Sequential Editing for Lifelong Training of Speech Recognition Models

Devang Kulshreshtha, Saket Dingliwal, Brady Houston, Nikolaos Pappas, Srikanth Ronanki

AWS AI Labs, USA

{kulshrde, skdin, hstbrady, nppappa, ronanks}@amazon.com

Abstract

Automatic Speech Recognition (ASR) traditionally assumes known domains, but adding data from a new domain raises concerns about computational inefficiencies linked to retraining models on both existing and new domains. Fine-tuning solely on new domain risks Catastrophic Forgetting (CF). To address this, Lifelong Learning (LLL) algorithms have been proposed for ASR. Prior research has explored techniques such as Elastic Weight Consolidation, Knowledge Distillation, and Replay, all of which necessitate either additional parameters or access to prior domain data. We propose *Sequential Model Editing* as a novel method to continually learn new domains in ASR systems. Different than previous methods, our approach does not necessitate access to prior datasets or the introduction of extra parameters. Our study demonstrates up to 15% Word Error Rate Reduction (WERR) over fine-tuning baseline, and superior efficiency over other LLL techniques on CommonVoice English multi-accent dataset.

Index Terms: speech recognition, lifelong learning, model editing, multi-accent ASR

1. Introduction

Recently, the field of speech recognition (and machine learning/AI in general) has trended toward large foundational models trained on very large, diverse datasets covering many domains. Despite this trend, it is still common in industrial settings that after a foundational/base model has been initially trained, to gradually train it on new domains or categories. In multidialect ASR, for example, these two situations would be improving the performance of the model on a single dialect (or subset of dialects) and adding a previously-unseen dialect/accent to the model. Both of these goals would typically be achieved by fine-tuning the base/foundational model, possibly with the addition/substitution of some model parameters, on new training data.

Fine-tuning often comes with a cost, which is that the model's performance on the domains seen during training can degrade due to catastrophic forgetting [1]. Returning to the example of adding a new dialect/accent to a multilingual ASR model, this degradation can be troublesome if the model is expected to perform well on the new dialect *and* on all previously-seen dialects. A common mitigation approach to the catastrophic forgetting observed when fine-tuning on a new dialect is to re-train the model with both the new data and the already-seen data. This, of course, can be extremely costly, especially in the era of very large models and training datasets. In addition, the previously-seen data is not always continually-available in practical settings. Lifelong learning (or continual learning) approaches have been shown to alleviate this catastrophic forgetting effect in a wide variety of machine learning models and tasks, including ASR.

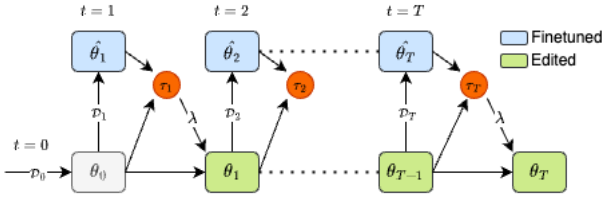

Figure 1: ***Sequential Model Editing for Lifelong ASR**: At each time step t, the current model θ_{t-1} is fine-tuned on data $\mathcal{D}_t$ to obtain $\hat{\theta}_t$. Then task vector τ_t is computed. Finally, new model is obtained my merging τ_t with θ_{t-1} as : $\theta_t = \theta_{t-1} + \lambda \cdot (\tau_t)$.*

The most simple lifelong learning approach is Experience Replay; when a new domain is added via fine-tuning, a subset of the original training is also included [2] (or possibly all of the training data [3]). However, this comes with the obvious downside of being inefficient, as each new domain being added requires more and more replay data. Adding multi-task training objectives to encourage the model to retain information on previous domains, as in Elastic Weight Consolidation [1] and Knowledge Distillation can also be effective in ASR [4] and other tasks [5], but they may show limited ability to scale to many new domains. Above approaches also require either additional parameters or previous domaind data to mitigate Catastrophic Forgetting (CF). More recently, several studies [6, 7, 8] investigated the manipulation and/or combination of fine-tuned model parameters with base model parameters for creating multi-task models, avoiding the need to re-use data or implement more complicated multi-task training approaches. However, these methods tend to degrade in quality when applied to a large number of tasks [8].

