Improved Meta Learning for Low Resource Speech Recognition

Satwinder Singh, Ruili Wang, Feng Hou

s.singh4@massey.ac.nz, ruili.wang@massey.ac.nz, f.hou@massey.ac.nz

School of Mathematical and Computational Sciences

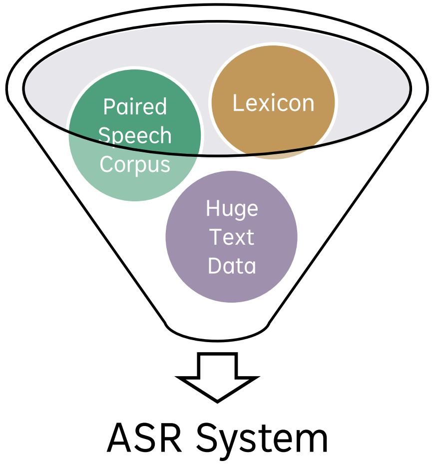
Massey University, Auckland, New Zealand

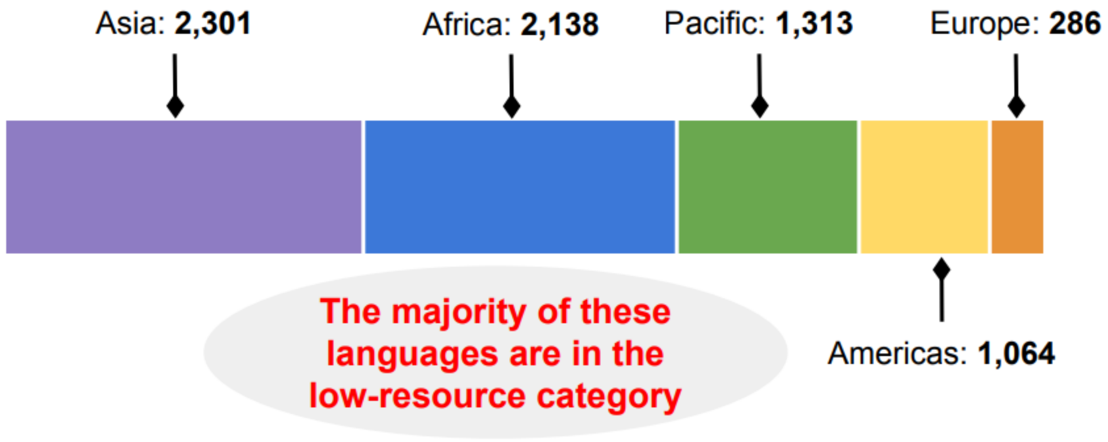
Outline

- Introduction to low resource languages
- Challenges
- Proposed approach
- Experimental setup and results
- Conclusions

