ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Sequential Continual Pre-Training for Neural Machine Translation

Niko Dalla Noce, Michele Resta, and Davide Bacciu *

University of Pisa - Computer Science Department
Largo Bruno Pontecorvo, 3, 56127, Pisa - Italy

Abstract. We explore continual pre-training for Neural Machine Translation within a continual learning framework. We introduce a setting where new languages are gradually added to pre-trained models across multiple training experiences. These pre-trained models are subsequently fine-tuned on downstream translation tasks. We compare mBART and mT5 pre-training objectives using four European Languages. Our findings demonstrate that sequentially adding languages during pre-training effectively mitigates catastrophic forgetting and minimally impacts downstream task performance.

1 Introduction

Since the advent of the transformer architecture [1], the field of Natural Language Processing (NLP) has witnessed a significant performance boost, thanks to the development of large pre-trained models like BERT and GPT. The pre-train-fine-tune paradigm has also proven beneficial in Neural Machine Translation (NMT), where it has contributed to enhancing translation quality through training on large amounts of data. The essence of this approach lies in training a model on a general task, such as reconstructing corrupted input, enabling it to learn a versatile hidden representation that can facilitate subsequent tasks. Several pre-training objectives have been proposed for NMT systems, including masking spans of consecutive words within sentences and training the model to reconstruct the original unmasked sentence or words [2, 3]. However, for NMT systems undergoing the pre-train fine-tune pipeline, expanding the range of supported languages can pose challenges and lead to the loss of previously acquired knowledge. This phenomenon, known as catastrophic forgetting (CF), is well-documented in neural networks, and methods and techniques to mitigate it are among the research endeavours of the Continual Learning (CL) field [4]. In this study, we investigate the impact of gradually introducing different languages during the pre-training phase on downstream translation tasks. The research question we want to answer is whether is possible to incrementally add languages to pre-trained models without hurting performances. This is particularly appealing as it would reduce total training time and enhancing energy efficiency, especially for models with a very high parameter count. We adopt an experimental setup akin to the CL class incremental scenario, where the learning process is divided into several experiences, each exposing the model to different

*Work supported by PNRR-M4C2 project FAIR Spoke 1 under the NextGeneration EU programme (PE00000013).

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

languages incrementally. We evaluate three distinct pre-training strategies to quantify the extent of catastrophic forgetting and assess the effectiveness of one of the simplest yet most effective CL strategies. Continual pre-training has been seldom explored in NLP [5], and to the best of our knowledge this is the first work exploring this setting in NMT.

The remainder of this paper is organized as follows: we formalize the problem and the mitigation strategies in Section 2 together with the the experimental setting. Section 3 describes experiment's outcome, together with a discussion of the results. Lastly, we conclude by highlighting several possible future expansions in Section 4.

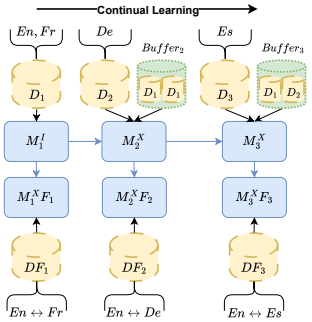
2 Methodology

The primary aim of this study is to evaluate the feasibility of incorporating additional languages during the pre-training phase of NMT models and the performance impact on the downstream translation task. Given the high computational cost associated with pre-training, the prospect of incrementally learning new languages is appealing, as it allows for the avoidance of re-initializing the model and starting a new pre-training process from scratch with new data. We consider four languages for both the pre-training and fine-tuning phases: English, French, German, and Spanish. We divide the training process into four different experiences comprising different languages. At each experience, the model is incrementally pre-trained on the corresponding dataset. After each pre-training experience, the model is fine-tuned on different translation datasets. We constrain the maximum amount of optimization steps for both the pre-training and the fine-tuning phase to be 180k and 100k, respectively. We compare two different pre-training objectives, namely mT5 [6] and mBART [3] to assess whether they exhibit differences in the aforementioned CL setting. The former, derived from T5 [2] consists of masking some text span within the sentences and training the model to reconstruct the missing parts. Instead, the pre-training objective of mBART focuses on reconstructing the entire original sentence. Additionally, mBART uses special tokens at the end of each sentence to indicate its language. We pre-train and fine-tune models according to 3 different training regimes: purely incremental, incremental with replay and jointly.

