Cocktail Party Problem for Bird Sounds

Jason Chou
Dep't of Physics, Stanford University
jasonhc@stanford.edu

Chun-Hao To
Dep't of Physics, Stanford University
chto@stanford.edu

Abstract

More often than not birds move in groups and make sounds together. Being able to recognize individual bird sounds in the natural mixture is thus of great practical values for bird lovers and researchers. We train Convolutional Neural Networks with different schemes on spectrograms of multi-species bird sounds to tackle this multi-class multi-label classification problem. Variations of ResNet combined with binary relevance method prove to be the most powerful, while changing loss functions and/or details of network architectures does not help improve the performance significantly. For a total of 10 species with 1–5 species present simultaneously in one segment of sound, our models achieve F1 score, precision and recall of 0.83, 0.90 and 0.85 respectively, considerably outperforming similar multi-label classification tasks on photo-images.

1 Introduction

The field of bird song recognition has recently seen thriving developments thanks to the rapid growth of the database for bird songs and calls. Not only are there regular workshops held internationally, but also downloadable mobile apps [1], both showing promising identification capabilities of current implementations. Despite outperforming (non-expert) humans in recognizing single bird species by sounds, all existing implementations flounder when it faces noisy background and/or multiple birds in the recordings – they fail when more than one species are to be singly identified within one recording.

Therefore, we endeavor to contribute to the community, the power of separating distinct species within the cacophonies, an avian analogue of solving the cocktail party problem. The success of the project can potentially benefit bird lovers and researchers by rendering a practical software that does not require recordings to be pre-processed or cleaned – you record, it outputs answers right away. The project is thus a multi-class multi-label classification problem for recognizing constituents from mixtures of classes. Although we call it cocktail party problem, we do not aim at "extracting" the sound components, but rather, only at being able to identify the individual species.

2 Related work

To our knowledge, the latest attempt for considerably similar problem is the work due to Briggs *et al.* (and references therein) [3] who use traditional machine learning algorithms to tackle multi-class multi-label classification problem of bird sounds with < 13 species, where on average $\approx$2 species appear simultaneously in a single recording. Multi-instance multi-label learning combined with k-nearest neighbors (MIML-kNN) and ensemble of classifier chains with random forest are reported to have the best performance. The Hamming loss achieves $\lesssim$0.05, which is equivalent to $\gtrsim$0.95 in accuracy. We should note, however, that since most ground-truth labels are zeros in an example (on average 2 non-zero labels for total of >10 classes), a non-informative classifier that only predicts zeros can easily achieve accuracy of 0.85. Furthermore, because the datasets used in the aforementioned works are highly unbalanced in species (some species appear more frequently), we consider it inappropriate to interpret the reported metrics as synonyms for performance, and thus we will not juxtapose our results with these models.

Indeed, multi-label classification is more often compared against less-biased metrics such as F1 score [15]. In the broader field of multi-label classification of images, the state-of-the-art results from deep learning-based algorithms can attain F1 score $\sim$ 0.6 (e.g. [16]) for >100 classes. Nevertheless, these studies often exploit either the dependencies

*Implementation of this project, in PyTorch [2], can be found at https://github.com/Mipanox/Bird_cocktail

Table 1. Dataset

Species	# of recordings	Total length (hr)	# processed spectrograms	RAW source
Acorn Woodpecker	154	1.24	2500	xeno-canto
American Robin	290	4.31	4317	xeno-canto
American Crow	201	2.44	6045	xeno-canto
American Goldfinch	176	2.23	4156	xeno-canto
Bewick's Wren	294	3.37	3109	xeno-canto
Fox Sparrow	294	4.80	5303	xeno-canto
Hermit Thrush	265	4.73	4318	xeno-canto
Song Sparrow	290	4.11	4737	xeno-canto
Spotted Towhee	295	3.31	3478	xeno-canto
White-throated Spa.	300	7.14	3332	xeno-canto

between the labels [17] (which appear concurrently) or the fact that objects in images have definite shapes [14], both of which are not applicable to our case. In other words, apart from the traditional approaches which are not directly comparable, we believe that our project is a novel venture.

