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

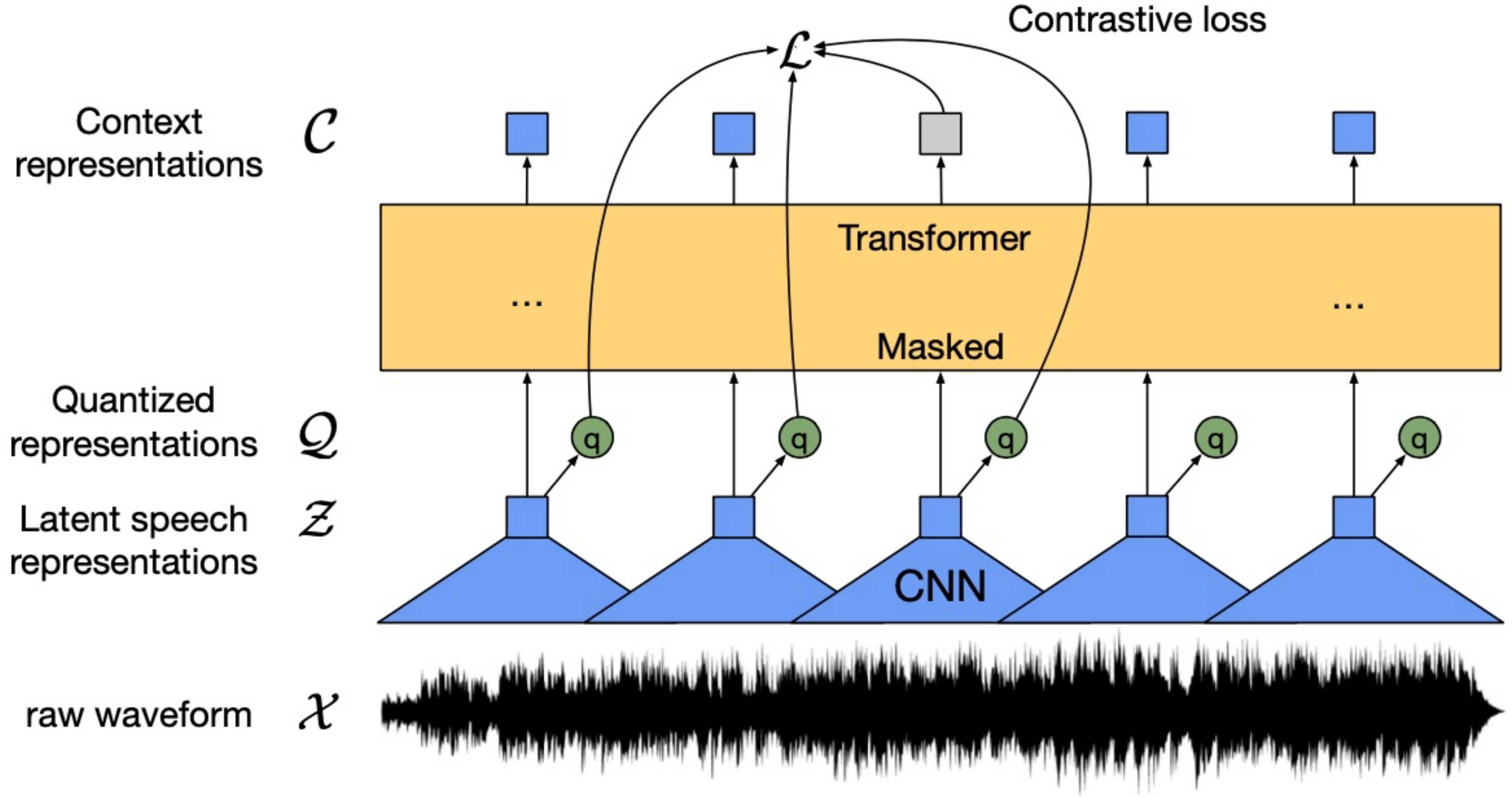
May 2022

Background

- Self-supervised learning provides an efficient way to utilize unlabeled data
  - Typical models include Wav2vec 2.0, HuBERT, WavLM, Data2vec
  - ASR models can be built with very small amounts of labeled data while maintaining very good accuracy

