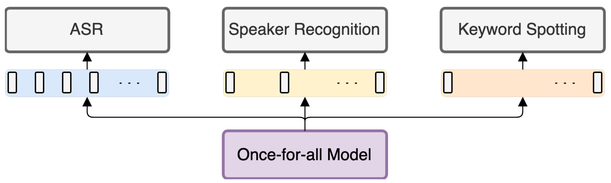
Once-for-All Sequence Compression for Self-Supervised Speech Models

Hsuan-Jui Chen*, Yen Meng*, Hung-yi Lee

National Taiwan University

Introduction

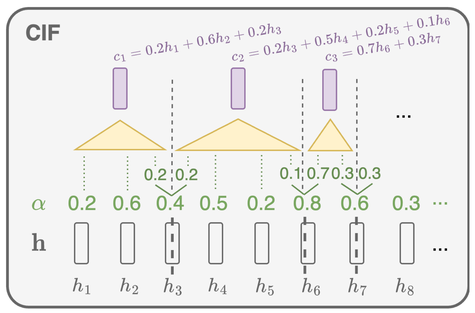
1. Reducing sequence length reduces computation.
2. Different tasks have different tolerance to sequence compressing.
3. Proposed a once-for-all (OFA) framework that supports different sequence compressing rates.



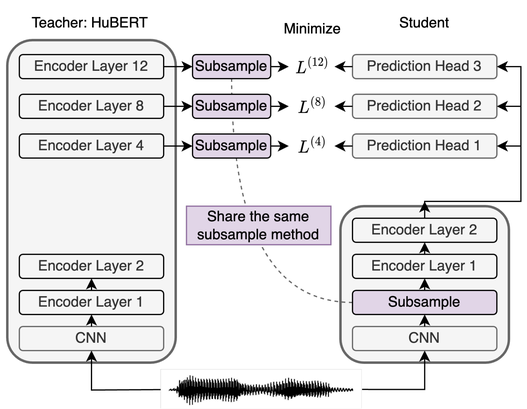
Background

Continuous Integrate-and-Fire (CIF) [Dong et al., 2020]

Fire when the accumulated sum of α reaches 1.

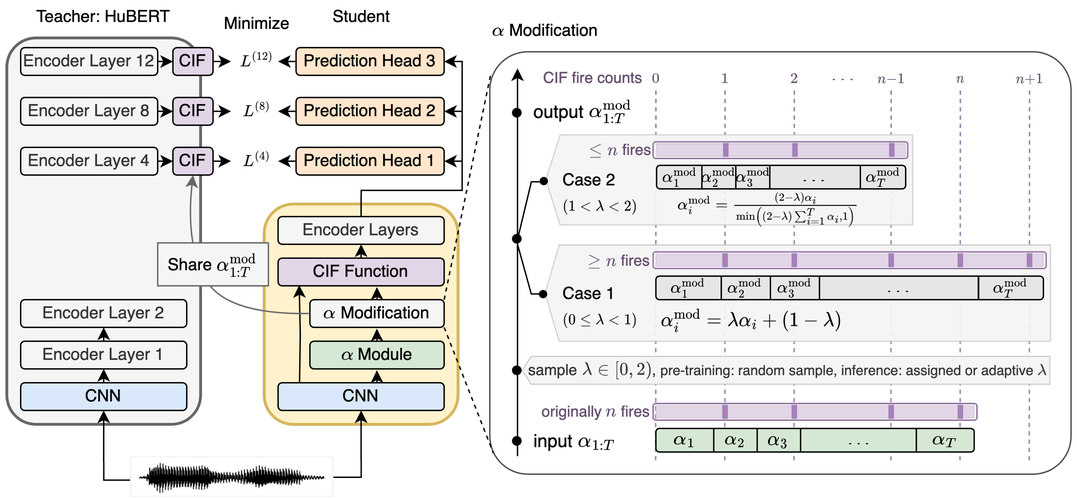


Variable-Length Subsampling [Meng et al., 2023]



