IMPROVED META LEARNING FOR LOW RESOURCE SPEECH RECOGNITION

Satwinder Singh *Ruili Wang* *Feng Hou**

School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

ABSTRACT

We propose a new meta learning based framework for low resource speech recognition that improves the previous model agnostic meta learning (MAML) approach. The MAML is a simple yet powerful meta learning approach. However, the MAML presents some core deficiencies such as training instabilities and slower convergence speed. To address these issues, we adopt multi-step loss (MSL). The MSL aims to calculate losses at every step of the inner loop of MAML and then combines them with a weighted importance vector. The importance vector ensures that the loss at the last step has more importance than the previous steps. Our empirical evaluation shows that MSL significantly improves the stability of the training procedure and it thus also improves the accuracy of the overall system. Our proposed system outperforms MAML based low resource ASR system on various languages in terms of character error rates and stable training behavior.

Index Terms— low resource languages, meta learning, MAML, automatic speech recognition

1. INTRODUCTION

Modern deep learning based end-to-end (E2E) models have lately become extremely popular in the speech community [1] and have achieved a significant milestone in terms of performance. These systems have been deployed under commercial domains as they have shown consistently lower word error rates that are close to 1-2% [2]. The modern ASR systems are mostly trained in end-to-end (E2E) fashion without requiring resources like a pronunciation dictionary and a language model as separate modules. These systems are able to achieve such a high degree of accuracy mainly because they are trained on various high performance large vocabulary datasets. However, these E2E systems tend to perform much worse for the languages that do not have such large quantities of annotated data.

Among roughly 7000 languages spoken across the world, there are only around 100 languages that have well-established speech recognition systems [3]. The rest of the languages are considered as low resource languages because they do not have a huge amount of annotated speech data, strong pronunciation dictionaries, and a huge collection of unpaired texts. A lot of progress has been made in low resource speech recognition, which includes efforts like transfer learning [4] and multilingual training [5]. Recently, a new paradigm, meta learning has been explored for low resource speech recognition [6]. Meta learning (also known as learning to learn) is a machine learning technique, where learning is done on two levels. On one level (inner loop) model acquires task specific knowledge, whereas the second level (outer loop) facilitates task across learning [7].

Previously, Hsu et al. [6] proposed a meta learning framework based on the MAML approach for ASR for low resource language. The proposed framework outperformed the no-pretraining and multi-lingual training settings. Similarly, Winata et al. [8] incorporated the MAML approach for the few shot accent adaptation task for the English. The MAML approach in general is a very straightforward and powerful approach. However, it is prone to numerous problems, including unstable training and slow convergence speeds. These issues also impact the generalizability of the model. Thus, to deal with these issues, in this paper we adopt the multi-step loss [7], which is introduced to stabilize the meta training procedure. The meta training approach with multi-step loss calculates the inner loss after every inner step updates and later computes the weighted sum of all the inner losses.

We evaluated our proposed approach on 10 different languages present in the Common Voice v7.0 dataset. All these languages are represented in form of a low resource setting where the language data ranges from 0.5 hours to 300 hours. We find that our approach indeed improves the training instabilities of the MAML approach, which in turn improves the overall accuracy of the model.

2. RELATED WORK

2.1. Meta Learning

Meta learning is not a new idea but begin to gain attention in recent times. Recently, in the context of deep learning, meta learning comes into the limelight due to its wide range of applications and advantages. Meta learning helps to generalize to various tasks faster with few steps and examples. Literature suggests the application of meta learning in two ways where the first is learning a better initialization of network parameters [9] and the second is learning a strategy or procedure for

*Corresponding author

updating the parameters of the network [10], [11].

Meta learning has been applied to a range of research domains including various computer vision tasks, natural language processing and recently automatic speech recognition. In the computer vision area, meta learning has been exploited for the few-shot image classification task [12], object detection [13] and video generation [14]. In the natural language processing domain, meta learning has shown promising results in neural machine translation (NMT) for resource constraint languages [15]. Apart from this, recently researchers have tried meta learning for speech processing tasks, such as automatic speech recognition [6], speaker adaptation [16], [17] and recognition [18], cross-lingual [19] and cross-accent adaption [8].

