Fraunhofer IAIS

Fraunhofer IAIS

Finetunning ASR models on low resource languages

30th August 2023

Krishna Teja Nallanukala

Supervisor
Mehmet Ali Tugtekin Turan

Introduction

- Automatic Speech Recognition(ASR) technology plays a pivotal role in our increasingly voice-driven digital landscape
- ASR has become an integral part of everyday life
- However, ASR models efficiency depends on the domain, amount of data and acoustic variability that the model was trained.
- ASR fine-tuning enhances the performance in specific contexts and efficient for real-world applications
- Fine-tuning refers to training a pretrained model further using small dataset

Motivation

- Low resource languages lack the amount of transcribed speech to train the models
- Many low-resource languages have diverse accents, dialects, and speaking styles that can vary significantly
- Adapting ASR models to these domains with limited data can be challenging
- Finetunning solves the problem by only requiring small amount of transcribed data [1]
- Therefore, this project work focuses on finetunning various ASR models and performing a comparative analysis on performance of models.

Background

Automatic Speech Recognition

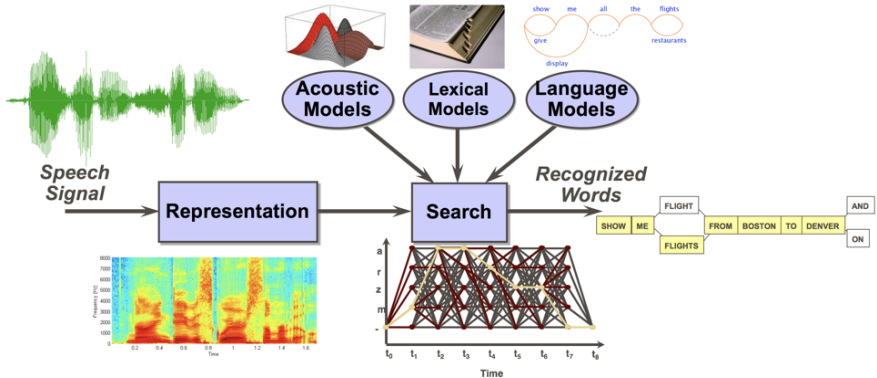

Figure 1: ASR pipeline

Background

- ASR models are classified into two types: self-supervised and supervised
- Self supervised models use the unlabeled data to train the model by auto generating the labels.
- Wave2vec uses the raw unlabeled speech data to train the ASR models to generate best speech representations possible [2]
- Unavailability of labelled dataset is solved by wave2vec
- Wav2vec model is trained in two phase: first phase in self-supervised approach and latter the pretrained model is trained with small amount labelled dataset for particular language.
- whisper model is a supervised model trained on 680,000 hours of labelled dataset combining various languages [1]
- unlike wave2vec, whisper models are ready to use

Datasets

- Common voice 11
- Fluer

Language	Fluers(hr)	avg duration(sec)	Total Words	Avg words/utterence	Vocab
Bengali	15.68	12.98	533986	123.37	124
Tamil	12	13	479055	143.64	92
Hindi	8.7	11.9	335932	120.96	107
Telugu	10.25	11.97	370384	120.05	110

Table 1: Fluer dataset statistics

Language	cv11 (hr)	avg duration(sec)	Total Words	Avg words/utterence	Vocab
Bengali	1268	6.14	2072917	61.9	75
Tamil	393	6.21	4217662	64.6	90
hindi	20	453	409349	43.39	98

Table 2: Fluer dataset statistics

Dataset preprocessing

- Audio data is sampled at 16kHz
- filtering the dataset samples with no ground truth transcriptions
- For wave2vec finetunning ground truth transcriptions are normalized by removing special characters and upper casings
- For Whisper models finetunning transcriptions are not normalized
- During inference all models predictions and GT transcriptions are normalized to compute Word Error Rate(WER)

