Continual-wav2vec2: an Application of Continual Learning for Self-Supervised Automatic Speech Recognition

Samuel Kessler[†*], Bethan Thomas[‡] and Salah Karout[‡]

[†]University of Oxford, [‡]Huawei R&D Cambridge

Objectives

- Learn a new self-supervised language representation more *efficiently* by re-using previous representations.
- Retain performance and prevent *forgetting* of the 1-st task when learning a new task.

Continual Learning

$$\mathcal{T}_1 \rightarrow \mathcal{T}_2 \rightarrow \mathcal{T}_3 \rightarrow \dots$$

wav2vec2.0

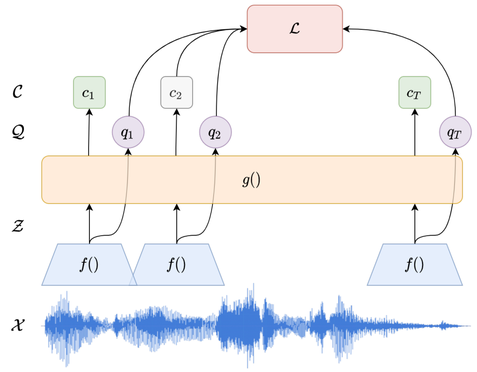
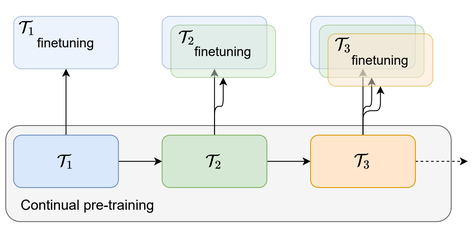


Figure 1:The wav2vec2.0 model takes in as input raw waveform $\mathcal{X}$, $f(\cdot)$ is a convolutional feature extractor and $g(\cdot)$ is a masked transformer encoder.

Continual Learning for SSL applied to ASR



wav2vec2.0 (cont.)

- SSL to learn speech representation then finetune on small labelled dataset [1].
- Expensive! Training takes 2 days on 64 GPUs.
- What if we want to learn a new language representation?
- Learning different language representations can be different $\mathcal{T}$.

Wav2vec2.0 components:

- CNN feature extractor $f : \mathcal{X} \rightarrow \mathcal{Z}$.
- MHSA context network $g : \mathcal{Z} \rightarrow \mathcal{C}$ [2].
- Discrete quantizer $\mathcal{Z} \rightarrow \mathcal{Q}$.

Continual-wav2vec2.0

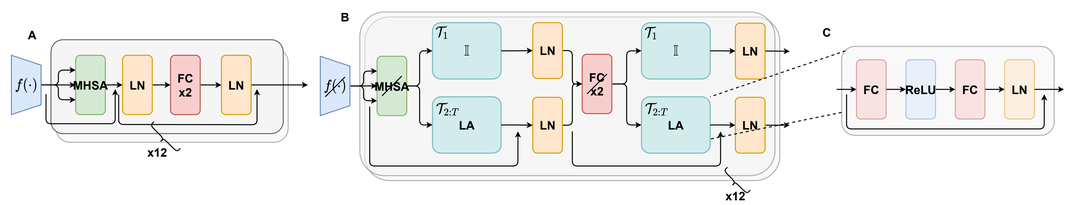
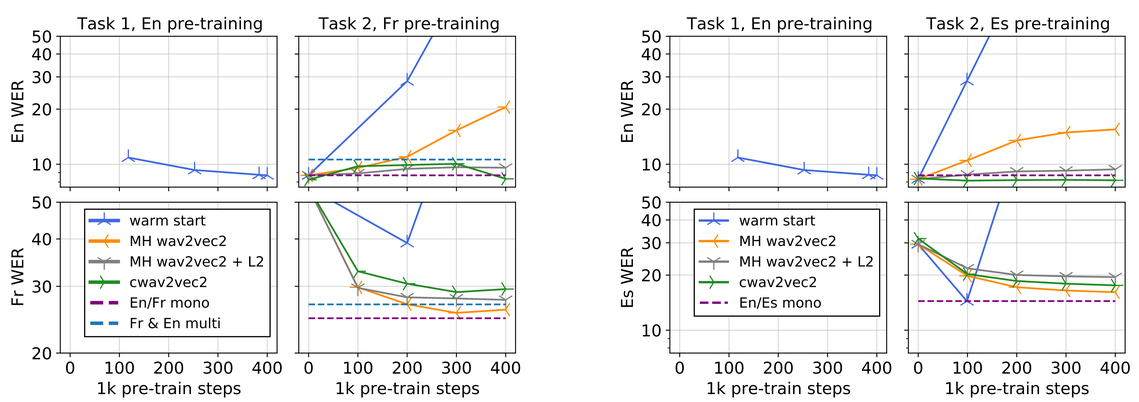


Figure 2: **A**, MHSA layer in the context network, $g(\cdot)$, of wav2vec2.0. **B**, cwav2vec2.0 layers of $g(\cdot)$ with LAs. **C**, LA module.

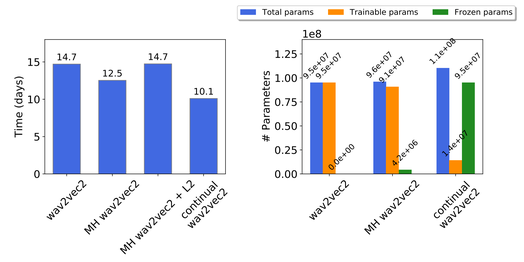
Main Results



Language Adapters

- Adapters are intermediate FC layers which are inserted into a deep network and allow adaptation to a new task.
- Language Adapters (LAs) have been shown to be a parameter efficient way to allow BERT to adapt to a new task [3].
- We also use LAs to allow wav2vec2.0 to learn a new self-supervised language representation efficiently.
- This is a *modular approach* to Continual Learning.

Efficiency



Conclusion

We have introduced continual-wav2vec2.0 which solves both of our original objectives. It **1)** is able to learn a new language representation successfully and do so efficiently by decreasing training times from around 15 to 10 days. Also **2)** it is able to completely prevent *forgetting*. Future directions are to scale to a another 1 or 2 tasks and explore mechanisms that allow LAs to combine.

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli.
wav2vec 2.0: A framework for self-supervised learning of speech representations.
arXiv preprint arXiv:2006.11477, 2020.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[3] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly.
Parameter-efficient transfer learning for nlp.
In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

*Work done during an internship at Huawei R&D Cambridge.