- Cross-lingual speech representation learning (XLSR)
  - Multilingual pre-training outperforms monolingual pre-training in low resource languages
  - It simplifies the procedure, with no need of training seed models for each language individually
  - For downstream multilingual applications, such as multilingual ASR and multilingual speech translation

Cross-lingual Speech Representation Learning (XLSR)



Language interference

XLSR extends Wav2vec 2.0 framework, and learns representation from different languages with a shared network

[1] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 2020, 33: 12449-12460.

Language Interference Problem

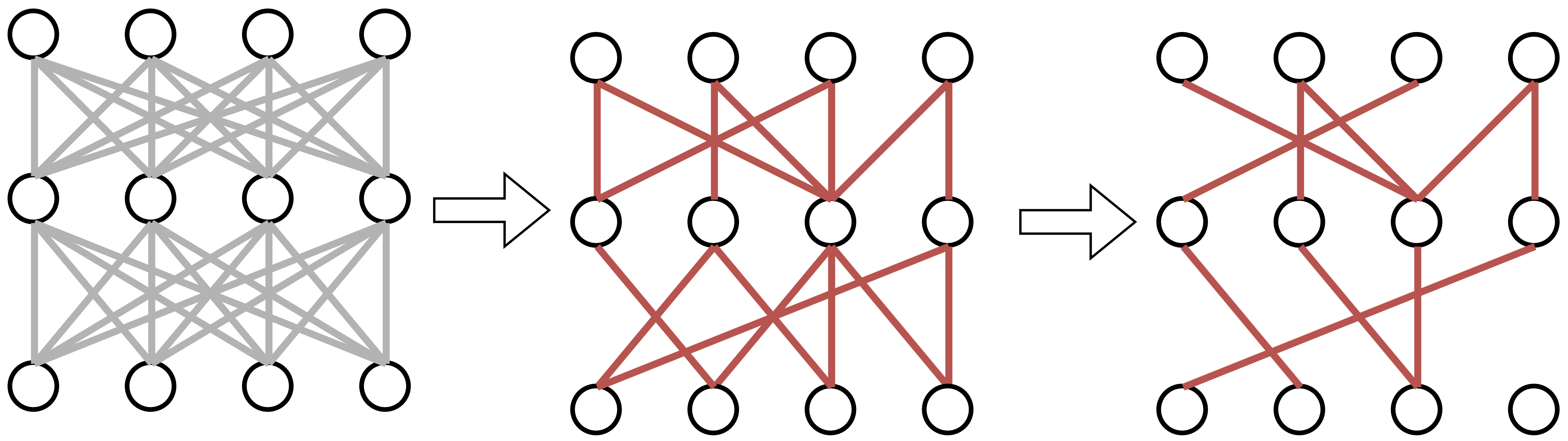
While multilingual pre-training enables better transfer to low resource languages, the model also needs to share its capacity across multiple languages, resulting in inferior performance on high-resource languages.

- Adaptation perspective
  - E.g. auxiliary LID features, LHUC, light weight adapters, decoupled multilingual encoder/decoder
  - The inserted module size, structure and injection position are all important factors to consider [2]

- Capacity perspective
  - 1B or even 10B parameters to accommodate multiple languages and vast amounts of data

[2] Gong X, Lu Y, Zhou Z, et al. Layer-Wise Fast Adaptation for End-to-End Multi-Accent Speech Recognition. Proc. Interspeech 2021, 2021: 1274-1278.

