BirdCLEF2021: An ornithology expert system

October 2, 2021

Edoardo De Matteis

Abstract

In this report I approached the BirdCLEF2021 bird classification task and came up with a model that uses only soundscape recording, thus making most of the data privileged, and getting good results. The code is available on GitHub at https://github.com/edodema/Birdcalls

1. Introduction

Birds are high up in the food chain, this means that are sensitive to changes in the ecosystem and monitoring them can give us informations about pollution and the environment as a whole. Small birds are often difficult to sight and easy to hear, these calls are generally characteristic of each species so throguh them we can identify them in a smart way. Our goal is to listen to ambient recordings and guess birds' species through their call, the task is set up as an image classification among $n + 1$ classes i.e. n species plus no bird being destected.

Dataset. The BirdCLEF2021 dataset contains both recordings of birds in a controlled environment and noisy soundscapes tracks, we will use only the latter. Tracks are split by 5 seconds windows and labeled with primary and secondary gold truths - among other privileged data - but for our purposes we will consider primary ones only; we are being conservative so that should not be a problem.

2. Related work

Many previous works exploit spectrograms to reduce audio tasks to image ones (Hamdy et al.) (Michelashvili & Wolf, 2020) (Xie et al., 2021), among them (Xie et al., 2021) uses recurrent layers to deal with sequential knowledge. With the same intent we employ attention layers in one of the basic blocks that will form our model, inspired by (Zhang et al., 2020). Some tweaks to stabilize image feature extraction have been taken by *ResNet* (He et al., 2016).

Email: Edoardo De Matteis <dematteis.1746561@studenti.uniroma1.it>.

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

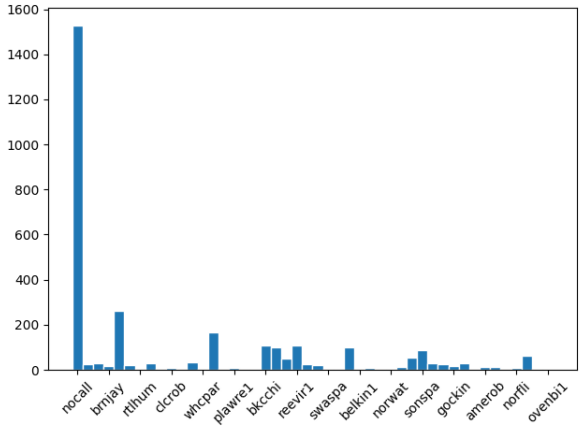

Figure 1. Primary label distribution in soundscape recordings.

3. Method

What happens when ornithologists recognize a birdcall, do they first detect a sound to then focus on classifying it or does their brain automatically do that subconsiously? It makes sense to believe it is possible to be so proficient to unconsciously recognize birds, as already happens with familiar sounds (Kirmse et al., 2009). I tried both the approaches but only the "subconscious" one is covered here due to computational issues with the other.

Preprocessing. The dataset is highly unbalanced (figure 1) so it was augmented by random oversampling and each 5 seconds frame is encoded as a spectrogram in mel scale. In audio samples there can be both plain ambient noise and birds chirping, thus transforming spectrograms (e.g. by random crops) can lead wrong labeling, besides that I wanted to avoid data losses due to them being sequential.

CNNAtt. This seq2seq layer expolits attention to grasp sequential knowledge: the spectrogram passes through three different convolution pipelines with each one being the query, key and value for the attention, it takes inspira-

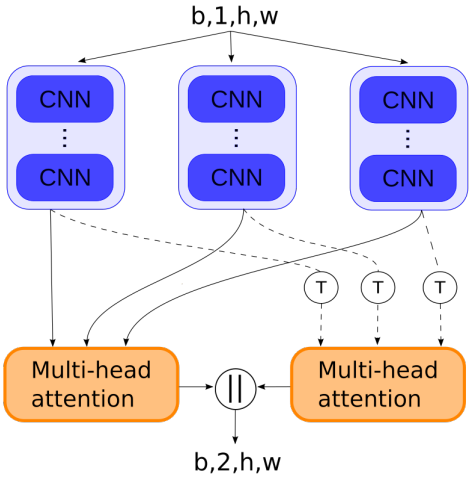

Figure 2. A CNNAtt block.