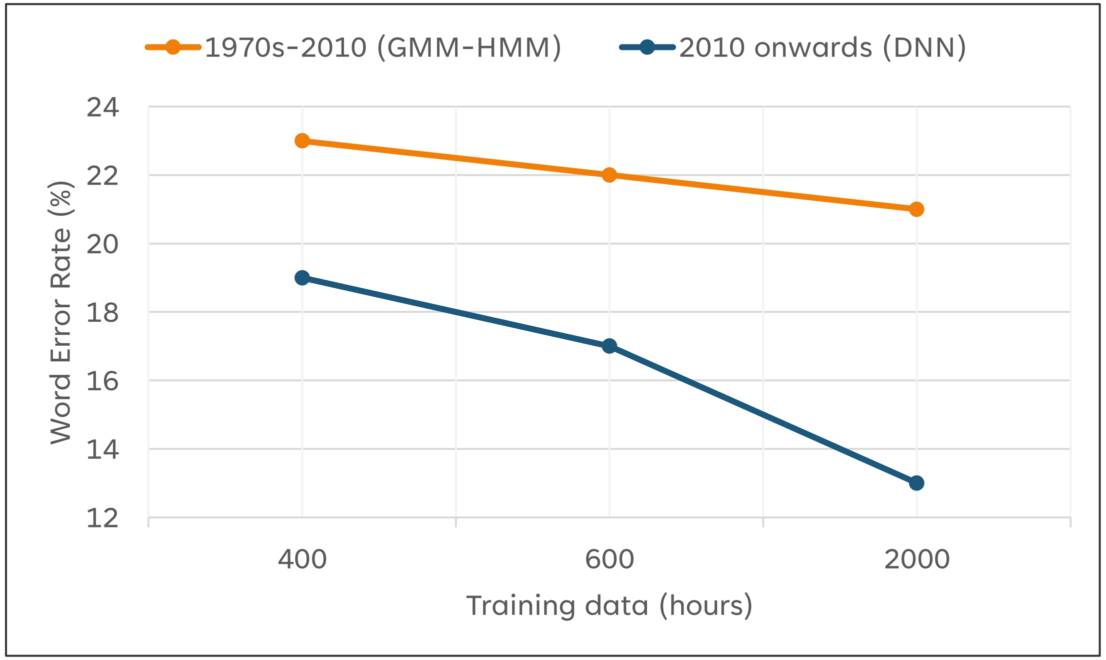
What are Low Resource Languages (LRL)?

- The languages that do not have enough linguistic resources are considered as low resource languages.
- There are approximately over 7000 languages being spoken around the world. (Precoda et al., 2004)
- Only around 100 languages have well established speech recognition systems.





Word Error Rate vs Available data



(Huang et al., 2014)

Challenges

- LRL may have few native speakers as only 400 languages have over one million speakers.
- It is tough to record diverse speech data, which is most expensive and time-consuming.
- Transcription process may also take a considerable number of efforts to produce accurate annotated data.
- Linguistic experts must be included in the process to create pronunciation dictionaries.

