Language Adaptive Cross-lingual Speech Representation Learning with Sparse Sharing Sub-networks

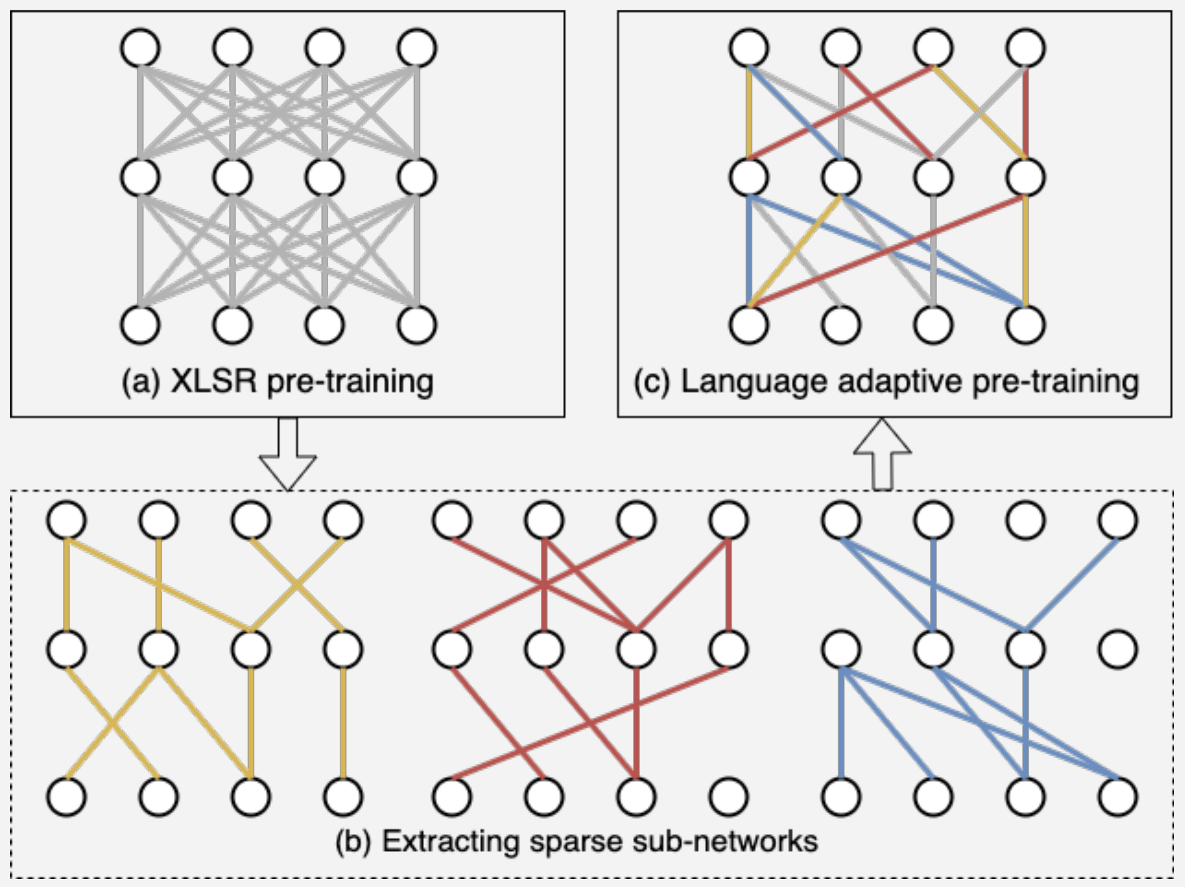
Yizhou Lu, Mingkun Huang, Xinghua Qu, Pengfei Wei, Zejun Ma
Speech & Audio Team, ByteDance AI Lab

#1773

Overview

- ❑ Standard XLSR model suffers from the language interference problem
  - ➢ Lacking language specific modeling ability
  - ➢ Limited model capacity
- ❑ We propose a sparse sharing sub-networks based language adaptive training approach
- ❑ The proposed S3Net achieves **9.8%/7.4%** relative improvements over XLSR base/large, without requiring additional learnable params

