Fine-tuning Strategies for Faster Inference using Speech Self-Supervised Models: A Comparative Study

Salah Zaiem[1,4] Robin Algayres[2] Titouan Parcollet[3] Slim Essid[1] Mirco Ravanelli[4]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France [2]COML, ENS-INRIA, PSL, Paris
[3]Samsung AI Research, Cambridge, UK [4] MILA, Montréal, Canada

Samsung AI Center-Cambridge

Motivation

- Using self-supervised representations seems crucial in low-resource scenarios.
- Most powerful models are large and induce long inference times.
- Can we, during fine-tuning, shrink the model or the inputs to enable faster inferences without a significant impact on the performance?

Global Setting

- SSL Model : WavLM Large, fine-tuned.
- Linear decoder head trained with character-level CTC loss. Results are shown with greedy or LM-rescored decoding.

Early-exit techniques and results

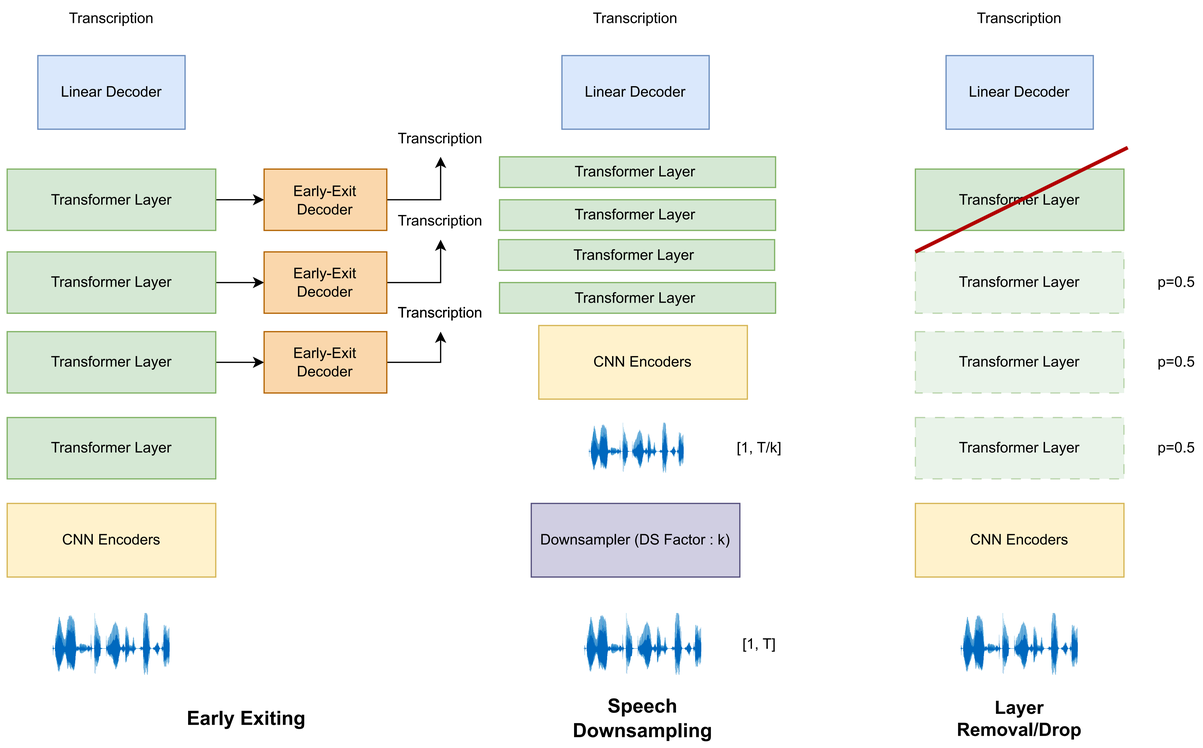
Two heuristics for threshold-based exiting :

- Entropy of probabilities of characters at early-exit decoder i.
- Cosine similarities between successive layers' representations.

