機械学習を用いた環境音分類に関する研究

The Research on environmental sound classification using machine learning

電子・機械技術部 ロボット・制御科　清野若菜

本研究では、ドローンの動作音を対象として、その機種を分類する機械学習モデルを構築した。畳み込みニューラルネットワークを用いた機械学習モデルでは、入力として時系列信号をスペクトログラムに変換し、画像分類した場合に正解率が高かった。また、モデルを軽量化し、データ数が少量であっても十分な性能を有することがわかった。

屋外環境音を混合したデータで推論を行った場合の正解率は、スペクトログラムを入力とした場合で 99.4[%]であり、ノイズのないデータで学習したモデルが、一定のノイズ環境下でも使用できる可能性があることがわかった。

Key words: 機械学習、多クラス分類、環境音、周波数解析

１．緒言

機械学習による画像認識技術の社会実装が拡大する中、機械学習を用いた環境音分析に関する研究開発も注目されてきている。そこで本研究では、ドローン動作音から機種を分類する課題を対象に、環境中から特定の音を認識し、分類する機械学習モデルを構築することで、音を用いた機械学習のノウハウを習得することを目的とする。

２．実験

２．１．ドローン動作音の収録

２．１．１．測定対象物

本研究では、測定対象をドローンの動作音とした。録音に用いたドローンを表１に示す。機種 No.0～2 は重量が 100[g]以下のトイドローン、No.3～5 は重量が 200[g]～320[g]の小型ドローンである。

２．１．２．実験装置及び実験環境

ドローン動作音の録音には、リニア PCM レコーダー（TASCAM DR-40X）を用いた。録音条件はサンプリング周波数 48[kHz]、量子化ビット数 24[bit]、チャンネル数 2[ch]とした。録音環境は半無響室とし、レコーダーの周囲でドローンを飛行させた状態の動作音を約３分間録音した。ただし、No.5 については無響室内での安定的な飛行が困難であったため、床面上でプロペラを回転させた際の動作音を録音した。

２．２．機械学習用データセットの作成

学習に用いたデータセットを表２に示す。モデルの性能評価を行う際のパラメータとして、学習用データの種類、信号長、データ数を設定した。

機械学習モデルに入力するデータは、時系列信号及びスペクトログラムの２種類とした。時系列信号は、ドローン動作音の録音信号を 0.25[s]、0.5[s]、1.0[s]間隔にそれぞれ分割し、48[kHz]16[bit]モノラルの wav ファイルとして出力した波形データである。スペクトログラムは、時系列信号を短時間フーリエ変換（STFT）し、横軸が時間、縦軸が周波数、強度を RGB 値で表した画像データである[1)]。データ数は、ドローン６機種分のデータの総数である。このうち、８割を学習用、２割を評価用としてそれぞれデータセットを作成した。なお、信号の解析及び処理は Python を用いて行い、環境はブラウザ上で Python プログラムを実行可能な Google Colaboratory を用いた。

表１　測定対象のドローン

ﾄﾞﾛｰﾝ No.	メーカー	機種名	サイズ (L×W×H)[mm]	重量 [g]
0	DJI	Tello	98×92.5×41	87
1	GFORCE	LEGGERO	102×136×36	60
2	HOLY STONE	HS170	133×133×34	42
3	DJI	Mavic Mini	245×289×55	199
4	Parrot	ANAFI	175×240×65	320
5	HOLY STONE	HS120	270×270×120	198

表２　データセットの一覧

ﾃﾞｰﾀｾｯﾄ	学習用データの種類	信号長 [s]	データ数
A	時系列信号	0.25	5037
B	時系列信号	0.25	959
C	時系列信号	0.5	961
D	時系列信号	1.0	958
E	スペクトログラム	1.0	958

２．３．機械学習モデル

本研究では、機械学習モデルとして畳み込みニュー

事業名「 「そだてる研究室」事業 」

ラルネットワーク（Convolutional Neural Network、以下 CNN）を用いた。ニューラルネットワークの構築及び学習には、ブラウザ環境又は Windows 環境で利用可能な SONY Neural Network Console[2)] を用いた。

構築した機械学習モデルを表３に示す。モデルは、サンプル数×1ch の時系列信号を入力とする波形分類モデルと、高さ 248×幅 387×3ch(RGB)のスペクトログラムを入力とする画像分類モデルの２種類とした。波形分類モデルは畳み込み層を１０層及び５層、画像分類モデルは５層とし、出力として各機種の確率を出力する多クラス分類モデルとした。学習の繰り返し回数を示すエポック数は波形分類モデルが 100、画像分類モデルは 20 とした。該当する学習用データセットを入力としてそれぞれモデルを学習し、評価用データセットを用いて学習終了時の正解率を算出した。

２．４．ノイズ混合音を用いた推論

屋外環境でのドローン動作音の分類を想定し、無響室で録音したドローン動作音に屋外環境音を合成した信号を入力として、ノイズのない学習モデルを用いて推論を行った。合成信号は、ドローン動作音及び屋外環境音のパワーを時間軸上すべてに対して正規化し、ＳＮ比を 0 として作成した[3)]。波形分類モデルの入力信号は信号長 0.25[s]の時系列信号、画像分類モデルの入力信号は信号長 1[s]のスペクトログラムとした。推論用のデータは、各機種 60[s]の合成音を信号長間隔に切り出したデータからランダムに 168 個抽出した。

３．結果及び考察

３．１．収録音の周波数解析

収録したドローン動作音のスペクトログラムを図１に示す。No.0_Tello は、広い周波数帯に一様なスペクトルがみられるが、その他５機種についてはそれぞれ特定のスペクトルにピークがみられた。これらのピークは周期的であることから、プロペラの回転数に起因するものであると考えられる。

３．２．学習結果及び性能評価

機械学習モデルの学習結果を図２に示す。各学習モデルの正解率はいずれも 90[%]以上であった。項目別にみると、時系列信号の信号長は 0.25[s]が最も正解率が高かったものの、信号の長さによる有意差はみられなかった。データ数では、多い方が正解率は高かったが、データ総数が 959、各機種平均で 160 個のデータでも正解率が 90[%]を超える結果となった。入力データでは、スペクトログラムを入力とした画像分類モデルの正解率が 100[%]であった。ニューラルネットワークの畳み込み層の数は、１０層から半減させても正解率の大きな減少はみられなかったことから、高性能な GPU を搭載していないコンピュータでも十分な性能を発揮する可能性があることがわかった。

表３　機械学習モデル

モデル	波形分類モデル (CNN)		画像分類モデル (CNN)
入力層	時系列信号 (サンプル数, 1)		スペクトログラム (3, 248, 387)
畳み込み層	10 層	5 層	5 層
エポック数	100		20
データセット	A, B, C, D	B	E

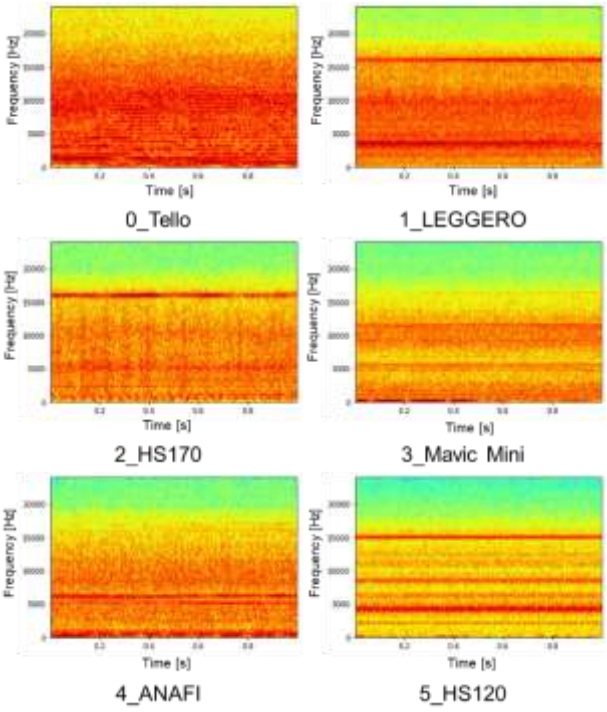

図１　ドローン動作音のスペクトログラム

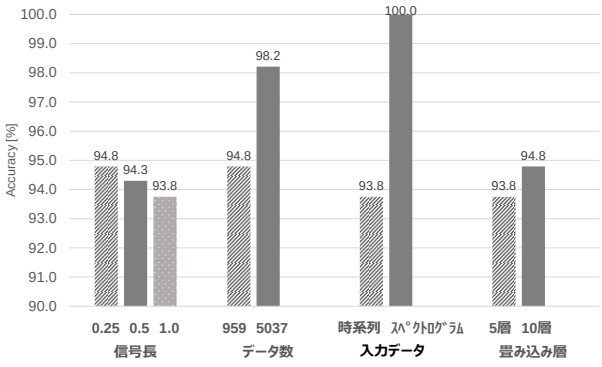

図２　学習モデル別の正解率

３．３．ノイズ混合音の推論結果

ノイズ混合音の推論結果を表４に示す。６機種全体の正解率は、時系列信号を入力とした波形分類モデルでは 76.2[%]、スペクトログラムを入力とした画像分類モデルでは 99.4[%]であった。ノイズ混合音の機種別適合率を図３に示す。波形分類モデルでは、No.1 と No.5 が 90[%]以上であり、最も低かったのは No.4 で 55.8[%]と機種間で差があった。一方画像分類モデルは、表３にあるように CNN の畳み込み層が５層、学習エポック数が 20 と波形分類モデルに比べ軽量である上、適合率は全機種において高い結果となった。これは、各ドローン動作音に周波数領域上で特徴的なスペクトルがあり、それが学習時の特徴量として有効に作用したためと考えられる。波形分類モデル及び画像分類モデルの推論結果の混同行列をそれぞれ図４、図５に示す。図４において、No.4_ANAFI は全 52 データのうち、23 データが No.3_Mavic Mini と誤分類されていた。屋外環境音を混合したことにより、ドローン動作音の特徴的な波形がマスクされ、スペクトルがより広帯域にわたる機種に分類されたためと考えられる。図５では、誤分類されたデータは全 168 個のうち１個のみであり、正解が No.2 のデータに対し No.3 と誤分類された。

４．結言

本研究では、ドローンの動作音を対象として、環境中から特定の音を認識し、その機種を分類する機械学習モデルを構築した。畳み込みニューラルネットワークを用いた機械学習モデルでは、入力として時系列信号をスペクトログラムに変換し、画像として分類した場合、正解率が高かった。また、モデルを軽量化し、データ数が少量であった場合も正解率が 90[%]以上となった。

屋外環境音を混合したデータで推論を行った場合の正解率は、スペクトログラムを入力とした場合で 99.4[%]であり、ノイズのないデータで学習したモデルが一定のノイズ環境下でも使用できる可能性があることがわかった。

参考文献

１） 小澤賢司. “ディジタル音の周波数解析”. ディジタル音響信号処理入門 –Python による自主演習–. コロナ社, 2022, p.54-91.

２） Sony Network Communications Inc. “Support”. SONY Neural Network Console. https://support.dl.sony.com/ja/, (参照 2023-02-21)

３） 森勢将雅. “時間領域での信号解析”. ひたすら楽して音響信号解析 MATLAB で学ぶ基礎理論と実装. コロナ社, 2021, p.17-38.

表４　ノイズ混合音の正解率

モデル	正解率［%］
波形分類モデル（時系列信号）	76.2
画像分類モデル（スペクトログラム）	99.4

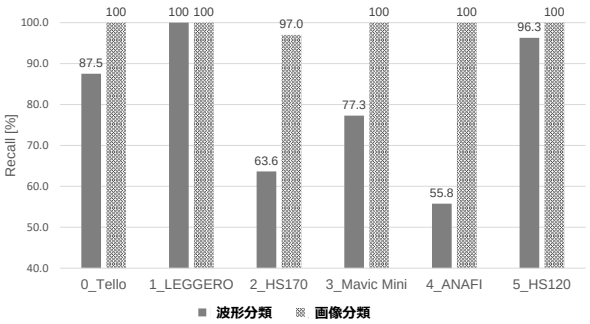

図３　ノイズ混合音の機種別適合率

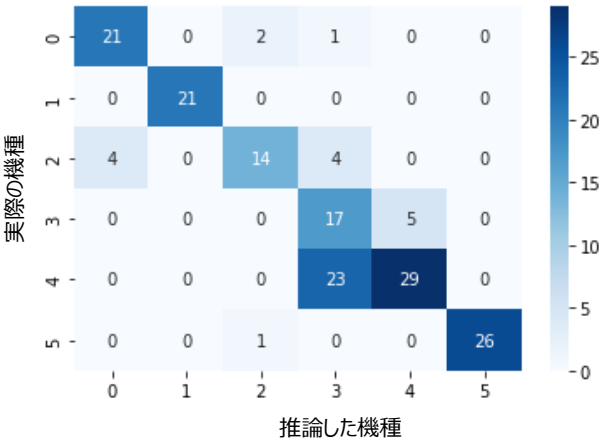

図４　推論結果の混同行列（波形分類モデル）

図５　推論結果の混同行列（画像分類モデル）