# Convolutional Neural Networks for Large-Scale Bird Song Classification in Noisy Environment

Bálint Pál Tóth, Bálint Czeba

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics,
Magyar Tudósok krt. 2., H-1117, Budapest, Hungary
toth.b@tmit.bme.hu, czbalint14@gmail.com

**Abstract.** This paper describes a convolutional neural network based deep learning approach for bird song classification that was used in an audio record-based bird identification challenge, called BirdCLEF 2016. The training and test set contained about 24k and 8.5k recordings, belonging to 999 bird species. The recorded waveforms were very diverse in terms of length and content. We converted the waveforms into frequency domain and splitted into equal segments. The segments were fed into a convolutional neural network for feature learning, which was followed by fully connected layers for classification. In the official scores our solution reached a MAP score of over 40% for main species, and MAP score of over 33% for main species mixed with background species.

**Keywords:** Convolutional Neural Network, Deep Learning, Classification, Bird Song, Audio, Waveform

## 1 Introduction

Identification and classification of bird species can greatly help to explore biodiversity and to monitor unique patterns in different soundscapes [1]. The LifeCLEF 2016 is a competition hosted by CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) [2]. BirdCLEF 2016 [3] is a part of the LifeCLEF competition and addresses the classification of 999 different bird species based on audio recordings of Xeno-canto collaborative database [4]. Whereas the original Xeno-canto database includes about 275,000 audio records covering 9450 bird species from all around the world, the BirdCLEF 2016 focuses on South-America (Brazil, Colombia, Venezuela, Guyana, Suriname and French Guiana) and contains 24607 audio recordings belonging to the 999 bird species. The test set included 8596 recordings from the BirdCLEF 2015 challenge extended by soundscape recordings. The latter means that the recordings are not focusing on specific bird species, but contains the environmental sounds with arbitrary number of singing birds. The length of the samples was widely diverse, in the training set the longest recording was ~45 minutes long, and the shortest length of recording was ~260 milliseconds. In the test set the longest was about 2 hours and 18 minutes, while the shortest ~700 milliseconds.

The LifeCLEF challenge allows manually aided solutions (like crowdsourcing), however we have chosen state-of-the-art deep learning techniques to address the problem. Our solution uses two dimensional convolutional neural networks, that is trained with preprocessed bird songs transformed to the frequency domain.

The outline of this paper is as follows. Section 2 briefly overviews the application of convolutional neural networks in speech recognition and sound classification, furthermore investigates some solutions for the previous BirdCLEF challenges. Section 3 describes the data preparation method we applied. Section 4 introduces the applied deep learning technique and neural network architectures for bird song classification. Section 5 presents our results and Section 6 draws conclusions.

## 2 Related Work

Besides image classification one of the main propelling force of deep learning is speech recognition. In speech recognition different deep learning techniques, like deep belief networks, deep neural networks and convolutional networks, are proven to surpass the accuracy of 'traditional' Gaussian Mixture Models [5]. Recurrent architectures, especially Long Short-Term Memory (LSTM) networks are successfully applied to speech recognition tasks as well [6]. Combining convolutional and LSTM-based recurrent networks the accuracy of speech recognition can be further improved [7].

The task of bird song classification with neural networks has been investigated even back in 1997 [8]. They have applied feedforward neural network with 3-8 hidden neurons to classify 6 bird species from 133 recordings. They have achieved 82% accuracy with neural nets, however Quadratic Discriminant Analysis reached significantly better results, namely 93%. Another approach is presented in [9]. In their work after noise reduction 13 dimensional Mel-Frequency Cepstral Coefficient (MFCC) features were extracted and their dynamic counterpart were calculated. This 26 dimensional vector of the current, the preceding and the following frames were fed into a feed forward neural network with one hidden layer and 10-160 hidden neurons. They reached 98.7% and 86.8% accuracy on classifying 4 and 14 bird species, respectively. In [10] a random forest based segmentation method is shown to select bird calls in noisy environments with 93.6% accuracy. The work introduced in [11] uses binned frequency spectrum, MFCC and Linear Prediction Coefficients (LPC) features, that are classified by an ensemble of logistic regression, random forests and extremely randomized trees. They achieved 4th place on NIPS4B bird classification challenge hosted on Kaggle.

