2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) | 979-8-3503-0067-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/APSIPAASC58517.2023.10317230

Effective Fine-tuning Method for Tibetan Low-resource Dialect Speech Recognition

Jiahao Yang*, Jianguo Wei†, Kuntharrgyal Khysru, Junhai Xu, Wenhuan Lu†, Wenjun Ke, Xiaokang Yang
* Tianjin International Engineering Institute, Tianjin University, Tianjin, China
† College of Intelligence and Computing, Tianjin University, Tianjin, China
E-mail: {jianguo,wenhuan}@tju.edu.cn

***Abstract*—Tibetan is a distinctive and culturally rich language spoken by millions of people across the Tibetan Plateau and surrounding regions. Exploring the application of speech recognition technology to Tibetan has special significance for preserving language diversity and fostering cultural integration. Moreover, Tibetan comprises a multitude of distinct dialects, which present a hurdle for reusing speech recognition models. In low-resource dialect tasks, conventional approaches endeavor to transfer well-trained models from linguistically akin languages to the target. However, recent studies have shown that an indiscriminate fine-tuning of all parameters may disrupt the feature extractor of the pre-trained model, leading to catastrophic forgetting. This paper introduces an innovative fine-tuning method grounded in model adaptation. Aimed at training automatic speech recognition (ASR) models within the constraints of limited training data and cross-dialect transfer, our novel approach refines a select group of language-specific parameters, leading to robust performance. These parameters, signified by a sparse binary mask identical to the model, circumvent the need for additional parameters. Experiments conducted on two downstream low-resource Tibetan languages show that our proposed methodology outperforms the traditional fine-tuning and adapter based fine-tuning.**

I. Introduction

Influenced by history and geography, Tibetan has gradually evolved various dialects and accents, and it is widely believed that there are more than twenty different spoken forms. Generally speaking, there are three main dialects of Tibetan language: Amdo Tibetan, Khampa Tibetan, and U-Gtsang Tibetan. Due to the linguistic diversity of Tibetan and the problem of limited data, the application of Tibetan speech recognition technology trails behind [1], [2].

Automatic Speech Recognition (ASR) has made remarkable progress with the development of deep learning. Transformer-based [3]–[5] models greatly benefit from a large number of training parameters. Despite this, in scenarios where data resources are restricted, the potential for enhancements remains somewhat hampered. For some low-resource languages, fine-tuning speech recognition models trained on other dialects within the same language family has become a more effective paradigm [6]–[9]. Meanwhile, Choosing an appropriate fine-tuning strategy holds significance in bolstering the model's generalization ability [10], [11].

The advantages of neural networks in transferring knowledge between linguistically related languages are evident, primarily due to the efficiency of adaptation depends on various factors, including the distribution differences between the source and target languages [12]. In the presence of insufficient target data that fails to capture the acoustic diversity, the conventional fine-tuning approach may suffer from issues such as low generalizability and catastrophic forgetting. These challenges arise due to the high complexity and data sensitivity of the Transformer-based architecture [13], [14].

To mitigate these limitations, researchers have explored alternative strategies to alleviate the issue of knowledge forgetting during the process of transfer learning. For instance, one approach is to use L2 regularization to constrain the divergence between the model's weights and the pre-trained weights [15]. Furthermore, substituting some model parameters stochastically with their corresponding pre-trained parameters can also produce effects similar to regularization [16]. Another prevalent technique hinges on adapters, which can add additional intermediate layer parameters and fine-tuned while keeping all the pre-trained parameters frozen [17]–[19]. Nonetheless, the disadvantages of adapters remain distinctly noticeable. On the one hand, adding adapters for each task will lead to a sharp increase in model size. On the other hand, the adapter module itself requires a large amount of target language data for training, rendering the incorporation of intricate network configurations superfluous in scenarios of data paucity. Recent endeavors within the realm of Natural Language Processing have turned to sparse fine-tuning techniques as a surrogate to adapters, abstaining from the introduction of any additional structure [20], [21]. This not only curtails computational expenditure but also retains the performance supremacy of pre-trained models.

