularization; Underfit
3	Dropout + Class Weighting	Weighted CrossEntropyLoss	~67%	Slight improvements; Early Stopping
4	BatchNorm + Class Weights + Adaptive Pool	Weighted CrossEntropyLoss	~80%	Best Performance + stable training

Table 1: Binary Classification Models comparison

In Attempt 4 the model achieved a test accuracy of ~80% meaning the model correctly predicted 81 out of every 100 test samples. This attempt builds on previous CNN design by introducing batch normalization after each convolution layer and replacing flattening with adaptive average pooling to reduce sensitivity to spatial dimensions.

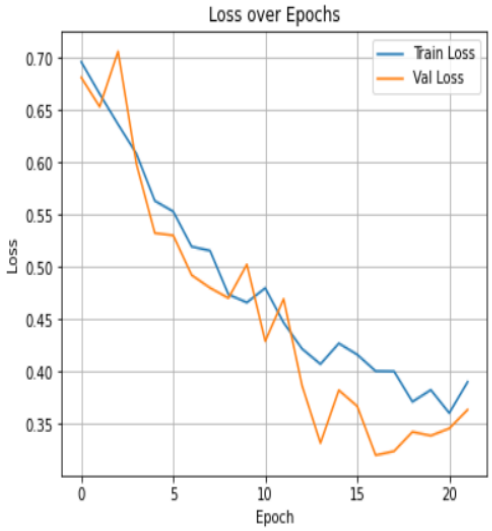

Fig. 2 Loss over Epochs (Binary)

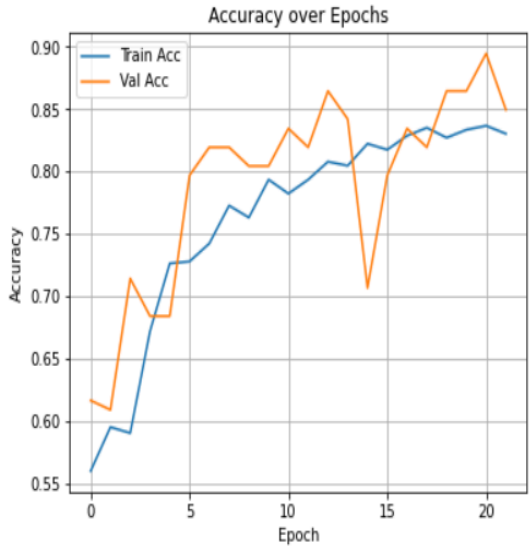

Fig. 3 Accuracy over Epochs (Binary)

Both Fig. 2 and Fig. 3 are for Attempt 4.

In Fig. 2 Both the training and validation loss decreased steadily. This shows the model is actually learning and is simply memorizing.

In Fig. 3 the model achieved training accuracy ~85% and validation accuracy close to ~89%. There is constant improvement in validation and eventually catches up to training and indicates no overfitting.

MultiClass Classification

The Number of samples for each species is shown below:

Species	No. of samples
American crow	66
Bewick's wren	144
Black-chapped Chickadee	45
Dark-eyed Junco	125

House Finch	84
House Sparrow	630
Northern Flicker	37
Red-winged Blackbird	187
Song sparrow	263
Spotted Towhee	137
White-crowned Sparrow	91
American robin	172

From the above table we can infer that there is class imbalance where Northern Flicker has least samples(37) and House Sparrow has the highest(630).

No.	Key Features	Dataset	Loss Function	Accuracy	Comments
1	Baseline CNN + Class Weights	Imbalanced	Weighted CrossEntropy	17.1%	Model biased towards frequent class
2	Depper CNN +Focal Loss + Weighted Simplifier	Imbalanced	Focal Loss	9.1%	Overfitted
3	Downsampled + Focal Loss	Balanced (Train only)	Focal Loss	18-19%	More stable but not generalized for all classes
4	Downsampled + Focal Loss + Separate Test Set	Balanced (Train + Test)	Focal Loss	35%	Best generalizations across all the models

Table 2 : Multi-Class Classification Models comparison

Across the four MultiClass classification attempts, there was an increase in performance as I adjusted class imbalance and scaled it. All models from Attempt 2 have on the fly augmentation using time and frequency masking. The baseline attempt 1 performed reasonably well but heavily favoured majority classes.

In Attempt 2, Focal Loss and data was downsampled but unfortunately led to poor performance due to model instability (early stopping).

In Attempts 3 and 4, the dataset was balanced. While attempt 3 improved training stability but failed to generalize to all classes attempt 4 combined the data with a held out test data and achieved the best overall accuracy of 35%

In both Binary and MultiClass setup early stopping was included based on validation loss.