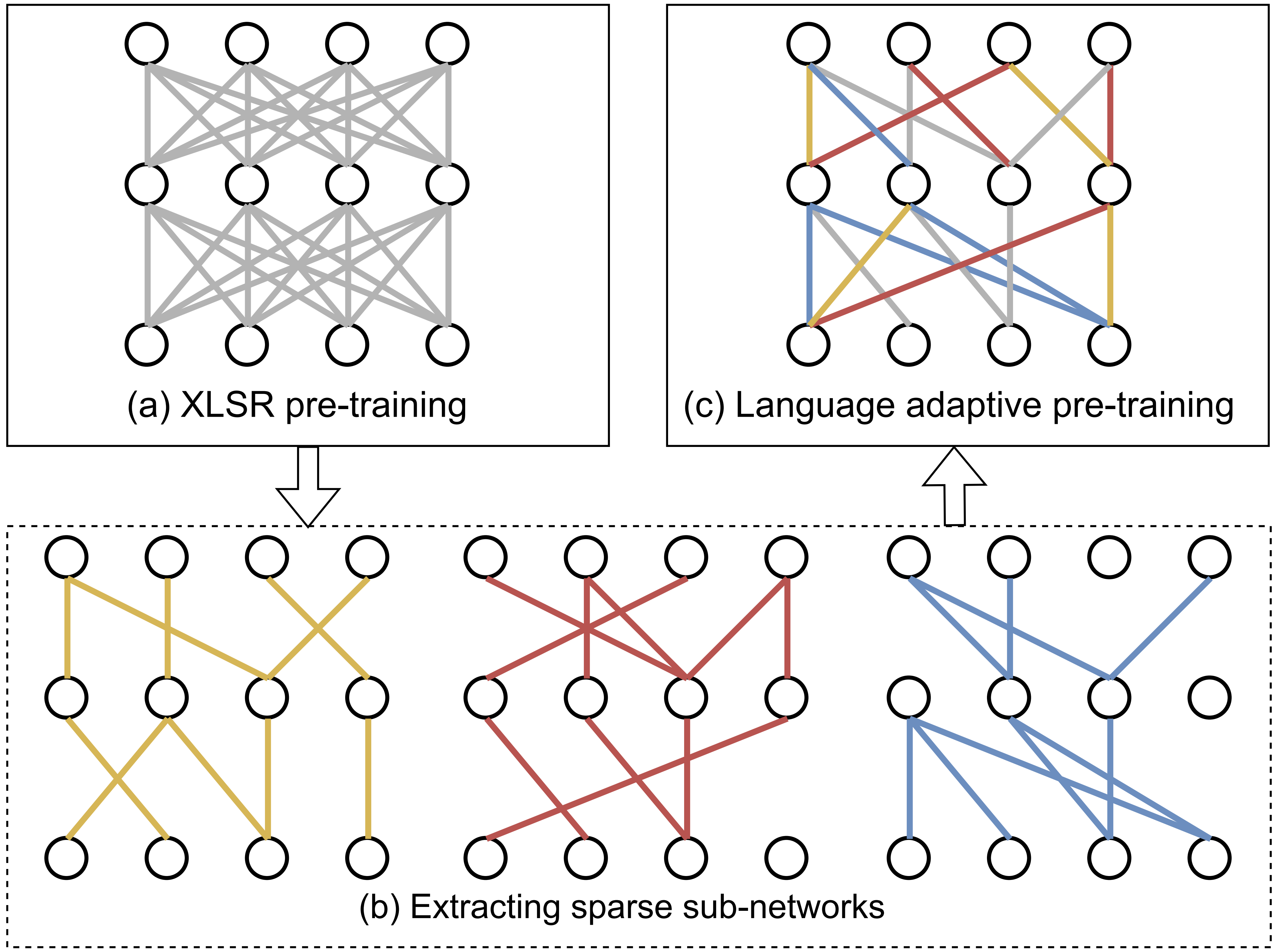
Motivation



The Lottery Ticket Hypothesis: *"A randomly-initialized, dense neural network contains a sub-network that is initialized such that, when trained in isolation, it can match the test accuracy of the original network after training for at most the same number of iterations."*

