Unsupervised Pretraining for Neural Machine Translation Using Elastic Weight Consolidation

Dušan Variš
Charles University,
Faculty of Mathematics and Physics
Malostranské náměstí 25
118 00 Prague, Czech Republic
varis@ufal.mff.cuni.cz

Ondřej Bojar
Charles University,
Faculty of Mathematics and Physics
Malostranské náměstí 25
118 00 Prague, Czech Republic
bojar@ufal.mff.cuni.cz

Abstract

This work presents our ongoing research of unsupervised pretraining in neural machine translation (NMT). In our method, we initialize the weights of the encoder and decoder with two language models that are trained with monolingual data and then fine-tune the model on parallel data using Elastic Weight Consolidation (EWC) to avoid forgetting of the original language modeling tasks. We compare the regularization by EWC with the previous work that focuses on regularization by language modeling objectives.

The positive result is that using EWC with the decoder achieves BLEU scores similar to the previous work. However, the model converges 2-3 times faster and does not require the original unlabeled training data during the fine-tuning stage.

In contrast, the regularization using EWC is less effective if the original and new tasks are not closely related. We show that initializing the bidirectional NMT encoder with a left-to-right language model and forcing the model to remember the original left-to-right language modeling task limits the learning capacity of the encoder for the whole bidirectional context.

1 Introduction

Neural machine translation (NMT) using sequence to sequence architectures (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017) has become the dominant approach to automatic machine translation. While being able to approach human-level performance (Popel, 2018), it still requires a huge amount of parallel data, otherwise it can easily overfit. Such data, however, might not always be available. At the same time, it is generally much easier to gather large amounts of monolingual data, and therefore, it is interesting to find ways of making use of such data. The simplest strategy is to use backtranslation (Sennrich et al., 2016), but it can be rather costly since it requires training a model in the opposite translation direction and then translating the monolingual corpus.

It was suggested by Lake et al. (2017) that during the development of a general human-like AI system, one of the desired characteristics of such a system is the ability to learn in a continuous manner using previously learned tasks as building blocks for mastering new, more complex tasks. Until recently, continuous learning of neural networks was problematic, among others, due to the catastrophic forgetting (McCloskey and Cohen, 1989). Several methods were proposed (Li and Hoiem, 2016; Aljundi et al., 2017; Zenke et al., 2017), however, they mainly focus only on adapting the whole network (not just its parts) to new tasks while maintaining good performance on the previously learned tasks.

In this work, we present an unsupervised pretraining method for NMT models using Elastic Weight Consolidation (Kirkpatrick et al., 2017). First, we initialize both encoder and decoder with source and target language models respectively. Then, we fine-tune the NMT model using the parallel data. To prevent the encoder and decoder from forgetting the original language modeling (LM) task, we regularize their weights individually using Elastic Weight Consolidation based on their importance to that task. Our hypothesis is the following: by forcing the network to remember the original LM tasks we can reduce overfitting of the NMT model on the limited parallel data.

We also provide a comparison of our approach with the method proposed by Ramachandran et al. (2017). They also suggest initialization of the encoder and decoder with a language model. However, during the fine-tuning phase they use the original language modeling objectives as an additional training loss in place of model regularization. Their approach has two main drawbacks: first, during the fine-tuning phase, they still require the original monolingual data which might not be available anymore in a life-long learning scenario. Second, they need to compute both machine translation and language modeling losses which increases the number of operations performed during the update slowing down the fine-tuning process. Our proposed method addresses both problems: it requires only a small held-out set to estimate the EWC regularization term and converges 2-3 times faster than the previous method.[1]

2 Related Work

Several other approaches towards exploiting the available monolingual data for NMT have been previously proposed.

Currently, the most common method is creating synthetic parallel data by backtranslating the target language monolingual corpora using machine translation (Sennrich et al., 2016). While being consistently beneficial, this method requires a pretrained model to prepare the backtranslations. Additionally, Ramachandran et al. (2017) showed that the unsupervised pretraining approach reaches at least similar performance to the backtranslation approach.

Recently, Lample and Conneau (2019) suggested using a single cross-lingual language model trained on multiple monolingual corpora as an initialization for various NLP tasks, including machine translation. While our work focuses strictly on a monolingual language model pretraining, we believe that our work can further benefit from using cross-lingual language models.

