Master 2 Internship Proposal

Advisors: Jules Cauzinille, Benoît Favre, Arnaud Rey

November 2023

Deep transfer knowledge from speech to primate vocalizations

Keywords: Computational bioacoustics, deep learning, self-supervised learning, transfer knowledge, efficient fine-tuning, primate vocalizations

1 Context

This internship takes part in a multidisciplinary research project aimed at bridging the gap between state of the art deep leaning methods developed for speech processing and computational bioacoustics. Computational bioacoustics is a relatively new research filed which proposes to tackle the study of animal acoustic communication with computational approaches Stowell [2022]. Recently, bioacousticians are showing increasing interest for the deep learning revolution embodied in transformer architectures and self-supervised pre-trained models, but much investigation still needs to be carried out. We propose to test the viability of self-supervision and knowledge transfer as a bioacoustic tool by pre-training models on speech and using them for primate vocalisation analysis.

2 Problem Statement

Speech based models are able to reach convincing performance on primate-related tasks including segmentation, individual identification or call type classification Sarkar and Doss [2023] as they are with many different downstream tasks (such as vocal emotion recognition Wang et al. [2021]). We have tested publicly available models such as HuBERT Hsu et al. [2021] and Wav2Vec2 [Schneider et al., 2019], two self-supervised speech-based architectures, on some of these tasks with Gibbon vocalizations 1. Our method involves probing and traditional fine-tuning of these models.

Figure 1: A Siamang Gibbon vocalizing

As to ensure true knowledge transfer from pre-training speech datasets to the downstream classification tasks, the goal of this internship will be to implement efficient fine-tuning methods in a similar fashion. These will allow to limit and control the amount of information lost in the fine-tuning process. Depending on the interests of the candidate, the methods can include prompt

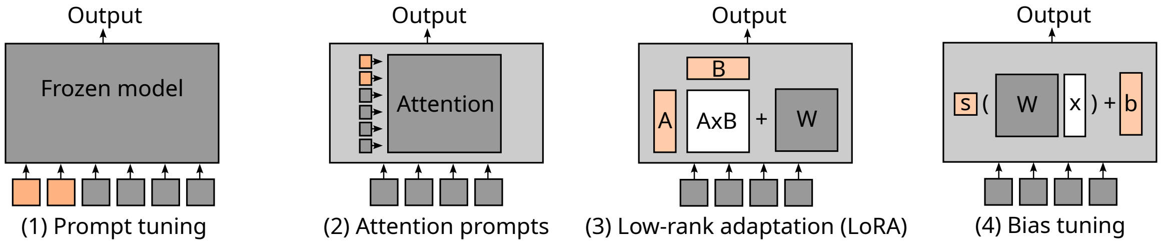

Figure 2: Some efficient fine-tuning methods

tuning Lester et al. [2021], attention prompting Gao et al. [2023], low rank adaptation Hu et al. [2021] or adversarial reprogramming Elsayed et al. [2018].

The candidate will also be free to explore other methods relevant to the question at hand, either on Gibbons or other species data-sets currently being collected.

3 Profile

The intern will propose and implement the efficient fine-tuning solutions on an array of (preferably self-supervised) acoustic models pre-trained on speech or general sound such as HuBERT, Wav2vec, WavLM, VGGish, etc. Exploring adversial re-programming of models pre-trained on other modalities (images, videos, etc.) could also be carried out. The work will be implemented using pytorch. The candidate must have the following qualities :

- Excellent knowledge of deep learning methods
- Extensive experience with PyTorch models
- An interest in processing bioacoustic data
- An interest in reading and writing scientific papers as well as some curiosity for research challenges

The internship will last 6 months at the LIS and LPC laboratories in Marseille during spring 2024. The candidate will work in close collaboration with Jules Cauzinille as part of his thesis on "Self-supervised learning for primate vocalization analysis". The candidate will also be in contact with the researchers community of the ILCB.

4 Contact

Please send a CV, transcripts and a letter of application to jules.cauzinille@lis- lab.fr, benoit.favre@lis-lab.fr, and arnaud.rey@cnrs.fr. Do not hesitate to contact us if you have any question (or if you want to hear what our primates sound like).

References

Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks, 2018.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1, 2021. doi: 10.1109/TASLP.2021.3122291.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

Eklavya Sarkar and Mathew Magimai Doss. Can Self-Supervised Neural Networks Pre-Trained on Human Speech distinguish Animal Callers?, May 2023. arXiv:2305.14035 [cs, eess].

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469, 2019. doi: 10.21437/Interspeech.2019-1873.

Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. 10:e13152, 2022. ISSN 2167-8359. doi: 10.7717/peerj.13152. URL `https://peerj.com/articles/13152`.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *CoRR*, abs/2111.02735, 2021. doi: 10.48550/arXiv.2111.02735.