Integrating pre-trained language models into Neural Machine Translation

Hoang Do[1,2†], Tu Le[1,2†], Long Nguyen[1,2⋆], and Dien Dinh[1,2]

[1] Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. We revisit established neural machine translation (NMT) models which utilize different pre-trained auto-encoding models and/or auto-regressive models and test their performance against less-resource datasets to determine if there is a significant increase in performance compared to other sequence-to-sequence models. We find that using a multilingual pre-trained auto-encoding model and an auto-regressive model achieves at least competitive results with other Transformer models in the task of translation from English to many Asian languages, even achieving the state-of-the-art result on the IWSLT'15 en-vi dataset without extra training data, while also evaluating NMT models incorporating pre-trained encoders without pre-training. We also raise some questions regarding the potential performance of such models specifically when they are utilized for the purpose of translating in the opposite direction.

Keywords: Neural machine translation · pre-trained models · model incorporation · multilingual translation

1 Introduction

Neural machine translation (NMT) has emerged as a powerful paradigm for automatic translation, achieving remarkable success in capturing the complexities of language and generating fluent translations. The introduction of the Transformer model (Vaswani et al., 2017) marked a significant milestone in the field by introducing a self-attention mechanism that captures global dependencies and enables parallelization, resulting in improved translation quality compared to traditional recurrent neural network-based approaches.

In recent years, research has explored the integration of pre-trained models based on the Transformer architecture into NMT systems (Zhu et al., 2020; Sun et al., 2021; Guo et al., 2020). Auto-encoding models, such as BERT (Devlin et al., 2019) and auto-regressive models such as GPT (Radford et al., 2018), have demonstrated exceptional language understanding and generation capabilities across a wide range of natural language processing (NLP) tasks. The idea is to

[†] These authors contributed equally to this work
[⋆] Corresponding author: nhblong@fit.hcmus.edu.vn

incorporate a pre-trained encoder and/or a pre-trained decoder into a traditional sequence-to-sequence NMT model as an additional component.

While NMT models demonstrate remarkable performance levels in the domain of translation, they encounter various challenges associated with datasets (Koehn and Knowles, 2017; Yang et al., 2020). These challenges encompass issues like inadequate training data availability (Koehn and Knowles, 2017) and the risk of overfitting when training large models with limited datasets (Yang et al., 2020), particularly in the context of low-resource language (LRL) pairs. In this paper, we investigate how pre-trained models can address these problems and how they contribute to the task of NMT. We conduct experiments with two established models which incorporate pre-trained encoders and/or decoders, BERT-NMT (Zhu et al., 2020) and Graformer (Sun et al., 2021). In summary, we make the following contributions:

1. In our study, we address the challenge of incorporating a pre-trained model into an NMT model and show how to overcome it through the utilization of an enhanced pre-trained model. To assess the performance, we conduct a thorough comparison between NMT models that integrate pre-trained encoder/decoder models and traditional NMT models. Our evaluation is based on the IWSLT'15 English-Vietnamese dataset (Cettolo et al., 2012).
2. Our results demonstrate that the pre-trained multilingual incorporated model surpasses the performance of the current state-of-the-art model on a language pair that the incorporated model has not encountered during the pre-training process without using any additional method.
3. Furthermore, our experiments on the ALT dataset (Thu et al., 2016), which comprises Asian low-resource language (LRL) pairs, reveal remarkable achievements by the pre-trained multilingual incorporated model. Notably, the model performs well on several language pairs that did not previously train, and it achieves these impressive results without showing signs of overfitting.

The remainder of this paper is organized as follows: Section 2 presents an overview of related work on incorporating pre-trained models into NMT architectures; Section 3 describes the methodology; Section 4 describes experimental setup, results and the analysis of our experiments; Section 5 discusses the limitations of our further experiments and suggests potential avenues for further research; and finally, Section 6 concludes the paper with a summary of our contributions and future directions.

