Report Date: 07/29/2022
To: ematson@purdue.edu, ahsmith@purdue.edu, lhiday@purdue.edu, and lee3450@purdue.edu
From: What is today's lunch?

- Ilmun Ku(mun90505@hufs.ac.kr)
- Seungyeon Roh(shtmddus99@konkuk.ac.kr)
- Gyeongyeong Kim(kky57389@sunmoon.ac.kr)
- Charles Taylor(taylo869@purdue.edu)

Summary

Tests with preprocessed datasets for Deep Learning models were conducted. The techniques of Deep Learning we utilized were Convolutional Neural Networks(CNN), Recurrent Neural Networks(RNN), and Convolutional Recurrent Neural Networks(CRNN). Data preprocessing methods that were employed are feature extraction and data augmentation. Also, the draft of the paper is written.

What 'What is today's lunch?' completed this week:

- **Experiment results of Deep Learning (DL) models**

Feature	CNN	RNN	CRNN
MFCCs	0.9493	0.8113	0.9174
Mel	0.9133	0.7247	0.9269
Chroma	0.7883	0.5743	0.8182
Tonnetz	0.7814	0.7162	0.8001
Contrast	0.5645	0.4256	0.7675

Table 1. Test result for Deep Learning models

  - Preprocessed datasets are fed into the DL models. The results for three kinds of DL models are in Table 1.

- **Paper draft**

  - Methodology
    - Dataset

      DJI Phantom 4 and an EVO 2 Pro were utilized to gather audio data. 1306 samples and 1229 samples were collected for each UAV, 2535 samples altogether, and 7.04 hours long in total, as shown in Table \ref{tab: data per label}. Moreover, 232 noise samples were collected in aggregate and 0.64 hours long. each audio recording is 10 second long.\\

      For each UAV, audio samples were recorded in three classes; UAV with two payloads, UAV with one payload, and unloaded UAV. The two payloads which are utilized in the study weighed 63g and 68g each. Only a payload of 68g was used when recording UAV with one payload class. 558 and 549 samples were collected for one payload class and two payload class, each with DJI Phantom 4. Also, 507 and 521 samples were gathered for each loaded class with EVO 2 Pro. Since the weight difference between one payload class and two payloads class is only 63g,

more data were collected for these classes than the other to classify those two classes.

	unloaded	1 payload	2 payloads	noise
DJI Phantom 4	199	558	549	
EVO 2 Pro	201	507	521	232
total	400	1065	1070	232

Table 2. Data per label

- Feature extraction

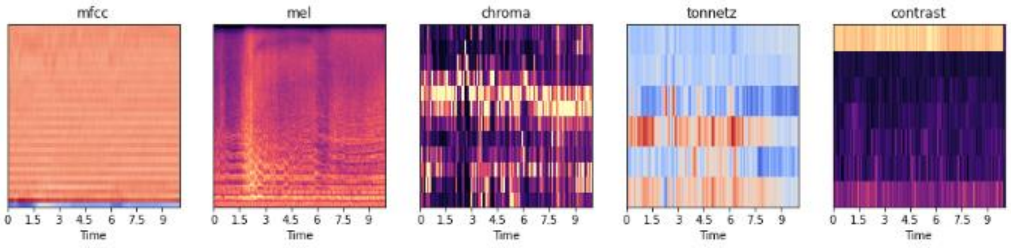

Fig. 1. Extracted Features

Features are values that represent the unique characteristics of sound that could be extracted from audio. The human perception system easily recognizes different audio. However, since the model does not have a human perception system, feature extractions are required so that model can understand the sound. In this study, the python library Librosa is employed to conduct feature extraction, providing the extraction method in the Table 2.

Mel-frequency cepstral coefficients (MFCCs), Mel, chroma, tonnetz, and contrast are the features that are used to classify the dataset. Mel-frequency cepstral coefficients (MFCCs) are values through cepstral analysis from the Mel spectrogram, which reflect the relationship between physical frequencies and human perceive frequencies. Therefore, MFCCs are commonly used for analyzing audio data since MFCCs are similar to human perception systems that are not linear but sensitive to low frequency, as mentioned [14][15].

Mel Spectrogram is the conversion of frequency to mel scale. Chroma means 12 pitches, C, C#, D, D#, E, F, F#, G, G#, A, A#, B. Tonnetz is the feature that represents the harmonic relationship of the sound. Contrast is a frequency power that can be measured at each timestamp.

To clearly distinguish the classes, It is better to preprocess the dataset with feature extraction than directly using the raw dataset. Additionally, it can be easily distinguished in spectrograms, which are visualization of features. The features extracted by librosa are saved as json files and entered as input to the model.

- Data augmentation

A large audio dataset is required to utilize DL models to solve audio classification problems. However, collecting enough audio datasets for DL rather than text or image is difficult. If datasets are insufficient, overfitting and poor generalization may occur in the classification process. Data augmentation technology is one of the solutions to cover the lack of a dataset problem. This method can obtain a new dataset by augmenting the original dataset. Therefore, Data augmentation methods

