COMPARISON OF 1D VS 2D CONVOLUTIONAL NEURAL NETWORKS FOR BIRD SOUND DETECTION

TAN PEI HONG

UNIVERSITI TEKNOLOGI MALAYSIA

COMPARISON OF 1D VS 2D CONVOLUTIONAL NEURAL NETWORKS FOR BIRD SOUND DETECTION

TAN PEI HONG

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Computer and Microelectronic Systems)

School of Electrical Engineering
Faculty of Engineering
Universiti Teknologi Malaysia

JULY 2022

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to acknowledge and give my warmest thanks to my supervisor Professor Madya Muhammad Mun'im Bin Ahmad Zabidi, for encouragement, guidance, critics, advices, motivation and friendship. Without his continued support and interest, this thesis would not have been the same as presented here.

My fellow postgraduate seniors should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Automatic acoustic detection system is useful to assist the bird naturalists on bird species monitoring and overall ecosystem health. Many birds are most easily discovered by their sounds, therefore passive acoustic monitoring is most appropriate. However, acoustic monitoring encounters practical limitations such as manual configuration requirement, highly dependent on sounds libraries, less accurate and less robust. In recent years, various machines learning techniques are proposed and detailed performance evaluation are conducted to determine how feasible the automatic acoustic detection system can be achieved. In this paper, we propose a 1D convolutional neural network (CNN) architecture for bird sound detection and compare it with Bulbul 2D CNN architecture which is the winner of Bird Audio Detection (BAD) challenge. The proposed 1D CNNs managed to learn a representation directly from the raw audio recordings. The preprocessing phase divides the audio signal into overlapping frames using a sliding window, thus it can handle audio streams of any duration. The sizes of each frame are compatible to the input layer of the 1D CNNs. On the other hand, the preprocessing phase of Bulbul 2D CNN architecture adopted two type of feature extraction methods, STFT spectrogram and Mel-scaled spectrogram to capture the amplitude of a signal as it changes over time and at various frequencies. The performance of the proposed 1D CNN model in detecting the bird sound was assessed on the warblrb10k dataset and the experimental results have shown that it achieves an accuracy lower than the Bulbul 2D CNN model. It was proven in a few previous 1D CNN state-of-the-art approaches outperform most of the other approaches that uses handcrafted features or 2D representations as input. Due to time constraint, several significant steps of promising high accuracy on 1D CNN model could not be done, such as aggregating the prediction result for all the audio frames belonging to the same audio recording with a majority rule or sum rule to determine the final prediction for presence of bird for the whole individual audio recording, thus lead to achieving low accuracy of 1D CNN model in this paper.

ABSTRAK

Sistem pengesanan akustik automatik berguna untuk membantu naturalis burung dalam pemantauan spesies burung dan kesihatan ekosistem keseluruhan. Banyak burung paling mudah ditemui melalui bunyinya, oleh itu pemantauan akustik pasif adalah paling sesuai. Walau bagaimanapun, pemantauan akustik menghadapi batasan praktikal seperti keperluan konfigurasi manual, sangat bergantung pada perpustakaan bunyi, kurang tepat dan kurang mantap. Dalam beberapa tahun kebelakangan ini, pelbagai teknik pembelajaran mesin dicadangkan dan penilaian prestasi terperinci dijalankan untuk menentukan sejauh mana sistem pengesanan akustik automatik boleh dicapai. Dalam kertas kerja ini, kami mencadangkan seni bina rangkaian neural konvolusi (CNN) 1D untuk pengesanan bunyi burung dan membandingkannya dengan seni bina Bulbul 2D CNN yang merupakan pemenang cabaran Pengesanan Audio Burung (BAD). CNN 1D yang dicadangkan berjaya mempelajari perwakilan secara langsung daripada rakaman audio mentah. Fasa prapemprosesan membahagikan isyarat audio kepada bingkai bertindih menggunakan tetingkap gelongsor, oleh itu ia boleh mengendalikan aliran audio dalam sebarang tempoh. Saiz setiap bingkai adalah serasi dengan lapisan input CNN 1D. Sebaliknya, fasa prapemprosesan seni bina CNN Bulbul 2D menggunakan dua jenis kaedah pengekstrakan ciri, spektrogram STFT dan spektrogram berskala Mel untuk menangkap amplitud isyarat apabila ia berubah mengikut masa dan pada pelbagai frekuensi. Prestasi model CNN 1D yang dicadangkan dalam mengesan bunyi burung telah dinilai pada set data warblrb10k dan keputusan percubaan telah menunjukkan bahawa ia mencapai ketepatan yang lebih rendah daripada model CNN Bulbul 2D. Ia telah terbukti dalam beberapa pendekatan terkini CNN 1D sebelum ini mengatasi kebanyakan pendekatan lain yang menggunakan ciri buatan tangan atau perwakilan 2D sebagai input. Disebabkan oleh kekangan masa, beberapa langkah penting untuk menjanjikan ketepatan tinggi pada model CNN 1D tidak dapat dilakukan, seperti mengagregatkan hasil ramalan untuk semua bingkai audio yang dimiliki oleh rakaman audio yang sama dengan peraturan majoriti atau peraturan jumlah untuk menentukan ramalan akhir untuk kehadiran burung untuk keseluruhan rakaman audio individu, dengan itu membawa kepada mencapai ketepatan model CNN 1D yang rendah dalam kertas ini.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	-	Area Under Curve
ARU	-	Acoustic Recording Unit
CNN	-	Convolutional Neural Network
FN	-	False Negative
FP	-	False Positive
FPR	-	False Positive Rate
GMM	-	Gaussian Mixture Model
HDF5	-	Hierarchical Data Format version 5
HMM	-	Hidden Markov Model
MB	-	Mega Byte
MFCC	-	Mel-Frequency Cepstral Coefficients
ML	-	Machine Learning
ROC	-	Receiver Operator Characteristic
SLVN	-	spectrogram-frame linear network
STFT	-	Short Time Fourier Transform
SVM	-	Support Vector Machine
TN	-	True Negative
TP	-	True Positive
TPR	-	True Positive Rate
XC	-	Xeno-Canto

