Audio Bird Classification with Inception-v4 extended with Time and Time-Frequency Attention Mechanisms

Antoine Sevilla, Hervé Glotin

AMU, Univ. Toulon, CNRS, ENSAM, LSIS UMR 7296, DYNI team, France
antoine-sevilla@etud.univ-tln.fr, herve.glotin@univ-tln.fr

Abstract. We present an adaptation of the deep convolutional network Inception-v4 tailored to solving bioacoustic classification problems. Bird sound classification was treated as if it were an image classification problem by a transfer learning of Inception. Inception, the state-of-the-art in image classification, was used together with an attention algorithm, to (multiscale) time-frequency representations or images of bird sounds. This has resulted in an efficient pipeline, that we call Soundception. Soundception scored highest on all tasks in the BirdClef2017 challenge. It reached 0.714 Mean Average Precision in the task that asked for classification of 1500 bird species. To our knowledge Soundception is currently the most effective model for biodiversity monitoring of complex soundscapes.

Keywords: Deep Learning, Inception-v4, Bird Species Classification, Transfer Learning, Attention Mechanism, Sound Detection.

1 Introduction

The main objective of our approach is to create an easy-to-use pipeline of an acoustic model from an image model. We want to stay in the same framework of the state-of-the art deep learning [7] Inception-v4, and to transfer it to soundscape classification. Inception-v4 has been pre-trained on imagenet, then we adapt its inputs and learn a bird acoustic activity detector thanks to the classification outputs joint to an attention mechanism. Then we process a transfer learning from pre-trained weights on imagenet to bird classification. The paper describes our methodology to efficiently build this model, that we call 'Soundception', in few weeks a reduced GPU resources. We show that Soundception gives to the best scores of the BirdClef 2017 challenge [5]. In the last section we discuss on perspectives to increase the accuracy of Soundception.

2 Audio featuring

2.1 Audio to tri-channel time-frequency image

The data representation is a crucial step in any learning process. In our approach, the representation must be scalable and based on image processing including

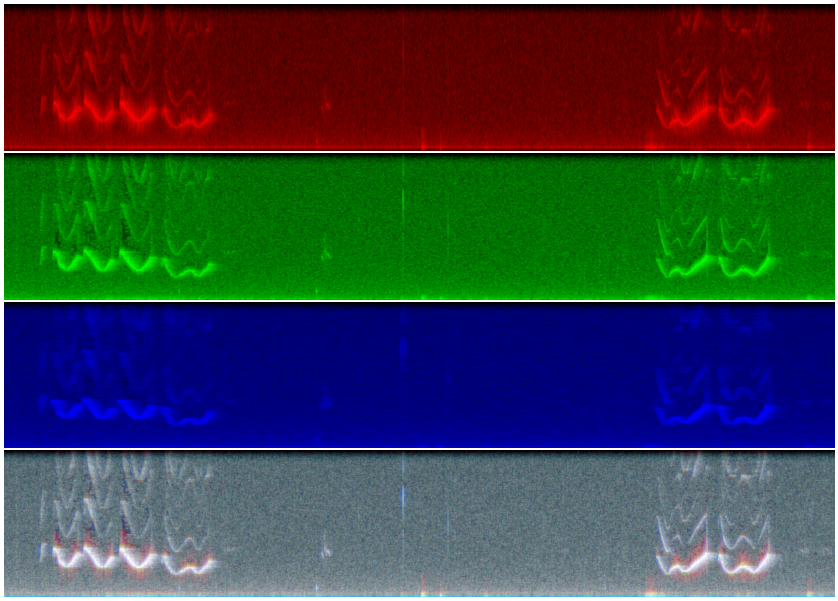

Fig. 1. From Top to Bottom, S1, S2, S3, resp. 128, 512, and 2048 bins FFT window spectrograms of a bird call. The fourth at the Bottom overlaps the three above and shows their complementarity.

some acoustic specificities, because we take advantage of Google Inception that is fed by a RGB image. Therefore, we generate three log-spectrograms by fast Fourier transform at three scales : window size of $w_{i\in\{0,1,2\}} = 2^{(2\times i)} \times 128$ (i.e. 128, 512, 2048). This fast computation approximates a compressed multi-scale representation of voices and chirps of birds. We think it improves usual spectral representations that deal with either temporal or frequency resolution (see usual representation in [4, 3]). Next, we reshape the three spectrograms by bilinear interpolation, into an optimal dimension for Inception inputs. In sum, our audio featuring is :