Another possible approach is to introduce an additional reordering (Zhang and Zong, 2016) or de-noising objectives, the latter being recently employed in the unsupervised NMT scenarios (Artetxe et al., 2018; Lample et al., 2017). These approaches try to force the NMT model to learn useful features by presenting it with either shuffled or noisy sentences teaching it to reconstruct the original input.

Furthermore, Khayrallah et al. (2018) show how to prevent catastrophic forgeting during domain adaptation scenarios. They fine-tune the general-domain NMT model using in-domain data adding an additional cross-entropy objective to restrict the distribution of the fine-tuned model to be similar to the distribution of the original general-domain model.

3 Elastic Weight Consolidation

Elastic Weight Consolidation (Kirkpatrick et al., 2017) is a simple, statistically motivated method for selective regularization of neural network parameters. It was proposed to counteract catastrophic forgetting in neural networks during a life-long continuous training. The previous work described the method in the context of adapting the whole network for each new task. In this section, we show that EWC can be also used to preserve only parts of the network that were relevant for the previous task, thus being potentially useful for compositional learning.

To justify the choice of the parameter constraints, Kirkpatrick et al. (2017) approach the neural network training as a Bayesian inference problem. To put it into the context of NMT, we would like to find the most probable network parameters θ, given a parallel data D_{mt} and monolingual data D_{src} and D_{tgt} for source and target languages, respectively:

$$p(\theta|D_{mt} \cup D_{src} \cup D_{tgt}) = \frac{p(D_{mt}|\theta)p(\theta|D_{src} \cup D_{tgt})}{p(D_{mt})} \tag{1}$$

Equation 1 holds, assuming datasets D_{mt}, D_{src} and D_{tgt} being mutually exclusive. The probability $p(D_{mt}|\theta)$ is the negative of the MT loss function and $p(\theta|D_{src} \cup D_{tgt})$ is the result of the unsupervised pretraining. We can assume that during the unsupervised pretraining, the parameters θ_{src} of the encoder are independent of the parameters θ_{tgt} of the decoder. Furthermore, we assume that the parameters of the encoder are independent of the target-side monolingual data and the parameters of the decoder are independent of the source-side monolingual data. Given these assumptions, we can express the posterior probability $p(\theta|D_{src} \cup D_{tgt})$ in the following way:

$$p(\theta|D_{src} \cup D_{tgt}) = p(\theta_{src}|D_{src})p(\theta_{tgt}|D_{tgt}) \tag{2}$$

Probabilities $p(\theta_{src}|D_{src})$ and $p(\theta_{tgt}|D_{tgt})$ are given by the pretrained source and target language models respectively. The true posterior probabilities given by the language models are intractable during fine-tuning, however, similarly to

[1] The speedup is with regard to the wall-clock time. In our experiments both EWC and the LM-objective methods require similar number of training examples to converge.

the work of Kirkpatrick et al. (2017), we can estimate $p(\theta_{src}|D_{src})$ as Gaussian distribution using Laplace approximation (MacKay, 1992), with mean given by the pretrained parameters θ_{src} and variance given by a diagonal of the Fisher information matrix F_{src}. Then, we can add the following regularization term to our loss function:

$$L_{ewc-src}(\theta) = \sum_{i,\theta_i \subset \theta_{src}} \frac{\lambda}{2} F_{src,i}(\theta_i - \theta^{\star}_{src,i})^2 \quad (3)$$

The model parameters not present during the language model pretraining are ignored by the regularization term. Analogically, the same can be applied for the target-side posterior probability $p(\theta_{tgt}|D_{tgt})$ giving a target-side regularization term $L_{ewc-tgt}$.

In the following section, we show that these regularization terms can be useful in a low-resource machine translation scenario. Since we do not necessarily need to preserve the knowledge of the original language modeling tasks, we focus on using them only as prior knowledge to prevent overfitting during the fine-tuning.

4 Experiments

In this section, we present the results of our experiments with EWC regularization and compare them with the previously proposed regularization by language modeling objectives.

