Joint Unsupervised and Supervised Training for Multilingual ASR

Junwen Bai[1] Bo Li[2] Yu Zhang[2] Ankur Bapna[2] Nikhil Siddhartha[2] Khe Chai Sim[2] Tara Sainath[2]

[1] Cornell University [2] Google

TLDR

- Multilingual ASR is concerned with dealing with multiple languages with one model
- Combine self-supervised losses with supervised loss to jointly train a powerful multilingual ASR system

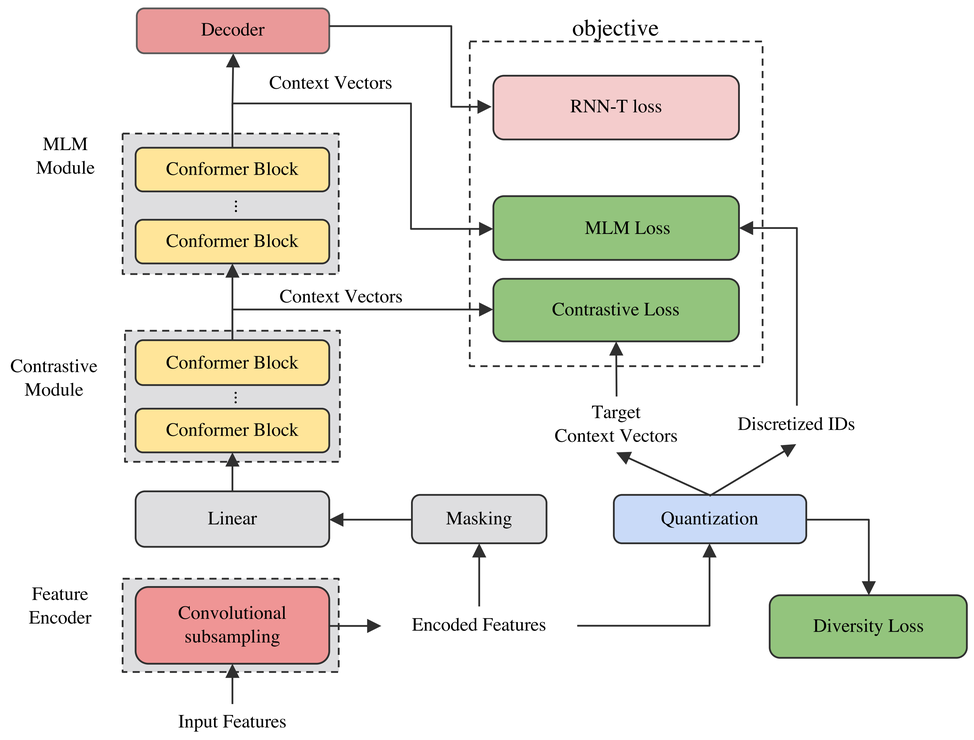
1. Motivations

- Training, deploying and maintaining one model per language, especially on long tail of low-resource languages, can quickly become cumbersome as the number of languages increases
- A single model for all languages can simplify the production pipeline significantly
- Training multilingual ASR models on a small set of similar languages can improve recognition performance
- Support the use case of code-switching

2. Dataset

- Multilingual LibriSpeech (MLS)
- English(en), German(de), Dutch(nl), Spanish(es), French(fr), Portuguese(pt), Italian(it), Polish(pl)
- Extremely imbalanced:
  - English has up to 44.6k hrs
  - Portuguese and Polish only have as low as ~100 hrs
- All the audio data are downsampled from 48kHz to 16kHz

3. JUST



Overview

- Self-supervision: contrastive loss, MLM
- Supervision: RNN-T

Feature encoder

- Conv downsample
- 2 Convolutional layers
- Extract latent representations from the surface features (log-mel filter bank)

Quantization

- Encoded features are passed to a quantizer without masking
- Quantizer “summarizes" all the latent speech representations to a finite set (codebook) of representative discriminative speech tokens
- Output both the quantized token + token ID

Contrastive Module

- A stack of Conformer blocks
- Read the encoded features with masking
- Extract context vectors from feature encoder output for computing the w2v2 contrastive loss

MLM Module

- Adapted from w2v-bert
- Continue to extract context vectors (from the contrastive module’s output) for computing the MLM loss
- Cross-entropy with ground-truth token IDs

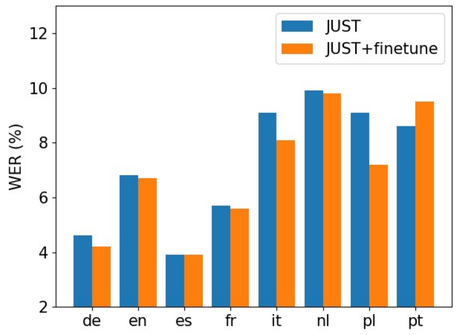
Decoder Module

- 2-layer LSTM
- RNN-T loss

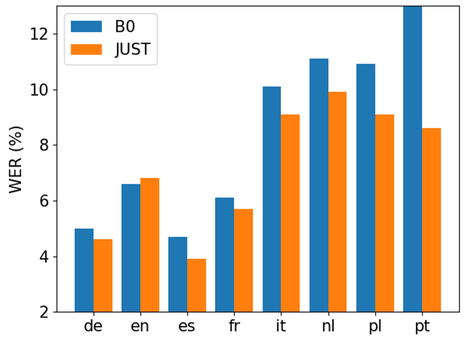
4. Experiments

- On average WER of all 8 languages, all JUST-based methods outperform previous works. In particular, JUST outperforms the monolingual baseline with 5-gram LM by 33.3%, XLSR-53 by 32.0%, B0 by 18.2%, E3 by 8.8%

JUST vs JUST+finetune



Per Language



Local vs Global Attention