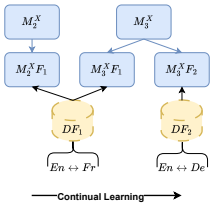
Purely Incremental pre-training (PIPT). In this regime, a model M_1^I is pre-trained using both English and French languages (dataset D_1). At the next experience, we pre-train M_1^I on German (dataset D_2). obtaining M_2^I. For the last experience, we add Spanish (dataset D_3) obtaining the M_3^I model, which has been incrementally pre-trained on all the languages examined. Figure 1a describes the pre-training schema.

Pre-training with replay. The setting matches the PIPT one, this time introducing random replay [7] with a fixed-size buffer. We fix the buffer size to be 5% of the total size of the datasets used in pre-training. At the end of each experience, the buffer is filled with random samples of the current experience in a uniform way ensuring that the buffer is balanced. As in PIPT, we obtain

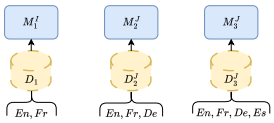
ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) Continual learning strategy.



(b) Control task schema.



(c) Joint training schema.

Fig. 1: On the left (a) the proposed CL schema comprising the pre-training and fine-tuning stages. The control task schema is at the top right (b) and the joint pre-training schema is below it (c). Models are delineated by a continuous blue line border, datasets by a yellow dashed line, and buffers by a green dotted line.

sequentially M_1^R, M_2^R, M_3^R at the end of the respective pre-training experience.

Joint pre-training. In CL scenarios training with all available data is usually considered as an upper bound on performance. We pre-train jointly on a larger set of data obtained by concatenating all the datasets used up to the experience i (fig. 1c). Model M_2^J is pre-trained on dataset D_2^J containing English, French and German languages. Model M_3^J is pre-trained on dataset D_3^J which contains all languages under study.

Fine-tuning. To assess the performance of each pre-trained model (according to the different schemes above) in the translation task, they are fine-tuned on parallel sentences. We denote these datasets as DF_i with i indicating the experience. Each model is fine-tuned only on languages seen during pre-training. Each model is bi-directional: it can translate from source to target and vice versa. Figure 1a illustrates the pre-training and fine-tuning phases combined. With the notation M_i^X, we indicate the exposure of a model M_i to any type of pre-training method, including incremental (I), joint (J), or Replay (R). These pre-trained models have all been fine-tuned following the same methodology.

Control task. The control task is designed to evaluate how much the incremental addition of new languages affects the representation learned by the pre-trained model. In this task we fine-tune the models M_2^X on dataset DF_1, and the models M_3^X on both DF_1 and DF_2, with $X \in \{I, J, R\}$. As a result, we get $M_2^X F_1$ after the former fine-tuning, $M_3^X F_1$ and $M_3^X F_2$ after the latter. We hypothesize that the performance of the fine-tuned model will be worse if the addition of new languages overwrites the models' learned knowledge.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Setup. We use CC100 as a pre-training dataset [8, 9]. We select the first 40 million sentences for each language of interest, for a total of 160 million samples. The fine-tuning datasets are a subset of CCMatrix [10, 11]. This corpus is composed of crawled sentences aligned according to LASER[1] score: a larger score indicates that two sentences are more likely to be a translation pair. For each language pair, we selected the top 35 million according to the LASER score. Each fine-tuning dataset covers both translation directions. We validate English-French and English-German on newstest2013 and use newstest2014[2] as the test set. For English-Spanish, we used a subset of CCMatrix as the validation set and newstest2013 as the test set. We excluded from the training set all the sentences that are also in the validation and test sets. For all the experiments, we use the implementation of the mBART and mT5 models provided by the Huggingface transformers library [12]. Both the models follow the Transformer [1] "base" architecture with 6 encoder/decoder layers with 8 attention heads, and a feed-forward size of 2048. We used tied embedding with a dimension of 512 and set Dropout to 0.1. Models are pre-trained using AdamW optimizer, a linear scheduler with an initial learning rate of 2e-4 for mT5 and 1e-4 for mBART. We select the best model with the lowest perplexity. For the fine-tuning phase, we set the initial learning rate to 4e-4 for mT5 and 6e-4 for mBART. Best models are chosen according to the average of the (sacre-)BLEU[3] scores in both translation directions computed on the validation split. The maximum input sequence length is limited to 128 tokens and the cumulative batch size is 256, distributed across two NVIDIA V100 with 16GB of VRAM each. Pre-train a single model required around 14h, while fine-tuning it approximately 24h.

