



Updating Only Encoders Prevents Catastrophic Forgetting of End-to-End ASR Models

Yuki Takashima¹, Shota Horiguchi¹, Shinji Watanabe^{2,3}, Paola García³, Yohei Kawaguchi¹

¹ Hitachi, Ltd. Research & Development Group, Japan

² Language Technologies Institute, Carnegie Mellon University, USA

³ Center for Language and Speech Processing, Johns Hopkins University, USA

yuki.takashima.ot@hitachi.com

Abstract

In this paper, we present an incremental domain adaptation technique to prevent catastrophic forgetting for an end-to-end automatic speech recognition (ASR) model. Conventional approaches require extra parameters of the same size as the model for optimization, and it is difficult to apply these approaches to end-to-end ASR models because they have a huge amount of parameters. To solve this problem, we first investigate which parts of end-to-end ASR models contribute to high accuracy in the target domain while preventing catastrophic forgetting. We conduct experiments on incremental domain adaptation from the LibriSpeech dataset to the AMI meeting corpus with two popular end-to-end ASR models and found that adapting only the linear layers of their encoders can prevent catastrophic forgetting. Then, on the basis of this finding, we develop an element-wise parameter selection focused on specific layers to further reduce the number of fine-tuning parameters. Experimental results show that our approach consistently prevents catastrophic forgetting compared to parameter selection from the whole model. **Index Terms:** Domain adaptation, end-to-end speech recognition, incremental learning

1. Introduction

End-to-end automatic speech recognition (ASR) has made remarkable progress in cases where a large amount of training data is available. However, in real applications, there is often an acoustic mismatch between the training environment and the operational environment where users utilize the ASR model, and it is difficult to collect a sufficient amount of training data in the operational environment beforehand. Therefore, we sometimes collect target domain data during operation and continuously adapt the model using the collected data.

A typical domain adaptation method [1, 2] aims to improve the performance on a target domain and thus usually suffers from performance degradation on a source domain, which is known as catastrophic forgetting [3]. Catastrophic forgetting becomes more severe when the model is continuously adapted to the target domain using sequentially arriving data. To mitigate catastrophic forgetting, several incremental learning methods have been proposed [4, 5, 6, 7]. Regularization-based methods [4, 5] introduce an additional loss during adaptation to make the current model close to the original one. Architecture-based methods [6, 8] dynamically expand a model architecture for new data. In the field of ASR, various studies [9, 10, 11] have demonstrated the benefits of the incremental learning approach. Fu *et al.* [12] proposed an incremental learning algorithm for end-to-end ASR that uses attention distillation and knowledge distillation [13] to prevent catastrophic forgetting. Such meth-

ods require retaining the previous model [6, 12] or adding extra parameters of the same size as the model for optimization [8, 9]. Therefore, adapting recent end-to-end ASR models with a huge amount of parameters is computationally expensive. A potential solution is pruning-based domain adaptation, which reduces the number of fine-tuning parameters by pruning [14, 15]. In [14], parameters of a subnetwork are fixed during adaptation to keep the performance of the source domain. The parameters to be fixed are determined on the basis of a pruning algorithm, since the parameters selected to be pruned are considered important for processing source domain data. However, the pruning algorithm is applied for the entire network, and there has been no investigation into the subnetwork-wise parameter freezing. Some studies on domain adaptation of ASR models have shown that updating only a part of the layers improves the performance on the target domain [16, 17]. However, there has been no research on the performance against catastrophic forgetting.

In this study, we investigate which parts of the end-to-end ASR model should and should not be adapted to the target domain to prevent catastrophic forgetting during incremental domain adaptation. For the experiments, we utilize two popular ASR models: a Transformer-based model and a recurrent neural network transducer (RNN-T). We use LibriSpeech [18] as the large-scale source domain dataset and the AMI meeting corpus [19] as the target domain dataset. We first investigate a module-wise parameter selection to be fine-tuned and show that adapting only the encoder can prevent catastrophic forgetting. Then, to further reduce the number of fine-tuning parameters, we develop an element-wise parameter selection from specific layers in the model. Experimental results demonstrate that our proposed parameter selection could select better parameters to prevent catastrophic forgetting compared to parameter selection from the entire network.