1. Resample the dataset to 22050 Hz sampling rate.
2. Let *min_duration* be the accepted minimum duration of audio sample.
3. Let *min_subduration* be the accepted minimum duration of the subimage.
4. Let $d(x)$ be the duration of the audio sample x.
5. While $d(x) < min\_duration$ self concatenate x.
6. Compute three log-spectrogram $S_i(x)$ with window sizes $w_i \in (128, 512, 2048)$.
7. Remove outliers of the $S_i(x)$ distribution to avoid quantification error.
8. Resize (bilinear interpolation) $S_i(x)$ to an optimal format for Inception: $height = 299$ pixels for the frequency dimension, $299 \times min\_subduration/1.5$ pixels for the time dimension.
9. Concatenate the three $S_i(x)$ into one 3 channels RGB multiscale image I.

2.2 Data augmentation

During the training stage, we run data augmentation. Therefore, we use standard transformations in computer vision. More precisely we run the Inception preprocessing on the spectrograms S_i as random hue, contrast, brightness, and saturation, plus random crop in time and frequency, as follows :

1. Random choice of an image I in the dataset.
2. Random crop a subimage I_c from I:
   - let hI_c = initial height of $I_c = 299$,
   - let dI_c = initial duration of $I_c = 15sec. \sim 299 \times 10$,
   - set random temporal dilatation factor of I_c uniformly sampled in $[0.95, 1.05]$,
   - set random_top of I_c uniformly sampled in $[0.96, 1] \times hI_c$,
   - set random_bottom of I_c uniformly sampled in $[0, 0.01] \times hI_c$,
   - crop and resize I_c from I with above parameters and random time offset.
3. Vision preprocessing of I_c by hue, contrast, brightness, saturation variations.
4. Add random noise or process local brightness to I_c.

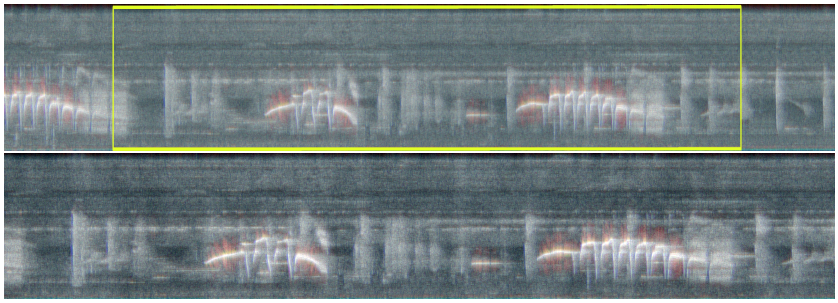

Fig. 2. Sample of a multiscale representation of bird activities, before (Top) versus after (Bottom) data augmentation.

3 Model specification

3.1 Transfer learning from Inception-v4 to Soundception

Inception-v4 is the state-of-the-art in computer vision [1]. There are several ways to adapt Inception-v4 to time-frequency analysis, as a simple average pooling on the time axis, or use recurrent layer at the top of the network or both. Here we adapt Inception-v4 to make it entirely convolutional on the time domain, the aim being to make it invariant to temporal translation, and to allow arbitrary width-sized image. Secondly we add a time and a time-frequency attention mechanisms into the branches as represented in the synopsis of Inception-v3 Fig 3.

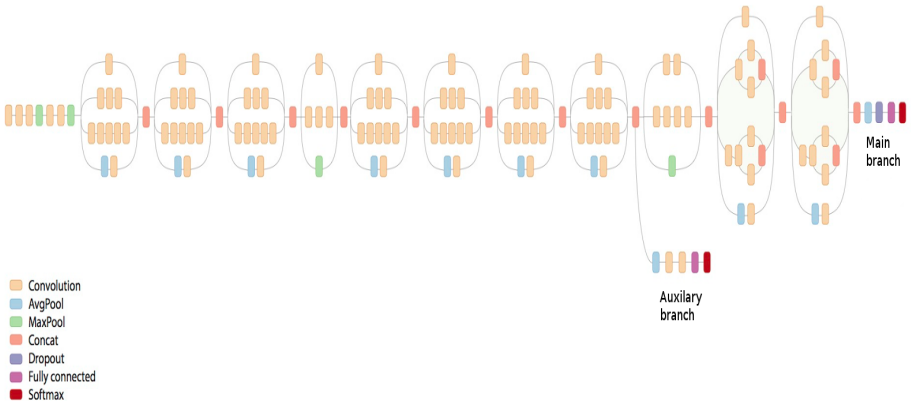

Fig. 3. Architecture of Inception-v3 (credit: [8]), similar to Inception-v4 model, showing the place of the branch where is connected each of the attention mechanism.

