Modeling Task-Aware MIMO Cardinality for Efficient Multilingual Neural Machine Translation

Hongfei Xu[1] **Qiuhui Liu**[2] **Josef van Genabith**[1] **Deyi Xiong**[3,4*]
[1]DFKI and Saarland University, Informatics Campus, Saarland, Germany
[2]China Mobile Online Services, Henan, China
[3]Tianjin University, Tianjin, China
[4]Global Tone Communication Technology Co., Ltd.
{hfxunlp, liuqhano}@foxmail.com, josef.van_genabith@dfki.de, dyxiong@tju.edu.cn

Abstract

Neural machine translation has achieved great success in bilingual settings, as well as in multilingual settings. With the increase of the number of languages, multilingual systems tend to underperform their bilingual counterparts. Model capacity has been found crucial for massively multilingual NMT to support language pairs with varying typological characteristics. Previous work increases the modeling capacity by deepening or widening the Transformer. However, modeling cardinality based on aggregating a set of transformations with the same topology has been proven more effective than going deeper or wider when increasing capacity. In this paper, we propose to efficiently increase the capacity for multilingual NMT by increasing the cardinality. Unlike previous work which feeds the same input to several transformations and merges their outputs into one, we present a Multi-Input-Multi-Output (MIMO) architecture that allows each transformation of the block to have its own input. We also present a task-aware attention mechanism to learn to selectively utilize individual transformations from a set of transformations for different translation directions. Our model surpasses previous work and establishes a new state-of-the-art on the large scale OPUS-100 corpus while being 1.31 times as fast.

1 Introduction

Multilingual translation between multiple language pairs with a single model (Firat et al., 2016a; Johnson et al., 2017; Aharoni et al., 2019) has some advantages compared to bilingual systems (Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017; Barrault et al., 2020), e.g., easy deployment, enabling transfer learning across languages and zero-shot translation.

Despite their advantages, multilingual systems tend to underperform their bilingual counterparts as the number of languages increases (Johnson et al., 2017; Aharoni et al., 2019). This is due to the fact that multilingual NMT must distribute its modeling capacity over different translation directions. Zhang et al. (2020) show that the model capacity is crucial for massively multilingual NMT to support language pairs with varying typological characteristics, and propose to increase the modeling capacity by deepening the Transformer.

However, compared to going deeper or wider, modeling cardinality based on aggregating a set of transformations with the same topology has been proven more effective when we increase the model capacity (Xie et al., 2017). In this paper, we efficiently increase the capacity of the multilingual NMT model by increasing the cardinality, i.e. stacking sub-layers that aggregate a set of transformations with the same topology.

Our main contributions are as follows:

- We propose to efficiently increase the capacity of the multilingual NMT model by increasing cardinality, and present a novel MIMO design that allows transformations in the subsequent layer to take different outputs from the current layer as their inputs, unlike previous studies (Xie et al., 2017; Yan et al., 2020) which feed the same input to several transformations and merge their outputs into one;
- We propose to learn a task-aware attention mechanism for the MIMO transformation, allowing the model to weigh different transformations of the set differently for specific translation directions;
- In our experiments on the OPUS-100 corpus, our approach outperforms previous work and

[*] Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 361–367
August 1–6, 2021. ©2021 Association for Computational Linguistics

