END TO END RAW AUDIO DEEP LEARNING OF TRANSIENTS, APPLICATION TO BIOACOUSTICS

Maxence Ferrari, Hervé Glotin, Ricard Marxer

END TO END RAW AUDIO DEEP LEARNING OF TRANSIENTS, APPLICATION TO BIOACOUSTICS

Maxence Ferrari[1] **Hervé Glotin**[1] **Ricard Marxer**[1]

[1] Univ. Toulon, Aix Marseille Univ. CNRS, LIS, DYNI, Marseille, France

maxence.ferrari@univ-tln.fr, glotin@univ-tln.fr, ricard.marxer@lis-lab.fr

1. INTRODUCTION

In this paper, we propose a raw audio deep learning approach to click classification. The advantage of having a neural network that works directly on the audio samples is that there is no need to find manually the best representation (spectrogram, MEL bands, ...), and avoids the work needed to tune all the hyperparameters that come with those representations (window size, stride ...). The way in which the information is extracted is learned from the data. The model proposed, named UpDimV2 was built to answer the challenge DOCC10 [1,2] that we created based on data from the DCLDE challenge [3] and the Sphyrna expedition [4].

2. MATERIAL AND METHOD

The network is composed of multiple UpDim blocks. As shown in figure 1, an UpDim block is composed of two ResNet blocks [5] followed by the UpDim operation, which adds a dimension (also known as unsqueeze or expand_dim). Thus a 1D example with a width and feature dimension with the respective size of w and f will become a 2D example with a height, a width and a feature dimension, with their respective sizes of w, f and 1. The UpDim operator can then be applied once again to turn this 2D example into a 3D example.

The DOCC10 dataset consists of clicks centered in a window of 8192 samples. This was motivated by the possibility of analysing clicks in a window of 4096 samples while being able to offset this shorter window. The combination of DCLDE and Sphyrna Odyssey brought this new dataset to a total count of 134,080, that we split into a training set of 113,120 clicks and a test set of 20,960 clicks for the DOCC10 challenge, which produces an approximately 85-15 split. The test set is balanced with 2096 clicks per class. For the challenge, the test set was split into a private test set (90%) and a public test set (10%). This split was done randomly, so that the classes are no longer perfectly balanced to prevent participant to use this information. The training set is also perfectly balanced with 11,312 clicks per class.

3. RESULTS

Since the release of the DOCC10 challenge in early 2020, 28 challengers have participated. The full up-to-date leaderboard can be found on the Challenge Data website (`https://challengedata.ens.fr/participants/challenges/32/ranking/public`). The top two scores were obtained by the same team, who used a semi-supervised approach on the test set, hence the score gap with the other participants.

4. DISCUSSION AND CONCLUSION

An alternate version of the DOCC10 dataset, called DOCC7, has been generated. It has the same samples, but restricted to only 7 species, which are Gg, Gma, La, Mb, Me, Pm, and Zc. The reason for the removal of UDA and UDB is more straightforward. When the DCLDE dataset was made, they used clustering methods to detect the various species. These two labels were then given to dolphin species that could not be identified. We decided to leave them in the DOCC10 challenge since they still represent clicks that belong to groups of dolphins, even if they do not represent only one species, unlike the other labels. These clusters are also useful to train a classifier that would be used after a click detector, and prevent it from classifying these dolphin clicks as another species. However, trained networks (with various architectures from various labs) have shown that, unlike the seven other classes in DOCC7, the trained networks had lower accuracy on the UDA and UDB labels. We believe that the networks' prediction might not be wrong, meaning that these classes have a higher label noise. Finally, the Ssp were also removed for two reasons. Firstly, Stenella is a genus and not a species unlike the other remaining classes. Secondly, there seems to be a large covariate shift between the training and test sets for this class. A slew of reasons could explain this difference between the training and test set, such as different species, different groups, different types of clicks, or mislabeling. This model was also tested on the DOCC7 database, achieving an accuracy of 95.09%. Applications of this model will be conducted in the CARIMAM project.