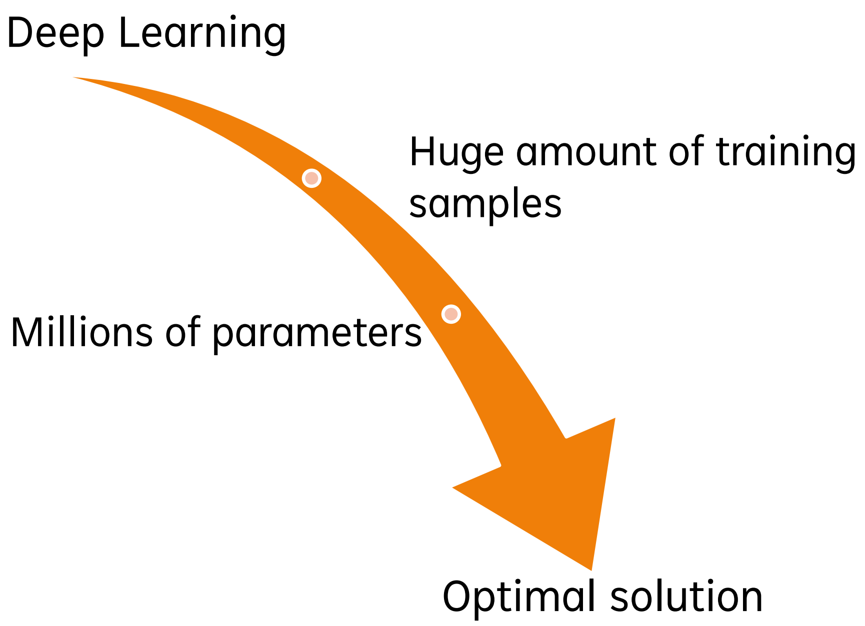
Proposed Solutions

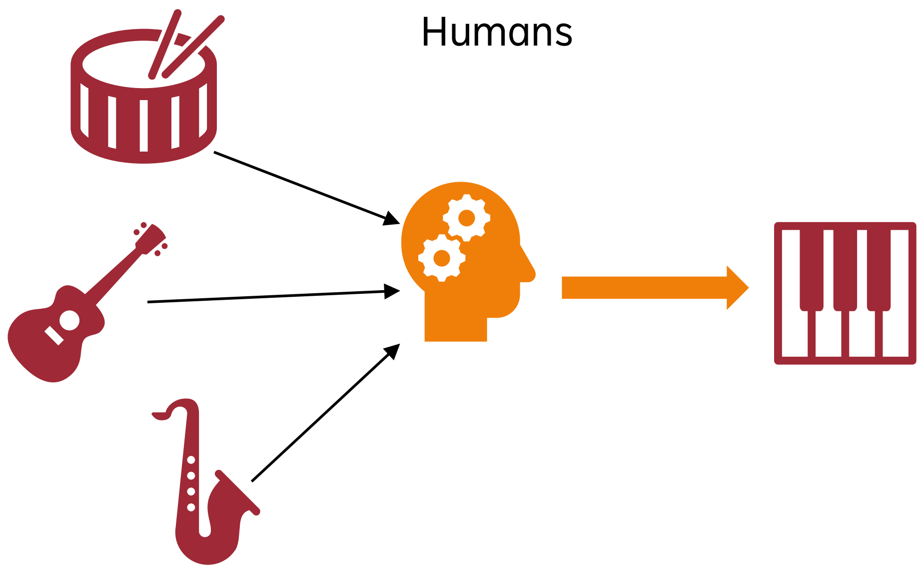
- Data augmentation
- Multilingual systems
- Cross lingual transfer learning
- Semi-supervised learning
- Meta learning

Meta learning

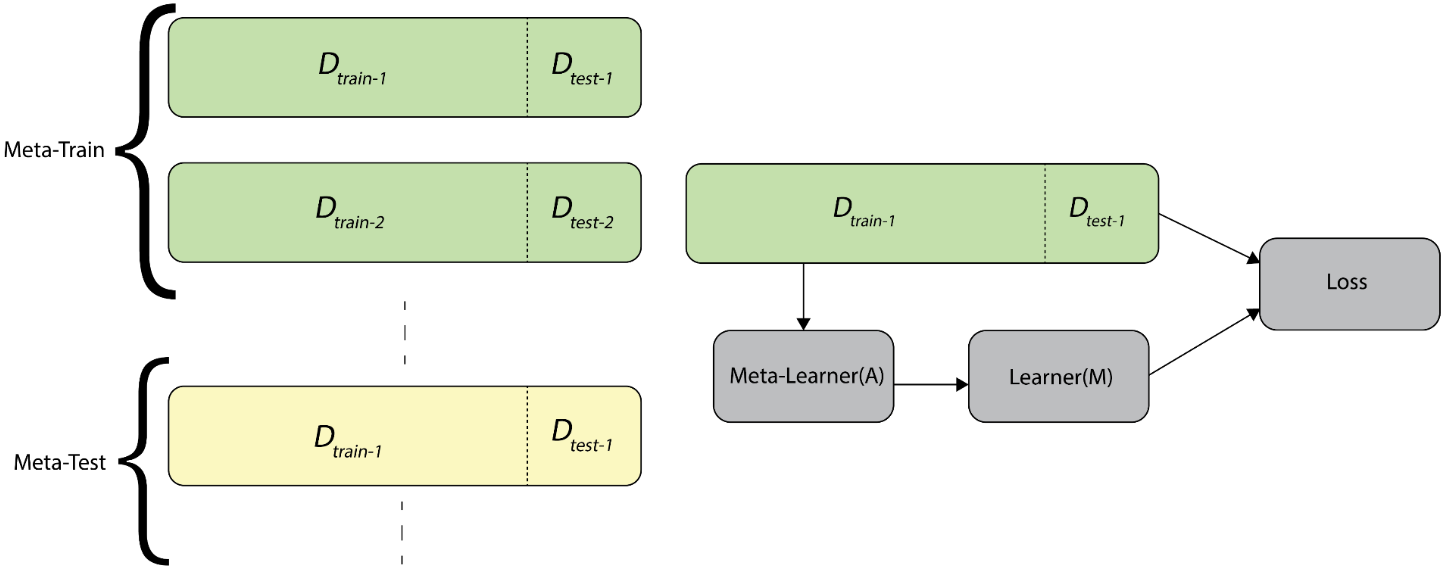
(Motivation)