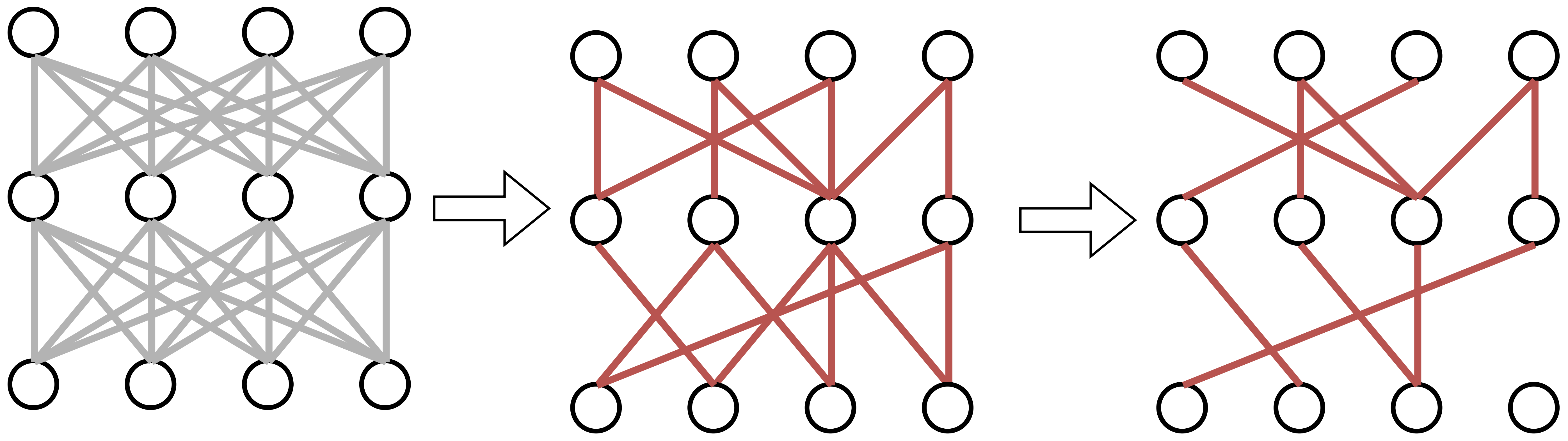
[3] Frankle, Jonathan, and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." *International Conference on Learning Representations*. 2018.