3 Dataset and Features

Table 1 summarizes our data. We use 98%/1%/1% for train/val/test split. We obtain `mp3` recordings of 10 common local "loud" bird species from the xeno-canto database [12], each capturing bird songs, calls or other representative sounds (e.g. woodpecker's drumming). We use magnitude spectrograms as inputs for models. This is done by first segmenting recordings into 3-sec chunks, then applying Short-Time Fourier Transform (SFTF) to convert them into spectrograms – a time versus frequency plot, sampled on a grey-scale 2D image of dimensions 128×192.

3.1 Pre-processing and Augmentation

We follow the heuristics in Kahl *et al.*'s work on single-label classification of bird sounds [10] for separating signals and noises. The algorithm does median blur, median threshold, spot removal and morphological closing, based on which it judges whether there is strong-enough signal contained in one spectrogram. Figure 1 illustrates the idea.

After this separation, we synthesize our own multi-species spectrograms from the un-processed signal and noise spectrograms. $1 - 5$[1] spectrograms are randomly chosen and weighted, then superposed together with randomly chosen noise sample, to form a single input for model. In this way, mathematics ensures almost countless possible combinations,

[1]For usual circumstances in the field, $\leqslant 5$ species are present simultaneously. Confirmed by realistic multi-species recordings such as the HJA Birdsong dataset [4]

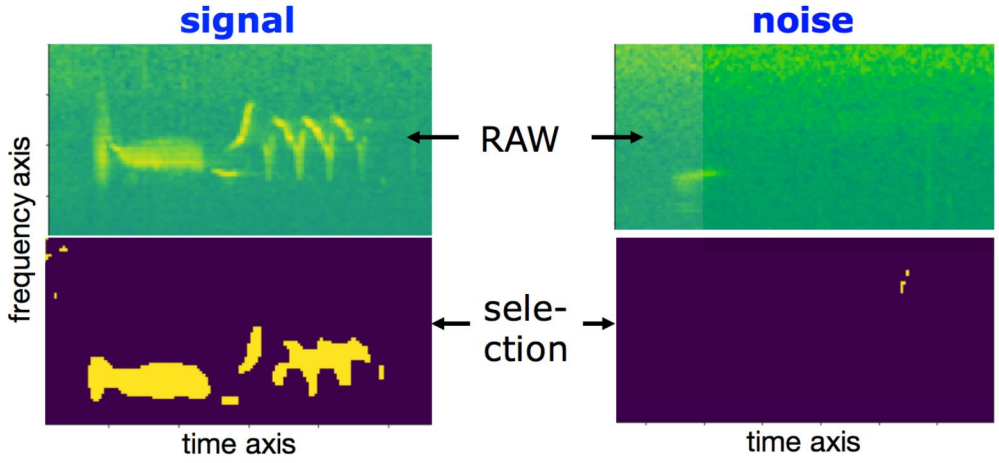

Figure 1 Example spectrograms of signal and background noise . If no significant coherent structures found in heuristics, the spectrogram is categorized as noise (bottom). *(Colors for visualization only)*