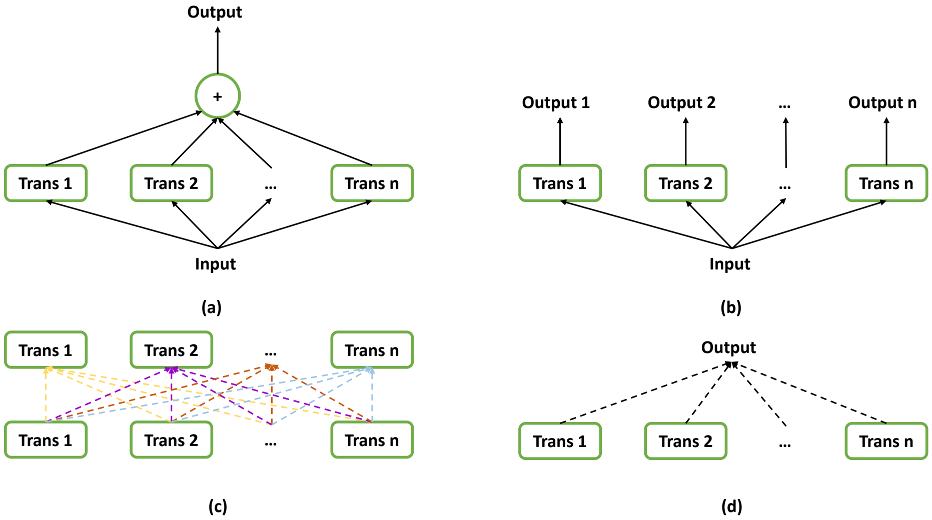

Figure 1: Block transformations. (a) takes the same input into a set of transformations, and adds up their outputs as the output of the block. (b) takes the same input and processes it with these transformations without merging their outputs. (c) is the MIMO architecture that combines weighted outputs of these transformations as inputs to the subsequent transformation set. (d) combines weighted outputs of these transformations into one. Dashed arrows indicate learned attention probabilities. Each "Trans" is a sub-layer that runs in the order of: transforming → dropout → residual connection → layer normalization, where the transforming unit can be either multi-head attention or FFN, as depicted in Figure 2. We aggregate the final output of layer normalization of each "Trans" in the block into the input fed to the next block in different ways (i.e., (a)-(d)).

achieves a new state-of-the-art while being 1.31 times as fast.

2 Preliminaries

Zhang et al. (2020) overcome the capacity bottleneck of multilingual NMT via deepening NMT architectures.

Xie et al. (2017) present a highly modularized network architecture for image classification. The network is constructed by repeating a building block that aggregates a set of transformations with the same topology. For a given input i, the block adopts n networks of the same topology $trans$ to process i and merges their outputs into the final output o of the layer:

$$o = \sum_{k=1}^{n} trans_k(i) \tag{1}$$

This design strategy exposes a new dimension, namely "cardinality" (the size of the set of transformations), as an essential factor in addition to the dimensions of depth and width. Xie et al. (2017) empirically show that increasing cardinality is more effective than going deeper or wider when we increase the capacity to improve classification accuracy.

Yan et al. (2020) present a multi-unit Transformer to efficiently improve the translation performance by increasing cardinality instead of depth. However, their work implements stacks of *input → performing multiple transformations → merging* blocks (as illustrated in Figure 1 (a)), is developed for bilingual sentence-level transformation, and requires the additional design of a biasing module and sequential dependency that guide and encourage complementariness among different units. By contrast, our work aims at efficiently increasing the capacity for multilingual translation, proposes the MIMO transformation (Figure 1 (c)) between stacked blocks, and naturally uses the translation task in attention form to guide individual transformations of the set to learn different representations for different translation directions.

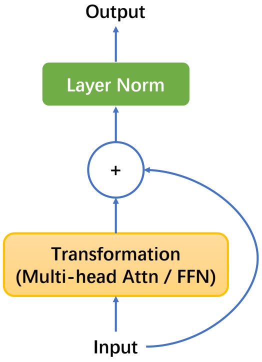

Figure 2: The "Trans" unit.

3 Our Approach

3.1 Multi-Input-Multi-Output (MIMO) Transformation

In contrast to previous approaches (Xie et al., 2017; Yan et al., 2020) that follow a stack of transformation-merging procedures (Figure 1 (a)) to increase cardinality, in our approach we allow our set of transformations to take different inputs. Compared to using the same input, this may encourage transformations to learn complementary representations. Furthermore, merging the outputs of different transformations into one is likely to incur information loss. This is avoided in our approach.

We employ a MIMO transformation between stacked layers (Figure 1 (c)) to enable each transformation of the block to selectively learn to operate on its own unique input.