2 Related works

2.1 Neural machine translation

The goal of NMT is to translate from a source language sequence into its equivalent in the target language. The popular approach to this sequence-to-sequence problem is using an encoder-decoder model architecture (Sutskever et al., 2014; Cho et al., 2014) that contains an encoder translating the input sequence to a hidden representation and a decoder translating this representation into the predicted output sequence. (Vaswani et al., 2017) proposed the self-attention mechanism along with Transformer model and achieved state-of-the-art in NMT tasks with a lower training cost than previous RNN-based models. Transformer is also the base model for many recent pre-trained models which were trained using only the encoder (Devlin et al., 2019; Liu et al., 2019), decoder (Radford et al., 2018), or the whole Transformer model (Raffel et al., 2020b) with more layers and massive corpora, which we will review in subsection 2.2 and show how they are used in the NMT task in subsection 2.3.

2.2 Pre-trained models

Pre-trained language models are trained on large amounts of unlabeled text data. This allows them to learn the universal representations of language. When these models are fine-tuned for specific downstream tasks, such as sentiment analysis or question answering, they can benefit from the knowledge they have learned from the unlabeled data. Two well-known examples of pre-trained language models are BERT (Devlin et al., 2019) and GPT (Radford et al., 2018).

BERT provides deep contextualized representations of words by leveraging the bidirectional context information from large-scale unlabeled corpora. These contextual embeddings capture fine-grained syntactic and semantic information, enabling a more accurate understanding and representation of source and target language sentences. The emergence of BERT has received much attention, leading to some BERT-based models with improvements. Liu et al. (2019) proposed RoBERTa, which used the dynamic masking strategy when training on bigger datasets and achieved state-of-the-art on a variety of NLP tasks. LIMIT-BERT was introduced in (Zhou et al., 2020), which can capture the linguistic information by using a linguistics-guided mask strategy.

On the other hand, the GPT model ocuses on generative tasks and displays exceptional proficiency in producing text sequences that are both coherent and contextually relevant. Through extensive training on massive and varied corpora of textual data, GPT learns to model the probability distribution of words and their interdependencies. This, in turn, empowers GPT to generate translations that are not only fluent but also maintain a high level of coherence.

2.3 Pre-trained models for NMT

Pre-trained models can capture the linguistic knowledge of target or source languages and improve downstream models without training from scratch. After training on monolingual corpora of source languages, BERT can be used to initialize the parameters of the encoder of the NMT model (Clinchant et al., 2019; Rothe et al., 2020) or provide a pre-trained contextual embedding model for the NMT model (Zhu et al., 2020). The BERT-fused model (or BERT-NMT) from (Zhu et al., 2020) extracts the representation from the input sentence and fuses it with the encoder and decoder of the NMT model through additional attention models. T5 (Raffel et al., 2020a) is also a recent pre-trained model that uses an encoder-decoder Transformer architecture. This architecture allows T5 to handle generative NLP tasks like machine translation. Sun et al. (2021) conducted the Graformer model by efficiently integrating both BERT and GPT into the NMT model on multilingual translation tasks. In order to investigate the impact of pre-trained models on the overall performance of NMT systems, we select Graformer and BERT-NMT as our chosen NMT models that integrate pre-trained encoder and/or decoder models. By incorporating these pre-trained models, we aim to delve into the influence and effectiveness of such models on the NMT task, thereby gaining insights into their contribution toward enhancing the performance of the entire NMT system.

2.4 Low-resource language pairs problem

The challenge of low-resource languages (LRL) in the field of NMT is not a new one. It is often difficult to find a substantial parallel corpus for each language pair, and when training on small datasets, large NMT models can easily be overfitting. To address this issue, various approaches have been proposed. Neubig and Hu (2018) introduced a similar-language regularization method, which involves training multilingual NMT models using both an LRL and a similar high-resource language (HRL) to mitigate overfitting. Lakew et al. (2019) expanded on this by proposing a data selection method that identifies HRLs similar to the LRL based on perplexity, and they also adopted the vocabulary from a pre-trained multilingual model trained on the LRL. Nevertheless, considering the diversity of Asian languages, we made the decision to refrain from employing these methods in our study. Instead, our approach involved the fine-tuning of the pre-trained model exclusively for the English-X language pair.

