4,500 Seconds: Small Data Training Approaches for Deep Drone Audio Classification

Andrew P. Berg[1], Dr. Qian Zhang[2], Dr. Mia Y. Wang[3]
School of Engineering, Computing, and Mathematics
College of Charleston
berga2@g.cofc.edu[1], zhangq@cofc.edu[2], wangy5@cofc.edu[3]

ABSTRACT

The usage of unmanned aerial vehicles (UAV) in all sectors has grown exponentially, with this is an urgent need for robust classification of UAVs to protect against global incidents. Using an extremely small 9-class dataset. This study implements and tests a parameter efficient fine-tuning process for pre-trained audio transformers. We compare our results to previous work using Convolutional neural networks (CNNs). Our results show that while encouraging initial runs, the CNN still outperforms the transformer approach by 1-2%, while still being computationally more efficient. These early findings show that there is more to explore with transformers to yield better results. Future works aims to scale the dataset and best understand the trade offs between the two approaches.

METHODOLOGY

In this study we used a 9-class UAV audio dataset consisting of one-hundred 5 second samples. Due to these data limitations we opted to use the pre-trained Audio Spectrogram Transformer (AST). We developed a data pipeline to normalize and augment our data; then rapidly fit our CNN and AST models. We used mel-spectrogram for feature extraction. To improve our training methods we hyper parameter tuned each model respectively. Specifically for the AST model we used additive adapters to further improve the fine-tuning process. Adapters are a type of Parameter efficient fine tuning (PEFT), that decrease the computational overhead and in some cases improve the performance of fine-tuned models on small datasets.

EXPERIMENTS & RESULTS

Because we evaluated the models for classification performance, we focused on the test accuracies and F1 scores.
We started by robustly tuning and tracking the optimal configurations for data augmentation and CNN and AST model fitting respectively.

We ran hyper parameter experiments respective to each model. Starting with AST model specifically, we focused on finding the most optimal adapter for our problem. We found that by far, the most consistent adapter was the ia3 adapter. Notably it added very few additional parameters to the pre-trained model.

Further, we experimented with the best augmentations for our dataset. We found that generally having more than one augmentation per sample improved performance, but only for the AST model; it had no consistent effect for the CNN model

Once we were satisfied with the hyper parameter tuning, we ran 5-fold cross validation runs with the best hyper parameter configurations. We found that the CNN generally outperformed the AST model while taking less time. Please refer to the figures section for precise metrics.

CONCLUSION

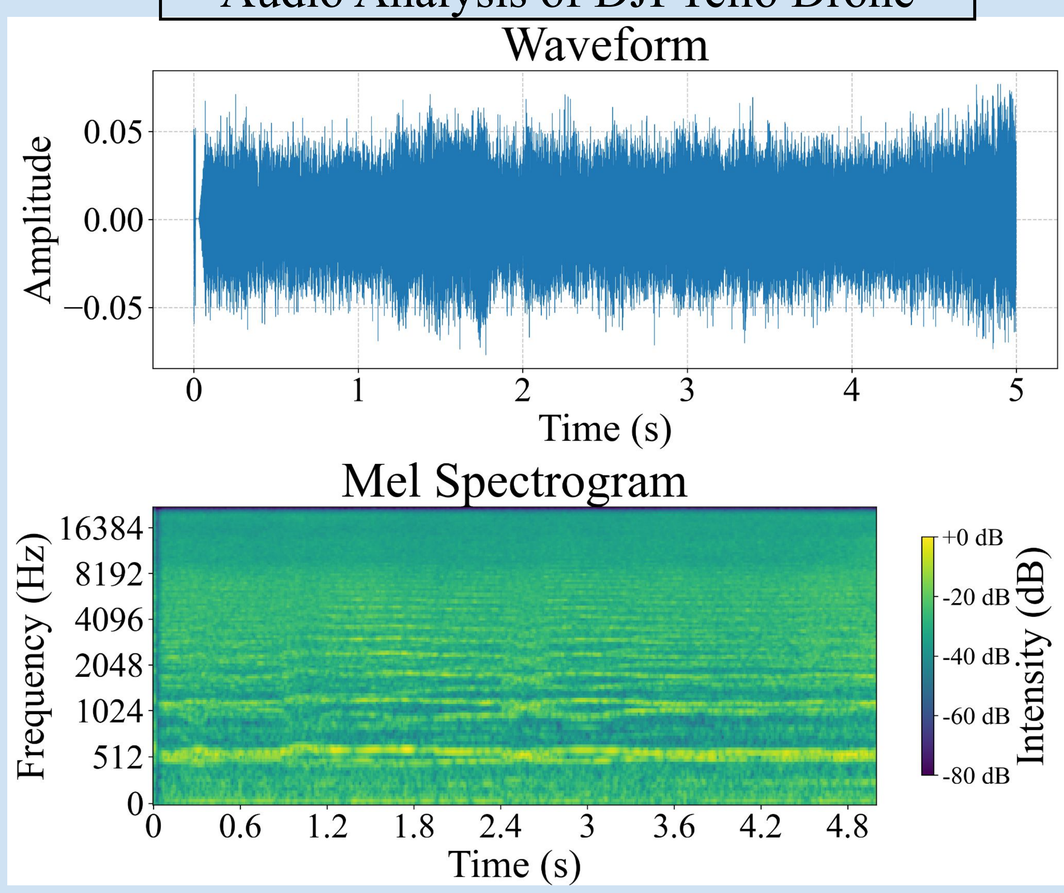
Given the results the CNN model clearly outperforms the transformer approach. However, it is not fully understood the implications of this scale of data using deep learning approaches. Future work will explore scaling the audioset, as well as experimenting on more types of CNN and transformer models.

FIGURES

5-Fold Cross Validation Results

Metrics	CNN	AST
Best Accuracy, No Augs (%)	95.43	96.31
Best Accuracy, w/ Augs (%)	**97.53**	95.71
Best F1, No Augs (%)	95.48	96.21
Best F1, w/ Augs (%)	97.53	95.54
Worst Time (mm:ss)	06:52	52:45
Best Time (mm:ss)	**01:30**	06:19
Mean Time (mm:ss)	04:06	26:50

Audio Analysis of DJI Tello Drone



February 2025