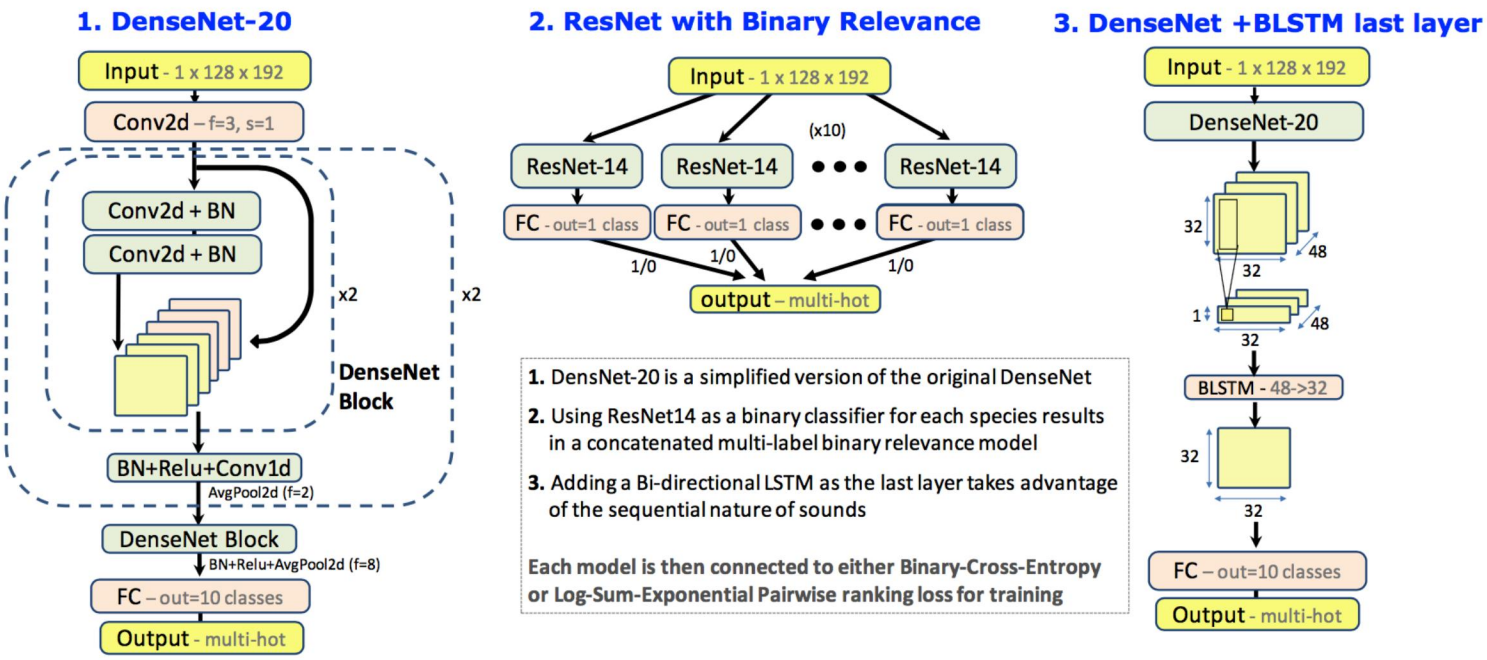

Figure 2 Our representative models are combinations of (simplified) ResNet, DenseNet, and BLSTM.

which means we can have arbitrary augmentation. Finally, a random cropping is optionally done to further augment the data.

4 Methods

Due to the similarity of our scenario to image classification tasks, we base our models on the state-of-the-art ResNet [7], DenseNet [8], among others [13, 5, 9]. The simplest way of addressing multi-label classification task is to replace output layer with a sigmoid followed by global thresholding, i.e. a multi-hot vector output. Contrarily, since the appearance of each species is not correlated by construction[2], each species can be treated independently, decomposing the problem into 10 parallel binary classification problem. This is known as binary relevance [18]. Alternatively, it might be tempting to take advantage of the sequential nature of sounds, by adding BLSTM layers before or after feature-extracting convolution blocks. Figure 2 depicts the structures of these three ideas.

In particular, ResNet accelerates training by feeding activations of former layers into later ones, which helps the network learn "simpler" functions (e.g. identities) more easily [7]. DenseNet, on the other hand, stacks these former layers with later layers, ending up with a more complex and densely connected network while not losing the connectivity between farther layers. Convolutions are basic building blocks for both networks. For learnable convolution operations, the model will learn the abstract features in an image that in some way constitute the representation of the "classes" we are interested in classifying.

Because a spectrogram is essentially a sequence of vectors that preserve temporal correlations of sounds, a bi-directional long-short-term-memory (BLSTM) can be utilized to "memorize" (or "forget") these correlations, in both directions of time. The gates in these layers are to be trained to realize how long and what correlations are critical for telling one class of sounds (spectrograms) from the others. Therefore, as long as the features learned from convolution blocks include useful temporal information, BLSTM should be helpful.

4.1 Loss Function

Two classes of loss functions are employed[3].

[2]Our synthesis guarantees independence between species.

[3]We follow the convention that y and $\hat{y}$ denote ground-truth target and output labels respectively. Also, m and N are the size of the batch and vector (in this case 10) respectively.

- Binary-cross-entropy (BCE) loss:

$$J := \frac{1}{m}\sum_{m'=1}^{m}\frac{1}{N}\sum_{n=1}^{N} y_n \cdot \log \hat{y}_n + (1-y_n)\cdot \log(1-\hat{y}_n) \tag{1}$$

  It assigns a large loss when y_n and $\hat{y}_n$ differ a lot (e.g. 1 v.s. 0) and vice versa.

