**Cornell Birdcall Identification**

**W251 Deep Learning in the Cloud and at the Edge**

**Summer 2020**

Flying Toasters – Abhinav Sharma, James Byrne, Carlos Capellan, Isaac Law

1. **Introduction**

Cornell Birdcall identification is a Kaggle competition. Birds can be found in nearly every environment, birds species in an environment is an excellent indicators of deteriorating habitat quality and environmental pollution. Main goal of the project is to identify a wide variety of bird vocalizations in soundscape recordings. And deploy the model in Nvidia Jetson TX2. This will help researchers better understand changes in habitat quality, levels of pollution, and the effectiveness of restoration efforts.

1. **Data**

The dataset is a labelled dataset consists of birdcall audio and the corresponding bird species. This dataset was collected by Cornell Lab of Ornithology’s Centre for Conservation Bioacoustics. There are 260 bird species in the data and 21,375 audio files.

We scaled down the dataset to 20 bird species for the project, below is the distribution of labels.

A close up of graphics

Description automatically generated

Our Dataset is a relatively balanced dataset, there are roughly 80 to 100 audio clips per each bird.

1. **Tools**
2. **IBM Cloud Object Storage**

We have all our 22GB audio data saved in IBM Cloud Object storage and this allows each team member to retrieve data for processing and training via s3fs-fuse.

1. **IBM Cloud Virtual Machine**

We provisioned VM with P100 GPU for model training. Our dataset scaled up by more than 4 times after data augmentation, model training would not be possible with the use of P100 GPU.

1. **Jetson TX2**

Our trained Tensorflow model is converted to TensorRT and deployed in Jetson TX2 for inference.

1. **Data Preparation and Augmentation**

The original dataset consists of raw recordings in different environment. There are lots of different noise in the raw recordings, e.g. airplane overflights, other type of birdcalls, human talking. Considering birdcalls are all with higher frequencies, we applied a high pass filter at 1400 Hz as mean of noise filtering.

***\*\*\*(For James to elaborate on Augmentation)\*\*\****

1. **Exploratory Analysis**

We first look at length distribution of audio.

A picture containing large, bus

Description automatically generated

Most of audio are with length 0 to 25 seconds. And by randomly listen to audio clips from different birds, we observed that the target birdcalls usually appear in first 10 seconds of audio. Therefore we focus on building models for classifying bird species base on 10 seconds of audio. At the same time, a 10 seconds audio interval is also a reasonable choice from inference standpoint.

Next, we explore the time series plot of 10 second audio of random sample for each bird.

A screenshot of a cell phone

Description automatically generated

Time series plot show the fluctuation of amplitude over time, we can see that there quite a lot information inside first 10 second of audio, this reinforces our choice of 10 second audio length for training.

However, there are 2 problems with this time series plot.

1. Time series plot only show fluctuation of amplitude over time, it does not give us any information about frequency spectrum of the birdcall.
2. Time series plot is a sparse representation of an audio. Assuming a common sampling rate of 22,050, a 10 second audio is a 1-Dimensional vector with length 220,500. It would be difficult for machine to learn and extract useful information.

To solve this problem, we use another common transformation technique in audio analysis, the Mel Spectrogram transformation. First we perform fast Fourier transform (FFT) on the audio data, this step transform audio into frequency representation. Next we stack these frequency representation over time (in Mel scale), and using colour scale to represent amplitude. Below is the Mel Spectrogram of our sample audio from 20 birds.

A picture containing kitchen, room

Description automatically generated

We can this representation of data is more meaningful than time series plot. For example, for bird Amered, it has a generally high pitch voice, and it always end its call with a lower frequency voice. And for bird Vigswa, it’s voice generally show a wide range of frequency.

By converting audio data into Mel Spectrogram, it solves the 2 problems we had before,

1. Mel Spectrogram represent the change of frequency (in Mel scale) over time, each bird has difference change in frequency in their call, thus by observing pattern in Mel Spectrogram, we should be able to classify birds.
2. Mel Spectrogram compress information from 1-D vector in 220,500 dimensional space into a 2-D vector of 224 pixels by 625 pixels. With a 2-D representation, we can perform the usual image classification technique, e.g. 2D convolutional network for training.
3. **Modelling and Result**

After transforming audio data into Mel Spectrogram, we perform image classification on these images. Our approach here is, we started with pre-trained efficientNetB3 as feature vector extractor, then we iterate through different pre-trained convolutional networks to see which network works the best. Then we use all our augmented data for final model training.

***\*\*\* (James and Abhi to talk about model selection and improvement from data augmentation) \*\*\****

1. **Error Analysis**

***\*\*\* (I am thinking to plot confusion matrix and perform some error analysis once we have final model) \*\*\****

1. **Edge Inference**

***\*\*\* (Carlos to talk about edge inference) \*\*\****

1. **Conclusion & Future Work**