3 Method

In order to comprehensively assess the influence of incorporating pre-trained models into the NMT model on translation performance, we conducted a series of experiments using BERT-NMT, Graformer, as well as their modified variants on the IWSLT'15 English-to-Vietnamese dataset.

To establish a fair and unbiased comparison between the BERT-NMT models and the Transformer baseline, our approaches do not employ warm-start training to the NMT module, contrary to the approach taken in the original paper by Zhu et al. (2020). Additionally, we explore the effects of different enhanced versions of BERT by substituting the BERT module of the BERT-NMT models with BERT, RoBERTa, and LIMIT-BERT, enabling us to thoroughly investigate the effects and implications of these variants on the overall performance of the models.

In our experiments on the Graformer models, we utilize the multilingual Graformer checkpoint from the paper's repository[†]. This multilingual Graformer model was trained on multilingual NMT tasks, thereby efficiently benefiting from

[†] https://github.com/sunzewei2715/Graformer

grafting techniques involving pre-trained multilingual BERT and GPT models, known as mBERT and mGPT. Additionally, we conducted an experiment on Graformer where only the grafting decoder was pre-trained, aiming to explore the specific contribution of the pre-trained decoder model. Considering the substantial parameter disparity between the Graformer models and both the Transformer baseline and BERT-NMT models, we conducted two comparative analyses. Specifically, we compared the Graformer model with the EnViT5-base model (Ngo et al., 2022) and the NLLB-200 model (NLLB Team et al., 2022), both of which possessed a comparable number of trainable parameters to the Graformer model and trained on multilingual NMT tasks. This comparison was crucial in elucidating the distinct contributions of integrating the pre-trained models into NMT models in our research.

Subsequently, we select the pre-trained model that attains the highest performance (e.g., based on BLEU score) from the initial task, then proceed with further experiments involving low-resource languages and languages that did not previously train. For this purpose, we opt for the ALT dataset (Thu et al., 2016) due to its inclusion of relatively smaller language pairs, making it an ideal choice for assessing the pre-trained model's generalization capabilities within the linguistically diverse Asian context. Once the model is chosen, we independently fine-tune and evaluate its performance in each English-to-X language pair, where X corresponds to a language present in the ALT dataset.

4 Experiment setup

4.1 Datasets

The experiment is first conducted using the default split of the IWSLT'15 English-Vietnamese dataset (Cettolo et al., 2012), which contains over 133000 sentences from TEDxTalk speech transcripts. We follow that with evaluating on the Asian Language Treebank (ALT) dataset (Thu et al., 2016), currently containing 20.106 sentences from English Wikinews, which is then translated into 12 other languages: Bengali, Filipino, Hindi, Bahasa Indonesia, Japanese, Khmer, Lao, Malay, Burmese, Thai, Vietnamese and Simplified Chinese. We use the standard split (18088, 1000, and 1018 sentences for the training, development, and test set respectively) to perform experiments in translation from English to each of the other 12 languages and report our results in the test set. BERT-NMT (Zhu et al., 2020) did not give us a desirable result (due to severe overfitting) on ALT so we omit it from this part of the experiment. This gives us more insight on the performance on low-resource environments.

4.2 Implementation and training details

Graformer (Sun et al., 2021) We use the code implementation of Graformer (Sun et al., 2021) and as-is, which is built on top of Fairseq (Ott et al., 2019) for training and evaluation. We fine-tune the original Graformer, whose encoder and decoder have 12 layers each with the model dimension of 1024 and intermediate dimension of 4096, separately on ALT and IWSLT'15 datasets. The model is fine-tuned until no significant increase in the maximum BLEU score (Papineni et al., 2002) (> 0.01) for 5 consecutive epochs is found on both ALT and IWSLT'15. Adam (Kingma and Ba, 2017) is used as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, 4000 warmup steps and initial learning rate $\eta = 10^{-3}$ combined with the inverse square root scheduler (Raffel et al., 2020b). We adopt a dropout rate of 0.1, unlike the original paper, which uses 0.3. We conduct the beam search with beam size 5 on both datasets for inference and use SacreBLEU (Post, 2018) for measuring case-sensitive corpus BLEU score. Following the original paper, we employ SentencePiece (Kudo and Richardson, 2018) as our tokenizer, utilizing a vocabulary size of 64000 tokens. Given the relatively low number of words in the ALT dataset that require substitution with unknown tokens, we retain the vocabulary from the original checkpoint rather than adapting it specifically to this dataset, allowing to consistently use this checkpoint in every translation task for each language of the ALT dataset.