Wave2vec2.0 Finetunning

Architecture

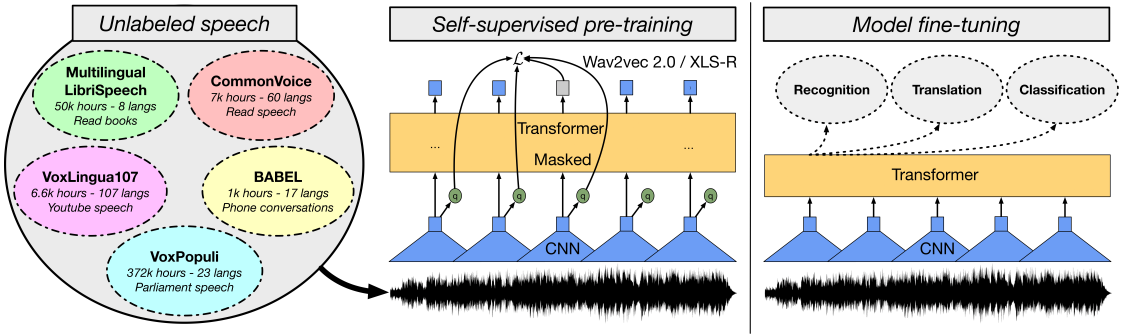

Figure 2: Wave2vec architecture

Wave2vec2.0 Finetunning

Finetunning details

- XLS-R cross lingual models used for finetuning:
  - wav2vec2-large-xlsr-53 (53 languages)
  - wav2vec2-xls-r-300m (128 languages)
- weights of feature encoder are not updated [3]
- classifier representing the output vocabulary of the respective downstream task on top of the model and train with a Connectionist Temporal Classification (CTC) loss [2]
- Sequential multi-lingual finetunning is not possible
- cross-lingual pretraining significantly outperforms monolingual pretraining [2]

Massive Multilingual Speech (MMS)

- MMS is based on Wave2vec2.0 and pretrained on more than 1,400 languages of unlabeled data and more than 500k hours of data
- Project creates models for over 1,100 languages [4]
- Finetunning only modifies adapter layers for particular languages
- "Pretrained MMS-1B checkpoint was further fine-tuned in a supervised fashion on 1000+ languages with a joint vocabulary output layer" [4]
- Pretrained models:
  - MMS-300m
  - MMS-1b
- Finetunned models:
  - mms-1b-fl102
  - mms-1b-l1107
  - mms-1b-all

XLS-R Results

Dataset	Pretrained models	Pretraining hours	Training time (hr)	GPUs used	Epochs	WER % Without LM	WER % With LM	Train batch size
Open SLR66	wav2vec2-large-xlsr-53 (53 languages)	56k	8	4	150	6.65%	5.56%	8
	wav2vec2-xls-r-300m (128 languages)	436k	7.6	4	100	4%	3.70%	8
	Hubert-large-ll60k	60k	5.4	4	100	17.39	11.05	8
IIIT_hyd	wav2vec2-large-xlsr-53 (53 languages)	56k	72	4	150	45%	36.70%	2
	wav2vec2-xls-r-300m (128 languages)	436k	125	4	100	38.06	35.51	2
	Hubert-large-ll60k	60k	137	4	100	55.72	42.9	2

Table 3: XLSR models finetunning on OpenSLR66 and IIThyd telugu dataset

Dataset	Model	Actual dataset (hr)	Used train dataset size (hr)	training time	steps	batchsize	WER %	Inference time
cv_bengali	xls-r-300-bengali	143	5hr	4hr	5000	8	53.09	1.1 min
	xlsr-300-bengali-tamil	218	10hr	4hr 43min	5000	8	74.29	1.32 min
	xlsr-300-hindi-bengali-tamil	223.5	15hr	5hr 6min	5000	8	90.91	1.08 min
cv_tamil	xls-r-300-tamil	75	5hr	3hr 45min	5000	8	77.205	1.08 min
	xlsr-300-bengali-tamil	128	10hr	4hr 43min	5000	8	99.7	1.32 min
	xlsr-300-hindi-bengali-tamil	223.5	15hr	5hr 6min	5000	8	109.93	1.11 min
cv_hindi	xls-r-300-hindi	5.45	5hr	3h	5000	8	54.82	1.08 min
	xlsr-300-hindi-bengali-tamil	223.5	15hr	5hr 6min	5000	8	100	1.11 min

