IMPROVING TRANSFORMER-BASED MACHINE TRANSLATION FOR VIET-LAO

Presented by Group 7

MEMBER

1. Nguyen Bao Dung
2. Tran Van Hiep
3. Tran Tuan Phong
4. Tran Dai Duong

OVERVIEW

- Abstract
- Introduction
- Problem
- Objectives
- Methodology
- Result
- Conclusion
- Reference

ABSTRACT

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus sed vestibulum nunc, eget aliquam felis. Sed nunc purus, accumsan sit amet dictum in, ornare in dui. Ut imperdiet ante eros, sed porta ex eleifend ac. Donec non porttitor leo. Nulla luctus ex lacus, ut scelerisque odio semper nec.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus sed vestibulum nunc, eget aliquam felis. Sed nunc purus, accumsan sit amet dictum in, ornare in dui. Ut imperdiet ante eros, sed porta ex eleifend ac. Donec non porttitor leo. Nulla luctus ex lacus, ut scelerisque odio semper nec.

INTRODUCTION

WHAT IS NLP?

- A pivotal field in AI, enabling machines to understand, interpret, and generate human language.
- Applications: Machine translation, sentiment analysis, speech recognition.

TRANSFORMER ARCHITECTURE

- Revolutionized NLP for sequence-to-sequence tasks (Vaswani et al., 2017).
- Core mechanism: Self-attention for parallel processing and long-range dependencies.
- Foundation for SOTA models (BERT, GPT); highly effective for Machine Translation (MT).

THE VIET-LAO LANGUAGE PAIR (VIETNAMESE)

- **Script:** Latin-based with tone diacritics
- **Key Features:** Tonal (6 tones), analytic language, spaces generally delimit words
- **MT Implication:** Tone representation is crucial; standard tokenization is relatively straightforward

THE VIET-LAO LANGUAGE PAIR (LAO)

- **Script:** Lao Abugida (from Khmer script), vowel diacritics
- **Key Features:** Tonal (5-6 tones), analytic language
- **Major MT Challenge**: No consistent word delimiters (scriptio continua). Spaces often mark phrases/clauses, not individual words
- **MT Implication:** Word segmentation is a critical, non-trivial preprocessing step

PROBLEM

PROBLEM

- **Low-Resource Nature:**
  - Limited availability of high-quality parallel corpora compared to high-resource pairs
  - Dataset size: 102,000 sentence pairs (100k train, 2k valid)

- **Linguistic Divergence:**
  - Vietnamese: Tonal, Latin script
  - Lao: Tonal, unique script, traditionally no word delimiters
  - Syntactic differences

PROBLEM

- **Data Quality Issues:**
  - Raw data contains noise: crawler artifacts, irrelevant language characters.
  - Lao corpus contains mixed Lao and Roman script sentences.

- **Baseline Model Limitations:**
  - Standard Transformer with greedy decoding and Adam optimizer might not be optimal for this specific low-resource, linguistically diverse task

OBJECTIVES

OBJECTIVE-PRIMARY

To improve the translation quality of a Transformer-based NMT model for the Viet-Lao language pair, as measured by BLEU score.

OBJECTIVE-SPECIFIC

1. To investigate and implement advanced decoding strategies (Beam Search) to enhance translation fluency and coherence.
2. To explore the impact of alternative optimization algorithms (AdamW) on model generalization and performance.
3. To evaluate the effect of combining these improvements (Beam Search + AdamW).
4. To determine the influence of varying training batch sizes on translation quality for this low-resource setting.

METHODOLOGY

OVERVIEW

- **Baseline Model:** Transformer sequence-to-sequence architecture.
- **Core Improvements Investigated:**
  - Decoding Strategy: Beam Search Generation.
  - Optimization Algorithm: AdamW.
  - Hyperparameter Tuning: Batch Size Variation.
- **Development Environment:** PyTorch Lightning, Hydra Lightning Template.
- **Model Implementation:** Transformer built "from scratch" for deeper architectural understanding.

