Birdcall Identification: The BirdCLEF 2021 Competition

May 18, 2022

Riccardo Pallucchi

Abstract

BirdCLEF 2021 is a code competition hosted by Kaggle, aimed at the development of machine learning algorithms to identify bird vocalizations. This work addresses the problem by exploiting convolutional and residual neural networks with different sets of hyperparameters and preprocessing techniques.

1. Introduction

The dataset provided for this competition (https://www.kaggle.com/c/birdclef-2021) consists of more than 676000 recordings of duration ranging between a few seconds to some minutes, with sampling rate 32 kHz, each paired with metadata information such as geographic position and date. The required task is to correctly classify the birds present in each 5 seconds segment (see Fig. 1) into 397 different species, to which one more class is added (*"nocall"*) corresponding to the absence of vocalizations. In the end, the final F1 score (Eq. 1) is obtained evaluating the performance on several 10 minutes soundscapes which constitute the hidden test set, while recording of similar length are publicly available for private tests.
The $F1$ score is defined as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$

where:

$$precision = \frac{tp}{tp + fp}; \qquad recall = \frac{tp}{tp + fn} \tag{2}$$

having defined tp as "true positive", fp as "false positive" and fn as "false negative"; in the multi-label, multi-class case, the row-wise micro-averaged $F1$ score considers the total number of tp, fp and fn.

Email: <pallucchi.1712592@studenti.uniroma1.it>.

Deep Learning and Applied AI 2021, Sapienza University of Rome, 2nd semester a.y. 2020/2021.

Approach. In order to exploit all the machinery already developed for image classification tasks, inherently $1D$ audio data are converted into $2D$ spectrograms (see Section 2); in this way it's possible to feed them directly to a convolutional neural network.

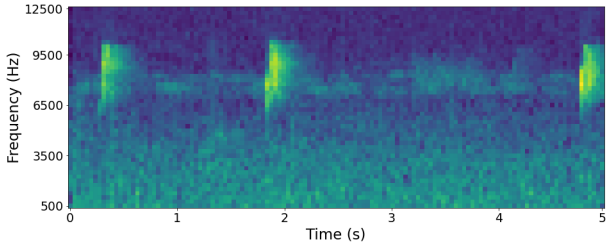

Figure 1. A 5s sample containing *Acadian Flycatcher* calls.

The main challenge of the competition is given by the nature of the dataset, which doesn't take into account the *nocall* class, assigning therefore a unique bird label to each recording even though large fractions of them lacks interesting signals. To overcome this obstacle various strategies are implemented, providing different ways to select valid data before training and during inference.

Convolutional and residual models are tested in various configurations; among them, ResNets gave the best performance. The code can be found at https://github.com/RiccardoPallucchi/BirdCLEF2021

2. Method

Data preprocessing. Firsts attempts to get valid training data were based on the *"rating"* associated with each sample. Other methods consisted in using only the first 5 seconds of each sample, or first and last 5 seconds, assuming that each available recording was suitably cut; also, two deterministic filters were designed: one is based on the presence or absence of sudden intensity variations, while the second relies on the difference between the signal and the average intensity at a given frequency in each spectrum, in order to discard samples before training or to create a brand new class for training (to allow more flexibility in the class boundaries). In the end the adopted strategy in-

volved using longer samples ($20s$ each) for training, providing therefore larger probabilities of meaningful signals, and repeated spectra ($5s \times 4$) at inference time.

Hence, training data are divided in $20s$ pieces, and then converted into mel spectrograms according to the formula:

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{3}$$

where m represents the frequency f in the new scale. The mel scale is a perceptual scale of pitches created in such a way humans perceive them at equal distance from one another; this conversion is needed since each recording was classified manually by experts, therefore the labels are conditioned by human perception of sounds. The final training sample is an image 48×512 where the mel frequencies are limited between 500 Hz and 12.5 kHz.

Data augmentation. The augmentations implemented are a combination of gaussian noise, time recombination (a transformation introduced to mix up temporal slices of a spectrogram, which also provides incomplete vocalizations), time and frequency masks.

The model. Initial efforts to carry out the classification task exploited a simple convolutional network in various configurations, from 3 to 5 layers; adding more layers was not worth the increase of computing time. Switching to residual networks to enhance the feature extraction capabilities led to visible gains in accuracy, and the deeper ResNet34 and ResNet50 gave better results than ResNet18.

It must be noted that, to take into account the temporal and spatial information encoded into metadata, after being converted in continuous scale and suitably normalized, they were attached to the flattened feature-rich representation produced by the convolutional part of the network.

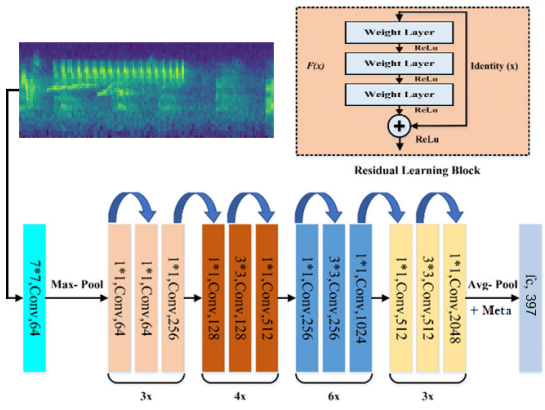

Figure 2. ResNet50 architecture scheme.

Classification. After the trained model is applied on the test set and a first classification is obtained, a method to discriminate between birds and *nocalls* is needed. In this regard, an efficient split is particularly important, since it is reasonable to assume that a significant portion of data will be attributable to plain background. The filters introduced at the beginning of this Section provide a partition which turns out to be more accurate than the methods based on the probability score alone, recognizing more than 75% of empty samples, and a similar percentage of bird vocalizations. Designing a binary classifier for this task using the test set was not feasible due to the low number of samples, while introducing a $398th$ class in the primary model populated by means of the filters didn't give promising results.

In addition to the filter, a probability threshold was imposed to discard inaccurate guesses, and the most problematic class was removed. Finally, an additional *nocall* label was assigned to samples with score below a fixed value; this is the only case involving more labels, since yielding more birds per sample (favouring recent species or species commonly found in the same sample) proved to be inaccurate.

3. Results

The highest scoring classifier (ResNet34) gets $F1 = 0.497$ on the private leaderboard, sensibly lower than the test set value $F1 = 0.570$. The gap is attributable to the difference in *nocall* fractions between the two; indeed, on test set the model is able to predict 7% of birds and 90% of *nocalls*.

4. Discussion and conclusions

The results look promising, but other ideas could be taken into consideration. In the first place, multilabel classification on the test set (choosing two or more birds with larger probabilities above a certain threshold) doesn't give the desired results; however, it would have been possible to exploit the "secondary label" feature of the train set to directly create a model suitable for this task, or even to combine more samples from different folders. Arguably, the multilabel approach mentioned earlier, based on recent birds and frequently paired species, would become useful for higher fraction of correct guesses. In addition, external information could be exploited to take into account the geographic distribution and to infer the habitat from background noise and preceding birds, in order to reduce the available label space; another way to proceed could be the production of different models to deal with different geographic regions.

Finally, the already large number of training samples, and the fundamental obstacle given by the mediocre quality of provided labels, suggested to rely upon data augmentation only; nonetheless, it could be interesting to expand the dataset through k-fold cross validation.