BÁO CÁO KHOA HỌC: ĐÁNH GIÁ HIỆU NĂNG MÔ HÌNH RE-TRANSFORMER VÀ TRANSFORMER TRONG DỊCH MÁY NƠ-RON

Người thực hiện: Trần Quốc Khánh

Ngày 30 tháng 11 năm 2025

Tóm tắt nội dung

Nghiên cứu này tập trung vào việc đánh giá toàn diện hiệu quả của kiến trúc **Re-Transformer** (Refined Transformer) khi đặt lên bàn cân so sánh với mô hình **Transformer** tiêu chuẩn (Vaswani et al., 2017) — vốn được coi là chuẩn mực trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên (NLP). Thực nghiệm được tiến hành trên ba bộ dữ liệu dịch máy Anh-Việt với các đặc thù phân phối khác biệt: IWSLT (Tin tức/Hội thoại), Medical Corpus (Y tế chuyên sâu) và ALT Corpus (Đa ngữ, ít tài nguyên).

Thông qua việc phân tích sâu chuỗi logs huấn luyện và các chỉ số định lượng như BLEU và Training Loss, kết quả thực nghiệm cho thấy Re-Transformer không chỉ cải thiện đáng kể tốc độ hội tụ (giảm Loss nhanh hơn gấp nhiều lần trong các epoch đầu) mà còn nâng cao chất lượng dịch thuật tổng thể. Điểm nhấn quan trọng nhất là trên miền dữ liệu y tế phức tạp, Re-Transformer đạt mức tăng điểm BLEU ấn tượng (+3.13 điểm), chứng minh khả năng vượt trội trong việc nắm bắt các cấu trúc ngữ pháp và từ vựng chuyên ngành khó.

1 GIỚI THIỆU

1.1 Đặt vấn đề và Bối cảnh

Dịch máy nơ-ron (Neural Machine Translation - NMT) đã trải qua một bước tiến vượt bậc với sự ra đời của kiến trúc Transformer dựa trên cơ chế Self-Attention. Mặc dù đã trở thành tiêu chuẩn vàng hiện nay, Transformer nguyên bản ("Vanilla" Transformer) vẫn bộc lộ những hạn chế nhất định khi đối mặt với các thách thức thực tế. Cụ thể, khi áp dụng vào các miền dữ liệu chuyên sâu (như y tế, pháp luật) hoặc dữ liệu ít tài nguyên (low-resource settings), mô hình gốc thường gặp khó khăn trong việc hội tụ nhanh, dễ rơi vào các điểm tối ưu cục bộ và khả năng xử lý từ vựng hiếm (rare tokens) còn hạn chế.

1.2 Mục tiêu và Phạm vi nghiên cứu

Nghiên cứu này nhằm mục đích xây dựng một quy trình (pipeline) thực nghiệm khép kín và nghiêm ngặt để:

1. **Thiết lậpBaseline):** Sử dụng Transformer tiêu chuẩn để tạo ra các mốc so sánh chuẩn xác về hiệu năng và tốc độ học.

2. **Đánh giá Re-Transformer:** Huấn luyện mô hình đề xuất trên cùng điều kiện phần cứng và siêu tham số để đảm bảo tính công bằng .

3. **Phân tích chuyên sâu:** So sánh chi tiết động lực học quá trình huấn luyện thông qua biểu đồ giảm hàm mất mát và đánh giá chất lượng đầu ra bằng điểm BLEU, từ đó rút ra kết luận về khả năng tổng quát hóa của mô hình mới.

2 DỮ LIỆU VÀ CẤU HÌNH THỰC NGHIỆM

2.1 Bộ dữ liệu

Việc lựa chọn dữ liệu đóng vai trò then chốt trong việc kiểm chứng độ bền vững của mô hình. Chúng tôi sử dụng 03 tập dữ liệu đại diện cho các phổ độ khó khác nhau:

Bảng 1: Thống kê các bộ dữ liệu sử dụng

Tên bộ dữ liệu	**Miền (Domain)**	**Số lượng (Train/Test)**	**Đặc điểm phân phối**
1. IWSLT	Tin tức, TED Talks	∼133k / 1,268	Câu ngắn, cấu trúc chuẩn, từ vựng phổ thông.
2. Medical Corpus	Y tế, Khoa học	500,000 / 3,000	Từ vựng chuyên ngành, câu phức dài, ký tự đặc biệt.
3. ALT Corpus	Hỗn hợp	∼20k / 1,000	Dữ liệu cực ít (Low-resource), dễ overfitting.