2.2. Low Resource Speech Recognition

The development of a speech recognition system for a low resource languages has been a very active research area for the past few years. The regular E2E ASR systems designed for resource rich languages seem not to work for low resource languages due to the lack of annotated speech data or other resources. There have been many attempts made to alleviate the scarcity of labelled speech data. These efforts include, speech data augmentation [20], transfer learning [4], multilingual [5], cross-lingual [19] and multi-task learning [6]. Recently, unsupervised cross-lingual wave2vec 2.0 XLSR model [21] shown a huge performance boost compared to other previous state-of-the-art models. Further, there have been recent attempts to explore a new research direction of meta learning for low resource languages. The idea is to extract meta parameters learned over multiple source languages and then bootstrap these learned meta parameters to fine-tune on the target languages. The whole process can be seen as learning a model that can perform fast adaptation to target languages with few epochs and data samples. As fine-tuning requires few training samples, this process of meta learning is totally aligned with our proposed framework of ASR for low resource languages.

3. PROPOSED SYSTEM

Our proposed system consists of two core components. The first is an ASR model that acquires language specific knowledge and the second is a multi-step loss based model agnostic meta learning algorithm.

3.1. The ASR Model

For our proposed system, we adopt the transformer ASR model [22] as our language specific model. The transformer model is a sequence-to-sequence model based on the encoder-decoder architecture. The proposed model extracts the input features using the learnable VGG based convolutional neural network (CNN) model [23]. The input embeddings produced

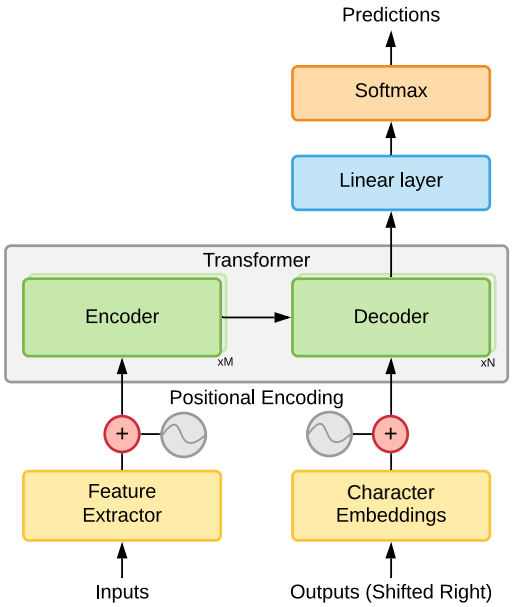

Fig. 1: The Transformer model for ASR

by the feature extractor are then fed to the encoder module through the positional encoding setup. The positional encoding setup generates a vector that is served as context for the symbols. Afterward, the outputs of the encoder module are passed on to the decoder module, where a multi-head attention mechanism is employed on these encoder outputs. The attention mechanism applies masking in the decoder block to restrict the attention layer from attending to any future tokens. Finally, the output of the decoder block goes through a linear and softmax layer and generates the predictions. The entire training process is optimized by maximizing the log probability using next step prediction based on the last output token. In the following equation, x, y_i and y'_{i-1} are the input character, next predicted character, and true label of the last character, respectively.

$$\max_{\theta} \sum_{i} \log P(y_i|x, y'_{i-1}; \theta) \tag{1}$$

3.2. Meta Learning Setup

In general, our proposed meta learning setup aims at learning the initial parameters for the model in a way that it can be quickly adapted to new languages with a fewer number of gradient descent steps. We adopt multi-step loss from MAML++ [7] procedure over standard MAML as MAML tends to have unstable training procedure. This can affect the overall speed of convergence, and also has a negative impact on the accuracy of the model. Figure 2 shows the computational graph for both MAML and MAML with multi-step loss. We select support samples (x_S, y_S) and validation data samples (x_T, y_T) from our source languages set. We start optimizing

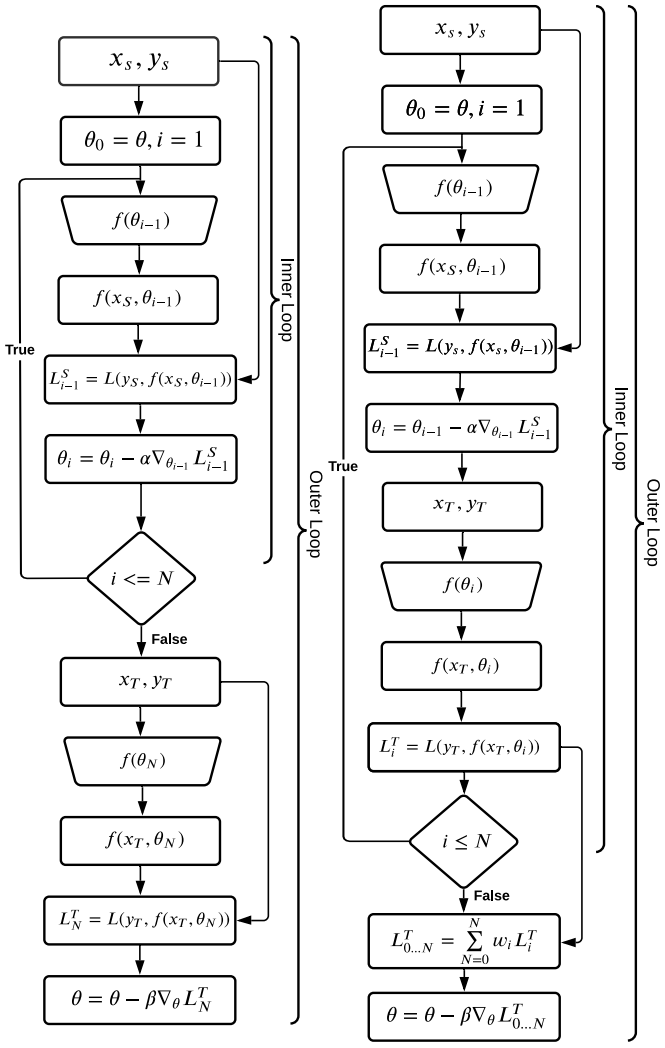

Fig. 2: MAML (Left) vs MAML with MSL (Right) (adopted from [7])

the inner loop by initializing our ASR model f with $\theta_0 = \theta$. Afterwards, the ASR model produces logits $f(x_S, \theta_{i-1})$ by using samples from training set and parameters θ_{i-1}. Here i represents i^{th} step of total N steps. In the next step, loss L^S_{i-1} is computed over true labels y_S and logits. Further, the L^S_{i-1} is utilized to update the current parameters of the model.

The inner optimization loop of our MSL MAML approach differs from MAML, where instead of using θ_N parameters for computing target set loss, our MSL MAML approach goes on using θ_i parameters. After completing the inner loop, we obtain N target set losses as in Eq. 2, which can be seen as a multi-step loss, where w_i is the weight of step i and specify the importance of per step target loss. Initially, all the losses have approximately the same importance, while later in training more importance is given to the losses on later steps. This way the model gradually steps towards the MAML loss, ensuring there is no issue of gradient degradation. Finally, these losses are combined together using a weighted sum of per step losses. The combined weighted loss is then used to update the outer loop parameters θ. The advantage of calculating per step loss is reducing the gradient vanishing and exploding problem of the original MAML. Following [8] and [6], we only compute first order approximation of θ.

$$L^T_{0...N} = \sum_{N=0}^{N} w_i L^T_i \tag{2}$$

4. EXPERIMENTAL SETUP

4.1. Dataset

For our experiments, we choose Common Voice dataset version 7 [24]. The data in Common Voice is a crowdsourced public dataset and contains many languages including resource rich and low resource languages. We select 10 low resource languages and the description is represented in Table 1. Some of the languages are very low resource having just a few hours of data. The audio from all the languages is downsampled to 16 kHz and labels are preprocessed to remove any kind of special symbols.

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

4.2. Methodology

Our model receives spectrogram as an input. These spectrogram inputs then go through a VGG based 6-layered CNN feature extractor. We use 2 encoder layers and 4 decoder layers with 8 multi-head attention layers. Our model produces input and output of dimension 512, whereas the inner layer has 2048 dimensions. We set the dropout value to 0.1 and keys and values dimensions to 64. We multilingually pretrain our model for 100K iterations on the source language set. We put together 3 source language sets where one set includes **fa,**

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

ar and, **ta**. The other set has **ar, mn** and, **lt** and the last set consists of **or, pa-IN, hi, ur,** and **as**. During the fine-tuning phase, we fine-tune the model on our target languages (**hi, mn, fa, ar** and **ta**) one by one for 10 epochs. The model is then evaluated on a test set of target language using beam search with a beam size of 5.

5. RESULTS AND DISCUSSION

5.1. Model's Accuracy Analysis

We evaluate the performance of our proposed MSL MAML approach on 10 languages from the Common Voice dataset. Our proposed approach showcases consistent improvement in character error rates (CER in %) over the standard MAML approach. The detailed results are presented in Table 2. On source languages set **[fa, ar, ta]** our approach achieves 70.47% and 60.52% of CER on Hindi and Mongolian language, respectively. Our proposed model shows around 1% of improvement over standard MAML on the Mongolian language. On set **[ar, mn, lt]** our approach slightly performs better than MAML on the Hindi language. On the same set, our approach outperforms the current MAML approach with 5.23% and 14.13% of relative improvement on Persian and Tamil language, respectively.

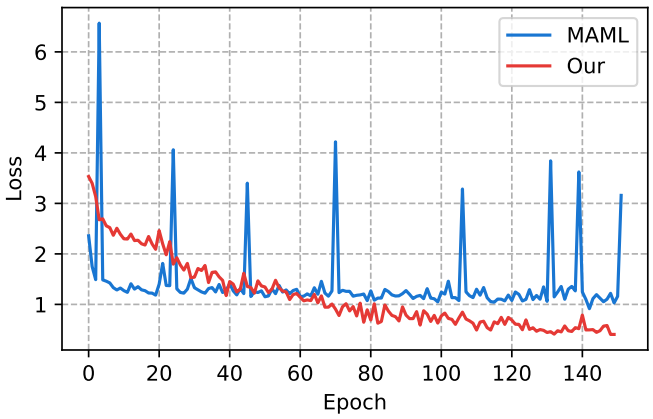

Fig. 3: Training curve of MAML vs our approach. The training loss curve for MAML shows unstable peaks whereas our approach shows more consistent loss curve.

Further, the Mongolian language demonstrates 4.43% of relative improvement over MAML on set **[or, pa-IN, hi, ur, as]**. Mostly, on this pretrain language set both MAML and our approach report similar results on Persian and Arabic languages. Interestingly, the MAML marginally outperforms our approach on the Tamil language. Overall, our approach shows consistent improvements across all the pretrain sets, where excellent performance is observed on **[ar, mn, lt]**.

5.2. Training Performance Analysis

The multi-step loss indeed stabilizes the training process of MAML as shown in Figure 3. The primary driver of instability in MAML is the gradient degradation problem while training deep network [7]. Our approach resolved this issue using multi-step loss, where the model is evaluated at each step against its validation set. Further, importance weight also makes sure later step loss has more importance. It also improves the convergence speed of the model as shown in Figure 3.

6. CONCLUSIONS

In this paper, we propose a multi-step loss based meta learning approach for speech recognition for low resource languages. The proposed system improves the inner loop optimization for the MAML algorithm, which results in a more stabilized training procedure. Our empirical results show that multi-step loss indeed improves the overall training procedure and also has a positive impact on the accuracy of the model. Apart from this, our model also trains faster as compared to MAML. In the future, we plan to conduct more experiments with more low resource languages. We would extend our experiments with different combinations of languages on the basis of their phonetic structures, geographic areas, and language family.

7. ACKNOWLEDGEMENT

This work is supported by the 2020 Catalyst: Strategic New Zealand - Singapore Data Science Research Programme Fund by Ministry of Business, Innovation and Employment (MBIE), New Zealand.

8. REFERENCES

[1] Satwinder Singh, Ruili Wang, and Yuanhang Qiu, "DeepF0: End-to-end fundamental frequency estimation for music and speech signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 61–65.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] Ekapol Chuangsuwanich, *Multilingual techniques for low resource automatic speech recognition*, Ph.D. thesis, MIT, Cambridge, United States, 2016.

[4] Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee, "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning," in *Proc. Interspeech*, 2019, pp. 2075–2079.

[5] Shiyu Zhou, Shuang Xu, and Bo Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[6] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, "Meta learning for end-to-end low-resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844–7848.

[7] Antreas Antoniou, Harrison Edwards, and Amos Storkey, "How to train your maml," in *International Conference on Learning Representations (ICLR)*, 2018.

[8] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung, "Learning fast adaptation on cross-accented speech recognition," in *Proc. Interspeech*, 2020, pp. 1276–1280.

[9] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel, "Meta-learning with temporal convolutions," *arXiv preprint arXiv:1707.03141*, vol. 2, no. 7, 2017.

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 1126–1135.

[11] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087.

[12] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[13] Aniwat Phaphuangwittayakul, Yi Guo, and Fangli Ying, "Fast adaptive meta-learning for few-shot image generation," *IEEE Transactions on Multimedia*, 2021.

[14] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro, "Few-shot video-to-video synthesis," in *International Conference on Neural Information Processing Systems*, 2019, pp. 5013–5024.

[15] Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho, "Meta-learning for low-resource neural machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622–3631.

[16] Ondřej Klejch, Joachim Fainberg, and Peter Bell, "Learning to adapt: A meta-learning approach for speaker adaptation," in *Proc. Interspeech 2018*, 2018, pp. 867–871.

[17] Ondřej Klejch, Joachim Fainberg, Peter Bell, and Steve Renals, "Speaker adaptive training using model agnostic meta-learning," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 881–888.

[18] Seong Min Kye, Youngmoon Jung, Hae Beom Lee, Sung Ju Hwang, and Hoirin Kim, "Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs," in *Proc. Interspeech*, 2020, pp. 2982–2986.

[19] Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki, "Meta-adapter: Efficient cross-lingual adaptation with meta-learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7028–7032.

[20] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[21] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[24] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.