

Dealing with Class Imbalance in Bird Sound Classification

Eduard Martynov¹, Yuuichiroh Uematsu²

¹Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation

²Ricoh Company, Ltd, 2-7-1, Izumi, Ebina-shi, Kanagawa 243-0460, Japan

Abstract

Recent achievements in machine learning have allowed to create fully-autonomous bird sound detection pipelines; however, they usually suffer from weak performance on underrepresented classes. We overcome this issue by proposing an approach which combines custom Convolutional Neural Networks and Pretrained Audio Neural Networks (PANNs). During training, we leverage pseudo labels as well as the hand labels for small classes. Moreover, we distribute classes between models and use different loss functions to train them. Our solution has achieved third place on private leaderboard in BirdCLEF 2022 challenge.

Keywords

Convolutional Neural Networks, Audio classification, BirdCLEF 2022, Sound Event Detection, Pretrained Audio Neural Networks, Computer Vision, CEUR-WS

1. Introduction

The BirdCLEF challenge plays an important role in developing biodiversity monitoring methods throughout the world. Previous BirdCLEF challenges [1, 2] utilize micro-like metrics such as F1-micro and class mAP. It has allowed participants to focus on improving top-line evaluation statistics on a common core set of species with large data amount, while species with less data available were primarily left uninvestigated.

The aim of BirdCLEF 2022 [3, 4] challenge is to enhance the detection performance of various bird calls, for which it is hard to acquire audio samples. In the task participants were given the dataset total of 15182 training samples, containing 152 classes. Only 21 of them were scored. Some classes had only one sample in the given dataset; non-scored ones were introduced to enrich dataset with other audios containing bird calls. Each training sample has associated primary and secondary labels. The primary label bird usually can be heard very clearly in the first and last 5 seconds of the audio clip. The birds from secondary labels can be heard anywhere in the audio.

The test set is larger by almost a factor of 10 than it was in the previous competition and contains 5500 soundscapes with duration of approximately one minute. For each 5-second


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ mart.eduard67@gmail.com (E. Martynov); yuuichiroh.uematsu@jp.ricoh.com (Y. Uematsu)

🆔 0000-0002-2122-0024 (E. Martynov)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

segment in every test audio, the participants were asked to predict whether each of 21 scored birds can be heard or not.

In our solution, we use custom CNN model proposed in [5] and PANN-like Sound Event Detection (SED) model proposed in [6] which showed good performance throughout BirdCLEF competition history. However, custom training techniques were required in order to achieve good performance, we'll discuss them below.

2. Dataset preparation

All of our models use 2D mel spectrograms as input. We used mel spectrogram transform implemented in torchaudio [7] library in order to convert raw audios to 2D images. This implementation enabled the conversion of audio clips directly on GPU, which boosted training speed by factor of 10 compared to similar approach with librosa [8] that executes conversion on CPU instead.

3. Model architecture

3.1. Convolutional Neural Network

One of the models that we used was custom CNN proposed in [5]. This model achieved second place in BirdCLEF 2021 challenge, so it is a good baseline which we decided to utilize not only because of its performance, but also due to the nature of the model.

We train this model using random 30-second crops, and for input audio that does not have this length, we pad it with zeros. As we can see from the Figure 1, before we feed mel spectrograms to the backbone, we reshape them from 30-second crops to 6 equal 5-second parts which essentially limits receptive field of this network to 5-second crops. We think that this architectural design allows the network to generalize to 5-second crops as well.

As a loss function we use BCE loss, and as targets we use union of primary and secondary labels. We believe that for this model precise localization of birds is not necessary, since random 30-second crops almost always contain target signal. To select models on validation, BCE loss

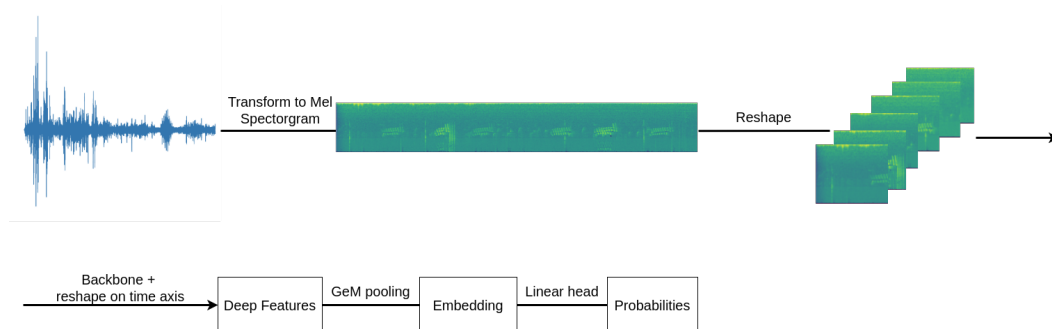


Figure 1: CNN training pipeline proposed in [5]. Input sample is a random 30-second crop from the original audio.

