Bird Call Recognition using Deep Convolutional Neural Network, ResNet-50

Mangalam Sankupellay and Dmitry Konovalov

Discipline of Information Technology, James Cook University, Queensland, Australia

ABSTRACT

Birds are an important group of animal that ecologist monitor using autonomous recordings units as an crucial indicator of health of an environment. There is not yet an adequate method for automated bird call recognition in acoustic recordings due to high variations in bird calls and the challenges associated with bird call recognition. In this paper, we use ResNet-50, a deep convolutional neural network architecture for automated bird call recognition. We used a publicly available dataset consisting of calls from 46 different bird species. Spectrograms (visual features) extracted from the bird calls were used as input for ResNet-50. We were able to achieve 60%-72% accuracy of bird call recognition using ResNet-50.

1 INTRODUCTION

Advancement of technology has made it easy to monitor the natural environment with autonomous recordings units (ARU). ARU make it convenient for ecologist to monitor the environment without wasting time on repeated visits to field sites as ARUs can be deployed in the field for weeks or months on end. In addition, ARU enable non-invasive, long-term Passive Environmental Monitoring (PAM), as multiple ARU can be deployed over large spatial and temporal scale, with minimal maintenance time and effort.

The acoustic recordings captured through PAM are used for different purposes by ecologists such as monitoring biodiversity (Gasc et al. 2013), environmental health (Kasten et al. 2012), threatened species (Campos-Cerqueira, Aide and Jones, 2016), invasive species (Hu et al. 2009), occupancy of animals (Kalan et al. 2015) and climate change (Farina, 2014). Commonly, ecologist detect and count the number of specific animal calls in an acoustic recording to track/monitor changes in the environment.

One important group of animal that ecologist monitor in acoustic recordings are birds. Birds are regarded as an important indicator of biodiversity as the number and diversity of bird species in an ecosystem can directly reflect biodiversity, ecosystem health and suitability of the habitat (Priyadarshani, Marsland, and Castro 2018). Birds are also susceptible to changes in the environment. Therefore, monitoring birds/bird calls in an environment provide vital information about changes in the environment itself (Priyadarshani, Marsland, and Castro 2018).

Despite the popular use of ARU in environmental monitoring, there is a lack of suitable tools for processing acoustics recording for automated bird call detections. Many ecologist still rely on manual identification creating a bottleneck in acoustic recording processing. Some tools (such as SoundID (Boucher, 2014)), use the semi-automated approach but these tools require high calibration time, require users to have considerable knowledge in signal processing to use them and the recognizers are tailored for specific calls and do not generalise to other calls (Priyadarshani, Marsland, and Castro 2018). The task of automatic bird call recognition in acoustic recordings is further impacted by the following challenges (Priyadarshani, Marsland, and Castro 2018) :-

- large inter- and intra-species bird call variability.
- environmental noise overlapping with bird calls.
- bird calls on top of each other, especially during dawn and dusk chorus.
- birds generating incomplete/quick calls or long calls in different situation, eg. birds generate quick calls in during breeding season as they are occupied by incubation and/or chick rearing.
- varying power in vocalisation due distance and angles of bird calls from the ALU's microphones

Due to the high variability in bird calls and the challenges associated with bird call recognition, Machine Learning (ML) based automated bird call recognition are favoured. Most of these ML approaches take their lead from automated speech recognition due to the commonalities between human speech and bird calls. These ML approaches include supervised neural networks (including deep learning neural networks) (Cai et al. 2007, Lopes et al. 2011, Chakraborty et al 2016, Qi et al. 2016, Sprengel et al. 2016, Goëau et al. 2016, Sevilla and Glotin, 2017, Fazekas et al. 2017, Kahl et al. 2017), unsupervised neural networks (Stowell and Plumbley, 2014), support vector machine (Dufour et al. 2013, Andreassen, Surlykke and Hallam, 2014, Nanni et al. 2016, Nanni et al. 2017), decision trees (Digby et al. 2013, Lasseck, 2015), random forests (Potamitis, 2014, Fodor, 2013), and hidden markov model (Kogan and Margoliash 1998, Kwan et al. 2004, Trifa 2006, Jančovič and Köküer, 2015, Jančovič and Köküer, 2016). Despite the significant amount of research into automated recognition of bird species, there is not yet an adequate method for field recordings and most works have limited their scope to analyse less noisy and carefully selected recordings, due to challenges associated with bird call recognition (Priyadarshani, Marsland, and Castro 2018).

