1-D AND 2-D CONVOLUTION NEURAL NETWORK FOR BIRD SOUND DETECTION

TEE YUN 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 2020

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 project report, I was in contact my supervisor and postgraduate students. They have contributed towards my understanding and thoughts. I wish to express my sincere appreciation to supervisor, Associate Professor Muhammad Mun'im bin Ahmad Zabidi for encouragement, guidance, critics and friendship. Without his continued support and interest, this project report would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Master study. My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have aided 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 members.

ABSTRACT

This research aimed to determine the most suitable audio input format to the Convolution Neural Network (CNN) model, to train a bird activity detector that is low in memory usage with decent accuracy. To enable this investigation, three types of CNN were developed, including one 1-D CNN and two architecturally identical 2-D CNNs that used two different input. 1-D CNN used wav as input, while these two 2-D CNNs used wav image and spectrogram image as input respectively. Accuracy, model size, and training time were used to determine the best model among these three types of CNN. Bird audio and Urban8k audio were used as positive datasets and negative dataset respectively. For each type of CNN model, the most suitable convolution filter size was determined first, before proceeding to determine the best model out of three models of different number of convolution layer. There was one winner for 1-D CNN, 2-D CNN using a wav image and 2-D CNN using a spectrogram image. These three winners were then being compared to determine the overall best model for bird activity detector. For this research, the overall best model was five layers 2-D CNN using a spectrogram image of filter size 5×5. The accuracy achieved was 97.12%, the model size was 6MB, and the training time was fourteen minutes. The additional arithmetic operations required in converting wav to spectrogram was deemed acceptable due to much better accuracy achieved. Spectrogram image was the most suitable audio input format to CNN to train a bird activity detector that is low in memory usage with decent accuracy.

ABSTRAK

Tujuan penyelidikan ini adalah untuk menentukan format input audio yang paling sesuai untuk model Convolution Neural Network (CNN), bagi melatih pengesan aktiviti burung yang rendah dalam penggunaan memori dengan ketepatan yang baik. Tiga jenis CNN dibina untuk membolehkan penyiasatan ini, termasuk satu CNN 1-D dan dua CNN 2-D yang serupa dari segi seni bina yang menggunakan dua jenis input yang berbeza. CNN 1-D menggunakan wav sebagai input, manakala dua CNN 2-D ini menggunakan gambar wav dan gambar spektrogram sebagai input masing-masing. Ketepatan, saiz model dan masa latihan digunakan untuk menentukan model terbaik dalam kalangan tiga jenis CNN ini. Audio burung digunakan sebagai data positif, manakala audio Urban8k digunakan sebagai data negatif. Untuk setiap jenis model CNN, saiz penapis konvolusi yang paling sesuai ditentukan terlebih dahulu, sebelum menentukan model terbaik dari tiga model yang berlainan dalam bilangan lapisan konvolusi. Terdapat satu pemenang untuk CNN 1-D, CNN 2-D yang menggunakan gambar wav dan CNN 2-D yang menggunakan gambar spektrogram. Ketiga-tiga pemenang ini dibandingkan untuk menentukan model terbaik keseluruhan untuk pengesan aktiviti burung. Dalam penyelidikan ini, CNN 2-D lima lapisan yang menggunakan gambar spektrogram dengan saiz penapis 5×5 adalah model terbaik keseluruhan. Ketepatan yang dicapai adalah 97.12%, saiz model adalah 6MB dan masa latihan adalah empat belas minit. Operasi aritmetik tambahan yang diperlukan untuk menukar wav menjadi spektrogram dianggap dapat diterima kerana mencapai ketepatan yang jauh lebih baik. Gambar spektrogram adalah format input audio yang paling sesuai untuk CNN bagi melatih pengesan aktiviti burung yang rendah dalam penggunaan memori dengan ketepatan yang baik.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1-D	-	One-Dimensional
2-D	-	Two-Dimensional
AUC	-	Area under the Receiver Operating Characteristic Curve
CVD	-	Cardiovascular Disease
CNN	-	Convolution Neural Network
CT	-	Chirplet Transform
ESC	-	Environmental Sound Classification
FFT	-	Fast Fourier Transform
GPU		Graphic Processing Unit
Hz	-	Hertz
MAP	-	Mean Average Precisions
MFCC	-	Mel Frequency Cepstral Coefficient
MFCT	-	Mel-frequency Cepstrum Transform
PCG	-	Phonocardiogram
PSD	-	Power Spectral Density
ReLU	-	Rectified Linear Unit
RGB	-	Red Green Blue
SGD	-	Stochastic Gradient Descent
STFT	-	Short Time Fourier Transform
SVM	-	Support Vector Machine
TF	-	TensorFlow
TPU	-	Tensor Processing Unit
UAR	-	Unweighted Average Recall
VGG	-	Visual Geometry Group

