BE5　深層学習を用いた鳥の種判別システムの開発

Development of Bird Species Identifier Using Deep Learning

地球循環共生工学領域　08E18029　佐藤匠 (Takumi SATO)

Abstract: As birds can be excellent indicators to assess the environmental quality and change, many projects are now addressing identifying the bird species by their songs. Converting the bird calls into spectrograms and inputting the images into the CNN-based classification model is the common method to identify the bird species. However, the transformer-based models have been gaining attention to strike the balance of the high performance and the low computational cost in the image classification task. In this study, I implemented the two types of 48 bird species identification systems, transformer-based and CNN-based, and compared the validation performance. The transformer-based and the CNN-based model achieved 75.8%, 74.3% in the F1-score metrics, respectively. The future task is to update the bird song sounds to increase the spectrograms.

Keywords: bird classification, acoustic monitoring, deep learning, transformer, convolutional neural network

１．背景と目的

鳥類は，生息地の質や環境状態を理解するための重要な指標となる [1]．そのため，鳴き声から種を判別して鳥類の生息状況を把握する試みが世界的に行われている．鳥類の種判別は，鳥の鳴き声を含む音声をスペクトログラムへ変換し，CNN (Convolutional Neural Network) モデルで画像分類タスクとして種判別を行う手法が一般的である [2]．しかし，画像分類の分野では高い精度と低い計算コストを両立するモデルとして Transformer をベースとした画像分類モデルが注目を集めている [3]．そこで本研究では，従前の CNN の画像分類モデルに加えて，Transformer の画像分類モデルを用いて，2 種類の鳥の種判別システムを構築し，画像分類モデル間の性能を比較し，種判別の性能を評価することを目的とした．

２．研究手法

２．１　スペクトログラムのデータセットの構築

鳥の種判別システムの構築では，2003~2005 年に千葉県立中央博物館が千葉県内の各地で録音した 48 種類の鳥のアノテーション付きの音声ファイル 7,504 を使用した (wave 形式，sampling rate: 22.05 kHz)．アノテーションには，鳴き声の開始時間，鳴き声の継続時間，鳥の種類の情報が含まれている．アノテーションをもとに鳴き声の音声を抽出した後，音声の長さを全て 5 秒で統一するために，5 秒以上継続する音声はその中からランダムに 5 秒を抽出，5 秒未満の音声は不足分を zero padding した．5 秒に調整された音声を短時間フーリエ変換でスペクトログラムに変換し，周波数をメル尺度 (N mels: 128) に変換した．鳥類 48 種と虫の鳴き声などのそれ以外の音 (Noise) の計 8,841 の音響イベントのスペクトログラムのデータセットを構築した．**図 1** はウグイスのスペクトログラムのサンプルである．時間幅は 5 s，周波数帯は 300 Hz から 10,010 Hz を示している．色の濃淡は各周波数帯の音圧レベルを示す．

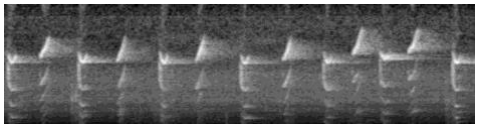

図 1　スペクトログラム (ウグイス)
横軸：時間情報，縦軸：周波数情報

２．２　交差検証によるモデルの訓練と種判別精度の評価

モデルには，CNN では ResNet50d，Transformer では AST (Audio Spectrogram Transformer) [4] を使用した．AST は，スペクトログラムの分類問題で CNN の性能を上回るモデルである．ResNet50d は ImageNet，AST は ImageNet と AudioSet で pre-trained されたモデルを使用した．クラスごとにデータ数の 80%を訓練用データ，20%を評価用データとした層化 5 分割交差検証を行い，評価用データに対して Overall accuracy, Precision, Recall, F1-score で交差検証精度を評価した (**表 1**).

表 1　交差検証に用いたパラメータ

Parameters	Value	
	ResNet50d	AST
Input image size	224 x 224	128 x 512
Optimizer	Adam	
Scheduler	CosineAnnealingLR	
Loss function	BCEWithLogitsLoss	
Learning rate	8.0×10^{-5}	2.0×10^{-5}
Minimum learning rate	1.0×10^{-5}	1.0×10^{-6}
Epochs	32	50
Batch size	64	32

