# DIVERGENCE BASED WEIGHTING FOR INFORMATION CHANNELS IN DEEP CONVOLUTIONAL NEURAL NETWORKS FOR BIRD AUDIO DETECTION

*Cemre Zor[b,a], Muhammad Awais[a], Josef Kittler[a], Miroslaw Bober[a], Sameed Husain[a], Qiuqiang Kong[a], Christian Kroos[a]*

[a]Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, United Kingdom, GU2 7XH

[b]Centre for Medical Image Computing (CMIC)
University College London, United Kingdom, WC1E 7JE

## ABSTRACT

In this paper, we address the problem of bird audio detection and propose a new convolutional neural network architecture together with a divergence based information channel weighing strategy in order to achieve improved state-of-the-art performance and faster convergence. The effectiveness of the methodology is shown on the Bird Audio Detection Challenge 2018 (Detection and Classification of Acoustic Scenes and Events Challenge, Task 3) development data set.

***Index Terms—*** Deep convolutional neural networks, bird audio detection, KL divergence, bulbul, layer weighting, layer initialisation

## 1. INTRODUCTION

The task of bird audio detection (BAD) is concerned with the labelling of an input sound recording for the presence or absence of a bird sound. A solution to the detection problem is expected to be useful to systems targeting wildlife monitoring, which measure the density of birds, analyse bird migrations, carry out species classification, etc. The research in this application domain is encouraged and supported by the BAD Challenge [1, 2], which provides a collection of diverse data sets every year to promote the development of classification methods including Deep Neural Networks (DNNs).

In this study, we initially introduce a new DNN architecture addressing the BAD problem, named BirdNet, by extending the state-of-the-art system, bulbul [3], that has won the BAD Challenge 2017 [1]. The learning of the network parameters is controlled by a novel weighting strategy for the information channels of the proposed architecture. We show that this innovation leads to a faster convergence of the network. The weighing methodology is based on a measure of information divergence between the positive and negative pattern distributions observable at different convolutional layer channels. We name this version of the BirdNet, aided by contextual channel weights, BirdNet-D.

Using multiple random splits by maintaining 80% / 20% training / test ratio on the development data set of BAD Challenge 2018, BirdNet is demonstrated to outperform the bulbul system by 6.55%. Furthermore, by employing BirdNet-D, we show that it is possible to obtain a better accuracy than that of BirdNet in every epoch and converge to optimum performance much earlier. Moreover, it is also possible to achieve a slight improvement in the performance.

This paper is organised as follows. In Section 2, we present the background information to the BAD problem. Section 3 provides details of the proposed methodology. Section 4 discusses the experimental validation of the proposed method. The conclusions are drawn in Section 5.

## 2. RELATION TO PRIOR WORK

In this section, a short review of the prior art related to the topics this study is presented, with a focus on BAD methodologies and weight initialisations for DNN layers.

### 2.1. BAD Approaches

Early research in BAD explored techniques such as Gaussian mixture models (GMMs), hidden Markov models (HMMs) and random forests [4, 5, 6, 7, 1] while using mel spectrum or mel frequency coefficients as features. More recently, DNN approaches, such as convolutional neural networks (CNNs), have been introduced for solving audio tagging problems including BAD [8]. Among these, the winner system of the BAD 2017 Challenge, bulbul [3], is an example.

Note that the BAD problem falls into the weakly-labelled data learning category, as the only annotation available is a label indicating the presence or the absence of a bird in a

This work was partially supported by the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/DSTL/MURI project EP/R018456/1.

recording, without the specification of the exact time of the bird sound occurrence. In our work, we use an extended version of the bulbul system architecture, augmented by further convolutional and pooling layers for better performance, and incorporating a weighting layer for faster convergence.

### 2.2. Weighting of CNN layers

Various methodologies for the initialisation of convolution and fully connected layer weights of CNNs [9, 10] are described in the literature. The aim of most of these weighting approaches is to obtain similar gradient scales across all layers. Recently in [11], by adopting a different point of view, a contextual information based scaling was proposed for addressing the problem of image retrieval under triplet loss. Specifically, the outputs of the convolution layers were weighted in a separate multiplication layer. Based on information theory, the proposed weighing gauges the importance of each convolved sub-region of the input image. For this purpose, Kullback-Leibler divergence (KL divergence) [12] was employed.

Although KL divergence is commonly used in the literature, the problems that limit its use cases are reported in [13, 14]. In this paper, we adopt a symmetric version of the KL divergence, overcoming some of its shortcomings, and apply the divergence based channel weighting strategy to a new network for audio data classification, using logistic regression on the deep features extracted.

## 3. METHODOLOGY

This section provides the details of the BirdNet and BirdNet-D systems, the features they take as input, and the proposed weighting strategy.