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1. Background of Study

Animal monitoring is important for the preservation and conservation of the environment. It allows researchers and environmental activists to assess the impacts of human activities on Earth. Among all the animals, birds are one of the most commonly monitored as the change of bird population could indicate any subtle change of the ecosystem of that particular region. The bird population is a good indicator of changes in biodiversity as scientists need to evaluate the impact of human activities on the population.

The traditional method of assessing bird population is by point count where an expert counts the number of birds seen or heard in a set time. However, this technique is labour-intensive and cannot be done at night or in poor weather. In the last few years, the use of autonomous recording units (ARUs) allows bird populations to be assessed by bioacoustics. The use of ARUs produces a large number of audio recordings also makes manual inspection not feasible. Hence, an automatic bird monitoring system is crucial and beneficial for the nature conservation movement, especially when bird sound is classified at the source. The main advantages of automatic bird sound recognition system are the long-term recording without requiring an observer, deployment in hard to access area is possible and recognition of nocturnal birds or birds with low vocal activity is possible.

For effective monitoring, these systems commonly have the capability of identifying the bird species on the spot using sound. However, most of these monitoring systems are embedded systems that are small in terms of memory space and run on battery. The process of species identification itself requires high computation power, which consumes a large amount of power to identify the bird species with high accuracy. To save power, it is crucial to have a bird activity detector that decides whether the received sound belongs to a bird or not, before turning on the main system to perform species identification.

1.2. Problem Statement

Similar to the actual species identification system, the bird activity detector itself should be able to decide on whether the detected sound does come from a bird or not with decent accuracy. This requires the bird activity detector to possess a certain level of intelligence, although not as high as the species identification system, making it a weak classifier. This can be achieved by using a convolutional neural network (CNN) to create the birds sound detection model. Since the monitoring system is an embedded system, which is low in terms of available memory space, the bird activity detector needs to have low memory usage as well. This requires the number of layers of the CNN model to be small. However, a small number of CNN layer will degrade the detection capability of the bird activity detector itself. Another way to achieve low memory usage is through the reduction of arithmetic operation performed on the input data. The additional arithmetic operation performed on the input increases the amount of useful information carried by each input, but this also increases the size of each input. This leads to an increase in the overall size of the CNN model due to the increase in the number of parameters. The reduction of such arithmetic operation helps to keep the CNN model small. On the other hand, the bird activity detector should also be low power, or else it fails to serve the purpose of conserving power. Both small number of CNN layer and small number of parameters do help to achieve the low power usage requirement

In short, the bird activity detector should have low memory space usage and low power consumption, while still being able to determine the source of detected sound with decent accuracy. With these two constraints, the audio input format has a significant impact. There are two possible types of audio input format to the CNN, either a one-dimensional (1-D) data, which is a raw audio waveform, or a two-dimensional (2-D) data, which is an image. In practice, conversion of 1-D data to 2-D data involves additional arithmetic operation, which is commonly done through Fast Fourier Transform (FFT). The trade-off between accuracy and power saving through the reduction of arithmetic operation performed on the input is the heart of the problem. Building CNN of different architectures and structures enabled this trade-off investigation.