2. End-to-end ASR architecture

In this section we describe the two architectures considered in this paper and explain how the layers in each are formulated. We explicitly define the parameters fed into each function to categorize them into the predefined layer types listed in Table 1. For simplicity, the bias parameters are omitted from the descriptions.

2.1. Transformer-based model

For the Transformer-based ASR model, we used a hybrid connectionist temporal classification (CTC)/attention architecture. It consists of an encoder and decoder network.

The encoder network converts a T_0 -length time sequence of F -dimensional acoustic features X_0 into a T -length sequence

Table 1: *Parameter splitting based on functionality. MHA denotes multi-head attention layers that include self-attention layers, source-target attention layers, or both.*

Module name	Transformer-based	RNN-T
Embeddings	$\Phi^{(\text{enc})} \cup W_{\text{emb}}$	$\Phi^{(\text{enc})} \cup W_{\text{emb}}$
Layer normalization	$\Psi^{(\text{enc})} \cup \Psi^{(\text{dec})}$	$\Psi'^{(\text{enc})}$
MHA & FFN in encoder	$\Theta^{(\text{enc})}$	–
MHA & FFN in decoder	$\Theta^{(\text{dec})}$	–
MHA & Conv & FFN in encoder	–	$\Lambda^{(\text{enc})}$
Prediction	–	$\Lambda^{(\text{pred})}$
CTC	W_{CTC}	–
Output	W_{out}	$\Lambda^{(\text{joint})}$

of D -dimensional hidden embeddings E_{out} , as

$$X = f_{\text{sub}}(X_0; \Phi^{(\text{enc})}) \in \mathbb{R}^{D \times T}, \quad (1)$$

$$E = f_{\text{enc}}(X; \Theta^{(\text{enc})}, \Psi^{(\text{enc})}) \in \mathbb{R}^{D \times T}. \quad (2)$$

$f_{\text{sub}}(\cdot)$ is the sub-sampling layer including the positional encodings [20] parameterized by $\Phi^{(\text{enc})} := \Gamma \cup \{W_0\}$, where Γ is the set of parameters of the convolutional sub-sampling and $W_0 \in \mathbb{R}^{D \times F'}$ is the subsequent linear projection matrix. X is a sequence of intermediate representations with a length of $T (< T_0)$. $f_{\text{enc}}(\cdot)$ is the N -stacked Transformer encoder parameterized by $\Theta^{(\text{enc})}$ and $\Psi^{(\text{enc})}$. $\Theta^{(\text{enc})}$ is a set of parameters of the linear projections in the encoders, namely,

$$\Theta^{(\text{enc})} := \bigcup_{i=1}^N \psi_{\text{self}}^{(i)} \cup \psi_{\text{FFN}}^{(i)}, \quad (3)$$

where $\psi_{\text{self}}^{(i)}$ is a set of weight matrices for the query, key, value, and output projections of the multi-head self-attention layer in the i -th encoder block, and $\psi_{\text{FFN}}^{(i)}$ is a set of weight matrices of the two-layer feed-forward network (FFN) layer in the i -th encoder block. $\Psi^{(\text{enc})}$ is a set of parameters for the element-wise affine transformation in the layer normalization layer [21].

The decoder network calculates the L -length target sequence $[\mathbf{y}_l]_{l=1}^L$ from the encoder output E in a sequence-to-sequence manner, as follows:

$$\mathbf{z}_l = f_{\text{emb}}(\mathbf{y}_l; W_{\text{emb}}) \in \mathbb{R}^D, \quad (4)$$

$$\mathbf{o}_l = f_{\text{dec}}([\mathbf{z}_1, \dots, \mathbf{z}_{l-1}], E; \Theta^{(\text{dec})}, \Psi^{(\text{dec})}), \quad (5)$$

$$\hat{\mathbf{y}}_l = f_{\text{out}}(\mathbf{o}_l; W_{\text{out}}) \in (0, 1)^V. \quad (6)$$