In this paper, we explore model editing approaches for sequential training of ASR models that overcome limitations in prior work in terms of requiring additional parameters or access to prior domain data. Specifically, we investigate sequential editing of the original model continually trained on all the previous data sets. Here, we focus on two representative methods, namely Task Arithmetic [7], which uses basic arithmetic operations to combine checkpoints from different tasks, and TIES-Merging [8] which addresses issues that arise when task vectors (i.e. models) are combined, such as sign conflicts and small weights. These approaches have been explored for out-of-domain generalization [7, 9], multi-task learning [7, 8], and transfer learning [10] but not for lifelong learning to the best of our knowledge. At every sequential step, we assume access only on the new data source which is a challenging setting for

Algorithm 1 Sequential Model Editing for Lifelong ASR

Require: Data sources $\mathcal{D} = [\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_T]$, λ, and k.
Ensure: Lifelong ASR model θ^*

Init: $\theta_0 \leftarrow$ Train on $\mathcal{D}_0$
for $t \leftarrow 1$ to T **do**
  $\hat{\theta}_t \leftarrow$ Fine-tune θ_{t-1} on $\mathcal{D}_t$
  $\tau_t \leftarrow \hat{\theta}_t - \theta_{t-1}$ ▷ Task Arithmetic
  if TIES-Merging **then**
    $\tau_t \leftarrow$ ties_merging_procedure(τ_t, k)
  end if
  $\theta_t \leftarrow \theta_{t-1} + \lambda \cdot (\tau_t)$
end for
$\theta^* \leftarrow \theta_T$

existing continual learning methods [11]. For evaluation, we focus on effectively learning multiple English dialects in an incremental fashion while preserving the performance on previously seen dialects on CommonVoice English data set. The main contributions can be summarized as follows:

- We propose a novel Sequential Model Editing approach that can be used for lifelong training of ASR models without relying on prior domain datasets or additional training and/or parameters.
- Our approach results in 15% WERR over the fine-tuning baseline on CommonVoice English multi-accent dataset, compared to 6% WERR achieved by previously proposed baselines.

2. Lifelong Learning for ASR

ASR systems often consist of a Conformer-based CTC model [12, 13] that takes in audio sequence and outputs a text sequence. These models are often trained with paired audio and text data.

Let θ_0 be the parameters of such a model that is trained on a large set of such audio-text pairs represented by $\mathcal{D}_0$. In many practical scenarios, the abilities of the model are expanded by targeting new domains/accents/languages. More formally, let $\mathcal{D}_1, \ldots, \mathcal{D}_T$ be a sequence of T data sources that are incrementally used to update the model with each dataset targeting a particular domain. As the capabilities of the ASR model expand to these new domains sequentially, it is desirable to retain the performance on the older domains. Also, with the recent use of massive volumes of datasets for training, it often becomes prohibitively difficult to store and maintain all the data sources. Similarly, for some publicly available models, training datasets are not released or are behind pay-walls, thereby making it challenging to adapt the model to new domains without catastrophic forgetting.

To address these practical challenges, we define the goal for Lifelong Learning (LLL) for ASR as learning the optimal model parameters θ^* that performs well on all the data sources $\mathcal{D}_0 \ldots \mathcal{D}_T$, where the data sources are obtained sequentially and at any time step t, only the data source $\mathcal{D}_t$ is accessible and no past or future data sources. This constraint makes the existing trivial multi-task solutions unusable, which assume access to all domains data simultaneously to train the model. Therefore, we propose a novel sequential model-editing based approach as summarized in the Algorithm 1 and Figure 1.

3. Sequential Model Editing

Model editing refers to the paradigm of adding new functionality and behaviors to pre-trained neural models by manipulating the parameters or outputs, without the need of expensive retraining. In particular, [7] defines editing neural networks based on task vectors, which encode the information necessary to do well on a given task. They obtain such vectors by taking the weights of a model fine-tuned on a task and subtracting the corresponding pre-trained weights. They showcase that performing simple arithmetic operations on these task vectors can adapt a model to a new task or negate an undesirable behavior. In this work, we propose a novel Sequential Model Editing approach that leverages task vectors to expand the abilities of our ASR model without the need to access old data or to introduce any training loss functions and/or additional model parameters.

At any given stage t of the update of the ASR model, we have access to the model parameters θ_{t-1} and the new domain/accent with data source $\mathcal{D}_t$. The objective is to learn new set of model parameters θ_t that performs well on the new domain/accent while maintaining its original capabilities. Therefore, the problem can be simplified to learning a task vector τ_t for the new domain/accent and then leveraging model editing to update the ASR model. The task vector τ_t will represent information specific to new domain that was missing in previous model checkpoint θ_{t-1}. First, we fine-tune the model on the new data source to arrive at an intermediate model checkpoint $\hat{\theta}_t$. Next, we explore two different ways of creating task vectors from this checkpoint. These two different versions of Sequential Model Editing are defined as follows:

(1) **Task Arithmetic [7]**: In this version, the task vector is defined by simply taking the element-wise difference between $\hat{\theta}_t$ and θ_{t-1}, i.e., $\tau_t = \hat{\theta}_t - \theta_{t-1}$.

(2) **TIES-Merging [8]**: Since the number of parameters in a model can be substantially large, the dimension of the task vector in the previous version will be equivalently large. Many of the values in this vector will be of low magnitude. Therefore, redundant parameters from τ_t are removed in this version. Specifically, the top-k% values are retained based on their magnitude, while the bottom (100-k)% are set to 0. Although the TIES-Merging procedure [8] involves more complex operations to create the final aggregate task vector when multiple tasks are involved, we omit those details in this work as our sequential model editing procedure involves only a single task at a time.

Finally, the task vector created by either of the two version is added back to the model parameters with help of a scaling factor λ to create the final model checkpoint θ_t:

$$\theta_t = \theta_{t-1} + \lambda \cdot (\tau_t)$$

Note that $\lambda = 1$ corresponds to the trivial baseline of fine-tuning the model which typically results in Catastrophic Forgetting (CF). On the other hand, $\lambda = 0$ corresponds to no update to the model. The choice of λ in Model Editing is determined using held-out validation sets and can then be used to create a balance between the two extremes.

4. Experiments

4.1. Data

We employ the CommonVoice English ASR data [14] and partition it based on accents. Our lifelong learning experiments involves incrementally improving the performance of the ASR model on six distinct accents: United States (US), England

Notation	Accent	Country	Train	Dev	Test
$\mathcal{D}_0$	US	United States	470	1.4	1.6
$\mathcal{D}_1$	ENG	England	152	1.2	1.2
$\mathcal{D}_2$	AUS	Australia	78	1	1.4
$\mathcal{D}_3$	IND	India	104	1.3	1.6
$\mathcal{D}_4$	SCO	Scotland	17	1	1.3
$\mathcal{D}_5$	IRE	Ireland	10	1	1.4

Table 1: *CV English dataset duration (hrs) per accent.*

(ENG), Australia (AUS), India (IND), Scotland (SCO), and Ireland (IRE). The initial model checkpoint θ_0 is obtained by training only on the US accented data, denoted as $\mathcal{D}_0$. Subsequently, we expand the capabilities of the ASR model to perform well on different accents incrementally and in the following sequence of stages: US→ENG→AUS→IND→SCO→IRE. This sequence is chosen as initial accent (US) has the largest training data, creating the strongest base ASR model. Subsequent accents are sequenced randomly to simulate real-world scenarios. The specifics of our datasets per accent are presented in Table 1.

4.2. Model Architecture

We use a 12-layer CTC Conformer model, incorporating 8 self-attention heads, a 1024-dimensional feedforward layer, and an input/output size of 80, following the approach [12]. The models are designed to directly predict subword targets, derived from a sentence-piece model trained on initial US dialect $\mathcal{D}_0$, with a total vocabulary size of 512. The initial training on D_0 data source spans 60 epochs with a learning rate of 5e-3. Subsequently, for every addition of new data for an accent, the models undergo an additional 10 epochs of training with a reduced learning rate of 5e-4. To enhance ASR inference, a 4-gram language model is trained on combined data from all accents, and it is employed during beam search. We use the ESPnet library [15] for ASR and KenLM [16] for LM training. All models are updated with the Adam optimizer with a weight decay of 0.1.

Model Editing: We assign $\lambda = 0.4$ for Task Arithmetic, and for TIES-Merging, we set $\lambda = 0.6$ and $k = 0.5$ consistently across all time steps. These specific values are determined through evaluation on the development set at stage $t = 1$, and we maintain them unchanged for future stages. Although we did not explore varying λ or k for each stage here, such an exploration remains a potential avenue for future research.

4.3. Baselines

- *Fine-tuning*: This involves fine-tuning previous checkpoint on new accent data, and is expected to be highly susceptible to catastrophic forgetting (CF).
- *Randomly Layer-wise (CLRL) Tuning [17]*: This approach suggests randomly fine-tuning only $M < N$ out of N encoder layers on new data while keeping the remaining $N-M$ layers frozen to mitigate CF. We set $N = 1$ as it yields optimal results based on the referenced paper.
- *Update Only Encoders (UOE) [18]*: This method involves updating only linear layers of Conformer encoders to prevent CF during incremental domain adaptation. Here, linear layers refer specifically to the weight matrices of the Feedforward Network and attention module within a Conformer block.

Other conventional LLL methods like Experience Replay [2] require access to the old data at every stage and therefore are not directly comparable to our methods (as our methods explicitly aim to relax this requirement). However, we do benchmark even stronger upper bounds (oracle): (1)*Multi-Task* model, which trains the ASR model on the pooled data from all accents $\mathcal{D}_{combined} = \cup_{t=0}^{T}\mathcal{D}_t$. This helps to better understand the gap between our methods and the best method when all the datasets are available at every stage. (2) *Sep. Model*, which trains separate ASR models for each accent dataset independently and hence uses more parameters than our methods.

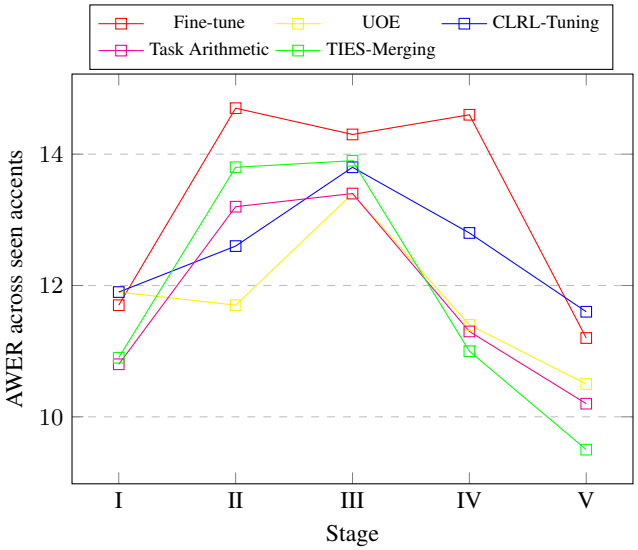

Figure 2: *Evolution of WER on seen accents for various approaches as new accents are added incrementally.*

4.4. Metrics

We report the WER per task, average WER across seen accents (AWER), and WER reduction (WERR %) across seen accents compared to the fine-tuning baseline. In this context, "seen" accents refer to those accents for which the corresponding data source has been used in any stage of training. For instance, at time step $t = 2$, the AWER is computed as the average of the baseline (US) and the next two accents (ENG, AUS) WER on the test sets $\mathcal{D}_0, \mathcal{D}_1, \mathcal{D}_2$, respectively.

5. Results

5.1. Catastrophic Forgetting - Model Editing vs Baselines

Table 2 presents the results of sequential lifelong learning experiments post the last step ($t = 5$), encompassing exposure to all six English tasks (accents).

The average WER (AWER) across all dialects of the oracle multi-task model achieved through training the ASR model on all data is 9.4, while sequential conventional fine-tuning results in a WER of 11.2. This highlights and underscores the existence of the catastrophic forgetting problem. The recently introduced Update Only Encoders (UOE) method [18] exhibits a 6.3% WERR improvement over the fine-tuning baseline. However, the CLRL-Tuning method [17] tends to perform below this baseline for most accents. Both these approaches however demonstrate efficacy in scenarios with fewer tasks (refer to section 5.2) but experiences degradation with the addition of more tasks.

