More Parameters? No Thanks!

by

Zeeshan Kha, Kartheek Akella, Vinay P. Namboodiri, C V Jawahar

in

The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)
: 1
-7

Report No: IIIT/TR/2021/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2021

More Parameters? No Thanks!

Zeeshan Khan[1], **Kartheek Akella**[1], **Vinay P. Namboodiri**[2], **C V Jawahar**[1]
[1]CVIT, IIIT-H
[2]University of Bath
{zeeshank606,sukruthkartheek}@gmail.com
vpn22@bath.ac.uk, jawahar@iiit.ac.in

Abstract

This work studies the long-standing problems of model capacity and negative interference in multilingual neural machine translation (MNMT). We use network pruning techniques and observe that pruning 50-70% of the parameters from a trained MNMT model results only in a 0.29-1.98 drop in the BLEU score. Suggesting that there exist large redundancies in MNMT models. These observations motivate us to use the redundant parameters and counter the interference problem efficiently. We propose a novel adaptation strategy, where we iteratively prune and retrain the redundant parameters of an MNMT to improve bilingual representations while retaining the multilinguality. Negative interference severely affects high resource languages, and our method alleviates it without any additional adapter modules. Hence, we call it parameter-free adaptation strategy, paving way for the efficient adaptation of MNMT. We demonstrate the effectiveness of our method on a 9 language MNMT trained on TED talks, and report an average improvement of +1.36 on high resource pairs. Code will be released here.

1 Introduction

Multilingual neural machine translation(MNMT) has seen various advances in recent years (Dong et al., 2015; Firat et al., 2016; Zoph et al., 2016; Tan et al., 2019; Aharoni et al., 2019; Arivazhagan et al., 2019). However, the core principle behind the effectiveness in terms of modelling multiple languages remains the same, i.e., sharing all the model parameters between all the languages (Johnson et al., 2017). Although highly scalable and effective, the performance on high resource languages decreases as more low resource languages are added in the model; this is called negative interference. To overcome this, recent works (Bapna and Firat, 2019; Philip et al., 2020; Zhang et al., 2020) proposed language-specific adapter modules, which provide extra parameters to learn language specific representations, and overcomes the effect of negative interference caused by a high degree of parameter sharing.

In this paper, we propose an alternative to adapter modules. Instead of adding more parameters, we show that the Transformer (Vaswani et al., 2017) has enough capacity to model multiple languages and overcome negative interference effectively. Inspired by the work of Mallya and Lazebnik (2018), we apply iterative pruning to free up the redundant parameters from an MNMT, and retrain them to learn language specific representations. We start with a trained MNMT model, and prune a fraction of the model parameters, we freeze the surviving parameters and retrain the free ones on a bilingual dataset. This process is iteratively applied for each bilingual pair to get bilingual masks over all the model parameters, as illustrated in figure 1. We show that using only a fraction of redundant parameters, significantly improves the performance on high resource languages. Also, we retain the multilinguality and the zero-shot translation ability after adaptation. By demonstrating the effectiveness of this approach, we open a potential research direction towards parameter-free adaptation in MNMT.

2 Related Work

Adding multiple tasks to a single network: Due to the over-parameterized nature of deep neural networks, prior works (Kirkpatrick et al., 2017; Lee et al., 2017; Li and Hoiem, 2017; Triki et al., 2017) aimed at developing methods to learn multiple tasks while avoiding catastrophic forgetting. Mallya and Lazebnik (2018) proposed an iterative pruning approach to free up parameters for adding new tasks and retain the previously trained parameters at the same time. Inspired by the concept,

