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

CVSSP Centre for Vision, Speech and Signal Processing
cmic Centre for Medical Image Computing

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, Guildford, GU2 7XH
[b] Centre for Medical Image Computing (CMIC), University College London, United Kingdom, WC1E 7JE
c.zor@surrey.ac.uk

PROPOSAL

- A new convolutional neural network architecture for the bird audio detection (BAD) problem
- A divergence based information channel weighing strategy to achieve faster convergence and improved state-of-the-art performance

DATA

- Bird Audio Detection Challenge 2018 (Detection and Classification of Acoustic Scenes and Events Challenge, Task 3) development data set
  - The audio clips are collected from 1) Field recordings around the world, gathered by the FreeSound project 2) Crowd-sourced smartphone audio recordings 3) Remote monitoring near Ithaca by the BirdVox project
  - Each clip consists of a 10 second single-channel recording sampled at the 44.1 kHz sampling rate, non-equantised to 16-bit resolution and stored as a PCM file

APPROACH

- We first introduce a new DNN architecture addressing the BAD problem by extending the state-of-the-art system, bulbul [3], that has won the BAD Challenge 2017
  - The proposed architecture is named **BirdNet**
- We then use a novel weighting strategy to learn the weights for the information channels of BirdNet
  - Based on information divergence between the positive and negative pattern distributions observable at different convolutional layer channels
  - This version of the BirdNet, aided by contextual channel weights, is named **BirdNet-D**

CONTRIBUTIONS

- BirdNet is demonstrated to outperform the bulbul system by 6.55%.
- By employing BirdNet-D, it is possible to obtain a better accuracy than that of BirdNet in every epoch and converge to optimum performance much earlier
  - It is also possible to achieve a slight improvement in the performance

METHODOLOGY

1) Feature Extraction

- The input signals are downsampled to 22.05 kHz and short-term Fourier transform spectra are calculated with a window size of 1024 and hop size of 220 samples
- From the spectra, logmel frequency coefficients are computed via 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 80x1000 (frequency by time)

2) Network Architectures

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

<u>BirdNet-D aims faster convergence than BirdNet</u>

- Different frequency bands or their convolved representations are of different importance to the task of bird audio detection
- Learning the weights that can act on each output channel is beneficial to underline their contributions to the prediction
- In BirdNet-D, we introduce a new layer (**Weight_f**) which assigns a weight for each convolved frequency channel across all feature maps (128 maps)

3) Weight Initialisation

- To initialise the Weight_f layer, BirdNet-D requires BirdNet to run on a small portion of the training data
  - We select this to be 1/5th of the original training set
- After the convergence of BirdNet using the subset, the output of the Pool_f layer is recorded for further analysis
  - A matrix of size 10x1x128 is generated for each training pattern, where 10 is no. convolved frequency bands, and 128 is no. feature maps
- For each frequency channel, PCA is applied for reducing the dimensionality of the feature maps from 128 to 4
  - Extracting the highest energy components while keeping the analysis tractable
- The distributions of the positive and negative class training samples in the 4D feature space are estimated, and the class separability is measured using information divergence
  - Using symmetric Kullback-Leibler (KLS) divergence
  - Let $P(x)$ and $Q(x)$ denote the probability distributions of the input data x in the positive and the negative classes

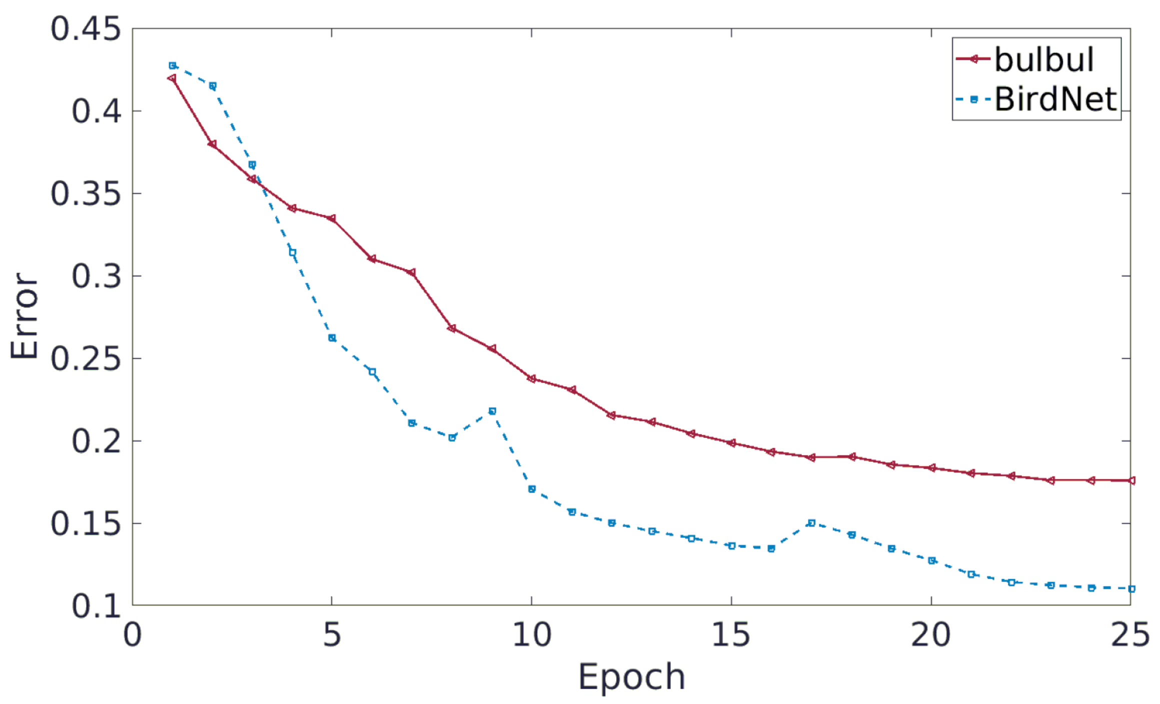
$$D_{KL}(P\|Q) = \sum_{x} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