Overview of the Proposed Method

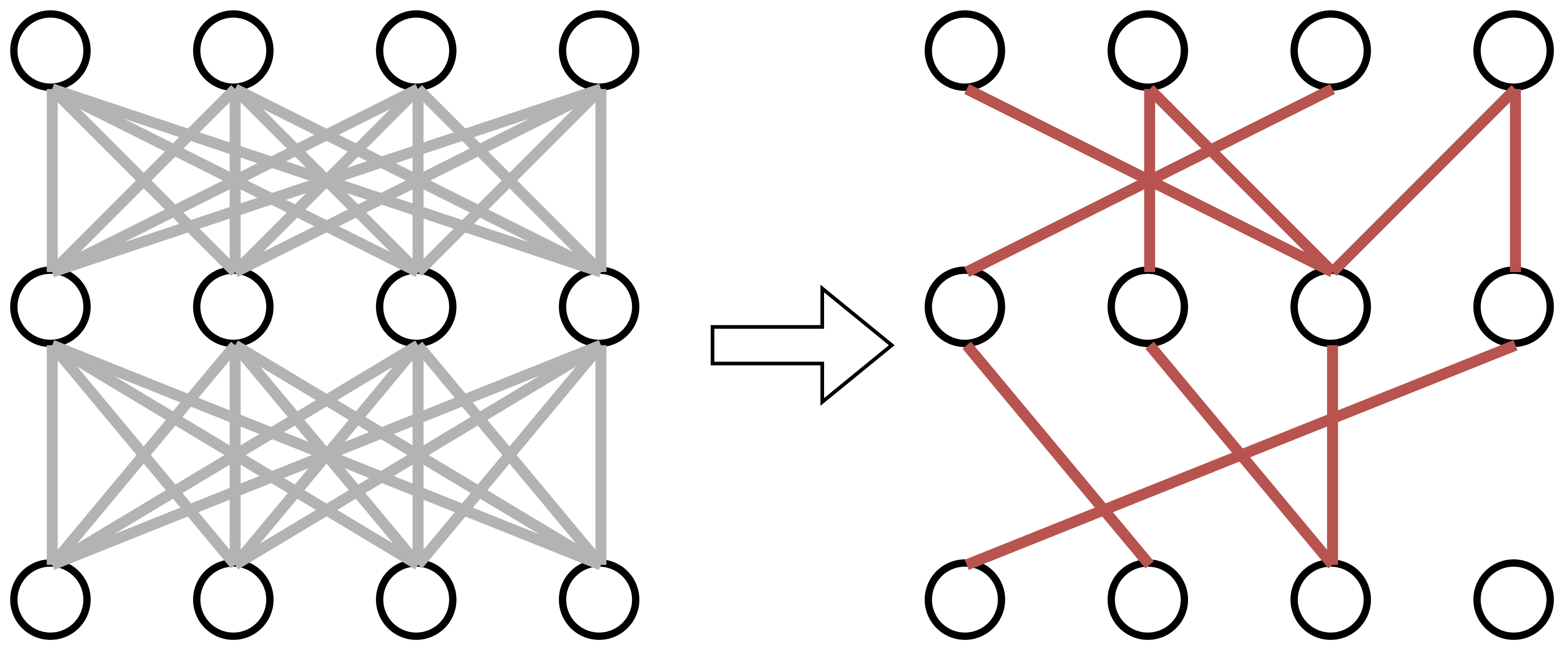


We extract a sub-network for each language, and all the sparsely shared sub-networks are jointly trained