Since there are 8 languages in the ALT dataset that Graformer is not trained on (such as Vietnamese, Lao, and Thai), we employ the multilingual translation scheme for languages pre-trained in Graformer, while the bilingual translation scheme is used for languages that the model is not trained on.

BERT-NMT (Zhu et al., 2020) For the BERT-NMT model and its variants mentioned in section 2.2, we follow the configuration from the training example from (Zhu et al., 2020), using a Transformer-base (Vaswani et al., 2017) model without pre-training as the core and change BERT (Devlin et al., 2019) with various pre-trained auto-encoding models, more specifically RoBERTa-base Liu et al. (2019) and LIMIT-BERT Zhou et al. (2020). The lack of pre-training is in order to observe performance in low-resource environments. We train the model with the same hyperparameters as Graformer (see above) till no significant increase in the maximum BLEU score for 5 consecutive epochs. For the inference phase, we utilize beam search with a size of five.

4.3 Baselines

On the IWSLT'15 en-vi dataset The Transformer-base model with BPE-Dropout sub-word regularization algorithm (Provilkov et al., 2020) is used as the baseline for the general comparison for BERT-NMT variants and two Graformer models, where one with pre-trained weights across all parameters, and the other one only utilizing the pre-trained decoder.

In addition, we compare the evaluation results of pre-trained multilingual translation Graformer fine-tuned on the IWSLT'15 en-vi dataset with the results of the following baseline models:

- EnViT5-base (Ngo et al., 2022): a NMT model based on the Text-To-Text Transfer Transformer (T5) (Raffel et al., 2020b) architecture, built specifically for English-Vietnamese translation. The model is initially trained on

CC100 (Conneau et al., 2020) then fine-tuned on PhoMT (Doan et al., 2021) and MTet (Ngo et al., 2022). Further fine-tuning of EnViT5-base on IWSLT'15 leads to the current state-of-the-art result[†].

– NLLB-200 (NLLB Team et al., 2022): a massively multilingual NMT model trained on FLORES-200 (NLLB Team et al., 2022), containing 200 languages. The reason we chose this model as a baseline is because, to the best of our knowledge, it achieved the best performance among those where no extra training data was used prior to the evaluation on IWSLT'15. This is aligned with our approach of fine-tuning Graformer, enabling a more objective comparison between Graformer and other NMT models.

On the ALT dataset We utilize the findings from (Dabre et al., 2019) as our baseline. The study examined the performance of the Transformer model (Vaswani et al., 2017) by training it separately on IWSLT'15 en-zh (Cettolo et al., 2012) and KFTT en-ja (Neubig, 2011) datasets, followed by multiple stages of fine-tuning on ALT. From the provided results, we select only the ones with the highest average BLEU score for both models.

4.4 Main results and Analysis

Model	BLEU	#Params (#Trainable)
Transformer with BPE-Dropout (baseline)	33.27	65M (65M)
BERT-NMT	25.82	174M (65M)
BERT-NMT + RoBERTa	26.34	189M (65M)
BERT-NMT + LIMIT-BERT	24.70	403M (68M)
Graformer (pre-trained mGPT only)	13.92	526M (241M)
Graformer	**47.55**	526M (241M)

Table 1. Results of en→vi translation on the IWSLT'15 dataset of the Transformer baseline with BPE-Dropout (Provilkov et al., 2020), original BERT-NMT, its variants where BERT is replaced with RoBERTa/LIMIT-BERT (indicated by the plus sign) and two versions of Graformer, which one is only pre-trained in the grafting decoder (using the checkpoint of mGPT) and another one is the whole multilingual Graformer checkpoint. **Bold** represents the highest BLEU score.