### 3.1. Feature Extraction

Each audio clip in the BAD Challenge data set consists of a 10 second single-channel recording sampled at the 44.1 kHz sampling rate, nonequantised to 16-bit resolution and stored as a PCM file. The features of these clips are extracted in a similar manner as in the bulbul system [3]: The input signals are downsampled to 22.05 kHz and short-term Fourier transform spectra are calculated with a window size of 1024 samples and hop size of 220 samples. From the spectra, log-mel frequency coefficients are computed using a filter bank of 80 triangular mel filters. The coefficients are normalised by subtracting the mean and dividing by the standard deviation per frequency band. The size of the resulting feature matrix for an input clip is therefore 80x1000, where the dimensions represent frequency and time, respectively.

| Input | 80 x 1000 x 1 |
|---|---|
| Conv (5x5) | 80 x 1000 x 32 |
| Pool (2x2) | 40 x 500 x 32 |
| Conv (3x3) | 40 x 500 x 64 |
| Pool (2x2) | 20 x 250 x 64 |
| Conv (3x3) | 20 x 250 x 128 |
| Pool (2x2) | 10 x 125 x 128 |
| Conv (3x3) | 10 x 125 x 128 |
| Pool_t (1x2) | 10 x 62 x 128 |
| Conv_t (1x3) | 10 x 62 x 128 |
| Pool_t (1x3) | 10 x 20 x 128 |
| Conv_t (1x3) | 10 x 20 x 128 |
| Pool_t (1x2) | 10 x 10 x 128 |
| Conv_t (1x3) | 10 x 10 x 128 |
| Pool_t (1x10) | 10 x 1 x 128 |
| Pool_f (10x1) | 1 x 1 x 128 |
| Dropout (0.5) | |
| Fully connected | 256 |
| Dropout (0.5) | |
| Fully connected | 64 |
| Dropout (0.5) | |
| Fully connected | 1 |

(a)

| Input | 80 x 1000 x 1 |
|---|---|
| Conv (5x5) | 80 x 1000 x 32 |
| Pool (2x2) | 40 x 500 x 32 |
| Conv (3x3) | 40 x 500 x 64 |
| Pool (2x2) | 20 x 250 x 64 |
| Conv (3x3) | 20 x 250 x 128 |
| Pool (2x2) | 10 x 125 x 128 |
| Conv (3x3) | 10 x 125 x 128 |
| Pool_t (1x2) | 10 x 62 x 128 |
| Conv_t (1x3) | 10 x 62 x 128 |
| Pool_t (1x3) | 10 x 20 x 128 |
| Conv_t (1x3) | 10 x 20 x 128 |
| Pool_t (1x2) | 10 x 10 x 128 |
| Conv_t (1x3) | 10 x 10 x 128 |
| Pool_t (1x10) | 10 x 1 x 128 |
| Weight_f | 10 x 1 x 128 |
| Pool_f (10x1) | 1 x 1 x 128 |
| Dropout(0.5) | |
| Fully connected | 256 |
| Dropout (0.5) | |
| Fully connected | 64 |
| Dropout (0.5) | |
| Fully connected | 1 |

(b)

**Table 1**. Deep neural network architectures of BirdNet (a) and BirdNet-D (b)

### 3.2. Network Architecture

The initial DNN architecture we propose for addressing the BAD task, namely BirdNet, is presented in Table 1-a. This network consists of seven convolutional and three fully connected layers. Each of these layers is followed by rectified linear unit (ReLU) nonlinearity.

The first three convolutional layers act on both the time and the frequency domains and result in 10x125 frequency-time feature maps. After this point, the network preserves the number of channels in frequency and only summarises time by applying a set of rectangular convolutional and pooling layers on the time axis. These operations lead to the representation of 10x10 frequency-time feature maps. The network then summarises the time component by applying a max-pool of filter size 1x10, leaving 10 convolved frequency channels. The rest of the network network has two fully connected layers, each one being associated with a dropout layer using a ratio of 0.5. At the output node, there is a logistic neuron giving the probability of the presence of a bird in the 10 second-long audio clip.

The second architecture, which aims to achieve faster convergence than BirdNet, named BirdNet-D, is depicted in Table 1-b. The main rationale behind this system is that different frequency bands or their convolved representations would be of different importance to the task of bird audio detec-

tion; therefore, it would be beneficial to learn the weights that can act on each output channel to underline their contributions to the prediction. For this purpose, we introduce a new layer which assigns a weight for each of the convolved frequency channel across all feature maps (128 feature maps in our setup), and name it Weight_f. Note that the remaining layers of the BirdNet-D are the same as those of BirdNet (two fully connected layers with dropout), followed by logistic regression for classification.

