Cocktail Party Problem of Bird Sounds

Jason Chou* and Chun-Hao To**

Department of Physics, Stanford University; *jasonhc / **chto @stanford.edu

Birds often appear in groups

and we wanted to be able to tell which species are present with confidence. While most (if not all) existing apps and projects[1] focus on single-label classification, our goal is to tackle the more challenging while more realistic multi-label multi-class classification problem.

10 "loud" species common in the Bay Area (and campus) were chosen, with the hope that this project will be helpful in identifying local birds!

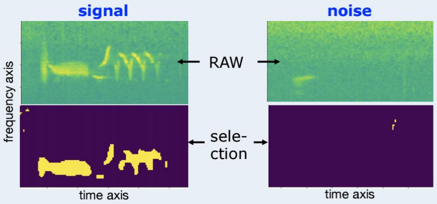
Online birdsound database provides hours of recordings

- We downloaded audios from xeno-canto[1] for the 10 species

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

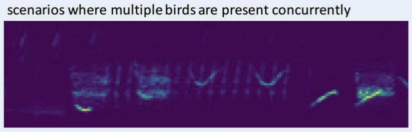
Spectrograms were extracted to represent data

- mp3 ⟶ 3-sec segments ⟶ magnitude spectrograms
- Separating noise: med. blur, spot removal, morpho. closing[2]



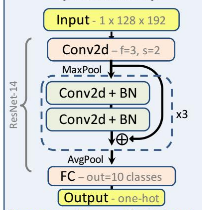
Synthesized multi-species spectrograms + noise = train/val/test datasets

- Randomly select and weight ⩽5 species to simulate scenarios where multiple birds are present concurrently

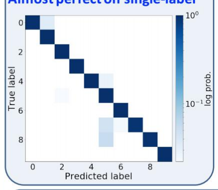


Example synthesized spectrogram of 3 birds singing together

Single-label model
(benchmark)

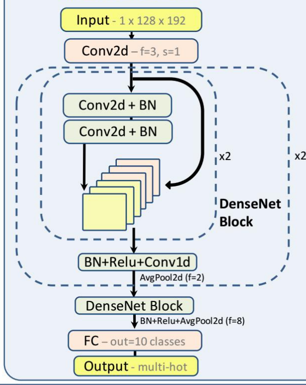


Almost perfect on single-label

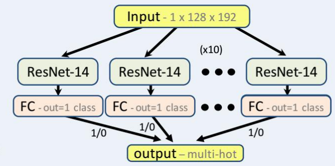


Our most successful models for multi-label task

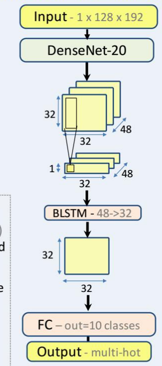
1. DenseNet-20



2. ResNet with Binary Relevance



3. DenseNet +BLSTM last layer



1. DensNet-20 is a simplified version of the original DenseNet[4]
2. Using ResNet14 (single-label architecture, simplified from [5]) as a binary classifier for each species results in a concatenated multi-label binary relevance model
3. Adding a Bi-directional LSTM as the last layer takes advantage of the sequential nature of sounds

Each model is then connected to either Binary-Cross-Entropy or Log-Sum-Exponential Pairwise ranking loss[6] for training

Binary relevance attains best F1 score while ranking loss proves valuable

F1* score is the reference metric for determining best models. Most models had gone through extensive hyperparameter searches - only best/notable ones are presented here

	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

Discussions and Conclusions

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 treated independently
4. Exploiting sequential nature of bird sounds with additional LSTM 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

Future Perspectives

Upon acquiring more computing resources, we wish to extend the datasets to 100+ species immediately, which will 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)

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

References

[1] e.g. BirdCLEF, EADM, BirdGenie; and proceedings/publications therein
[2] Xeno-canto database of bird sounds: https://www.xeno-canto.org
[3] Kahl *et al.*, Working Notes of BirdCLEF (2017). Github: kahst/BirdCLEF2017
[4] Huang *et al.*, Densely Connected Convolutional Networks (2016). arXiv: 1608.06993
[5] He *et al.*, Deep Residual Learning for Image Recognition (2015). arXiv: 1512.03385
[6] Li *et al.*, Improving Pairwise Ranking for Multi-label Image Classification (2017). arXiv: 1704.03135
[7] Szegedy *et al.*, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning (2016). arXiv: 1602.07261