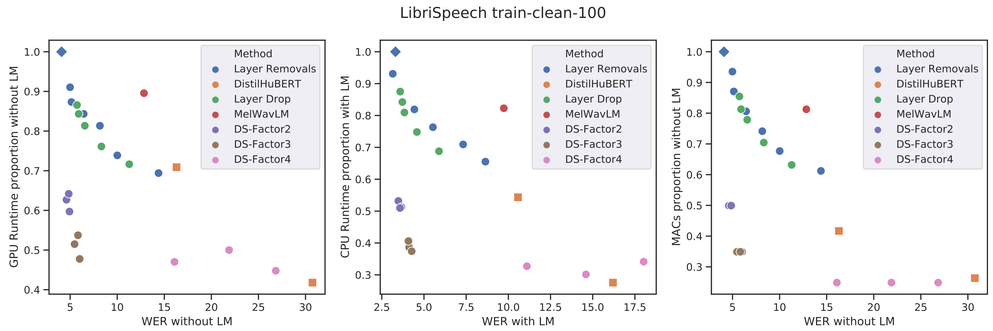
Technique		WER ↓	GPU (s)	CPU (s)
Baseline	**Full Model**	4.09	134	1121
Early Exit : Entropy Threshold	**Mean Exit Layer**			
0.06	13.80	12.08	96	757
0.03	17.61	7.67	116	974
0.025	20.52	6.66	128	1127
0.01	23.98	6.20	142	1275
Early Exit : Layer Sim. Threshold	**Mean Exit Layer**			
0.92	15.97	10.23	99	812
0.95	17.18	8.78	104	850
0.965	21.44	6.79	120	1070
0.98	24.00	6.20	128	1153
Two Steps EE : Layer Sim. Threshold	**Mean Exit Layer**			
0.955	13.97	25.29	95	798
0.96	14.52	21.95	102	866
0.97	21.46	6.17	126	1138
0.98	23.0	4.54	130	1175

⟶ The early-exits using the proposed heuristics severely harm the downstream performance.
⟶ Training the early-exits simultaneously with the model weights leads to poor final-exit performance.

Compared approaches

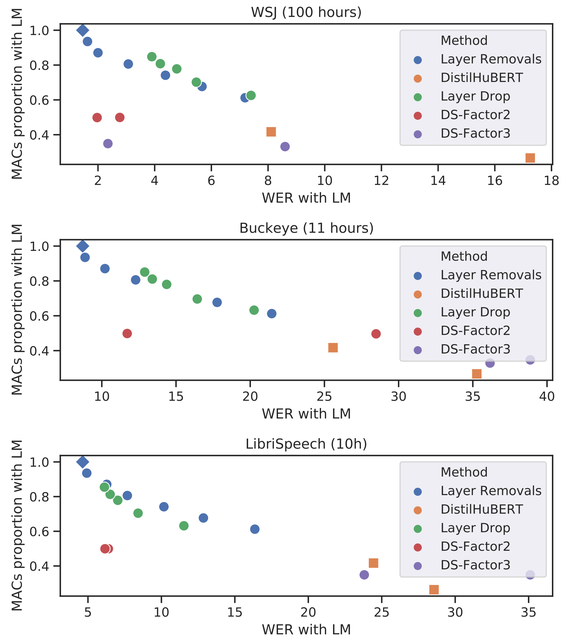


Results on LibriSpeech *train-clean-100*



⟶ Downsampling approaches lead to high inference time gains with low performance drop: 61.3% MACs drop with an WER increase of only 0.81.
⟶ Preferable to the use of distilled/smaller SSL models (not true for memory issues).

Robustness to dataset change



⟶ Downsampling performance is the most harmed by smaller available fine-tuning annotated data.

Datasets

- LibriSpeech *train-clean-100* and 10-hour splits.
- Buckeye, 11 hours of spontaneous english.
- Wall Street Journal (100 hours sample of mix between WSJ0 and WSJ1).

Take-home messages

- With a reasonable amount of annotated data, downsampling your inputs allows substantial efficiency gains with low performance drops.
- Code is available on github and within the SpeechBrain library for replication and further investigations.