INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

CLRL-Tuning: A Novel Continual Learning Approach for Automatic Speech Recognition

Zhihan Wang, Feng Hou, Ruili Wang**

School of Mathematical and Computational Sciences, Massey University, New Zealand
{z.wang4, f.Hou, ruili.wang}@massey.ac.nz

Abstract

In this paper, we propose a novel **C**ontinual **L**earning approach, which is **R**andomly **L**ayer-wise Tuning (CLRL-Tuning) of a pre-trained Automatic Speech Recognition (ASR) model. CLRL-Tuning tackles the randomness of subsequent datasets by updating the parameters of randomly selected encoder layers of the pre-trained model (such as wav2vec 2.0) for every training epoch. CLRL-Tuning is different from the previous approaches in that it neither uses previous datasets, nor expands/runs previous models. Furthermore, we perform experiments to evaluate our approach compared with four strong baselines, including *Knowledge Distillation* and *Gradient Episodic Memory*. Our approach achieves significant improvements over the baselines in average word error rate (WER) for the wav2vec 2.0 model. Additionally, we implement ablation studies for our approach by tuning one, three, six and full encoder layers of the model, and experimental results show only tuning one encoder layer of the model at each training epoch is the most effective way to mitigate catastrophic forgetting.

Index Terms: automatic speech recognition, continual learning, lifelong learning, pre-trained model, partial layers tuning, self-adaptation, fine-tuning.

1. Introduction

Continual/Incremental Learning is a process that trains a model sequentially on multiple incrementally collected datasets rather than trains it once on a whole combined dataset [1]. However, the differences in data distributions of datasets may induce *catastrophic forgetting* [2] of knowledge learned from previous datasets. For computer vision and automatic speech recognition (ASR), many methods have been developed to mitigate catastrophic forgetting. These methods can be categorized into three approaches: the parameter-based approach, data-based approach and regularization-based approach.

Parameter-based methods dynamically increase model capacity to maintain previous knowledge by either creating sub-networks for different tasks [3–5] or increasing the numbers of layers and/or neurons [6, 7]. Here are two examples of creating sub-networks for speech recognition: (i) Factorizing the likelihood model of an HMM-DNN-based ASR model into sub-models for different domains of datasets [8]; (ii) Inserting adapters into pre-trained ASR models for learning new languages [9]. Nevertheless, parameter-based methods will significantly increase model sizes and make it difficult to tune hyper-parameters of the models, especially when a new dataset size is unknown.

Data-based methods [10–13] replay data either from stored previous datasets or generated examples. However, the data used for the previous tasks may not be permanently stored where data storage is limited or where there are legal restrictions on data storage. (e.g., for privacy concerns [14]).

Regularization-based methods [15–19] incorporate additional regularization terms to consolidate the knowledge from a previously trained model. This approach has been applied to a multi-dialect acoustic model [20], which is continually trained with regularization-based methods (e.g, Learning without forgetting [15] and Elastic Weight Consolidation [17]). This approach has also been applied to the continual training of CTC-based ASR models [21] and pre-trained ASR models [22]. However, regularization-based methods require running previous models and can restrict the freedom of the model to learn new tasks.

2. Motivation

Recently, pre-trained language models with parameter-efficient prompt tuning has made significant progress, that learning lightweight task-specific embeddings while freezing parameters of a pre-trained language model [23–25]. This allows the model to quickly adapt to a new task with minimal additional training, while still leveraging the knowledge and capabilities of the pre-trained model. Intuitively, we hypotheses a large pre-trained ASR model contains prior knowledge learned from previous datasets and large-scale of unlabelled speech utterances. As a result, most of the pre-trained ASR model parameters trained on different speech datasets are likely to be reusable in solving the continual learning problem.