Currently, deep learning convolutional neural network (CNN) based architectures have reported the most success in benchmarked automatic bird call recogniser challenges. One such benchmarking challenge is the LifeCELF Bird Challenge (BirdCELF), an annual challenge to evaluate the state-of-the-art of audio based identification systems at scale. In BirdCELF 2016, the 3 out of the 5 participating research group who reported their working notes, were based on CNN, specifically shallow-CNN (< 18 layers) (Goëau et al. 2016). To manage more complex image recognition tasks and to increase accuracy of recognition, deep-CNN (> 50 layers) architectures were developed. The BirdCELF 2017 winner, used a pre-trained deep-CNN, the Inception-v4, for high accuracy in bird call recognition, reaching a 0.714 Mean Average Precision (MAP) on 1500 bird species (Sevilla and Glotin, 2017).

The Inception-v4 architecture is an architecture that utilizes residual learning (Szegedy et al, 2016). In CNN, as the layers get larger, training of deep-CNN becomes difficult and the accuracy starts to saturate and then degrade. Residual learning help solve this degrading accuracy problem (He et al. 2016). Residual learning uses shortcut connections as a training method to directly connect input to some other subsequent layers (not just to the next adjacent layer), to train deep-CNN. (Figure 1 shows building block of residual learning, with shortcut connections (He et al. 2015)). The success of Inception-v4 on bird call recognition was attributed to the use of residual learning in training (Szegedy et al, 2016).

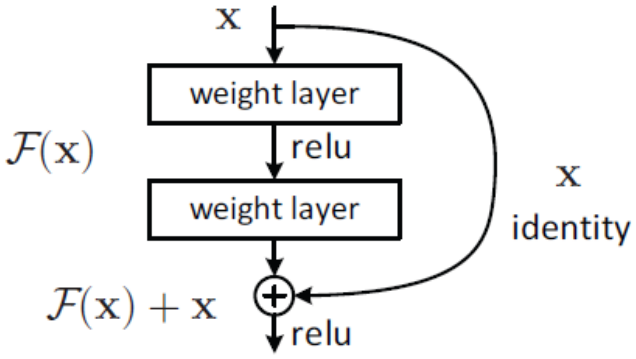


Figure 1: A building block of residual learning (He et al. 2015)

The ResNet-50 (a 50 layer deep-CNN architecture), is the first deep-CNN architecture that utilized residual learning in 2015 (He et al. 2015). ResNet-50 have been successful in increasing accuracy in computer vision benchmarking challenges, winning first prize in the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC, 2015) and Microsoft Common Objects in Context 2015 (MS COCO, 2015) (He et al. 2015). The ResNet-50 model was trained on 1.28 million training images in 1000 classes and reaching an average of 5.25% of top-5 error (He et al. 2015).

Based on the success of ResNet-50 architecture in the computer vision domain, in this research we will utilize the ResNet-50 architecture for automated bird call recognition. We want to take advantage of ResNet-50 use of residual learning for training deep-CNN, similar to Inception-v4. However, we are not using Inception-v4 (winner of BirdCELF 2017) as at the time of writing it was not available in the user-friendly Keras ML platform (Chollet, F. et al. (2015). It is hoped that using the off-the-shelf ImageNet-pretrained ResNet-50 CNN architecture will make the model more acceptable to the ecological research community, and the reported results could be reproduced more easily. Furthermore, if the easy-to-use Keras ResNet-50 is applied to other bird sounds, the results could be compared more meaningfully. To our knowledge there has not been any reported research into the application of ResNet-50 in bird call recognition.

2 METHODS

In this research we use the ResNet-50, a deep-CNN based architecture for automatic bird call recognition in acoustic recordings.

2.1 Data

In BirdCELF 2016 and BirdCELF 2017 challenges, the dataset was derived from Xeno-Canto containing bird calls from 1500 bird species. In our research, we will use a publicly available dataset that is a subset of Xeno-Canto dataset used in the BirdCELF challenges. We will use a dataset developed by Nanni et al. 2016, from the Xeno-Canto site, which has bird calls within a radius of up to 250 km of the city of Curitiba, in the South of Brazil. Nanni et al. 2016 removed all bird species with less than 10 samples. After these filters, 2814 audio samples, representing 46 bird species remained in the dataset and was made available at Nanni et al. 2016 homepage (http://www.din.uem.br/yandre/birds/bird_songs_46.tar.gz). The sample rate of the audio files were 22.05KHz.

