Final **Report** - FieldworkAnimalDetector

Jianzhong (Ken) Shi, Student #: 1005108811
Daniel Sheen, Student #: 1004923238
Yutong (Wendy) Wang, Student #: 1005075561
Word Count: 1927, Permissions at the end of the report

December 6, 2020

1 Introduction

Field Research involves data collection within a natural environment and while this is an effective method in producing insight that is unobtainable in a lab, there are numerous dangers as well [1]. One of these dangers include wildlife and in Africa, lions are one of the most aggressive predators. The safest option when dealing with lions, is to avoid them as well as the areas they preoccupy because they are highly territorial animals [2]. This requires the field researchers to be able to identify the animals beforehand, which is a difficult task to do by relying on visuals. As a result, our team created an animal sound classifier, which can classify whether a given sound is coming from a lion or a hyena, the latter being less threatening because they feed primarily off of carrion [3]. Neural networks have proven themselves to be effective in classifying animals by sound, indicated by a paper released this year that uses a convolutional neural net to classify over 20 animals, based on a given animal sound [4].

2 Background and Related Work

Most researchers approach the animal sound classification problem by training a neural network with extracted audio features from the sound files obtained either through web scraping or available dataset online. [4][5] Some of the common audio features used in neural network training are Mel Frequency Cepstral Coefficients(MFCC, excitation type, spectral variation and perceptual spectral variation. [6]) These approaches take advantages of image classification by representing the audio features as image inputs to address the audio classification problem. Some common neural network models that have been proved to complete the task are CNN, KNN, and RNN. For this project, the team has implemented a similar approach, which will train the model to analyze and classify sound files based on their MFCC, delta MFCC, and delta-delta MFCC features.

3 Illustration

The figure below demonstrates the overall structure of the project:

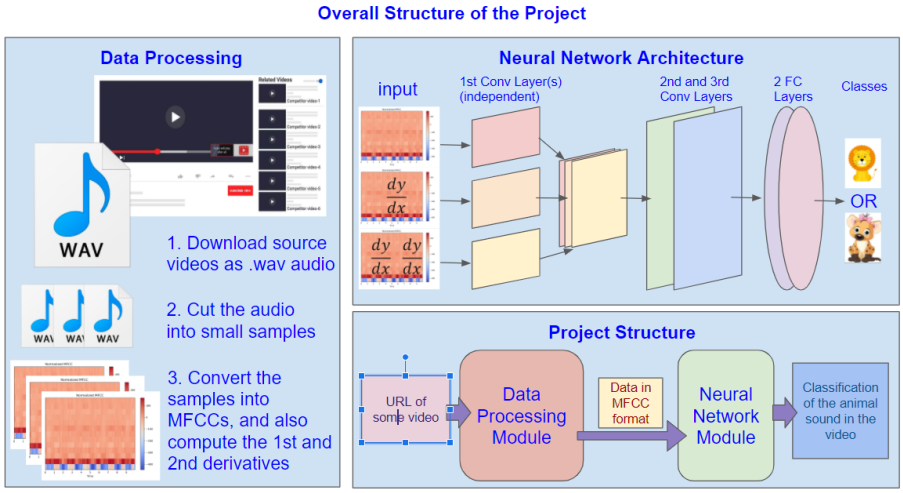

Figure 1: The Overall Structure of the Project

The project splits into two main module, which the first module takes in video URLs, and outputs the MFCC of the audio samples of the video; the second module takes the MFCC data and outputs a classification using the Neural Network structure. More details about the two modules are all in the figure, as well as the later sections.

4 Data and Data Processing

The data used for this project all came from Youtube videos. Most of the training and validation data comes from two 10 hours videos of each animal roaring [7] [8]. These two videos are combinations of sound collected from different sources, so the team was confident in the generalization of the data. The testing data comes from various sources of shorter videos online (typically one or two minutes) Although the team was aware the danger of using different sources for training and testing data, the later experiment proved the model generalized well.

The URLs of the source videos were fed into a pipeline program written by the team, which will download the videos as .wav audio files with Youtube-DL [9], and then trimmed by ffmpeg [10] at **random time stamps** for a fixed length to produce a number of samples. The samples are then sent into a conversion program (written by the team) to be converted into MFCC data, which is used

in further computation processes (including computing derivatives). A diagram below visualizes the above process.