To better adapt pre-trained model to target low-resource Tibetan dialects, we proposed an effective fine-tuning method inspired by the Lottery Ticket Hypothesis (LTH) [22], which only fine tune the most task-related parameters within a subnetwork of the full model. The crux of the matter currently resides in discerning these precise parameters. In particular, there are two variations of this method: the first one discovers the subnetwork by measuring the differences between the upstream and downstream tasks, while the second one selects the most important parameters for the downstream task based on their gradient information. In comparison to the original idea, our approach diverges in two main facets. Firstly, LTH primarily focuses on the compression of large models, while

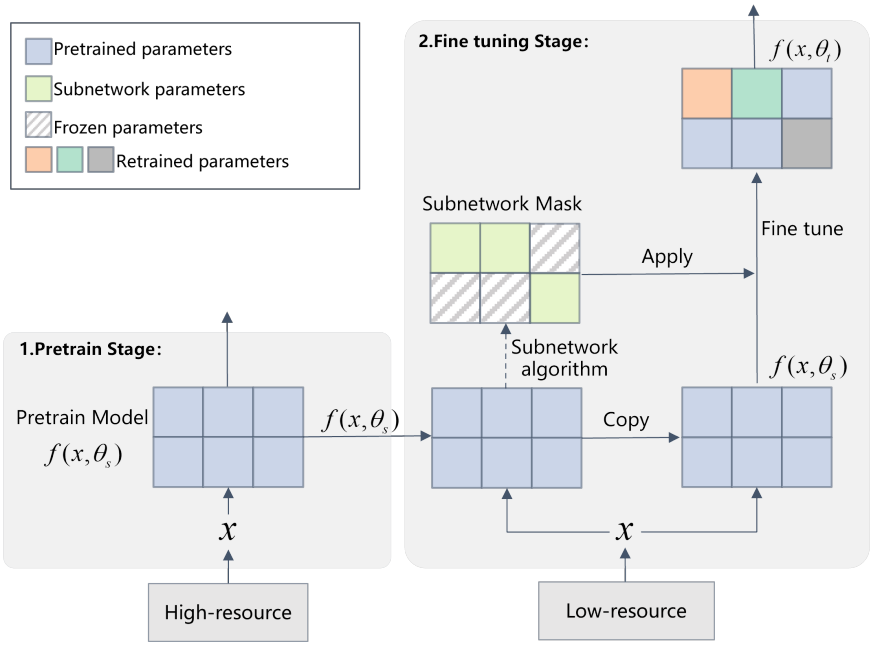

Fig. 1. Overview of the fine-tuning algorithm proposed in this work: Firstly, we train the model on high-resource corpus, and the pre-trained model is represented as function $f(x, \theta)$. Then, we obtain a mask to represent the corresponding language-specific subnetwork, where the green blocks symbolize trainable neurons, and the white blocks denote frozen ones. Finally, retrain the network using mask.

our proposed approach preserves all parameters of the network and avoids any pruning that may affect model accuracy. Furthermore, the parameters of the subnetwork extracted by the original LTH are determined by their significance in the pre-training task, which may not be applicable in transfer learning due to the potential mismatch between upstream and downstream tasks.

The rest of this paper is organized as follows. In Section 2, the proposed two fine-tuning strategies are introduced. Section 3 shows the experimental settings and evaluation results. The conclusions are in Section 4.

II. Proposed Methods

In this section, we mainly introduce our proposed cross-lingual transfer-learning methods based on language specific sparse fine-tuning, as shown in Figure 1.

A. Lottery Ticket Hypothesis

The Lottery Ticket Hypothesis (LTH) illustrates that a sparse subnetwork within a fully dense network is capable of being trained in isolation from initialization to achieve high accuracy, as shown in Figure 2.

Recent works have leveraged unstructured pruning technology to compress large pre-trained models based on the Lottery Ticket Hypothesis [22]. These methods only need to update a specific sparse sub-network and prune the remaining weights, achieving high sparsity with minimal performance degradation [23]–[25]. In different tasks or models, it is necessary to conduct multiple experimental iterations to identify an optimal "winning tickets" subnetwork. For instance, during the fine-tuning stage, a mask is typically applied, consisting of absolute values such as a set of ones and zeros. Usually, the subnetwork is identified by training a model on the source language and pruning the weights with the smallest magnitudes. This iterative pruning and re-training process may need to be executed numerous times to achieve an optimal compressed model.