By using a publicly available dataset that has been tested with an existing algorithm, we will be able to benchmark the ResNet-50 architecture with the work of Nanni et al 2016. In addition, using a subset of the dataset from BirdCELF dataset, would enable us to quickly validate the use of ResNet-50 without extended training time.

2.2 Spectrogram

The bird calls were converted into spectrogram images where the spectrum of frequencies (vertical y-axis, Hz) varied according to time (horizontal x-axis, sec). The intensity of each pixel represent the amplitude of the bird call. Spectrograms were calculated using Fast Fourier Transformation (FFT) with a Hamming window with a frame length of 256x4=1024 samples and a (256-32)x4=896 samples (87.5%) overlap between subsequent frames.

2.3 ResNet-50

In this work, we used ResNet-50, a deep-CNN architecture for bird call recognition. First introduced by He et al. 2015, the ResNet-50 architecture has become a seminal model to demonstrate that deep networks can be trained with residual learning and help solve the degrading accuracy problem in deep-CNN architectures. We used ResNet-50 model that was available in the high-level neural network Application Programming Interface (API) Keras (Chollet et al 2015) in the machine learning Python package, TensorFlow backend (Abadi, et al. 2015).

The ResNet-50 model, in Keras was with pre-trained weights, that was trained to recognise the 1,000 different ImageNet (Russakovsky et al. 2015) object classes. The original ImageNet trained architectures was modified to classify the 46 bird species. This was achieved by replacing the last densely-connected 1,000 neuron layer with a 46 neuron fully-connected layer. Specifically, let r and c denote the number of spectrogram pixel rows and columns, respectively. The network then accepted an r x c x 3 input image, where the grayscale spectrogram image was loaded into each of the three color channels expected by the ResNet CNN. After removing the ImageNet 1,000 classification layers, the ResNet-50 network outputs had the r_a x c_a x f_a shape, where f_a was the number of extracted features for each r_a x c_a spacial location. The spatial average pooling layer was used to convert the fully-convolutional r_a x c_a x f_a output to the f feature vector, which was then densely connected to the final 46 bird call classification layer. Sigmoid activation function was used in the classification layer since in practice multiple bird calls could be present in the same image. Hence each bird call specific sigmoid-activated neuron could independently detect the bird call it was trained for in a given spectrogram. Minimal preprocessing was done to the training images: the [0,255]-range pixel color intensity was divided by 125.5 and per-image mean-value was subtracted. The training images were augmented by multiplying each image by a scaling factor randomly selected from the [0.75,1.25]-range.

All 2814 labelled spectrograms were randomly partitioned into a 75%-25% split of training and validation subsets. The training subset is made up by 75% of the randomly selected spectrograms, while 25% were used as the validation subset to monitor the training process and to estimate the predictive accuracy of the CNN. Prior to training, the ResNet-50 model was loaded with the corresponding ImageNet-trained weights, available within Keras. It is more accurate and faster to train ResNet-50 model with ImageNet-trained weights than to train randomly initialised ResNet-50 model (Oquab et al. 2014). For the newly created 46-neuron fully-connected (i.e. the output layer), the weights were initialised using uniform random distribution specified in Glorot and Bengio, 2010. The standard binary cross-entropy loss function was used for training.

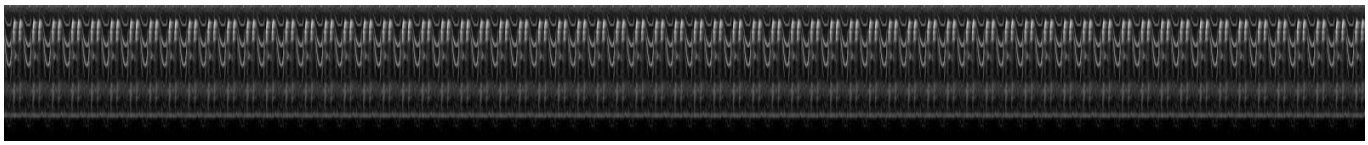
To train the ResNet-50 model, the Keras implementation of Adam (Kingma and Ba, 2014), a first-order gradient-based method for stochastic optimisation, was used. The initial learning-rate (*lr*) was set to $lr = 1 \times 10^{-6}$, which was very low to allow the ImageNet-trained weight to adjust gradually. It was then successively halved every time the validation loss did not decrease after 10 epochs, where the validation loss refers to the classification error computed on the validation subset of images. While training, the model with the smallest running validation loss was continuously saved, in order to restart the training after an abortion. The training was performed in batches of 4 spectrograms, and aborted if the validation loss did not decrease after 32 epochs. In such cases, the training cycle is repeated two more times with the initial learning rates $lr = 0.5 \times 10^{-6}$ and $lr = 0.25 \times 10^{-6}$, respectively.