4.1 Model Description

In all experiments, we use the self-attentive Transformer network (Vaswani et al., 2017) because it is the current state-of-the-art NMT architecture, providing us with a strong baseline. In general, it follows the standard encoder-decoder paradigm, with encoder creating hidden representations of the input tokens based on their surrounding context and decoder generating the output tokens autoregressively while attending to the source sentence token representations and tokens it generated in the previous decoding steps.[2]

We use Transformer with 6 layers in both encoder and decoder. We set the dimension of the hidden states to 512 and the dimension of the feed-forward layer to 2048. We use multi-head attention with 16 attention heads. To simplify the pretraining process, we use a separate vocabulary for source and target languages, each containing around 32k subwords. We use separate embeddings in the encoder and decoder. In the decoder, we tie the embeddings with the output softmax layer (Press and Wolf, 2017). During both pretraining and fine tuning, we use Adam optimizer (Kingma and Ba, 2014) and gradient clipping. We set the initial learning rate to 3.1, use a linear warm-up for 33500 training steps and then decay the learning rate exponentially. We set the training batch size to a maximum of 2048 tokens per batch together with sentence bucketing for more efficient training. We set dropout to 0.1. During the final evaluation, we use beam search decoding with beam size of 8 and length normalization set to 1.0.

When pretraining the encoder and decoder, we use identical network parameters. We train each language model to maximize the probability of each word in a sentence using its leftward context. To pretrain the decoder, we use the decoder architecture from Transformer with encoder-attention sub-layer removed due to the lack of source sentences. Later, we initialize the NMT decoder with the language model weights and the encoder-attention weights by a normal distribution. We reset all training-related variables (learning rate, Adam moments) during the NMT initialization.

For simplicity, we apply the same approach for the encoder pretraining. In Section 4.2, we discuss the drawbacks of our encoder pretraining and suggest possible improvements. In all experiments, we set the weight λ of each EWC regularization term to 0.02.

The model implementation is available in Neural Monkey[3] (Helcl and Libovický, 2017) framework for sequence-to-sequence modeling.

4.2 Dataset and Evaluation

In our experiments, we focused on the low-resource Basque-to-English machine translation task featured in IWSLT 2018.[4] We used the parallel data provided by IWSLT organizers, consisting of 5,600 in-domain sentence pairs (TED Talks) and around 940,000 general-domain sentence pairs. During pretraining, we used Basque Wikipedia for source language model and News-

[2] For more details about the architecture, see the original paper.

[3] https://github.com/ufal/neuralmonkey

[4] https://sites.google.com/site/iwsltevaluation2018/TED-tasks

		SRC	TGT	ALL
Baseline	15.68	–	–	–
Backtrans.	19.65	–	–	–
LM best	–	13.96	15.56	16.83
EWC best	–	10.77	**15.91**	14.10
LM ens.	–	15.16	16.60	17.14
EWC ens.	–	10.73	**16.63**	14.66

Table 1: Comparison of the previous work (LM) with the proposed method (EWC). We compared models with only pretrained encoder (SRC), pretrained decoder (TGT) and both (ALL). All pretrained language models contained 3 layers. We compared both single best models and ensemble (using checkpoint averaging) of 4 best checkpoints. Results where the proposed method outperformed the previous work are in bold.

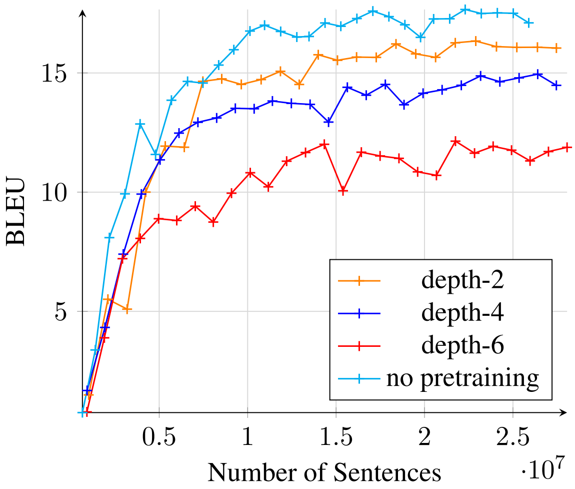

Figure 1: Performance of MT models where only the encoder was initialized by the language model of varying depths and then regularized by EWC. We include the performance of the MT system that was not pretrained for comparison.

Commentary 2015 provided by WMT[5] for target language model. Both corpora contain close to 3 million sentences. We used UDPipe[6] (Straka and Straková, 2017) to split the monolingual data to sentences and SentencePiece[7] to prepare the subword tokenization. We used the subword models trained on the monolingual data to preprocess the parallel data.

During training, we used development data provided by IWSLT 2018 organizers which contains 1,140 parallel sentences. To approximate the Fisher Information Matrix diagonal of the pretrained Basque and English language models, we used the respective parts of the IWSLT validation set. For final evaluation, we used the IWSLT 2018 test data consisting of 1051 sentence pairs.