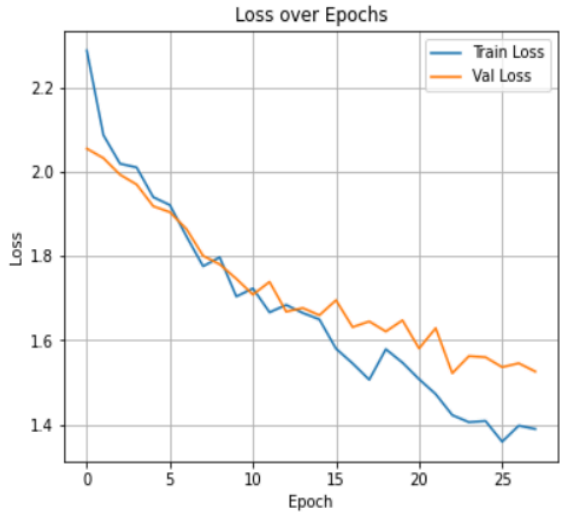

Fig. 4 Loss over Epochs (MultiClass)

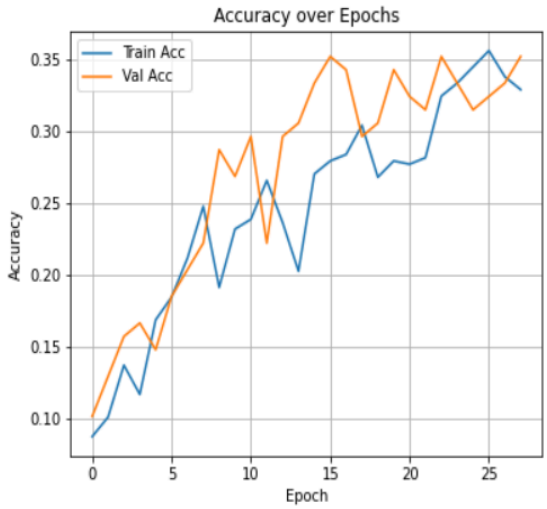

Fig. 5 Accuracy over Epochs (MultiClass)

Fig. 4 and Fig. 5 shows Loss and Accuracy over Epochs respectively.

The training and validation show steady improvements in both loss and accuracy. The validation reached its highest at 35% which is close to the training curve showing no signs of over fitting.

Discussion

One of the main limitations I encountered in this homework was Class Imbalance found in the original dataset. Several species had hundreds of samples while others had around 50 and some had less than 50 making it difficult for the model to learn or identify patterns accurately.
This was solved through downsampling and class-weighting or focal loss, but these techniques came with trade-offs - such as model instability or failure to classify species.

Another challenge I faced was the limited size of the dataset even after downsampling. With only 28 samples per class in the final version, the model lacked diversity making it even difficult to classify.
Since the pre-processed shape was already fixed there were less modifications allowed and hence it was difficult to add temporal context (ex. RNN).

Training time was reasonable overall. Binary Classification trained under 1 hour and Multi Class trained around 2 hours. Deeper or regularized models with focal loss or augmentation took more 10-15 mins. Since I used on the fly data augmentation, memory usage remained efficient, but training time increased.

From both Classification report and concussion matrix, it is clear that certain species were significantly harder to predict than others. Class 0 (House Sparrow) was consistently predicted well and achieved high precision and recall. In contrast, Class 1 (Song Sparrow), Class 9 (Spotted towhee) and class 11(White-crowned Sparrow) have very low F1 scores and sometimes close to 0.

After listening sample and reviewing spectrograms I observed that:

- These challenging classes tend to have shorter and more irregular calls.
- The frequencies overlap with common species making it harder to classify.
- Spectrograms for these classes had less consistent vertical or horizontal structure which CNNs rely on to detect patterns.

Furthermore species that were confused may be due to similar structure or unwanted background noise. For instance class 1 and class 9 may be due to overlapping frequency.

While CNNs are a natural choice due to their ability to extract patterns from images other models that could have been explored are:

- CNN+RNN : CNN captures spatial information whereas RNN captures temporal features making it well suited for Bird sound recognition.
- Pretrained audio models: Leverage transfer learning potentially improving low data.

Despite this CNN is still the best choice because:

- Sectograms resemble closely to images making it reasonable for CNN to extract spatial features
- They allow end-end training which gives us freedom to manual engineer features
- With techniques like Batch Normalization, Dropout, and focal loss they adapt really well to imbalance data.

Conclusion

The most striking finding of this report was how much class imbalance and augmentation impacted the models ability to correctly identify bird species. Species with clearer and more distinct vocal patterns were recognized easily , while other overlapping or inconsistent classes led to frequent misclassifications.

The data used for this work consisted of preprocessed, high quality clips from a curated dataset. It would be interesting to rerun these models on continuous or real-time audio, where environmental noise and overlapping species present a greater challenge.

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R (2nd ed.). Springer.
2. Combination of RNN and CNN
3. CNN
4. Neural Networks-A beginners guide
5. Layers in CNN
6. Training a CNN from scratch using data augmentation
7. Machine Learning with imbalanced data
8. Focal Loss