- Log-sum-exponential pairwise ranking loss [6, 11]:

$$J := \frac{1}{m}\sum_{m'=1}^{m} \log\left(1 + \sum_{u}^{c_+}\sum_{v}^{c_-} \exp \hat{y}_v - \hat{y}_u\right) \tag{2}$$

  where c_+ (positive label), c_- (negative label) are label index of each $N-$long output where ground-truth is 1 and 0 respectively. For a given pair of indices, if the model outputs a higher $\hat{y}$ for a negative label than for a positive label, the loss penalizes the model by giving an exponential loss of their difference. In contrast, if the "ranking" between the indices is correct, the exponent is negative, thus a small contribution.

5 Results

In figure 3, we demonstrate that the single-label problem can be solved almost perfectly by a single ResNet14 with L2 regularization. After ten epochs of training, we are able to obtain accuracy of 94.2% on validation set. This result shows that our neural networks can indeed recognize the sounds of individual birds without other birds singing at the same time.

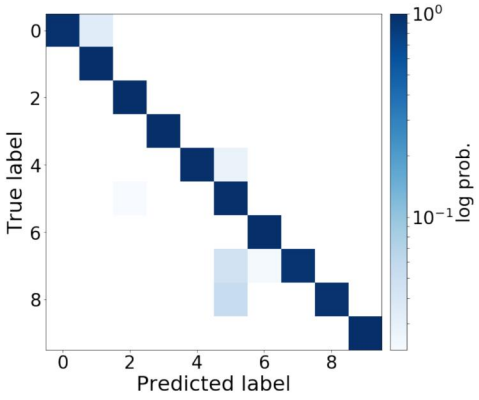

Figure 3 Confusion matrix for single-label classification. True label corresponds to the species of bird we put in, and the predicted label corresponds to the species of bird predicted by the ResNet14.

For multi-label classification problem, we summarize the result in Table 2. All networks are trained according to the following steps. First, the data is separated into mini-batches with batch size 32. The batch size is chosen in compromise between memory usage and extra noise injected by this procedure. We then find the best learning rate and learning rate decay parameters to make sure the losses are steadily decreasing. Finally, if there is a strong sign of overfitting, we add an L2 regularization. If L2 regularization does not work, early stopping is used to stop training the network.

To determine the metric to fairly evaluate the performance of the model, we need to find a metric that is less sensitive to the threshold beyond/below which a given output label is converted to 1/0. F1 score is hence chosen as the primary metric, which is the harmonic mean of precision and recall. In table 2, we find that using ResNet with Binary Relevance model performs the best among all models we have tried. One of the possible explanation for this is the fact that, here each ResNet only focuses on one bird species, while in other models the networks have to take all 10 different species into account at the same time. The decomposition into relative simple tasks helps the neural network learn faster and better. Besides, by comparing ResNet with LSEP loss and ResNet with BCE loss, we find that, though mildly, LSEP loss improves the performance.

Another noticeable thing we find is that adding BLSTM does not improve the performance, contradicting our intuition that BLSTM excels at classifying sequential data. Our hypothesis is that because the length as well as the pattern of bird sounds varies across distinct recordings. Furthermore, since there is no correlation between species, our multi-species spectrograms involving multiple patterns which complicates the task on finding the coherence of sound along time axes.

Table 2 Summary of F1 score, precision, and recall of each neural network. The best values among all neural networks are highlighted in red.

	Loss	Data	F1	Precision	Recall	Regularization
ResNet (14L) + Binary Relevance	BCE†	Training Set	0.851	0.937	0.815	None
		Validation Set	**0.831**	**0.904**	0.804	
DenseNet (20L)	LSEP††	Training Set	0.782	0.854	0.721	Early Stopping
		Validation Set	0.735	0.856	0.686	
DenseNet (20L) + LSTM	LSEP	Training Set	0.671	0.677	0.620	L2
		Validation Set	0.654	0.651	0.579	
ResNet (14L)	LSEP	Training Set	0.787	0.833	0.747	L2
		Validation Set	0.793	0.830	0.759	
ResNet (14L)	BCE	Training Set	0.800	0.729	0.873	L2
		Validation Set	0.770	0.741	**0.854**	
DenseNet (20L)	BCE	Training Set	0.756	0.810	0.724	None
		Validation Set	0.703	0.734	0.729	
InceptionNet[7]	BCE	Training Set	0.491	0.745	0.454	None
		Validation Set	0.360	0.419	0.351	