We set $dI_c = 15$ seconds, $min\_d = 60$ sec., scale = 1.5 sec. for 299 pixels in respect to baseline Inception inputs of 299×299 pixels, and the available VRAM per GPU (12 Go). We do not use specific bird detection in order to avoid handcrafted detection which could weaken the complete pipeline. The detection is processed by an attention mechanism as presented in the next section. It runs on a large time window ($dI_c = 15sec.$) to increase the probability of bird activity in the image. The main process results into one time frequency RGB image (as Fig. 1) per audio file, thus 36492 in BirdClef 2017.

3.2 Attention mechanisms in time and time-frequency

Attention mechanisms are gaining popularity. The goal is to focus attention somewhere or on something. We can learn detection from classification by adding a soft attention mechanism [2]. Thus, we add to the Inception model two attention mechanisms : a temporal attention into the auxiliary branch, and a time-frequency attention in the main branch.

Each attention mechanism is an element wise product of the detector feature map with the feature maps of the previous Inception layers. These mechanisms learn how to pass the information and thus play the role of bird activity detectors.

The first attention mechanism is the temporal attention defined by a sigmoid activation because the bird temporal activities are expected to be binomial in time.

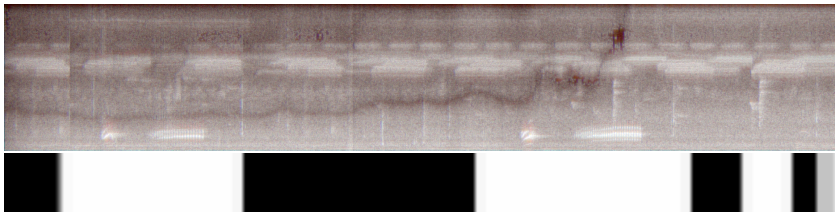

Fig. 4. Example of temporal attention on spectrogram (Top). The temporal attention is shown in white segments (Bottom).

The second attention mechanism is defined by the softmax of the outputs[1], yielding to smooth neuron time-frequency activity distribution.

Next, based on the sigmoid or softmax of the outputs, we compute the element wise product between the feature maps of the signal and the feature maps of the detector.

In Fig. 4 we show how Soundception focuses in time on bird activities. In Fig. 5 we show that in time-frequency Soundception indeed focuses on the loudest formant/frequency of the bird call.

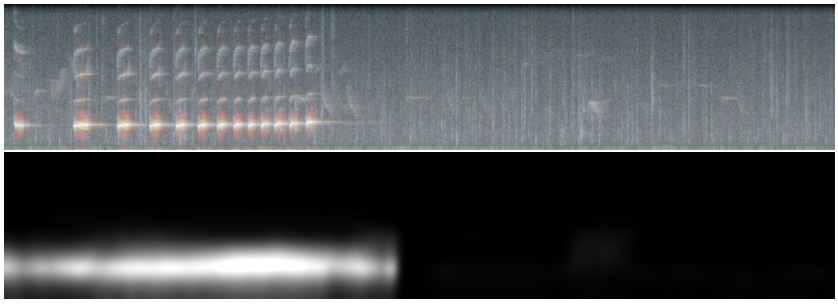

Fig. 5. Example of time-frequency attention : the spectrogram image (Top), the time-frequency attention (Bottom), in grey scale (highest attention is white, lowest is black).

3.3 Training stage

The training of the model took several days depends of hyper-parameters, using randomly split training (90%) and (10%) validation sets as in [4]. For the transfer learning, we first train only the top layer of Soundception, then we fine tune all layers. We split training in different stages with different batch sizes, according to the available GPU memory :

[1] Normalization with softmax is: $(\exp n_u)/\sum_v \exp n_v$.

1. Train the model with time window of $dI_c = 15sec.$,
2. Train last layers and detectors with mini-batch size of 8,
3. Fine tune all layers with mini-batch size of 4.

There are different options to evaluate the prediction on the development set. We could consider the audio file transform into images I as previously described with arbitrary temporal size. Here, we optimize the model according to the average score of the predictions from the subimages I_c of the main images I.