Extracting Sub-networks with Lottery Ticket Hypothesis

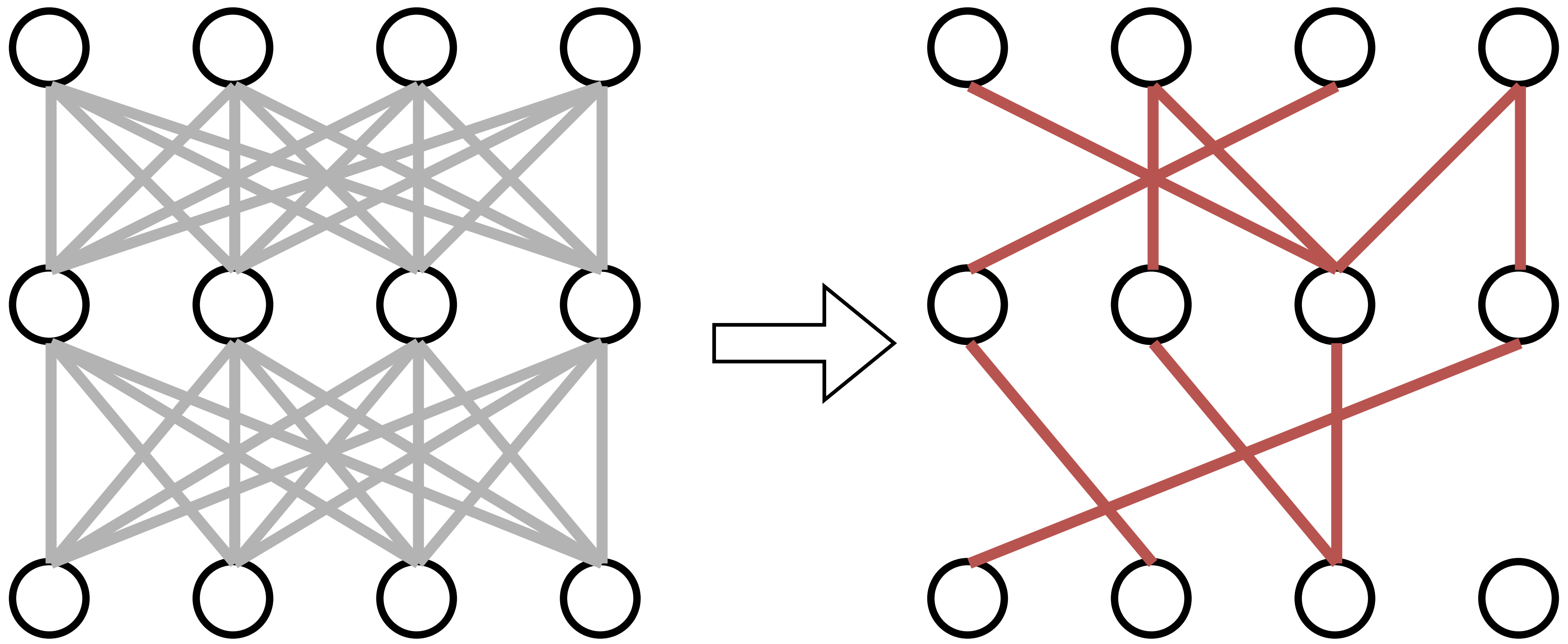


Iterative Magnitude Pruning: Training -> Pruning -> Resetting -> … -> Training -> Pruning



We adopt a simple one-shot magnitude pruning instead, and start from a pre-trained XLSR model

Extracting Sub-networks with Taylor Expansion

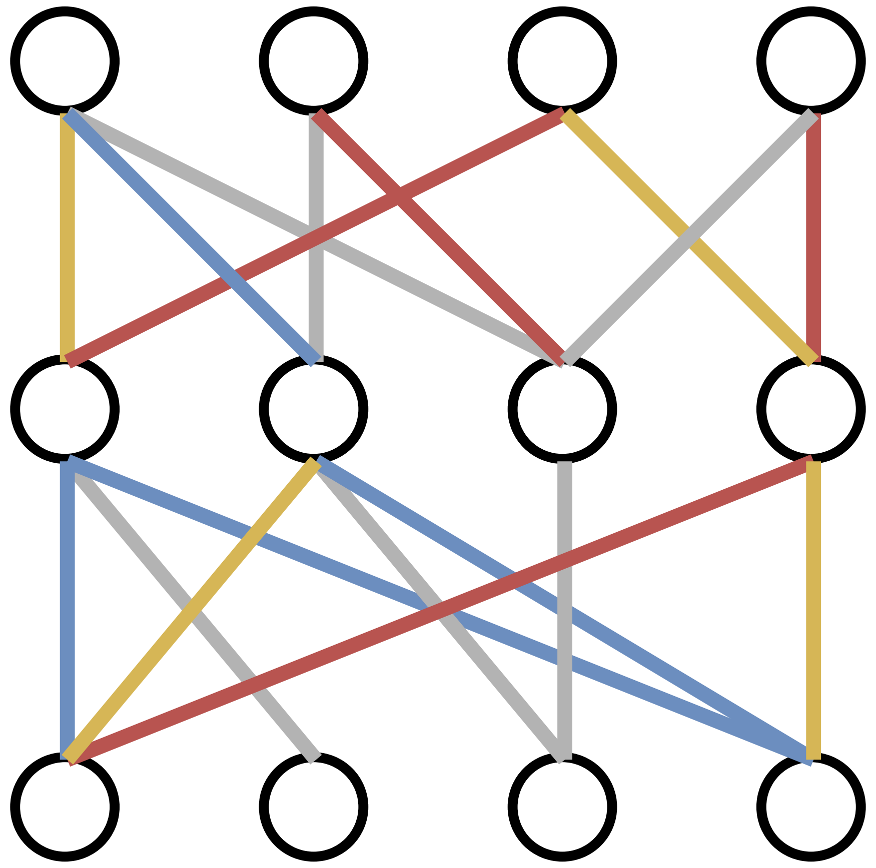


One-shot Magnitude Pruning is still computationally expensive as we have to deal with ten languages...