In this paper, we propose a **C**ontinual **L**earning approach by **R**andomly **L**ayer-wise Tuning (CLRL-Tuning) to utilize prior knowledge of a pre-trained ASR model, i.e., wav2vec 2.0. Following the principle of lightweight tuning, which is to freeze most of the pre-trained parameters and only tune a smaller set of parameters [23–25], CLRL-Tuning updates the parameters of a randomly selected encoder layer of the model for every epoch. In Figure 1, we demonstrate a theoretical analysis schematically to compare our proposed CLRL-Tuning approach with the previous approaches. Our proposed CLRL-Tuning approach is different from the previous approaches in that it neither uses previous datasets, nor expands/runs previous models. We perform experiments to evaluate our approach compared with four strong baselines, including *Knowledge Distillation* [15] and *Gradient Episodic Memory* [10]. Our approach achieves significant improvements over the baselines in average word error rate (WER) for the pre-trained wav2vec 2.0 [26] model. Additionally, we implement ablation studies for our approach by tuning one, three, six, and full encoder layers of the model. Experimental results show only tuning one encoder layer of the model is the most effective way to mitigate catastrophic forgetting.

10.21437/Interspeech.2023-503

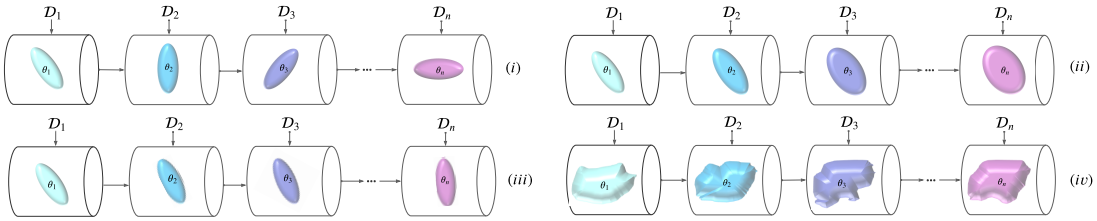

Figure 1: *Continual learning with different methods of mitigating catastrophic forgetting illustrated in schematic parameter spaces, where the nth model with parameters $\boldsymbol{\theta_n}$ is trained on the nth dataset $\mathcal{D}_n$ sequentially. (i) Direct fine-tuning method updates the model parameters to fit a new dataset, leading to the problem of catastrophic forgetting where the model parameters diverge from the region of the previous parameters; (ii) Parameter-based methods incrementally add new parameters to the model to adapt to new dataset, but this can make it difficult to tune hyper-parameters of the model; (iii) To mitigate the issue of model parameters deviating too far from the previous parameters, Data-based methods employ the replay of previous data samples to remind the model of its prior training on previous datasets, and Regularization-based methods add regularization terms to the loss function that penalizes changes to the parameters. (iv) CLRL-Tuning makes minor adjustments to the parameter space through layer-wise tuning of a large pre-trained model. The advantage of using a pre-trained model is that it has already learned a generic and large shape of the parameter space from larges-scale of unlabeled data. Consequently, most of the pre-trained model parameters can be shared across all datasets.*

3. Methodology

3.1. Continual Learning by Randomly Layer-wise Tuning of a Pre-trainded Model (CLRL-Tuning)

Suppose a pre-trained ASR model (e.g., wav2vec 2.0 [26], HuBERT [27]) is continually trained on T datasets $\mathcal{D}_1, \mathcal{D}_2 ... \mathcal{D}_T$ with different data distributions, the model parameters trained with these datasets is denoted as $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T$, respectively. We denote the current model as $\boldsymbol{\theta}_t$ trained on the tth dataset $\mathcal{D}_t$, and $t \in [1, \ldots, T]$. Let the transformer-based pre-trained ASR model comprises of M encoder layers (i.e., an encoder layer consists of multi-heads self-attention sub-layer and fully connected feed forward sub-layer). Before the start of every epoch to train the current dataset $\mathcal{D}_t$ on the current model $\boldsymbol{\theta_t}$, we randomly set a fixed number N encoder layers for parameters updating, and the rest of the $M - N$ encoder layers are set to be frozen. For each step of the training epoch, a batch of input speech data is passed to the current model $\boldsymbol{\theta}_t$ to generate a text prediction. We use the Connectionist Temporal Classification (CTC) [28] loss as the *objective* function to calculate the loss between the generated texts from the model $\boldsymbol{\theta}_t$ and the ground truth of the labeled text from the current dataset $\mathcal{D}_t$. This CTC loss ($\mathcal{L}_{CTC}$) is used to update the parameters of the current model $\boldsymbol{\theta}_t$.