5. ACKNOWLEDGEMENTS

This research was partly funded by AID Agency, Region Haut de France, ANR-18-CE40-0014 SMILES, MARITTIMO FEDER GIAS projects on advanced studies on whale-ship anti-collision system, and ANR-20-CHIA-

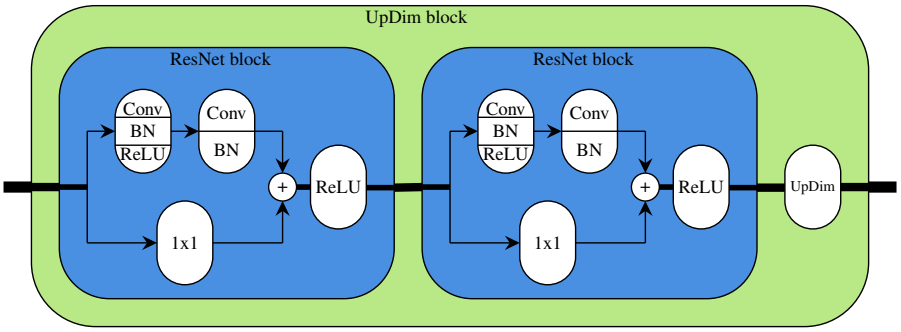

Figure 1. Architecture of an UpDim block

Layer name	Input size	Kernel	Strides	Out features
Conv-1D	$N * 4096 * 1$	3	1	32
Conv-1D	$N * 4096 * 32$	3	2	32
Skip	$N * 4096 * 1$	1	2	32
Conv-1D	$N * 2048 * 32$	3	2	64
Conv-1D	$N * 1024 * 64$	3	2	128
Skip	$N * 2048 * 32$	1	4	128
Conv-2D	$N * 1024 * 128 * 1$	3*3	1*1	32
Conv-2D	$N * 1024 * 128 * 32$	3*3	2*2	32
Skip	$N * 1024 * 128 * 1$	1*1	2*2	32
Conv-2D	$N * 512 * 64 * 32$	3*3	2*2	64
Conv-2D	$N * 256 * 32 * 64$	3*3	2*2	128
Skip	$N * 512 * 64 * 32$	1*1	4*4	128
Conv-3D	$N * 128 * 16 * 128 * 1$	3*3*3	1*2*1	32
Conv-3D	$N * 128 * 8 * 128 * 32$	3*3*3	2*2*2	64
Skip	$N * 128 * 8 * 128 * 1$	1*1*1	2*4*2	64
Conv-3D	$N * 64 * 4 * 64 * 64$	3*3*3	2*2*2	128
Conv-3D	$N * 32 * 2 * 32 * 128$	3*3*3	2*2*2	256
Skip	$N * 64 * 8 * 64 * 64$	1*1*1	4*4*4	256
Softmax	$N * 16 * 1 * 16 * 256$	16*1*1		
MaxPool	$N * 16 * 1 * 16 * 256$	16*1*1		
Flatten	$N * 1 * 1 * 16 * 256$			
Dense	$N * 4096$			1024
Dense	$N * 1024$			512
Dense	$N * 512$			7

Table 1. Topology of UpDimV2 model
Dimensions are given in NHWC order. Horizontal lines separate each residual block.

0014-01 national Chair in Artificial Intelligence for Bioacoustics ADSIL (H. Glotin).

6. REFERENCES

[1] M. Ferrari, H. Glotin, R. Marxer, and M. Asch, "Docc10: Open access dataset of marine mammal transient studies and end-to-end cnn classification," in *IJCNN*, 2020.

[2] M. Ferrari, *Study of a Biosonar Based on the Modeling of a Complete Chain of Emission-Propagation-Reception with Validation on Sperm Whales*. PhD thesis, Université de Picardie Jules Verne, Amiens, France, Sept. 2020.

[3] H. Glotin, J. Hildebrand, K. Dunleavy, and M. Roch, "Dclde challenge," pp. 134, Proceedings of DCLDE2018, Sorbonne Université, Université de Toulon Ed., sabiod.org, 2018.

[4] M. Ferrari, M. Poupard, P. Giraudet, R. Marxer, J.-M. Prévot, T. Soriano, and H. Glotin, "Efficient artifacts filter by density-based clustering in long term 3d whale passive acoustic monitoring with five hydrophones fixed under an autonomous surface vehicle," in *OCEANS 2019-Marseille*, pp. 1–7, IEEE, 2019.

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