1.3. Research Aim and Objective

This research aims to determine the most suitable audio input format to the CNN model, to train a bird activity detector that is low in memory usage with decent accuracy. There are three objectives in this research, which includes:

- to investigate the feasibility of arithmetic operation reduction and search for the most suitable CNN architecture for bird detection,
- to determine the optimum number of neural network layer for both 1-D and 2-D CNN,
- to evaluate the impact of audio input format on the performance of bird activity detector.

1.4. Scope of Work

To achieve the research objective, several scopes had been outlined. This research aims to determine the most suitable audio input format to the CNN model. Bird audio and non-bird audio were collected and pre-processed to produce three types of dataset, namely wav dataset, wav image dataset, and spectrogram image dataset. Four 1-D CNNs and four 2-D CNNs of various depths were developed. By using a smaller number of dataset, the most suitable filter size was determined for both 1-D and 2-D CNN separately. This was to ensure better performance of CNN during the actual training. This was followed by best model determination for 1-D and 2-D CNN separately using full dataset. Finally, the best CNN was determined among the best 1-D CNN and the best 2-D CNN.

1.5. Significance of Study

By completing this research, the feasibility of arithmetic operation reduction, the most suitable architecture, the optimum number of neural network layer of CNN, as well as impact of audio input format on the performance of bird activity detector can be determined. This allows future researchers to construct and fabricate the required hardware that incorporates this CNN model, which will eventually be applied in an actual embedded system. With this bird activity detector, the bird monitoring system will be able to function for an extended period of time.

LIST OF REFERENCE

[1] “Convolutional neural network,” Wikipedia, 15-Nov-2019. [Online]. Available: https://en.wikipedia.org/wiki/Convolutional_neural_network. [Accessed: 24-Nov-2019].

[2] "A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way", Medium, 2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. [Accessed: 24-Nov-2019].

[3] Kaggle.com. 2018. Rectified Linear Units (Relu) In Deep Learning. [online] Available at: https://www.kaggle.com/dansbecker/rectified-linear-units-relu-in-deep-learning [Accessed: 24-Nov-2019].

[4] Brownlee, J., 2019. A Gentle Introduction To The Rectified Linear Unit (Relu). [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ [Accessed: 24-Nov-2019].

[5] Ackermann, N., 2018. Introduction To 1D Convolutional Neural Networks In Keras For Time Sequences. [online] Medium. Available at: https://blog.goodaudience.com/introduction-to-1d-convolutional-neural-networks-in-keras-for-time-sequences-3a7ff801a2cf [Accessed: 26-Nov-2019].

[6] Green, G., 2018. Seeing Sound: What Is A Spectrogram? - Sound And Vision Blog. [online] Blogs.bl.uk. Available at: https://blogs.bl.uk/sound-and-vision/2018/09/seeing-sound-what-is-a-spectrogram.html [Accessed: 29-Nov-2019].

[7] Kiranyaz, Serkan & Avci, Onur & Abdeljaber, Osama & Ince, Turker & Gabbouj, Moncef & Inman, Daniel. (2019). 1D Convolutional Neural Networks and Applications: A Survey.

[8] Bardeli, Rolf & Wolff, Daniel & Kurth, Frank & Koch, Martina & Tauchert, Klaus-Henry & Frommolt, Karl-Heinz. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognition Letters. 31. 1524-1534. 10.1016/j.patrec.2009.09.014.

[9] P. Rai, V. Golchha, A. Srivastava, G. Vyas and S. Mishra, "An automatic classification of bird species using audio feature extraction and support vector machines," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-5. doi: 10.1109/INVENTIVE.2016.7823241

[10] S. Abdoli, P. Cardinal, and A. L. Koerich, “End-to-end environmental sound classification using a 1D convolutional neural network,” Expert Systems with Applications, vol. 136, pp. 252–263, 2019.