- Meta-learning, also known as learning to learn, focuses on improving the learning efficiency based on previous experiences on wide variety of tasks.





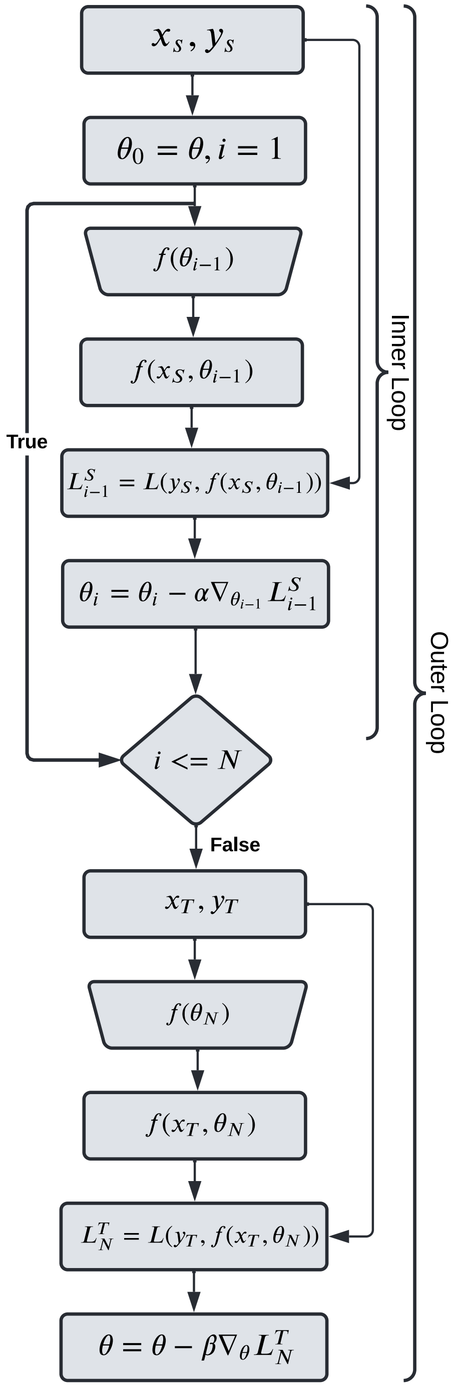
Meta learning
(Formulation)



- **Input:** Meta-training set $\mathscr{D}_{meta-train} = \{(D^{(n)}_{train}), (D^{(n)}_{test})\}^{N}_{n=1}$
- **Output:** Parameters θ algorithm A (Meta-learner)
- **Objective:** Good performance on $\mathscr{D}_{meta-test} = \{(D'^{(n)}_{train}), (D'^{(n)}_{test})\}^{N'}_{n=1}$

Model Agnostic Meta learning (MAML)

(Finn et al., 2017)



MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

What has been Done?

- MAML for low resource ASR (Hsu et al., 2020)
  - Outperformed no-pretraining and multilingual training settings
- MAML for accent adaptation (Winata et al., 2020)
  - Outperformed joint training setting across various English accents in few shot scenarios

