Dynamic ASR Pathways: An Adaptive Masking Approach Towards Efficient Pruning of A Multilingual ASR Model

Jiamin Xie, Ke Li, Jinxi Guo, Andros Tjandra, Yuan Shangguan, Leda Sari, Chunyang Wu, Junteng Jia, Jay Mahadeokar, Ozlem Kalinli
University of Texas at Dallas, Meta AI

Introduction

- Neural network pruning is effective for compressing ASR models.
- Pruning a multilingual ASR model entails several rounds of pruning and re-training.
- Propose an **adaptive masking** approach to **efficiently prune** a multilingual ASR model.
- Proposed method **adapts a pruning mask with data** in training.
- Compare performance with existing methods in two scenarios:
  1. Sparse monolingual models for each language.
  2. One sparse multilingual model for all languages.

Recap: Existing Pruning Methods

- Suppose a dense neural net $f(x;\theta)$ with a binary pruning mask $m \in \{0,1\}^{|\theta|}$,
- **Iterative magnitude pruning (IMP) [1]**
  - *Initialization*: $\theta = \theta_0$ *and* $m = \mathbf{1}$, where θ_0 are the pre-trained weights

    Repeat
    1. Train $f(x; m \odot \theta)$ for T steps to obtain $f(x; m \odot \theta_T)$.
    2. Prune $p\%$ of total weights that has small magnitudes from $m \odot \theta_T$
    3. Assign θ_T to θ for the next iteration.

    Until m reaches the target sparsity
- **Lottery ticket hypothesis (LTH) [2]**
  - Rewind the sub-network by assigning θ_0 to θ in step 3 instead.
- **ASR Pathways [3]**
  - Stage (1): Identify **language-specific** sub-networks by IMP or LTH.
  - Stage (2): Fine-tune each pathway with a monolingual batch.

Drawback in Existing Pruning Methods

- The sub-network structure remains **fixed** throughout training.
  - May commit early to a sub-optimal choice.
  - May propagate errors to further fine-tuning stages.

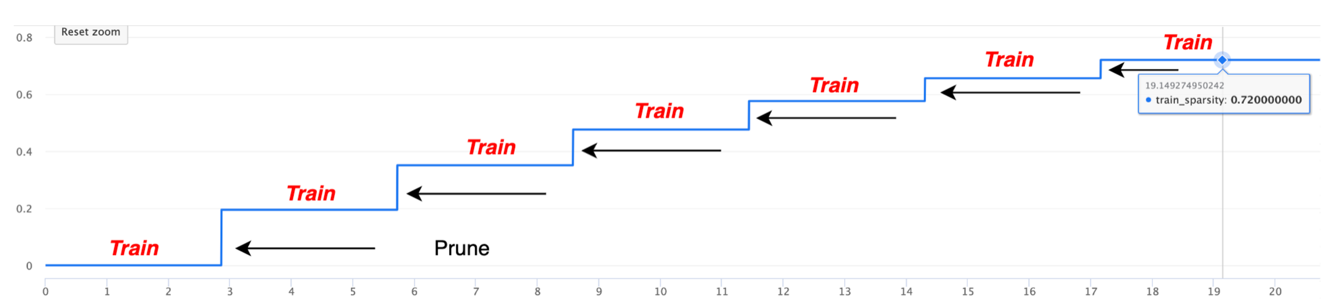

Figure 1. Progressive pruning schedule [4]: prune a network at a low sparsity and incrementally steps up to the *target* sparsity. The pruning mask can be fixed for the training cycles at any sparsity level.

Proposed Adaptive Masking *(monolingual)*

- **Masked-out: prune "softly"**
  - Prune weights in the network by setting them to zero.
  - Keep pruned weights trainable.
- **The adaptation step n**
  - Re-rank the magnitude of weights after n steps of training.
  - Note $n < T$, where T is the pruning interval.