There have been a number of competitive approaches in the BirdCLEF challenges of previous years, however deep learning was not applied in the BirdCLEF competition before. The winning solution of 2014 used a robust feature extraction (including MFCC, fundamental frequency, zero crossing rate, energy features, etc. - altogether 6669 features per recordings), feature selection (reducing the number of features from 6669 to 1277) and template matching. The last year's challenge was won by the same competitor. His work described in [12] downsamples the spectrograms for faster feature extraction, applies decision trees for feature ranking and selection and bootstrap aggregating for classification.

## 3 Data preparation

As a first step, we downsampled every audio file to 16 kHz frequency in order to reduce the size of the training data. Following the preprocessing steps of [10], first a Hamming window and then a short-time FFT were applied with a frame length of 512 samples and 256 samples overlap between subsequent frames. Next we implemented and applied a filtering method to extract the essential parts of the spectrogram, that contains bird calls. Some previous work (e.g. [11]) filters frequencies below 1 kHz, however in the current dataset we found useful information also in this range (see Figure 1), so we only applied low-pass filter with cutoff frequency of 6250 Hz.

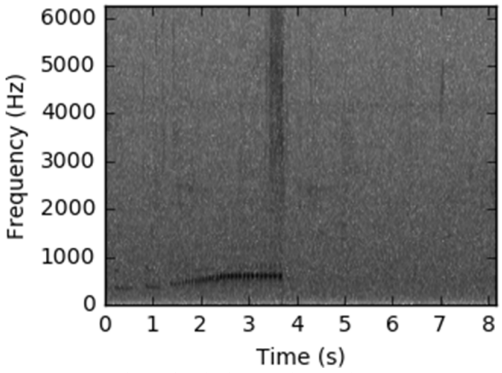


**Fig. 1.** Example of useful information (bird call) below 1 kHz.

As a result, the vertical dimension (frequency) of the spectogram was 200, and the horizontal dimension (time) depended on the length of the recording. In the time domain (horizontal axis) we split the spectograms into 30 sample long columns (that corresponds ~0.5 seconds) and in the frequency domain (vertical axis) we split the spectrograms into 10 sample high rows. As a result, every spectogram was split into 30×10 sized cells. We used these cells to remove the irrelevant parts (that is likely not to contain any bird call) of the spectrogram based on the mean and the variance. We calculated the mean and variance of every 10 sample high row (that corresponds a frequency range). If a cell's mean is less than 1.5 times the addition of mean plus variance of the actual row, than we dropped the cell. In case of Run 1, 3 and 4 we also removed those parts of the filtered spectrogram where 95% of the column vectors were zeros (see Figure 2). This step was skipped at our second submission (referred to as 'BME TMIT Run 2' in the official results; see Table 1). After these preprocessing steps we split the remaining parts of the spectrogram to five seconds long pieces. Thus the dimensions of the resulting arrays were 200×310 (310 samples corresponds to five seconds). We used this as the input of the convolutional neural network.

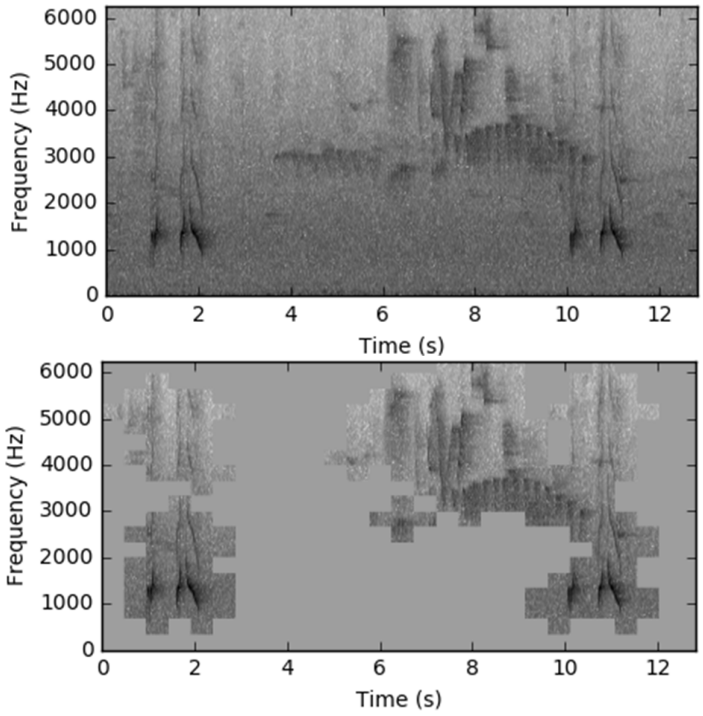


**Fig. 2.** Example of a spectrogram before (above) and after (below) preprocessing, when zero elements were kept (Run 2).