f_{emb} is the embedding layer that converts a one-hot encoding $\mathbf{y} \in \{0, 1\}^V$ into the corresponding embedding $\mathbf{z}_n \in \mathbb{R}^D$ using the projection matrix $W_{\text{emb}} \in \mathbb{R}^{D \times V}$, where V is the vocabulary size. f_{dec} is M -stacked Transformer decoders parameterized by $\Theta^{(\text{dec})}$ and $\Psi^{(\text{dec})}$. $\Theta^{(\text{dec})}$ is a set of parameters of the linear projections in the decoders, denoted as

$$\Theta^{(\text{dec})} := \bigcup_{i=1}^M \phi_{\text{self}}^{(i)} \cup \phi_{\text{src}}^{(i)} \cup \phi_{\text{FFN}}^{(i)}, \quad (7)$$

where $\phi_{\text{self}}^{(i)}$ and $\phi_{\text{src}}^{(i)}$ are sets of parameters of the multi-head self-attention layer and multi-head source-target attention layer in the i -th decoder block, and $\phi_{\text{FFN}}^{(i)}$ is a set of weight matrices of two-layer FFNs in the i -th decoder block. Each of $\phi_{\text{self}}^{(i)}$ and $\phi_{\text{src}}^{(i)}$ consists of four weight matrices, similar to $\psi_{\text{self}}^{(i)}$. A

set of parameters $\Psi^{(\text{dec})}$ consists of weight vectors for element-wise affine transformation in the layer normalization layers. The classification layer f_{out} converts the output from the Transformer decoder into the vocabulary-wise posterior probability. It consists of a linear layer parameterized by $W_{\text{out}} \in \mathbb{R}^{V \times D}$ followed by a softmax function.

The entire network is optimized by minimizing the negative log-likelihood and the CTC loss [20]. The negative log-likelihood is calculated between the target sequence and the output sequence from the decoder. The CTC loss is determined between the target sequence and the sequence $[\bar{\mathbf{y}}_t]_{t=1}^T$, each of which is calculated from the encoder output $E = [e_t]_{t=1}^T$ as follows:

$$\bar{\mathbf{y}}_t = f_{\text{CTC}}(e_t; W_{\text{CTC}}) \in (0, 1)^{V \times T}, \quad (8)$$

where $f_{\text{CTC}}(\cdot)$ is the CTC function using the projection matrix $W_{\text{CTC}} \in \mathbb{R}^{V \times D}$. The CTC layer maps a sequence of the frame-wise label probabilities into a target label sequence to calculate the likelihood of a target sequence (for more details, see [22]).

2.2. Recurrent neural network transducer

The RNN-T ASR model utilizes a Conformer encoder, a long short-term memory (LSTM)-based prediction network, and a joint network. Similar to the Transformer-based model, the input acoustic features are sub-sampled before being fed into the encoder using Eq. (1).

The encoder converts a T -length D -dimensional sequence into the same dimensional hidden embedding sequence as

$$[e'_1, \dots, e'_T] = f'_{\text{enc}}(X; \Lambda^{(\text{enc})}, \Psi'^{(\text{enc})}) \in \mathbb{R}^{D \times T}, \quad (9)$$

where e'_t is the t -th hidden embedding. $f'_{\text{enc}}(\cdot)$ is the N -stacked Conformer encoder [23] parameterized by $\Lambda^{(\text{enc})}$ and $\Psi'^{(\text{enc})}$. $\Lambda^{(\text{enc})}$ is a set of parameters of the linear projections in the encoders, namely,

$$\Lambda^{(\text{enc})} := \bigcup_{i=1}^N \lambda_{\text{pre}}^{(i)} \cup \lambda_{\text{self}}^{(i)} \cup \lambda_{\text{conv}}^{(i)} \cup \lambda_{\text{post}}^{(i)}, \quad (10)$$

where $\lambda_{\text{self}}^{(i)}$ is a set of parameters of the multi-head self-attention layer in the i -th encoder, and $\lambda_{\text{conv}}^{(i)}$ is a set of parameters of convolutional kernels including pointwise and depthwise convolutions in the i -th encoder. $\lambda_{\text{pre}}^{(i)}$ and $\lambda_{\text{post}}^{(i)}$ are sets of parameters of two-layer FFN layers before the self-attention layer and after the convolution layer, respectively. Ψ is the set of parameters of the affine transformations in the layer normalization layers.

