BirdCLEF 2021: building a birdcall segmentation model based on weak labels

Maxim V. Shugaev [1 †], Naoya Tanahashi [2 †], Philip Dhingra [3], Urvish Patel [4]

[1] *Intelligent Automation, Inc, 15400 Calhoun Drive, Suite 190, Rockville, Maryland 20855, USA*
[2] *Research and Development Group, Hitachi, Ltd. 1-280, Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan*
[3] *2443 Fillmore St. #380-8802, San Francisco, CA 94115, USA*
[4] *Pirimid Fintech, B-503 Mondeal Heights, SG Hwy, near Wide Angle Cinema, Ahmedabad, 380015, India*

Abstract

Deep-learning-based approach is applied to identify bird species in soundscapes. Two distinct model concepts targeted at audio classification and segmentation are considered. An audio segmentation model incorporating global multi-head self-attention to account for the interaction between different parts of audio is proposed, and peculiarities of building a segmentation model based on weak labels are discussed. To overcome the key challenges of BirdCLEF 2021, label weakness and domain mismatch, we developed a multistep training procedure with generation of segmentation pseudo labels and intelligent sampling of train audio, performed hand annotation of a part of the data, and used the noise from external data to mitigate the domain mismatch and improve model performance on soundscapes containing a substantial level of background noise. Our solution has reached 0.6605 F_1 score at the private leader board and achieved top-18 among 816 teams and 1001 competitors of BirdCLEF 2021.

Keywords

Bird identification, BirdCLEF 2021, Deep Learning, Convolutional Neural Network, multi-head self-attention, audio segmentation, weak labels, domain shift, log-sum-exp aggregation

1. Introduction

Each of us has heard bird songs in the everyday life. They are not only beautiful and sometimes inspiring but also may uncover the effect of human activity on nature. Changes in bird population and their behavior may indicate more global changes in the entire ecosystem, resulted, for example, by environmental pollution or global warming. To track these changes the ornithology community is collecting many hundreds of hours of audio recording every day. However, since there are more than 10,000 bird species, the task of bird classification based on their calls becomes almost impossible for the public and very difficult even for experts. Recent progress in the field of deep learning has revolutionized such areas as computer vision, natural language processing, and even generation and classification of audio. To harness these novel advancements for the improvement of an automatic bird classification system, Cornell Lab of Ornithology has hosted BirdCLEF 2021 competition at Kaggle [1], the world's largest data science competition platform. The overview of this challenge is provided in refs. [2, 3].

The objective of BirdCLEF 2021 is identification of bird calls in soundscapes collected in several places around the world. The model performance is evaluated base on row-wise micro averaged F_1 score computed for 80 10-minute-long soundscapes split as 35/65 ratio for public and private leader board (LB). The provided data includes the following [4]:

[1] CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: mvs9t@virginia.edu (A. 1); na.tanahashi@gmail.com (A. 2); philipkd@gmail.com (A. 3); urvishp80@gmail.com (A. 4)
ORCID: 0000-0002-1841-3677 (A. 1)
[†] These authors contributed equally to this work.

CEUR Workshop Proceedings (CEUR-WS.org)

(1) *train short audio* is a collection of short records of individual birds (62874 audios) uploaded to *xenocanto.org* website [5] (these files have been down-sampled to 32 kHz to match test set audio and converted to ogg format). The provided metadata contains a variety of information including primary and secondary labels: the main bird present in the recording and birds present in the background, respectively. The total number of considered classes is 397.
(2) *train soundscapes* are a collection of 20 soundscapes similar to ones used for model performance evaluation. The labels are provided for each 5-second chunk and may contain several bird songs simultaneously as well as nocall parts.

In the 2020 Cornell Birdcall Identification challenge [6] the best performance has been achieved for a model based on a convolutional neural network (CNN) applied to audio files converted into mel-scale spectrograms. Therefore, our approach is based on this method, and we considered two distinct model concepts targeted at audio classification and segmentation. Since the first of them is widely applied for bird classification, in our manuscript we focus on discussion of the segmentation approach, we proposed working on BirdCLEF 2021, and only briefly touch the key ideas used for building the classification model. The main challenges that arose in this competition include (1) *label weakness and noise* and (2) *domain mismatch between train and test data*. These challenges are addressed with the use of multistep training procedure including generation of pseudo labels (PL) for segmentation, sampling based on PL, performing domain mitigation as a separate step, and use of noise from external datasets as well as a smart selection of training audio chunks and manual labeling the part of the data. As the result, among 816 teams and 1001 competitors in this challenge, our team took 18th place by achieving a row-wise micro averaged F_1 score of 0.6605 at private LB. In this working note, we introduce our solutions and findings from the competition.