Table 4: Wave2vec2-xlsr-300m model finetunning results

MMS results

Dataset	Model	Used dataset size (hr)	training time	steps	batchsize	WER %	Inference time	
cv_bengali	mms-300m	5hr	3 hr50 min	5000	8	62.97	1.116	
	mms-1b	5hr	3hr 20min	2000	8	49.4	2.94	
	mms-1b-all	-	-	-	-	32.98	2.94	
cv_tamil	mms-300m	5hr	3 hr50 min	5000	8	72.39	1.116	
	mms-1b	5 hr	3hr 20 min	2000	8	60.912	3.06	
	mms-1b-all	-	-	-	-	45.11	2.94	
cv_hindi	mms-300m	5hr	2 hr49 min	5000	8	52.106	1.2	
	mms-1b	5hr	3hr 10min	2000	8	103.31	1.2	
	mms-1b-all	-	-	-	-	44.39	3	

Table 5: MMS finetunning results

Whisper Finetunning

Architecture

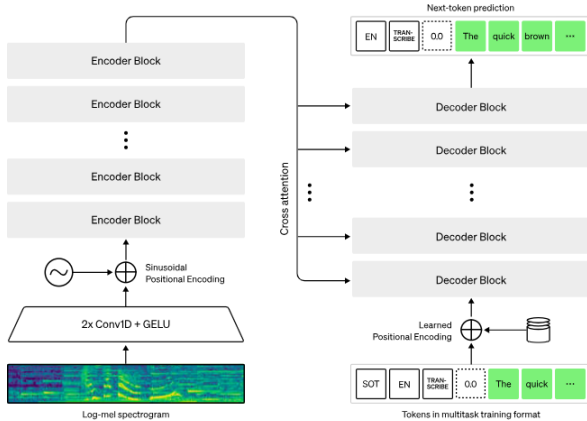

Figure 3: Whiper sequence-to-sequence model [1].

Whisper Finetunning

Finetunning details

- The whisper has models of various sizes as shown in Figure 4
- Trained on 680,000 hours of data from internet
- models can be used for multiple tasks: Language identification, timestamps prediction, transcription and translations [1]
- whisper can do multilingual finetunning as well as sequential finetunning

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Figure 4: Whisper models [1]

whisper results

Dataset	models	Training time	Train batch size	GPUs used	Optimization steps	WER % (before finetunning	WER % (finetunned)
Openslr	Tiny	7.69	8	2	4000	307.9	44.24
	Base	11.22	8	4	4000	207.2	19.16
	Small	26.2	8	2	4000	121.38	12.93
IIIT_hyd	Tiny	127	1	4	55800	372.01	51.28
	Base	150	1	3	55800	276.1	45.72
	Small	200	1	4	55800	189.69	33.85

Table 6: whisper models finetunning on OpenSLR and IIT-hyd telugu dataset

Dataset	Model	Dataset	Training time	steps	batchsize	WER %	Inference time	Before finetunning WER
cv2_hi	whisper-small (ta+hi)	10	13hr 8min	5000	8	34.68	0.134	-
	whisper-small (bn+hi)	10	15 hr 31m	5000	8	34.84	0.134	-
	whisper-small (ta+hi+bn)	15	1d5hr	5000	8	34.92	0.139	-
	whisper-small (hi)	5hr	10 hr 26min	5000	8	59.06	0.129	112.18
	whisper-small-hindi-bengali-seq-tamil	5hr	10hr	5000	8	80.18	0.149	-
	whisper-small-hindi-bengali-seq-tami-concat(30min hndi-30min bengali)	7hr	10hr	5000	8	37.84	0.131	-
cv2_bn	whisper-small (ta+hi+bn)	15	1d5hr	5000	8	49.65	0.237	-
	whisper-small (bn+ta)	10	1d 4h	5000	8	48.45	0.242	-
	whisper-small (bn+hi)	10	15 hr 31m	5000	8	51.49	0.248	-
	whisper-small (bn)	5hr	11hr 12min	5000	8	55.98	0.24	160.69
	whisper-small-hindi-bengali-seq-tamil	5hr	10hr	5000	8	78.43	0.229	-
	whisper-small-hindi-bengali-seq-tami-concat(30min hndi-30min bengali)	7hr	10hr	5000	8	55.36	0.23	-
cv2_ta	whisper-small (ta+hi+bn)	15	1d5hr	5000	8	54.22	0.13	-