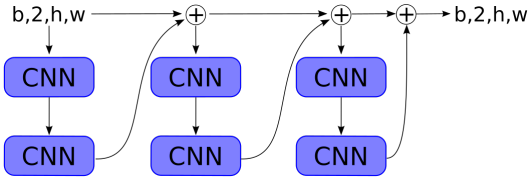

Figure 3. A CNNRes block.

tion by a *ResNeSt* block (Zhang et al., 2020) with 3 fixed cardinals. The RNNs in the attention layers work on tensor rows independently, looking at frequencies with no connection between them, thus a second attention layer works on spectrograms feature images' transpose; the two representations are then concatenated to a 2 channel image. When we refer to a CNNAtt$k$ convolutions have all the same kernel size k.

CNNRes. Residual networks help reducing errors in deep networks, this architecture employs six convolutions of the same kernel size with three skip connections between them (figure 3), n layers can be stacked one on top of the other and we call it CNNRes$n$. As a rule the number of kernels stays the same when the input image and the feature map have the same size, it is doubled when the latter is half the first (He et al., 2016). Between two different CNNRes block there are no connections thus no dealing with different dimensionalities.

The afore mentioned layers build up the feature extraction backbone, plus a convolution between CNNAtt and CNNRes to deal with the number of filters, after that we have m bilinear GRU layers to extract causal knowledge, and last a linear layer to output logits. We can ignore these two last heads when defining our architecture, due to them being common to all models.

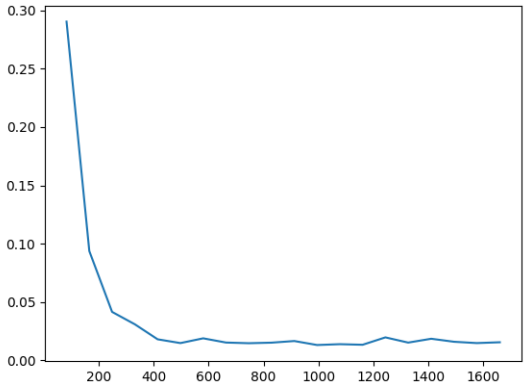

Figure 4. Validation loss for CNNRes2GRU1FC1.

4. Results and conclusions

In our experiments we will use CNNRes2 (with kernel sizes 3 and 5) and CNNAtt5CNNRes2 models, according to the previously defined rules, we also fine-tune pretrained models *ResNet18* and *ResNet50*, in which we add a convolution and a linear layer to match *ResNet*s input and output. Optimization is done by *Adam* algorithm on a cross entropy loss with small batches, both to gain a regularizing effect and because of memory limitations.

Results are reported at table 1, a small learning rate is needed to stabilize descent due to the high variance of 8-sampled batches. Since datasets have been balanced and all classes are equally important it makes sense to use accuracy.

Table 1. Performances for the validation set on 20 epochs with fixed seed, the test accuracy is reported for CNNRes2 only.

Model	LR	Loss ↓	Acc. ↑
ResNet50	1e-4	.52	82.3%
ResNet18	1e-4	.21	92.2%
ResNet18 (frozen)	todo	todo	todo%
CNNRes2	1e-4	**.02**	**99.1%**
CNNAtt5 + CNNRes2	1e-4	.04	98.7%
Test	-	-	99.4%

From the results we can confirm that convolutional neural networks do wonders when priors apply, while training attention is not worth the hassle. Occam's razor holds and simpler residual network seem to be better than fine-tuned models, yet we are transferring knowledge so we cannot say that with certainty.

Future works. Adding a threshold to consider secondary labels too, train on the whole dataset to see if it yields some

improvement and train the "conscious" model (section 3) to compare results.

References

Hamdy, A., Vedula, P. K., and Konduru, M. V. J. Audio separation and isolation: A deep neural network approach.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kirmse, U., Jacobsen, T., and Schröger, E. Familiarity affects environmental sound processing outside the focus of attention: An event-related potential study. *Clinical neurophysiology*, 120(5):887–896, 2009.

Michelashvili, M. and Wolf, L. Speech denoising by accumulating per-frequency modeling fluctuations. *arXiv preprint arXiv:1904.07612*, 2020.

Xie, J., Aubert, X., Long, X., van Dijk, J., Arsenali, B., Fonseca, P., and Overeem, S. Audio-based snore detection using deep neural networks. *Computer Methods and Programs in Biomedicine*, 200:105917, 2021.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., and Smola, A. Resnest: Split-attention networks, 2020.