Taylor Expansion based pruning with importance score, the importance of a parameter can be quantified by the error induced by removing it:

$$\mathcal{I}_i^l = \left[\mathcal{L}(\mathcal{D}^l, \boldsymbol{\theta}) - \mathcal{L}\left(\mathcal{D}^l, \boldsymbol{\theta} \mid \theta_i = 0\right)\right]^2 \Longrightarrow \mathcal{I}_i^l \approx (g_i^l \theta_i)^2$$

Language Adaptive Training with S3Net



Once we have extracted all sub-networks, we re-started from the pre-trained XLSR model. Only the sub-network from the corresponding language will participate the forward computation and be updated.

Experiments

Model	Pre-trained data	es	fr	it	ky	nl	ru	sv	tt	zh	Avg
Number of unlabeled audio data		168h	353h	90h	17h	29h	55h	3h	17h	50h	
Baselines from XLSR [10]											
XLSR-Monolingual	CV-Mono*	6.8	10.4	10.9	29.6	37.4	11.6	63.6	21.4	31.4	24.8
XLSR-10	CV-Multi*	9.4	13.4	13.8	8.6	16.3	11.2	21.0	8.3	24.5	14.1
XLSR-10 (Large)	CV-Multi*	7.7	12.2	11.6	7.0	13.8	9.3	20.8	7.3	22.3	12.4
Re-run baselines and our models											
XLSR-10	CV-Multi	10.8	12.8	15.1	8.5	15.4	11.8	22.1	8.1	24.2	14.3
S3Net-TE	CV-Multi	9.9	12.0	14.4	7.8	14.7	11.3	22.1	7.7	23.9	13.8
S3Net-LTH	CV-Multi	**8.7**	**10.8**	**12.4**	**7.5**	**14.1**	**10.1**	**22.0**	**7.2**	**22.9**	**12.9**
XLSR-10 (Large)	CV-Multi	9.0	10.6	12.7	6.8	12.8	10.1	19.9	6.6	21.5	12.2
S3Net-TE (Large)	CV-Multi	8.4	10.5	12.4	6.7	12.5	10.1	19.6	6.3	21.6	12.0
S3Net-LTH (Large)	CV-Multi	**7.3**	**9.2**	**10.4**	**6.3**	**12.1**	**9.4**	**19.5**	**6.1**	**21.5**	**11.3**

High resource languages achieve more improvements

Table 1. Evaluation results on CommonVoice dataset. The last column is the averaged PER on nine languages. Re-run baselines and our models are all pre-trained on ten languages, and evaluated on nine languages with shared vocabulary using CTC criterion. *: They use different version of the CommonVoice dataset, but the data size is the same as ours.

We reproduce similar results as XLSR, and both S3Net-TE and S3Net-LTH consistently outperforms XLSR model

Experiments

Table 2. Comparison of different adaptation methods. Multilingual evaluation results are averaged on high resource languages (High), low resource languages (Low) and all nine languages (Avg).

Model	#Params	CV-Eval		
		High	Low	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 all other adaptation methods, while requiring fewer parameters

Table 3. Analysis of different sub-networks. Models are trained with base structure and prune rate is set to 0.4 throughout the experiments.

Model	#Mask	Type	Strategy	CV-Eval		
				High	Low	Avg
XLSR-10	N/A	N/A	N/A	12.9	15.0	14.3
S3Net	1	Global	LTH	13.0	15.3	14.5
	5	Global	LTH	10.8	15.0	13.6
	10	Global	LTH	10.8	**14.0**	13.0
	10	Global	Random	14.2	16.9	16.0
	10	Layerwise	TE	12.1	14.6	13.8
	10	Layerwise	LTH	**10.6**	**14.0**	**12.9**