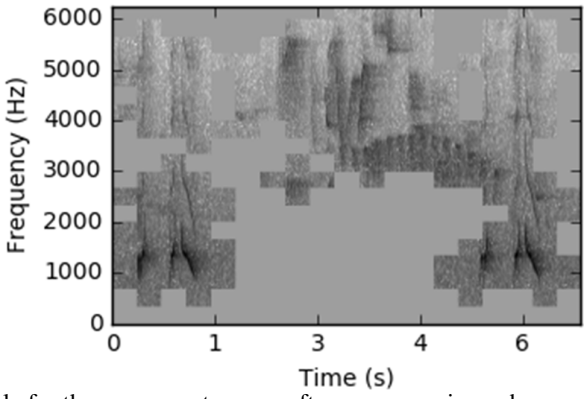


**Fig. 3.** Example for the same spectrogram after preprocessing, when mostly-zero cells were removed (Run 1, 3, 4).

## 4 Deep learning based classification

For classifying the bird songs we used convolutional neural networks. The resulting 200×310 arrays of the spectograms after data preparation were fed into the convolutional neural network and was treated like grayscale images. We used two different CNN architectures: the first one was inspired by the winner architecture of 2012 ImageNet competition [14] (AlexNet [15]), the second convolutional neural network was inspired by audio recognition systems.

In the first type of neural network we modified the shape of the input and the convolutional layers of AlexNet. We also added batch normalization layers before the max-pooling layers. Experiments show that with batch normalization significantly better accuracy can be achieved on MNIST and ImageNet datasets with faster convergence [16]. This network is referred to as CNN-Bird-1.

The second type of neural network used a simpler architecture, it consisted four convolutional layers and the fully connected layers had less neurons. We used ReLUs as activation functions [17] and batch normalization layers were also applied. The number of parameters of the second network was much less, thus the network was learning faster. This network is referred to as CNN-Bird-2. The proposed networks are shown in Figure 4.

To train the model we used RMSProp adaptive algorithm as optimizer [18] with mini-batch learning. Early stopping with a patience of 100 epochs was applied.

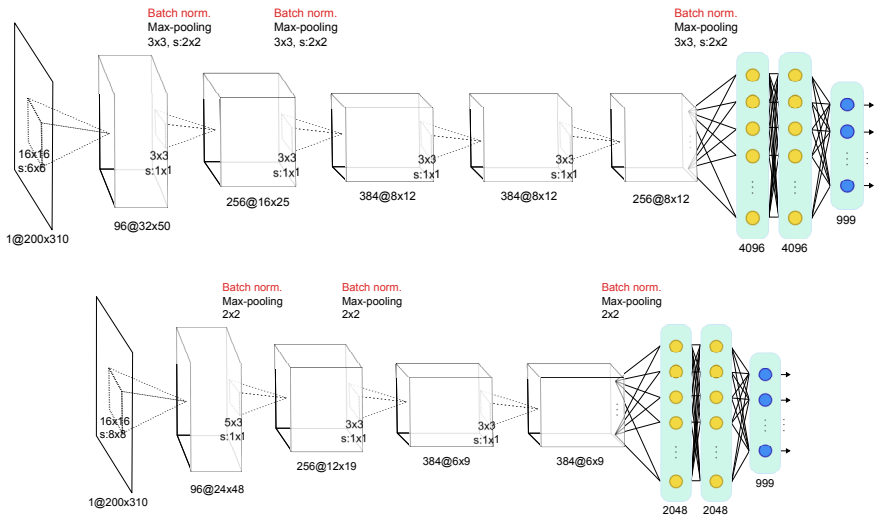


**Fig. 4.** CNN-Bird-1 (above) and CNN-Bird-2 (below) convolutional neural networks for bird species identification based on the spectogram of bird song recordings. (A@BxC refers to A number of planes with size BxC. The DxD refers to the kernel size.)

Due to the fact that we split each audio file to smaller pieces (that were fed to the CNNs) if a recording was longer than five seconds we had to combine the multiple predictions of the neural network. In case of 'BME TMIT Run 1, 2, 3' we simply calculated the mean of the classification results. In case of 'BME TMIT Run 4' we used a custom calculation method for submitting the classification results: if the recording was split into more parts than we calculated the variance of the CNN's outputs of each predicted class throughout the 5 seconds long split parts. Next the six predictions with the highest variance were selected. The predicted bird species came from the mean of these predictions.