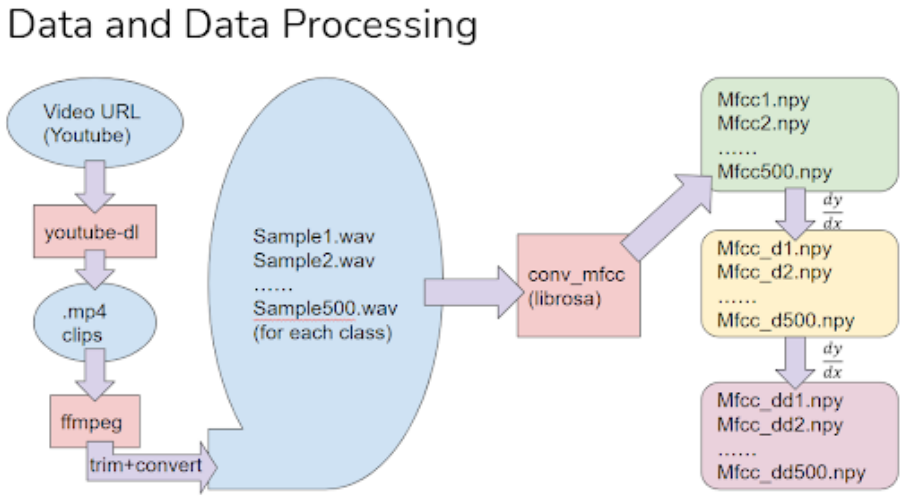

Figure 2: A demonstration of the data processing procedure.

The team attempted samples of 1.5, 3.0 and 10.0 seconds, and decided to use 10.0s samples for the project. For the training and validation data, 1000 data from each 10-hour video corresponding to each animal was trimmed. For testing data, 20 data was generated from 4 or 5 different videos for each class. An example of the sample .wav and MFCC data is shown below:

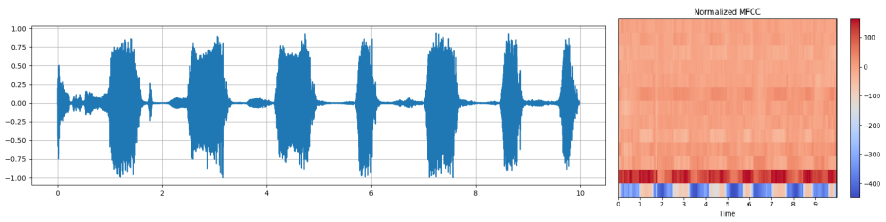

Figure 3: An example of .wav and MFCC data from a random sample.

As it turns out, the source variety of the two 10-hour videos helped the model to generalize very well. The final model was able to reach a testing accuracy of 95%. Since the amount of testing data was not as convincing, the team also wrote a pipeline program that takes the URL of any youtube video, and generates a dynamic subtitle that shows the classification result. A good demonstration is available online. [11] After applying the program to many other videos, the team is more confident about the results from the mode.

5 Neural Network Architecture

Our final model used a total of 5 convolutional layers and two fully connected layers. There were three unique inputs, with identical shapes: the mel frequency cepstrum coefficients (MFCCs) and its corresponding first and second gradient. Only the first 13 MFCCs, and corresponding gradients were considered, because majority of the audio's information is contained within these [12], which led to each input having a dimensions of 13x430.

Each input was then fed into an independent convolutional layer with three kernels of size 13×10 with a stride of 5 to convert a single input into three, one dimensional tensors with a length of 85. The tensors from all three of our inputs were then stacked on top of each other such that we're left with three tensors of shape 3×85; the first row of one of these tensors was obtained from the MFCCs, the second row was obtained from the first gradient and the third row was obtained from the 2nd gradient. This helped preserve the time-dependent relationship between the MFCCs and its gradients.

After, we applied two convolutional layers, the first having 10 kernels of size 3×10 and a stride of 3, and the second having 5 kernels of size 1×10 and a stride of 1. This resulted in 5, one-dimensional tensors of length 17, which was flattened so it could be fed into two fully connected layers with 85 and 10 hidden units respectively. This results in a single numerical output, to which sigmoid was applied to obtain a value within range of 0 (hyena) or 1 (lion).