2. Methods

2.1. Data preprocessing

The audio files, before use as an input, are converted into mel-scale spectrograms with 128 frequency channels. We considered two Fast Fourier Transform (FFT) window sizes n_{fft} of 1024 and 2048. In our work, the larger value of n_{fft} is observed to mitigate the domain mismatch between train and test data. This effect may result from a decrease in the signal from short noisy sounds due to averaging them within a wider FFT window. Meanwhile, since the duration of the bird call is typically much larger than n_{fft} divided by the sampling rate, the birdcall signal preserves the amplitude. Despite we only considered two values of n_{fft}, and the effect of n_{fft} on the mitigation of the domain mismatch requires an additional study, in the final models we used $n_{fft} = 2048$ because of the better performance. The models are trained on 5-second chunks, which corresponds to 128×313 and 128×512 mel-scale spectrograms image size for classification and segmentation models, respectively, because of the use of different stride size. The training is performed on short audio, while the train soundscapes are used as a cross-validation (CV) dataset as well as a source of audio segments with noise.

For augmentation we used MixUp modified to include labels from both audios with values of 1 and a large value of alpha of 4-5 to ensure mixing the spectrograms in approximately the same portion. Since sounds overlay each other rather than overlap, MixUp based augmentation is natural for audio data and provides a considerable performance boost. To mitigate the domain mismatch between train and test data, white and pink noises are added during training. Beyond that, noise extracted from train soundscapes and records from the Rain Forest dataset [7] is added during training of segmentation models, which provided important insights about the peculiarities of soundscape data as well as enforced the model to distinguish actual birdcalls from sounds similar to birdcalls (for example from frogs).

2.2. Classification models: mitigation of the effect of weak labels

In BirdCLEF 2020 we used two distinct approaches: audio classification and segmentation. In this section, we explain the key details of the first of them. After the conversion of the input data into mel-scale spectrograms, it is passed to a CNN [8] for bird call classification. As a backbone we used

ImageNet pre-trained ResNeSt50 network [9, 10]. Below we explain 3 main ideas used to build the model.

(1) *Smart crop selection*. The provided data has various lengths, and before passing it to a model, the data should be aligned to a particular length. We have chosen 5-second segments because it corresponds to the chunk length required for model evaluation. The simplest way to generate a 5-second chunk is a random crop of an input clip. However, as pointed out above, one of the challenges related to the provided data is label weakness: labels are assigned to entire clips. Therefore, after the selection of a random chunk, it may end up containing only a part where the bird is not calling, and the global label assigned to the chunk is not correct anymore. To improve the cropping algorithm, we came up with a simple hypothesis: before uploading audio to the *xenocanto* website, authors delete parts of audio when there is no song at the beginning or at the end of the clip. This hypothesis was confirmed by the Exploratory Data Analysis (EDA) that we conducted during this competition. Therefore, during training, we assumed that the first 5 seconds and the last 5 seconds of the audio almost certainly contain a birdcall. Specifically, instead of random cropping in 20% cases we selected a chunk from the beginning or the end of the file. In addition, we randomly shift the selected parts by 0-2 seconds to avoid using the same 5-second chunks. We have chosen this strategy because selection of only the first and the last 5-second chunks does not provide sufficient diversity.

(2) *Use only short clips (length 60 seconds or less)*. The method of smart crop selection, described above, is not perfect, and there is a possibility of selection of a chunk with nocall. To decrease this probability, we have shortened the length of the audio used: for longer audio longer nocall parts may be present. Selection of 60-second clips or shorter helps to choose audio chunks with actual birdcalls. However, it has an obvious disadvantage of reducing the total number of training clips and the diversity of the data. In our case, the decrease of the number of clips from 62874 to 45423 did not hurt training, and more proper selection of 5-second chunks overcame this negative effect resulting in an increase in the model performance in comparison to our baseline.

(3) *Hand labeling*. We manually labeled time ranges of nocall and "human voice" in a part of short train audio clips. Labeling individual birdcalls is very time-consuming. Therefore, instead of labeling each birdcall we only marked regions with sufficiently long nocall parts, as depicted in Figure 1a. To avoid mixing the regions with primary and secondary labels and ensure that parts that are not marked as nocall actually include calls from primary birds, we only considered audio that does not have secondary labels and has a length of 16 seconds or less. These audios are unlikely to contain more than one bird. During labeling we realized that some of the audio files contain digitized human voices, as illustrated in Figure 1b. It appeared that a particular author included the voice when edited the data. And we went through all corresponding audio and marked parts with a human voice.

