Meta-Adapter: Efficient Cross-lingual Adaptation with Meta-learning

Wenxin Hou, Yidong Wang, Shengzhou Gao, Takahiro Shinozaki

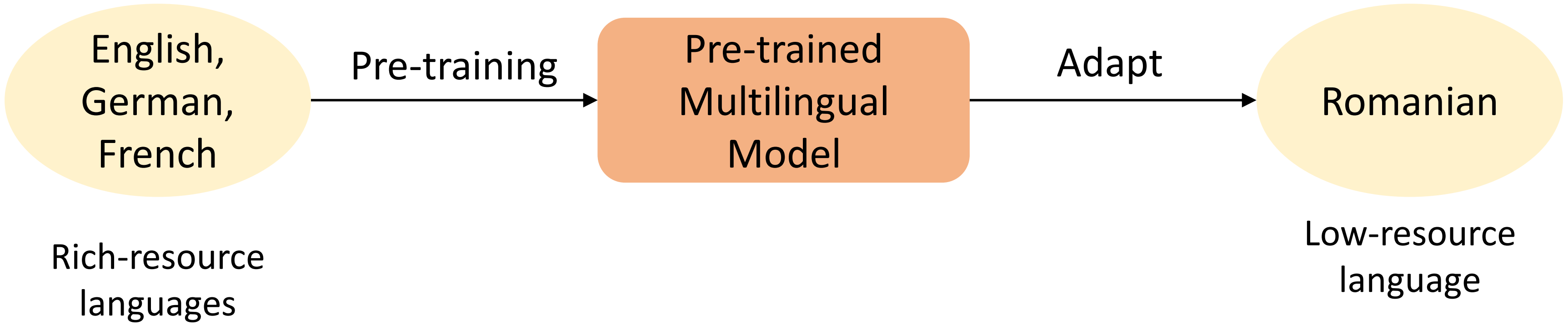
Tokyo Institute of Technology

IEEE ICASSP 2021

Background

- Low-resource automatic speech recognition (ASR) is a challenge for data-hungry end-to-end (E2E) models

- Cross-lingual ASR: adapt or extend a pre-trained multilingual model to a new unseen language



Conventional Methods

- Multilingual Joint Pre-training
  - Pre-train the ASR model on multiple languages
  - Fine-tune the pre-trained model on the target language

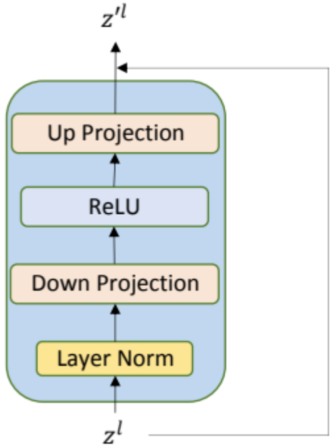
✗ Catastrophic forgetting
✗ Low parameter-efficiency
✗ Overfitting problem

- Multilingual Meta Pre-training
  - Pre-train the ASR model using Model-Agnostic Meta-Learning (MAML) on multiple languages
  - Fine-tune the pre-trained model on the target language

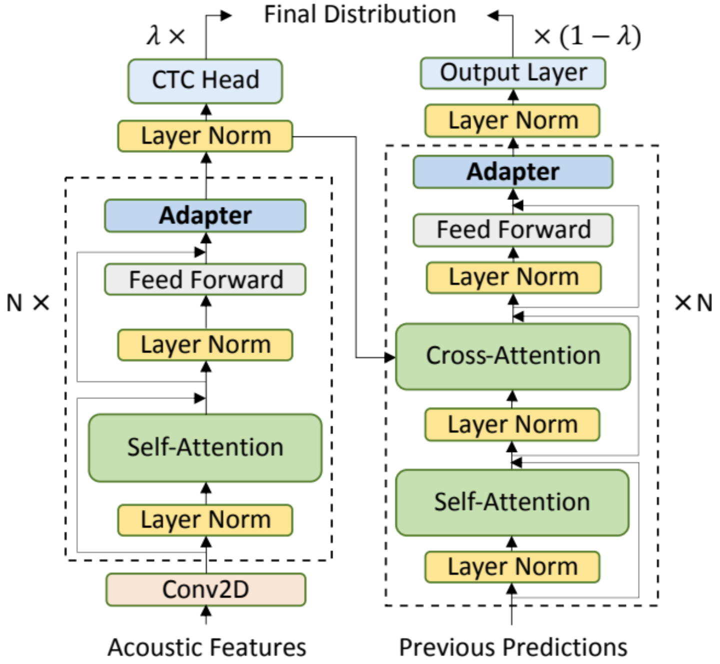
✗ Catastrophic forgetting
✗ Low parameter-efficiency
✗ High computational cost

