Investigating Transformer Enhancements for Low-Resource Vietnamese-Lao Neural Machine Translation

Quan Xuan Truong
Faculty of Information Technology
VNU University of Engineering and Technolog
Hanoi, VietNam
22028031@vnu.edu.vn

***Abstract*—Neural Machine Translation (NMT) based on the Transformer architecture has demonstrated remarkable capabilities. However, achieving high performance for low-resource language pairs like Vietnamese (Vi) and Lao (Lo) remains a significant challenge. This paper details the development and evaluation of Transformer models for Vi-Lo translation, focusing on the impact of several architectural enhancements. We integrate Rotary Positional Embedding (RoPE) for improved relative positional encoding, utilize the GeGLU activation function within feed-forward networks, and adopt Pre-Layer Normalization for enhanced training stability. Unlike many state-of-the-art setups, our models are trained with a fixed learning rate using the AdamW optimizer and employ label smoothing. Decoding is performed using greedy search. We investigate the effect of vocabulary size on translation quality, comparing models with 8000 and 16000 subword units. Our enhanced models demonstrate notable improvements in BLEU score over a strong Transformer baseline, with the 16000 vocabulary size model achieving a BLEU score of 21.76, showcasing the benefits of the architectural modifications and vocabulary selection for this specific low-resource task.**

***Index Terms*—Neural Machine Translation, Transformer, Low-Resource Languages, Vietnamese, Lao, Rotary Positional Embedding, GeGLU, Pre-Layer Normalization, Vocabulary Size.**

I. Introduction

Neural Machine Translation (NMT) has become the dominant approach for automatic translation [1], with the Transformer architecture at its core. While highly effective for resource-rich language pairs, translating low-resource languages such as Vietnamese (Vi) and Lao (Lo) presents substantial difficulties, primarily due to the scarcity of parallel training data.

Vietnamese and Lao, despite their geographical proximity, exhibit distinct linguistic characteristics. Vietnamese is an SVO, isolating language with a rich tonal system. Lao, also tonal and SVO, belongs to the Tai-Kadai language family, featuring a different script and grammatical nuances. The limited availability of high-quality parallel corpora for Vi-Lo pairs further complicates the development of robust NMT systems.

This work focuses on constructing Transformer-based NMT models for Vi-Lo translation from scratch. We systematically incorporate several modern architectural enhancements: Rotary Positional Embedding (RoPE) [2], GeGLU activation functions [3] in the feed-forward networks (FFNs), and Pre-Layer Normalization (Pre-LN) [9]. In contrast to complex learning rate schedules often employed, we investigate the efficacy of a fixed learning rate in conjunction with these improvements. We also explore the impact of vocabulary size on translation quality. Decoding is performed using a standard greedy approach.

Our contributions are:

- Development of custom Transformer models for Vi-Lo translation incorporating RoPE, GeGLU, and Pre-LN.
- Training these models with a fixed learning rate and label smoothing [4] using the AdamW optimizer [5].
- Evaluation of the impact of different SentencePiece vocabulary sizes (8000 vs. 16000) on translation performance.
- Comparison of our enhanced models against a strong baseline Transformer, demonstrating improvements in BLEU score [6].

The paper is organized as follows: Section II provides a brief overview of related research. Section III details the data preparation, model architecture, and the implemented enhancements. Section IV describes the experimental setup, datasets, metrics, hyperparameters, and presents the results. Section V concludes the paper and discusses potential future directions.

II. Related Work

The Transformer [1] has set the standard for NMT. Key architectural improvements include alternative positional encodings like RoPE [2], which offers a relative way to inject positional information. Advanced activation functions for FFNs, such as GeGLU (a GLU variant [3], [10]), have shown benefits in large models. Pre-Layer Normalization [9] is often adopted for more stable training compared to the original Post-LN.

Training strategies for Transformers typically involve label smoothing [4] and specialized optimizers like AdamW [5]. While complex learning rate schedules with warmup are common [1], fixed learning rates can also be effective, particularly when other stabilizing elements like Pre-LN are present.

Greedy decoding is a simple and fast inference method, though beam search [11] is often used for higher quality at the cost of speed. For low-resource NMT, techniques like data augmentation [13] and vocabulary optimization are critical.

III. Methodology

Our Vi-Lo NMT system is built upon the Transformer architecture with specific enhancements.

A. Data Preparation and Tokenization

We use a Vietnamese-Lao parallel corpus, divided into training, development, and test sets. Tokenization is performed using SentencePiece [7] with the BPE algorithm. Separate tokenizers are trained for Vietnamese and Lao. We experiment with two target vocabulary sizes for each language: 8000 and 16000 subword units. Special tokens '[PAD]', '[BOS]', '[EOS]', and '[UNK]' are consistently defined.

B. Enhanced Transformer Architecture

The model adheres to the encoder-decoder structure. Both encoder and decoder consist of 3 layers. Each layer incorporates multi-head self-attention (4 heads, $d_{model} = 256$) and a position-wise FFN.

1) Rotary Positional Embedding (RoPE): We replace standard absolute positional embeddings with RoPE [2]. RoPE applies rotational transformations to query and key vectors based on their absolute positions, implicitly encoding relative positional information within the attention mechanism.

2) GeGLU Activation in FFN: The FFN blocks utilize the GeGLU activation function [3]: $FFN_{GeGLU}(x) = W_{down}((GELU(W_{gate}x)) \odot (W_{up}x))$. The hidden dimension of the GeGLU FFN is set to 680

3) Pre-Layer Normalization (Pre-LN): We employ Pre-LN [9], applying Layer Normalization before each sub-layer (self-attention, cross-attention, FFN) and its residual connection. A final Layer Normalization is also applied at the output of the encoder and decoder stacks.

