**Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội**

—----------


# BÁO CÁO BÀI TẬP LỚN
# XỬ LÝ NGÔN NGỮ TỰ NHIÊN

## Đề tài: Xây dựng và Đánh giá Hệ thống Dịch máy Nơron Anh-Việt trong Lĩnh vực Y khoa

**Nhóm** : 727
**Thành viên:** Lê Vũ Hiếu - 23020365
Nguyễn Duy Hải Bằng - 23020335

**Tóm tắt**

Bài báo cáo này trình bày quá trình thực hiện một hệ thống dịch máy nơron (NMT) chuyên biệt cho cặp ngôn ngữ Anh-Việt trong lĩnh vực y khoa. Hai phương pháp tiếp cận chính đã được triển khai và so sánh. Phương pháp thứ nhất là xây dựng một mô hình từ đầu dựa trên kiến trúc Transformer gốc, huấn luyện trực tiếp trên bộ dữ liệu y khoa được cung cấp. Phương pháp thứ hai, hiện đại hơn, là áp dụng kỹ thuật học chuyển giao bằng cách tinh chỉnh (finetuning) mô hình ngôn ngữ lớn ViT5-base, vốn đã được huấn luyện trước trên một kho dữ liệu tiếng Việt khổng lồ. Hiệu năng của cả hai hệ thống được đánh giá một cách định lượng thông qua điểm BLEU và định tính thông qua việc phân tích các ví dụ dịch cụ thể trên tập kiểm tra. Kết quả thực nghiệm cho thấy mô hình ViT5 sau khi finetune đạt điểm BLEU vượt trội (43.94) so với mô hình Transformer huấn luyện từ đầu (15), qua đó khẳng định sức mạnh và tính hiệu quả của phương pháp học chuyển giao trong các bài toán NLP chuyên ngành.

Github

# Chương 1: Giới thiệu

## 1.1. Bối cảnh và Tầm quan trọng của NMT

Dịch máy nơron (Neural Machine Translation - NMT) đã tạo ra một cuộc cách mạng trong lĩnh vực xử lý ngôn ngữ tự nhiên, mang lại những bản dịch có chất lượng cao, tự nhiên và mạch lạc hơn hẳn so với các phương pháp thống kê trước đây. Trong bối cảnh toàn cầu hóa, nhu cầu dịch thuật các tài liệu chuyên ngành, đặc biệt là trong lĩnh vực y khoa, ngày càng trở nên cấp thiết để phổ biến kiến thức và cải thiện chất lượng chăm sóc sức khỏe.

## 1.2. Giới thiệu kiến trúc Transformer và ViT5

Kiến trúc Transformer, được giới thiệu vào năm 2017, đã trở thành nền tảng cho hầu hết các mô hình ngôn ngữ hiện đại nhờ vào cơ chế chú ý (attention) hiệu quả, cho phép xử lý song song và nắm bắt các phụ thuộc xa trong câu. Kế thừa và phát triển từ Transformer, ViT5-base là một mô hình ngôn ngữ lớn thuộc họ T5 (Text-to-Text Transfer Transformer), được VinAI Research huấn luyện riêng cho tiếng Việt, sở hữu kiến thức nền sâu rộng về ngôn ngữ.

## 1.3. Mục tiêu của Bài tập

Bài tập này đặt ra hai mục tiêu chính:

Xây dựng và huấn luyện một mô hình NMT Anh-Việt từ đầu dựa trên kiến trúc Transformer để hiểu sâu về cách hoạt động của nó.

Áp dụng kỹ thuật finetuning trên mô hình ViT5-base để giải quyết cùng bài toán.

So sánh, phân tích và đánh giá hiệu năng của cả hai phương pháp để rút ra kết luận về tính hiệu quả của học chuyển giao.

## 1.4. Cấu trúc của Báo cáo

Báo cáo được trình bày theo các chương: Giới thiệu, Cơ sở Lý thuyết, Phương pháp Thực nghiệm, Kết quả và Thảo luận, và cuối cùng là Kết luận.

# Chương 2: Cơ sở Lý thuyết

## 2.1. Kiến trúc Transformer

Mô hình Transformer bao gồm một Encoder và một Decoder. Encoder xử lý chuỗi đầu vào và tạo ra một chuỗi các biểu diễn ngữ cảnh. Decoder sử dụng các biểu diễn này cùng với chuỗi đã được dịch trước đó để sinh ra từ tiếp theo. Điểm đột phá của Transformer là hoàn toàn loại bỏ các mạng nơ-ron hồi quy (RNN) và chỉ dựa vào cơ chế Self-Attention, giúp tăng tốc độ huấn luyện và nắm bắt ngữ cảnh tốt hơn.

