A Baseline for Large-Scale Bird Species Identification in Field Recordings

Stefan Kahl[1], Thomas Wilhelm-Stein[1], Holger Klinck[2], Danny Kowerko[3], and Maximilian Eibl[1]

[1] Chair Media Informatics,
Chemnitz University of Technology, D-09107 Chemnitz, Germany
[2] Junior Professorship Media Computing,
Chemnitz University of Technology, D-09107 Chemnitz, Germany
[3] Bioacoustics Research Program, Cornell Lab of Ornithology,
159 Sapsucker Woods Road, Ithaca, NY 14850, USA
{stefan.kahl, thomas.wilhelm-stein, danny.kowerko,
maximilian.eibl}@informatik.tu-chemnitz.de, Holger.Klinck@cornell.edu

Abstract. The LifeCLEF bird identifcation task poses a difficult challenge in the domain of acoustic event classification. Deep learning techniques have greatly impacted the field of bird sound recognition in recent years. We discuss our attempt of large-scale bird species identification using the 2018 BirdCLEF baseline system.

Keywords: Bioacoustics · Bird Sounds · Deep Learning · BirdCLEF.

1 Motivation

Large-scale bird sound identification in audio recordings is the foundation of long-term species diversity monitoring. Aiding this labor intensive task with automated systems that can recognize multiple hundreds of species has been the focus in recent years. As part of the 2018 LifeCLEF workshop [1], the BirdCLEF bird identification challenges [2] provide large datasets containing almost 50.000 recordings to assess the performance of various systems attempting to push the boundaries of automated bird sound recognition.

2 Related Work

In 2016, Sprengel et al. [3] demonstrated the superior performance of convolutional neural networks (CNN) for the classification of bird sounds. Following that approach, we were able to improve the performance on a larger dataset containing 1500 different species with our 2017 BirdCLEF participation [4]. This year, we present an implementation of a streamlined workflow built on the most fundamental principles of visual classification using CNN. We published the code repository as baseline system complementing the 2018 BirdCLEF challenge [5]. The following workflow design, training scheme and submission results are entirely based on that system, establishing a good overall baseline for future comparisons and improvements.

3 Workflow

The key stages of our workflow include dataset pre-processing, spectrogram extraction, CNN training and evaluation. We adopted our last year's attempt and focused mainly on basic deep learning techniques, keeping the code base as simple and comprehensible as possible, while maintaining a good overall performance.

3.1 Dataset Handling

Using convolutional neural networks for the classification of acoustic events proved to be very effective despite the fact that these techniques are tailor made for visual recognition. Representing audio recordings as spectrograms overcomes this gap between the two domains of audio and image. We decided to use MEL-scale log-amplitude spectrograms which have been effectively used in similar approaches (e.g. [6]). A more detailed description of the extraction and pre-processing process can be found in [5].

3.2 Training

Our baseline training process supports multiple shallow and deep model architectures, extensive dataset augmentation, learning rate scheduling, model pre-training and result pooling. We implemented two basic CNN concepts: Fully-convolutional architectures with simple layer sequences and ResNet variations with shortcut connections. We also provide eBird[4] checklist metadata for both soundscape locations in Peru and Columbia along with the basline repository.

3.3 Model Distillation

Most CNN implementations are computationally expensive and rely on power-hungry hardware. Future applications of automated bird sound recognition will include field recorders capable of not only of recording, but also analyzing audio data in real-time. In those cases, battery life becomes an issue. In recent years, (semi-) mobile hardware - mostly used for IoT-applications - has been designed to aid this task. However, those hardware platforms are not yet suited for deep learning inference using complex models.

In 2015, Hinton et. al [7] presented an approach to distill knowledge in neural networks. We followed that scheme of model distillation and implemented a basic variant of teacher-student learning. Our baseline system allows to replace binary training targets with log-probability predictions of either single models or entire ensembles. We designed a simple shallow model that can predict species probabilities of one-second audio chunks in less than one second running on a Raspberry Pi 3+. The resulting scores are slightly lower than those of large single models, but still above the initial capabilities of the tiny CNN model.