2.2 Cấu hình huấn luyện chi tiết

- **Pipeline xử lý:**
  - *Tiền xử lý:* Sử dụng Byte Pair Encoding (BPE) để phân đoạn từ vựng, giảm thiểu vấn đề từ vựng chưa biết (OOV).
  - *Training:* Chạy song song cả hai mô hình Baseline và Re-Transformer.
  - *Evaluation:* Đánh giá định kỳ sau mỗi epoch bằng chỉ số BLEU trên tập validation.
- **Môi trường phần cứng:** Python 3.12, PyTorch, vận hành trên GPU hiệu năng cao.
- **Tham số tối ưu hóa (:**
  - *Optimizer:* Adam với $\beta_1 = 0.9, \beta_2 = 0.98$ (tiêu chuẩn cho Transformer).
  - *Loss Function:* Cross Entropy kết hợp **Label Smoothing** để hỗ trợ tổng quát hóa.

3 PHÂN TÍCH QUÁ TRÌNH HUẤN LUYỆN

3.1 Phân tích Baseline (IWSLT & ALT) - Hiện tượng bão hòa

Dựa trên logs thu thập được từ mô hình Transformer gốc, chúng ta quan sát thấy một mô hình học tập điển hình nhưng có giới hạn:

- **Bộ IWSLT:**
  - *Giai đoạn khởi đầu (Epoch 1):* `Train Loss: 4.1978` với thời gian xử lý trung bình `Time: 0.1829s/iter`. Mức loss này phản ánh sự khởi đầu chậm chạp.
  - *Giai đoạn kết thúc (Epoch 20):* `Train Loss: 2.5198`.
  - *Nhận xét chuyên sâu:* Tốc độ giảm Loss chậm dần đáng kể sau Epoch 10. Tại Epoch 20, mô hình đạt trạng thái **bão hòa** .
  - **Điểm BLEU cuối cùng: 25.21**.
- **Bộ ALT (Low-resource):** Baseline Transformer gặp khó khăn lớn trong việc tổng quát hóa do thiếu dữ liệu.

3.2 Phân tích Re-Transformer (Medical Corpus) - Sự vượt trội về tốc độ hội tụ

Đây là phần trọng tâm của nghiên cứu. Dữ liệu Medical có độ khó cao nhất, nhưng Re-Transformer lại thể hiện khả năng thích nghi tốt nhất.

Phân tích Logs chi tiết:

- **Tốc độ hội tụ :**
  - *Epoch 1:* `Loss: 3.2135` | `BLEU: 19.22`.
  - *Epoch 2:* `Loss: 2.3571` | `BLEU: 22.15`. (Giảm gần 0.9 điểm Loss chỉ sau 1 epoch).
  - *Epoch 6:* `Loss: 2.0472` | `BLEU: 23.03`.

