Continual Learning for On-Device Speech Recognition using Disentangled Conformers

Anuj Diwan¹, Ching-Feng Yeh², Wei-Ning Hsu², Paden Tomasellos², Eunsol Choi¹, David Harwath¹, Abdelrahman Mohamed²
¹University of Texas at Austin, ²Meta Inc.

1. Overview

ASR models deployed in households encounter ever-changing speaker distributions. Given a base ASR model (trained on a general-purpose dataset), we would like to build and evaluate models that can continually adapt as new speaker-specific data is received, in an efficient manner (for on-device adaptation). Our contributions are two-fold:

1. **Overview**

   - **Evaluation**: Our LibriContinual ASR benchmark
   - **Modelling**: Our DisConformer model with NetAug for Base ASR training and DisentangledCL for Continual Learning

2. **DisConformer**

   DisConformer splits the parameters of the FFN, Self-Attention and Conv modules of the Conformer into core and augment parameters.

   1. **Base ASR Training with NetAug**
      - Pass inputs through just core subset (term 1) as well as core + random subset of augment (term 2)

   2. **Continual Learning with LibriContinual**
      - Freeze core, finetune only a fixed, small, random subset of augment

   Use \( W_{\text{core}} \) for general-purpose and \([W_{\text{core}}, W'_{\text{aug}}]\) for speaker-specific ASR

3. **LibriContinual & Evaluation Metrics**

   - **Data Source**: 118 diff. speakers reading LibriVox books; transcripts generated by wav2vec2.0
   - **Data Splits**: Train: 10m, 30m, 1h, 2h, 5h, 10h ; Val: ~3.13h ; Test: ~2.66h for every speaker
   - **Increasingly-sized train data simulates continual interaction**

   **Evaluation Framework**
   - **1. Base ASR Training**: Train a base ASR model \( M \) on a general-purpose dataset (Librispeech)
   - **2. Continual Learning**: Given a continual learning algorithm \( A \), run it on the base ASR model using the LibriContinual train set of every speaker \( s \) to obtain 118 different ASR models \( M^{(s)} \)

   **Evaluation Metrics**
   - **1. #Params**: # Avg. trainable parameters modified by the CL algorithm \( A \) (proxy for efficiency)
   - **2. LibriContinual WER**: Median WER of model \( M^{(s)} \) on its respective speaker \( s \) test set
   - **3. Librispeech WER**: Median WER of model \( M^{(s)} \) on Librispeech; tests catastrophic forgetting

4. **Key Results**

   **DisCo-* models** disentangle each module type individually. **Base-* baselines** are DisCo-* models with just the core

   1. **NetAug trains better base ASR models**

      | Model   | LibriSpeech | LibriContinual |
      |---------|-------------|----------------|
      |         | test-c      | test-o         | val | test |
      | Base-FF | 4.02        | 10.16          | 7.92 | 8.36 |
      | DisCo-FF| 3.75        | 9.82           | 7.41 | 7.82 |

   **2. DisCL outperforms CL baselines on Librispeech**

   **3. DisCL outperforms parameter-matched CL baselines, and even performs as well as fully-finetuned baselines on LibriContinual**

5. **Conclusion & Future Work**

   LibriContinual reveals that current base ASR models underperform on speaker-specific data and current baseline CL algorithms are parameter-inefficient and catastrophically forget general-purpose data; on the other hand, our DisConformer with NetAug and DisCL is parameter-efficient and has high performance across the board! We invite future work on continual learning in absence of labelled data, multi-speaker adaptation, and more!