## 5 Evaluation

The hardware we used for training were a NVidia GTX 970 (4 GB) and a NVidia Titan X (12 GB) GPU card hosted in two i7 servers with 32 GB RAM. Ubuntu 14.04 with Cuda 7.5 and cuDNN 4.0 was used as general software architecture. For data preparation, training and evaluating deep neural networks the Keras [19] framework with Theano [20] backend was used. For calculating area under the precision-recall curve (AUROC) values we used the sklearn Python package. The differences in data preparation (see Section 3), in the architectures, in the combination method of the predictions (see Section 4) and the epochs needed to reached the maximum AUROC measured on validation set are summarized in Table 1. The AUROC values throughout the training of Run 1, 2, 3 and 4 are shown in Figure 5. The database sizes, the data preparation and CNN training times are shown in Table 2.

**Table 1.** The experimental setup of the submitted runs.

| Run | Data preparation | CNN architecture | Combination of the predictions | Epochs |
|---|---|---|---|---|
| BME TMIT Run 1 | 'Zero' parts removed | CNN-Bird-1 | Mean | 124 |
| BME TMIT Run 2 | 'Zero' parts not removed | CNN-Bird-1 | Mean | 121 |
| BME TMIT Run 3 | 'Zero' parts removed | CNN-Bird-2 | Mean | 104 |
| BME TMIT Run 4 | 'Zero' parts removed | CNN-Bird-2 | Mean of top six predictions with highest variance | 104 |

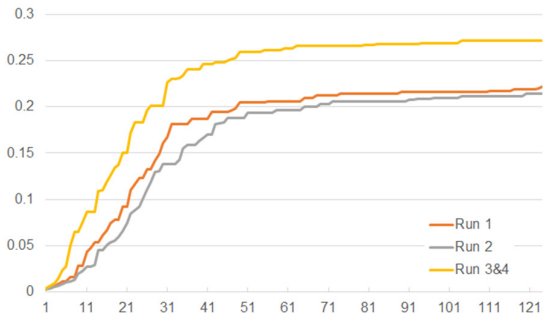

**Fig. 5.** The value of AUROC throughout the training for Run 1, Run 2 and Run 3&4.

We investigated the accuracy of the model on a separated test set. The least average precisions (AP) were achieved by *Ochre-rumped Antbird* (AP=0.00067), *Santa Marta Antpitta* (AP=0.00136) and *Rufous-breasted Leaftosser* (AP=0.0015) bird calls. *Yellow-eared Parrot* (AP=0.692), *Lesser Woodcreeper* (AP=0.796) and *Spillmann's Tapaculo* (AP=0.899) species scored the best in the test. Furthermore, a lot of bird calls were misclassified to *Orange-billed Nightingale-Thrush* (AP=0.229). Analyzing the waveforms and the spectrograms of these species we couldn't find any particular feature. Hence we suppose the significant difference in AP and the misclassification are generally caused by some shortcomings of the proposed CNN architectures.

The MAP (Mean Average Precision) values of our submission in the official results are shown in Table 3. The first MAP value corresponds to the recordings in which there was a dominant singing bird in the foreground with some other ones in the background. The second MAP is for recordings with only one singing bird. And the third MAP value is for the soundscape audio, that was not targeting specific species and these recordings might have contained an arbitrary number of singing birds. The results show that the smaller convolutional neural network (CNN-Bird-2; Run 3 and 4), which was faster to train performed similarly as the bigger CNN. However, the gain in AUROC on the validation database is not reflected in the official results (MAP values) in case of Run 3 and 4. Moreover the difference in the combination methods of Run 3 and 4 could be measured on the validation set, but in the official results Run 4 didn't outperform our other approaches. According to the official results we resulted the 4th place out of 6. It should be noted that we joined the competition only on April and we had no previous experience with bird call recognition.