[4] www.ebird.org/explore

The prediction performance of this approach is promising and model distillation may have significant impact on the field of mobile real-time species diversity assessment.

4 Results

We tried to cover different basic training and prediction schemes with our run submissions, including single baseline models, large and diverse model ensembles, metadata assisted attempts with species pre-selection and knowledge distillation training of tiny models. Table 1 provides an overview of selected results from our submissions.

Table 1. Selected submission results (run IDs in brackets, not all runs listed for clarity). Large ensembles including different net architectures and dataset splits perform best. Pre-selecting species for specific locations does not improve the results. Model distillation helps to reduce computational costs and can maintain results that are comparable with the performance of large single nets.

	Monophone Task		Soundscape Task	
Run Description	MRR Foreground	MRR Background	c-mAP Peru	c-mAP Columbia
Best Single	0.487 (1)	0.448 (1)	-	-
Best Ensemble	0.644 (5)	0.588 (5)	0.086 (6)	0.117 (6)
Pre-Selection	-	-	0.081 (7)	0.052 (8)
Raspberry Pi	0.425 (4)	0.385 (4)	0.077 (9)	0.083 (9)

The results show that our baseline attempt yields competitive results considering the complex evaluation task. Most results did match our expectations for the audio-only classification of field recordings. The key takeaways of the analysis of the submission results are:

- Diverse model ensembles covering different net architectures and dataset splits outperform single neural nets by a significant margin. This comes as no surprise in the domain of metric-centered competitions, but might not be applicable to real-world scenarios due to increased computational costs.

- Pre-selecting species did not improve the overall performance as expected. In some cases, selecting species based on time of the year and location helps to reduce training time. Using metadata as post-filter to eliminate false detections or as input during model training might lead to better results.

- Model distillation is a powerful tool to increase the classification performance of tiny neural networks. The results show comparable performance in the soundscape domain despite much smaller model architectures, when compared to model ensembles.

We published our entire code repository[5] and encourage future participants and interested research groups to build upon our results and improve the performance for the analysis of complex soundscapes - the most crucial aspect of species diversity monitoring.

5 Future Work

Assessing high-quality field recordings for the presence of bird species using convolutional neural networks is an effective application of deep learning techniques to the domain of acoustic event detection. Considering our own scores and those of other participants, current machine learning algorithms yield very strong results for this task. However, the 2018 BirdCLEF evaluation showed that the transfer of knowledge extracted from monophonic community recordings to the domain of long-term soundscape recordings is still very difficult. Hardly any improvements over last year's result have been accomplished. Future research should specifically focus on this task. Additionally, power-hungry hardware and computationally expensive algorithms are not well-suited for real-world applications such as mobile reorders. Improving techniques to shrink the size of neural networks while maintaining the overall performance will greatly help the field of long-term species diversity assessment.

References

1. Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Planqué, R., Vellinga, W.-P., Müller, H.: Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of AI. In: Proceedings of CLEF 2018 (2018).
2. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.-P., Kahl, S., Joly, A.: Overview of BirdCLEF 2018: monophone vs. soundscape bird identification. In: CLEF working notes 2018 (2018).
3. Sprengel, E., Martin Jaggi, Y. K., Hofmann, T.: Audio based bird species identification using deep learning techniques. In: Working notes of CLEF 2016 (2016).
4. Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M.: Large-Scale Bird Sound Classification using Convolutional Neural Networks. In: CLEF working notes 2017 (2017).
5. Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: Recognizing Birds from Sound - The 2018 BirdCLEF Baseline System. arXiv preprint arXiv:1804.07177 (2018).
6. Grill, T., Schlüter, J.: Two convolutional neural networks for bird detection in audio signals. In: Signal Processing Conference (EUSIPCO), 2017 25th European, pp. 1764–1768, IEEE (2017).
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).

[5] https://github.com/kahst/BirdCLEF-Baseline