The training framework for our proposed CLRL-Tuning approach is summarized in Algorithm 1. Our CLRL-Tuning approach trains a pre-trained ASR model on various datasets sequentially, and mitigates catastrophic forgetting by randomly selecting and tuning partial encoder layers of the model and freezing the rest of them during every training epoch. Randomly selecting and tuning partial encoder layers resembles the randomness of the future datasets distributions. Thus, our approach is more effective for mitigating catastrophic forgetting. Moreover, our approach is simple to implement and avoids the disadvantages of the parameter-based approach, the data-based approach, and the regularization-based approach, that require modifying neural network structure (or expanding neural network capacity), reusing of previous datasets (i.e., $\mathcal{D}_{t-1}$) and reusing of previous ASR models (i.e., $\boldsymbol{\theta}_{t-1}$).

Algorithm 1: CLRL-Tuning

Inputs: number of stages T, number of epochs for each stage K, number of training steps for each stage S, pre-trained model $\boldsymbol{\theta}$, number of total encoder layers M, number of encoder layers selected for tunning N, datasets $\mathcal{D} = [\mathcal{D}_1, \ldots, \mathcal{D}_T]$

$\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}$
for $t \leftarrow 1$ *to* T **do**
  $\boldsymbol{\theta}_t^0 \leftarrow \boldsymbol{\theta}_{t-1}$
  for $k \leftarrow 1$ *to* K **do**
    randomly select N indices of encoder layers
    load $\boldsymbol{\theta}_t^{k-1}$ and freeze the rest $M - N$ layers
    $\boldsymbol{\theta}_t^k \leftarrow$tune the model with S steps on $\mathcal{D}_t$
  end
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^K$
end
return $\boldsymbol{\theta}_T$

4. Experiments And Results

4.1. Datasets

We use four open source datasets, LibriSpeech-960h [29], TIMIT-5h [30], LJSpeech-26h [31], and VCTK-44h [32], denoted as $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ respectively, for our experiments. More details about the four datasets with the configuration of their training set, test set, and validation set are shown in Table 1.

Table 1: *Datasets for the continual learning experiments*

Data set	Training Set	Test Set	Validation Set
LibriSpeech $\mathcal{D}_1$	train-clean-100	test-clean	dev-clean
TIMIT $\mathcal{D}_2$	90% subset of TRAIN	10% subset of TRAIN	TEST
LJSpeech $\mathcal{D}_3$	80% subset of LJSpeech	10% subset of LJSpeech	10% subset of LJSpeech
VCTK $\mathcal{D}_4$	80% subset of VCTK	10% subset of VCTK	10% subset of VCTK

Table 2: *Experiment test sets results (WER %) on various methods.*

Method \ Datasets	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	Avg
Individual	5.8	20.1	3.1	3.2	8.1
Fine-tuning (i.e., CLRL-Tuning, $N = 12$)					
$\boldsymbol{\theta}_2$	6.8	12.0			9.4
$\boldsymbol{\theta}_3$	11.3	23.6	2.5		12.5
$\boldsymbol{\theta}_4$	16.4	27.3	11.4	2.2	14.3
Knowledge Distillation [15]					
$\boldsymbol{\theta}_2$	6.8	11.7			9.3
$\boldsymbol{\theta}_3$	8.9	16.7	2.5		9.4
$\boldsymbol{\theta}_4$	13.4	21.8	9.5	3.0	11.9
Gradient Episodic Memory [10]					
$\boldsymbol{\theta}_2$	6.7	11.8			9.3
$\boldsymbol{\theta}_3$	9.0	16.3	2.3		9.2
$\boldsymbol{\theta}_4$	13.2	20.2	8.5	3.2	11.3
Our Approach (CLRL-Tuning, $N = 1$)					
$\boldsymbol{\theta}_2$	6.9	11.7			9.3
$\boldsymbol{\theta}_3$	7.8	15.6	2.4		8.6
$\boldsymbol{\theta}_4$	9.8	16.7	6.4	2.6	8.9

