PhD Proposal: Efficient and Effective SSL Models for Speech

Yannick Estève and Marco Dinarelli and Titouan Parcollet

Juillet 2022

Starting date: Flexible.
Application deadline: Running until the position is filled.
Salary: 2 240 € gross/month (social security included)
Mission: research oriented (teaching possible but not mandatory)

Keywords: natural and spoken language processing, self-supervised learning, deep learning efficiency.

1 Context

The ANR project E-SSL (Efficient Self-Supervised Learning for Inclusive and Innovative Speech Technologies) will start on November 1st 2022. Self-supervised learning (SSL) has recently emerged as one of the most promising artificial intelligence (AI) methods as it becomes now feasible to take advantage of the colossal amounts of existing unlabeled data to significantly improve the performances of various speech processing tasks.

2 PhD Objectives

This PhD program aims at providing efficient deep neural network architectures training to approach SSL with the speech modality while keeping model deployment in mind, e.g. some use-cases may require streaming capabilities. In this extent, this objective is decoupled in two axes with a gradually increasing difficulty. First, the PhD candidate will leverage the existing and large literature from deep architecture efficiency to quickly, yet systematically, document the effects of carefully selected architecture alterations with the latest SSL architectures for speech. Second, the candidate will develop novel deep architectures with respect to different efficiency parameters including local storage or compute capability. The impacts of the latter goals will consistently be evaluated both at training and deployment stages to cover the whole life-cycle of the model.

2.1 Research Directions

The latest and best performing SSL models for speech rely on the transformers neural networks. Hence, the PhD candidate will start with replacing blocks of transformers that have been identified as being particularly computation demanding with more efficient alternatives. A vanilla implementation of the self-attention layer induces a quadratic scaling of the computation and memory complexities of the model with respect to the input sequence length. This is particularly problematic for speech as typical training utterances vary a lot in the time domain. To tackle this phenomenon, two directions can be followed:

1. Reduce the time-dependency of the scoring part of the self-attention layer.

2. Reduce the floating point operation complexity of the self-attention layer.

The former concept will be implemented by drawing inspirations from Luna [3]. Hence, the candidate will integrate existing or novel modules into the self-attention layer that decouple the inner query and key products from the time dimension reducing it to a fixed and controllable dimension. The second axis will be evaluated following low-rank decomposition of the self-attention cell methods like used in the Linformer [5], or entire novel attention cells like adopted in FNet [1], that are built with linear complexity in mind.

The candidate will also investigate alternatives to the transformers architecture. Indeed, and according to a recent trend in the literature, transformers seem to not be all we need. Not only self-attention layers are more computationally demanding than, for instance, convolutional layers, but they may also achieve degraded performance. For instance, mixing tokens based on the Fourier transform [1], mixing linear layers [4] or convolutional layers [2] defeated transformers with much lower computational and memory burdens: MLP-Mixer [4] has a five times higher images-per-second throughput than the closest transformer. We hypothesize that there is no reason for the speech SSL community to stick with highly inefficient transformers. Thus, the candidate will carefully design both novel convolutional and mixing SSL models leveraging recent advances like EfficientNetV2 and MLP-Mixer [4].

3 Skills

- Master 2 in Speech Processing, Natural Language Processing, Computer Science or Data Science.

- Good mastering of Python programming and deep learning frameworks.

- Previous experience in Self-Supervised Learning, acoustic modeling or ASR would be a plus.

- Very good communication skills in English.

- Knowledge of French would be a plus but is not mandatory.

4 Scientific Environment

The thesis will be conducted within the Speech and Language Group (SLG) of the the LIA laboratory and the Getalp team of the LIG laboratory. The GETALP team and the LIA have a strong expertise and track record in Natural and Spoken Language Processing. The recruited person will be welcomed within the teams which offer a stimulating, multinational and pleasant working environment. The means to carry out the PhD will be provided both in terms of missions in France and abroad and in terms of equipment. The candidate will have access to the GPU cluster of both laboratories. Furthermore, access to the National supercomputer Jean-Zay will enable to run large scale experiments. The PhD position will be co-supervised by Prof. Yannick Estève (LIA, Avignon) and Dr. Marco Dinarelli (LIG, Université Grenoble Alpes). Joint meetings are planned on a regular basis and the student is expected to spend time in both places. Moreover, the PhD student will collaborate with several team members involved in the project in particular the two other PhD candidates from the E-SSL project, and the partners from LIA, LIG and Dauphine Université PSL, Paris. Finally, the project will involve Dr. Titouan Parcollet, co-creator of SpeechBrain, with whom the candidate will interact closely.

5 Instruction for applying

Applications must contain: CV + letter/message of motivation + master notes + possibly one or more recommendation letters (not mandatory, but it is a plus); and must be addressed to (all recipients): Titouan Parcollet (titouan.parcollet@univ-avignon.fr), Yannick Estève (yannick.esteve@univ-avignon.fr) and Marco Dinarelli (marco.dinarelli@univ-grenoble-alpes.fr).

References

[1] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. FNet: Mixing Tokens with Fourier Transforms. *arXiv preprint arXiv:2105.03824*, 2021.

[2] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *To appear in CVPR 2022*, 2022.

[3] Xuezhe Ma, Xiang Kong, Sinong Wang, Chunting Zhou, Jonathan May, Hao Ma, and Luke Zettlemoyer. Luna: Linear unified nested attention. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.