Cross-Lingual Transfer Learning for Automatic Speech Recognition

Rasta Tadayon rasta@ucsb.edu
Shinda Huang shinda@ucsb.edu
Nawel Alioua nawel@ucsb.edu

Problem Description

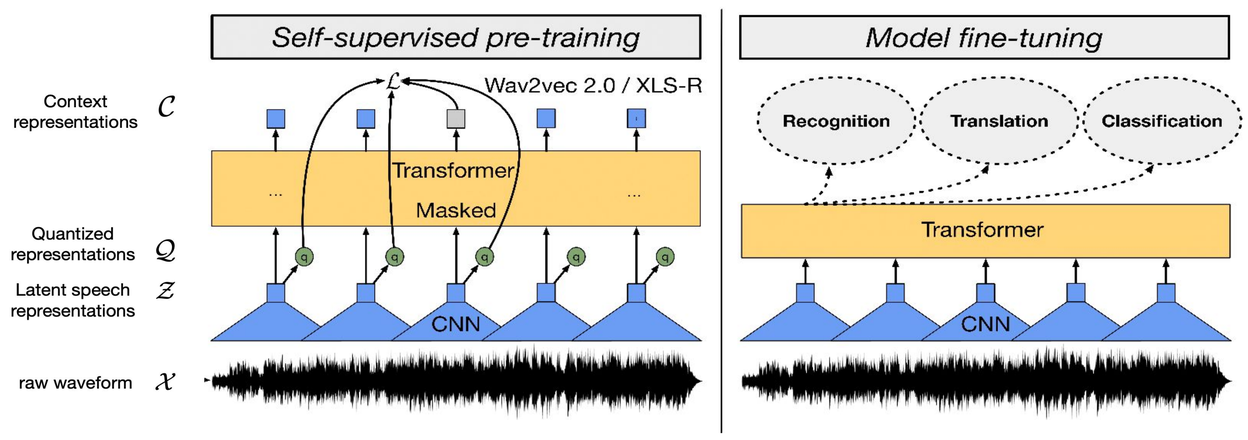
- Automatic Speech Recognition (ASR) : The task of translating spoken language into text

- **Challenge 1:** Limited volume of labeled data
  - ➢ Solution: Self-supervision, which is a training method that can learn from unlabeled data
  - ➢ Speech recognition models wav2vec 2.0 and XLS-R use self-supervision for audio representation learning

- **Challenge 2:** Low resource languages
  - ➢ Solution: Transfer learning

Question: How well will monolingual wav2vec 2.0 perform for cross-lingual transfer learning?

wav2vec 2.0



Model

$g : \mathcal{Z} \mapsto \mathcal{C}$ Contextual representation
$f : \mathcal{X} \mapsto \mathcal{Z}$ Feature encoder
$\mathcal{Z} \mapsto \mathcal{Q}$ Quantization module

Quantization module

G codebooks with V entries $\mathbf{e} \in \mathbb{R}^{V \times d/G}$
$\mathbf{z}$ is mapped to $\mathbf{l} \in \mathbb{R}^{G \times V}$
Linear transformation $\mathbb{R}^d \mapsto \mathbb{R}^f$ to obtain $\mathbf{q} \in \mathbb{R}^f$

$$p_{g,v} = \frac{\exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} \exp(l_{g,k} + n_k)/\tau}$$

Masking

- Sample without replacement proportion p of Latent representation vectors Z.
- For each chosen sample consecutive M time steps are masked.

Objective

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d$$

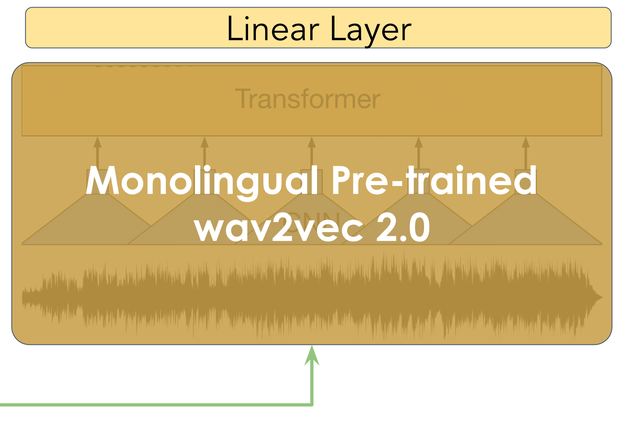
$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)} \qquad \mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

Transfer Learning

Preprocessing steps: Sample rate; Language Vocabulary

- Dutch and French languages
- European languages with different phonetic and lexical similarities to English