On the IWSLT'15 en-vi dataset Table 1 presents the BLEU scores on the test set of the IWSLT'15 en-vi dataset for various models: the Transformer baseline from Provilkov et al. (2020), BERT-NMT variants with additional encoders such as BERT, RoBERTa, LIMIT-BERT, and Graformer models (Sun et al., 2021). The Graformer models consist of one variant pre-trained only on the grafting decoder and the other that was fully pre-trained across all parameters.

[†] https://research.vietai.org/mtet/

Despite the increased complexity of the BERT-fused models and Graformer with a pre-trained decoder, they achieve inferior results compared to the Transformer baseline. This can be attributed to the fact that these pre-trained NMT models were not trained on any parallel corpus, limiting their ability to effectively utilize the knowledge from the incorporated pre-trained encoder/decoder. Consequently, they face the challenge of both integrating the pre-trained components into the core NMT model and learning the translation task simultaneously, whereas the Transformer baseline solely focuses on translation. Another factor is that models without training on any parallel corpus tend to overfit as the number of parameters increases. LIMIT-BERT, despite having a similar size to BERT-large, obtained a lower BLEU score compared to BERT-NMT models, which utilize BERT-base models (BERT and RoBERTa) as the BERT module. This discrepancy becomes more apparent with Graformer, as the variant pre-trained on the grafting encoder, is the largest experimented model with three times the number of trainable parameters compared to the baseline, yielding the worst result, which is twice as low as the baseline.

Interestingly, when fine-tuned from a multilingual Graformer checkpoint, Graformer achieves the highest BLEU score of 47.55, surpassing the baseline score of 14.28. To ensure that this improvement is a result of incorporating pre-trained models rather than solely due to its larger capacity and multilingual pre-training process, we conduct two comparisons with the NLLB-200 model and the EnViT5-base model in this dataset. Both of these models possess comparable numbers of trainable parameters to the Graformer model, with EnViT5 trained on the Vietnamese dataset before testing on the IWSLT'15 English-to-Vietnamese dataset, and NLLB-200 being a multilingual NMT model trained on a dataset comprising 200 languages, while multilingual Graformer checkpoint just trained on 45 languages that do not include Vietnamese.

Model	#Params (#Trainable)	Extra training data		BLEU
		Dataset	#Pairs	
EnViT5-base	275M (275M)	PhoMT + MTet	6.2M	40.2
NLLB-200	**250M** (250M)	-	0	34.8
Graformer	526M **(241M)**	-	0	**47.6**

Table 2. Results of en→vi translation on the IWSLT'15 dataset for further comparisons with Graformer. The number of parameters in parentheses is the number of trainable parameters (those which are not frozen). **Bold** represents the most beneficial numerical figures in each column (lowest for the total number of parameters and the number of trainable parameters, highest otherwise).

Table 2 displays the BLEU scores of Graformer, NLLB-200, and EnViT5-base (Ngo et al., 2022) on the test set of the IWSLT'15 en-vi dataset. It also provides the total number of parameters and trainable parameters for each model. Despite Graformer having the largest parameter count, with 526M in total, only approximately half of them (241M) are trainable, which is still lower compared to NLLB-200 and EnViT5-base, both exceeding 250M parameters. Notably, while EnViT5-base is additionally trained on two English-Vietnamese translation datasets, Graformer outperforms EnViT5-base without any exposure to Vietnamese data during the pre-training process. This achievement represents a new state-of-the-art in English-to-Vietnamese translation on the IWSLT'15 dataset, with a notable improvement of 7.4 BLEU score. Furthermore, Graformer's performance surpasses that of NLLB-200, which is a pre-trained multilingual model in 200 languages, including both Vietnamese and English, by 12.8 BLEU. The remarkable improvement of Graformer can be attributed to its capacity and pre-training on multilingual translation tasks, as well as its ability to leverage linguistic knowledge from incorporated models. The multilingual training scheme enables Graformer to effectively extract language information from mBERT and mGPT, thereby enhancing translation tasks.