Table 3: *Test sets results (WER %) for the ablation study with N randomly selected encoder layers for CLRL-tuning.*

Method \ Datasets	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	Avg
CLRL-Tuning ($N = 12$) (i.e.,Fine-tuning)					
$\boldsymbol{\theta}_2$	6.8	12.0			9.4
$\boldsymbol{\theta}_3$	11.3	23.6	2.5		12.5
$\boldsymbol{\theta}_4$	16.4	27.3	11.4	2.2	14.3
CLRL-Tuning ($N = 6$)					
$\boldsymbol{\theta}_2$	6.9	11.9			9.4
$\boldsymbol{\theta}_3$	9.2	21.9	2.3		11.1
$\boldsymbol{\theta}_4$	13.2	25.2	10.8	2.2	12.9
CLRL-Tuning ($N = 3$)					
$\boldsymbol{\theta}_2$	6.8	11.8			9.3
$\boldsymbol{\theta}_3$	8.8	19.3	2.2		10.1
$\boldsymbol{\theta}_4$	12.4	21.9	8.5	2.4	11.3
CLRL-Tuning ($N = 1$)					
$\boldsymbol{\theta}_2$	6.9	11.7			9.3
$\boldsymbol{\theta}_3$	7.8	15.6	2.4		8.6
$\boldsymbol{\theta}_4$	9.8	16.7	6.4	2.6	8.9

4.2. Experimental Settings

To evaluate our proposed CLRL-Tuning approach, we adopt a commonly used transformer-based ASR model: **wav2vec 2.0** [26] pre-trained on the full LibriSpeech-960h [29] dataset. wav2vec 2.0 comprises of *twelve* encoder layers with *twelve* attention heads for each encoder layer. We extract raw waveform with 16000hz for the proposed datasets $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$. We use Adam [33] to optimise the learning rate when training the ASR model for the experiments. When decoding the prediction of the ASR model, we use the beam search with a beam size of 10.

The wav2vec 2.0 model is continually trained using three schemes: (i) **Individual** training : wav2vec 2.0 is trained on the speech datasets $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ individually to observe the performances of the single models; (ii) **Fine-tuning**: wav2vec 2.0 is continually trained on the datasets sequentially to show catastrophic forgetting; (iii) **Mitigating Forgetting**: wav2vec 2.0 is continually trained on the datasets sequentially with the *Knowledge Distillation* [15] method, the *Gradient Episodic Memory* [10] method (two strong baselines used in the lifelong learning of end-to-end ASR studies [21]), and our proposed approach CLRL-Tuning to compare with their performances.

We continually train the pre-trained wav2vec 2.0 model with four training stages in the order of $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$, and the model is trained for 50 epochs for each stage. We use the current model trained on $\boldsymbol{\theta}_t$ to evaluate them on the test sets of the datasets $\mathcal{D}_t, \mathcal{D}_{t-1}, \ldots, \mathcal{D}_1$. For example, we evaluate the current model $\boldsymbol{\theta}_3$ on the current training dataset $\mathcal{D}_3$, and the previous datasets $\mathcal{D}_2$ and $\mathcal{D}_1$. To show the catastrophic forgetting of sequential **fine-tuning** and compare the effectiveness of **mitigating forgetting** methods, we need to start from a shared fine-tuned model on $\mathcal{D}_1$. Thus, the first model $\boldsymbol{\theta}_1$ in Algorithm 1 is fine-tuned directly on the dataset $\mathcal{D}_1$, and we apply mitigating forgetting methods to the subsequent datasets.