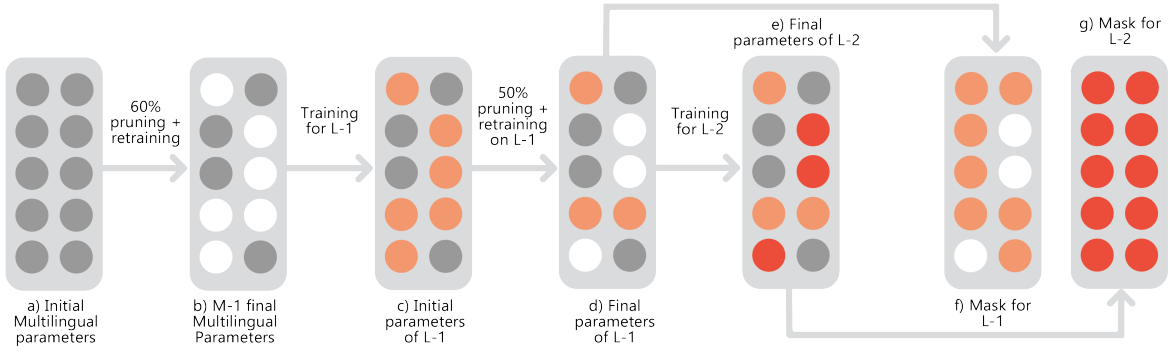

Figure 1: (Better seen in colour.) Illustration of the evolution of model parameters. (a) shows the multilingual parameters in grey. Through 60% pruning and retraining, we arrive at (b), here white represents the free weights with value=0. The surviving weights in grey will be fixed for the rest of the method. Now, we train the free parameters on the first bilingual pair (L-1) and arrive at (c), which represents the initial parameters of L-1 in orange, and share weights with the previously trained multilingual parameters in grey. Again, with 50% pruning and retraining on the current L-1 specific weights in orange, we get the final parameters for L-1 shown in (d) and extract the final mask for L-1 in (f). We repeat the same procedure for all the bilingual pairs and extract the masks for each pair.

we show that an MNMT Transformer model can be heavily pruned and the freed up parameters can be retrained to improve bilingual performance, while retaining the multilinguality.

Adapting multilingual model to a new language pair and domain adaptation: Prior works on adaptation (Neubig and Hu, 2018; Variš and Bojar, 2019; Stickland et al., 2020; Escolano et al., 2020; Akella et al., 2020; Bapna and Firat, 2019; Philip et al., 2020; Zhang et al., 2020) aims at improving language specific performance by either fine-tuning the same MNMT model or adding language specific modules. While being effective, these methods either lose their multilinguality or introduce additional parameters. Sharing the same objective, we propose a method to adapt an MNMT, without adding language-specific modules, while retaining the multilinguality at the same time. Another line of work (Thompson et al., 2018; Wuebker et al., 2018), proposed training of subnetworks and freezing the rest for domain adaptation.

3 Method

The central idea of our method is to use magnitude pruning to free up parameters in the model and learn bilingual specific representations. Figure 1 depicts the evolution of model weights during the training procedure, with (a) representing the initial multilingual weights in grey. We prune away a fraction of parameters using the one-shot magnitude pruning technique (Han et al., 2015), which results in a compressed multilingual representation. We further train the survived multilingual weights for a few more epochs on the multilingual dataset to compensate for extreme pruning, now the multilingual parameters will remain fixed. Then, we use the free parameters to learn the first language-specific representations. We select the first bilingual dataset and train the free parameters. Next we again prune a fraction of weights from the current bilingual parameters only, to accommodate more bilingual representations. We repeat the same procedure for all the existing bilingual pairs. A point to note is that during a forward pass data flows through all the shared and specific weights, while during the backward pass only the current bilingual-specific parameters get updated. Hence, the accuracy is retained for all the previously trained bilingual pairs and it enables a high degree of sharing and specificity at the same time.

Pruning Approach: We perform magnitude pruning (Han et al., 2015) over the weights of all layers. For simplicity, we do not use the more sophisticated pruning methods (Frankle and Carbin, 2019; Michel et al., 2019; Voita et al., 2019). We do not perform pruning over biases and layer normalization parameters, since they correspond to less than 1% of the total parameters, which is insignificant. Also, we do not prune the embeddings, as they are data specific parameters. All are kept fixed after training the multilingual model.

Inference: After finishing the training for each bilingual pair, we get the final mask over all the parameters of the model. Values of the mask range

from $1 \rightarrow N$, where N is the total number of bilingual pairs. Each model parameter is masked according to the bilingual pair of interest. To predict a translation for the t^{th} pair, all the parameters learned for languages $1 \rightarrow t$ will be used, as shown in figure 1(f) and (g).

4 Experiments

4.1 Datasets

We use the TED talks (Qi et al., 2018) in all our experiments, and all the numbers are BLEU (Papineni et al., 2002) scores over the test set[1]. Here we have chosen to train on 8 English centring language pairs[2] en-xx covering a spectrum of sizes from high resource *Ar (Arabic)*, 214K to low resource *Be (Belarusian)*, 4.5K.

4.2 Training

Architecture: We use Transformer architecture (Vaswani et al., 2017), implemented in `fairseq` (Ott et al., 2019), which was modified to include the pruning and masking modules. We train a joint BPE model (Sennrich et al., 2016) on all languages to the vocabulary size of 40K. The Transformer (Vaswani et al., 2017) architecture used in this work[3] has 8 attention heads, 6 encoder and decoder layers, an embedding size of 512, and a feed-forward dimension of 2048. We set the dropout to 0.3.

MNMT Training: We train a standard MNMT model following similar settings as Johnson et al. (2017). A single many-to-many model is trained on all the English-centric data, using a source-side control token to indicate the target language. We use Adam (Kingma and Ba, 2015) with an inverse square root schedule, with 4500 warm-up updates and a maximum learning rate of 0.0003. We set the maximum batch size per GPU to 3050 tokens and train on 4 GPUs. Like Arivazhagan et al. (2019), to avoid the size imbalance, we use the temperature-based sampling strategy with $T = 5$. The MNMT is trained for 40 epochs over 8 English-centric language pairs, i.e., 16 directions. As shown in table 1, we train a strong parent MNMT baseline.

Pruning MNMT: We prune 50% of parameters from a fully converged MNMT model, and retrain the surviving parameters on the same multilingual dataset for ten more epochs, to compensate for the lost parameters.

[1]Scores reported are SacreBLEU (Post, 2018)
[2]`ar, az, be, de, gl, he, it, sk`
[3]`transformer in fairseq`

Adapting MNMT to bilingual specific representations: After pruning the MNMT model, we select each bi-direction datasets (en-xx and xx-en) in the descending order of dataset sizes. We use the original source side control token, reset the learning rate scheduler and train all the free parameters for 20 epochs. Then, we prune 75% of parameters from the current bilingual specific parameters and retrain for ten more epochs to compensate for heavy pruning.

Pruning ratios are decided based on the trade off between the accuracy lost and the space left to adapt all the languages. We prune 50-70% of parameters from the parent MNMT and observe that it leads to a drop of 0.29-1.98 Bleu score. Therefore, we select 50% to be the first pruning ratio, and is kept constant in all the experiments. The second pruning ratio is kept 75% such that the last language pairs get at least 2-5% of parameters. More variations in the second pruning ratio is demonstrated in section 5.4.

5 Results and Discussions

5.1 Overcoming interference for high resource pairs:

In table 1, we present a comparative study of a high resource language scenario, severely affected by negative interference. Adapted MNMT outperforms the parent MNMT on all the 8 directions, with an average improvement of +1.40 on xx-en, and +1.32 on en-xx directions, and closes the gap with high performing bilingual baselines.

Analysing model capacity and negative interference: Now, we expound on the problems of model capacity and interference. As shown in table 1, pruning 50% of parameters from the parent MNMT model leads to an average loss of just 0.29 BLEU points. This observation confirms, that there exists large redundancies even in a 9-language MNMT model. The drop in the performance of an MNMT over its counterpart bilingual models is loosely associated with the lack of capacity. As can be seen in figure 2, by using only a fraction of parameters for each bilingual pair, we can significantly improve the performance over the parent MNMT. Our results demonstrate the ability of parameter-free adaptation to fight negative interference, and improve the performance of severely affected high resource language pairs.

		xx→ en				*en→ xx*			
		Ar	*De*	*He*	*It*	*Ar*	*De*	*He*	*It*
(1)	Aharoni et al. (2019)	27.84	30.50	34.37	33.64	12.95	23.31	23.66	30.33
	Philip et al. (2020)	32.99	37.36	39.00	39.73	**17.22**	**29.94**	**27.47**	35.42
	Our Bilingual	**33.11**	**39.01**	**39.11**	**41.40**	16.79	29.73	26.80	**36.23**
(2)	Aharoni et al. (2019)	28.32	32.97	33.18	35.14	14.25	27.95	24.16	33.26
	Philip et al. (2020)	30.68	36.53	36.00	38.77	15.40	28.60	24.53	34.02
	Parent MNMT	31.33	37.13	36.86	39.54	15.71	26.32	24.60	33.91
	50% Pruned MNMT	30.84	37.10	36.29	39.44	15.41	26.20	24.06	33.70
	Adapted MNMT	32.68	38.41	38.31	41.04	16.72	27.63	25.76	35.76

Table 1: BLEU scores of our models on the TED test sets compared to the literature, (1) - Bilingual baselines. (2) - Multilingual models scores. Here Aharoni et al. (2019) and Philip et al. (2020) are trained on 59 and 20 languages respectively. Parent MNMT is our multilingual model trained till convergence on 9 languages. 50% pruned MNMT is the compressed parent MNMT. Adapted MNMT is the proposed model.