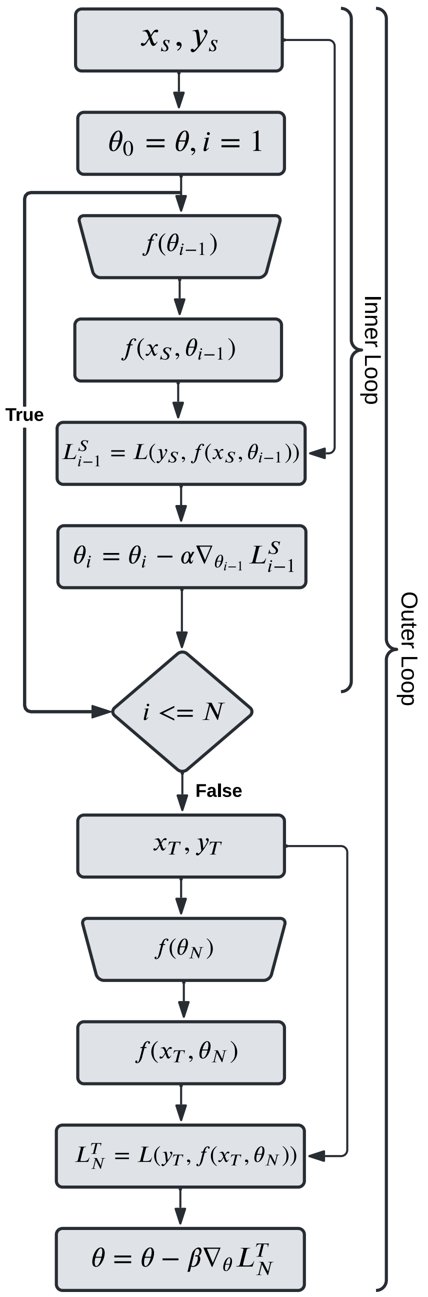
Issues with MAML

- Inconsistent training behaviour
- Slower convergence speed

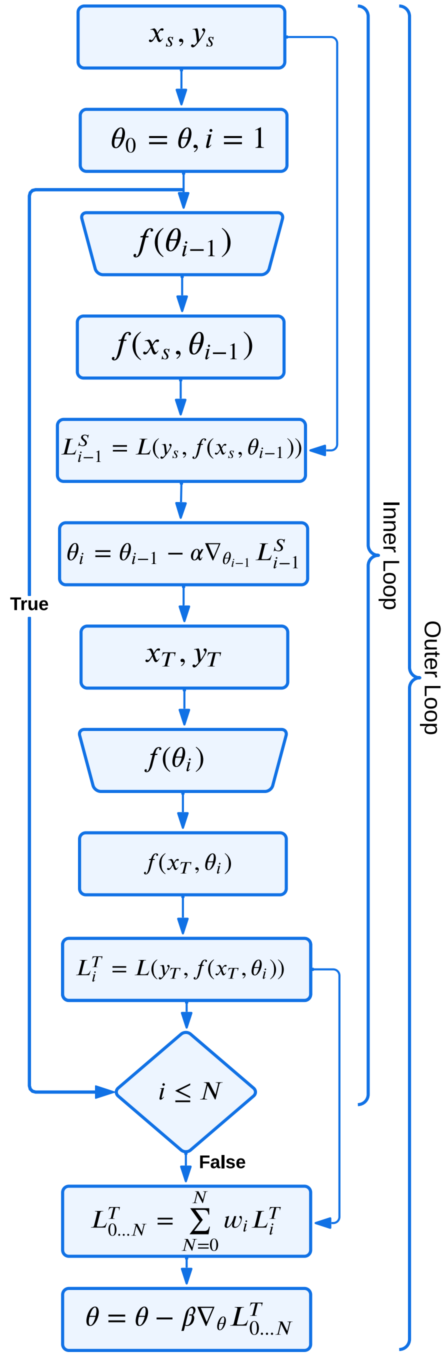
Proposed Solution

- Multi-step loss (MSL) (Antoniou et al., 2018)
  - Originally, proposed for the image classification task.
  - It calculates the inner loss after every inner step updates.
  - Then computes the weighted sum of all the inner losses.