Notably, both of our sequential editing methods show improvement over the baseline. Task Arithmetic and TIES-

Method		US	ENG	AUS	IND	SCO	IRE	AWER	WERR (%)
Baselines	Fine-tune	13.2	11.5	8.9	16.9	9	7.8	11.2	-
	UOE	12.3	10.9	8.4	15.4	**8.2**	7.5	10.5	6.3
	CLRL-Tuning	12.9	12	9.7	18	9.2	8	11.6	-3.5
Model Editing	Task Arithmetic	12.1	9.8	9	14.8	9.1	6.4	10.2	9.1
	TIES-Merging	**11.3**	**8.8**	**8.2**	**14.2**	8.8	**5.9**	**9.5**	**15**
Oracle	Sep. Model	12.9	9.8	6.3	12.1	7.7	7.1	9.3	16.8
	Multi-task	13	9.6	6.2	13.4	7.3	7.2	9.4	15.7

Table 2: *WER (↓) on the CV English testset after learning the six tasks (i.e. accents) in sequence.*

Model/T	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Intermediate ($\hat{\theta}_t$)	11.7	14.2	14.1	12.8	10.8
Edited (θ_t)	10.9	13.8	13.9	11	9.5
WERR (%)	3.4	2.8	1.5	14	12

Table 3: *AWER at every timestep for intermeditate fine-tuned vs edited checkpoint.*

Merging yield 9.1% and 15.0% WERR, respectively, almost reaching the performance of the oracle upper bound methods. The superiority of TIES-Merging, with lower overall WER than Task Arithmetic, underscores the importance of employing additional steps during merging of task vectors as proposed in [8].

5.2. Stage-wise Analysis

We evaluate the performance of various approaches at each time step of task addition, ranging from $t = 1$ (introducing $\mathcal{D}_1$) to $t = 5$ (introducing $\mathcal{D}_5$). The AWER at each time step is computed, and Figure 2 visually represents the results.

Our observations indicate that previously proposed methods, such as UOE and CLRL-Tuning, exhibit superiority over the baseline fine-tuning and even match the performance of our model editing approach in the initial stages up to $t = 3$. However, with the incorporation of additional accented data, these approaches start encountering the issue of forgetting, and the model editing approaches consistently outperform all baselines. This underscores the scalability of these model-editing approaches for sequential lifelong learning, with the potential for further enhancements as more accents are introduced.

5.3. Incremental improvements in Model Editing

To assess the incremental enhancements in model editing at each time step t, we compare the intermediate fine-tuned checkpoint $\hat{\theta}_t$ with the edited checkpoint θ_t, where $\theta_t = \theta_{t-1} + \lambda(\tau_t)$. Table 3 presents the AWER for various time steps, comparing both the fine-tuned and edited (TIES-Merging) checkpoints. Note that this fine-tuned checkpoint is different than conventional fine-tuning baseline, since the former is fine-tuned on previously edited checkpoint. The table shows that incorporating task vectors consistently enhances performance at all stages compared to fine-tuning, with WERR gains between 1.5-14%.

5.4. Choosing the optimal λ to mitigate CF

To analyze the impact of the scaling factor λ in the Task Arithmetic technique, we conducted an ablation study by varying λ during $t = 2$. The results, illustrated in Figure 3, reveal impacts across previous accents, new accents, and the average of both.

Notably, $\lambda = 1$, or full fine-tuning, results in catastrophic forgetting for previous accents, leading to the worst performance. Conversely, this setting yields the best performance for the new accent. Intriguingly, a λ value of 0.2 emerges as the most optimal, in contrast to the fixed value of 0.4 used for all Task Arithmetic-based model merging. This finding suggests the potential for fine-tuning λ differently for each time step and even tailoring it for different task vectors. We leave this avenue for further exploration in future work.