Sparse Sharing Sub-networks


(a) XLSR pre-training
(c) Language adaptive pre-training
(b) Extracting sparse sub-networks

Training procedure of the proposed S3Net:

(a) XLSR pre-training (**Optional**)

(b) Extracting subnet for each language
- ➢ Each subnet shall be able to maintain the full network's accuracy

(c) Language adaptive training with S3Net
- ➢ Sparse sharing structure automatically distributes both shared and language specific parameters at each layer

Extracting Sub-networks

We experiment with two approaches of extracting sparse sub-networks:
- ➢ Lottery Ticket Hypothesis (**Accurate!**)
- ➢ First Order Taylor Expansion (**Efficient!**)

Extracting subnets with LTH

- ➢ Start from a pre-trained XLSR model or from scratch, denote the starting point as θ
- ➢ For each language l, train model θ with specific language data D^l for a few steps to get language specific model $\hat{\theta}^l$
- ➢ One-shot magnitude pruning on $\hat{\theta}^l$, those parameters with lowest magnitude are pruned out, the structure is denoted with a binary mask m^l, with $\theta^l = m^l \odot \theta$
- ➢ One can also apply iterative pruning strategy for a more accurate subnet

Extracting subnets with TE

The importance of a parameter can be quantified by the error induced by removing it:

$$\mathcal{I}_i^l = [\mathcal{L}(D^l, \theta) - \mathcal{L}(D^l, \theta | \theta_i = 0)]^2$$

The above equation can be approximated with first order Taylor Expansion:

$$\mathcal{I}_i^l \approx (g_i^l \theta_i)^2$$

where $g_i^l = \frac{\partial \mathcal{L}(D^l, \theta)}{\partial \theta_i}$ is the gradient for θ_i that can be efficiently calculated with backward propagation

Language Adaptive Training

Once we obtain all masks $m_1, m_2, \dots, m_L$, we apply language adaptive training:
- ➢ Each batch only contain utterances from one language
- ➢ Multilingual batches are sampled with a multinomial distribution: $p_l \sim (\frac{n_l}{N})^\alpha$
- ➢ For each batch, only $\theta^l = m^l \odot \theta$ participate the forward and backward calculation

Experiments

For pre-training and finetuning, we follow the setup in XLSR paper:
- ➢ We use Common Voice dataset for pre-training
- ➢ We adopt CTC criterion and evaluate the multilingual performance of pre-trained model

Model	es	fr	it	ky	nl	ru	sv	tt	zh	Avg
Number of audio data	168h	353h	90h	17h	29h	55h	3h	17h	50h	-
XLSR-10	10.8	12.8	15.1	8.5	15.4	11.8	22.1	8.1	24.2	14.3
S3Net-TE	9.9	12.0	14.4	7.8	14.7	11.3	22.1	7.7	23.9	13.8
S3Net-LTH	**8.7**	**10.8**	**12.4**	**7.5**	**14.1**	**10.1**	**22.0**	**7.2**	**22.9**	**12.9**
XLSR-10 (Large)	9.0	10.6	12.7	6.8	12.8	10.1	19.9	6.6	21.5	12.2
S3Net-TE (Large)	8.4	10.5	12.4	6.7	12.5	10.1	19.6	6.3	21.6	12.0
S3Net-LTH (Large)	**7.3**	**9.2**	**10.4**	**6.3**	**12.1**	**9.4**	**19.5**	**6.1**	**21.5**	**11.3**

- ✓ S3Net-LTH models perform better than S3Net-TE, achieve **9.8%/7.4%** relative improvements over XLSR models
- ✓ S3Net achieves more improvements on high resource languages, with **17.8%/16.7%** improvements for base/large

Comparison with other adaptation methods

Model	#Param	CV-Eval High	CV-Eval Low	CV-Eval Avg
XLSR-10	95M	12.9	15.0	14.3
+ Gating Network	95M	12.2	14.7	13.9
+ Adapter	143M	11.5	14.1	13.2
S3Net-LTH	95M	**10.6**	**14.0**	**12.9**
XLSR-10 (Large)	317M	10.8	13.0	12.2
+ Gating Network	317M	10.4	12.8	12.0
+ Adapter	444M	10.4	12.9	12.1
S3Net-LTH (Large)	317M	**9.0**	**12.5**	**11.3**