Specifically, we keep n outputs of the set of transformations to produce multiple inputs for the next layer instead of merging them into one. The input i_k^j to the kth transformation of the jth layer $trans_k^j$ is a weighted accumulation of the outputs o^{j-1} of the layer $j-1$.

$$i_k^j = \sum_{m=1}^{n} p_m^j * o_m^{j-1} \tag{2}$$

where p_m^j are softmax-normalized learnable parameters to model translation task-aware attention for multilingual NMT described in Section 3.2.

o_k^j is produced by $trans_k^j$ with i_k^j as its input:

$$o_k^j = trans_k^j(i_k^j) \tag{3}$$

In the case of a Transformer for multilingual NMT, $trans_k^j$ can be either the multi-head attention or the feed-forward neural network. We adopt a one-to-many transformation (Figure 1 (b)) for the self-attention layer in the first encoder/decoder layer to project one input from the embedding layer to multiple inputs to subsequent layers, and perform a many-to-one transformation (Figure 1 (d)) with the outputs of the feed-forward layer of the last decoder layer to build a single input for the classifier.

3.2 Task-Aware Attention

Rather than separating the multilingual NMT model into 2 parts: 1) the shared part for all language pairs trained on the full dataset; 2) the language isolated part which will only be activated in the corresponding translation task and trained on the part of the whole dataset specifically for the language, we compute all transformations of each block regardless of the translation task, thus all model parameters can utilize and benefit from the whole training set. At the same time, we introduce a task-aware attention mechanism to utilize different transformations of the block differently for specific translation directions.

Specifically, we learn an embedding v for each translation direction (i.e., to X (e.g., en, zh, de)) for each transformation to weightedly aggregate multiple outputs of the block below. v is first normalized into a probability p:

$$p = \text{softmax}(v) \tag{4}$$

Next, p is used in Equation 2 for weighted aggregation. p is expected to assign a higher weight to corresponding transformations of the block which are more important for the translation direction.

3.3 Discussion

Increasing model capacity via increasing cardinality is more efficient than deepening a model or widening it (Xie et al., 2017; Yan et al., 2020). Compared to widening a model, increasing cardinality removes connections between hidden units and reduces both parameters and computation. Compared to deepening a model, increasing cardinality allows to parallelize the computation of all transformations of a set, accelerating both training and decoding.

4 Experiments

4.1 Settings

We conducted our experiments on the challenging massively many-to-many translation task on the OPUS-100 corpus (Tiedemann, 2012; Aharoni et al., 2019; Zhang et al., 2020). We followed Zhang et al. (2020) for experiment settings. We implemented our approaches based on the Neutron implementation (Xu and Liu, 2019) of the Transformer translation model. Parameters were initialized under the Lipschitz constraint (Xu et al., 2020). We adopted BLEU (Papineni et al., 2002) for translation evaluation with the SacreBLEU toolkit (Post, 2018).[1] We report average BLEU over 94 language pairs BLEU$_{94}$, win ratio WR (%) compared

[1] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a +version.1.4.1

Models	Direction	BLEU$_{94}$	WR	BLEU$_4$	BLEU$_{zero}$	Speed-Up
Zhang et al. (2020)	En→xx	23.36	-	19.45	14.08	1.00
	xx→En	30.98		26.78		
Ours	En→xx	**24.17**	78.72	**20.08**	**14.71**	**1.31**
	xx→En	**32.19**	87.23	**27.92**		

Table 1: Main results

Models	BLEU$_{94}$	
	En→xx	xx→En
Full	**24.17**	**32.19**
-MIMO	23.78	31.61
-MIMO-Task Attention	23.54	31.27

Table 2: Ablation on the MIMO and task-aware attention.

#Layers	#Trans.	BLEU$_{94}$	
		En→xx	xx→En
4	6	23.92	31.76
6	4	**24.17**	**32.19**
8	3	24.08	31.94

Table 3: Results of different configurations.

to Zhang et al. (2020), average BLEU over 4 selected typologically different target languages (de, zh, br, te) BLEU$_4$, and average BLEU for zero-shot translation BLEU$_{zero}$.

4.2 Main Results

For fair comparison, we use a 6-layer model where each attention/FFN block contains 4 transformations, which leads to a similar number of parameters compared to the 24-layer model of Zhang et al. (2020). Results are shown in Table 1.

Table 1 shows that our approach achieves better performance in all evaluations while being 1.31 times as fast.

4.3 Ablation Study

We study removing MIMO transformations and task-aware attention. Results are shown in Table 2.

Table 2 verifies that both mechanisms contribute to the performance.

We also examine different combinations of depth and cardinality. Results are shown in Table 3.