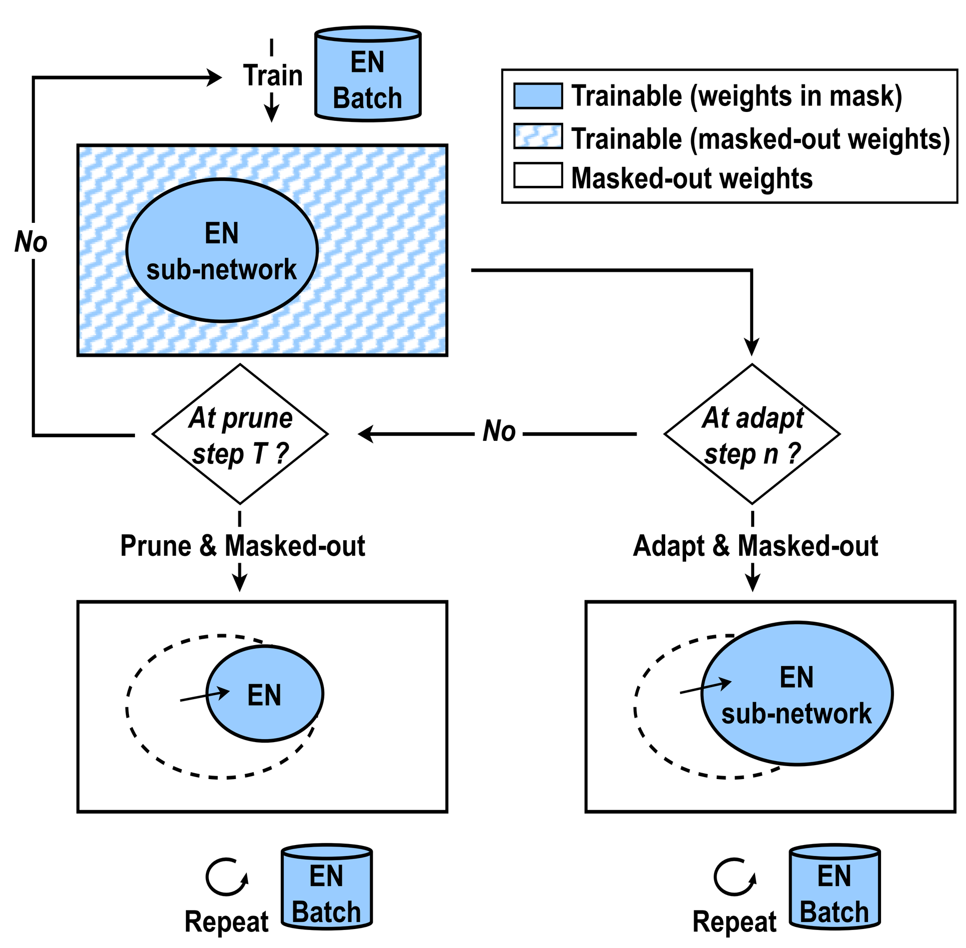

Figure 2. Flowchart of the training and pruning process with adaptive masking enabled for monolingual data

Proposed Adaptive Masking *(multilingual)*

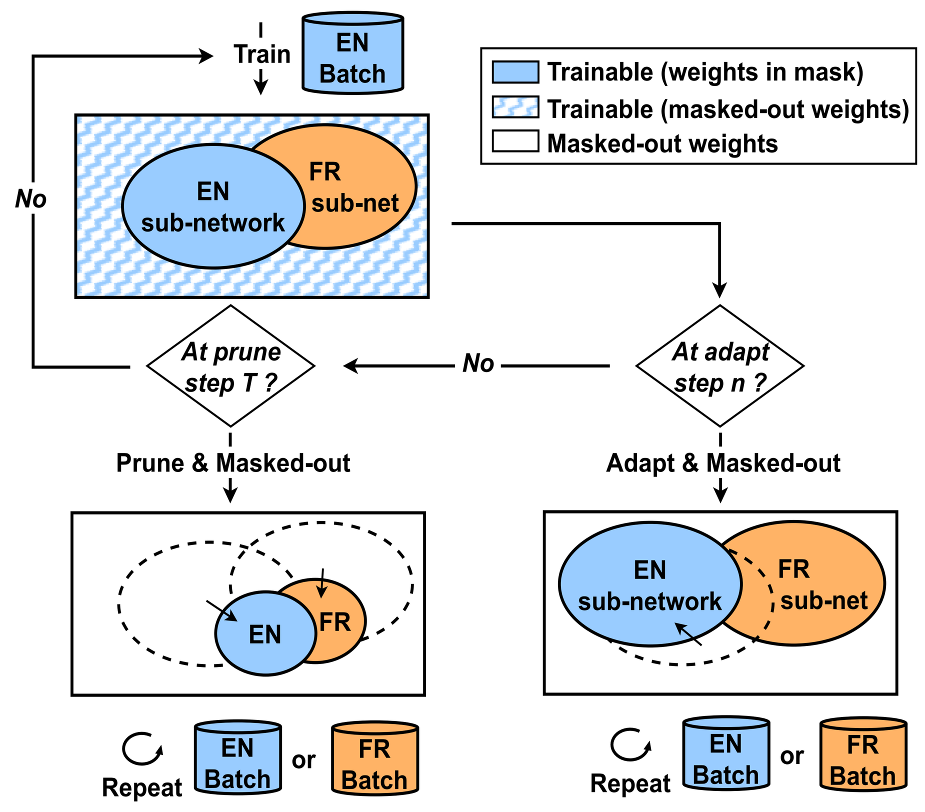

Figure 3. Flowchart of the training and pruning process with adaptive masking enabled for multilingual data