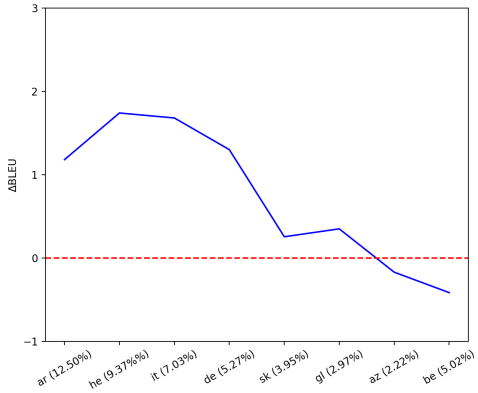

Figure 2: Absolute difference in the BLEU scores, with the parent MNMT, for 8 bilingual pairs. Each bilingual pair is the average over both the English-centric directions. The languages are arranged in the exact order of the training sequence. Numbers on the x-axis are percentages of the bilingual specific parameters used.

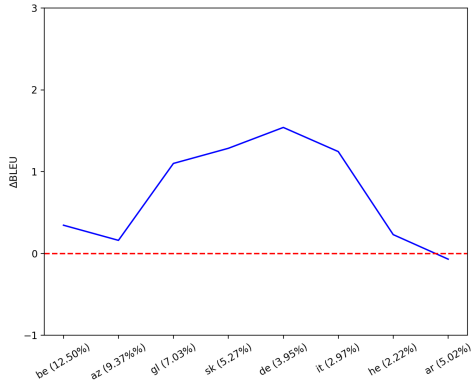

Figure 3: Same as figure 2, trained in the reverse order

5.2 Analysing differences in the adaptation of high and low resource pairs:

To understand the impact of parameter-free adaptation on both the high and low resource language pairs in an unbiased setting. We train two models in opposite orders of adding bilingual pairs. First, we train in the order of high to low resource languages (Ar to Be). Second, we train in the order of low to high resource languages (Be to Ar). Now, we assign the same proportion of parameters, to the high and low resource languages (Ar, He) in case 1, and (Be, Az) in case 2 respectively. As evident from figure 2 and 3, the improvements in Ar and He in case 1 is significantly more, than the improvement in Be and Az in case 2. This observation agrees with the fact that negative interference severely affects the high resource languages in an MNMT, and it needs adaptation to be improved. But, the performance of low resource languages in an MNMT, is already near saturation due to the positive transfer from high resource languages. Hence, to extract the most out of parameter-free adaptation, it is better to prune and retrain the network in the order of high to low resource languages. This assigns high proportion of parameters to high resource pairs, to effectively overcome negative interference.

5.3 Zero-shot Translation:

Zero-shot translation in the context of MNMT, refers to inference between pairs that are not seen directly during the training phase xx-xx. We show that we retain this important ability in our adapted MNMT. Adapted MNMT consists of 50% pruned MNMT weights and 50% language specific weights. The pruned MNMT weights are used to evaluate on zero-shot pairs, just like a traditional MNMT by appending the source side language control token

		xx→ en				*en→ xx*			
		Ar	*He*	*It*	*De*	*Ar*	*He*	*It*	*De*
(Bilingual)	Full-FT	33.89	39.66	41.64	40.00	17.38	27.50	36.72	29.87
	Ar only	33.01	-	-	-	16.80	-	-	-
	Ar-He	33.21	38.26	-	-	16.72	25.52	-	-
(Multilingual)	Ar-He-It	32.99	38.45	41.03	-	16.61	26.00	35.56	-
	Ar-He-It-De	32.68	38.43	41.14	38.39	16.72	25.89	36.08	27.25

Table 2: Full-FT represents the bilingual models derived from finetuning the full parent MNMT. Rest are the adapted MNMTs adapted over 50% free parameters of the pruned MNMT. 1) Ar only with 50% parameters, 2) Ar, He with 25% each, 3) Ar, He, It with 16.6% each, and 4) Ar, He, It, De with 12.5% each.

Johnson et al. (2017). As shown in figure 4, adapted MNMT performs as good as the parent MNMT on all the 56, xx-xx directions even with only 50% of the total parameters.