*F1 score is the harmonic mean of Precision and Recall
† BCE: Binary Cross-Entropy Loss
†† LSEP: Log-sum-exponential Pairwise Ranking Loss

6 Discussion and Conclusion

1. Simplified ResNet is more-than-sufficient for single-label classifying 10 species
2. Binary relevance method is conceptually more intuitive and tops most evaluation metrics in multi-label task by treating every species independently
3. Log-sum-exponential pairwise ranking loss is most useful when learning multi-label task for single network where labels are not treated independently
4. Exploiting sequential nature of bird sounds with additional BLSTM layer does not hurt or improve noticeably the performance of multi-label classification
5. Variations of ResNet are preferred due to their lower memory consumption and faster training

7 Future Work

Upon acquiring more computing resources, we wish to extend the datasets to 100+ species, which would give us the most realistic situation to test the validity of our approaches (on Stanford campus for example, there are roughly 100 year-round species.) We are optimistic on scaling ResNet with Binary Relevance model to 100 species dataset. Since all the ResNets are computed in parallel, we expect the computation time to grow linearly with respect to number of species in the dataset.

8 Acknowledgement

We are grateful to Amazon Web Services and the teaching staff for their generous support of computing resources and guidelines for project as well as the extremely helpful example codes and tutorials.

9 Contributions

Both authors share equal contributions in designing and developing various models. While Jason Chou (JC) was more interested in literature research, which manifests as coming up with new ideas, Chun-Hao To (CHT) played a larger role in fine-tuning each model, reducing variance and obtaining ever-improving performance. JC was responsible for acquisition of the data, pre-processing, and maintaining the Github repository. Experiments were mainly summarized and rendered in nice formats by CHT.

References

[1] e.g. BirdCLEF, EADM, BirdGenie, Bird Song Id USA; *and proceedings/publications therein*.

[2] Official PyTorch tutorials, documentation, and source code. www.xeno-canto.org. Accessed: Feb 22, 2018.

[3] F. Briggs, X. Z. Fern, and J. Irvine. Multi-Label Classifier Chains for Bird Sound. *arXiv e-prints*, April 2013.

[4] Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z. Fern, Raviv Raich, Sarah J. K. Hadley, Adam S. Hadley, and Matthew G. Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012.

[5] Remi Cadene. Pretrained models for Pytorch. https://github.com/bamos/densenet.pytorch. Accessed: Mar 12, 2018.

[6] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep Convolutional Ranking for Multilabel Image Annotation. *ArXiv e-prints*, December 2013.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *ArXiv e-prints*, December 2015.

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. *ArXiv e-prints*, August 2016.

[9] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *ArXiv e-prints*, February 2016.

[10] Stefan Kahl, Thomas Wilhelm-Stein, Hussein Hussein, Holger Klinck, Danny Kowerko, Marc Ritter, and Maximilian Eibl. Large-scale bird sound classification using convolutional neural networks. *Working notes of CLEF*, 2017.

[11] Y. Li, Y. Song, and J. Luo. Improving Pairwise Ranking for Multi-label Image Classification. *ArXiv e-prints*, April 2017.

[12] Numerous recordists for all 10 species we used. Xeno-canto database for bird sounds. www.xeno-canto.org. Accessed: Feb 22, 2018.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *ArXiv e-prints*, February 2016.

[14] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1901–1907, 2016.

[15] X.-Z. Wu and Z.-H. Zhou. A Unified View of Multi-Label Performance Measures. *arXiv e-prints*, September 2016.

[16] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang. Learning Deep Latent Spaces for Multi-Label Classification. *arXiv e-prints*, July 2017.

[17] L. Zhang, S.K. Shah, and I.A. Kakadiaris. Hierarchical multi-label classification using fully associative ensemble learning. *Pattern Recognition*, 70:89 – 103, 2017.

[18] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, Nov 2017.