- For a language z, define a "free-zone" sub-network mask $\boldsymbol{m_r} = \mathbf{1} - \cup_{l \text{ in } L,\, l \neq z}\, \boldsymbol{m_z}$.
  - Prune "softly" from weights in $\boldsymbol{m_{z,r}} \odot \theta$, where $\boldsymbol{m_{z,r}} = \boldsymbol{m_z} \cup \boldsymbol{m_r}$.
  - The procedure is **language-specific** by maintaining monolingual batches.
- For pruning on all languages after T steps, prune "softly" for each language.
  - The procedure is **language-agnostic,** sharing weights newly trained.

Experiment Setup & Results

- Dense model: a streaming RNN-T model [5].
- Dataset: Multilingual Librispeech (MLS) dataset [6].
- **Scenario 1: a consistent <u>5.3% relative</u> WER reduction.**
  - **One less training stage** compared to ASR Pathways.

Stage	Model	Mask can change?	Sparsity	Monolingual or Multilingual training?	EN	FR	IT	NL	Avg.
Ref.	56M Dense	/	0%	Monolingual	12.15	16.00	27.62	23.23	19.75
(1)	187M Dense	/	0%	Multilingual	12.91	10.90	16.94	17.56	14.58
(2)	LAP	No	70%	Multilingual	13.82	11.98	27.71	19.32	18.21
	IMP	No	70%	Monolingual	10.74	11.26	17.90	18.38	14.57
	LTH	No	70%	Monolingual	10.80	10.38	18.44	17.48	14.28
(3)	ASR Pathways (IMP-70%)	No	70%	Multilingual	11.15	10.68	17.53	16.90	14.06
	ASR Pathways (LTH-70%)	No	70%	Multilingual	11.39	10.20	17.58	15.84	13.75
(2) *Proposed*	IMP	Yes	70%	Monolingual	**10.07**	10.90	17.21	16.98	13.79
	LTH	Yes	70%	Monolingual	10.54	**9.91**	**17.06**	**16.63**	**13.53**

Table 1. WER (%) results on the MLS test set, pruning a dense multilingual ASR model. The proposed approach allows the mask to change in training and is compared to other pruning methods for **monolingual training scenario**.

- **Scenario 2: a better <u>5.8% relative</u> WER reduction**.
  - **Efficient pruning** starting from a language-agnostic pruning (LAP) mask.
  - **Strong extensions** to more languages.

Model	Initialization	Mask change?	Sparsity	FR	NL	Avg.
ASR Pathways	LTH-70%	No	70%	10.73	16.23	13.48
ASR Pathways	LAP-70%	No	70%	11.98	19.32	15.65
Dynamic ASR Pathways	LTH-70%	Yes	70%	11.31	15.55	13.43
	LTH-50%	Yes	70%	**10.48**	**14.92**	**12.70**
	LTH-20%	Yes	70%	10.99	16.17	13.58
Dynamic ASR Pathways	LAP-70%	Yes	70%	10.98	16.54	13.76
	LAP-50%	Yes	70%	**10.82**	**16.25**	**13.54**
	LAP-20%	Yes	70%	10.88	16.43	13.65

Table 2. WER (%) results on the MLS test set, utilizing language- specific pruning masks. The proposed approach is compared to an existing method for **bilingual training scenario**.

Model	Initialization	Sparsity	EN	FR	IT	NL	Avg.
ASR Pathways	LTH-70%	70%	**13.56**	10.53	17.10	16.37	14.39
Dynamic ASR Pathways	LTH-50%	70%	14.84	**10.35**	**16.10**	**15.15**	**14.11**

Table 3. WER (%) results on the MLS test set, utilizing language- specific pruning masks. The proposed approach is compared to an existing method, **extending to four languages**.

References

[1] Song Han, Jeff Pool, John Tran, and William Dally, "Learning both weights and connections for efficient neural network," in NeuralPS 2015.
[2] Jonathan Frankle and Michael Carbin, "The lottery ticket hy- pothesis: Finding sparse, trainable neural networks," in ICLR 2019.
[3] Mu Yang, Andros Tjandra, Chunxi Liu, David Zhang,Duc Le, and Ozlem Kalinli, "Learning asr pathways: A sparse multi- lingual asr model," in ICASSP 2023.
[4] Michael H. Zhu and Suyog Gupta, "To prune, or not to prune: Exploring the efficacy of pruning for model compression," in ICLR 2018.
[5] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in ICASSP 2021.
[6] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Syn- naeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," arXiv preprint arXiv:2012.03411, 2020

Take a photo to learn more: