CONTINUALLY LEARNING NEW LANGUAGES

Ngoc-Quan Pham¹  Jan Niehues¹  Alex Waibel¹,²

¹Interactive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany
²Carnegie Mellon University, Pittsburgh PA, USA

ABSTRACT

Multilingual speech recognition with neural networks is often implemented with batch-learning, when all of the languages are available before training. An ability to add new languages after the prior training sessions can be economically beneficial, but the main challenge is catastrophic forgetting. In this work, we combine the qualities of weight factorization, transfer learning and Elastic Weight Consolidation in order to counter catastrophic forgetting and facilitate learning new languages quickly. Such combination allowed us to eliminate catastrophic forgetting while still achieving performance for the new languages comparable with having all languages at once, in experiments of learning from an initial 10 languages to achieve 27 languages.

Index Terms: speech recognition, multilingual, transformer, continual learning, incremental learning

1. INTRODUCTION

Neural sequence-to-sequence models are applied to automatic speech recognition with a great success [1] and this model can be easily extended for the multilingual scenario in which the training data consists of multiple languages [2, 3, 4]. These models are able to improve quality especially for the low-resource languages since the acoustic representation can benefit from sharing with other languages. However, the common assumption in such multilingual works so far is that, the datasets for all languages are available before the training process, so that the model has access to the combination of the data for training. The standard batch-training process involves collecting data and sampling batches to the model often tries to balances the amount of data per language given to the model.

In practice, it is also possible that only a subset of the languages is available at first, and the data for new languages can be added after the training process. Moreover, the data for previously trained languages might be discarded for storage or privacy reasons. In such case, the typical batch-training scenario often resorts to two options, either to fine-tune the models on the new datasets to obtain new models that are capable of transcribing new languages, or to combine the old and new datasets to construct new models that fit all languages. Naturally, both of these options are not optimal. Fine-tuning on new languages poses a threat for the previously learned languages to be forgotten, known as catastrophic forgetting [5] that happened when the parameters of the neural networks are shifted towards optimizing the loss function for the new dataset, and far away from the optimal points with respect to the old ones. On the other hand, training all languages together can potentially obtain the best performance for all languages, but is costly since the training time of neural networks can scale depending on the amount of training data. Furthermore it is not possible when the previous languages are no longer available.

The objective of this paper, therefore, is to find new training strategy for multilingual speech recognition in such continual learning scenario. The desiderata of continual learning involves a number of factors:

- Forward transfer: adding new languages to the current multilingual model can ideally obtain the performance similar to when having them in the initial training.
- Backward transfer: catastrophic forgetting is avoided for the previously learned languages, ideally adding the new languages should not affect the performance for the previously learned ones.
- Optimal training cost: the process of learning new languages should be economically better than re-training all languages from the beginning, in terms of training speed and storage.

In order to realize the desiderata, the key idea is to combine three different techniques: transfer learning, weight factorization and elastic weight consolidation, by the following reasons. On the one hand, neural sequence-to-sequence models rely on a fixed vocabulary output, which is likely to be changed when the models are exposed with new languages. Multilingual pretrained language models [6] can relax this problem, by providing the syntax knowledge of the new languages, so that the output layer can be easily applied for new languages. On the other hand, Weight factorization [4] facilitates multilingual models by decomposing each matrix in the neural network into a shared component and per-language additive and multiplicative factors. Motivated by progressive neural networks [7], this modeling scheme can prevent catastrophic forgetting by offloading information in the language
specific weights. Meanwhile, elastic weight consolidation [8] can be used to find the “empty space” in the shared components of the networks to accommodate new languages.

The combination of all three techniques is applied in our experiments with 27 languages involved with the Transformer based model [9][10]. In the process of learning these languages in different iterations, we were able to minimize the effect of catastrophic forgetting for previously learned languages, while the performance of the new languages are still comparable or potentially competitive with training all languages, without the effect of catastrophic forgetting for previously learned languages in different iterations, we were able to minimize the performance of the new languages.

Experiments with 27 languages involved with the Transformer components of the networks to accommodate new languages.

We can use the knowledge of previous tasks to facilitate learning new task is an important topic in machine learning that has been investigated in computer vision or reinforcement learning. There are three common approaches in continual learning: regularization, progressive architecture and replaying from memory. The regularization approach is model agnostic and focuses on designing objective functions that punish weights that tend to be shifted too far from the original positions, where the optimal state with respect to the previous tasks is achieved. The important weights can be identified by importance [8] or memory synapses [11]. Besides, the network can also be designed to isolate the weights and module of each task, while allocating new weights for new tasks [7][12]. It is also possible to store examples of previous tasks as memory replaying [13] to ensure that the gradient updates in the new tasks do not have negative effect over the previous datasets.

In Automatic Speech Recognition, continual learning or incremental learning has been explored in a number of monolingual scenarios. The hybrid HMM models were explored in continual learning by learning different datasets such as World Street Journal, Reverb, Librispeech and Chime4 consecutively [14]. In a similar manner, the sequence-to-sequence model can also be trained on different datasets with the goal of evaluating the performance in each domain after training on another [15]. Recently, the replaying from memory approach has been applied to online continual learning [16] without a clear boundary within task.

Compared to the related works, continual learning new languages in multilingual ASR has a clear task separation due to the difference between languages, compared to monolingual setups. The weight factorization method can be classified into the architectural approach, by assigning new network parameters for new languages. In our work, we combine both architectural and regularization approaches to cover forward and backward transfers in the desiderata.

3. CONTINUALLY LEARNING METHOD

3.1. Weight factorization

Neural networks are composed by layers containing matrices of weights being multiplied with input vectors/matrices to generate output features for the next layers. Therefore, weight allocation per task can be manifested at this fundamental level, by factorizing a basic \( Y = WX \) equation into:

\[
\begin{align*}
Y &= (W_S \odot W_M + W_B)X \\
\end{align*}
\]

where \( W_S \in \mathbb{R}^{D_{in} \times D_{out}} \) is the normal projection weight shared for all languages. \( W_M \in \mathbb{R}^{D_{in} \times D_{out}} \) and \( W_B \in \mathbb{R}^{D_{in} \times D_{out}} \) are the multiplicative and bias terms that are language exclusive.

In order to keep the number of parameters in check as well as encouraging the model to share more information between languages instead of partitioning into the exclusive terms, each language-dependent matrix \( W_M \) or \( W_B \) is further factorized into dot-products of vectors \( r \in \mathbb{R}^{D_{in}} \) and \( v \in \mathbb{R}^{D_{out}} \):

\[
\begin{align*}
W_M &= r_m \odot v_m \\
W_B &= r_b \odot v_b
\end{align*}
\]

The capacity of each factor is controlled via an additional hyperparameter \( k \) that increases the rank of \( W_M \) and \( W_M \) via adding multiple 1-rank matrices:

\[
W = \sum_i r_i \odot v_i
\]

With the value of \( k \ll D_{in} \) or \( D_{out} \), the cost of adding each language is \( \frac{2kD_{in}D_{out}}{D_{in}D_{out}} \) number of parameters, assuming \( D_{in} = D_{out} \).

3.2. Elastic weight consolidation

The main idea of EWC is to add a regularization term based on Bayesian framework seeking to approximate the posterior distribution of the model parameters \( \theta \) conditioned on a prior dataset \( T^0 \) and the current dataset \( T^1 \). The objective function would involve two different terms:

\[
\log p(\theta|D, D^0) = \log p(D^1, \theta) + \log p(\theta|D^0)
\]

The left hand side is simply the negative likelihood of the output sequences given the input acoustic signal. The second term is intractable, and by approximation using a second-order Taylor expansion, we can obtain

\[
\frac{1}{2} \sum_{j=1}^{d} f_i(\theta_j - \theta^{0}_j)
\]

1. Its actually much lower than that, because the network may contain layers that do not need to be factorized, such as the output layer, or layer normalization.
in which $f_i$ is from the diagonal Fisher Information matrix approximating the expected negative Hessian (also impractical to compute). The diagonal Fisher can be considered as the importance metric of the weight given the prior datasets.

After each learning iteration, we can compute the Fisher Information on the new dataset, and add it to the current Fisher \[17\]. As a result, this approach stays at a constant memory cost (equalling the number of trainable parameters in the network).

### 4. EXPERIMENTAL SETUP

**Dataset** The currently available multilingual datasets provided an ideal experimental ground for life-long learning new languages. We use 27 languages from Mozilla Common-voice [18] and Europarl [19] for our language pools as shown in Table 1.

**ASR Model** We used the Transformer model as the base network for speech recognition. The architecture configuration is based on the wav2vec 2.0 model for encoder transfer learning [20] and MBART50 for decoder initialization [6, 21]. As mentioned before, one of the main reasons to use transfer learning, apart from achieving better performance than random initialization [22, 23], is to ensure that the output layer can contain the languages to be added without additional word embeddings. Despite the experiments being limited with the languages covered by the pretrained language models, it still remains practical thanks to the current coverage reaching nearly 200 languages [24].

### 5. RESULTS

#### 5.1. Learning without ‘catastrophic’ forgetting

In Table 1 we presented a simple continual learning scenario with 1 iteration starting from a base model trained with 10 languages. The main purpose of this table is to measure the short term impact of the methods, including the Factorized Weights, EWC and when all languages are present at once.

Without any regularization or weight factorization, the models trained on new languages quickly overfit to the new ones while reaching more than 100% WER for the prior languages. Having EWC can only slow down the catastrophic forgetting process and made the prior error rates 6.5 times faster, while also make learning new languages worse. Meanwhile, weight factorization with frozen shared weights, by nature, totally prevents catastrophic forgetting, so the prior error is almost the same with having all languages (the last column).

The combination of EWC and weight factorization is probably the best in this case. The EWC regularization is the middle ground between freezing all all shared units (limited in learning new languages) and fine-tune everything (catastrophic forgetting since most of the information is still in the main shared weights).

| Lg    | Hrs | E   | E+F | F   | FFT | All  |
|-------|-----|-----|-----|-----|-----|------|
| (de)  | 1050| 53.3| 7.5 | 7.2 | 39.3| 7.2  |
| (fr)  | 800 | 19.7| 11.6| 11.4| 37.4| 11.2 |
| (es)  | 400 | 14.6| 7.1 | 6.8 | 19.9| 6.2  |
| (it)  | 325 | 26  | 7.3 | 6.8 | 24  | 6.5  |
| (fa)  | 293 | 80  | 4.1 | 3.7 | 36.2| 4    |
| (ta)  | 198 | 91.4| 20.5| 18.2| 44.9| 21   |
| (pt)  | 120 | 37  | 7.6 | 7   | 23.1| 6    |
| (ru)  | 148 | 40.5| 6.4 | 5.3 | 26.7| 5.4  |
| (pl)  | 145 | 40.8| 8.4 | 7.7 | 30.9| 7.6  |
| (th)  | 133 | 94.4| 3.5 | 3.2 | 15.5| 3.2  |
| Avg   |     | 49.8| 8.4 | 7.7 | 29.8| 7.8  |
| (nl)  | 150 | 10.6| 7.3 | 7.6 | 7.3 | 6.8  |
| (ar)  | 85  | 24.8| 19  | 17.7| 17.3| 18.2 |
| (zh)  | 63  | 23.6| 14.7| 15.8| 14.7| 14.7 |
| (uk)  | 56  | 21.7| 8.1 | 8.7 | 7.8 | 7.4  |
| (cs)  | 49  | 19.8| 9.2 | 9.6 | 9.4 | 8.4  |
| (ro)  | 45  | 28  | 11.2| 11.7| 11.4| 10.1 |
| (sv)  | 35  | 28.3| 12.4| 13.2| 12.5| 12.3 |
| (et)  | 32  | 31.5| 13  | 14.3| 13.6| 13.2 |
| (tr)  | 30  | 19.6| 7.8 | 8.3 | 7.4 | 8.4  |
| (ja)  | 26  | 20.5| 7.5 | 8.5 | 7.4 | 7.9  |
| (id)  | 23  | 17.6| 6.8 | 6.9 | 6.7 | 6.7  |
| (lt)  | 16  | 56.1| 26.9| 27.7| 27.4| 25.5 |
| (mn)  | 12  | 51.2| 24.5| 25.7| 25.1| 24.3 |
| (sl)  | 9   | 41.3| 10.1| 10.2| 10.5| 9.1  |
| (hi)  | 8   | 70.5| 27.9| 28.9| 28.5| 27.6 |
| (gl)  | 7   | 16.7| 11.7| 10.7| 14.5| 9.8  |

| Avg   | 30.1| 13.6| 15  | 13.8| 13.1|      |

#### 5.2. Continually learning in Multiple iterations

In this experiment, the interest is to measure the degradation rate of the combination of EWC and weight factorization over time, which is qualitatively confirmed in the previous experiment. Therefore, we divided the test languages into three groups, 5 languages in each and measure the degradation after learning in 3 iterations. All of these models in this section contain weight factorization. For the prior languages, their error rates dropped from 8.1 to 11.7 percent in average. This level of performance is still acceptable, but it is 52% higher than the original error rate after training the first 10 languages (7.7%).

Table 1. Performance(WER↓) on the CommonVoice-Europarl testsets. The models included are EWC (E), EWC combined with Factorization (E+F), Factorized with main weights frozen (FFr) and Factorized with all weights fine-tuned (FFt). Models are first trained on 10 languages (top) and then learned on 17 languages (bottom).
### Table 2. Continual learning with EWC and Factorization (EWCF) or Frozen Factorized (FFr) for three iterations (the first session is with the top 10 languages and the next sessions are separated by middle rules.

| Lg  | E+F | F   | E+F | F   | E+F | F   |
|-----|-----|-----|-----|-----|-----|-----|
|     | Iter 1 | Iter 2 | Iter 3 | Iter 1 | Iter 2 | Iter 3 |
| (de)| 7.4  | 7.23 | 8.7  | 7.23 | 10.5 | 7.23 |
| (fr)| 11.5 | 11.4 | 13.1 | 11.4 | 15   | 11.4 |
| (es)| 6.74 | 6.74 | 8.4  | 6.7  | 9.8  | 6.74 |
| (it)| 7.1  | 6.8  | 9.1  | 6.8  | 11.1 | 6.8  |
| (fa)| 4.1  | 3.7  | 4.9  | 3.7  | 6.5  | 3.7  |
| (ta)| 19.7 | 18.2 | 23.3 | 18.2 | 28.5 | 18.2 |
| (it)| 7.3  | 7    | 9.2  | 7    | 11   | 7    |
| (pt)| 6    | 5.3  | 7.4  | 5.3  | 9.6  | 5.3  |
| (th)| 8.1  | 7.72 | 9.4  | 7.72 | 11.3 | 7.72 |
|     | 3.4  | 3.2  | 4    | 3.2  | 4.6  | 3.2  |
| Avg | 8.1  | 7.7  | 9.8  | 7.7  | 11.7 | 7.7  |
| (nl)| 15.9 | 19.4 | 16   | 19.4 | 17.4 | 19.4 |
| (ar)| 14.8 | 17.7 | 15.7 | 17.7 | 16.9 | 17.7 |
| (uk)| 7.9  | 10.4 | 9.1  | 10.4 | 12.6 | 10.4 |
| (cs)| 9.3  | 10.6 | 10.3 | 10.6 | 13.9 | 10.6 |
| Avg | 11   | 13.2 | 11.8 | 13.2 | 14   | 13.2 |
| (ro)| -    | -    | 11.6 | 12   | 12.8 | 12   |
| (sv)| -    | -    | 12.1 | 14.8 | 14.7 | 14.8 |
| (et)| -    | -    | 12.1 | 16.8 | 14.7 | 16.8 |
| (tr)| -    | -    | 7.5  | 9.4  | 9.5  | 9.4  |
| (ja)| -    | -    | 7.5  | 9.5  | 8.3  | 9.5  |
| Avg | -    | -    | 10.2 | 12.5 | 12   | 12.5 |
| (id)| -    | -    | -    | 7.9  | 8    |
| (lt)| -    | -    | -    | 28.5 | 29.3 |
| (mn)| -    | -    | -    | 27.7 | 28   |
| (sl)| -    | -    | -    | 11   | 12.3 |
| (hi)| -    | -    | -    | 29.7 | 30.8 |
| (gl)| -    | -    | -    | 12.3 | 10.9 |
| Avg | -    | -    | -    | 19.5 | 19.9 |

Moving to the new languages in the second set (nl, ar, zh, uk, cs), the error rate of the EWCF model is still lower than the FFr after one lifelong learning iteration (from Iter 1 to Iter 2). Two iterations is required for the Frozen model to surpass this, naturally because one model is inevitably getting worse after each iteration, while the other does not forget despite a worse starting point. This pattern is also observed in the third series participating from Iteration 2. It is notable that certain languages learn significantly better with EWC + Factorization, such as Japanese, Chinese, Swedish and Estonian. It is possible that these languages contain unique compared to the previously learned languages.

In the last iteration, the performance of the new languages with EWC is quite similar than freezing the body of the network. This is the indication that the model has reached the capacity limit that EWC can find. Overall, the results in this table indicated that, for a short amount of iteration the combination of EWC and Factorization is more effective, and the compromise between learning new languages is acceptable. If the model has to learn in many iterations, using only Factorization might be preferable.

#### 5.3. Analysis

We extracted the Fisher information after the initial training stage, and other two iterations in the second set of experiments. For each Fisher diagonal matrix, the number of weights with high Fisher value (with a threshold of 0.5) is counted. For all layers in the network, the number of important parameters is increased over the iterations.

The explanation for the ineffectiveness of EWC probably comes from the derivation into the final equation of the regularization loss term. From the theoretical analysis [17], EWC originates from replacing the log posterior \( \log p(\theta|T_1) \) with its Taylor expansion form, that requires the optimal value \( \theta^* \) during optimizing the model for the data \( T_1 \). The SGD algorithm is not guaranteed to achieve the exact optimal value (averaging checkpoints often yield better model parameters).

The approximation is further "approximated" by the fact that the Hessian in the Taylor expansion is approximated by the diagonal of the Fisher Information matrix. Furthermore, the prior is also assumed to be a zero-mean isometric Gaussian [17] which is rather a simple assumption [25]. From such approximation, it is understandable that EWC might be only effective when the new task/data is somewhat close to the original task which is unlikely in language learning.

#### 6. Conclusion

Learning new languages without forgetting falls into the intersection between automatic speech recognition and continual learning/incremental learning areas. In this paper, we presented a model for this specific task with the combination of Elastic Weight Consolidation and Weight factorization. This model exhibits an interesting property when learning languages sequentially, that it can learn the new languages very effectively compared to other regularized baselines, while maintaining the performance for the previously learned languages to an acceptable extent. We provided an analysis to empirically explain the effectiveness (and ineffectiveness) of EWC and aim at improving the weaknesses of the model to improve further on this task. One of the missing ingredients is to use distillation from the previous models as a compression of the data that is promising, which is shown in many concurrent work in image generation [26].
7. REFERENCES

[1] Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel, “Super-human performance in online low-latency recognition of conversational speech,” 2020.

[2] Yun Zhu et al., “Multilingual Speech Recognition with Self-Attention Structured Parameterization,” in Interspeech, 2020.

[3] Vineel Pratap et al., “Massively multilingual asr: 50 languages, 1 model, 1 billion parameters,” 2020.

[4] Ngoc-Quan Pham, Tuan-Nam Nguyen, Sebastian Stüker, and Alex Waibel, “Efficient Weight Factorization for Multilingual Speech Recognition,” in Proc. Interspeech 2021, 2021, pp. 2421–2425.

[5] Robert M French, “Catastrophic forgetting in connectionist networks,” Trends in cognitive sciences, vol. 3, no. 4, pp. 128–135, 1999.

[6] Yinhua Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer, “Multilingual denoising pre-training for neural machine translation,” Transactions of the Association for Computational Linguistics, 2020.

[7] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell, “Progressive neural networks,” 2016.

[8] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., “Overcoming catastrophic forgetting in neural networks,” Proceedings of the national academy of sciences, 2017.

[9] Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel, “Very Deep Self-Attention Networks for End-to-End Speech Recognition,” in Proc. Interspeech 2019, 2019, pp. 66–70.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017.

[11] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars, “Memory aware synapses: Learning what (not) to forget,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 139–154.

[12] Xin Guo et al., “Continual learning long short term memory,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1817–1822.

[13] David Lopez-Paz and Marc’Aurelio Ranzato, “Gradient episodic memory for continual learning,” Advances in neural information processing systems, vol. 30, 2017.

[14] Samik Sadhu and Hynek Hermansky, “Continual learning in automatic speech recognition,” in Interspeech, 2020, pp. 1246–1250.

[15] Heng-Jui Chang, Hung yi Lee, and Lin shan Lee, “Towards Lifelong Learning of End-to-End ASR,” in Proc. Interspeech 2021, 2021, pp. 2551–2555.

[16] Muqiao Yang, Ian Lane, and Shinji Watanabe, “Online continual learning of end-to-end speech recognition models,” 2022.

[17] Ferenc Huszár, “On quadratic penalties in elastic weight consolidation,” arXiv preprint arXiv:1712.03847, 2017.

[18] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, “Common voice: A massively-multilingual speech corpus,” arXiv preprint arXiv:1912.06670, 2019.

[19] Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan, “Europarl-st: A multilingual corpus for speech translation of parliamentary debates,” in ICASSP, 2020.

[20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[21] Yuqing Tang et al., “Multilingual translation with extensible multilingual pretraining and finetuning,” 2020.

[22] Ngoc-Quan Pham, Alex Waibel, and Jan Niehues, “Adaptive multilingual speech recognition with pretrained models,” in Proc. Interspeech 2022, 2022.

[23] Xian Li et al., “Multilingual speech translation with efficient finetuning of pretrained models,” arXiv preprint arXiv:2010.12829, 2020.

[24] Marta R et al., “No language left behind: Scaling human-centered machine translation,” 2022.

[25] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling, “Improved variational inference with inverse autoregressive flow,” Advances in neural information processing systems.

[26] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” 2021.