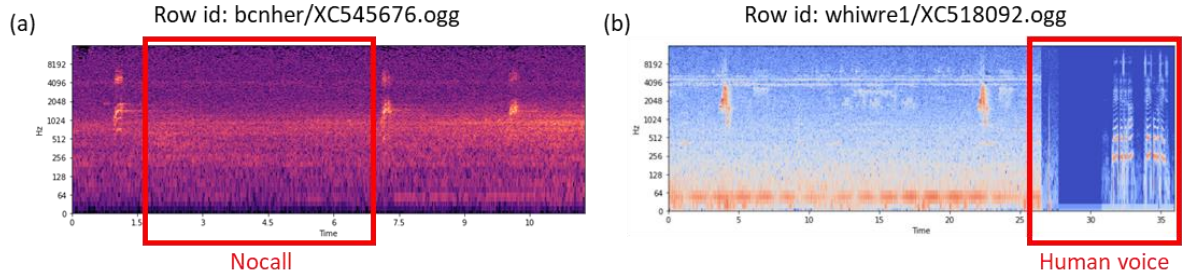

Figure 1: Examples of spectrograms with hand labels for nocall (a) and human voice (b) time ranges.

Table 1: Example of hand annotation

row_id	human_voice	nocall	t_range
bcnher/XC545676.ogg	0	1	[[1.5,7]]
whiwre1/XC518092.ogg	1	0	[[25,1000]]
…	…	…	…

Our hand-labeled data is available at ref. [11], and below we explain how to use it. An example of our annotation is listed in Table 1. The row_id column contains the primary label and filename concatenated by a slash, while the human_voice and nocall columns indicate whether the data contains human voice and nocall, respectively. In the provided example we can see that for the data with the row_id of *whiwre1/XC518092.ogg*, the second element of its t_range is set to 1000, well beyond the length of the audio. This indicates that the human voice is present until the end of the audio.

The above procedures are targeted at mitigating the effect of weak labels by ensuring that a selected chunk contains the actual primary label rather than nocall. The data preparation in our training procedure can be described as the following. When the audio data is loaded, first we check whether the data contains human voice based on the row_id and cut the audio based on t_range to exclude the voice part. Next, we apply Smart Crop Selection (1) step and generate a 5-second chunk. If the loaded audio is indicated to have nocall parts, we check whether the selected chunk is included in any listed nocall intervals, and if it is the case, we reset the original label to nocall. We keep nocall parts to teach the model to deal with nocall chunks present in the evaluation set. Despite improvement in the model performance, the above-described procedure has the disadvantage of reducing the size of training dataset. In addition, the procedure of hand annotation of bird calls is very time-consuming. Therefore, in parallel, we have developed a multistep training procedure, based on label self-distillation, and applied it to train a birdcall segmentation model (Section 2.4).

2.3. Audio segmentation models

Similar to common audio classification methods [8], in our approach the generated mel-scale spectrograms are passed through a convolutional neural network extracting features. However, in contrast to these methods, where the feature extractor is followed by pooling and generation of a prediction, we built a head performing a segmentation task and utilizing global all-versus-all attention to account for similarity and interactions of different parts of the produced audio sequence. Our model is schematically illustrated in Figure 2. In our study, we use ResNeXt50 backbone pre-trained in a semi-supervised manner on one billion images [12], which provides both a high level of computational efficiency and accuracy. For training at 5-second audio segments, the generated feature map has dimensions of 2048×4×16 (number of channels, frequency/temporal dimensions of feature maps). We compress the feature map with 4×1 convolution to 512×1×16 sequence representing a variation of audio feature over time with a resolution of approximately 0.3 seconds. This sequence is passed to 2 sequential multi-head self-attention blocks (8 heads) with skip connections followed by 1×1 convolution to generate 16×*num classes* sequence of predictions. In the initial model development performed by Dr. Shugaev, one of the authors of this manuscript, within the Cornell Birdcall Identification challenge [13], it was observed that adding batch normalization, a nonlinearity, and a fully connected part to multi-head self-attention blocks inhibits convergence on the competition data [14]. Therefore, they are excluded from our model.

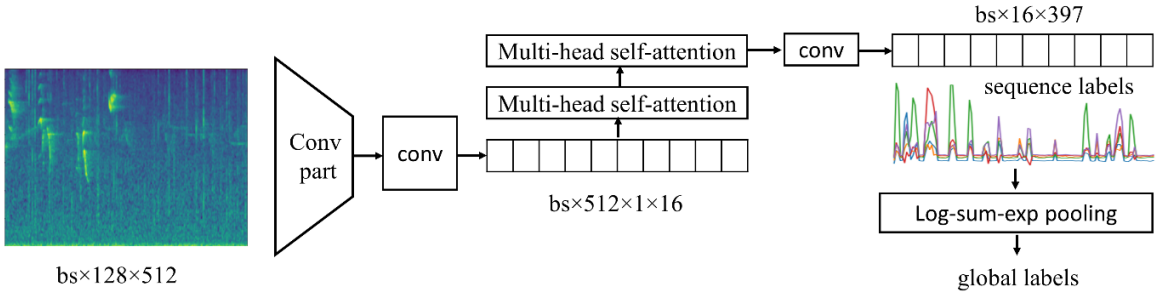

Figure 2: Schematic illustration of the proposed model architecture.