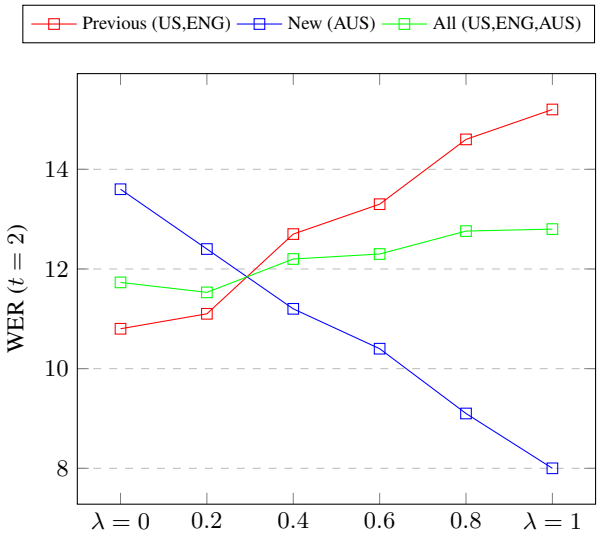

Figure 3: *Variation in WER for previous seen accents vs new accent for $t = 2$ for different λ.*

6. Conclusion

We address the challenge of adapting Automatic Speech Recognition (ASR) to new domains by introducing Lifelong Learning (LLL) algorithms. Traditional methods face computational inefficiencies and concerns about Catastrophic Forgetting (CF) during fine-tuning. While previous LLL techniques exist, our study propose Sequential Model Editing, a novel approach that does not require previous datasets or additional parameters. Empirical results showcase up to a 15% Word Error Rate Reduction (WERR) over the fine-tuning baseline and superior efficiency compared to other LLL techniques on the CV English multi-accent dataset. This approach effectively mitigates CF and maintains high performance across diverse domains.

One avenue for future research involves experimenting with varying values of the scaling factor λ at different time steps, potentially yielding enhanced improvements, as illustrated in Section 5.4. Another avenue is exploring into the theoretical foundations that contribute to the superior performance of model editing in the context of lifelong learning. Additionally, we intend to explore recently proposed editing techniques, including Drop And REscale [19] and Soft Merging of Experts [20].

7. Acknowledgements

We thank Veera Raghavendra Elluru for his constant feedback during the course of work, as well as rebuttal phase of the paper.

8. References

[1] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[2] D. Isele and A. Cosgun, "Selective experience replay for lifelong learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[3] B. Li, R. Pang, Y. Zhang, T. N. Sainath, T. Strohman, P. Haghani, Y. Zhu, B. Farris, N. Gaur, and M. Prasad, "Massively multilingual asr: A lifelong learning solution," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6397–6401.

[4] B. Houston and K. Kirchhoff, "Continual learning for multi-dialect acoustic models," 2020.

[5] M. H. Phan, T.-A. Ta, S. L. Phung, L. Tran-Thanh, and A. Bouzerdoum, "Class similarity weighted knowledge distillation for continual semantic segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 845–16 854.

[6] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," 2022.

[7] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," 2023.

[8] P. Yadav, D. Tam, L. Choshen, C. Raffel, and M. Bansal, "Ties-merging: Resolving interference when merging models," 2023.

[9] X. Jin, X. Ren, D. Preotiuc-Pietro, and P. Cheng, "Dataless knowledge fusion by merging weights of language models," *ArXiv*, vol. abs/2212.09849, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254877510

[10] M. Matena and C. Raffel, "Merging models with fisher-weighted averaging," *ArXiv*, vol. abs/2111.09832, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244345933

[11] 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*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/15825aee15eb335cc13f9b559f166ee8-Paper.pdf

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[14] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[15] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[16] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan, Eds. Edinburgh, Scotland: Association for Computational Linguistics, Jul. 2011, pp. 187–197. [Online]. Available: https://aclanthology.org/W11-2123

[17] Z. Wang, F. Hou, and R. Wang, "Clrl-tuning: A novel continual learning approach for automatic speech recognition."

[18] Y. Takashima, S. Horiguchi, S. Watanabe, P. García, and Y. Kawaguchi, "Updating only encoders prevents catastrophic forgetting of end-to-end asr models," 2022.

[19] L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li, "Language models are super mario: Absorbing abilities from homologous models as a free lunch," 2024.

[20] M. Muqeeth, H. Liu, and C. Raffel, "Soft merging of experts with adaptive routing," 2023.