C. Training and Decoding

1) Training Details: The models are trained using the AdamW optimizer [5] with a fixed learning rate of 1×10^{-4}. We apply label smoothing [4] with a factor of 0.1. The dropout rate is set to 0.1. Models are trained for 30 epochs with a batch size of 128.

2) Decoding: During inference and evaluation, we use greedy decoding, where the token with the highest probability is selected at each step to form the translation.

IV. Experiments

A. Dataset and Baseline Model

Our experiments utilize a VLSP2023 Vietnamese-Lao parallel corpus. The dataset was divided into training, development (validation), and test sets. The training set consists of approximately 100,000 sentence pairs, the development set contains 2,000 pairs, and the test set comprises 1,000 pairs.

Our baseline is a Transformer model with 12 encoder layers and 6 decoder layers, 8 attention heads, $d_{model} = 512$, and a dropout of 0.1. It was trained with a batch size of 64 using the Adam optimizer. The learning rate was set to 0.4 [Note: This is unusually high; confirm if this is a base LR for a scheduler not fully described, or a fixed LR for Adam] with 8000 warmup updates and label smoothing of 0.1. The baseline achieved a BLEU score of 16.29 on the test set.

B. Evaluation Metric

We evaluate translation quality using the BLEU score [6], computed with SacreBLEU [8] for standardized comparison.

C. Implementation Details

Our models were implemented in PyTorch [12]. The enhanced models consistently use 3 encoder and 3 decoder layers, 4 attention heads, $d_{model} = 512$, and a dropout of 0.1. The GeGLU FFN hidden dimension is 680. Training was performed on 1 A100 80GB. The model checkpoint yielding the lowest validation loss was used for test set evaluation.

TABLE I
Key Hyperparameters for Enhanced Models

Parameter	Value
Vocab Sizes (Vi/Lo)	8000 / 16000
Embedding Size (d_{model})	256
Attention Heads	4
Encoder Layers	3
Decoder Layers	3
FFN Hidden Dim (GeGLU)	680
Dropout	0.1
Max RoPE Length	512
Optimizer	AdamW
Learning Rate (Fixed)	1×10^{-4}
Label Smoothing ϵ_{ls}	0.1
Batch Size	128
Training Epochs	30
Decoding Method	Greedy Search

D. Results

Table II presents the BLEU scores of our enhanced Transformer models with different vocabulary sizes, compared to the baseline.

TABLE II
BLEU Scores

Model	BLEU
Baseline Transformer (12 Enc, 6 Dec layers)	16.29
Enhanced Model (3 Enc, 3 Dec layers):	
+ RoPE, GeGLU, Pre-LN (Vocab 8000)	18.03
+ RoPE, GeGLU, Pre-LN (Vocab 16000)	**21.76**

Our enhanced model with RoPE, GeGLU, and Pre-LN significantly outperforms the baseline, even with fewer encoder layers (3 vs. 12) and decoder layers. The model utilizing a vocabulary size of 8000 achieves a BLEU score of 18.03, an improvement of 1.74 BLEU points over the baseline. Increasing the vocabulary size to 16000 further boosts performance, resulting in a BLEU score of 21.76. This represents

a substantial gain of 5.47 BLEU points over the baseline and 3.73 BLEU points over our enhanced model with the smaller vocabulary.

E. Discussion

The results demonstrate the collective benefit of the architectural enhancements (RoPE, GeGLU, Pre-LN) for the Vi-Lo NMT task. RoPE likely provides more effective positional representations. GeGLU's gated FFN may capture more nuanced features. Pre-LN contributes to training stability, allowing effective training even with a fixed learning rate.

The considerable impact of vocabulary size is noteworthy. The jump from 8000 to 16000 subword units yields a significant BLEU improvement. This suggests that for Vi-Lo, a larger vocabulary allows for better representation of words and common phrases, reducing out-of-vocabulary issues and potentially leading to more precise translations, despite the low-resource nature of the data. The BPE algorithm's ability to create subword units is crucial here, but a larger base vocabulary seems to provide a better trade-off between sequence length and token expressiveness for this pair.

While our enhanced models have fewer encoder layers than the baseline, they still achieve superior performance, highlighting the efficiency of the chosen modifications. The baseline's high learning rate of 0.4 (if fixed) is unusual for Adam and might indicate specific tuning or a different optimizer behavior in its original setup; our fixed 1×10^{-4} learning rate with AdamW proved effective for our enhanced architecture.

V. Conclusion

This paper investigated the application of several architectural enhancements to Transformer models for low-resource Vietnamese-Lao NMT. By integrating Rotary Positional Embedding, GeGLU activation, and Pre-Layer Normalization, and training with a fixed learning rate and label smoothing, we achieved significant improvements over a strong baseline. Our findings also highlight the substantial impact of vocabulary size, with a 16000-unit vocabulary yielding the best BLEU score of 21.76.

These results underscore the potential of carefully selected architectural modifications and hyperparameter tuning, such as vocabulary size, to enhance NMT performance in data-constrained scenarios. Future work could explore the impact of these enhancements when combined with data augmentation techniques, or investigate the use of even larger vocabularies if computational resources permit. Further analysis of common translation errors could also guide targeted improvements.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[2] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[3] N. Shazeer, "GLU variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[5] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[6] 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 (ACL)*, 2002, pp. 311–318.

[7] 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 (EMNLP Demo)*, 2018, pp. 66–71.

[8] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT)*, 2018, pp. 186–191.

[9] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Tao, Y. Chen, J. Liu, and M. R. Lyu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning (ICML)*, 2020, pp. 10524–10533.

[10] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 933–941.

[11] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[12] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8026–8037.

[13] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 1341–1351.