4.2.1. Knowledge Distillation (KD)

For the Knowledge Distillation [15] method, a regularization-based approach, we calculate a KL divergence [34] between the output distributions of the previous model $\boldsymbol{\theta}_{t-1}$ and the current model $\boldsymbol{\theta}_t$ as a regularization term to constrain parameters shifting from the previous model $\boldsymbol{\theta}_{t-1}$. The regularization term is summed with the CTC loss as the *objective* function for the training of the current model θ_t.

4.2.2. Gradient Episodic Memory (GEM)

The *Gradient Episodic Memory method* [10] is a data-based approach. To update the gradients of the current model $\boldsymbol{\theta}_t$, we use stored samples with a length of 30 minutes from the previous dataset $\mathcal{D}_{t-1}$, and select samples close to the median lengths of utterance of the dataset $\mathcal{D}_{t-1}$, that is an optimal set up for the GEM method proposed in the study of [21].

4.2.3. Our Approach (CLRL-Tuning)

We apply our CLRL-Tuning approach with four settings: tuning one ($N = 1$), three ($N = 3$), six ($N = 6$) and full ($N = 12$) encoder layers and freeze the rest of them to continually train wav2vec 2.0 on the datasets sequentially. As an ablation study presented in Section 4.4, we compare the effect of the number of encoder layers being tuned for the model.

4.3. Experimental Results

The results of **Individual** training, **Fine-tune** training, and **Mitigating forgetting** are summarized in Table 2. For a fair and reasonable comparison, we use the average word error rate (Avg WER) [8, 21] as the evaluation metric to compare our proposed *CLRL-Tuning* approach with the baselines. We present the optimal result for our approach, i.e., randomly tuning one encoder layer ($N = 1$) of wav2vec 2.0 during each epoch. Comparing the results of individual training and fine-tune training, we can see that continual training with direct fine-tuning (i.e., tuning the full ($N = 12$) encoder layers of the wav2vec 2.0 model) causes catastrophic forgetting. Using our proposed CLRL-Tuning approach can retain knowledge more effectively than using the KD method and the GEM method.

We then use the validation sets of $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ to plot the continual learning curves of our proposed CLRL-Tuning approach and the baselines in Figure 2. The curves in four different colours represent the four methods used in our experi-

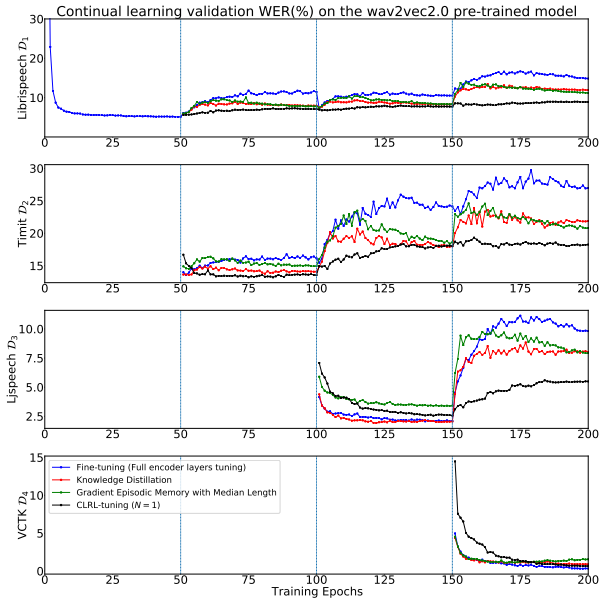

Figure 2: *Validation sets learning curves for the wav2vec 2.0 pre-trained model with the methods of Fine-tuning, KD, GEM and our CLRL-Tuning* $(N = 1)$.