The prediction network f_{pred} calculates the L -length hidden representations $[\mathbf{h}_l]_{l=1}^L$ in a sequence-to-sequence manner, as follows:

$$\mathbf{h}_l = f_{\text{pred}}([\mathbf{z}_1, \dots, \mathbf{z}_{l-1}]; \Lambda^{(\text{pred})}) \in \mathbb{R}^D, \quad (11)$$

where $\Lambda^{(\text{pred})}$ is the parameters of LSTM [24]. \mathbf{z}_l is the hidden representation calculated using Eq. (4).

The joint network $f_{\text{joint}}(\cdot)$ integrates each frame of the encoder output e'_t and the output \mathbf{h}_l from the prediction network using a feed-forward network as follows:

$$\bar{\mathbf{y}}_{t,l} = f_{\text{joint}}(e'_t, \mathbf{h}_l; \Lambda^{(\text{joint})}) \in (0, 1)^V, \quad (12)$$

where $\Lambda^{(\text{joint})}$ is a set of weight matrices for the encoder-subspace, prediction-subspace, and subspace-output projections.

The likelihood of the target sequence is calculated as the objective function by summing the possibilities of all possible alignments for the target sequence [25] as well as CTC.

Table 2: Statistics of adaptation, development, and test sets.

		#Utterances	Duration (h)
Adaptation sets	AMI-stage1	25,904	18.3
	AMI-stage2	28,746	19.7
	AMI-stage3	24,265	18.0
Development set	AMI	13,059	8.9
Test sets	LibriSpeech-clean	2,620	5.4
	LibriSpeech-other	2,939	5.3
	AMI-scenario	7,503	5.7
	AMI-non-scenario	5,109	3.0

3. Method

3.1. Module-wise parameter selection

Recent end-to-end ASR models consist of several types of layers, as mentioned in Section 2. These layers capture different characteristics and are typically utilized to transfer their characteristics to the target-domain model [16, 17]. With this observation in mind, we categorize them into eight modules, as shown in Table 1. We assume that each module differs in adaptability to the target domain and a forgetting property of the source domain. Our first method, module-wise parameter selection, aims to clarify whether we can choose the optimal modules to be updated to achieve end-to-end ASR models that perform well on both source and target domains. Given a model well-trained on the source domain dataset, we select several modules (shown in Table 1) to be fine-tuned. The parameters of the remaining modules are kept unchanged during adaptation. We conducted multiple stages of adaptation using subdivided data of the target domain. Note that each subdivided data is used only in one of the stages and is not carried over between stages.

3.2. Element-wise parameter selection for specific layers

Each module in Table 1 still has a large number of parameters, especially in the encoder and decoder. To further reduce the number of fine-tuning parameters, we propose an element-wise parameter selection for specific modules. Given a model trained on the source domain, we first select a subset of parameters within specific modules. In this work, we determine these specific modules on the basis of the investigation in Section 3.1 to prevent catastrophic forgetting. For parameter selection, we utilize random selection and magnitude-based selection, the latter of which is commonly used in network pruning [15, 26]. Then, we fine-tune only the selected parameters during adaptation.

4. Experiments

4.1. Conditions

We conducted incremental domain adaptation experiments using the LibriSpeech dataset [18] as the source domain and the AMI meeting corpus [19] as the target domain. For the AMI corpus, we selected the official Full-corpus-ASR partition and randomly divided the scenario meetings in the training set into three subsets to conduct three stages of adaptation. AMI-stage{1,2,3} were used in this order. We also divided the test set into two subsets, scenario meetings and non-scenario meetings, which were used to evaluate the performance on target domain data and out-of-domain data, respectively. Additionally, we used the test sets of the LibriSpeech dataset to measure performance degradation on the source domain. The statistics of

each dataset are listed in Table 2.

We used 80-dimensional log-mel features with a 25 ms frame length and 10 ms frame shift as input for the end-to-end ASR models. We used a 5k vocabulary based on the subword segmentation algorithm [27] with 5,002 output dimensions V including subword tokens. For adaptation, we utilized the Adam optimizer [28] with a fixed learning rate of 1×10^{-3} . For decoding, we used a beam width of 10.