3 RESULTS

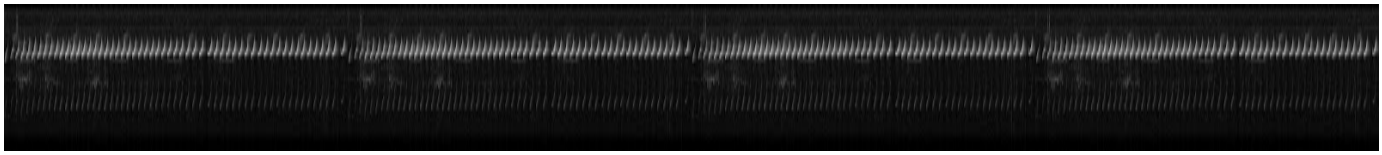
We used ResNet-50, a deep-CNN architecture for automated bird call recognition on a dataset consisting of bird calls from 46 bird species.

3.1 Spectrogram

A total of 2814 spectrogram of bird calls was generated. Figure 2 shows sample spectrograms of 2 different bird species (Bird Species No. 7 and Bird Species No. 46).



(a) Bird Species No. 7



(b) Bird species No. 46

Figure 2: Sample spectrograms of Bird Calls

3.2 ResNet-50

Figure 4 shows the learning process for ResNet-50 model on both the training and validation bird call subsets. Initially, the ResNet-50 was trained on 512 (height) x 1024 (width) images randomly cropped from the spectrograms. With this training crop-size, the network reached about 65% training accuracy and 57% validation accuracy. The validation images were augmented and pre-processed identically to the training images. After the two restarts, the accuracy began to plateau after 300 epochs. An additional experiment was then performed, where the model was continue to be trained (including the two restarts) but with 512 x 512 size cropped images. In addition, a uniform random noise in the range of [0,25] was added to each cropped image. This improved the accuracy to about 72% training accuracy and 65% validation accuracy. The accuracy began to plateau after 400 epochs.

The training accuracy in both instances exceeds the validation accuracy by only a small amount (< ~7%). This is indicated of a network that is not underfitting or overfitting to the training data. It took approximately 100 hours to train the ResNet-50 model on an Nvidia GTX 1070 GPU.

