MULTI RESOLUTION FEATURE FUSION FOR SUPERVISED DETECTION OF STRONGLY-LABELLED WHALE CALLS

Technical Report

Matija Marolt, Eva Boneš

University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia
{matija.marolt, eva.bones}@fri.uni-lj.si

ABSTRACT

We present the outline of our deep architecture for supervised detection of strongly- labelled whale calls. Our architecture is based on three parallel Swin transformers [1],that process the input audio on multiple time scales. The input audio is windowed into approx. 16 second long normalized chunks, which are processed by the learnable frontend for audio classification - LEAF [2] to obtain a time-frequency representation. Three different representations are calculated, using the same number of frequency channels but different window sizes. Each is processed by a different three-layer Swin transformer (using 4x4 patch sizes with 2x2 stride to obtain initial input patch tokens, and with patch merging between layers [3]). Output feature maps with different time resolutions are upscaled to the same time resolution and fused with a downscaling convolution in feature space. The final feature map is upscaled in time to the original time resolution, followed by the fully-connected classification layer. Trained on the provided training set and evaluated on the validation set, the architecture yields a micro-F1 score of 0.58 and a macro-F1 score of 0.64. To obtain the output for the task, the input context window is shifted by 0.5 seconds along the time axis and the outputs of the model are averaged for the same time instances. Two versions of the model's output are submitted - one using the softmax activation in the output layer (multi-class), and another using the sigmoid activation (multi-label) - the latter should result in greater recall with lower precision than the former.

1. REFERENCES

[1] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[2] N. Zeghidour, O. Teboul, F. D. C. Quitry, and M. Tagliasacchi, "Leaf: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[3] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.