### 3.3. Weight Initialisation

In order to initialise the Weight_f layer, BirdNet-D requires an initial run of BirdNet on a small portion of the training data. In our experiments, this portion is selected to be equal to a random 1/5 portion of the original training set. After the convergence of BirdNet using the subset, the output of the Pool_f layer (see Table 1-a for details) is recorded for further analysis. Specifically, this layer generates a matrix of size 10x1x128 for each training pattern, where 10 is the number of convolved frequency bands, and 128 is the number of feature maps.

For each one of the 10 frequency channels, Principle Component Analysis (PCA) is applied for reducing the dimensionality of the associated feature maps from 128 to 4. This is so as to extract the highest energy components while keeping the analysis tractable. The distributions of the positive and negative training samples in the 4D feature space for positive are estimated, and the class separability is measured using information divergence.

A common information measure to gauge the similarity of two distributions is the Kullback-Leibler (KL) divergence [12]. If the distributions are identical, or similar, the measure will tend to zero. A high value of the measure would indicate differences and therefore high discrepancy. Let $P(x)$ and $Q(x)$ denote the probability distributions of the input data, $x$, in the positive and the negative classes. By considering one of the distributions as a reference, the KL divergence between $P$ and $Q$ can be measured as

$$D_{KL}(P||Q) = \sum_{x} P(x) \log\left(\frac{P(x)}{Q(x)}\right). \quad (1)$$

It can be observed from Equation 1 that KL divergence is not symmetric, i.e $D_{KL}(P||Q) \neq D_{KL}(Q||L)$: The value of the divergence changes depending on the choice of the reference distribution. In order to avoid this problem, we make use of the symmetric KL measure (KLS), which is formulated as

$$D_{KLS}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P). \quad (2)$$

The KLS values computed for each frequency channel are passed onto BirdNet-D as the starting points of the Weight_f layer: A high KLS value would mean a high separability between the positive and negative classes, and the weighting layer would promote the associated frequency band accordingly. Note that we allow the network to learn and fine-tune these weights during training.

## 4. EXPERIMENTS

In this section, after introducing the data sets used for the experimental analysis, the experimental setup is described and the results obtained are presented.

### 4.1. Data set

In this study, we use the three development data sets provided by BAD Challenge 2018, which are composed of 10 second-long WAV files totalling 100 hours. The audio clips collected from three diverse sources are categorised as: 1) Field recordings around the world, gathered by the FreeSound project 2) Smartphone audio recordings crowd-sourced by users of the bird recognition app, Walblr 3) Remote monitoring recordings collected near Ithaca, NY, USA by the BirdVox project. In our experiments, we carry out 3 random splits of the union of these data sets with 80% / 20% training / test ratio for the performance analysis.

### 4.2. Setup

In addition to bulbul, BirdNet and BirdNet-D architectures, two more variants of the proposed network have been implemented in order to assess the impact of using KLS for assigning the frequency channel weights. For this, the Weight_f layer is re-initialised using: 1) KL divergence 2) Uniform random distribution in the interval [0,1] scaled by a factor of $2/\sqrt{n_c}$, where $n_c$ is the number of frequency channels.

All convolution and fully connected layers of the networks are initialised by Xavier initialisation using 3 independent runs, each time having a different random seed. During training, the learning rate is made to decrease with a constant factor starting from $10^{-3}$ at the first epoch and ending with $10^{-6}$ at epoch number 25. Note that all networks have been observed to have converged to their optima by the time training reaches 25 epochs. Therefore, the results will be presented for the first 25 epochs only, for brevity.

### 4.3. Results

The aim of the first set of experiments is to compare the bulbul and BirdNet architectures using 9 independent runs (3 data splits x 3 Xavier initialisations). The mean error rates obtained by both networks for 25 epochs are presented in Figure 1, from which it can be observed that after the third epoch BirdNet outperforms bulbul, converging to an error rate of 11.04% at 25 epochs, while this rate is equal to 17.59% for bulbul.

Secondly, in Figure 2, we compare BirdNet, BirdNet-D, and the two variants based on using KL divergence and scaled