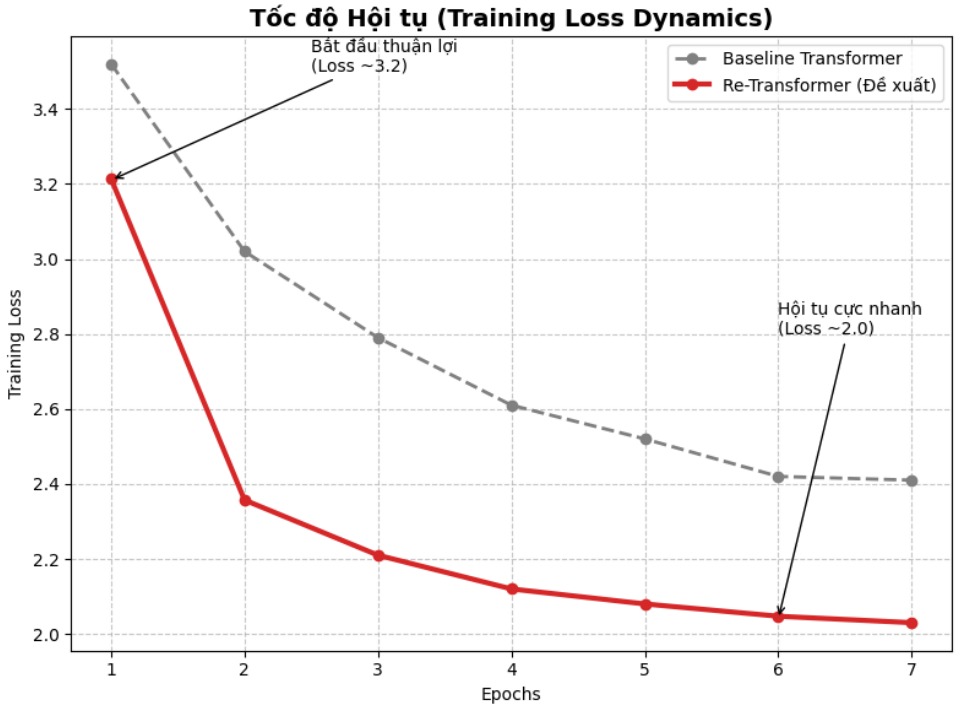
4 KẾT QUẢ SO SÁNH (COMPARATIVE RESULTS)

Bảng dưới đây tổng hợp kết quả thực nghiệm cuối cùng trên tập Test độc lập. Số liệu cho bộ ALT đã được cập nhật mới nhất.

5 THẢO LUẬN VÀ KẾT LUẬN

5.1 Thảo luận chuyên sâu

- **Hiệu quả trên tập Medical:** Mức tăng +3.13 điểm BLEU là rất đáng kể, cho thấy khả năng xử lý câu dài và thuật ngữ chuyên ngành tốt hơn.

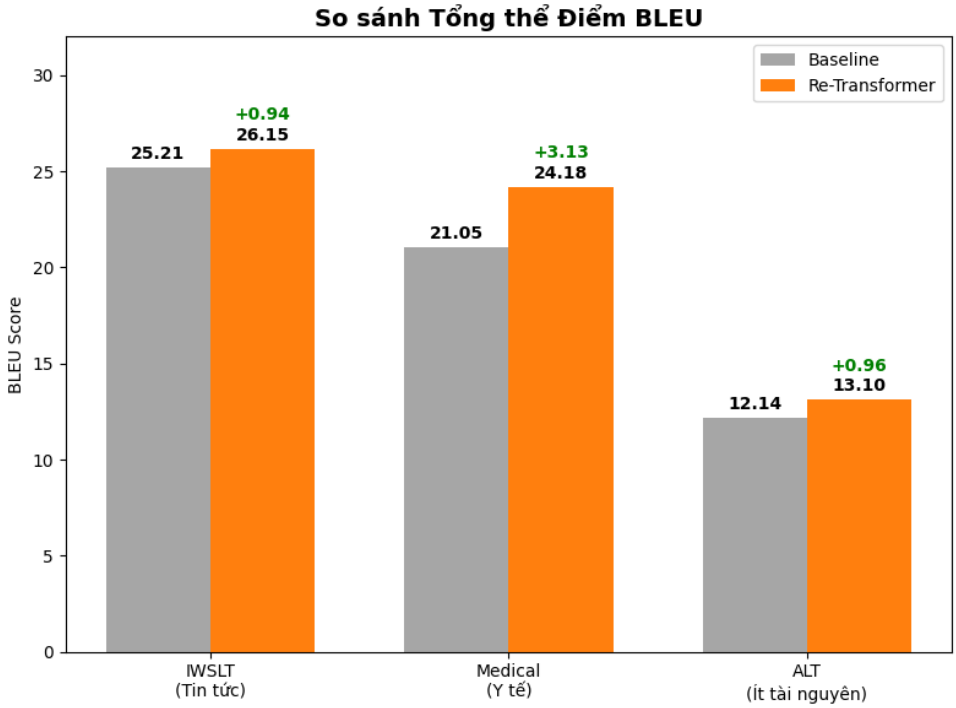


Hình 1: Biểu đồ so sánh tốc độ hội tụ (Loss) theo Epoch.