Model	Trained languages				Untrained languages							
	bn	hi	ja	zh	id	khm	lo	ms	my	th	vi	fil
Transformer en-zh	10.30	N/A	20.08	N/A	27.24	28.66	N/A	33.19	N/A	28.22	35.34	N/A
Transformer en-ja	10.77	N/A	22.60	N/A	28.89	30.03	N/A	34.75	N/A	28.62	37.06	N/A
Graformer	**29.12**	36.93	**30.17**	31.46	**45.67**	**34.10**	41.92	**52.47**	50.95	**47.31**	**52.59**	40.55

Table 3. Result of en→x translation direction on the ALT dataset (Bengali (bn) is denoted *bg* in ALT), where Transformer en-zh is the Transformer model pre-trained on IWSLT'15 en-zh dataset and Transformer en-ja is pre-trained on KFTT en-ja dataset, both of which are then fine-tuned for multiple stages as described in (Dabre et al., 2019). The best BLEU score for each target language is highlighted in bold. Languages without available results from (Dabre et al., 2019) are excluded from the comparison, however presented nonetheless for potential future evaluations.

On the ALT dataset The remarkable BLEU score achieved by Graformer in English-Vietnamese translation, despite not being trained on any Vietnamese corpus beforehand, raises an intriguing question. We wonder if a pre-trained multilingual NMT model, with its incorporated encoder/decoder, can capture general language features and adapt to languages that are previously untrained or have limited resources (LRLs). To explore this, we conducted an experiment on the ALT dataset, which includes LRLs, some of which the Graformer checkpoint is not trained on. The training results of Graformer (Sun et al., 2021) on the ALT dataset are presented in Table 3. Graformer demonstrates a significant improvement over the baselines for all tested language pairs, with an average BLEU increase of 14.09. Even in languages where baseline results were not available (Dabre et al., 2019), Graformer achieves high performance. The largest increase was observed in the en→th language pair, surpassing the best baseline by 18.69 BLEU points, and the smallest increase was seen in the en→khm language pair, with a BLEU score 4.07 higher than the best baseline. Notably, Graformer achieves BLEU scores of 50 or higher in Malay (50.95), Burmese (52.47), and

Vietnamese (52.59). These findings provide a clear answer to our earlier question: training a model that incorporates pre-trained models in a multilingual translation task empowers the model to extract essential general language features, thereby mitigating overfitting and facilitating adaptation to untrained languages, even when data availability is limited.

5 Limitations

Our experiments focus on the translation direction from English to other Asian languages. Although we have not extensively examined the opposite direction, as we have raised a question regarding how Graformer achieves remarkable BLEU scores even in languages where the pre-trained encoder, decoder, and the model in general are not specifically trained, we trained Graformer on the vi→en translation direction using the ALT dataset, which has led to intriguing findings.

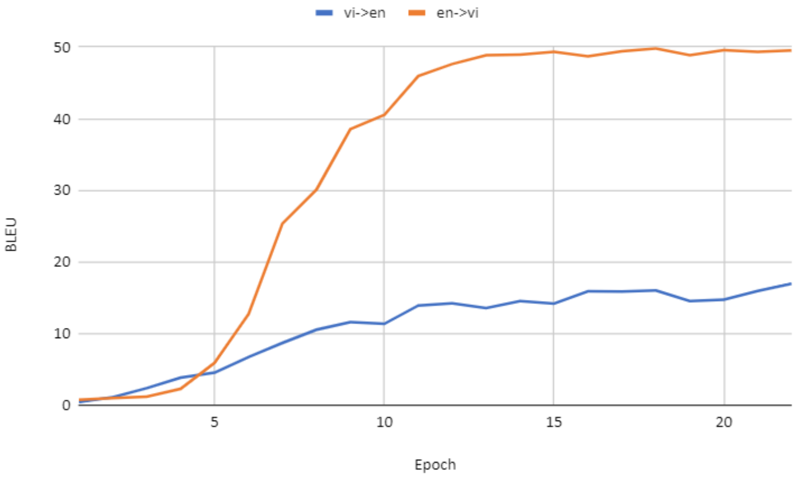

Fig. 1. The graph demonstrating BLEU scores of Graformer on ALT en-vi, on both directions (en→vi and vi→en) after each of the first 22 epochs of training.