Table 7: whisper small models finetunning

comparison and discussion

Language	Models	Training time	WER %	Inference time
cv_bengali	mms-300m	3 hr50 min	62.97	1.116
	mms-1b	3hr 20min	49.4	2.94
	whisper-small (ta+hi+bn)	13hr 8min	49.65	14.22
	whisper-small (bn+ta)	15 hr 31m	48.45	14.52
	whisper-small (bn+hi)	1d5hr	51.49	14.88
	whisper-small (bn)	10 hr 26min	55.98	14.4
	whisper-small-hindi-bengali-seq-tamil	10hr	78.43	13.74
	whisper-small-hindi-bengali-seq-tami-concat(30min hndi-30min bengali)	10hr	55.36	13.8
	xls-r-300-bengali	4hr	53.09	1.1 min
	xlsr-300-bengali-tamil	4hr 43min	74.29	1.32 min
	xlsr-300-hindi-bengali-tamil	5hr 6min	90.91	1.08 min

Table 8: Comparision of commonvoice bengali finetunned models

comparison and discussion

Language	Models	Training time	WER %	Inference time
cv_tamil	mms-300m	3 hr50 min	72.39	1.116
	mms-1b	3hr 20 min*	60.912	3.06
	whisper-small (ta+hi+bn)	1d5hr	54.22	7.8
	whisper-small (ta+hi)	13hr 8min	53.14	7.62
	whisper-small (bn+ta)	1d 4h	54.46	7.56
	whisper-small (ta)	9hr 17m	60.86	7.56
	whisper-small-hindi-bengali-seq-tamil	10hr	51.75	7.68
	whisper-small-hindi-bengali-seq-tami-concat(30min hndi-30min bengali)	10hr	52.7	7.44
	xls-r-300-tamil	3hr 45min	77.205	1.08 min
	xlsr-300-bengali-tamil	4hr 43min	99.7	1.32 min
	xlsr-300-hindi-bengali-tamil	5hr 6min	109.93	1.11 min

Table 9: Comparision of commonvoice tamil finetunned models

comparison and discussion

language	Models	Training time	WER %	Inference time
cv_hindi	mms-300m	2 hr49 min	52.106	1.2
	mms-1b	3hr 10min	103.31	1.2
	whisper-small (ta+hi)	13hr 8min	34.68	8.04
	whisper-small (bn+hi)	15 hr 31m	34.84	8.04
	whisper-small (ta+hi+bn)	1d5hr	34.92	8.34
	whisper-small (hi)	10 hr 26min	59.06	7.74
	whisper-small-hindi-bengali-seq-tamil	10hr	80.18	8.94
	whisper-small-hindi-bengali-seq-tami-concat(30min hndi-30min bengali)	10hr	37.84	7.86
	xls-r-300-hindi	3h	54.82	1.08 min
	xlsr-300-hindi-bengali-tamil	5hr 6min	100	1.11 min

Table 10: Comparision of commonvoice hindi finetunned models

Conclusion

- whisper multilingual lingual finetunned models gave better performance than monolingual
- whisper sequential finetunning with some data of languages achieve better WER.
- whisper models take more time for inference and finetunning compared to wave2vec models
- Overall whisper models showed better performance in transcribing data
- XLS-R models performance drops with multilingual finetunning
- Additional language models certainly improves the performance of models

Challenges faced

- Finetunning of large and medium models on nm-cluster.
- 8-bit quantization is not supported on nm-cluster GPUs to perform large and medium models finetunning even using Parametric Effecient Finetunning(PEFT) [5].
- Ctranslate2 [6] and Faster whisper are used for large models inference but the WER is not optimal.
- PEFT adalora [5] model can finetune large models however produces "!" marks in prediction.
- Main reason for such behaviour is also due to 8bit quantization not supported in cluster GPUs

Future Work

- Finetune whisper large and medium models using PEFT on high CUDA compte capability GPUs
- Finetune multi-lingual xlsr models on high resource language to observe the perforance of multilingual models
- Compare Nemo models with whisper and wave2vec
- Resolving the MMS adapter issue

References

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.

A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” **CoRR**, vol. abs/2006.13979, 2020. [Online]. Available: https://arxiv.org/abs/2006.13979

“MMS Finetunning Blog,” https://huggingface.co/blog/mms_adapters, accessed: 2023-08-29.

“MMS,” https://ai.meta.com/blog/multilingual-model-speech-recognition/, accessed: 2023-08-29.

“PEFT,” https://huggingface.co/blog/peft, accessed: 2023-08-29.

“Faster-whisper,” https://github.com/guillaumekln/faster-whisper, accessed: 2023-08-29.

Thank you for the attention!
Questions?