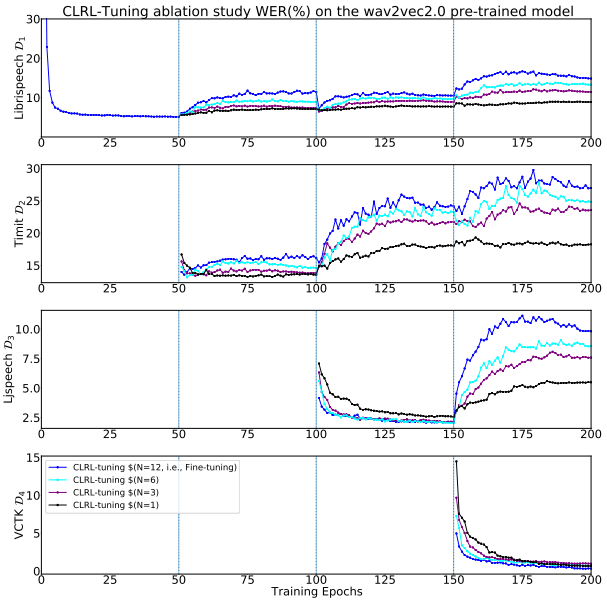

Figure 3: *Validation sets learning curves for the wav2vec 2.0 pre-trained model with CLRL-tuning of various number of encoder layers* $(N = 1, N = 3, N = 6, N = 12)$.

ments, and they are plotted in four windows to denote the four training stages on Librispeech $\mathcal{D}_1$, Timit $\mathcal{D}_2$, Ljspeech $\mathcal{D}_3$, and VCTK $\mathcal{D}_4$. At each training stage, the learning curves of the Fine-tuning, the KD and the GEM methods (plotted in blue, red and green in Figure 2, respectively) jump up when the model is trained on the new dataset, while the learning curve of our CLRL-Tuning (i.e., the black curve and $N = 1$) grows remarkably more gentle than theirs. The learning curves plotted in Figure 2 also indicate that our method is more effective for mitigating catastrophic forgetting than the Fine-tuning, the KD and the GEM methods, since the Fine-tuning method results in significant deviation of the model parameters from the region of the previous parameters (as demonstrated in Figure 1 (i)). Conversely, the KD and the GEM methods can mitigate this deviation and alleviate catastrophic forgetting. (as shown in Figure 1 (ii)).

4.4. Ablation Study for CLRL-Tuning

In this section, we inspect the effect of the number of encoder layers being tuned for the wav2vec 2.0 model on the datasets $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ sequentially. We attempt to tune one $(N = 1)$, three $(N = 3)$, six $(N = 6)$ and full $(N = 12)$ encoder layers (i.e., full encoder layers tuning is equivalent to the Fine-tuning approach) of the model. Table 3 shows the less encoder layers being tuned, the more effective for the model to alleviate catastrophic forgetting. Figure 3 shows the learning curves on the validation sets of the datasets for the four training stages. We can see that tuning less encoder layers of the model makes the learning curves grow more gentle at each training stage. Therefore, tuning one encoder layer $(N = 1)$ of the model can most effectively mitigate catastrophic forgetting.

The results for the ablation study show that a pre-trained ASR model with *good prior knowledge* can efficiently learn similar speeches patterns with only minor adjustments to its parameters, since most parameters of the model are shareable across all the datasets (as depicted in Figure 1 (iv)). Tuning one encoder layer $(N = 1)$ of the model results in minimal changes to the parameter, while still leveraging the shared parameters for good generalization across all datasets. However, tuning a larger number of encoder layers $(N = 3, N = 6, N = 12)$ results in greater deviation of the model parameters from the previous parameters, which causes catastrophic forgetting.

5. Conclusions and Future Work

In this paper, we propose a continual learning approach that tunes partial encoder layers of a pre-trained ASR model for effectively retaining the knowledge learned from previous datasets. Experimental results show our approach can utilize the prior knowledge of a pre-trained ASR model (i.e., trained on large amount of unlabelled data) to alleviate the catastrophic forgetting. Notably, our approach has neither memory restriction, nor extra parameter tuning, nor privacy considerations compared with previous approaches.