was used as a metric. For given audio from validation set, we crop first 30 seconds and use it as an input to the model.

For the inference we use 5-second crops directly.

3.2. Pre-trained audio neural network

Additionally, in our solution we used a PANN-based [6] model architecture, which achieved high accuracy in last year competition. We used several versions of this model which differ in backbone. These models were trained using two stages.

At first stage we split the dataset into 4 folds and train models using 10-15 second random crops in cross validation manner. Each of the 4 models learned underlying structure of the data and obtained the ability to distinguish between different bird calls. However, since during first stage we used random crops, it's not absolutely necessary that model received proper gradients. Sometimes, random crop can contain no bird call at all, in this case the label provided to the model is wrong. This can introduce unwanted noise and can be dealt with by using pseudo labels.

As for our approach to fight weak labels, at the second stage we used pseudo labels obtained from predictions on out-of-fold data. For given audio, we selected all segments that contained bird calls according to the predictions of the model from the first stage. We also dropped secondary labels, if our model was not confident enough about them. Of course, we zeroed-out any probabilities from pseudo-labels which correspond to the birds that weren't included neither in the union of primary label and secondary labels for the given audio clip.

We think that for good performance of SED models pseudo-labels are crucial, since the input samples are 10-15 second long. We believe that training without pseudo-labels can affect the performance of these models.

We then re-train these models using pseudo-labels using BCE loss and Focal loss [9] for different models.

The checkpoint selection was based on F1-macro metric in both stages.

4. Augmentations

The difference between training and testing data is big, since training data is a collection of human recorded audios of birds from xeno-canto web library [10], while the test data is a set of automatically recorded soundscapes. To fight the domain shift and make sure that our models generalize well to the unseen data, we use the set of the following augmentations:

- Mixup [11]
We applied mixup augmentation with probability of 1.0 and alpha = 1.0, this augmentation stabilized the training and when applied with cosine annealing learning rate schedule allowed models to converge even when trained on whole training dataset.
- Cutmix [12]
We also applied cut-mix augmentation to further improve stability. Applied after mix-up, it didn't introduce any noticeable changes.

- Background noise
To simulate the noisy environment of the nature, we also added background noise to all audio samples with different SNR. To not accidentally add noise containing bird call, we selected no-call samples from freefield1010 [13], BirdVox-DCASE-20k [14] and previous year challenge data and used them as background noise.
- Random power
We randomly raised mel spectrograms to a power varying from 0.5 to 3.
- Spec-Augment [15]
We randomly dropped 2 time stripes and 8 frequency stripes from mel spectrograms during training to enrich training dataset.
- Gaussian SNR
We also added the gaussian noise to allow further generalization of the model.

5. Oversampling and hand labels

Adding hand-crafted labels was one of our approaches to enhance performance on underrepresented classes. First we manually extracted segments from audio data with the target bird singing and then split the audio to increase the amount of training samples. This work took 4-5 hours, as it was done only for small classes.

Second, we increased the amount of training samples of some class up to 10 by random oversampling, provided this class had less than 10 samples. This number was chosen to balance between underfitting and overfitting to certain classes during training.