Random pruning experiment demonstrates the effectiveness of the proposed method

Experiments

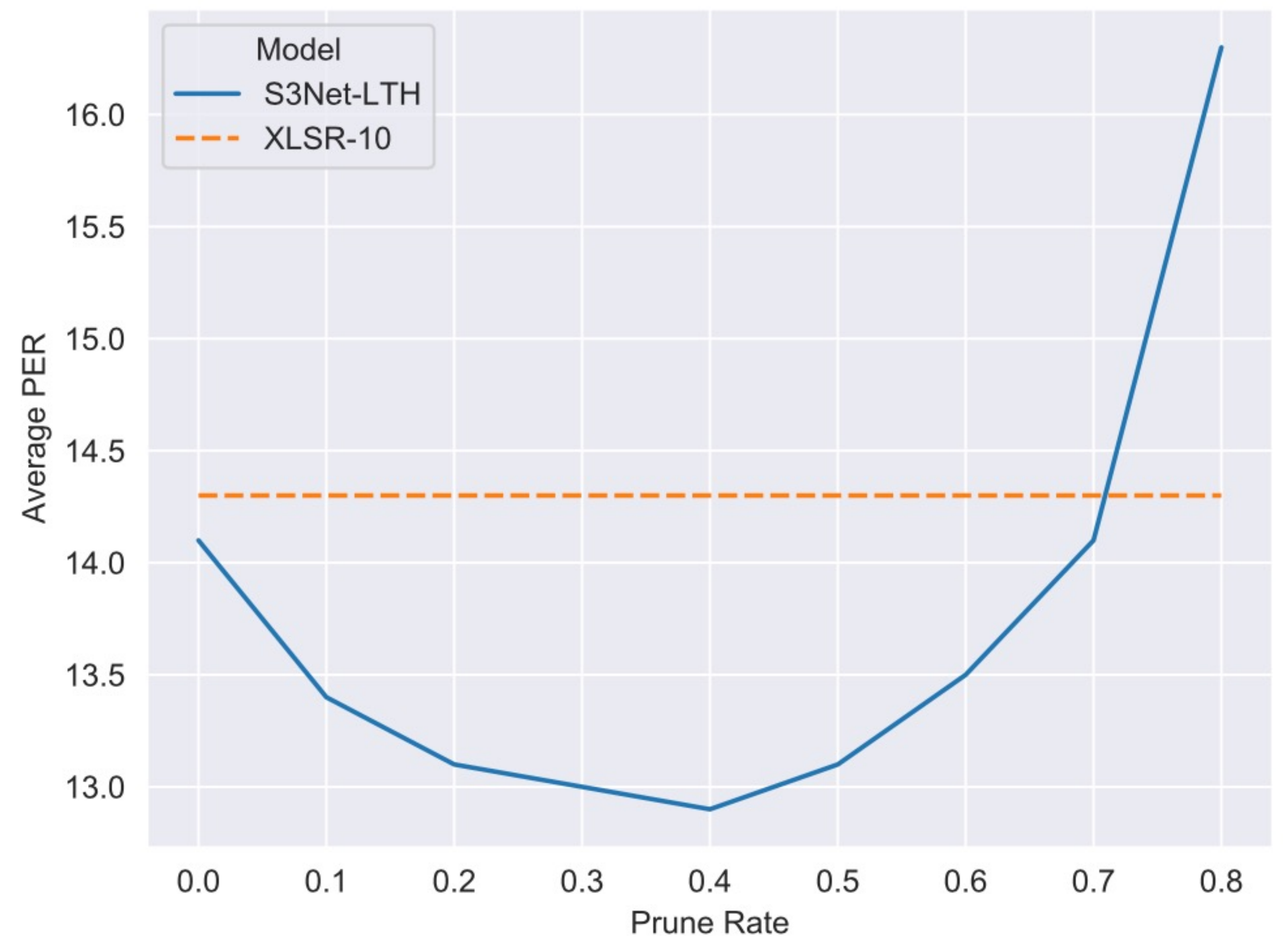


Fig. 2. Evaluation results of different prune rate for S3Net-LTH.

As the pruning rate increases from 0.0 to 0.4, the language interference problem is gradually alleviated;
But when it continues to increase, the sub-networks can not maintain the full network's accuracy, thus start to degrade

Conclusion and Future Work

- Our proposed S3Net helps alleviating the language interference problem, especially for high resource languages
- We experiment with two different approaches of extracting sub-networks: LTH and TE
- Our proposed S3Net outperforms other adaptation methods while requiring fewer parameters
- In the future, we plan to study structured sparsity and N:M sparsity for network acceleration