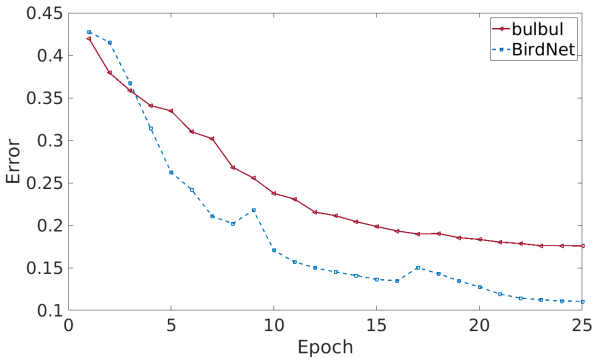


**Fig. 1**. Performance comparison of bulbul and BirdNet architectures

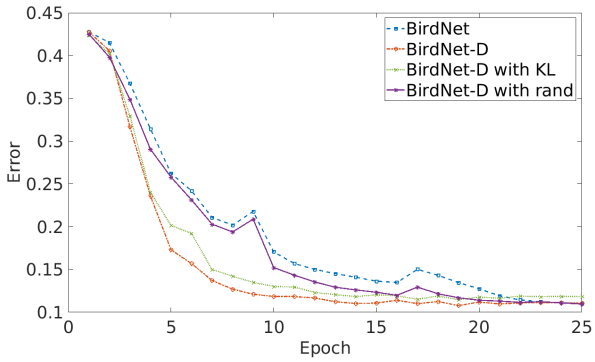


**Fig. 2**. Performance comparison of the architectures BirdNet, BirdNet-D, and the two variations, BirdNet-D with KL and BirdNet-D with rand.

random weights during the initialisation of the Weight_f layer. The variants are named "BirdNet-D with KL" and "BirdNet-D with rand", respectively. Figure 2 shows that the best overall result in terms of performance per epoch are obtained by BirdNet-D after the second epoch. In other words, BirdNet-D achieves the fastest overall convergence. Although the KL variant of the system follows closely, it fails to converge to the minimum exhibited by BirdNet-D. The minimum error rates obtained by all systems including bulbul, and their associated epoch indices are given in Table 2. From Table 2, it can also be seen that the best error rate achieved by BirdNet-D is also better than those of the other systems.

It should be underlined that having a weighting layer for the frequency channels always generates improved results, compared to those of the original system, BirdNet, which does not include the weighting layer in its architecture. This can be visualised by comparing BirdNet against BirdNet-D with random scaled weight initialisations, latter of which reveals higher accuracy for all epochs.

Finally, note that the standard deviations of the errors achieved by the bulbul, BirdNet and BirdNet-D systems at their best are $0.775, 0.538$ and $0.388$ respectively, showing the superiority of the BirdNet-D in terms of stability as well.

| System | Min Error (%) | Best Epoch Index |
|---|---|---|
| bulbul | 17.59 | 25 |
| BirdNet | 11.04 | 25 |
| BirdNet-D | 10.78 | 19 |
| BirdNet-D with KL | 11.46 | 19 |
| BirdNet-D with rand | 10.94 | 25 |

**Table 2**. Minimum error rates (%) and the corresponding epoch indices obtained for all systems

## 5. CONCLUSIONS

In this paper, we proposed a new deep convolutional neural network architecture to address the problem of bird audio detection (BAD). Using the development data set of BAD Challenge 2018 (as part of the Detection and Classification of Acoustic Scenes and Events Challenge), the proposed network, BirdNet, has been shown to achieve better than the state-of-the-art detection performance by 6.55%.

Secondly, we have introduced a weighting layer for the convolved frequency information outputs of the BirdNet system. The motivation behind the weighing approach was to identify frequency components that are more informative for the task of BAD, and use this information to control the training of the detection system. To achieve this goal, we used a divergence based criteria to measure the class separability afforded by each frequency channel and weight its contribution to the final decision. We experimented with the classical Kullback-Leibler divergence and showed that its symmetric version produces better results. The resulting network has been named BirdNet-D, and is shown to exhibit much faster convergence and better accuracy compared to the rest of the networks, including BirdNet.

This study shows the effect of an informed network design on the final performance and the convergence rate of the detection system. For application fields where computationally expensive training in terms of time and resources is unavoidable, the proposed design technique offers a favourable alternative by achieving similar performance in shorter time, or better performance in a fixed time.

## 6. REFERENCES


[1] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[2] D. Stowell, Y. Stylianou, M. Wood, H. Pamuła, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge," *Methods in Ecology and Evolution*, 2018.

[3] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *Signal Processing Conference (EUSIPCO)*, 2017, pp. 1764–1768.

[4] C. Kwan, 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," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, vol. 5, pp. V–289.

[5] P. Jančovič and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 982–936, 2011.

[6] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, vol. 80, pp. 1–9, 2014.

[7] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, pp. e488, 2014.

[8] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," 2016.

[9] Kaiming H., Xiangyu Z., Shaoqing R., and Jian S., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[10] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 13–15 May 2010, vol. 9, pp. 249–256.

[11] S. Husain and M. Bober, "REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval," *submitted to IEEE Trans. Image Processing*, 05 2018.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 03 1951.

[13] J. Kittler and C. Zor, "Delta divergence: A novel decision cognizant measure of classifier incongruence," *IEEE Transactions on Cybernetics*, pp. 1–13, 2018.

[14] J. Kittler, C. Zor, I. Kaloskampis, Y. Hicks, and W. Wang, "Error sensitivity analysis of delta divergence - a novel measure for classifier incongruence detection," *Pattern Recognition*, vol. 77, pp. 30–44, 2018.