Table 1 compares the performance of the models fine-tuned using the LM objective regularization and the EWC regularization. First, we can see that using EWC when only the decoder was pretrained slightly outperforms the previous work. On the other hand, our method fails when used in combination with the encoder initialization by the source language model. The reason might be a difference between the original LM task that is trained in a left-to-right autoregressive fashion while the strength of the Transformer encoder is in modelling of the whole left-and-right context for each source token. The learning capacity of the encoder is therefore restricted by forcing it to remember a task that is not so closely related to the sentence encoding in Transformer NMT. Figure 1 supports our claim: the deeper the pretrained language model and therefore more layers regularized by EWC, the lower the performance of the fine-tuned NMT system. We think that this behaviour can be mitigated by initializing the encoder with a language model that considers the whole bidirectional context, e.g. a recently introduced BERT encoder (Devlin et al., 2018). We leave this for our future work.

In addition to improving model performance, EWC converges much faster than the previously introduced LM regularizer. Figure 2 shows that the model fine-tuned without LM regularization converged in about 10 hours, while it took around 20-30 hours to converge when LM regularization was in place. Note, that all models converged after seeing a similar number of training examples, however, computing the LM loss for regularization introduces an additional computation overhead. The main benefit of both EWC and LM-based regularization is apparent here, too. The non-regularized model quickly overfits.

As the comparison to the model trained on the backtranslated monolingual corpus shows, none of our regularization methods can match this simple but much more computationally demanding benchmark.

[5] http://www.statmt.org/wmt18/translation-task.html

[6] http://ufal.mff.cuni.cz/udpipe

[7] https://github.com/google/sentencepiece

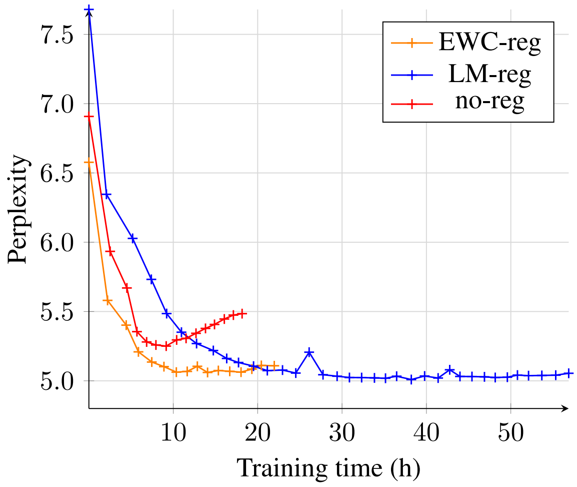

Figure 2: Comparison of relative convergence times (measured by perplexity) of models where only the decoder was pretrained. The models were regularized using EWC, LM objective or were not using any regularization (no reg.). All models were trained on the same number of training examples (∼27M sentences). All used a pretrained LM with 3 Transformer layers.

5 Conclusion

We introduced our work in progress, and exploration of model regularization of NMT encoder and decoder parameters based on their importance for previously learned tasks and its application in the unsupervised pretraining scenario. We documented that our method slightly improves the NMT performance (compared to the baseline as well as the previous work of LM-based regularization) when combined with a pretrained target language model. We achieve this improvement at a reduced training time.

We also showed that the method is less effective if the original language modeling task used to pretrain the NMT encoder is too different from the task learned during the fine-tuning. We plan to further investigate whether we can gain improvements by using a different pretraining method for the encoder and how much this task mismatch relates to the learning capacity of the encoder.

Acknowledgments

This work has been in part supported by the project no. 19-26934X (NEUREM3) of the Czech Science Foundation, by the Charles University SVV project number 260 453 and by the grant no. 1140218 of the Grant Agency of the Charles University.

References

R. Aljundi, P. Chakravarty, and T. Tuytelaars. 2017. Expert gate: Lifelong learning with a network of experts. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 7120–7129.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jindřich Helcl and Jindřich Libovický. 2017. Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*, (107):5–17.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 13:3521–3526.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer.

David J. C. MacKay. 1992. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

Michael McCloskey and Neil J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:104–169.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks*, volume 2, pages 486–491, Stroudsburg, PA, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Prajit Ramachandran, Peter J. Liu, and Quoc V. Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 383–391.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3987–3995.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.