Proposed Method

- Introduce Adapter [Bapna+, EMNLP-IJCNLP, 2019] module to improve parameter-efficiency



$$\mathrm{Adapter}(\mathbf{z}^l) = \mathbf{z}^l + \mathbf{W}_u^l \mathrm{ReLU}\left(\mathbf{W}_d^l\left(\mathrm{LayerNorm}\left(\mathbf{z}^l\right)\right)\right)$$



- Introduce meta-learning for fast adaptation
  - MAML [Finn+, ICML, 2017]

$$\mathcal{L}_{S_i^{val}}(f_{\theta'_{a,i}}) = \mathcal{L}_{S_i^{val}}(f_{\theta_a - \epsilon \nabla_{\theta_a} \mathcal{L}_{S_i^{tra}}(f_{\theta_a})}) \qquad \theta_a = \theta_a - \gamma \sum_{S_i^{val} \sim p(S^{val})} \nabla_{\theta_a} \mathcal{L}_{S_i^{val}}(f_{\theta'_{a,i}})$$

  - Reptile [Nichol+, arXiv, 2018]

$$\theta_{a,i_{k+1}} = \theta_{a,i_k} - \epsilon \nabla \mathcal{L}_{D_i}(f_{\theta_{a,i_k}}) \qquad \theta_a = \theta_a + \gamma \sum_{S_i \sim p(S)} (\theta_{a,i_K} - \theta_a)$$

Experimental Setup

- Dataset: Common Voice Corpus 5.1 [Ardila+, LREC, 2020]
- Data amount:

Lang.	Train Dur.(hrs)	#Train Utt.	#Test Utt.
or	0.45	319	84
hsb	1.48	808	379
br	2.84	3684	1953
ga-IE	2.10	2338	497
ro	3.04	2789	1372

- Baselines
  - Head-FT: fine-tune the language-specific heads only
  - Vanilla-Adapter: inject and train the adapters with random initialization
  - MOL-Adapter: pre-train the adapters on multiple source languages as initialization
- Metric: Word error rate (WER)