[11] Sigtia, S., Stark, A., Krstulovic, ´ S., & Plumbley, M. (2016). Automatic environmental sound recognition: Performance versus computational cost. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11), 2096–2107

[12] Lee, Jongpil & Kim, Taejun & Park, Jiyoung & Nam, Juhan. (2017). Raw Waveform-based Audio Classification Using Sample-level CNN Architectures.
[13] Sharath Adavanne and Tuomas Virtanen. Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network. In DCASE Workshop, 2017.
[14] Sharath Adavanne, Konstantinos Drossos, Emre Çakır, and Tuomas Virtanen. Stacked convolutional and recurrent neural networks for bird audio detection. In EUSIPCO, 2017.
[15] Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, and Roman Jarina. Stacked convolutional and recurrent neural networks for music emotion recognition. In Sound and Music Computing Conference (SMC), 2017.
[16] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2721-2725. doi: 10.1109/ICASSP.2017.7952651
[17] Sachin Chachada and C-C Jay Kuo, "Environmental sound recognition: A survey," APSIPA Transactions on Signal and Information Processing, vol. 3, pp. e14, 2014.
[18] Dan Stowell and Mark D Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," PeerJ, vol. 2, pp. e488, 2014.
[19] B. Xiao, Y. Xu and X. Bi et al., Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption, Neurocomputing, https://doi.org/10.1016/j.neucom.2018.09.101
[20] J. Herzig, A. Bickel, A. Eitan, N. Intrator, Monitoring cardiac stress using features extracted from S1 heart sounds, IEEE Trans. Biomed. Eng. 62 (2015) 1169–1178.
[21] D.B. Springer, L. Tarassenko, G.D. Clifford, Logistic regression-HSMM-based heart sound segmentation, IEEE Trans. Biomed. Eng. 63 (2016) 822–832.
[22] Á. Incze, H. Jancsó, Z. Szilágyi, A. Farkas and C. Sulyok, "Bird Sound Recognition Using a Convolutional Neural Network," 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2018, pp. 000295-000300. doi: 10.1109/SISY.2018.8524677
[23] D. Gupta. (2017) Transfer learning and the art of using pre-trained models in deep learning. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/06/transfer-learning-the-art-of-fine-tuning-a-pre-trained-model/
[24] R. Narasimhan, X. Z. Fern and R. Raich, "Simultaneous segmentation and classification of bird song using CNN," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 146-150. doi: 10.1109/ICASSP.2017.7952135
[25] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Timefrequency segmentation of bird song in noisy acoustic environments," ICASSP, pp. 2012–2015, 2011.
[26] J. Xie, C. Ding, W. Li and C. Cai, "Audio-only Bird Species Automated Identification Method with Limited Training Data Based on Multi-Channel Deep

Convolutional Neural Networks", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1803.01107. [Accessed: 19- Nov- 2019].

[27] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," 2017 25th European Signal Processing Conference (EUSIPCO), Kos, 2017, pp. 1764-1768. doi: 10.23919/EUSIPCO.2017.8081512

[28] Berger, F., Freillinger, W., Primus, P. and Reisinger, W., 2018. Bird Audio Detection-DCASE 2018. [online] Semanticscholar.org. Available at: https://www.semanticscholar.org/paper/Bird-Audio-Detection-DCASE-2018-Berger-Freillinger/72fbb94984dd90d0ebdf7e94b2410ed702dadf3f#paper-header [Accessed 20 March 2020].

[29] E. Sprengel, M. Jaggi, Y. Kilcher and T. Hofmann, "Audio Based Bird Species Identification using Deep Learning Techniques", Semanticscholar.org. [Online]. Available: https://www.semanticscholar.org/paper/Audio-Based-Bird-Species-Identification-using-Deep-Sprengel-Jaggi/42ffd303b6a8373300a965da4327439575d23131. [Accessed: 20- Nov-2019].

[30] Joly, A., Go¨eau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Champ, J., Planqu´e, R., Palazzo, S., M¨uller, H.: Lifeclef 2016: multimedia life species identification challenges. In: Proceedings of CLEF 2016 (2016)

[31] Xeno-canto.org. (2019). xeno-canto: Sharing bird sounds from around the world. [online] Available at: https://www.xeno-canto.org/ [Accessed 16 Dec. 2019].

[32] Salamon, J., Bello, J. and Jacoby, C. Urban Sound Datasets. (2019). UrbanSound8K. [online] Available at: https://urbansounddataset.weebly.com/urbansound8k.html [Accessed 16 Dec. 2019].