3 Experimental Results

Table 1 present our experimental result. We report the BLEU scores obtained by fine-tuning pre-trained models on translation datasets. As a reference (REF), we utilize incrementally pre-trained models from the *i-th* experience (M_i^I), fine-tuned with the corresponding translation dataset (DF_i). For instance, in the third experience, the REF model is M_3^I, fine-tuned on DF_3. The most challenging scenario (denoted with *) arises when the pre-trained model from the final experience is fine-tuned on DF_1 as the initial representation learned may be hindered with subsequent updates from new languages. For models pre-trained incrementally without any CL strategy the loss for the most challenging control task (REF*) is 1.02 BLEU points for mT5 on English→French and 0.31 on French→English for mBART. To assess whether pre-training is still effective in this worst-case scenario, we fine-tuned a model on DF_1, starting from random weights, and called it M_{rnd}. As shown in Table 1, this model exhibits the worst performance among those fine-tuned in English-French. This holds for both mBART and mT5, indicating that pre-training remains an effective ap-

[1] `https://github.com/facebookresearch/LASER`

[2] `https://www.statmt.org/wmt14/` - `https://www.statmt.org/wmt13/`

[3] `https://github.com/mjpost/sacrebleu`. ("intl" tokenization type.)

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

proach for enhancing translation performance in this challenging context. Using a replay buffer almost entirely mitigates performance loss when fine-tuning on languages seen in the first pre-training experience, with slight enhancements observed in English ↔ German, Spanish. This applies to both mT5 and mBART models. Jointly pre-trained models exhibit the highest performance drop compared to REF*: 1.18 BLEU points for M_3^J with mT5 objective, while mBART experiences a milder drop of 0.36 BLEU points on average. However, performance in English ↔ German, Spanish remains unaffected, with scores slightly outperforming the benchmark except for mT5, exhibiting decreased score in English → Spanish direction. From the experiments, the pre-training with replay regimes performs best, followed by the joint pre-training approach. This is somewhat surprising as it is usually an upper bound in CL. We hypothesize that language interference may explain this observation. While incremental pre-training exhibits the worst absolute performance, the average loss is negligible for both mBART and mT5: -0.26 and -0.28 BLEU points respectively. Notably, mBART consistently outperforms mT5 in all experiments. These results indicate the robustness of pre-training objectives to catastrophic forgetting, performing strongly even without mitigation strategies.