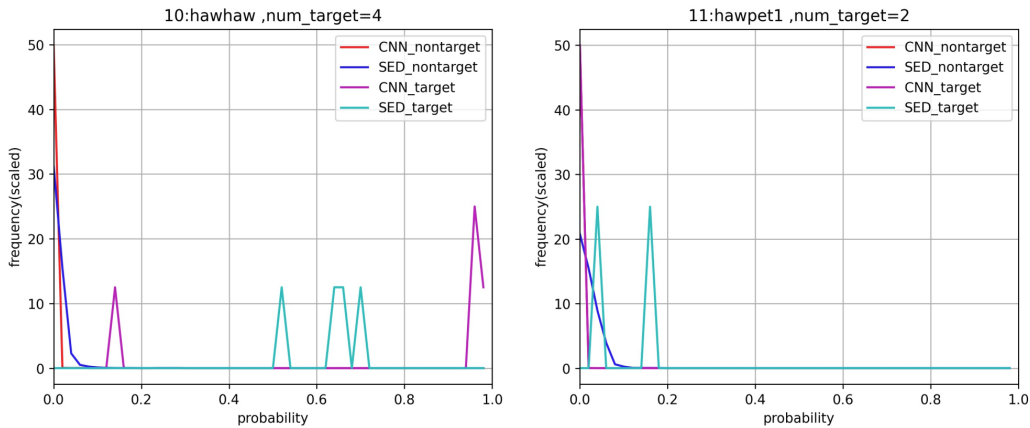


Figure 2: SED advantage over CNN. This figure shows that CNN can sometimes miss samples from small classes, however, SED doesn't lack the ability to detect them.

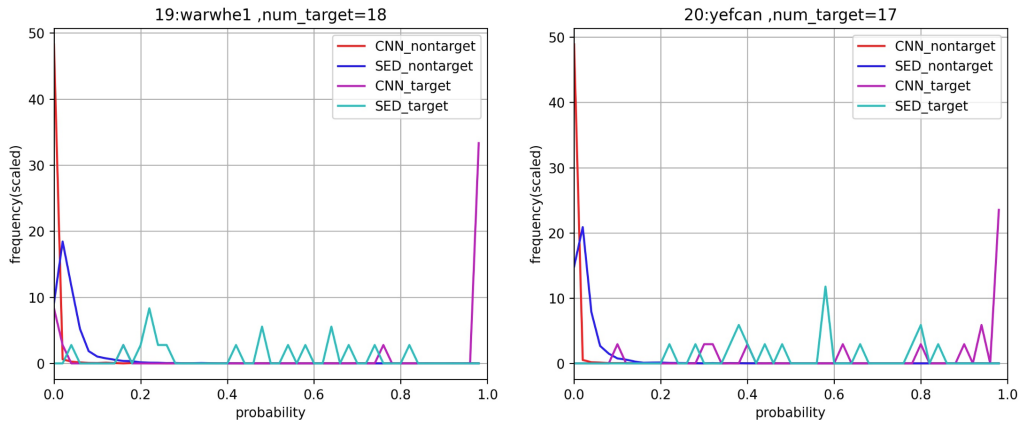


Figure 3: CNN advantage over SED. In this figure we can clearly see that CNN captures the information about well-represented classes better than SED.

6. Ensembling

6.1. Bird split

During ensembling stage we figured out that simple averaging of prediction of SED trained with Focal loss and CNN trained with BCE loss does not improve the competition metric value. To reason this we inspected probability distributions of predictions on both target and non-target data for every bird. It appeared that SED model showed tendency to make more conservative predictions, however it did not miss bird calls belonging to underrepresented classes as can be seen on Figure 2. On the other hand, CNN model was always confident and can divide the data very well for large classes, predicting probabilities for non-target data close to 0 and for target data close to 1. It allowed CNN model to reduce number of false positive detections while having the benefit of obtaining high probabilities for the actual bird calls, Figure 3.

To address that, we split birds into two groups; the first one contains 7 underrepresented birds - 'crehon', 'ercfra', 'hawgoo', 'hawhaw', 'hawpet1', 'maupar', 'puaioh', which we call this set of birds Group One; and the second group contains other 14 scored birds, we call it Group Two.

In the final ensemble we ended up using SED models trained with Focal loss to predict birds from Group One, while for the birds from Group Two we used CNN models and SED models both trained with BCE loss.

6.2. Postprocessing

We applied so called time-smoothing as in [5] post-processing to probabilities obtained for birds from Group Two, which is essentially a sliding window weighted average of probabilities applied on the time axis. It can be seen as a soft way of lowering the thresholds for the model, since only probabilities with high-scored neighbours receive a considerable gain. This postprocessing generates missed true positive detections with inconsiderable amount of newly introduced false positives.