We used the pre-trained Transformer-based model with roughly 78.4M parameters in ESPnet [29], which has 12 Transformer encoder blocks and six Transformer decoder blocks. The number of attention heads was eight. The sub-sampling factor and the embedding dimensionality were set to 4 (*i.e.*, $T = T_0/4$) and $D = 512$. For the convolutional sub-sampling layer, $\Phi^{(enc)}$ has two convolution layers with a 3×3 filter followed by a ReLU activation with a channel size of 512. The model was optimized by minimizing the negative log-likelihood and CTC loss with an interpolation weight 0.3. For inference, we implemented two decoding methods: joint CTC/attention decoding and CTC decoding.

For the RNN-T model, we also used the pre-trained model with roughly 88.7M parameters in ESPnet, which has 12 Conformer encoder blocks with four attention heads. The embedding dimensionality was also set to $D = 512$. For the convolutional sub-sampling layer, $\Phi^{(enc)}$ has four convolution layers with a 3×3 filter followed by a ReLU activation with channel sizes 64, 64, 128, 128. The prediction network has an embedding layer of 1,024 units and one LSTM layer of 512 units. The joint network is composed of FFNs with 512 hidden units.

For the investigation of the optimal adapting parameters of the model, we split the parameters of the Transformer-based model and the RNN-T model into six and five modules, respectively, as shown in Table 1.

4.2. Results

Figure 1 shows the word error rates (WERs) for the Transformer-based model and the RNN-T model. In this figure, the bottom left is better, and going right means catastrophic forgetting. Fine-tuning all the parameters (‘fine-tune’) caused catastrophic forgetting. As shown in Fig. 1a, fine-tuning weight matrices in the Transformer (③&④) achieved the best performance on both domains. In particular, weight matrices in the encoder (③) contributed to improving the performance on the target domain while those in the decoder (④) contributed to preventing catastrophic forgetting. On the other hand, fine-tuning the output layer (⑥) caused severe catastrophic forgetting even though this layer has only 3.3% of the model parameters. We can see that modules related to language characteristics (except for the output module) tend to retain the performance of the source domain. We hypothesize that the model could have learned sufficient language characteristics from the LibriSpeech dataset. Moreover, the embedding module achieved a small improvement on the target domain compared to other modules as well as the previous work [15], in which the embedding layer did not yield consistent differences for adaptation. Figure 1b shows the results of CTC decoding, where the results of fine-tuning the decoder (④) and the output layer (⑥) are not plotted because these results are identical to the result without adaptation. Here, we can see similar trends with CTC/attention decoding.

In the case of the RNN-T model shown in Fig. 1c, we also see similar trends to the Transformer-based model. Fine-tuning the joint network (④) caused catastrophic forgetting with poor

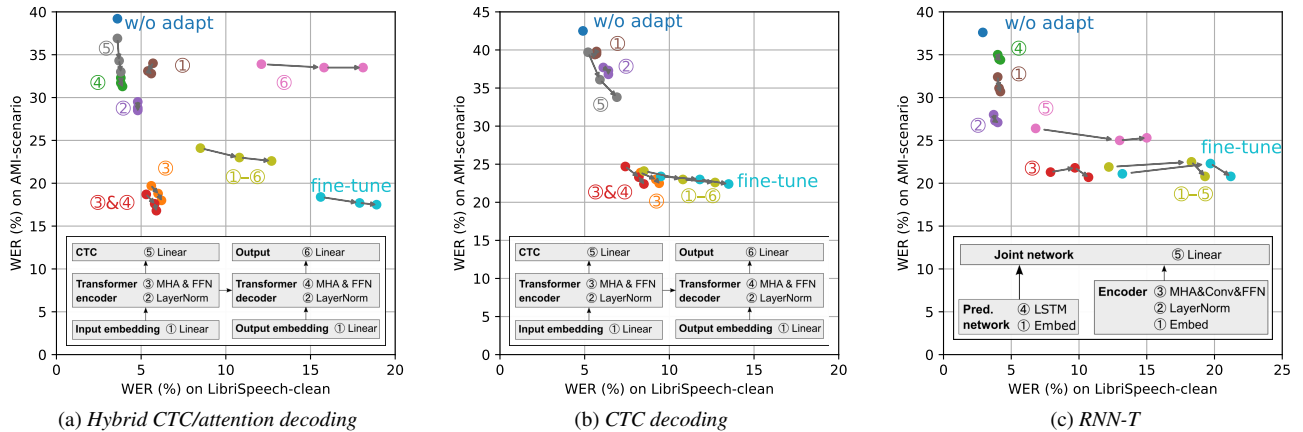


Figure 1: WERs (%) with different adapting modules. (a) and (b) show the results using different decoding methods in the Transformer-based model. (c) shows the results of the RNN-T model. The arrows indicate the trajectory of adaptation.