As demonstrated in figure 1, on the en→vi direction, the test split BLEU score quickly reached sub-50 over 22 epochs (and achieved over 50 after epoch 23). This observation holds true when training Graformer on other languages with similar translation directions. On the other hand, it achieves below 20 BLEU (16.95) in the opposite direction (vi→en) after the same number of epochs, which did not improve after. And while ja→en translation also achieves a lower result than en→ja, the discrepancy is much less apparent (21.38 compared to 30.17). This phenomenon can be generalized on other language pairs as demonstrated in table 4, where with the exception of fil↔en, there are no pairs between

English and an untrained language which performs equally well as other pairs between English and other pre-trained languages in terms of BLEU discrepancy between the two translation directions. Most notably, Lao, Burmese and Thai achieved astonishingly low BLEU scores, all of which were under 10. Note that we trained Graformer on x→ en tranlation in the same manner as the original setup mentioned in Section 4. We have a suspicion that training NMT models from the aforementioned languages to English may converge at a slower rate compared to the opposite direction, particularly when the pre-trained encoder handles untrained languages. This raises questions about the potential performance of the model if more resources from other languages are included in the pre-training process. Additionally, since Graformer demonstrates proficiency in dealing with untrained languages as translation targets (but not the other way around), further investigation is warranted to explore the capabilities of well-trained autoregressive models incorporated into NMT systems.

Direction	Trained languages				Untrained languages							
	bn	hi	ja	zh	id	khm	lo	ms	my	th	vi	fil
en→x	29.12	36.93	30.17	31.46	45.67	34.10	41.92	52.47	50.95	47.31	52.59	40.55
x→en	15.03	25.50	21.38	20.84	29.40	14.11	2.07	31.93	0.76	7.08	16.95	28.53
BLEU difference	14.09	11.43	8.98	10.62	16.27	19.99	39.85	20.64	50.19	40.23	35.64	12.02

Table 4. Result of en→x and x→en translation directions on the ALT dataset on Graformer, along with the differences in BLEU scores in-between.

Furthermore, the training process on IWSLT'15 en-vi and ALT also raises some interesting questions that we currently are not able to explain. Training Graformer on IWSLT'15, we achieved the highest performance (47.55 BLEU) after 5 epochs, and this score was not surpassed even after another 10 epochs of training. However, it took 67 epochs of training on ALT from English to Vietnamese, despite ALT being much smaller in size, to achieve the highest BLEU score on its test set. This phenomenon was also observed when training on other languages, with Japanese requiring the longest time to reach the highest BLEU score (after 73 epochs), as shown in figure 2.

6 Conclusion and Future works

This paper presented a comprehensive analysis of the challenges encountered when training NMT models that incorporate pre-trained components for translation tasks. We addressed these challenges by leveraging additional multilingual translation training and devised an effective strategy for training such NMT models, even when dealing with low-resource or untrained languages. As part of our future work, we plan to employ this strategy to tackle other NMT issues, including out-of-domain translation, and in turn thoroughly investigate the en→x translation direction. Additionally, we aim to extend our approach to smaller