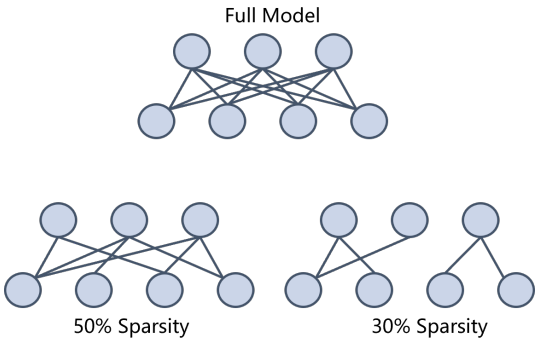

Fig. 2. Compared to a full model, a sparse subnetwork has only a few parameters trained.

B. Sparse Fine-tuning Method

This work investigates the generalization capability of sparse fine-tuning techniques in cross-dialect transfer learning, specifically in the domain of speech recognition for languages with limited resources. Consider a generic neural network parameterized by function $f(x, \theta)$ with initial parameters $\theta = \theta_0$ and input x. The output from the forward pass is $y = f(x, \theta)$. The parameters in full fine-tuning strategy would be updated

as follow:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \tag{1}$$

Where $\mathcal{L}(\theta_t)$ is the loss function. $\frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t}$ represents the gradients corresponding to the model parameters θ_t, and η is the learning rate. Our objective is to determine a subnetwork with a sparsity $S \in [0, 1]$. To achieve this, we consider a pruning mask consisting of a series of ones representing the subnetwork and replace the rest with zeros. This is illustrated as follows:

$$M_t^i = \begin{cases} 1 & \text{if } \theta^i \in P^t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

P^t represents a series of weight parameters considered to be activated at iteration t. i is the i-th weight, and M is the subnetwork mask used to update parameters at each backpropagation. Usually, P^t is obtained from the top-K proportion of weights θ_t based on their magnitude. This process can be described as follows:

$$P^t = \left\{ i \mid \theta_t^i \in \mathrm{TopK}(\theta_t) \right\} \tag{3}$$

Then we update the parameters with the following equation:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \odot M_t \tag{4}$$

C. Language Specific Transfer

The selection of subnetworks is often crucial for sparse fine-tuning methods. In this paper, we proposed two variants methods for obtaining a set of target language-specific parameters from the backbone network, as shown in Algorithm 1. One method is based on weight amplitude difference, and the other is based on the gradient of the neural parameter.

Algorithm 1 Framework of our fine-tuning method.

Require:

$\mathcal{L}()$: learning objective of model;
θ_s : the pretrain neural model weights;
S : Sparsity of subnetworks;
subnetwork($TopK(|\theta_s - \theta_t|), S$) : different strategies of subnetwork extraction amied at downstream language are implemented (if using fisher method, the first input parameter is $TopK(\mathbf{H})$;

Ensure:

Initialize model, Start fine-tuning stage:
while in epochs **do**
  if Calculate mask **then**
    M_t= subnetwork($TopK(|\theta_s - \theta_t|), S$).
  end if
  Update θ_t using Mask.
  $\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \odot M_t$
end while

Weight-Difference Fine-tuning(differ): First, the pre-trained model is thoroughly fine-tuned on the downstream language. θ_s on the source language and θ_t on the target language represent the neural network parameters before and after fine-tuning. The parameters are sorted by importance according to specific rules, which can be calculated by the greatest absolute difference $|\theta_t^i - \theta_s^i|$. i is the i-th weight. Those parameters with the largest difference in the above formula are trained in the next step on downstream language, and the remaining parameters will be frozen. We use a mask M to distinguish the two types of parameters. Precisely, the K in Top-K function corresponds to the sparsity setting S at the beginning.

Second, the model re-fine-tuning on the target language, and this time, the mask M given by the previous step is used. Parameter updating is similar to equation (4). Notice that the L1 regularization is applied in this work to encourage more robust parameter updating.

