GemST: Continual Learning for End-to-End Speech-to-Text Translation

Pranav Karande
IIT Indore
pranav.3943@gmail.com

Balaram Sarkar
IIT Indore
balarakar@gmail.com

Chandresh Kumar Maurya
IIT Indore
ckm.jnu@gmail.com

Abstract

Effective cross-lingual communication remains a significant challenge in today's rapidly globalizing world. Developing Speech-to-Text Translation (S2T) systems using artificial intelligence presents various difficulties, such as the unavailability of all language pairs for simultaneous model training. Additionally, when a model is trained on a new language, it often loses its ability to remember previously learned tasks, a phenomenon known as catastrophic forgetting. This paper explores the application of Gradient Episodic Memory (GEM) to address these challenges. Our study investigates the effectiveness of GEM in enhancing S2T model performance across sequentially introduced language pairs. Experimental results demonstrate that GEM can significantly reduce forgetting by **24.8%** and boost translation accuracy by **44.5%** as compared to baseline, offering a promising approach for scalable and efficient multilingual-continual S2T systems.

1 Introduction

Speech-to-text (S2T) translation is a technology that bridges language barriers by converting spoken language into written text in a different language. This capability is increasingly vital in our globalized world, where effective and seamless communication across diverse linguistic communities is essential. Traditional S2T translation systems like (Bansal et al., 2017; Le et al., 2021; Sarkar et al., 2023) typically require large, diverse datasets for training and are often retrained from scratch whenever new language pairs are introduced. This process is not only computationally expensive and time-consuming but also environmentally unsustainable due to the high energy consumption involved.

Continual learning, also known as lifelong learning, offers a promising solution to these challenges. In the realm of S2T translation, continual learning allows models to adapt incrementally to new languages and dialects without forgetting previously learned ones (Bremner et al., 2013; Rusu et al., 2022). Traditional methods suffer from Catastrophic Forgetting (McCloskey and Cohen, 1989), where a model's performance on previously learned tasks degrades as it learns new ones. Continual learning mitigates this issue by enabling S2T models to retain past knowledge while incorporating new information, thus maintaining high performance across all tasks. This approach not only improves the efficiency and scalability of multilingual S2T systems but also ensures they remain effective as new language data becomes available.

To this end our contributions include studying continual learning for end-to-end (E2E) S2T to mit-

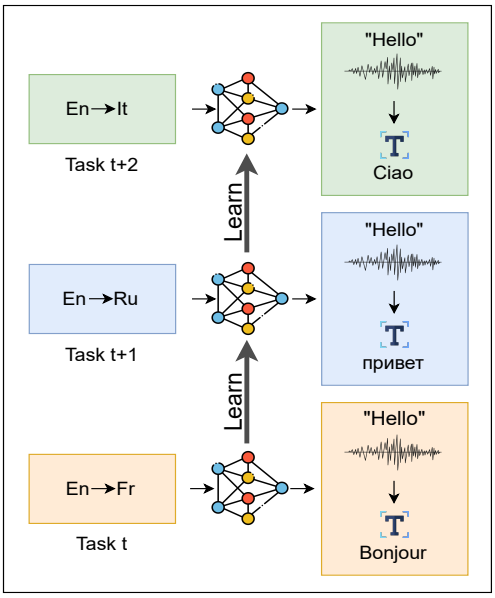

Figure 1: Task diagram of proposed work where new task (S2T for English to Russian, t+1) is trained using the model from previous task (S2T for English to French, t). And similarly for English to Italian (task t+2 trained from model of task t+1).

igate the catastrophic forgetting. To the best of our knowledge, no prior research has been conducted on this specific domain.

2 Related Work

Recently, several studies have applied continual learning to automatic speech recognition. (Sadhu and Hermansky, 2020) sequentially trained an HMM-DNN model on four different tasks using the Wall Street Journal, Reverb, Librispeech , and Chime4 corpora. (Chang et al., 2021) developed an end-to-end ASR model in which they first pre-trained it on the WSJ corpus then on Librispeech and finally on the Switchboard corpus (Godfrey et al., 1992) tested the model's performance on different speech recognition tasks after each update. As real-world data distributions vary a lot from one task to another, it becomes quite essential to know when the training data is presented with a different task than the one it was trained for. In this respect, (Zeno et al., 2019) came up with a Bayesian approach to continual learning that does not require knowledge at the time of transition from one task to another. Similarly, (Mai et al., 2021) introduced the concept of online continual learning over image classification, assuming that the emergence of new classes or instances of images may include a variety of online data streams.

Applications of continual learning have been successfully extended to various tasks such as computer vision (Aljundi et al., 2017) and automatic speech recognition (Eeckt and hamme, 2023; Fu et al., 2021). This has not been investigated with respect to Speech-to-Text Translation so far.

3 Problem Statement

The continual learning of S2T models is defined as follows. First, we assume we have an initial model that has been trained on a given dataset (D_0). This model serves as a seed model on which a sequence of continual learning updates is applied. Second, we have a set of labeled datasets $D = \{D_i\}_{i=1}^{N}$ which become available sequentially over time for model training. N represents the total number of language pairs available to train the model. Retraining the S2T model from scratch each time a new dataset D_n becomes available incurs substantial computational costs. Hence a replay-based continual learning method retains few samples from previous tasks to minimize the L2 distance between gradients of new and old data, thereby preserving past knowledge while learning new tasks:

$$\mathcal{L}_{total} = \mathcal{L}_{new} + \lambda \sum_{i=0}^{n-1} \|g_{new} - g_{old}^{i}\|_2^2 \quad (1)$$

Where, $\mathcal{L}_{new}$ is the loss on current task, g_{new} is the gradient of the loss with respect to the new data, g_{old}^{i} is the gradient of the loss with respect to the samples from the i-th previous task and λ is a regularization parameter that controls the importance of preserving old knowledge.

4 Methodology

The S2T baseline used is a transformer-based encoder-decoder model (Vaswani et al., 2023). The hypothesis is that the model being trained for future tasks be optimized by comparing the gradients of previous tasks. To that end, we aspire to employ the approach originally proposed for visual recognition task handling continual learning using gradient episodic memory (GEM) (Lopez-Paz and Ranzato, 2022). Motivated by its recent application in computer vision tasks, we ask the following research question: Will the same approach be able to learn in an S2T setting? We confirm that using GEM in S2T setting, we are able to mitigate the catastrophic forgetting.

4.1 Gradient Episodic Memory (GEM)

GEM is a replay-based continual learning method that retains samples from past data in its memory. When the model encounters data from a new task, it minimizes the L2 distance between the gradients of the new data and the old data. To minimize the total loss $\mathcal{L}_{\text{total}}$, we ensure that the update to the model parameters does not significantly change the gradients computed for the old tasks. This constraint helps to prevent catastrophic forgetting. The new gradient is calculated as:

$$g_{\text{new}} = \nabla_\theta \mathcal{L}_{\text{new}} \quad (2)$$

where θ represents the model parameters, and $\mathcal{L}_{\text{new}}$ is the loss function for the new task. The gradients from the stored examples are then calculated as:

$$g_{\text{old}}^{i} = \nabla_\theta \mathcal{L}_{\text{old}}^{i} \quad \text{for} \quad i = 0, \ldots, n-1 \quad (3)$$

where $\mathcal{L}_{\text{old}}^{i}$ is the loss for the samples from the i-th previous task stored in the episodic memory. The gradient ω to prevent catastrophic forgetting is then defined as:

$$\omega = G^{\top} v + g_{\text{new}} \quad (4)$$

where $G = (g_{\text{old}}^1, \ldots, g_{\text{old}}^{n-1})$ is the matrix of gradients for the previous tasks, and v is a vector obtained by solving the quadratic programming problem that ensures the constraints on gradient alignment.

4.2 S2T Transformer

The Transformer model is an adaptation of the Transformer architecture, specifically designed to handle speech representations as input. These features are inputted into the S2T encoder, which is composed of several layers utilizing self-attention mechanisms. These mechanisms allow the model to process various segments of the input sequence, thereby efficiently capturing long-range dependencies. The self-attention mechanism calculates attention weights to emphasize key features during the decoding process. In the training phase, the model is fine-tuned to align with the ground truth target text by optimizing the following loss function:

$$\mathcal{L}_{ST} = -\sum_{n} \log P(x_n|y_n) \tag{5}$$

Here, $\mathcal{L}_{ST}$ represents the label-smoothed cross-entropy loss on speech and target language text pairs, x is the input speech and y is the target text. This loss is calculated by updating the model parameters θ such that it doesn't change the gradients of previous tasks g_{old}. The S2T Transformer generates a sequence of predicted tokens that articulate the translated textual representation.

5 Experiment

In this section, we detail the following components: (a) datasets, (b) baselines, (c) training and testbed and (d) evaluation metrics.

5.1 Dataset

We conduct experiments on four pairs of translation directions available in **MuST-C**[1] (Di Gangi et al., 2019) dataset: English (En) to German (De), French (Fr), Russian (Ru) and Dutch (Nl). It contains audio, transcript and translation from TED talks for each direction. The statistics of the dataset is shown in Table 1.

5.2 Baselines

As there is no previous continual learning baseline for S2T models, we create two baselines of our own. First is to simply $fine\text{-}tune$ the model on new tasks and the second baseline is a setup where all task's datasets are available together during training as it is a $joint$ approach. In this experiment, we consider the $fine\text{-}tune$ to be a lower bound and $joint$ to be an upper bound for the performance of the model.

[1]We use v1.0. `https://ict.fbk.eu/must-c/`

MuST-C Dataset					
En →	Hours	#Sents	Train	Val	Test
De	408	274K	269K	1.5K	2.8K
Fr	492	280K	275K	1.4K	2.6K
Ru	489	270K	265K	1.3K	2.5K
Nl	442	253K	248K	1.4K	2.6K

Table 1: Train, test and validation splits for MuST-C.

5.3 Training and Testbed

In this study, we utilized the FAIRSEQ S2T toolkit (Wang et al., 2020) to implement our method. The core architecture is an S2T Transformer encoder-decoder model. Both the encoder and decoder consist of 6 self-attention layers, each featuring 8 attention heads. Due to limitations in training resources, the encoder and decoder are of the $small$ configuration, comprising of 256 hidden units. Data augmentation is performed using SpecAugment (Park et al., 2019), and the GELU activation function is employed to enhance convergence, normalization and training stability. The S2T model is trained with label-smoothed cross-entropy loss, with a label smoothing factor set at 0.1. The Adam optimizer is used, featuring a learning rate of 1e-4, and the learning rate schedule follows an inverse square root pattern.

5.4 Performance Metric

Case-sensitive detokenized BLEU using sacreBLEU (Post, 2018) is used to report the performance of the model. All experiments are repeated with three different random seeds, and we report the average BLEU on the MuST-C `tst-COMMON` set.

6 Results

We measure the performance of the system across four tasks sequentially as shown on Table 2. T-1 was conducted on De, T-2 on Fr, T-3 on Ru, and T-4 on Nl. The goal is to retain model performance on previous tasks while performing T-2, T-3 and T-4. The results are given in terms of BLEU scores

	T-1	T-2			T-3				T-4					Agg.
	De	De	Fr	Avg	De	Fr	Ru	Avg	De	Fr	Ru	Nl	Avg	Avg
Fine-Tune	23.85	0.5	30.1	15.3	0.2	0.2	17.23	5.87	0.3	0.2	0.1	28.78	7.34	7.37
Forg		98%			99.2%	99.3%			98.8%	99.3%	99.4%			98.8%
Joint	26.02	26.02	36.05	31.03	26.02	36.05	18.23	26.76	26.02	36.05	18.23	29.78	27.52	27.84
GEM	23.85	5.42	26.71	16.07	5.59	6.33	14.41	8.78	4.88	5.77	4.32	24.86	9.96	**10.65**
Forg		77%			76.5%	76.3%			79.5%	78.4%	70%			**74.3%**

Table 2: Task-wise average BLEU score and forgetting on four pairs of MuST-C data. Fine-tune and Joint are the baselines whereas GEM is the proposed method for continually learning S2T models. Forg denotes the forgetting on that method. Here, T-1, T-2, T-3, and T-4 are tasks where we train the model on De, Fr, Ru and Nl language pairs sequentially. Agg Avg is overall average.

and Forgetting in percentages, which quantify the retention of tasks learned before.

6.1 Automatic Evaluation

As seen in Table 2 for the Task 1 with $fine\text{-}tune$, the BLEU score for De is 23.85 whereas it significantly lowered in next subsequent tasks. In Task 2, $fine\text{-}tune's$ BLEU score on De lowers to 0.5 and further goes even worse down to 0.2 after Task 3 and after Task 4 to 0.3. It shows similar result with other language pairs as the number of task increases. Conversely, GEM demonstrates quite smooth performance with BLEU score of 23.85 for Task 1 on De, and a score of 5.42 for De after Task 2, 6.33 after Task 3, and 4.32 after Task 4, showing that the model is able to remember the previous task. It follows a similar score for other languages as well. In Figure 2, although Nl in Task 4 is trained for the first time in both $fine\text{-}tune$ and GEM, the increase in BLEU score can be explained due to the forward transfer experienced by the S2T model using GEM. The result shows that GEM is able to preserve previous knowledge at an average BLEU score of 10.65 across all tasks compared to the baseline $fine\text{-}tune$ with an average of 7.37.

Forgetting: One of the main challenges of continual learning is forgetting, which means that the performance on the tasks learned earlier in the run deteriorates upon the introduction of new tasks. From Table 2 we see $fine\text{-}tune$ baseline has very high forgetting rates of 98% on Task 1, 99.2% on Task 2, 98.8% on Task 3, and 99.4% on Task 4. However, this effect of forgetting is considerably reduced if applied GEM: 77% for Task 1, 76.5% for Task 2, 76.3% for Task 3, and 70% for Task 4. This reduces the average rate of forgetting for GEM to be 74.3%.

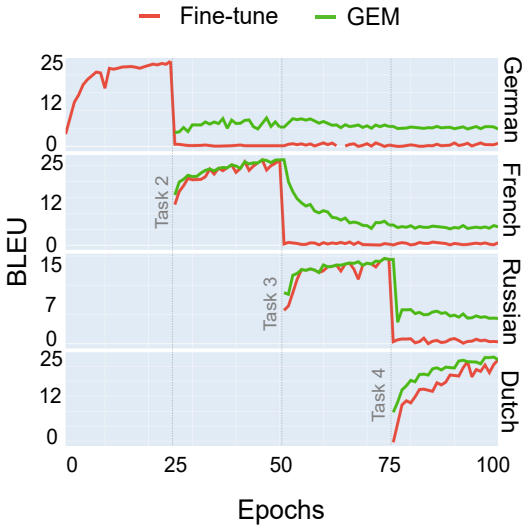

Figure 2: Epoch wise BLEU score on all tasks trained sequentially.

7 Conclusion

In this paper, we propose GemST, a new method for continually learning Speech-to-Text models. Results obtained from our experiments on the MuST-C dataset indicate that GEM not only improves the BLEU scores of multiple tasks compared to the baseline, but it also causes a requisite massive drop in the forgetting rates. Hence it demonstrate GEM's efficacy toward the development of robust S2T systems that learn tasks introduced sequentially without suffering from the so-called catastrophic forgetting. This development paves the way for future research and development on continual learning methodologies within the S2T domain.

Limitations

While our proposed method demonstrates superior performance compared to the baseline, a few limitations should be noted: (1) While GEM effectively retains knowledge from previous tasks, there is potential to further minimize the forgetting. Developing more advanced methods could lead to greater reductions in forgetting, enhancing the overall performance of the model, (2) As this study presents the first application of continual learning to S2T, there is a lack of established baselines. Future work could develop and compare additional continual learning baselines to provide a more comprehensive evaluation. Nevertheless, our primary objective was to initiate research in this area.

References

Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. 2017. Expert gate: Lifelong learning with a network of experts.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 474–479, Valencia, Spain. Association for Computational Linguistics.

Andrew Bremner, David Lewkowicz, and Charles Spence. 2013. Multisensory development.

Heng-Jui Chang, Hung yi Lee, and Lin shan Lee. 2021. Towards lifelong learning of end-to-end asr.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven Vander Eeckt and Hugo Van hamme. 2023. Continual learning for monolingual end-to-end automatic speech recognition.

Li Fu, Xiaoxiao Li, Libo Zi, Zhengchen Zhang, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Incremental learning for end-to-end automatic speech recognition.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824, Online. Association for Computational Linguistics.

David Lopez-Paz and Marc'Aurelio Ranzato. 2022. Gradient episodic memory for continual learning.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. 2021. Online continual learning in image classification: An empirical survey.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, interspeech$_2$019.$ISCA$.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2022. Progressive neural networks.

Samik Sadhu and Hynek Hermansky. 2020. Continual Learning in Automatic Speech Recognition. In *Proc. Interspeech 2020*, pages 1246–1250.

Balaram Sarkar, Chandresh K Maurya, and Anshuman Agrahri. 2023. Direct speech to text translation: Bridging the modality gap using SimSiam. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 250–255, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. 2019. Task agnostic continual learning using online variational bayes.