CHAPTER 1

INTRODUCTION

1.1 Introduction

This chapter introduces the project background, followed by the problem statement that leads to defining the research objectives, project scope and close with the project outline that explains the organization of this paper.

1.2 Project Background

Birds are frequently utilized as bio-indicators of determining environmental quality and ecosystem changes. Acoustic monitoring has been verified as one of the effective methods used by the ecologist to examine the population trend of bird species in a habitat especially a high dense forest. Birds are easily to be detected by sounds as they communicate through vocalizations, thus the recorded sounds can be used for identifying the catch rates and management advice for the following year.

Autonomous sound unit (ARU) are commonly deployed by the ecologist in various random point in the forest and farmland to determine the presence of birds. However, the limitation of ARU includes limited detection distance of the ARU, low accuracies on the target detection, limited digital storage capacity and battery life constraint of the ARU where it limits the onsite operating time.

Deep learning has been a well-known approach recently as it could perform label classification and detection at a higher accuracy, but it was labelled with the fact of heavy computing and monitoring behaviours. It would be a challenge of deploying deep learning algorithm on an embedded system such as an onboard ARU that could support high computational deep learning algorithm. Hence, the research direction would be targeting on delivering an algorithm with low complexity and sufficient accuracy.

1.3 Problem Statement

Bird sound detection in audio is contributing to the task of autonomous wildlife monitoring, citizen science, and the administration of audio libraries. However, acoustic monitoring is frequently constrained by real-world issues such as the requirement of human configuration, reliance on sample sound libraries, lack of precision, low robustness, and generalizability to new acoustic situations.

In modern day, various deep learning techniques have been proposed by the researchers to perform acoustic bird detection, with the use of presented acoustic monitoring dataset, pretraining the model for targeting the acoustic characteristics, and without the need of manual calibration.

In recent years, deep learning algorithm such as convolution neural network (CNN) based approaches reported relatively high accuracy in many audio classification and detection task, especially 2D CNN are widely adopted as it takes in 2D representation comprises of high-dimensional waveforms, although it resulted in high computational power. There are many 2D CNN models submission for bird sound detection in Bird Audio Detection (BAD) challenge 2018 where it has showcased excellent classification capability of 2D CNN. In our opinion, the complexity of 2D CNN was still a crucial issue although it possesses high accuracy in detection task. Besides, 1D CNN is also a well-known method and widely adopted in many audio detection and recognition task. It learns the representation directly from a raw audio signal and was feed into a compact 1D CNN architecture that reduces the quantity of data needed for training and the cost of computing. In previous state of art, 1D CNN achieved excellent performance in various environmental sound classification and detection task but it had not been explored in the task of bird sound detection yet.

1.4 Research Goal

This study aims to compare the model performance of 1D and 2D CNN on identifying the presence or absence of the bird sounds. The 1D CNN learns the representation directly from a raw audio signal while conversion is needed from raw audio signal to 2D frequency representations such as Short-time Fourier transform (STFT) spectrogram and Mel-scaled spectrogram before feeding into the 2D CNN. In this study, the proposed 1D CNN model is referring to the previous state of art from

Abdoli with different amount of convolution layers implemented while Bulbul architecture of 2D CNN model implemented was the winning model architecture of BAD challenge. Based on accuracy, model size, and training duration, the optimal strategy was chosen.