Gradient based Fine-tuning(fisher): Fisher Information [26] determines the update direction in natural gradient descent, which is a nonlinear function of the weights and data [27]. As described below, the Fisher is usually given by a hessian matrix:

$$\mathbf{H} = E_X[J^T J], J = \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t}, \tag{5}$$

where H is Hessian Matrix, and J denote the gradient matrix of the model. Given the task-specific training data D, we use the diagonal elements of the empirical Fisher Information Matrix (FIM) to estimate the task-related importance of the parameters , as described in [21]. We derive the Fisher information for the i-th parameter as shown in (6):

$$\mathbf{H} = \sum_{j=1}^{|D|} \left(\frac{\partial \log p(y_j \mid \mathbf{x}_j; \theta)}{\partial \theta^{(i)}} \right)^2 \tag{6}$$

We rank all parameters based on their Fisher Information, assuming that those with higher Fisher Information are more important for the target task. Then, we use the Top-K function to select the highly relevant parameters to form the subnetwork. During the back propagation process, only the parameters in the subnetwork are updated by gradients, while the other non-subnetwork parameters are frozen.

TABLE I
STATISTICS OF ANNOTATED LANGUAGE DATASETS.

Dialect	Size (Hours)			Speakers	Utterances
	Train	Dev	Test	Total	Total
Am-do	133.1	5.8	2.1	114	113643
U-Gtsang	4.6	2.2	2.3	8	7892
Khampa	5.1	2.5	2.4	8	7996

III. TASK DESCRIPTION AND BASELINE SYSTEMS

A. Dataset

We employ a comprehensive Tibetan dataset, including Amdo, U-Gtsang, and Khampa. The corpus is recorded in a serene environment, and sentences are employed in everyday discourse. The data acquisition process employed a monophonic

recording setup with a sampling rate of 16 kHz and 16-bit quantization. The speakers who participated in the recordings were youthful denizens hailing from Tibetan regions, aged between 18 and 23. The duration of each audio is generally within ten seconds. Subsequently, the datasets were segregated into 'train', 'dev' and 'test' splits for training and validation correspondingly for all datasets. The size details of the dataset are presented in Table I.

B. Baseline system description

We utilize Transformer[3] based ASR model implemented in SpeechBrain [28]. In order to better evaluate the performance of our proposed method on models with different numbers of parameters, we designed two Transformer models with different model size. For the transformer-small, the encoder has 12 layers and the decoder has 6 layers with width 256. For the transformer-large, the encoder has 16 layers and the decoder has 8 layers with width 512, and other parameters are exactly the same. Each layer includes multi-head attention and fully connected layer. We extract 80 dimensions fbank features (windows with 25ms size and 10ms shift). Except for the mentioned parameters, all other settings are exactly the same as in this work.

IV. Experimental results

TABLE II
WORD ERROR RATE (WER) FOR CROSS-DIALECT FINE-TUNING FROM AM-DO TO U-GTSANG AND KHAMPA.

Model	Params (M)	Method	Khampa		U-Gtsang		Avg
			Dev	Test	Dev	Test	
Transformer (Small)	27.3	vallina	14.08	14.22	23.60	24.46	19.09
		fisher	14.18	14.43	24.27	24.86	19.43
		diff	14.02	14.20	22.14	22.47	**18.20**
Transformer (Large)	87.5	vallina	12.62	13.30	23.45	24.91	18.57
		fisher	10.30	11.65	21.20	22.28	16.35
		diff	10.26	10.93	19.76	20.02	**15.24**

We compare our proposed two fine-tuning strategies: diff and fisher, to vanilla fine-tuning (update all parameters). As shown in Table II, our proposed methods outperform vanila fine-tuning, achieving a relative reduction in WER ranging from 5% - 18%. In particular, the 'diff' method exhibits superior efficacy on U-Gtsang and Khampa compared to 'fisher'. Upon comparing the experimental results of the 'Transformer (Small)' model with those of the 'Transformer (Large)' model, , it can be observed that as the model's parameter size increases (from 27.3M to 87.5M), the advantages of the sparse fine-tuning method become more pronounced.