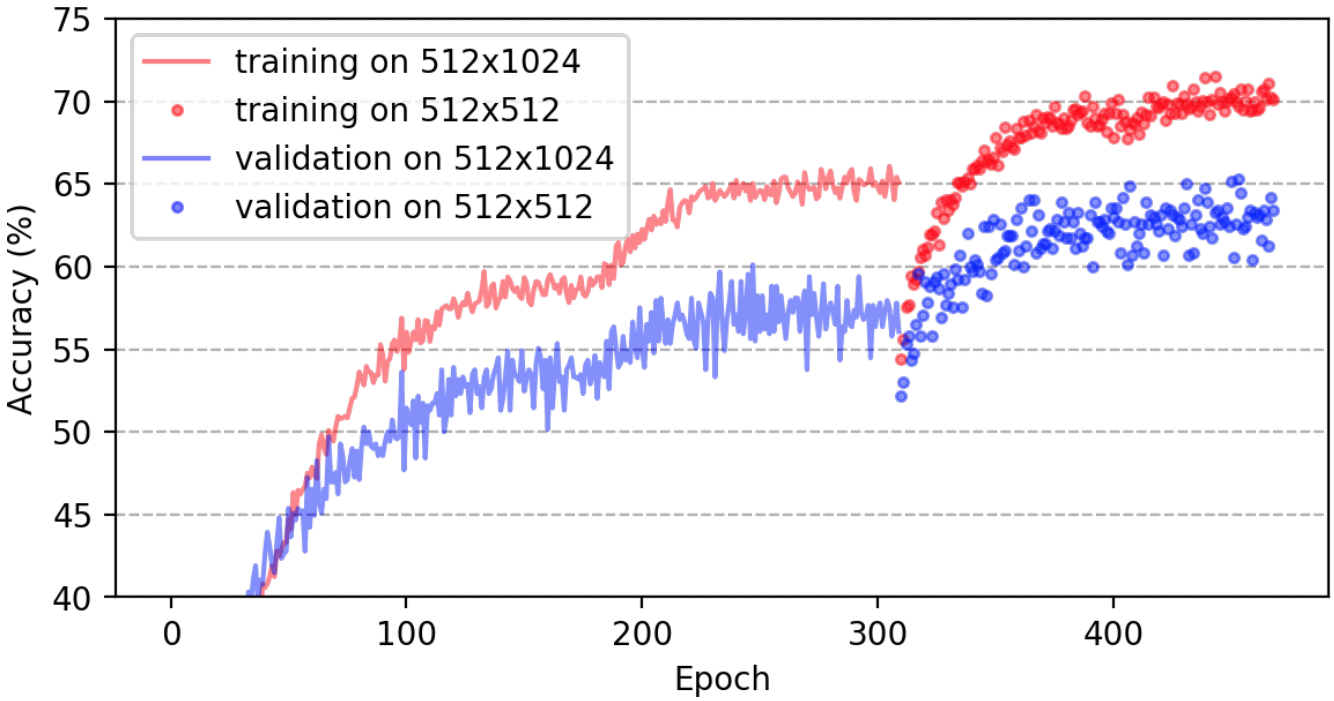


Figure 3: The training and validation accuracy of ResNet-50

4 DISCUSSION AND CONCLUSION

In this work, we trained the ResNet-50 model to automatically recognise bird calls in acoustics recording. We selected the ResNet-50 model due its success in the ImageNet Large Scale Visual Recognition Challenge in 2015. ResNet-50 uses residual learning to train deep-CNN models and overcome the degrading accuracy problem that occurs when training deep-CNN. We applied the ResNet-50 model to training and recognise bird calls from 46 bird species, from a publicly available dataset.

ResNet-50 training accuracy was about 65% when the input was 512 (height (frequency)) x 1024 (length (time)). When the input data is selected at 1024 length (time), in affect the input spectrogram to ResNet-50 is 5.94 sec long (1024 * 32 * 4 / 22050 Hz) call. ResNet-50 training accuracy improved to 72% when the selected input spectrogram was reduced to 512 (height (frequency)) x 512 (length (time)), effectively reducing the input bird call to 2.97 sec long calls. The improvement of ResNet-50 accuracy indicated the CNN learned shorter bird calls at a more accurate level.

Typically bird calls can be split into segments called syllables (Catchpole and Slater, 2003). In Figure 4, below, shows a close up of a spectrogram of a bird call from Bird Species No. 7 consist of 2 syllables. Bird call are better characterized by syllabes (Lopes et al. 2011) and syllables can be seen as basic recognition units. In our work, ResNet-50 is able to achieve higher level of accuracy when the input calls (spectrograms) are shorter, indicating that the convolutional pools within ResNet-50 are extracting and using smaller features (syllables) for bird call recognition. This is inline with the recommendation by Nanni et al. 2016 to use basic features of a bird call syllable for automated recognition. In future work, we will explore the use of shorter input spectrograms to improve the accuracy of ResNet-50 model.

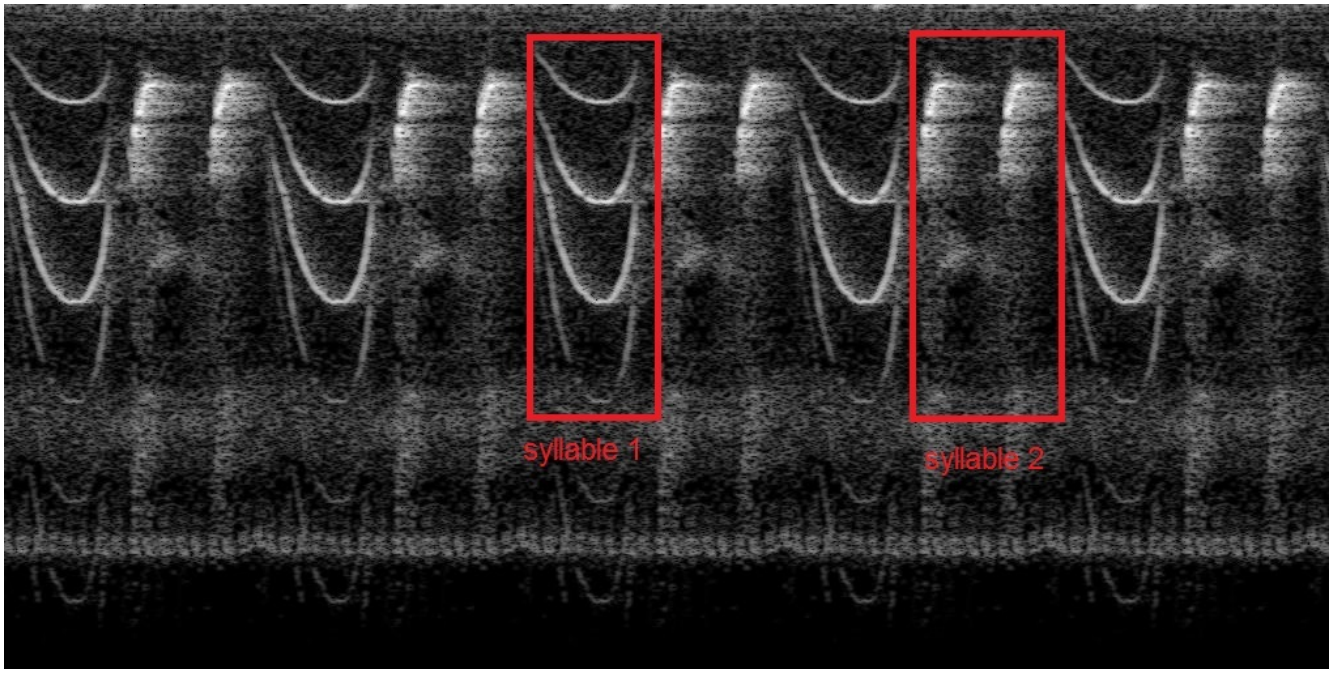


Figure 4: Spectrogram of Bird Species No. 7 consisting of two syllables

The database used in our research is a publicly available dataset made available by Nanni et al. 2016 to train support vector machine models for automatic recognition of bird calls. Nanni et al. 2016 used a combination of visual and acoustic features from bird calls to achieve 94.5% accuracy. (Nanni et al. 2016 acknowledges that one of the drawback of their analysis is the increased computational cost for extracting multiple features.) In our future work, we will investigate the use of audio and visual features (as used by Nanni et al. 2016) as training input for ResNet-50 to increase the accuracy of bird calls recognition.

REFERENCES

Abadi, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous system. Software available from https://www.tensorflow.org/

Antoine Sevilla, Herv´e Glotin. 2017. Audio Bird Classification with Inception-v4 extended with Time and Time-Frequency Attention Mechanisms, – CLEF (working notes) http://ceur-ws.org/Vol-1866/paper_177.pdf

Botond Fazekas, Alexander Schindler, Thomas Lidy, and Andreas Rauber. 2017. A Multi-modal Deep Neural Network approach to Bird-song identification CLEF (working notes) http://ceur-ws.org/Vol-1866/paper_179.pdf

Boucher, N. J. 2014. SoundID version 2.0.0 documentation. – SoundID.

Catchpole, C. K.. Slater, P. J. B. *Bird Song: Biological Themes and Variations*. Cambridge University Press, 2018.

Cai, Jinhai, Dominic Ee, Binh Pham, Paul Roe, and Jinglan Zhang. "Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition." *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, 2007. doi:10.1109/issnip.2007.4496859.

Cai, Jinhai, Dominic Ee, Binh Pham, Paul Roe, and Jinglan Zhang. "Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition." *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, 2007. doi:10.1109/issnip.2007.4496859.

Campos-Cerqueira, Marconi, and T. Mitchell Aide. "Improving Distribution Data of Threatened Species by Combining Acoustic Monitoring and Occupancy Modelling." *Methods in Ecology and Evolution* 7, no. 11 (07, 2016): 1340-348. doi:10.1111/2041-210x.12599.