To perform training on labels assigned to a clip, the predicted sequence, corresponding to segmentation of the clip, should be collapsed. We use log-sum-exp (LSE) based aggregation. Following ref. [15], we introduced a temperature of 5 to LSE. In addition, since bird calls appear only within a short time frame, we enforced a soft minimum of the predicted sequence (LSE with the temperature of

-1) for all predictions, even ones having positive labels. This component is added to the total loss with a weight of 0.3. An additional important thing is using sufficiently short audio segments during training. Long segments help to overcome label weakness because the probability of finding a bird within a segment is increasing with its length. However, we realized that a model trained on long segments prefers to learn easy sounds ignoring the rest: it is often enough to find an easy example in a long sequence to assign the corresponding label. To enhance the discriminative ability of the model we used 5-second chunks during training, while inference is performed on longer, 10 min clips.

2.4. Multistep training

To addressed the key challenges related to the data (label weakness and domain mismatch) we proposed 3 step train-segment-shift approach outlined below.

(1) We begin with training a segmentation model based on the original labels assigned to entire clips. The main difficulty is that clip labels are not equal to chunk labels because clips may have no bird calls in particular parts or bird calls from other species. So direct training of a model on clip labels assigned to chunks cannot produce a sufficiently accurate model but gives PL for the next step. As an objective, we use focal loss applied to global predictions generated with LSE pooling with temperatures 5 and -1, as described above.

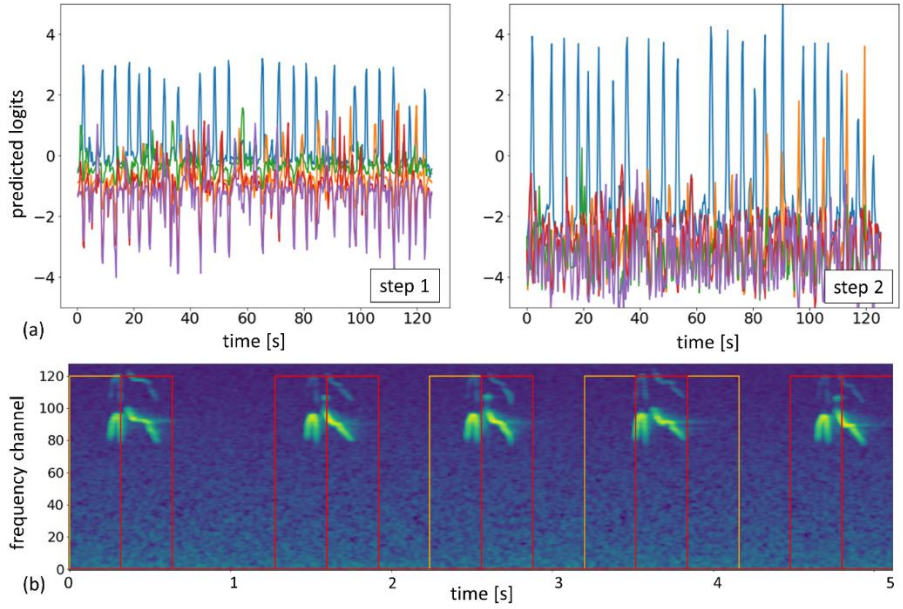

Figure 3: (a) illustration of segmentation predictions for top-5 classes on XC165287 file for models trained at step-1 and step-2 setups (each spike corresponds to a birdcall). Blue color represents primary label prediction, and orange corresponds to one of the secondary labels listed in the provided metadata. (b) An example of a 5-second mel-scale spectrogram extracted from XC449522 file with high and low confidence predictions for a step-2 model outlined by red and orange bars, respectively.

(2) Next, we run a similar setup as at the previous step but perform sampling of clips based on PL segmentation making sure that the primary label bird call is included in each selected audio segment. In addition, we apply segmentation loss to regions having high confidence pseudo labels (both positive and negative cases), while masking regions where confidence is not sufficient in segmentation loss evaluation. The total loss is composed of equal contributions of global and segmentation loss components. The segmentation component guides the model and improves the quality of produced segmentation labels, as illustrated in Figure 3a. Figure 3b, meanwhile, provides an example of mel-scale spectrogram with high and low confidence PL of the primary class marked by red and orange bars, respectively. At step (2) we were able to build a model reaching ~0.86 F_1 CV based on out-of-fold predictions on short train audio [4] evaluated by taking the maximum values for predictions within the entire length of a file. However, when we tried to apply this model directly to provided train soundscapes (which are similar to test data), we reached only 0.67 CV (on soundscapes), which

indicates the domain mismatch between short audio and soundscapes. The PL produced at this step are shared at ref. [16].