TABLE III
COMPARED TO OTHER SPARSE FINE-TUNING METHODS

Method	Khampa		U-Gtsang		Avg
	Dev	Test	Dev	Test	
vallina	12.62	13.30	23.45	24.91	18.57
topk	13.37	13.92	23.38	24.26	18.73
random	20.17	21.45	31.34	37.64	27.65
adapter	18.04	19.11	28.66	30.58	24.09
diff	10.26	10.93	19.76	20.02	**15.24**

We further compare our proposed method with other sparse fine-tuning techniques, such as topk (selecting subnetwork from the pre-trained model), random (randomly selecting parameters as subnetwork), and adapter-based methods, employing Transformer (Large) as the foundational model. As exhibited in Table III, the experimental results indicate 'diff' method is superior to other sparse fine-tuning methods. Moreover, it suggests that obtaining task-specific sparse subnetworks from the target domain enables a more precise depiction of the target domain's characteristics, thereby avoiding the necessity for extensive data transformation or alignment between the source and target domain.

TABLE IV
THE EXPERIMENTAL RESULTS OF FUSING KNOWLEDGE REPRESENTED BY VARIOUS METHODOLOGIES.

Method	Combination Type	Khampa		U-Gtsang		Avg
		Dev	Test	Dev	Test	
vallina	-	12.62	13.30	23.45	24.91	18.57
fisher	-	10.30	11.65	21.20	22.28	16.35
diff	-	10.26	10.93	19.76	20.02	**15.24**
diff + fisher	union	10.41	11.09	20.05	20.53	15.52
	intersection	10.40	11.23	20.51	21.38	15.88
diff(30%) + topk(10%)	union	10.02	10.27	19.82	20.13	**15.06**
	intersection	10.73	11.49	21.14	21.68	16.26

To analyze whether there is a correlation between the knowledge encapsulated in the subnetworks extracted using different sparse fine-tuning methods, we conducted experiments utilizing distinct combination techniques (union and intersection) and employ Transformer (Large) as the foundational model. As shown in Table IV, compared to using the diff method alone, there is a slight degradation in performance when combining the two variants of the method proposed by us (diff and fisher). However, by conjoining the knowledge derived from the source domain ('topk') and the target domain ('diff'), the efficacy of the model can be further enhanced.

As Figure 3, to investigate how diverse levels of sparsity may impact the fine-tuning efficacy, we varied our target sparsity rate to span from 10% to 100%. Notice that the sparsity denotes the trainable weights compared to the entire model. We analyze the effect of sparsity on both low-resource datasets and found that lower degrees of sparsity result in greater enhancements compared to higher levels of sparsity. Meanwhile, the model achieved its highest accuracy in low-

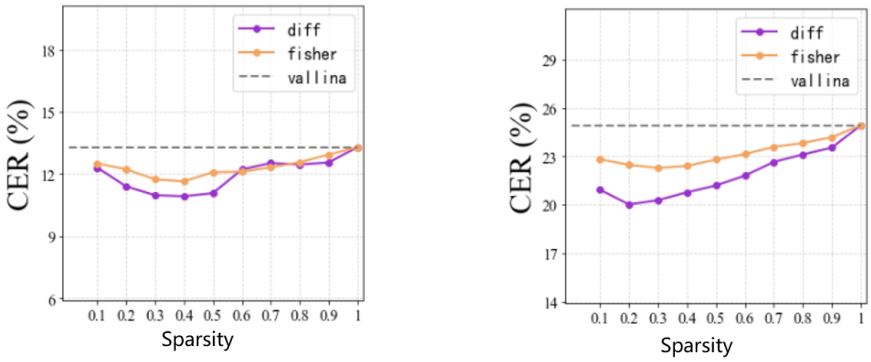

Fig. 3. Performance across two dialects of limited resources, contemplating the range of sparsity from 0.1 to 1. Sparsity denotes the remaining weights compared to the full model.

resource languages when the sparsity ranged from 10% to 30%.

V. Conclusions

Conventional methodologies of transfer learning often suffer from the issue of knowledge forgetting, leading to suboptimal performance in Tibetan dialect transfer learning. In this paper, we propose two sparse fine-tuning techniques based on the lottery ticket hypothesis, which evaluates the importance of each parameter for the target low-resource dialect. By refining the language-specific subnetwork and solidifying the remaining parameters, we mitigate knowledge loss, thereby achieving a more potent transfer of knowledge among Tibetan dialects.

Acknowledgment

Thanks to the National Natural Science Foundation of China (No. 62261045), National Key R&D Program of China (No. 2020YFC2004103) and National Natural Science Foundation of China (No. 61876131, U1936102).