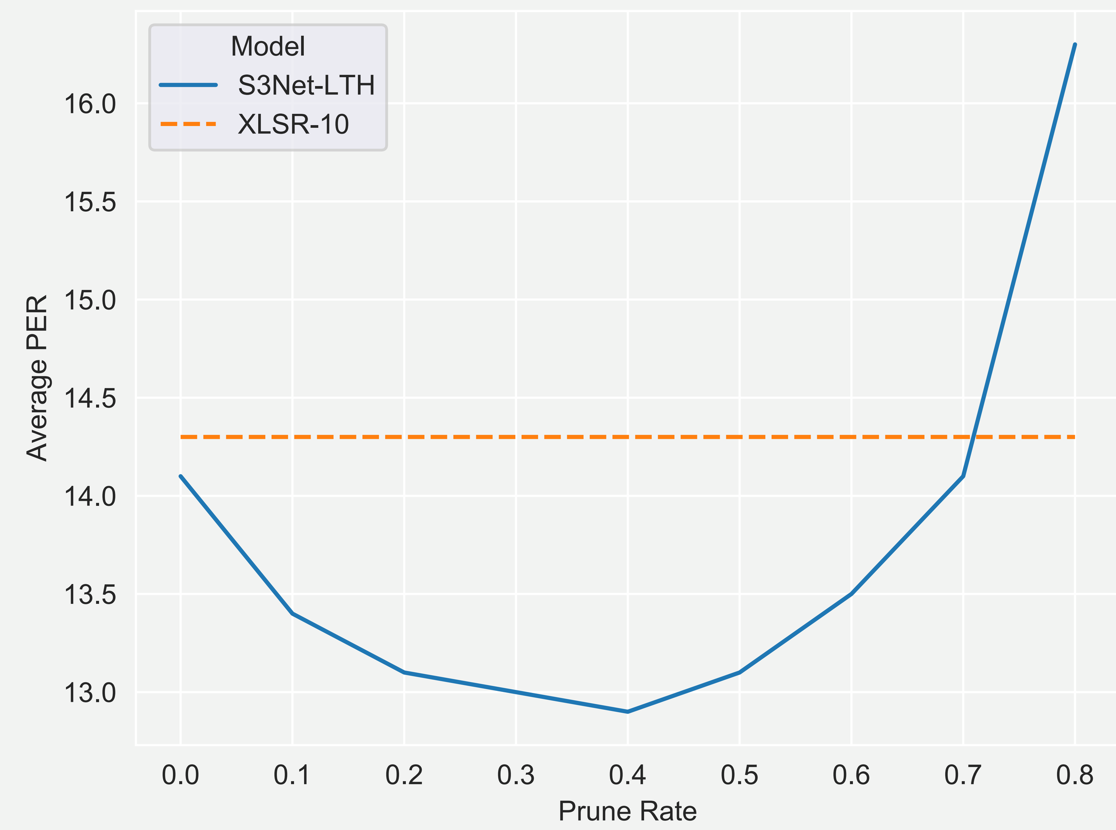
- ✓ S3Net-LTH outperforms other adaptation methods while requiring fewer parameters

Ablation studies

Model	Type	Strategy	CV-Eval High	CV-Eval Low	CV-Eval Avg
XLSR-10	N/A	N/A	12.9	15.0	14.3
S3Net	Global	LTH	10.8	**14.0**	13.0
	Global	Random	14.2	16.9	16.0
	Layerwise	TE	12.1	14.6	13.8
	Layerwise	LTH	**10.6**	**14.0**	**12.9**

- ✓ Layerwise pruning slightly outperforms global pruning
- ✓ Random pruning demonstrates the effectiveness of proposed methods

Pruning rate curve



Conclusion and Future Work

Language adaptive pre-training with S3Net
- ❑ S3Net alleviates language interference problem
- ❑ Two different pruning strategies are explored: TE & LTH
- ❑ S3Net outperforms other adaptation methods while requiring fewer parameters

Future Work:
- ❑ Structured sparsity and N:M sparsity for network acceleration

icassp 2022
Singapore, China, Online