(3) To accommodate for the domain mismatch, we need one additional step in building a model. When we plot the spectrograms for train soundscapes, we realized a presence of a substantial fraction of noise and background sounds, which likely degrade the model performance on the soundscapes. As discussed in the next section in more details, at this step the original signal from short train audio clips is mixed with pink noise, noise extracted from train soundscapes (using proper CV split to avoid the use of parts with extracted noise for model evaluation), and sounds from the Rain Forest challenge dataset [7]. This approach boosted the single model performance to 0.7598/0.6486 at public/private challenge LB. We would like to highlight the importance of step (2) and accurate segmentation PL for step (3). Without noise the model is able to localize the areas with corresponding bird calls. Meanwhile, if extensive noise is added, the model needs hints on localization of birdcalls and ignoring the rest. Segmentation PL help to guide training under conditions of strong noise and background sounds added to the train data.

It should be noted that the use of a label self-distillation-based approach [17] to correct label errors and handle weak labels (which is at the core of the proposed multistep training method) is not a new idea but rather a common procedure to a specific set of tasks. For example, Born-Again Network [18] based approach has already been used in earlier BirdCLEF competitions [19] to mitigate the effect of weak labels.

Additional training details: training at steps (1) and (2) is performed for 64 epochs with n_{ftt}=1024, Ranger-Larese-Lookahead optimizer [20]. Step (3) is performed for 28-64 epochs with the best model selection with n_{fft}=1024 and n_{fft}=2048. We apply a cosine annealing schedule without warmup with the initial learning rate of 10^{-3} for the backbone and 10^{-2} for the head (at step (3) we reduced the maximum learning rates twice). Focal loss is used in all experiments, but the total objective function includes contributions from global soft maximum and soft minimum as well as segmentation loss based on PL, as described above.

2.5. Postprocessing

We have used 3 following postprocessing methods to improve the predictions of our models.

(1) *Global correction*. The test data is recorded at the same location within 10 minutes. Therefore, we assumed that if a bird call is detected with high confidence in one of the chunks, it is likely to have the same species call in other parts of the record. To account for this effect, we multiply the model predictions for a particular class by 1.3 if the maximum predicted value over the entire length of the record for this class is exceeding the threshold of 0.5. The values 1.3 and 0.5 are determined based on the maximization of the validation score.

(2) *Power average*. The predictions of individual models are combined according to the following equation: $p = \sqrt{\sum_N p_i^2 / N}$, where N is the number of models, p_i is the prediction of i-th model, and p is the final prediction.

(3) *Sliding window*. Birdcalls may be located at chunk boundaries, which results in difficulties with assigning it to a specific chunk, both in provided audio annotation and in model predictions, as well as boundary effects if the inference is performed on 5-second chunks. To mitigate these effects, in addition to the original sequence of chunks we considered chunks shifted by 2.5-second and averaged our predictions according to the following rule:
$p(t) = 0.5 \cdot \tilde{p}(t) + 0.25 \cdot \tilde{p_s}(t - 2.5) + 0.25 \cdot \tilde{p_s}(t + 2.5)$, where $\tilde{p}(t)$ is the chunk of the original sequence corresponding to the moment of time t, and $\tilde{p_s}$ are predictions for a shifted sequence. For example, a postprocessed prediction for (5,10) second chunk is generated as an average of the model prediction on (5,10) chunk with a weight of 0.5 and (2.5,7.5) and (7.5,12.5) chunks with weights of 0.25. This postprocessing naturally works during the aggregation step in segmentation models but requires a generation of two sets of predictions for classification models.

In our study, all 3 postprocessing methods are applied to classification model predictions. While only sliding windows postprocessing is appeared to be effective for segmentation models because global attention is already performing global correction of predictions. We also did not consider the

information from metadata about recording location and time, which could help to filter out specific species and slightly boost the model performance.

3. Results and discussion

We started our experiments by exploring the performance of a simple classification-based approach. The first interesting observation we want to point out is the importance of secondary labels provided with the metadata. Including the labels to our objective with the value of 0.995 has boosted the performance of a model at public LB from 0.59 to 0.67. A "secondary label" means that other birds can be heard in the background at a lower volume. There is no distinction, however, between a bird that is in the foreground or the background, and, therefore, both of them should be included. However, one of the key difficulties related to training data is the presence of long segments without primary bird calls. For secondary labels this difficulty becomes even more pronounced since secondary birdcalls often appear only in several places of the audio files. Moreover, in our analysis we noticed that sometimes secondary birdcalls are not assigned or assigned but not present in the corresponding files. Therefore, we proposed two strategies outlined below: use of only short audio with additional hand annotation, where the impact of secondary birdcalls is minimum, and perform multistep training procedure with an iterative generation of PL.

The use of hand-annotated labels and chunk selection strategies described in Section 2.2 improved the performance of our classification models to 0.7316 and 0.7157 for CV and public LB, respectively. This result is produced based on the power average of predictions of models trained on 5 CV splits of the original short train clips with approximately the same distribution of records for each class (stratified 5-fold split). To improve model performance, we considered *sliding window* and *global correction* postprocessing (Section 2.5). The first one has boosted our CV to 0.7393 and public LB score to 0.7161, while *global correction* has improved CV to 0.7531 and public LB score to 0.7188. So, both of the considered postprocessing techniques increase the prediction accuracy, and their combination results in the final CV of 0.7611 and public LB score of 0.7479 (Table 2).