4 Results

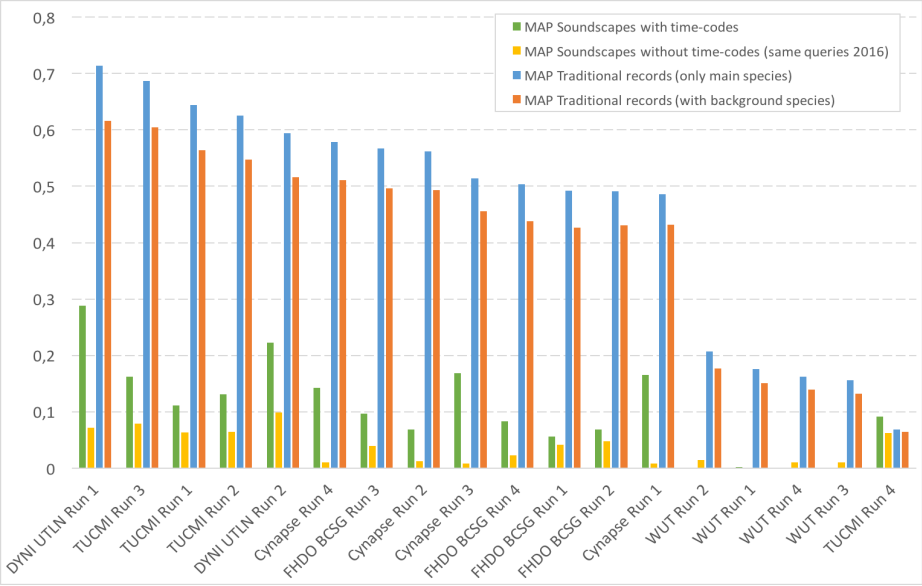

Fig. 6. Results of all the challengers in the BirdClef2017 international contest. The DYNI UTLN RUN1 is the best in three task categories : Soundscapes with time-code, and the two traditional classification tasks. It is third in the 'Soundscape without time-codes'. Complementary, the DYNI UTLN Run2 yields to lower scores in average but to the best score in the other task 'Soundscape without time-codes'.

We report Fig. 6 the official scores on the four tasks of each of the challenger. Our model Soundception wins this challenge in the four tasks. The DYNI UTLN RUN1 depicted in this paper is the best model in three of the tasks, with 0.714 Mean Average Precision (MAP) on the 1500 species 'traditional records' task, 0.616 MAP with 'background species' task, and 0.288 MAP on the 'Soundscapes with time-codes' task. It is third in the 'Soundscape without time-codes' task, for

which our other run (DYNI UTLN Run2) which explored different parameters is first.

These results are good despite the fact that we had not time to completely train Soundception on different topologies and to develop associated preprocessings in the four weeks of the challenge.

5 Conclusion and future work

In this paper we show how we transferred Inception-v4 from image to the acoustic domain, and how it learns bird sound detection by itself using attention models. The results show that it is possible to tackle state-of-the-art sound classification by the transfer learning of efficient pre-existing image classification model. This strategy can be useful to tackle other challenge without pre-segmentation.

We had not been able to let completely converge the training stage of Soundception due to the huge computation and GPU needs, however it reaches the best results in the BirdClef 2017 challenge. There is a lot to be done in this area. Our current work also explores different scalable optimizations to learn audio to image representations instead of pseudo multi-scale FFT spectrograms. We currently develop a model with stacked GRU at the top of the network.

Acknowledgements. We thank Laura Bessone for her support in this paper. We thank XenoCanto, LifeClef team, EADM GDR CNRS MADICS, SABIOD.org and Amazon Explorama Lodges with Lucio Pando, P. Bucur and M. Trone for the co-organization of this challenge. This research is also supported by STIC-AmSud BRILAAM for South American bioacoustics. We thank TPM, CG83, UTLN for their support in the Captile project on soundscape analysis. We thank V. Roger for setting the GPU, and S. Paris for lending them. We are grateful to the anonymous reviewers. Antoine Sevilla has set up the experimentations and ideas in this article.

References

1. C. Szegedy, S. Ioffe, V. Vanhoucke: Inception-v4, inception-resnet and the impact of residual connections on learning, arXiv:1602.07261 (2016)
2. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio. Show, attend & tell: Neural image caption generation with visual attention, ICML (2015)
3. H. Goeau, H. Glotin, WP. Vellinga, R. Planque, A. Joly. LifeCLEF Bird Identification Task 2016: The arrival of Deep learning, CLEF (2016)
4. E. Sprengel, YK. Martin Jaggi, T. Hofmann: Audio based bird species identification using deep learning techniques, CLEF (2016)
5. A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, J.-C. Lombardo, R. Planqué, S. Palazzo, H. Müller: LifeCLEF 2017 Lab Overview: multimedia species identification challenges (2017)
6. Xeno Canto Foundation: Sharing bird sounds from around the world, www.xeno-canto.org (2012)

7. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, et al.: TensorFlow: Large-scale machine learning on heterogeneous systems, tensorflow.org (2015)
8. https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html