MAML vs MAML with MSL



MAML

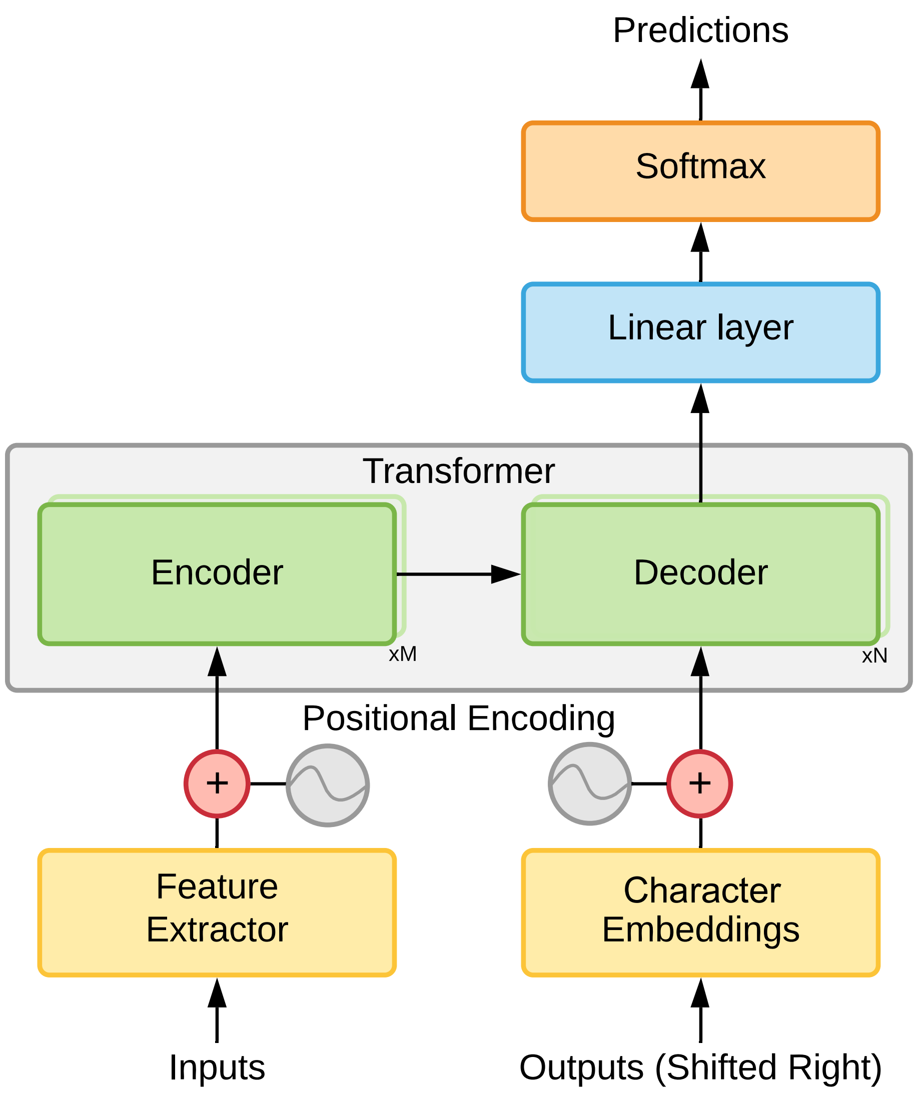


MAML with MSL

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

The ASR model

- Transformer based model
  - 6 layered VGG extractor
  - 2 encoder layers
  - 4 decoder layers
  - 8 heads for multi-head attention



MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Experimental Setup

Datasets

- Common Voice V7.0
- Source language sets
  - [fa, ar, ta], [ar, mn, lt], [or, pa-IN, hi, ur, as]
- Target language set
  - hi, mn, fa, ar, ta

Methodology

- Pretrain
  - 100K iterations on 3 source sets
- Fine-tune
  - Fine-tune for 10 epochs with beam size of 5.