**Table 2.** Database sizes, data preparation (left) and CNN training (right) times.

| Method | Preparation time [minutes] | Size [GB] |
|---|---|---|
| 'Zero' parts removed | 103 | 41.8 |
| 'Zero' parts not removed | 123 | 48.76 |

| CNN architecture | Training time [hours] |
|---|---|
| CNN-Bird-1 | 37.8 |
| CNN-Bird-2 | 48.8 |
| CNN-Bird-3 & 4 | 27.6 |

**Table 3.** Official results: MAP values of our submissions.

| Run | MAP (with background species) | MAP (only main species) | MAP ('soundscape' recordings) |
|---|---|---|---|
| BME TMIT Run 1 | 0.323 | 0.407 | 0.054 |
| BME TMIT Run 2 | 0.338 | 0.426 | 0.053 |
| BME TMIT Run 3 | 0.337 | 0.426 | 0.059 |
| BME TMIT Run 4 | 0.335 | 0.424 | 0.053 |

## 6 Conclusions

In this paper a deep learning based approach was presented for large-scale bird species identification based on their songs. In the data preparation process the spectogram for every recording was calculated and the irrelevant parts were removed. The resulting spectogram was sliced into five seconds long segments, these segments were used as input of the CNN. Two different types of CNNs were used that achieved about the same accuracy, while one of them had much less parameters. At the final step the predictions of the slices were combined. The results show that the deep learning based approach is well suitable for the task, however fine-tuning is necessary to reach better accuracy, like separating time and frequency in the CNN feature learning part and applying recurrent architectures, e.g. Long Short-Term Memory (LSTM).

## Acknowledgement

Bálint Pál Tóth gratefully acknowledges the support of NVIDIA Corporation with the donation of an NVidia Titan X GPU used for his research.

## References

1. Frommolt, K. H., Bardeli, R., and Clausen, M. Computational bioacoustics for assessing biodiversity. In Proceedings of the International Expert meeting on IT-based detection of bioacoustical patterns, BfN-Skripten, No. 234. (2008)
2. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W. P., Champ, J., Planqué, R., Palazzo, S., Müller, H., LifeCLEF 2016: multimedia life species identification challenges, Proceedings of CLEF 2016 (2016)
3. Goëau, H., Glotin, H., Planqué, R., Vellinga, W. P., Joly, A., LifeCLEF Bird Identification Task 2016, CLEF working notes 2016 (2016)
4. Xeno-canto Foundation. (2012). Xeno-canto: Sharing bird sounds from around the world.
5. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. and Kingsbury, B., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6), pp.82-97. (2012)
6. Graves, A., Mohamed, A.R., and Hinton, G. Speech recognition with deep recurrent neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 6645-6649. (2013)
7. Sainath, T.N., Vinyals, O., Senior, A. and Sak, H., Convolutional, long short-term memory, fully connected deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015), pp. 4580-4584. (2015)
8. McIlraith, A.L. and Card, H.C., Bird song identification using artificial neural networks and statistical analysis. In IEEE Canadian Conference on Electrical and Computer Engineering, Engineering Innovation: Voyage of Discovery, Vol. 1, pp. 63-66. (1997)
9. Cai, J., Ee, D., Pham, B., Roe, P. and Zhang, J., December. Sensor network for the monitoring of ecosystem: Bird species recognition. In IEEE 3rd International Conference on Intelligent Sensors, Sensor Networks and Information (ISSNIP 2007), pp. 293-298. (2007)

10. Leng, Y.R. and Dat, T.H., December. Multi-label bird classification using an ensemble classifier with simple features. Asia Pacific Signal and Information Processing Association (APSIPA), pp. 1-5. (2014)
11. Neal, L., Briggs, F., Raich, R. and Fern, X.Z., Time-frequency segmentation of bird song in noisy acoustic environments. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), pp. 2012-2015. (2011)
12. Lasseck, Mario. Improved automatic bird identification through decision tree based feature selection and bagging. In Working notes of CLEF 2015 Conference. (2015)
13. Lasseck, Mario. Large-scale Identification of Birds in Audio Recordings. In Working Notes CLEF, pp. 643-653. (2014)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Hunag, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., & Fei-Fei, L. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252. (2015)
15. Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097-1105. (2012)
16. Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015)
17. V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In Proc. 27th International Conference on Machine Learning, pp. 807-814. (2010)
18. Tieleman, T., & Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 4, 2. (2012)
19. Chollet, F. Keras: Theano-based deep learning library. Code: https://github.com/fchollet. Documentation: http://keras.io. (2015)
20. The Theano Development. A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688 (2016)
21. Hochreiter, S., & Schmidhuber, J. Long short-term memory. Neural computation, 9(8), 1735-1780. (1997)