Future work will test CLRL-Tuning on larger wav2vec 2.0 models with more extensive unlabeled data, such as the wav2vec 2.0 model pre-trained on the LirbiVox 60K dataset [35], to examine whether larger networks and more unlabeled data can further enhance the performance of our approach. Additionally, we propose to investigate tuning more detailed sub-layers of the pre-trained ASR models, such as recent advances in speaker adaptation for the conformer transducer model [36].

6. Acknowledgements

This work is supported by the 2020 Catalyst: Strategic NZ-Singapore Data Science Research Programme Fund, MBIE, New Zealand. Ruili Wang* and Feng Hou* are the corresponding authors.

7. References

[1] G. I. Parisia, R. Kemkerb, J. L.Part, C. Kanan, and S. Wermtera, "Continual lifelong learning with neural networks: A review," in *Neural Networks*, vol. 113, 2019, pp. 54–71.

[2] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, 1989, pp. 109–165.

[3] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7765–7773.

[4] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *Proc. International Conference on Machine Learning (ICML)*, 2018, pp. 4548–4557.

[5] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "Pathnet: Evolution channels gradient descent in super neural networks," in *arXiv preprint arxiv 1701.08734*, 2017.

[6] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7120–7129.

[7] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," in *arXiv preprint arxiv 1606.04671*, 2016.

[8] S. Sadhu and H. Hermansky, "Continual learning in automatic speech recognition," in *Proc. Interspeech*, 2020, pp. 1246–1250.

[9] S. Kessler, B. Thomas, and S. Karout, "An Adapter Based Pre-Training for Efficient and Scalable Self-Supervised Speech Representation Learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3179–3183.

[10] D. Lopez-Paz and M. A. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 6467–6476.

[11] S. Rebu, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2001–2010.

[12] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 532–547.

[13] R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, and L. Page-Caccia, "Online continual learning with maximal interfered retrieval," in *Advances in Neural Information Processing Systems (NIPS)*, 2019, pp. 11 849–11 860.

[14] P. McClure, C. Y. Zheng, J. R. Kaczmarzyk, J. A. Lee, S. S. Ghosh, D. Nielson, P. Bandettini, and F. Pereira, "Distributed weight consolidation: a brain segmentation case study," in *Advances in Neural Information Processing Systems (NIPS)*, 2018, pp. 4097–4107.

[15] Z. Li and D. Hoiem, "Learning without forgetting," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 614–629.

[16] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 831–839.

[17] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, and A. Grabska-Barwinska, "Overcoming catastrophic forgetting in neural networks," in *Proc. National Academy of Sciences*, vol. 114, 2016.

[18] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.

[19] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," in *arXiv preprint arXiv:1710.10628*, 2017.

[20] B. Houston and K. Kirchhoff, "Continual learning for multidialect acoustic models," in *Proc. Interspeech*, 2020, pp. 576–580.

[21] H. Chang, H. Lee, and L. Lee, "Towards lifelong learning of end-to-end asr," in *Proc. Interspeech*, 2021, pp. 2551–2555.

[22] J.-H. Lee, C.-W. Lee, J.-S. Choi, J.-H. Chang, W. K. Seong, and J. Lee, "CTRL: Continual Representation Learning to Transfer Information of Pre-trained for WAV2VEC 2.0," in *Proc. Interspeech*, 2022, pp. 3398–3402.

[23] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2021.

[24] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 3045–3059.

[25] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2021, pp. 61–68.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, 2020, pp. 12 449–12 460.

[27] W. Hsu, Y.Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, p. 6533–6537.

[28] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[30] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1992.

[31] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[32] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," https://doi.org/10.7488/ds/2645, 2017.

[33] P. K.Diederik and B. Jimmy, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[34] S. Kullback and R. A. Leibler., "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22 (1), pp. 79–86, 1951.

[35] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, and P. E. M. et al., "Libri-light: A benchmark for asr with limited or no supervision," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673.

[36] Y. Huang, G. Ye, J. Li, and Y. Gong, "Rapid Speaker Adaptation for Conformer Transducer: Attention and Bias Are All You Need," in *Proc. Interspeech*, 2021, pp. 1309–1313.