are pivotal for the smooth and continuous improvement of audio classification performance while utilizing these DL algorithms. This study uses time-stretching, pitch scaling, time masking, and frequency masking to augment audio data. Time stretching and Pitch scaling are raw audio augmentation technology, whereas time masking and frequency masking is spectrogram augmentation technology that treats augmentation as a visual rather than an audio problem. Raw augmentation is a method of augmenting the audio file itself. It is proven helpful in enhancing accuracy for LSTM-based RNN using raw audio data [16]. There are time-shifting, time stretching, pitch scaling, noise addition, impulse response addition, low/high/pass-band filters, and so forth. Time stretching changes the speed without changing the pitch, which can change the speed of the sound, slow or fast, not impacting the frequency. As opposed to time stretching, pitch scaling changes the pitch without changing the speed. For example, C major is changed to D major if a signal is up to 2. It is recognized to be advantageous when exercised in advancing CNN accuracy [17]. Spectrogram augmentation is data augmentation with a spectrum rather than a raw audio file, including time masking and frequency masking. Time masking masks a particular part of the spectrum, and those are masked with 0 or minimum value. It cuts off a portion of the time domain, which is the x-axis of the spectrum. Frequency masking is an inverse version of time masking, which cuts off a part of the frequency domain of a spectrogram by masking a specific part with 0 or minimum value. They are widely known to improve the network performance without the extra arrangement for the network, or hyperparameter [18]. They help the network robust against deformation and prevent overfitting by presenting corrupted data.

- Deep Learning Models

  CNN has demonstrated itself very effective not only for image classification but also has been shown to produce promising results for audio classification [19]. CNN can extract the local information from audio signal representation such as MFCCs or Mel spectrogram.

  RNN is proper to process sequential data such as text or sound data. However, RNN has its downside: gradient exploding and gradient vanishing [20]. Long Short Term Memory (LSTM) networks were applied to improve this issue. In LSTM layers, self-recurrent weights make the cell in the memory block retain previous information [21] [22]. CRNN is one of the neural network architectures that combines CNN and RNN. CRNN imposes upon local information and the longer temporal context. The local information is extracted with the help of the CNN, and the longer temporal context is captured by the RNN [23]. In acoustic classification, CRNN has proven noticeable results over the years [5].

- Conclusion and future work

  In this paper, UAV payloads classification was applied with three deep learning techniques; CNN, RNN, CRNN. Deep learning models assorted the audio dataset into four classes; noise, unloaded, one payload, and two payloads. The models not only distinguish whether a UAV is loaded or not, but categorize even the number of payloads. Additional equipment was not demanded when collecting the dataset since the MacBook pro's built-in microphone was operated to record the sounds. Through

this method, cost-effective and straightforward UAV payload detection can be conducted. Dataset was collected with two UAVs, DJI Phantom 4 and Evo 2 pro, and two payloads, dummy explosives. Collected datasets are 7.13 hours long in total, consisting of 2567 samples. Feature extraction and Data augmentation methods were conducted in preprocessing stage. The results of preprocessing were put into the DL models; CNN, RNN, and CRNN. In the CNN, MFCC produced the best performance with 94.93% accuracy. Also, Mel presented an excellent result of 92.69% accuracy in CRNN. MFCC has reasonable accuracy of 81.13% with RNN. The limitation of this study is that only two UAVs were used to collect audio samples. Furthermore, one kind of payload was exploited. Additional studies are expected to collect different kinds of data with manifold UAVs and payloads to generalize the UAV payload detection system

Things to do by next week

- Preparing final presentation and finishing the paper.

Problems or challenges:

- Out of memory when data is uploaded on RAM to feed DL models.

References

[1] M. al Kibsi, “Houthi drone targets senior yemeni officers, kills five soldiers,” Available at https://www.aljazeera.com/news/2019/1/10/houthi-drone-targets-senior-yemeni-officers-kills-five=soldiers (2022/05/18).
[2] W. Ripley, “Drone with radioactive material found on japanese prime minister’s roof,” Available at https://www.cnn.com/2015/04/22/asia/japan-prime-minister-rooftop-drone/index.html (2022/05/18).
[3] E. Kelly, “Venezuela drone attack: Here's what happened with Nicolas Maduro,” Available at https://www.usatoday.com/story/news/politics/2018/08/06/venezuela-drone-attack-nicolas-maduro-assassination-attempt-what-happened/913096002 (2022/05/18).
[4] Y. Wang, F. E. Fagiani, K. E. Ho, and E. T. Matson, “A feature engineering focused system for acoustic uav payload detection,” the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022), pp. 470–475, 2022.
[5] S. Wei, S. Zou, F. Liao et al., “A comparison on data augmentation methods based on deep learning for audio classification,” 2019 2nd International Conference on Computer Information Science and Artificial Intelligence(CISAI 2019), pp. 1–8, 2020.
[6] S. Al-Emadi, A. Al-Ali, A. Mohammad, and A. Al-Ali, “Audio based drone detection and identification using deep learning,” pp. 459–464, 2019.
[7] S. Jamil, M. Rahman, A. Ullah, S. Badnava, M. Forsat, S. S. Mirjavadi et al., “Malicious uav detection using integrated audio and visual features for public safety applications,” Sensors, vol. 20, no. 14, p. 3923, 2020.
[8] Y. Wang, F. E. Fagian, K. E. Ho, and E. T. Matson, “A feature engineering focused system for acoustic uav detection,” pp. 125–130, 2021.
[9] L. Shi, I. Ahmad, Y. He, and K. Chang, “Hidden markov model based drone sound recognition using mfcc technique in practical noisy environments,” Journal of Communications and Networks, vol. 20, no. 5, pp. 509–518, 2018.
[10] V. Kartashov, V. Oleynikov, I. Koryttsev, S. Sheiko, O. Zubkov, S. Babkin, and I. Selieznov, “Use

of acoustic signature for detection, recognition and direction finding of small unmanned aerial vehicles," pp. 1–4, 2020.
[11] U. Seidaliyeva, M. Alduraibi, L. Ilipbayeva, and A. Almagambetov, "Detection of loaded and unloaded uav using deep neural network," pp. 490–494, 2020.