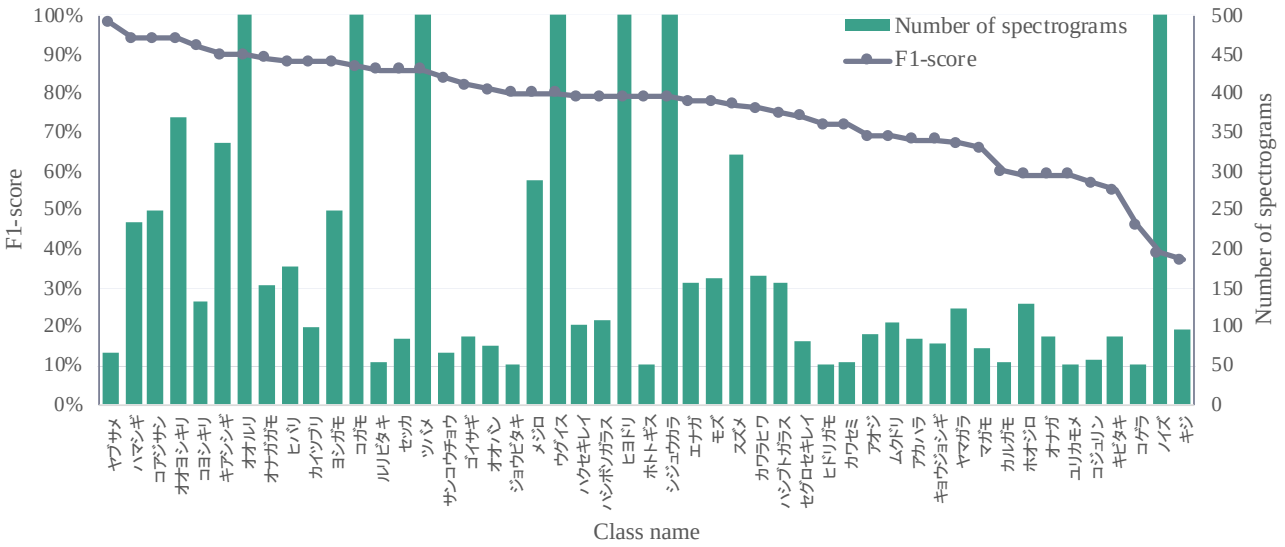

図 2　AST のクラスごとの F1-score とデータ数

３．結果と考察

３．１　種判別の精度評価

図 2 に AST のクラスごとの F1-score とデータ数，**表 2** に 5 分割交差検証の精度の結果を示す．まず Overall accuracy, Recall, F1-score では AST が ResNet50d を上回る精度を示し，Transformer のモデルが鳥類の鳴き声の分類タスクでも高い判別性能を持つことがわかる．また鳥の種類別の精度では，AST は 14 種の鳥で 80%以上の精度を達成した．ここで 8 種の 60%以下の精度の鳥はデータ数が少ない傾向があり，モデルが鳴き声の特徴を十分に捉えることができていない可能性がある．そのため，さらなる判別精度の向上のためにはデータ数の拡張が必要である．

表 2　5 分割交差検証の精度の結果

	AST	ResNet50d
Overall accuracy	**78.8** ± 0.3	**75.7** ± 1.8
Precision	**76.8** ± 0.6	**77.2** ± 0.7
Recall	**75.7** ± 0.5	**72.8** ± 0.4
F1-score	**75.8** ± 0.3	**74.3** ± 0.5

３．２　種判別の精度向上に向けた考察

AST の予測では，全スペクトログラムのうち誤判別は (1,872 枚, 21.2 %) となった．この誤判別は (1) 鳥類の鳴き声と Noise の誤判別 (593 枚, 6.7 %)，(2) 鳥類の鳴き声間での誤判別 (1279 個, 14.5 %) に分類することができる．(1) の誤判別の主な要因は，鳥類の鳴き声とアノテーションされた音声の SN 比が低く，鳴き声の特徴量がスペクトログラムに明確に保持されず誤判別をしたと考えられる．(2) の誤判別の主な要因は，複数種が同時に鳴いていることが考えられる．本研究は 1 枚のスペクトログラムに 1 種類の音響イベントがあるタスクを想定したため，複数種の鳴き声の特徴量が 1 枚のスペクトログラムにある場合，モデルがどの種を予測すべきかを判断できなかったために誤判別が生じたと予想される．

４　今後の課題

今後の課題として，Noise が含まれていても頑健な鳥の種判別を可能にするため，(1) Gaussian Noise や Pink Noise 等を追加することで音声データへの Data Augmentation を行うこと，(2) Noise かどうかの判別をした後に鳥の種判別を行うという 2 段階の予測モデルへ変更すること，(3) 同時に複数種が鳴いた時でも判別を可能にするため，マルチラベル分類タスクに更新することが有効である可能性がある．

参考文献

1) O'Connell, T.J., Jackson, L.E. and Brooks, R.P. (2000), BIRD GUILDS AS INDICATORS OF ECOLOGICAL CONDITION IN THE CENTRAL APPALACHIANS. Ecological Applications, 10: 1706-1721. https://doi.org/10.1890/1051-0761(2000)010[1706:BGAIOE]2.0.CO;2
2) Stowell, D. (2021). Computational bioacoustics with deep learning: a review and roadmap. ArXiv, abs/2112.06725.
3) Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.
4) Gong, Y., Chung, Y., & Glass, J.R. (2021). AST: Audio Spectrogram Transformer. ArXiv, abs/2104.01778.