WER and CER performance metrics



- Freeze CNN Layers
- CTC loss for ASR fine-tuning

Experiments

Data:
- Training (fine-tuning) data: CommonVoice ~11 hours of recorded speech
- Testing data: CommonVoice ~6 hours

Baseline XLS-R: Wav2Vec2-XLS-R-300M. 300 million parameters, pretrained on 436k hours of unlabeled multilingual speech

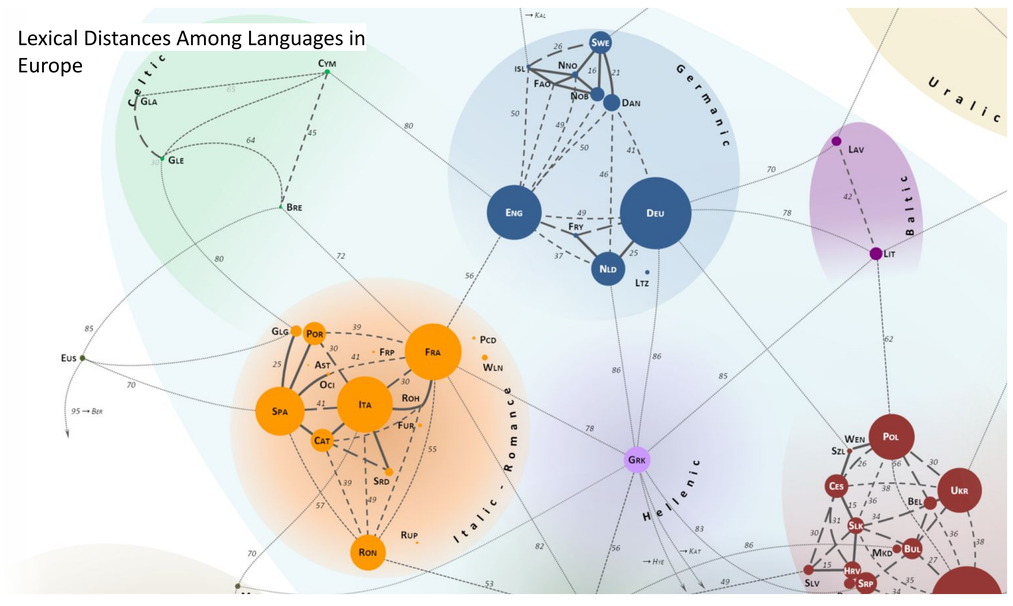
Pre-trained wav2vec 2.0 model: Wav2Vec2-Base. 95 million parameters, 53k hours of unlabeled English speech

	CommonVoice FR		CommonVoice NL		Timit EN		VoxPopuli EN	
	WER	CER	WER	CER	WER	CER	WER	CER
XLS-R	0.333	0.110	0.257	0.081	–	–	–	–
wav2vec 2.0	0.461	0.164	0.387	0.129	0.268	0.087	0.253	0.103

	Ground Truth	XLSR Prediction	wav2vec 2.0 Prediction
French	un vrai travail intéressant va enfin être mené sur ce sujet	un vrai travail intéréssant va enfin être mener sur ce sujet	un vrai travaillintéressant va enfin être mener sur ce sujet
	un comité interministériel du handicap s'est tenu il y a quelques semaines	un cuanmité intelm-nistérial du hendicap s'étenu il y a guelte se-maine	un camite entaminitéréal du randécapes s'es tenu il y a quelque semaine
Dutch	de schoonmaakploeg was net gepasseerd in de vrouwentoiletten	de schoonmakploeg was let gebaseerd in de vrouwentwaliten	de schoonmaak bloeg was lit gebaseerd in de vrouwe tweleten
	door de wind moest ze zich goed vasthouden aan de reling	door de wind moest e zich goed vasthouden aan de rijbing	door de wind moet zo zeer goed vasthouden aan de reping
English	artificial intelligence is for real	–	art official intelligence is for real
	the nearest synagogue may not be within walking distance	–	the nearest synnegu may not be within walk in distance

Analysis

- wav2vec 2.0 performs slightly better on Dutch than on French.
- The results can be explained with the higher similarity of Dutch to English.
- French and Dutch are non-phonetic languages → difficult to infer the right spelling from pronunciation alone.
- Grammar and spelling are <u>not considered</u>.



https://alternativetransport.files.wordpress.com/2015/05/lexical-distance-among-the-languages-of-europe-2-1-mid-size.png

Conclusion & Future Work

- Using a language model to improve the grammatical accuracy.
- Extending the transfer learning framework to more low resource languages

References

[1] Baevski, A, etc "wav2vec 2.0: A framework for self-supervised learning of speech representations." .NeurIPS 2020

[2] Babu, A, etc. XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint 2021