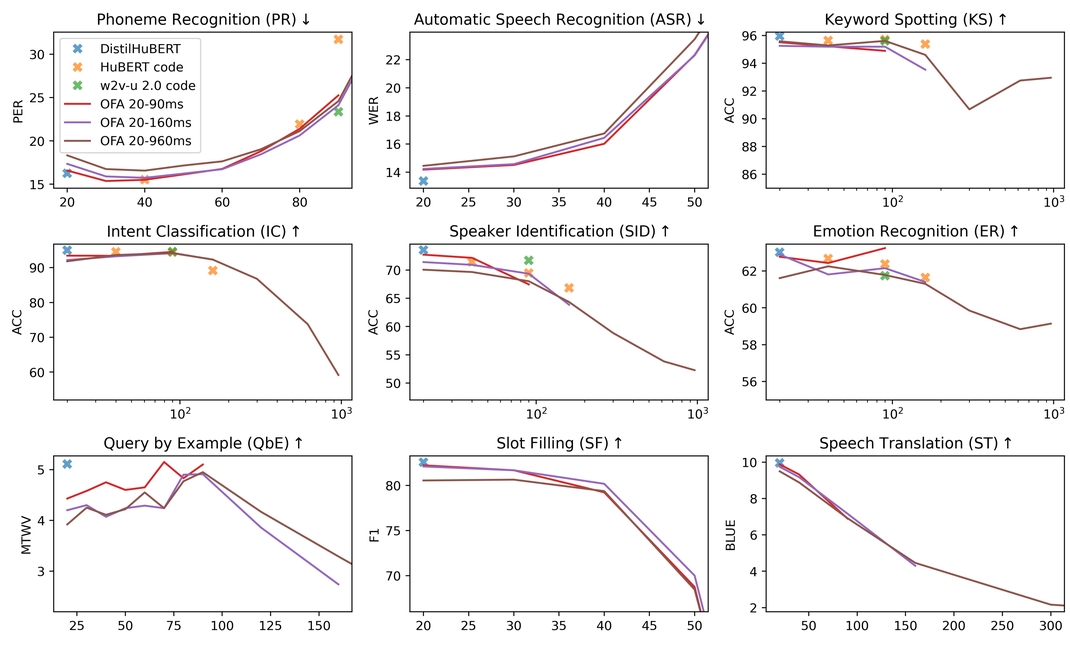
Once-for-all Sequence Compression

An α modification module is added to control the compressing rate. At each pre-training step, the compressing rate (controlled by λ) is randomly sampled.



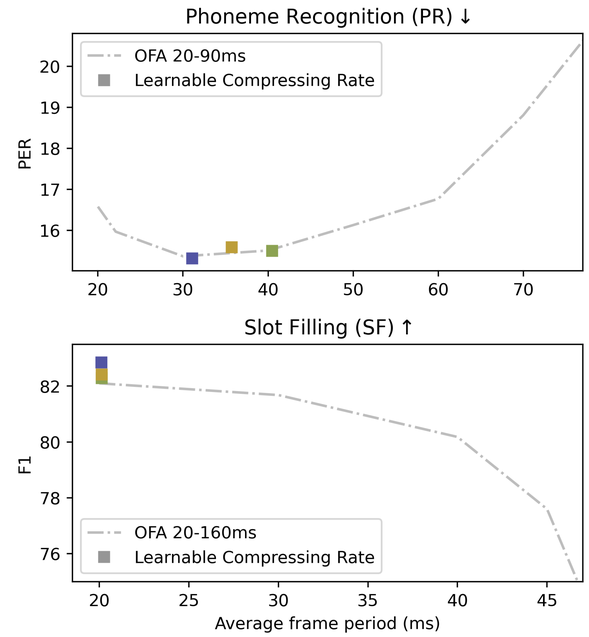
Main Results

Evaluation on a subset of the SUPERB benchmark [Yang et al., 2021] with manually selected λ. The *crosses* are models from previous work [Chang et al., 2022, Meng et al., 2023] with fixed compressing rates.



Adaptive Compressing Rate Learning

Treat λ as a downstream tuneable parameter.



Discussions

Main Results

1. The OFA models perform on par with single compressing rate models from previous works.
2. Different downstream (CTC, seq2seq, pooling) have different tolerance to sequence compressing.

Adaptive Compressing Rate Learning

1. With adaptive compressing rate learning, an overall best result can be obtained without grid-search.

Acknowledgments

We thank the Taiwan Web Service and the National Center for High-performance Computing (NCHC) for providing computing and storage resources.