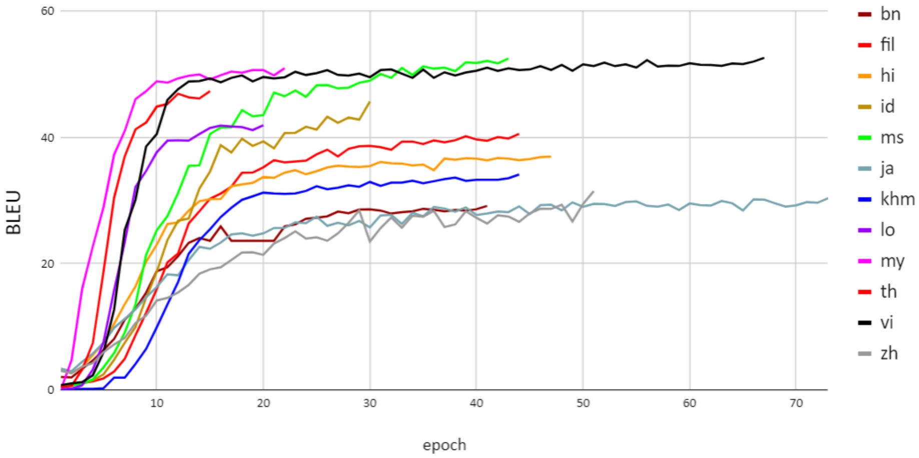

Fig. 2. BLEU scores achieved by Graformer trained on each language in the ALT dataset before the maximum BLEU score, which is then not surpassed after the next 5 consecutive epochs. English-Japanese takes 73 epochs to reach the highest BLEU, followed by English-Vietnamese taking 67.

models that incorporate either an encoder or a decoder, such as BERT-NMT with pre-trained multilingual translation.

7 Acknowledgments

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

M. Cettolo, C. Girardi, and M. Federico. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy, May 28–30 2012. European Association for Machine Translation. URL https://www.aclweb.org/anthology/2012.eamt-1.60.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179.

S. Clinchant, K. W. Jung, and V. Nikoulina. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural*

Generation and Translation, pages 108–117, Hong Kong, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5611. URL https://aclanthology.org/D19-5611.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020. URL https://aclanthology.org/2020.acl-main.747.

R. Dabre, A. Fujita, and C. Chu. 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, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1146. URL https://aclanthology.org/D19-1146.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1:4171–4186, 2019. URL https://aclanthology.org/N19-1423.

L. Doan, L. T. Nguyen, N. L. Tran, T. Hoang, and D. Q. Nguyen. PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503, 2021. URL https://aclanthology.org/2021.emnlp-main.369.

J. Guo, Z. Zhang, L. Xu, H.-R. Wei, B. Chen, and E. Chen. Incorporating bert into parallel sequence decoding with adapters, 2020. URL https://dl.acm.org/doi/10.5555/3495724.3496634.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018. doi: 10.18653/v1/d18-2012.

S. M. Lakew, A. Karakanta, M. Federico, M. Negri, and M. Turchi. Adapting multilingual neural machine translation to unseen languages. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, Nov. 2-3 2019. Association for Computational Linguistics. URL https://aclanthology.org/2019.iwslt-1.16.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

G. Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

G. Neubig and J. Hu. 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, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1103. URL https://aclanthology.org/D18-1103.

C. Ngo, T. H. Trinh, L. Phan, H. Tran, T. Dang, H. Nguyen, M. Nguyen, and M.-T. Luong. Mtet: Multi-domain translation for english and vietnamese. *arXiv preprint arXiv:2210.05610*, 2022.

NLLB Team, M. R. Costa-jussÃă, J. Cross, O. ÃĞelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. GuzmÃąn, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

M. Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

I. Provilkov, D. Emelianenko, and E. Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.170.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020b. URL https://jmlr.org/papers/volume21/20-074/20-074.pdf.

S. Rothe, S. Narayan, and A. Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280, 2020. doi: 10.1162/tacl_a_00313. URL https://aclanthology.org/2020.tacl-1.18.

Z. Sun, M. Wang, and L. Li. Multilingual translation via grafting pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2735–2747, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.233. URL https://aclanthology.org/2021.findings-emnlp.233.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Y. K. Thu, W. P. Pa, M. Utiyama, A. Finch, and E. Sumita. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1249.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. 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*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

S. Yang, Y. Wang, and X. Chu. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*, 2020.

J. Zhou, Z. Zhang, H. Zhao, and S. Zhang. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.399. URL https://aclanthology.org/2020.findings-emnlp.399.

J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu. Incorporating BERT into neural machine translation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyl7ygStwB.