Chakraborty, Deep, Paawan Mukker, Padmanabhan Rajan, and A. D. Dileep. "Bird Call Identification Using Dynamic Kernel Based Support Vector Machines and Deep Neural Networks." *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 12 2016. doi:10.1109/icmla.2016.0053.

Chollet, F. et al. (2015). Keras. https://github.com/fchollet/keras

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, arXiv:1602.07261

COCO 2015 Object Detection Task, http://cocodataset.org/#detection-2015

Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization, 2014. arXiv:1412.6980

Digby, Andrew, Michael Towsey, Ben D. Bell, and Paul D. Teal. "A Practical Comparison of Manual and Autonomous Methods for Acoustic Monitoring." *Methods in Ecology and Evolution* 4, no. 7 (05, 2013): 675-83. doi:10.1111/2041-210x.12060.

Dufour, Olivier, Thierry Artieres, Herv Glotin, and Pascale Giraudet. "Clusterized Mel Filter Cepstral Coefficients and Support Vector Machines for Bird Song Identification." *Soundscape Semiotics - Localisation and Categorisation*, 03, 2014. doi:10.5772/56872.

Fagerlund, Seppo. "Bird Species Recognition Using Support Vector Machines." *EURASIP Journal on Advances in Signal Processing* 2007, no. 1 (05, 2007). doi:10.1155/2007/38637.

Farina, Almo. *Soundscape Ecology: Principles, Patterns, Methods and Applications*. Springer, 2014.

Fodor, Gabor. "The Ninth Annual MLSP Competition: First Place." *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 09 2013. doi:10.1109/mlsp.2013.6661932.

Gasc, Amandine, Jérôme Sueur, Sandrine Pavoine, Roseli Pellens, and Philippe Grandcolas. "Biodiversity Sampling Using a Global Acoustic Approach: Contrasting Sites with Microendemics in New Caledonia." *PLoS ONE* 8, no. 5 (05, 2013). doi:10.1371/journal.pone.0065311.

Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. & Titterington, M. (eds.) *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, vol. 9 of Proceedings of Machine Learning Research*, 249–256, PMLR, Chia Laguna Resort, Sardinia, Italy, 2010.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2016. doi:10.1109/cvpr.2016.90.

Hervé Goëau, Hervé Glotin, Willem-Pier Vellinga, Robert Planqué, Alexis Joly. LifeCLEF Bird Identification Task 2016: The arrival of Deep learning. *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, Sep 2016, Evora, Portugal. pp.440–449, 2016.

Hu, Wen, Van Nghia Tran, N. Bulusu, Chun Tung Chou, S. Jha, and A. Taylor. "The Design and Evaluation of a Hybrid Sensor Network for Cane-toad Monitoring." *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.* doi:10.1109/ipsn.2005.1440984.

Jancovic, Peter, and Munevver Kokuer. "Acoustic Recognition of Multiple Bird Species Based on Penalised Maximum Likelihood." *IEEE Signal Processing Letters*, 2015, 1. doi:10.1109/lsp.2015.2409173.

Jančovič, Peter, and Münevver Köküer. "Recognition of Multiple Bird Species Based on Penalised Maximum Likelihood and HMM-Based Modelling of Individual Vocalisation Elements." *Interspeech 2016*, 09, 2016. doi:10.21437/interspeech.2016-669.

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun, 2015. Deep Residual Learning for Image Recognition. https://arxiv.org/pdf/1512.03385.pdf

Kalan, Ammie K., Roger Mundry, Oliver J.j. Wagner, Stefanie Heinicke, Christophe Boesch, and Hjalmar S. Kühl. "Towards the Automated Detection and Occupancy Estimation of Primates Using Passive Acoustic Monitoring." *Ecological Indicators* 54 (07 2015): 217-26. doi:10.1016/j.ecolind.2015.02.023.

Kasten, Eric P., Stuart H. Gage, Jordan Fox, and Wooyeong Joo. "The Remote Environmental Assessment Laboratory's Acoustic Library: An Archive for Studying Soundscape Ecology." *Ecological Informatics* 12 (11 2012): 50-67. doi:10.1016/j.ecoinf.2012.08.001.