6 Baseline Model

A naive and heuristic baseline model that we chose was to characterize lion and hyena sounds by distinct audio features and then classify the sound by ear. The main distinction that we found between lion and hyena sounds was that hyena sounds generally increased in pitch and sounded clear, similar to how a wolf howls, while lions maintained the same pitch and sounded more guttural. This baseline simulates how a normal field researcher would distinguish between the two animals in a natural environment.

7 Quantitative Results

	Loss	Accuracy
Training	0.5236	96.63%
Validation	0.5141	98.50%
Testing	0.5281	95%

Table 1: Table of Loss and Accuracy for Training, Validation and Testing dataset

A summary of the quantitative result from training, validating and testing is provided in Table 1. The team would use the loss computed to analyze how well the model is doing and use accuracy to analyze the model's performance by finding the percentage of data sample with correct prediction. The CNN model is trained with 500 sound files from each class and is tested with 20 sound files from each class in the end. After tuning the CNN model, the accuracy for all training, validating, and testing datasets reach over 95% after 15 epochs. It is observed that with batch size being 30, the model does not have significant improvement in accuracy after 15 epochs.

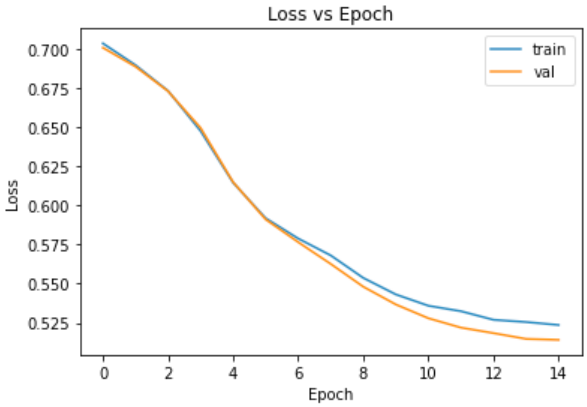

Figure 4: The graph of loss v.s. epochs for training and validation data.

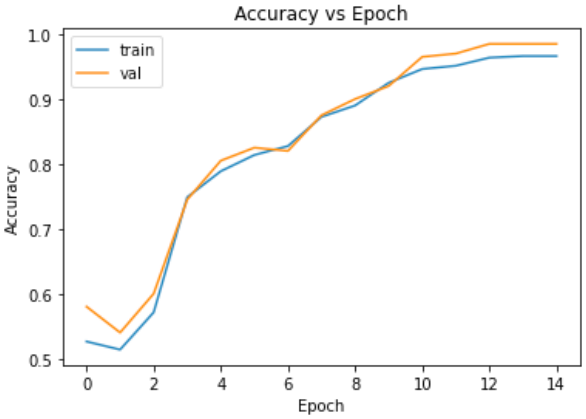

Figure 5: The graph of accuracy v.s. epochs for training and validation data.

The loss curve and accuracy curve of validation and training datasets are provided in Figure 4 and 5. From the curves, it can be observed that the CNN model is not overfitting, as performance of the model with validation

dataset is roughly the same as training dataset. The CNN model is clearly not underfitting, as classifying by random chances should give an accuracy of around 50%. From the confusion matrix obtained, it can be observed that the CNN model has successfully classified all hyena audio inputs and misclassified two lion audio inputs.

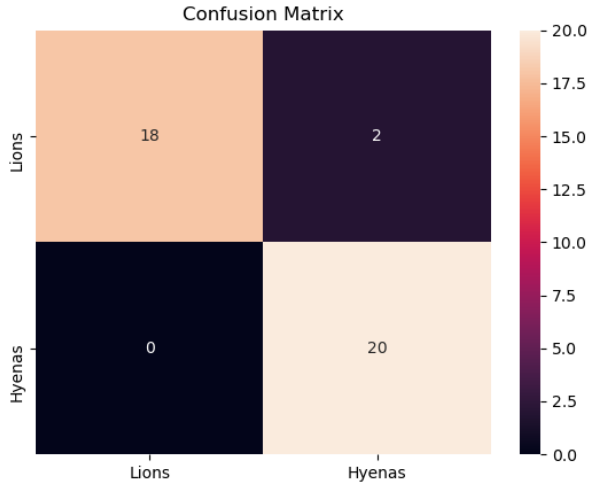

Figure 6: The confusion matrix for the testing results

8 Qualitative Results

The below are some .wav samples from the dataset:

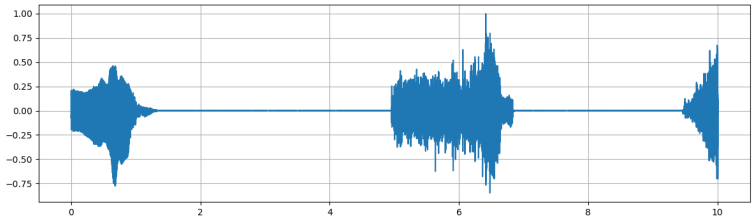

Figure 7: Sample .wav file 01

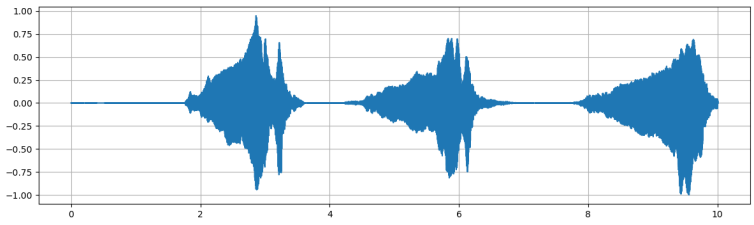

Figure 8: Sample .wav file 02

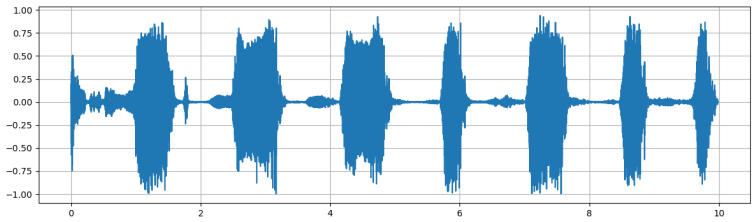

Figure 9: Sample .wav file 03

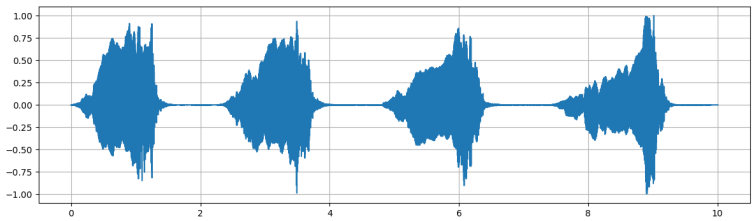

Figure 10: Sample .wav file 04

The cases which the prediction from the model failed are sample 01 and 02 compared to sample 03 and 04, which are the cases the model can successfully classify the sound of lion and hyenas. By inspection, the audio signal in sample 01 and 02 have roughly the same shape as the audio signal in sample 04. Sample 01 and 02 are cases in which the model classifies the sound as hyenas but are in fact lions' sounds whereas sample 04 is arbitrarily chosen as a representative case for hyenas' sound. It is possible that the current CNN model cannot distinguish lions and hyenas when the audio features extracted are very similar to each other.

9 Discussion and Learning

Based on the results from Table 1, the end test loss and test accuracy of 95% has demonstrated that CNN models can successfully classify lions and hyenas for most of the time. Combining the quantitative and qualitative results, it can be concluded that the CNN model might fail to classify sound files which have very similar audio features but are, in fact, in different classes. This limitation could be resolved by training our model on a larger dataset that covers all possible hyenas and lions sounds.

One important feature for the success of the model is good selection of the dataset. The dataset used for training and validating come from compositions of videos from different sources, which guarantees generalization of the model. The length of each data sample is also critical. The team has previously attempted to use 1.5s and 3s long audio files, but the model failed to learn. The audio features the team extracted from these videos are not enough for the model to obtain enough information to learn the pattern of hyena sounds and lions sounds and classify the sound signals. Another important feature for the success of the model is good selection of hyperparameters. After various trials and testing, the team noticed that having more convolutional layers(i.e. Three to four layers after the independent layers) will not increase the performance of the model, but increasing kernel size will. The kernel will then be able to search for patterns more efficiently.

In the future, the team can make little modification to the model to allow it to perform multinomial classification. The input audio files will still be obtained from YouTube and the model will be classifying animals based on the audio features(MFCC, delta MFCC, and delta-delta MFCC) extracted from the audio files. Other future improvements could be training the model to classify audio files even much more noise are presented and how to optimize the model to use less computational power and training time when given a large dataset.