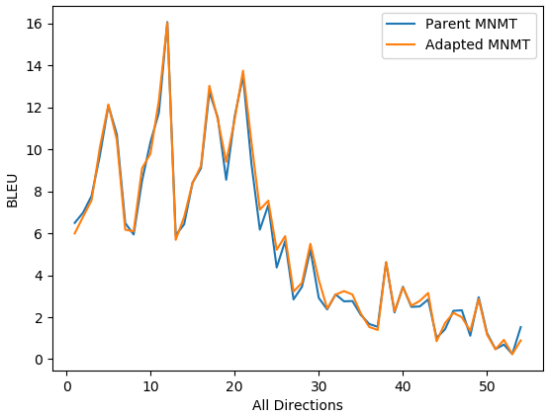

Figure 4: Absolute BLEU scores for the parent and the adapted MNMT on all the 56 zero-shot xx-xx pairs arranged from high to low resource.

5.4 Adapting to a subset of languages and retaining the multilinguality:

Due to limited and fixed number of parameters, we cannot adapt to arbitrary number of languages. However, this framework allows high flexibility in adapting the parent MNMT to only the languages of interest, while retaining the multilinguality simultaneously. We adapt the parent MNMT to four models: 1) Ar, 2) Ar, He, 3) Ar, He, It and, 4) Ar, He, It, De. This way, we can assign all the free parameters to only the languages of interest and increase their capacities. The first pruning ratio is set to 50% for all four models. The second pruning ratio is set such that each language receives equal proportion of parameters. From the results in table 2, we observe that assigning more parameters improve the performance marginally. The four adapted MNMTs have similar performances, even with a significant difference in the proportion of parameters assigned for each language. The 4^{th} model, with only 12.5% parameters reserved for Ar, performs competitively with the 1^{st} model with 50% parameters for Ar. This implies, that a small fraction of parameters can effectively overcome negative interference, hence allowing space to adapt to multiple languages. To infer on the remaining languages which are not adapted, we can use 50% pruned MNMT weights, as done for zero-shot translation in the previous section, hence retaining the multilinguality.

In table 2, we also compare the results of the four adapted MNMTs, with naive finetuning of the full parent MNMT to bilingual pairs (Full-FT). The difference between naive finetuning and the proposed adaptation approach is that the former uses all the 100% of model parameters and the embeddings to adapt to a single bilingual pair, thus the multilinguality is lost. While in our approach, the pruned MNMT weights and the embeddings are fixed, and we only retrain the free parameters very efficiently, allowing to adapt to multiple languages. As can be seen in table 2, adapted MNMTs perform competitively with Full-FT while retaining the multilinguality.

6 Conclusion

We investigate the problems of model capacity and negative interference in multilingual neural machine translation. We show that even a 9 language MNMT has a large proportion of redundant parameters, which are efficiently retrained to overcome interference. We propose a parameter-free adaptation strategy. Where, we use iterative pruning and retraining to improve bilingual representations, without any additional parameters. We hope that our work will attract more attention to practical and efficient ways of adapting an MNMT.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Kartheek Akella, Sai Himal Allu, S. Ragupathi, Aman Singhal, Z. Khan, Vinay P. Namboodiri, and C. Jawahar. 2020. Exploring pair-wise nmt for indian languages. In *arXiv preprint arXiv:2012.05786*. ArXiv.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. In *arXiv preprint arXiv:1907.05019*. ArXiv.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and M. Artetxe. 2020. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *arXiv preprint arXiv:2004.06575*. ArXiv.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*. OpenReview.net.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1135–1143.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

James Kirkpatrick, Razvan Pascanu, Neil 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*, 114(13):3521–3526.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4652–4662.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. In *arXiv preprint arXiv:1606.09282*. ArXiv.

Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7765–7773. IEEE Computer Society.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, pages 14014–14024. Curran Associates, Inc.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

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.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2020. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *arXiv preprint arXiv:2004.14911*. ArXiv.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 124–132, Brussels, Belgium. Association for Computational Linguistics.

Amal Rannen Triki, Rahaf Aljundi, Matthew B. Blaschko, and Tinne Tuytelaars. 2017. Encoder based lifelong learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1329–1337. IEEE Computer Society.

Dušan Variš and Ondřej Bojar. 2019. Unsupervised pretraining for neural machine translation using elastic weight consolidation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 130–135, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Kaiser Lukasz, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 881–886, Brussels, Belgium. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.