As an alternative, we performed training using the train-segment-shift concept, proposed in the previous section. The initial attempt of training a model at step (3) with white noise only resulted in quite a low CV of 0.7064 in comparison with our classification model, Table 2. To boost the model performance and reduce the domain gap between train and evaluation data, we mixed training data with noise extracted from train soundscapes (parts labeled as nocall), which resulted in the CV of 0.7778. This value is substantially higher than the CV of the classification model, but the obtained public LB score is lower, 0.7231. Since the private LB performance of the model was not available to us during experiments (which appeared to agree with CV), we assumed the presence of implicit leaks appearing due to a similarity between the provided train soundscapes.

We attribute the observed public LB mismatch between our classification models, not using train soundscape noise, and segmentation models to a lower value of $n_{ftt} = 1024$ used in the second case. While this parameter is not playing a substantial role for training on clean data, it appeared to be important for noisy data, and a larger value of $n_{ftt} = 2048$ mitigates the domain mismatch. In addition, to eliminate possible implicit leaks resulted from the use of train soundscapes we started using data provided with the Rain Forest challenge [7]. This data contains multiple background sounds and sounds similar to bird calls from nonbird species, *i.e.*, frogs. While this data also includes several bird species, we assumed that they are geographically different from the ones considered in the BirdCLEF 2021. This additional data has helped the model to learn how to ignore sounds similar to bird calls as well as bird calls from unknown species. The effect of both of these modifications has boosted the performance of our segmentation model to 0.7790 and 0.7598 for local CV and public LB, respectively (Table 2).

The inference for segmentation models is performed on full-length, 10 min, clips, which enables global all-versus-all self-attention to account for the interaction between different parts of the audio file and find similar patterns. Inference of our segmentation model on separated 5-second chunks gives only 0.7499 F_1 CV, while inference on full-length audio increases this value to 0.7654. *Sliding window* postprocessing further boosts the value to 0.7790 CV. This observation suggests that looking into the entire audio clip is important to make a decision about a specific part. We also verified that the performance of the model is not changed significantly between recording locations. The CV evaluated

on COR location is 0.8073 at the fraction of nocall of 0.699. Meanwhile, CV for SSW location, after reweighting the model score to the same fraction of nocall as in COR soundscapes, has 0.7935 F_1.

Table 2: results of key experiments

Setup	local CV	public LB	private LB
Initial classification model with secondary labels	0.7138	0.6767	0.5874
Classification model with hand label annotation + PP 2	0.7316	0.7157	0.6036
Classification model with hand label annotation + PP 2,3	0.7393	0.7161	0.6023
Classification model with hand label annotation + PP 1,2	0.7531	0.7188	0.6058
Classification model with hand label annotation + all PP	0.7611	0.7479	0.6188
Segmentation model step (3): n_{fft}=1024, no noise	0.7064	-	-
Segmentation model step (3): n_{fft}=1024, train noise	0.7778	0.7231	0.6401
Segmentation model step (3): n_{fft}=2048, rain forest noise	0.7790	0.7598	0.6486
Segmentation + classification model	0.8004	0.7735	0.6605

The combination of our final segmentation and final classification models with weights of 0.7 and 0.3, respectively, resulted in 0.8004 CV and 0.7735 and 0.6605 public and private LB scores. This significant improvement is resulted from the inherently different nature of the considered approaches. However, our initial experiments have demonstrated that a combination of nine models having a public LB performance of 0.67 produces a prediction scored at 0.70 F_1. Therefore, we can expect that incorporation of additional classification and segmentation models (trained with a different split and slight variation of the training procedure) in the ensemble will easily result in a further performance boost.