Table 3 shows that using 6 layers with 4 transformations in each block leads to the best performance.

Main	en	de	fr	ar	zh	ru
1	rw	sv	pt	he	ja	sh
2	yi	da	it	mt	ko	lt
3	gd	nn	ca	fa	th	sr
4	de	nb	es	ga	vi	mk
5	xh	no	mt	yo	bn	lv

Table 4: Languages with similar task-aware attention weights.

4.4 Task-Aware Attention Weight Analysis

To verify whether task-aware attention learns to aggregate similar languages together, we extract the learned task-aware attention probabilities, flatten them into vectors, and select the languages with the top-5 cosine similarity. Results for several languages are shown in Table 4.

Table 4 shows that close languages are aggregated together.

5 Related Work

Multilingual NMT includes one-to-many (Dong et al., 2015), many-to-many (Firat et al., 2016a) and zero-shot (Firat et al., 2016b) scenarios. A simple solution is to insert a target language token at the beginning of the input sentence (Johnson et al., 2017).

Multilingual NMT has to handle different languages in one joint representation space, neglecting their linguistic diversity, especially for massively multilingual NMT (Aharoni et al., 2019; Zhang et al., 2020; Freitag and Firat, 2020). Most studies focus on how to mitigate this representation bottleneck (Zoph and Knight, 2016; Blackwood et al., 2018; Wang et al., 2018; Platanios et al., 2018; Wang et al., 2019a; Tan et al., 2019b; Wang et al., 2019b; Tan et al., 2019a; Bapna and Firat, 2019; Zhu et al., 2020; Lyu et al., 2020).

There are also studies on the trade-off between shared and language-specific parameters (Sachan and Neubig, 2018; Zhang et al., 2021), on the training of multilingual NMT (Al-Shedivat and Parikh, 2019; Siddhant et al., 2020; Wang et al., 2020b,a), and on analyzing translations from multilingual NMT (Lakew et al., 2018) or the trained model (Kudugunta et al., 2019; Oncevay et al., 2020). Transferring a pre-trained multilingual NMT model can help improve the performance of downstream language pairs (Kim et al., 2019; Lin et al., 2020), especially for low-resource scenarios (Dabre et al., 2019). Multilingual data also has been proven useful for unsupervised NMT (Sen et al., 2019; Sun et al., 2020).

6 Conclusion

We propose to efficiently increase the capacity for multilingual NMT by increasing the cardinality. We present a MIMO architecture that allows each transformation of the block to have its own input. We also present a task-aware attention mechanism to learn to selectively utilize individual transformations from a set of transformations for different translation directions.

Our model surpasses previous work and establishes a new state-of-the-art on the large scale OPUS-100 corpus while being 1.31 times as fast.

Acknowledgments

We thank anonymous reviewers for their insightful comments. Hongfei Xu acknowledges the support of China Scholarship Council ([2018]3101, 201807040056). Josef van Genabith is supported by the German Federal Ministry of Education and Research (BMBF) under funding code 01IW20010 (CORA4NLP). Deyi Xiong is partially supported by the joint research center between GTCOM and Tianjin University and the Royal Society (London) (NAF\R1\180122).

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.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource 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 1410–1416, 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.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. 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.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 550–560, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

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.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Sungwon Lyu, Bokyung Son, Kichang Yang, and Jaekyoung Bae. 2020. Revisiting Modularized Multilingual NMT to Meet Industrial Demands. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5905–5918, Online. Association for Computational Linguistics.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. 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.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Brussels, Belgium. 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.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 2827–2835, Online. Association for Computational Linguistics.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019a. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. Association for Computational Linguistics.

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

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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 5998–6008. Curran Associates, Inc.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019a. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020a. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960, Brussels, Belgium. Association for Computational Linguistics.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019b. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223, Florence, Italy. Association for Computational Linguistics.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020b. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hongfei Xu and Qiuhui Liu. 2019. Neutron: An Implementation of the Transformer Translation Model and its Variants. *arXiv preprint arXiv:1903.07402*.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Jingyi Zhang. 2020. Lipschitz constrained parameter initialization for deep transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 397–402, Online. Association for Computational Linguistics.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2020. Multi-unit transformers for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1047–1059, Online. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

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.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.