References

[1] L. Pan, S. Li, L. Wang, and J. Dang, "Effective training end-to-end asr systems for low-resource lhasa dialect of tibetan language," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2019, pp. 1152–1156.

[2] S. Qin, L. Wang, S. Li, J. Dang, and L. Pan, "Improving low-resource tibetan end-to-end asr by multilingual and multilevel unit modeling," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–10, 2022.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "A neural network for large vocabulary conversational speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, 2016.

[5] X. Song, Z. Wu, Y.Huang, C. W, D. Su, and H. Meng, "Non-autoregressive transformer asr with ctc-enhanced decoder input," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5894–5898, 2021.

[6] A. Kuznetsova, A. Kumar, J. D. Fox, and F. M, "Curriculum optimization for low-resource speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8187–8191, 2022.

[7] S. Ueno, T. Moriya, M. Mimura, *et al.*, "Encoder transfer for attention-based acoustic-to-word speech recognition," *INTERSPEECH*, pp. 2424–2428, 2018.

[8] T. Moriya, R. Masumura, T. Asami, *et al.*, "Progressive neural network-based knowledge transfer in acoustic models," *Asia-Pacifific Signal and Information Processing Association Annual Summit and Confer ence (APSIPA ASC)*, pp. 998–1002, 2018.

[9] A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoyanov, "Emerging cross-lingual structure in pretrained language models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6022–6034.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[11] F. Zhuang, Z. Qi, K. Duan, *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[12] V. Joshi, R. Zhao, R. R. Mehta, K. Kumar, and J. Li, "Transfer learning approaches for streaming end-to-end speech recognition system," *arXiv preprint arXiv:2008.05086*, 2020.

[13] Z. Wang, Z. C. Lipton, and Y. Tsvetkov, "On negative interference in multilingual models: Findings and a meta-learning treatment," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4438–4450.

[14] A. Aghajanyan, A. Shrivastava, A. Gupta, N. Goyal, L. Zettlemoyer, and S. Gupta, "Better fine-tuning by reducing representational collapse," *arXiv preprint arXiv:2008.03156*, 2020.

[15] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu, "Recall and learn: Fine-tuning deep pretrained language models with less forgetting," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7870–7881.

[16] C. Lee, K. Cho, and W. Kang, "Mixout: Effective regularization to finetune large-scale pretrained language models," in *International Conference on Learning Representations*, 2020.

[17] A. Ansell, E. M. Ponti, J. Pfeiffer, *et al.*, "Mad-g: Multilingual adapter generation for efficient cross-lingual transfer," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4762–4781.

[18] A. Kannan, A. Datta, T. N. Sainath, *et al.*, "Large-scale multilingual speech recognition with a streaming end-to-end model," *arXiv preprint arXiv:1909.05330*, 2019.

[19] W. Hou, H. Zhu, Y. Wang, *et al.*, "Exploiting adapters for cross-lingual low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 317–329, 2021.

[20] A. Ansell, E. Ponti, A. Korhonen, and I. Vulić, "Composable sparse fine-tuning for cross-lingual transfer," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1778–1796.

[21] R.Xu, F.Luo, Z.Zhang, *et al.*, "Raise a child in large language model: Towards effective and generalizable fine-tuning," *Association for Computational Linguistics(ACL)*, pp. 9514–9528, 2021.

[22] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.

[23] Z. Wu, D. Zhao, Q. Liang, J. Yu, A. Gulati, and R. Pang, "Dynamic sparsity neural networks for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6014–6018.

[24] Z.Lin, L. Wu, M.Wang, and L. Li, "Learning language specific sub-network for multilingual machine translation," *Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pp. 293–305, 2021.

[25] D. Guo, A. Rush, and Y. Kim, "Parameter-efficient transfer learning with diff pruning," *Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pp. 4884–4896, 2021.

[26] J.Pennington and P.Worah, "The spectrum of the fisher information matrix of a single-hidden-layer neural network," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5410–5419, 2018.

[27] D. G. S. B. Hassibi and G. J. Wolff, "Optimal brain surgeon and general network pruning," *IEEE International Conference on Neural Networks*, pp. 293–299, 1993.

[28] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624. arXiv: 2106.04624.