2C-04

環境音からの野鳥の鳴き声を検出する畳み込みニューラルネットワークの構築

齋藤　宥杜†　　三上　剛†

苫小牧工業高等専門学校 創造工学科 情報科学・工学系†

1. はじめに

ラムサール条約登録湿地である北海道苫小牧市にあるウトナイ湖は，サハリン，シベリアへ向かう渡り鳥の中継地として知られている．ここで観測される野鳥は 270 種に及び，その中には環境省が指定する絶滅危惧種 15 種が含まれており，日本野鳥の会の調査員が毎日生息調査を行っている[1]．野鳥の生息調査に関する問題点として，一般に以下の点が指摘されている[2].

- 冬季，夜間などは調査員への負担が大きい
- 調査費用の制約から十分な期間と範囲の調査が難しい

これらの改善を目的として，本研究ではウトナイ湖における野鳥の生息調査の高度化を目指したシステムの構築を行っている．システムは屋外に設置する複数台の IoT デバイスと 1 台のサーバによって構成されている．IoT デバイスは AI を搭載し，環境音から野鳥の音声のみを切り出して録音し，4G 回線を用いて定期的にサーバに音声データを送信する．サーバは送られてきたデータから野鳥の種類を識別し，調査記録を自動的に作成する．IoT デバイスをウトナイ湖周辺の複数箇所に設置することにより，夜間の調査や広域調査が可能になると考えられる．

本論では，IoT デバイス側に搭載する「環境音から野鳥の音声のみを切り出して録音する」機能の実現を目指し，音声認識を行う畳み込みニューラルネットである VGGish[3]を転移学習することで様々な環境音と鳥の音声との弁別を行ったので，その結果について報告する．

2. 手法

2-1. 音声データ

野鳥の鳴き声を識別するデータサイエンスコンペティションである BirdCLEF [4]で公開されている 397 種類 19,447 個のサウンドデータを使用した．野鳥以外の音声としては，Google が提供している AudioSet [5]より 20,000 個のデータを用意した．これらのデータから学習用に 80%，評価に 20%のデータを使用した．それぞれのデータについて 10 秒間の時間幅でトリミングし，125Hz～7.5kHz の帯域幅のメルスペクトログラムを計算した．その際，サンプリング周波数は BirdCLEF が 32kHz，AudioSet が 44.1kHz であるため，それぞれダウンサンプリングして 16kHz に変換した．

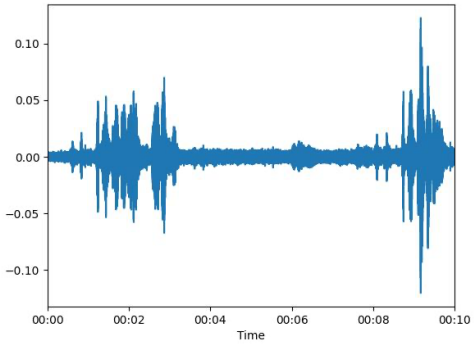

図 1: 野鳥（キバシリ）の鳴き声の音声データ

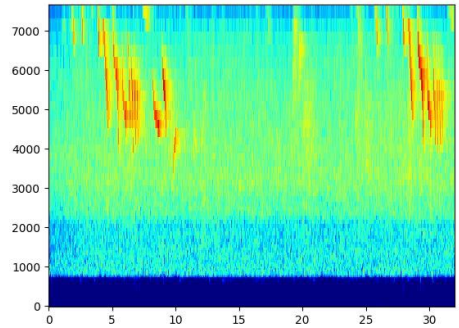

図 2: 図 1 の音声のメルスペクトログラム

2-2. 畳み込みニューラルネットワーク

VGGish は環境音認識を行う畳み込みニューラルネットワークであり，入力は横軸を時間，縦軸を周波数とする音声スペクトログラムを画像として入力して音声識別を行う．学習済みの VGGish のモデルは公開されているため，これを用いて転移学習により野鳥の音声を識別するよ

Detection of birdsongs from environmental sounds using convolutional neural network
†Yuto Saito, Tsuyoshi Mikami
†National Institute of Technology, Tomakomai College

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

うに構築した．VGGish のアーキテクチャを図 1 に示す．入力するデータはメルスペクトログラムであり，時間軸方向に 64 分割，周波数軸方向に 96 分割された値をマトリクスとして入力する．その後，畳み込みとプーリングを繰り返す．VGGish の学習済みモデルとしては Global Average Pooling 層までが公開されているため，この後，全結合層を 2 層接続して野鳥の鳴き声か否かを出力することにした．尚，畳み込み層，全結合層の活性化関数は ReLU であり，最終出力は Softmax とした．学習はミニバッチ 1024，エポック数 50 と設定した．VGGish は画像分類を行う VGG-16/19 モデルをもとにし，音声のスペクトログラムに適用した学習済みモデルである．VGG-16/19 と同様に畳み込みを行った後に MaxPooling を行う流れの繰り返しであるが，層の数は VGG-16/17 より少なく軽量なモデルである．

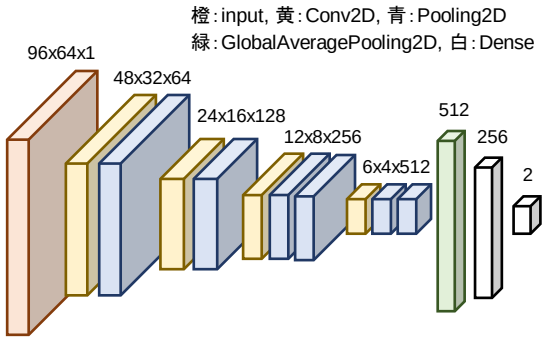

図 3: 本研究で用いた畳み込みニューラルネット

3. 結果と考察

実行環境は表 1 に示すとおりである．50 エポックの学習が終了するまでの時間は 335 秒であった．Accuracy と Loss の学習曲線を図 4, 5 に示す．検証データの識別率が最終的には 95%程度までに到達し，高い精度で野鳥の鳴き声を識別することが可能となった．

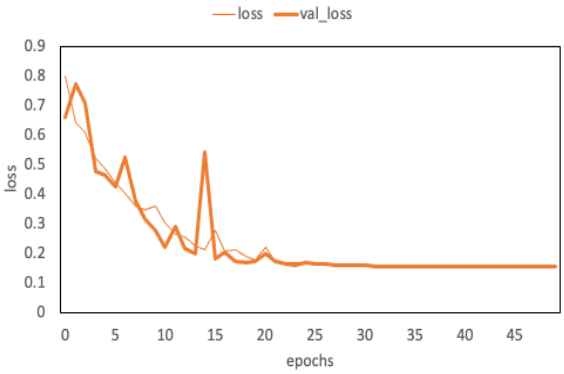

図 4: 学習データと検証データを用いた損失関数

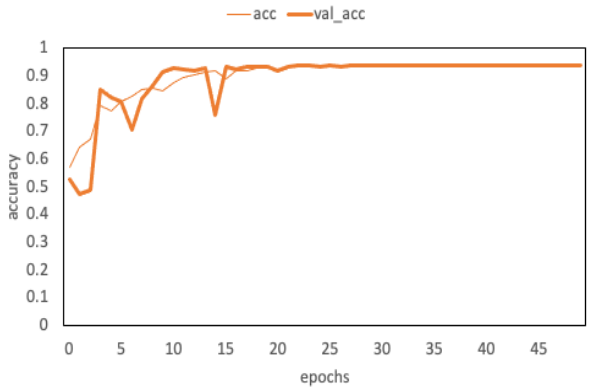

図 5: 学習データと検証データを用いた精度

表 1: 実験で用いた計算機環境

CPU	Intel Corei9-12900K
OS	Ubuntu 20.04
GPU	NVIDIA GeForce RTX3090
CUDA	version 11.5
コンテナ実行環境	NVIDIA Container Toolkit tensorflow:21.03-tf2-py3

4. おわりに

VGGish を転移学習することにより，野鳥の鳴き声のみを環境音から弁別することが可能となった．より実際の状況を考慮すると，以下の点が問題となる

1. 複数の同種の野鳥の鳴き声が同時に集音された場合
2. 複数の異種の野鳥の鳴き声が同時に集音された場合
3. 風，虫，他の動物などの環境音が同時に集音された場合

これらは，サーバ側の識別処理にも関わってくるので，今後実用面での利用を考慮しながら検討していく．

参考文献

[1] ウトナイ湖サンクチュアリ http://park15.wakwak.com/~wbsjsc/011/
[2] 斎藤他, ”音声解析技術の活用による生物の生息調査方法”, 環境アセスメント学会 2019 年講演要旨集
[3] Hershey, S., et al, “CNN Architecture for Large-scale Audio Classification”, ICASSP 2017
[4] BirdCLEF2021 -Birdcall Identification-, Identify bird calls in soundscape recordings, https://www.kaggle.com/c/birdclef-2021/
[5] AudioSet: A large-scale dataset of manually annotated audio events, https://research.google.com/audioset/

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.