1.5 Research Objective

The following were the research objective of this paper:

1. To review existing approaches for bird sound detection such as Random Forest (RF), Support Vector Machine (SVM) and 2D Convolutional Neural network.
2. To implement the 1D Convolutional Neural Network for Bird Sound Detection task by learning direct representation from raw audio signal.
3. To implement the 2D Convolutional Neural Network Bulbul architecture for Bird Sound Detection task by learning 2D representation such as spectrogram from audio signal.
4. To compare the performance gap between 1D CNN and 2D CNN inbird sound detection.

1.6 Project Scope

- The work will be based on derivatives of the 2D CNN Bulbul architecture designed by Thomas Grill and Jan Schlüter and 1D CNN architecture designed by Abdoli et al.
- The algorithms will be trained and tested on the Warblr Bird Call Dataset, which is an UK-based crowd-sourcing platform for smart-phone recording. It has around 8000 ten-second audio recording which cover around 22 hours of recording.
- The algorithm will be developed and tuned on Google Colab, a free Google Platform that allows written and execution of Python code in the browser.

- A computer with internet access is needed in this project to access the Google Colab platform.
- Programming work will be done by using Python, TensorFlow and Keras Libraries.

1.7 Report Organization

This report can be divided into 5 chapters. The Chapter 1 gives a brief introduction to autonomous bird sound detecting system, as well as a problem statement, research goal, objectives, and the project scope. Followed by Chapter 2, Literature Review which is summary of the current state of art and the proposed approach for bird sound detection task and other audio application task using deep learning and non-deep learning method. The methodology and structure for this research are described in Chapter 3 while Chapter 4 shows the experimental result and discussion of research. Finally, Chapter 5 concludes the work and provides the future work.

REFERENCES

1. Cakir, Emre, et al. "Convolutional recurrent neural networks for bird audio detection." 2017 25th European signal processing conference (EUSIPCO). IEEE, 2017.
2. Sevilla, Antoine, and Hervé Glotin. "Audio Bird Classification with Inception-v4 extended with Time and Time-Frequency Attention Mechanisms." CLEF (Working Notes). Vol. 1866. 2017.
3. Bravo, Carlos J. Corrada, Rafael Álvarez Berríos, and T. Mitchell Aide. "Species-specific audio detection: a comparison of three template-based detection algorithms using random forests." PeerJ Computer Science 3 (2017): e113.
4. Ludena-Choez, Jimmy, Raisa Quispe-Soncco, and Ascension Gallardo-Antolin. "Bird sound spectrogram decomposition through Non-Negative Matrix Factorization for the acoustic classification of bird species." PloS one 12.6 (2017): e0179403.
5. Qian, Kun, et al. "Active learning for bird sounds classification." Acta Acustica united with Acustica 103.3 (2017): 361-364.
6. Labao, Alfonso B., Mark A. Clutario, and Prospero C. Naval. "Classification of bird sounds using codebook features." Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2018.
7. Nanni, Loris, et al. "Combining visual and acoustic features for bird species classification." 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2016.
8. Grill, Thomas, and Jan Schlüter. "Two convolutional neural networks for bird detection in audio signals." 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017.
9. Solomes, Alexandru-Marius, and Dan Stowell. "Efficient bird sound detection on the bela embedded system." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

10. Sankupellay, Mangalam, and Dmitry Konovalov. "Bird call recognition using deep convolutional neural network, ResNet-50." Proceedings of ACOUSTICS. Vol. 7. No. 9. 2018.

11. Lasseck, Mario. "Acoustic bird detection with deep convolutional neural networks." DCASE. 2018.

12. Abdoli, Sajjad, Patrick Cardinal, and Alessandro Lameiras Koerich. "End-to-end environmental sound classification using a 1D convolutional neural network." Expert Systems with Applications 136 (2019): 252-263.

13. Allamy, Safaa, and Alessandro Lameiras Koerich. "1D CNN architectures for music genre classification." 2021 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2021.

14. Pons, Jordi, et al. "End-to-end learning for music audio tagging at scale." arXiv preprint arXiv:1711.02520 (2017).

15. Adesuyi, Tosin Akinwale, Byeong Man Kim, and Jongwan Kim. "Snoring Sound Classification Using 1D-CNN Model Based on Multi-Feature Extraction." International Journal of Fuzzy Logic and Intelligent Systems 22.1 (2022): 1-10.