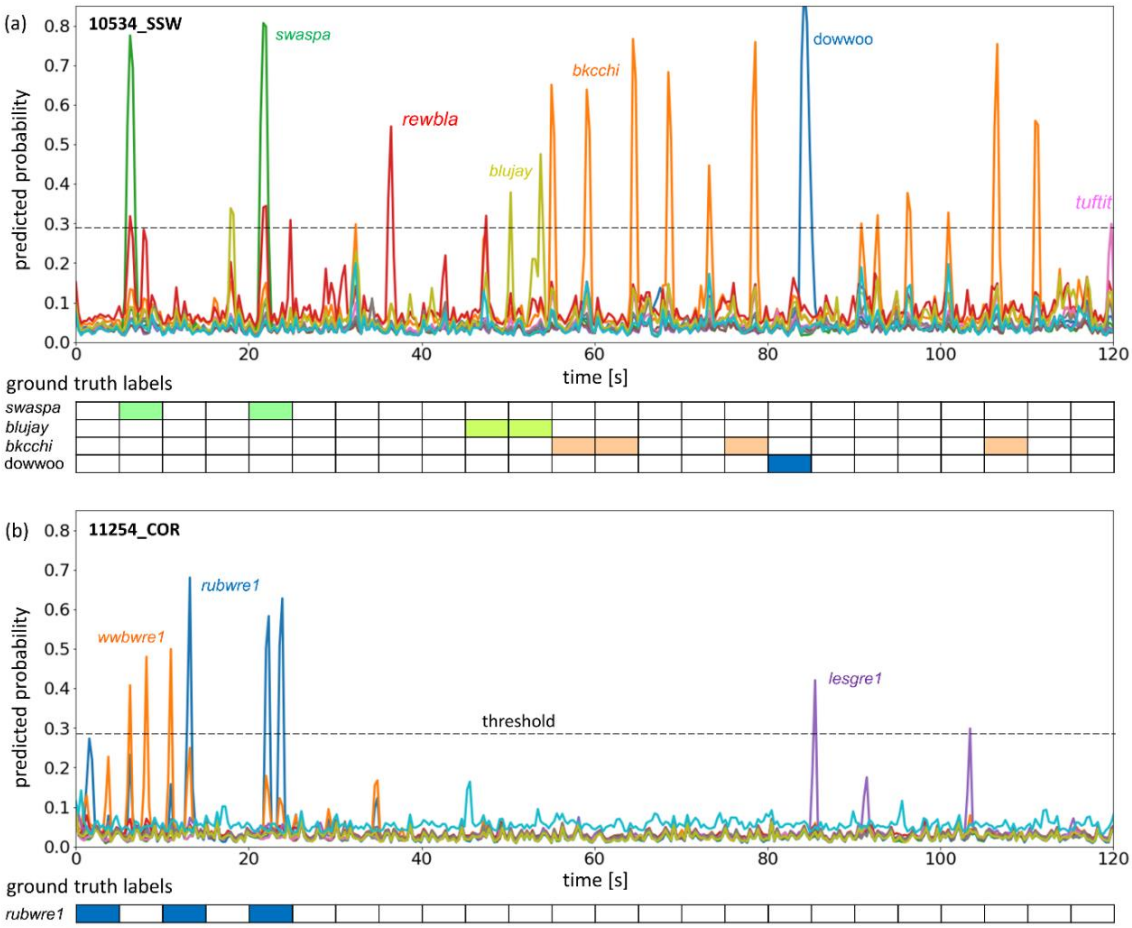

Figure 4: Illustration of the top 10 model predictions on the first two minutes of 10534_SSW (a) and 11254_COR (b) train soundscapes. The ground truth labels are listed on the bottom. The dashed line outlines the threshold.

The predictions of the segmentation model for the first two minutes of 10534_SSW train soundscape are depicted in Figure 4a for the top-10 predicted classes. The model is quite confident in the detection of nocall. However, sometimes the model may be confused with a prediction of particular cases. For example, a *swaspa* call with some noticeable probability may correspond to *rewbla*. Also, we noticed a number of places where the model predicts a birdcall with sufficient confidence, but there are no corresponding ground truth labels. Listening to the audio, we could recognize the presence of birdcalls at 18, 36, 43, 66, 73, 98, 111, and 120 seconds, as suggested by the model, but since we are not experts in bird classification, we cannot make a conclusion about the correctness of the model predictions. Consideration of audio from a different location, 11254_COR, in Figure 4b does not show any peculiarities in the model performance. The model predicts annotated birdcalls as well as a number of birdcalls at 6, 8, 85 seconds that could be recognized by listening to the audio but do not have an annotation. We understand that human labeling of soundscapes is a very challenging task given a variety of bird species as well as the complexity of vocalizations for specific species. However, the observation from Figure 4 may suggest that some parts of the provided soundscapes with bird calls were labeled by expects as nocall because of insufficient confidence or significant background noise. To exclude missing annotation and improve the accuracy of the ground truth labels, iterative AI (artificial intelligence) assisted labeling may be utilized. In this case, a human expert compares his/her labels with AI-generated predictions and decides on labels correction.

Finally, we would like to discuss possible ways to improve our approach. Since our efforts on the development of the classification and segmentation models were performed in parallel, step (1) training is done with random chunk selection. Incorporation of a more elaborate procedure, used for building a classification model, would improve the quality of the model. Next, in our study use of both noise from train soundscapes and noise from the Rain Forest dataset resulted in a noticeable improvement in model performance. However, we did not have a chance to perform an additional study targeted at finetuning a model trained with Rain Forest data with noise extracted from train soundscapes. Since the latter one includes the information about artifacts specific to devices used for soundscape data acquisition, such finetuning would likely further improve the performance. Regarding the segmentation model, we used short, 5-second, chunks to improve the discriminative ability of the model. Meanwhile, longer chunks could help to teach the model to consider the interaction between different parts of the audio. Another option would be consideration of the head from PANNs models [21] and comparison it with LSE-based aggregation. Finally, the performance of our model at the evaluation set could be improved with more careful threshold selection, performed with bootstrapping and accounting for the variation of the fraction of nocall chunks.