10 Ethical Framework

The primary stakeholders of our project are: the field workers and the wildlife, which includes lions and hyenas. Our project only has a single use: classifying a sound as a lion or hyena so field workers can make a safe and well-informed decision. This does not respect the autonomy of field workers because if a sound is classified as coming from a lion, the field worker will be pressured to make a safe decision. On the other hand, this respects the autonomy of the wildlife because it is not intrusive or invasive towards the lions or hyenas, allowing them to continue living normally.

While our project applies principles of beneficence for field workers, it fails to do so for the wildlife. The software that our team has created acts solely in the benefit of field workers, to provide critical information such that the field workers can make safe decisions. It does not provide any benefits for the wildlife, because they do not have a method of using it.

Fortunately, the principles of non-maleficence are applied for both of our primary stakeholders. Interactions between wildlife and humans increases the chance of harm coming to both stakeholders, because humans are a foreign existence in a natural environment. Our project aims to keep fieldworkers safe and as described in the introduction, the safest option is to avoid lions, which helps minimize the chance of either stakeholder harming each other.

Finally, the principle of justice is not applied since it is primarily the field workers who are benefitting. However, there are no costs associated with either stakeholder with use of our software, making this a nonissue.

References

[1] "What is Field Research: Definition, Methods, Examples and Advantages," *QuestionPro*, 24-Sep-2018. [Online]. Available: `https://www.questionpro.com/blog/field-research/`. [Accessed: 06-Dec-2020].

[2] T. R. E. C. S. with U. T. A. Joshua T Powell, E "nvironmental Health & Safety," *Safety Guidelines for Field Researchers | Environmental Health & Safety*. [Online]. Available: `https://ehs.utexas.edu/training/field-guide.php`. [Accessed: 06-Dec-2020].

[3] "Spotted Hyena: National Geographic," *Animals*, 21-Sep-2018. [Online]. Available: `https://www.nationalgeographic.com/animals/mammals/s/spotted-hyena/`. [Accessed: 06-Dec-2020].

[4] N. Singh, "Classification of Animal Sound Using Convolutional Neural Network," *ARROW@TU Dublin*. [Online]. Available: `https://arrow.tudublin.ie/scschcomdis/203/`. [Accessed: 06-Dec-2020].

[5] E. Şaşmaz and F. B. Tek, "Animal Sound Classification Using A Convolutional Neural Network," *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, 2018, pp. 625-629, doi: 10.1109/UBMK.2018.8566449.

[6] S. Gunasekaran and K. Revathy, "Content-Based Classification and Retrieval of Wild Animal Sounds Using Feature Selection Algorithm," *2010 Second International Conference on Machine Learning and Computing,* Bangalore, 2010, pp. 272-275, doi: 10.1109/ICMLC.2010.11.

[7] "10 Hours of Lion Sounds", *Youtube*. [Online]. Available: `https://www.youtube.com/watch?v=9mizXaQYUd0`. [Accessed: 06-Dec-2020].

[8] "10 Hours of Hyena Sounds", *Youtube*. [Online]. Available: `https://www.youtube.com/watch?v=k1kVc4qsspY`. [Accessed: 06-Dec-2020].

[9] "youtube-dl," *youtube-dl*. [Online]. Available: `http://ytdl-org.github.io/youtube-dl/`. [Accessed: 06-Dec-2020].

[10] "About FFmpeg," *FFmpeg*. [Online]. Available: `https://ffmpeg.org/about.html`. [Accessed: 06-Dec-2020].

[11] Ken Shi, "Animal Classifier: Demo Edited footage", *Youtube*.`https://www.youtube.com/watch?v=nJQFQZxhpgk`. [Accessed: 06-Dec-2020]

[12] "Cepstral Coefficient," *Cepstral Coefficient - an overview | ScienceDirect Topics*. [Online]. Available: `https://www.sciencedirect.com/topics/computer-science/cepstral-coefficient`. [Accessed: 06-Dec-2020].

11 Permissions to publish

I, Jianzhong (Ken) Shi, grant permissions to the course instructor to post videos, final report and source code to the course website.

I, Daniel Sheen, grant permissions to the course instructor to post videos, final report and source code to the course website.

I, Yutong (Wendy) Wang, grant permissions to the course instructor to post videos, final report and source code to the course website.