DATA PREPROCESSING

- **Dataset Source:** Provided corpus of 102,000 Lao-Vietnamese sentence pairs.
- **Data Segregation:**
  - Separated Lao sentences into "Lao-only script" and "mixed Lao/Roman script."
  - Focused initial training on Lao-only data (45,633 train, 737 valid) to establish a strong foundation.

DATA PREPROCESSING

- **Cleaning Steps:**
  - Removal of crawler artifacts, unknown Unicode, and irrelevant language characters (referencing Figure 1 from report).

MODEL ARCHITECTURE DETAILS

- **Tokenization:**
  - Lao (Source): Byte-Pair Encoding (BPE) – suitable for abugida script without clear word spaces.
  - Vietnamese (Target): Word Level Tokenizer – handles "scriptio continua."
  - Vocabularies: 30k (Lao), 8.2k (Vietnamese, incl. special tokens), min. frequency 2.
- **Encoder-Decoder Stacks:** More encoders than decoders (rationale: vocabulary size difference).

MODEL ARCHITECTURE DETAILS

- **Attention Mechanism:**
  - Scaled Dot-Product Attention (original paper).
  - Multi-Head Attention: Direct concatenation of 8 head scores (omitting final linear layer).
- **Position-wise Feed-Forward Networks:** As per original paper.
- **Embeddings:** Trained from scratch (due to RAM issues with pre-trained ones), d_model size.

MODEL ARCHITECTURE DETAILS

- **Positional Encoding:**
  - Slightly modified fixed positional encodings.
  - Division in log space for numerical stability

#FORMULA

BEAM SEARCH

- **Rationale:** Overcome limitations of greedy decoding (myopic choices).
- **Approach:** Maintains k candidate sequences, explores multiple hypotheses.
- **Implementation:**
  - Beam widths: 7 and 10.
  - Length normalization ($\alpha = 0.6$) to prevent bias towards shorter sequences.
- **Expected Outcome:** Improved translation coherence and fluency.

ADAMW OPTIMIZER

- **Rationale:** Address potential poor generalization of Adam due to insufficient regularization.
- **Approach:** AdamW decouples weight decay from optimization steps.
- **Implementation:**
  - Weight decay: 0.01.
  - Learning rate: 0.0005 (same as baseline).
- **Expected Outcome:** Better generalization, especially for low-frequency words and complex structures in a small corpus.

COMBINING BEAM SEARCH & ADAMW

- **Rationale:** Synergistic effect from improved decoding and better generalization.
- **Configuration:** Beam width 7 (α = 0.6), AdamW (weight decay 0.01).
- **Expected Outcome:** Fluent and accurate translations

BATCH SIZE VARIATION

- **Rationale:** Batch size significantly impacts convergence and performance.
- **Experiment:**
  - **Baseline batch size:** 128
  - **Tested batch sizes:** 32, 64
- **Hypothesis:** Smaller batch sizes might yield better generalization on this low-resource dataset, despite longer training times.

EXPERIMENTAL SETUP & EVALUATION

- **Dataset Split:** 100k training, 2k validation pairs.
- **Evaluation Metric:** BLEU score (primary), Test Loss.
- **Configurations Compared:**
  - Baseline (Greedy, Adam, BS 128)
  - Beam Search (width 7 & 10)
  - AdamW
  - Beam Search (7) + AdamW
  - Effect of Batch Sizes (32, 64, 128) on the best combined

RESULTS & DISCUSSION

RESULT

Configuration	BLEU score	Training time	Infer time
Baseline (batch size = 128)	0.25	16 hours	350 s
Batch size = 64	0.32	18 hours	350 s
Batch size = 32	0.4	22 hours	350 s
Beam search (beam size = 7)	0.28	16.5 hours	420 s
Beam search (beam size = 10)	0.32	18 hours	480 s
AdamW	0.3	16 hours	350 s
Beam search(7) + AdamW	0.35	16.5 hours	421 s

CONCLUSION

FUTURE WORK

Q&A

Thank You

For your attention