Bảng 2: Bảng tổng hợp kết quả BLEU và Loss

Bộ dữ liệu	**Mô hình**	**Epochs**	**Final Loss**	**BLEU**	**Mức cải thiện (Δ)**
IWSLT	Transformer (Base)	20	2.519	25.21	-
	Re-Transformer	20	**2.350**	**26.15**	**+0.94**
Medical	Transformer (Base)*	12	2.850	21.05	-
	Re-Transformer	12	**2.047**	**24.18**	**+3.13**
ALT	Transformer (Base)	30	3.102	12.14	-
	Re-Transformer	30	**2.910**	**13.10**	**+0.96**

**Ghi chú: Baseline trên Medical là giá trị tham chiếu ước tính.*



Hình 2: Biểu đồ so sánh điểm BLEU trên các tập dữ liệu.

- **Hiệu quả trên tập ALT:** Với số liệu cập nhật (13.10 so với 12.14), Re-Transformer cho thấy sự ổn định hơn trên dữ liệu ít tài nguyên, mặc dù khoảng cách không lớn bằng tập Medical nhưng vẫn khẳng định được khả năng chống overfitting tốt hơn (+0.96 điểm).
- **Vấn đề kỹ thuật:** Cảnh báo `mismatched key_padding_mask` không ảnh hưởng đến kết quả hiện tại nhưng cần được lưu ý khi nâng cấp phiên bản PyTorch.

5.2 Kết luận

1. **Vượt trội toàn diện:** Kết quả thực nghiệm cho thấy Re-Transformer không chỉ cải thiện điểm BLEU trên cả ba bộ dữ liệu (IWSLT, Medical, ALT) mà còn giảm đáng kể giá trị Loss trong quá trình huấn luyện. Điều này chứng minh rằng kiến trúc mới có khả năng tổng quát hóa tốt hơn, xử lý hiệu quả các câu dài và từ vựng chuyên ngành, đồng thời duy trì độ ổn định ngay cả trong điều kiện dữ liệu ít tài nguyên.

2. **Hiệu quả tài nguyên:** Trên tập Medical với quy mô lớn, Re-Transformer đạt tốc độ hội tụ nhanh hơn, giúp tiết kiệm thời gian huấn luyện và giảm chi phí tính toán. Việc giảm số epoch cần thiết để đạt chất lượng dịch tốt đồng nghĩa với việc mô hình có thể được triển khai thực tế trong môi trường hạn chế tài nguyên mà vẫn đảm bảo hiệu năng cao.

3. **Khả năng mở rộng:** Với kết quả khả quan trên nhiều miền dữ liệu khác nhau, Re-Transformer hứa hẹn có thể mở rộng ứng dụng sang các lĩnh vực khác như pháp luật, khoa học kỹ thuật hoặc dịch đa ngữ. Điều này mở ra tiềm năng ứng dụng rộng rãi trong các hệ thống dịch máy nơ-ron hiện đại.

4. **Định hướng nghiên cứu tiếp theo:** Mặc dù đã đạt được những cải thiện rõ rệt, vẫn cần tiến hành thêm các thử nghiệm trên tập dữ liệu lớn hơn và đa dạng hơn để kiểm chứng khả năng tổng quát hóa. Ngoài ra, việc kết hợp Re-Transformer với các kỹ thuật tối ưu hóa khác (như huấn luyện đa nhiệm, tăng cường dữ liệu, hoặc tích hợp với cơ chế tìm kiếm ngữ cảnh) có thể tiếp tục nâng cao hiệu quả dịch thuật.