4. Summary

This report describes a deep-learning-based approach for identification of birdcalls in soundscapes. Convolutional neural networks are applied to mel-scale spectrograms to perform audio classification and segmentation. To overcome the main challenges of BirdCLEF 2021, label weakness and domain mismatch between train and test data, we have proposed a multistep train-segment-shift training procedure producing segmentation PL first and then performing mitigation of the domain mismatch with using noisy data but cleaned labels. As a source of noise, we considered train soundscapes and Rain Forest dataset [7], containing background noise and sounds from nonbird species, *i.e.*, frogs. Use of the noise has improved the model performance on soundscapes. In addition, we experimented with an intelligent selection of audio chunks during training and hand labeling of a part of the data to build a classification model. One of the interesting observations in our work is the effect of n_{ftt} on the model robustness to domain mismatch, and a larger value of $n_{ftt} = 2048$ improves model performance on soundscapes with background noise. Our findings have helped us to build a model that achieved 0.6605 F_1 score at the private leader board, which corresponds to the top-18 among 816 teams and 1001 competitors of BirdCLEF 2021.

5. References

[1] BirdClef 2021- Birdcall Identification, URL: https://www.kaggle.com/c/birdclef-2021/

[2] A. Joly *et al.*, Overview of LifeCLEF 2021: a System-oriented Evaluation of Automated Species Identification and Species Distribution Prediction. Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021), lifeclef2021, 2021.
[3] S. Kahl, T. Denton, H. Klinck, H. Glotin, H. Goëau, W.-P. Velling, R. Planqué and A. Joly, Overview of BirdCLEF 2021: Bird call identification in soundscape recordings, Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, birdclef2021, 2021.
[4] BirdClef 2021 – Birdcall Identification data, URL: https://www.kaggle.com/c/birdclef-2021/data
[5] Xeno-Canto Sharing bird sound from around the world, URL: https://www.xeno-canto.org/
[6] Cornell Birdcall Identification challenge, URL: https://www.kaggle.com/c/birdsong-recognition/
[7] Rainforest Connection Species Audio Detection data, URL: https://www.kaggle.com/c/rfcx-species-audio-detection/data
[8] K. Choi, G. Fazekas, M. Sandler, Automatic tagging using deep convolutional neural networks, arXiv preprint arXiv:1606.00298v1, 2016.
[9] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, A. Smola, ResNeSt: Split-Attention Networks, arXiv preprint arXiv:2004.08955, 2020.
[10] R. Wightman, Pytorch image model, URL: https://github.com/rwightman/pytorch-image-models, Accessed on June 2, 2021.
[11] N. Tanahashi, BirdCLEF 2021 Handlabel_data (Version 1), URL: https://www.kaggle.com/naoism/birdclef-2021-handlabel-data, Accessed on June 9, 2021.
[12] I. Zeki Yalniz, H. Jégou, K. Chen, M. Paluri, D. Mahajan, Billion-scale semi-supervised learning for image classification, arXiv preprint arXiv:1905.00546v1, 2019.
[13] M. V. Shugaev, Cornell Birdcall Identification discussion 39th place solution [top1 at public], URL: https://www.kaggle.com/c/birdsong-recognition/discussion/183258, Accessed on June 9, 2021.
[14] Cornell Birdcall Identification Challenge data, URL: https://www.kaggle.com/c/birdsong-recognition/data, Accessed on June 28, 2021.
[15] P. O. Pinheiro, R. Collobert, From Image-level to Pixel-level Labeling with Convolutional Networks, arXiv preprint arXiv:1411.6228, 2015.
[16] M. Shugaev, BirdCLEF 2021 OOF, URL: https://www.kaggle.com/iafoss/birdclef-2021-oof, Accessed on June 28, 2021.
[17] Z. Zhang, M. R. Sabuncu, Self-Distillation as Instance-Specific Label Smoothing, arXiv preprint arXiv:2006.05065, 2020.
[18] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, A. Anandkumar, Born Again Neural Networks, arXiv preprint arXiv:1805.04770, 2018
[19] J. Schlüter, Bird Identification from Timestamped, Geotagged Audio Recordings, URL: http://www.ofai.at/~jan.schlueter/pubs/2018_birdclef.pdf, Accessed on June 28, 2021.
[20] M. Grankin, Over9000, URL: https://github.com/mgrankin/over9000, Accessed on June 9, 2021.
[21] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D Plumbley, PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition, arXiv preprint arXiv:1912.10211v5, 2020