6.3. Threshold selection

Since metric of BirdCLEF 2022 challenge is threshold-dependent, we had to accurately select the thresholds for all birds using different properties of this year dataset:

- The probability distribution of out-of-folds predictions for non-target data when using Focal loss differs depending on the bird, so we set the threshold for each bird from Group One depending on this distribution. We adopt the value of 91 percentile of the probability distribution for these birds.
- For birds from Group Two, we set the threshold to 0.05, except for "skylar", for this bird the threshold was set to 0.35, since it was obvious from validation that our models recognize this bird very well. We found out that for our models threshold 0.05 is the best on public leaderboard.

7. Results

As our best submission, we used ensemble of CNN and SED models with proper bird split between models. We show our results in Table 1.

Individual models performed competitively, but we had to use ensemble to get the third place.

Table 1

Model results. This table highlights the performance of different ensembles of the models.

Models	Public LB	Private LB
CNN model without augmentations	0.7715	0.7278
CNN model with augmentations	0.7761	0.7359
Best CNN ensemble	0.8327	0.7898
SED model (4 folds)	0.8339	0.7823
SED + CNN ensemble using two groups of birds	0.8532	0.8052
Add SED trained with BCE loss to birds from Group Two	0.8750	0.8126
Lower the thresholds for Group Two birds (0.05 -> 0.03)	0.8707	0.8274

8. Conclusion and future work

During this challenge we explored various models and found out that it was necessary to carefully choose training strategies to get the best performance. Different techniques such as pseudo labeling, oversampling and hand labeling were tested and performance was verified, as well as the smart ensembling of various models. Moreover during training we used BirdCLEF 2022 competition dataset along with no-call samples from freefield1010 [13], BirdVox-DCASE-20k [14] and previous year challenge data, each of these datasets was a good source of background audio samples.

As for the future work, we would like to inspect the impact of random-crop length as well as impact of pseudo-labels for CNN models, since we think that smaller crops can benefit the

model during training and make it easier to learn signal; however, it is harder to acquire correct training samples in this case.

References

- [1] S. Kahl, T. Denton, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of birdclef 2021: Bird call identification in soundscape recordings, 2021, pp. 1437–1450. URL: <http://ceur-ws.org/Vol-2936/paper-123.pdf>.
- [2] S. Kahl, M. Clapp, W. A. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga, A. Joly, Overview of birdclef 2020: Bird sound recognition in complex acoustic environments, 2020. URL: <http://ceur-ws.org/Vol-2696/paper-262.pdf>.
- [3] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, I. Bolon, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc, H. Müller, Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022.
- [4] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of birdclef 2022: Endangered bird species recognition in soundscape recordings, Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022).
- [5] C. Henkel, P. Pfeiffer, P. Singer, Recognizing bird species in diverse soundscapes under weak supervision, 2021. URL: <https://arxiv.org/abs/2107.07728>. doi:10.48550/ARXIV.2107.07728.
- [6] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition, 2019. URL: <https://arxiv.org/abs/1912.10211>. doi:10.48550/ARXIV.1912.10211.
- [7] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhersch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélaïr, Y. Shi, Torchaudio: Building blocks for audio and speech processing, arXiv preprint arXiv:2110.15018 (2021).
- [8] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: Proceedings of the 14th python in science conference, volume 8, 2015.
- [9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2017. URL: <https://arxiv.org/abs/1708.02002>. doi:10.48550/ARXIV.1708.02002.
- [10] X. Canto, Sharing bird sounds from around the world, 2022. URL: xeno-canto.org.
- [11] H. Zhang, M. Cissé, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, CoRR abs/1710.09412 (2017). URL: <http://arxiv.org/abs/1710.09412>. arXiv:1710.09412.
- [12] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy

to train strong classifiers with localizable features, CoRR abs/1905.04899 (2019). URL: <http://arxiv.org/abs/1905.04899>. arXiv:1905.04899.

- [13] D. Stowell, M. D. Plumbley, freefield1010 - an open dataset for research on audio field recording archives, in: Proceedings of the Audio Engineering Society 53rd Conference on Semantic Audio (AES53), Audio Engineering Society, 2014.
- [14] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, J. Bello, Birdvox-full-night: A dataset and benchmark for avian flight call detection, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 266–270. doi:10.1109/ICASSP.2018.8461410.
- [15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, in: Interspeech 2019, ISCA, 2019. URL: <https://doi.org/10.21437/Interspeech.2019-2680>. doi:10.21437/interspeech.2019-2680.