Table 1: The selected low resource languages from the Common Voice dataset v7.0 and the total amount of speech data in terms of hours.

ID	Languages	Hours
ar	Arabic	85
as	Assames	1
hi	Hindi	8
lt	Lithuanian	16
mn	Mongolian	12
or	Odia	0.94
fa	Persian	293
pa-IN	Punjabi	1
ta	Tamil	198
ur	Urdu	0.59
Total		615.53

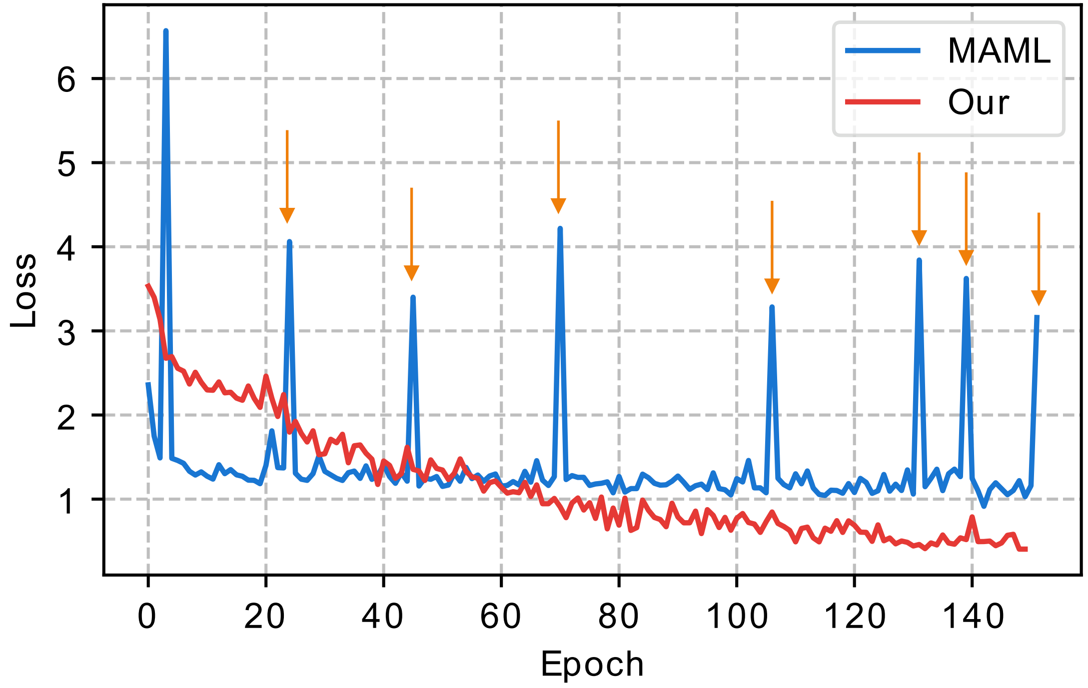
Experimental Results

Table 2: The average experimental results in terms of character error rate (CER in %) on 5 target languages. We have not fine-tune our model on the languages that are present in the pretrain source language sets. These cells are represented by hyphen (-).

Pretrain languages	Finetune									
	Hindi		Mongolian		Persian		Arabic		Tamil	
	MAML	Our	MAML	Our	MAML	Our	MAML	Our	MAML	Our
[fa, ar, ta]	70.51	**70.47**	61.05	**60.52**	-	-	-	-	-	-
[ar, mn, lt]	71.61	**71.37**	-	-	47.96	**45.45**	-	-	40.96	**35.17**
[or, pa-IN, hi, ur, as]	-	-	62.26	**59.50**	52.42	**52.41**	**36.00**	36.09	**45.96**	46.60

Training performance (MAML vs MAML with MSL)

- MAML approach with MSL improves the training consistency.



Conclusions

- Multi step loss indeed improves the training stability.
- It also has positive impact on the overall accuracy of the model.
- In the future, we plan to conduct more experiments with more low resource languages.

Thank you!