Table 3: WERs (%) for different methods on four test sets. #Params denotes the number of fine-tuning parameters. ‘③&④’ corresponds to the numbers shown in Fig. 1.

Test set	Method	#Params	Stage1	Stage2	Stage3
LibriSpeech-clean	EWC	78.4M	5.7	6.0	6.0
	Fine-tuning ③&④	62.9M	5.3	5.8	5.9
LibriSpeech-other	EWC	78.4M	13.5	14.6	15.0
	Fine-tuning ③&④	62.9M	12.5	13.9	14.4
AMI-scenario	EWC	78.4M	18.4	17.7	17.4
	Fine-tuning ③&④	62.9M	18.7	17.6	16.8
AMI-non-scenario	EWC	78.4M	29.5	28.8	28.4
	Fine-tuning ③&④	62.9M	31.0	29.1	27.9

performance on the target domain. Fine-tuning the Conformer encoder (③) achieved competitive WERs with fine-tuning all the parameters on the target domain while keeping degradation on the source domain small. From these results, we conclude that the encoder in the end-to-end ASR model is an optimal choice for adaptation. The encoder is a counterpart of the acoustic model, and our conclusion is in good agreement with the well-known traditional acoustic model adaptation [30]. We also conclude that adaptation of the encoder contributes to the prevention of catastrophic forgetting.

Next, we conducted comparative experiments with our approach and the conventional incremental learning, EWC [5]. We utilized the Transformer-based model with hybrid CTC/attention decoding shown in Fig. 1a. As shown in Table 3, our approach outperformed EWC on all test sets in the final stage. However, in stage 1 and 2, the WERs of our approach on AMI-non-scenario were slightly higher than those of EWC. We hypothesize that EWC was able to adapt the model to the AMI dataset well in the early stages of adaptation because it fine-tuned all the parameters of the model. Our approach has the advantage of not requiring any pre-computation or additional parameters, while EWC requires the Fisher information matrix calculated on the source domain. These results demonstrate that the simple fine-tuning approach is effective for incremental domain adaptation.

Finally, we show the results of the element-wise parameter selection. We again utilized the Transformer-based model with hybrid CTC/attention decoding. We evaluated three parameter selection approaches: random, smaller-magnitude, and larger-magnitude. We applied parameter selection to weight matrices

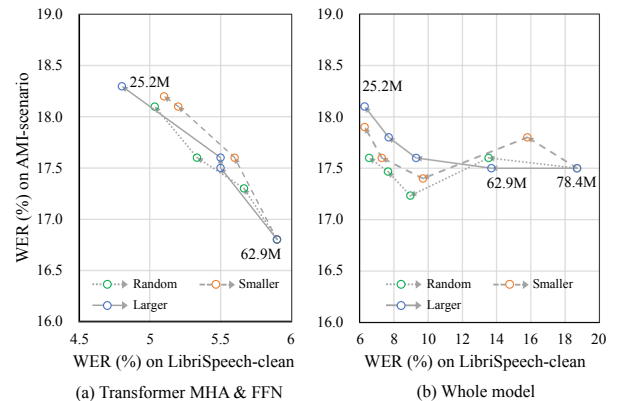


Figure 2: WERs (%) at the third stage of adaptation with various numbers of adapting parameters: 62.9M, 50.3M, 37.8M, and 25.2M. 62.9M corresponds to the number of parameters of Transformer MHA & FFN. The WERs of the random selection are mean values across three trials.

in the Transformer and the whole model that correspond to ③&④ and ‘fine-tune’ in Fig. 1. As shown in Fig. 2, parameter selection with only 37.8M parameters achieved competitive results on the target domain compared with fine-tuning all the parameters while preventing catastrophic forgetting. However, we could not observe a large difference among different parameter selections. In the case of the whole model, we obtained inconsistent WERs among different numbers of fine-tuning parameters, except for larger-magnitude selection. These results show that fine-tuning a part of the parameters within the Transformer achieves a consistent performance regardless of the difference in the parameter selection.