mBART							
BLEU	$En \to Fr$ (DF_1)	$Fr \to En$ (DF_1)	$En \to De$ (DF_2)	$De \to En$ (DF_2)	$En \to Es$ (DF_3)	$Es \to En$ (DF_3)	Avg.
M_1^I	38.88 (REF*)	35.12 (REF*)	—	—	—	—	—
M_{rnd}	38.18 (−0.70)	34.87 (−0.25)	—	—	—	—	−0.48
M_2^I	38.74 (−0.14)	35.11 (−0.01)	25.35 (REF)	30.63 (REF)	—	—	−0.08
M_3^I	38.74 (−0.14)	34.81 (−0.31)	25.15 (−0.20)	30.25 (−0.38)	33.41 (REF)	32.06 (REF)	−0.26
M_2^R	38.97 (+0.09)	35.37 (+0.25)	26.23 (+0.88)	30.69 (+0.06)	—	—	+0.32
M_3^R	38.87 (−0.01)	35.54 (+0.42)	26.11 (+0.76)	30.83 (+0.20)	33.44 (+0.03)	32.01 (−0.05)	+0.23
M_2^J	38.65 (−0.23)	35.10 (−0.02)	26.04 (+0.69)	30.68 (+0.05)	—	—	+0.12
M_3^J	38.37 (−0.51)	34.91 (−0.21)	25.91 (+0.56)	30.89 (+0.26)	33.41 (+0.00)	32.17 (+0.11)	+0.04
mT5							
M_1	38.94 (REF*)	35.00 (REF*)	—	—	—	—	—
M_{rnd}	36.78 (−2.16)	33.75 (−1.25)	—	—	—	—	−1.71
M_2^I	37.94 (−1.00)	34.80 (−0.20)	26.11 (REF)	31.07 (REF)	—	—	−0.60
M_3^I	37.92 (−1.02)	34.76 (−0.24)	26.13 (+0.02)	31.20 (+0.13)	29.56 (REF)	31.09 (REF)	−0.28
M_2^R	38.44 (−0.50)	34.89 (−0.11)	25.62 (−0.49)	31.07 (+0.00)	—	—	−0.28
M_3^R	38.66 (−0.28)	34.99 (−0.01)	26.14 (+0.03)	31.40 (+0.33)	29.04 (−0.52)	31.23 (+0.14)	−0.05
M_2^J	37.81 (−1.13)	34.75 (−0.25)	26.50 (+0.39)	31.53 (+0.46)	—	—	−0.13
M_3^J	37.76 (−1.18)	35.66 (+0.66)	26.76 (+0.65)	31.60 (+0.53)	23.71 (−5.85)	31.88 (+0.79)	−5.06

Table 1: BLEU score on the test sets obtained by mBART (top) and mT5 (bottom). Values within parentheses are the delta w.r.t. the reference (REF) model, i.e., the one on top of each column.

4 Conclusions

In this study, we explored the impact of continual pre-training on NMT models. In the proposed CL scenario, we demonstrated that the sequential addition of new languages during pre-training incurs in negligible catastrophic forgetting in downstream translation tasks. Notably, both mBART and mT5 objectives showed resilience to forgetting even without traditional CF mitigation strategies like Replay. While our findings are promising on a small scale, further investigation across more languages and experiences could offer compelling avenues for

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

incremental language acquisition without the necessity of starting pre-training from scratch.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 2017.

[2] C. Raffel, N. Shazeer, A. Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21, 2020.

[3] Y. Liu, J. Gu, N. Goyal, et al. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8, 2020.

[4] R. M. French. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In *Advances in Neural Information Processing Systems 6, 7th NIPS Conference, 1993*, 1993.

[5] A. Cossu, T. Tuytelaars, A. Carta, et al. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*, 2022.

[6] L. Xue, N. Constant, A. Roberts, et al. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL-HLT*. Association for Computational Linguistics, 2021.

[7] L. Pellegrini, G. Graffieti, V. Lomonaco, and D. Maltoni. Latent replay for real-time continual learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*. IEEE, 2020.

[8] G. Wenzek, M.-A. Lachaux, A. Conneau, et al. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, 2020. European Language Resources Association.

[9] A. Conneau, K. Khandelwal, N. Goyal, et al. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. ACL.

[10] H. Schwenk, G. Wenzek, S. Edunov, et al. Ccmatrix: Mining billions of high-quality parallel sentences on the web, 2020.

[11] A. Fan, S. Bhosale, H. Schwenk, et al. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48, 2021.

[12] T. Wolf, L. Debut, V. Sanh, et al. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019.