Kogan, Joseph A., and Daniel Margoliash. "Automated Recognition of Bird Song Elements from Continuous Recordings Using Dynamic Time Warping and Hidden Markov Models: A Comparative Study." *The Journal of the Acoustical Society of America* 103, no. 4 (04 1998): 2185-196. doi:10.1121/1.421364.

Kwan, C., G. Mei, X. Zhao, Z. Ren, R. Xu, V. Stanford, C. Rochet, J. Aube, and K.c. Ho. "Bird Classification Algorithms: Theory and Experimental Results." *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/icassp.2004.1327104.

Large Scale Visual Recognition Challenge 2015 (ILSVRC2015) http://image-net.org/challenges/LSVRC/2015/results

Lasseck, Mario. "Towards Automatic Large-Scale Identification of Birds in Audio Recordings." *Lecture Notes in Computer Science Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 2015, 364-75. doi:10.1007/978-3-319-24027-5_39.

Lasseck, Mario. "Towards Automatic Large-Scale Identification of Birds in Audio Recordings." *Lecture Notes in Computer Science Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 2015, 364-75. doi:10.1007/978-3-319-24027-5_39.

Lopes, Marcelo T., Lucas L. Gioppo, Thiago T. Higushi, Celso A.a. Kaestner, Carlos N. Silla Jr., and Alessandro L. Koerich. "Automatic Bird Species Identification for Large Number of Species." *2011 IEEE International Symposium on Multimedia*, 12 2011. doi:10.1109/ism.2011.27.

Lopes, Marcelo Teider, Carlos Nascimento Silla Junior, Alessandro Lameiras Koerich, and Celso Antonio Alves Kaestner. "Feature Set Comparison for Automatic Bird Species Identification." *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 10 2011. doi:10.1109/icsmc.2011.6083794.

Nanni, L., Y.m.g. Costa, D.r. Lucio, C.n. Silla, and S. Brahnam. "Combining Visual and Acoustic Features for Bird Species Classification." *2016 IEEE 28th International Conference on Tools with Artificial Intelligence*

Nanni, L., Y.m.g. Costa, D.r. Lucio, C.n. Silla, and S. Brahnam. "Combining Visual and Acoustic Features for Audio Classification Tasks." *Pattern Recognition Letters* 88 (03 2017): 49-56. doi:10.1016/j.patrec.2017.01.013.

Oquab, Maxime, Leon Bottou, Ivan Laptev, and Josef Sivic. "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks." *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 06 2014. doi:10.1109/cvpr.2014.222.

Potamitis, Ilyas. "Automatic Classification of a Taxon-Rich Community Recorded in the Wild." *PLoS ONE* 9, no. 5 (05, 2014). doi:10.1371/journal.pone.0096936.

Priyadarshani, Nirosha, Stephen Marsland, and Isabel Castro. "Automated Birdsong Recognition in Complex Acoustic Environments: A Review." *Journal of Avian Biology* 49, no. 5 (05 2018). doi:10.1111/jav.01447.

Qi, Simeng, Zheng Huang, Yan Li, and Shaopei Shi. "Audio Recording Device Identification Based on Deep Learning." *2016 IEEE International Conference on Signal and Image Processing (ICSIP)*, 08 2016. doi:10.1109/siprocess.2016.7888298.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115, no. 3 (04, 2015): 211-52. doi:10.1007/s11263-015-0816-y.

Stefan Kahl, Thomas Wilhelm-Stein, Hussein Hussein, Holger Klinck, Danny Kowerko, Marc Ritter, and Maximilian Eibl. 2017 Large-Scale Bird Sound Classification using Convolutional Neural Networks CLEF (working notes) http://ceur-ws.org/Vol-1866/paper_143.pdf

Sprengel, E., Jaggi, M., Kilcher, Y. and Hofmann, T. 2016. Audio based bird species identification using deep learning techniques – CLEF (working notes), pp. 547–559.

Stowell, Dan, and Mark D. Plumbley. "Automatic Large-scale Classification of Bird Sounds Is Strongly Improved by Unsupervised Feature Learning." *PeerJ* 2 (07, 2014). doi:10.7717/peerj.488.

Trifa, V. 2006. A framework for bird songs detection, recognition and localization using acoustic sensor networks. – Master's thesis, École Polytechnique Fédérale de Lausanne.