Enhanced Transformer for Vietnamese-Lao Neural Machine Translation

Quan Xuan Truong

UET

May 16, 2025

Outline

Introduction: Neural Machine Translation (NMT)

- NMT: Dominant paradigm for automated translation.
- Transformer architecture: Cornerstone of modern NMT systems [?].
- **Challenge:** Performance heavily relies on large parallel datasets.

Transformer architecture enables high-quality translation.

The Challenge: Vietnamese-Lao Translation

Low-Resource Language Pair:

- Scarcity of high-quality parallel data.

Linguistic Differences:

- Vietnamese (Vi): SVO, tonal, Latin script.
- Lao (Lo): SVO, tonal, distinct script (Tai-Kadai).

The Challenge: Vietnamese-Lao Translation

Low-Resource Language Pair:

- Scarcity of high-quality parallel data.

Linguistic Differences:

- Vietnamese (Vi): SVO, tonal, Latin script.
- Lao (Lo): SVO, tonal, distinct script (Tai-Kadai).

Building robust NMT for Vi-Lo is demanding.

🏁 Vietnamese 🏁 Lao

Distinct scripts and structures.

Objectives & Contributions

Main Objective: Build and evaluate enhanced Transformer models for Vi-Lo NMT. **Key Contributions:**

1. Integrate architectural enhancements:
   - Rotary Positional Embedding (RoPE)
   - GeGLU Activation in Feed-Forward Networks (FFN)
   - Pre-Layer Normalization (Pre-LN)
2. Train with fixed learning rate (10^{-4}) & label smoothing.
3. Investigate impact of vocabulary size (8k vs. 16k).
4. Compare against a strong Transformer baseline.
5. Evaluate using BLEU score.

Methodology Overview

Vi-Lo Parallel Corpus → *SentencePiece BPE Tokenization* **Enhanced Transformer Model**

- RoPE
- GeGLU in FFN
- Pre-LN

Training: Fixed LR, AdamW, Label Smoothing → *Decoding:* Greedy Search *Evaluation:* BLEU Score

Architectural Enhancement 1: Rotary Positional Embedding (RoPE)

- **Problem with traditional PE:** Encodes absolute positions.
- **RoPE Solution:** Efficiently encodes relative positions.
  - Applies rotational transformations to Query (Q) and Key (K) vectors based on their absolute positions.
  - Implicitly injects relative positional information into self-attention.
- **Benefit:** Improved understanding of token relationships based on distance.

Architectural Enhancement 2: GeGLU Activation in FFN

- **Traditional FFN:** Typically uses ReLU or GELU.
- **GeGLU (GELU Gated Linear Unit):**
  - $FFN_{GeGLU}(x) = W_{down}((GELU(W_{gate}x)) \odot (W_{up}x))$
  - Introduces a gating mechanism to control information flow.
- **Benefit:** Potentially better representational power and feature selection.

Architectural Enhancement 3: Pre-Layer Normalization (Pre-LN)

- **Post-LN (Original Transformer):** $x + Dropout(Sublayer(x))$
- **Pre-LN (Our approach):** $x + Dropout(Sublayer(LayerNorm(x)))$
  - Layer Normalization is applied *before* the sub-layer (Attention, FFN) and its residual connection.
- **Benefit:** More stable training dynamics, can allow for higher learning rates or simpler warmup.

Experimental Setup

- **Dataset:** Vietnamese-Lao parallel corpus. [Add sentence counts if concise]
- **Tokenization:** SentencePiece (BPE), vocab sizes 8000 and 16000.
- **Our Enhanced Models (3 Enc, 3 Dec layers):**
  - $d_{model} = 256$, 4 Heads, Dropout 0.1
  - Fixed LR 10^{-4}, AdamW, Label Smoothing 0.1
  - RoPE, GeGLU, Pre-LN
- **Baseline Model (12 Enc, 6 Dec layers):**
  - $d_{model} = 512$, 8 Heads, Dropout 0.1
  - Adam, LR 0.4 (unusually high, from provided info), Warmup 8000 steps, Label Smoothing 0.1
- **Evaluation Metric:** BLEU score (SacreBLEU).

Results: Vi-Lo Translation (BLEU Score)

Table: BLEU Scores on the Vi-Lo

Model Configuration	**BLEU**
Baseline Transformer (12 Enc, 6 Dec)	16.29
Enhanced Model (6 Enc, 6 Dec):	
+ RoPE, GeGLU, Pre-LN (Vocab 8k)	18.03
+ RoPE, GeGLU, Pre-LN (Vocab **16k**)	**21.76**

- Enhanced models (6 layers) outperform stronger baseline (12 layers).
- Vocab 16k yields best result: +5.47 BLEU over baseline.
- Vocab 16k shows +3.73 BLEU over vocab 8k within enhanced models.

Discussion of Results

- **Effectiveness of Architectural Enhancements:**
  - RoPE, GeGLU, and Pre-LN collectively provide significant gains, even with fewer parameters (3 vs. 12 encoder layers).

Discussion of Results

- **Effectiveness of Architectural Enhancements:**
  - RoPE, GeGLU, and Pre-LN collectively provide significant gains, even with fewer parameters (3 vs. 12 encoder layers).
- **Impact of Vocabulary Size:**
  - Larger vocabulary (16k) crucial for Vi-Lo, offering better token representation.
  - Suggests a good balance between subword granularity and sequence length for this pair.

Discussion of Results

- **Effectiveness of Architectural Enhancements:**
  - RoPE, GeGLU, and Pre-LN collectively provide significant gains, even with fewer parameters (3 vs. 12 encoder layers).
- **Impact of Vocabulary Size:**
  - Larger vocabulary (16k) crucial for Vi-Lo, offering better token representation.
  - Suggests a good balance between subword granularity and sequence length for this pair.
- **Fixed Learning Rate:** Viable with Pre-LN and AdamW for stable training.
- **Comparison Note:** Our models achieve superior results despite fewer encoder layers than the baseline. The baseline's high LR (0.4) is noteworthy.

Conclusion

- Successfully developed enhanced Transformer models for Vi-Lo NMT.
- Key improvements from:
  - Architectural changes (RoPE, GeGLU, Pre-LN).
  - Optimized vocabulary size (16k proving most effective).
- Achieved a BLEU score of **21.76** with the 16k vocabulary model.
- Demonstrates that targeted enhancements can significantly boost performance in low-resource scenarios.

Future Work

- **Data Augmentation:** Implement back-translation to expand training data.
- **Decoding Strategies:** Evaluate Beam Search for potentially higher quality translations.
- **Learning Rate Scheduler:** Experiment with standard Transformer LR schedules (e.g., inverse square root decay with warmup).
- **Vocabulary Exploration:** Test even larger vocabulary sizes if resources permit.
- **Detailed Error Analysis:** Identify common error types to guide further improvements.
- **Transfer Learning:** Investigate leveraging pre-trained multilingual models.

Thank You! Questions?

Code/Report

available at: https://github.com/quanxuantruong/Enhanced-Transformer-for-Vietnamese-Lao-Neural-Machine-Translation

Email: 22028031@vnu.edu.vn

Appendix: Key Hyperparameters (Enhanced Models) I

Table: Hyperparameters for the Enhanced Transformer 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