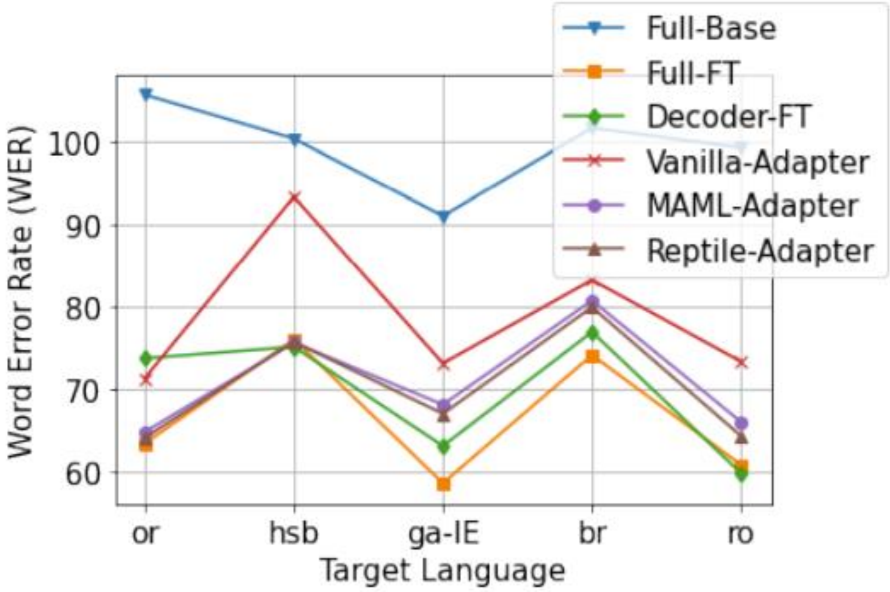
Results

- Quantitative Analysis

Method	or	hsb	ga-IE	br	ro
Head-FT	95.1	100.5	82.6	91.8	86.4
Vanilla-Adapter	71.3	93.3	73.1	83.2	73.3
MOL-Adapter	77.3	89.7	68.2	82.2	67.5
MAML-Adapter	64.8	**75.6**	68.1	80.7	66.0
Reptile-Adapter	**64.1**	75.7	**67.0**	**79.9**	**64.3**

- Impact of Trainable Parameters

Method	#Parameters
Full-Base & Full-FT	27,235K
Decoder-FT	9,550K
Adapters	381K



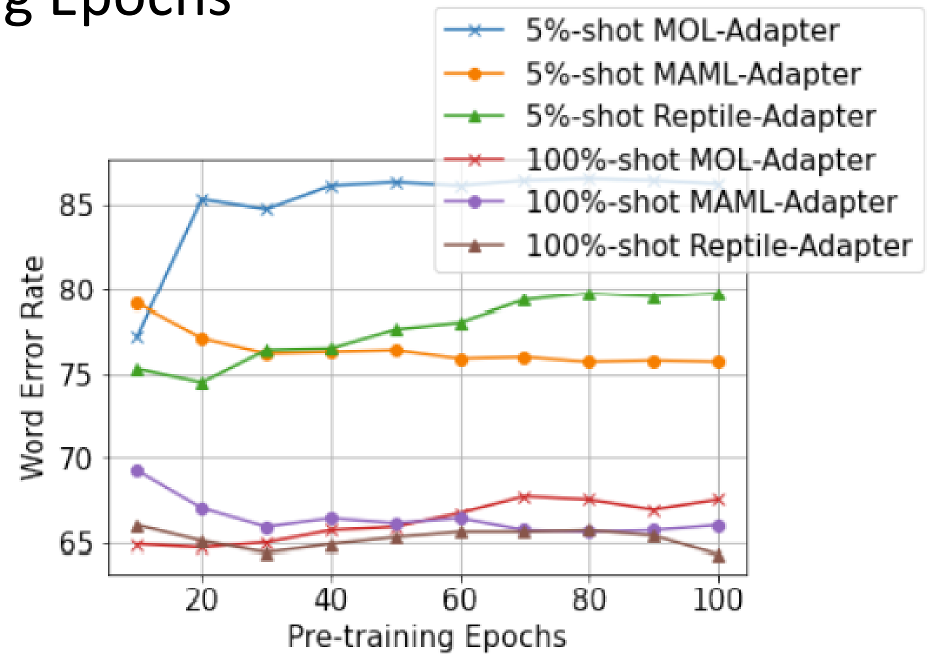
Additional Experiments

- Impact of Adaptation Data Size

Method	5%	10%	15%	30%	100%
Decoder-FT	87.9	70.5	**64.7**	**60.7**	**59.8**
Full-FT	77.3	73.2	67.8	65.7	60.8
Vanilla-Adapter	84.2	78.3	76.7	73.9	73.3
MOL-Adapter	86.2	78.4	72.6	69.1	67.5
MAML-Adapter	**75.7**	**69.9**	66.8	65.1	66.0
Reptile-Adapter	79.7	71.0	67.9	65.2	64.3

<u>Meta-Adapters are more robust to the adaptation data size</u>

- Impact of Pre-training Epochs



<u>MAML and Reptile hardly overfit</u>

Summary

- Combining meta-learning and adapters can result in fast and parameter-efficient cross-lingual adaptation for E2E ASR

- Meta-adapters achieve significant improvement compared with other parameter-efficient methods

- Future work aims to close the gap between parameter-efficient methods and full-model fine-tuning