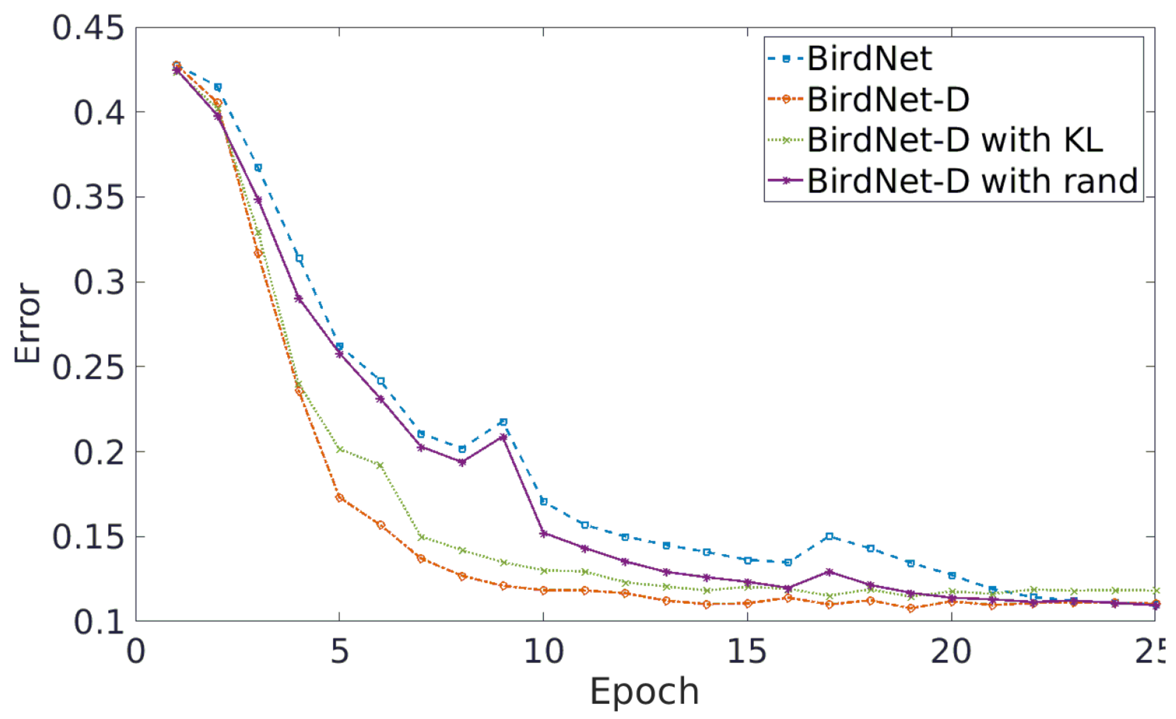
$$D_{KLS}(P\|Q) = D_{KL}(P\|Q) + D_{KL}(Q\|P)$$

  - If the distributions are similar, the measure tends to zero. A high value would indicate high discrepancy.
- The KLS values computed for each frequency channel are passed onto BirdNet-D as the starting points of the Weight_f layer
  - High KLS value → High separability between the positive and negative classes, high weighting for associated frequency bands

EXPERIMENTS

- Comparisons are made between bulbul, BirdNet, BirdNet-D, and two more variants of BirdNet-D where Weight_f layer is initialized using 1) KL divergence 2) Uniform random distr. scaled by $2/\sqrt{n_c}$





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

Minimum error rates (%) and the corresponding epoch indices obtained for all systems

RESULTS

- ✓ BirdNet achieves better than the state-of-the-art detection performance by 6.55%.
- ✓ BirdNet-D identifies frequencies that are more informative for the task of BAD, and initialise their weights accordingly. It exhibits much faster convergence and better accuracy compared to the rest of the networks, including BirdNet.
- ✓ KLS class separability measure is most effective in setting the layer weights.
- ✓ We show the effect of informed network design on the performance and the convergence rate of a detection system.