5. Conclusions

In this paper, we presented a simple but effective incremental domain adaptation strategy to prevent catastrophic forgetting for end-to-end ASR models. Our experiments revealed that parameter selection can sufficiently prevent catastrophic forgetting without the complex procedures used in the standard incremental domain adaptation methods [4, 5]. In future work, we will apply our findings to conventional incremental domain adaptation methods.

6. References

- [1] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. NIPS*, 2006, p. 137–144.
- [2] S. Mirsamadi and J. H. Hansen, "On multi-domain training and adaptation of end-to-end RNN acoustic models for distant speech recognition," in *Proc. INTERSPEECH*, 2017, pp. 404–408.
- [3] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.
- [4] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, p. 2935–2947, 2018.
- [5] 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.
- [6] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," in *Proc. ICLR*, 2018.
- [7] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. NIPS*, 2017, pp. 6467–6476.
- [8] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.
- [9] S. Ghorbani, S. Khorram, and J. H. L. Hansen, "Domain expansion in DNN-based acoustic models for robust speech recognition," in *Proc. ASRU*, 2019, pp. 107–113.
- [10] S. Sadhu and H. Hermansky, "Continual learning in automatic speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 1246–1250.
- [11] B. Li, R. Pang, Y. Zhang, T. N. Sainath, T. Strohmaier, P. Haghani, Y. Zhu, B. Farris, N. Gaur, and M. Prasad, "Massively multilingual ASR: A lifelong learning solution," in *Proc. ICASSP*, 2022, pp. 6397–6401.
- [12] L. Fu, X. Li, L. Zi, Z. Zhang, Y. Wu, X. He, and B. Zhou, "Incremental learning for end-to-end automatic speech recognition," in *Proc. ASRU*, 2021, pp. 320–327.
- [13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS workshop*, 2015.
- [14] S. Gu, Y. Feng, and W. Xie, "Pruning-then-expanding model for domain adaptation of neural machine translation," in *Proc. NAACL-HLT*, 2021, pp. 3942–3952.
- [15] H. Zhu, Z. Wang, H. Zhang, M. Liu, S. Zhao, and B. Qin, "Less is more: Domain adaptation with lottery ticket for reading comprehension," in *Proc. EMNLP*, 2021, pp. 1102–1113.
- [16] S. Ueno, T. Moriya, M. Mimura, S. Sakai, Y. Shinohara, Y. Yamaguchi, Y. Aono, and T. Kawahara, "Encoder transfer for attention-based acoustic-to-word speech recognition," in *Proc. INTERSPEECH*, 2018, pp. 2424–2428.
- [17] V. N. Sukhadia and S. Umesh, "Domain adaptation of low-resource target-domain models using well-trained asr conformer models," *arXiv preprint arXiv:2202.09167*, 2022.
- [18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [19] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. MLMI*, 2005, pp. 28–39.
- [20] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on Transformer vs RNN in speech applications," in *Proc. ASRU*, 2019, pp. 449–456.
- [21] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *Proc. NIPS 2016 Deep Learning Symposium*, 2016.
- [22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets," in *Proc. ICML*, 2006, p. 369–376.
- [23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.
- [24] M. Huang, J. Zhang, M. Cai, Y. Zhang, J. Yao, Y. You, Y. He, and Z. Ma, "Improving RNN transducer with normalized jointer network," *arXiv preprint arXiv:2011.01576*, 2020.
- [25] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.
- [26] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems*, vol. 28, 2015.
- [27] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proc. ACL*, 2018, pp. 66–75.
- [28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
- [29] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.
- [30] S. Sagayama, K. Shinoda, M. Nakai, and H. Shimodaira, "Analytic methods for acoustic model adaptation: A review," in *Proc. ISCA ITR-Workshop*, 2001, pp. 67–76.