## 2.2. Học Chuyển giao và Finetuning

Học chuyển giao là một kỹ thuật trong đó một mô hình được huấn luyện trước (pre-trained) trên một lượng lớn dữ liệu tổng quát, sau đó được tinh chỉnh (finetuned) trên một tập dữ liệu nhỏ hơn, chuyên biệt cho một tác vụ cụ thể. Điều này giúp mô hình tận dụng được kiến thức nền đã học và đạt được hiệu năng cao với thời gian và dữ liệu ít hơn.

## 2.3. Thước đo Đánh giá: BLEU

BLEU (Bilingual Evaluation Understudy) là thước đo phổ biến nhất để đánh giá chất lượng dịch máy. Nó tính toán độ chính xác của các n-gram (cụm từ gồm n từ) trong câu dịch của máy so với một hoặc nhiều câu tham khảo của con người, sau đó kết hợp chúng lại và áp dụng một "hệ số phạt" (brevity penalty) cho các câu dịch quá ngắn.

# Chương 3: Phương pháp Thực nghiệm

## 3.1. Mô tả Tập dữ liệu

Bộ dữ liệu được sử dụng là Medical Parallel Corpus với 500,000 cặp câu huấn luyện và 3,000 cặp câu kiểm tra. Đây là một bộ dữ liệu có kích thước lớn, chứa nhiều thuật ngữ y khoa phức tạp và các câu có độ dài đa dạng, tối đa lên tới hơn 500 từ.

## 3.2. Tiền xử lý Dữ liệu

Quy trình tiền xử lý chung bao gồm: chuẩn hóa văn bản (viết thường), huấn luyện tokenizer BPE với vocab size 30,000, và chia 10% tập huấn luyện thành tập xác thực.

## 3.3. Thiết lập Thí nghiệm 1

***Huấn luyện Transformer từ đầu***

Cấu hình Mô hình: *N=6, d_model=256, num_heads=8, d_ff=1024, seq_len=900.*

Cấu hình Huấn luyện: *Môi trường Kaggle GPU T4, batch_size=2, gradient_accumulation_steps=16, optimizer Adam với learning rate schedule, loss CrossEntropy với label_smoothing=0.1.*

## 3.4. Thiết lập Thí nghiệm 2

***Finetuning ViT5-base***

Mô hình: *Sử dụng vietai/vit5-base từ Hugging Face Hub.*

Cấu hình Finetuning: *Môi trường Kaggle GPU T4, batch_size=4, gradient_accumulation_steps=4, learning_rate=2e-5, num_epochs=15, MAX_INPUT_LENGTH=512.*

Toàn bộ quy trình được quản lý bằng Seq2SeqTrainer của Hugging Face với tính năng EarlyStopping (patience=2) và tự động lưu lại checkpoint tốt nhất dựa trên điểm BLEU của tập validation.

# Chương 4: Kết quả và Thảo luận

Chương này trình bày và phân tích chi tiết kết quả thực nghiệm của cả hai hệ thống NMT. Hiệu năng được đánh giá trên cả phương diện định lượng (điểm BLEU) và định tính (phân tích các ví dụ dịch cụ thể).

**4.1. Kết quả Đánh giá Định lượng**

Bảng dưới đây so sánh điểm BLEU trên tập kiểm tra của cả hai mô hình.

| **Mô hình** | **BLEU** |
|---|---|
| Transformer from scratch | 15 |
| ViT5 finetuned | 43.94 |

=> Rõ ràng, mô hình ViT5 finetuned cho kết quả vượt trội, cho thấy hiệu quả rõ rệt của việc tận dụng kiến thức từ một mô hình đã được pre-train.

**4.2. Phân tích kết quả dịch từ ViT5 Finetuned**

Để hiểu sâu hơn về khả năng của mô hình, chúng tôi tiến hành phân tích các ví dụ cụ thể được dịch bởi mô hình ViT5 finetuned, vốn cho kết quả tốt hơn.

**Ví dụ 1: Xử lý câu dài và thuật ngữ phức tạp**

**Source:** Conclusion: Our study has shown that the percutaneous closure of large secondary atrial septal defects in the 20 - 37 mm diameter range under intracardiac echocardiography guidance can be performed safely and effectively.

**Reference:** Kết luận: Nghiên cứu của chúng tôi đã chỉ ra rằng việc đóng lỗ thông liên nhĩ thứ phát lớn có đường kính từ 20 - 37 mm qua da dưới sự hướng dẫn của siêu âm tim có thể được thực hiện an toàn và hiệu quả.

**NMT (ViT5):** Kết luận: Nghiên cứu của chúng tôi đã cho thấy việc đóng thông liên nhĩ 20-37 mm trong khoảng đường kính 20-37 mm dưới hướng dẫn siêu âm tim nội tâm mạc có thể thực hiện an toàn và hiệu quả.
**Phân tích:** Mô hình không chỉ dịch đúng các thuật ngữ (intracardiac echocardiography -> siêu âm tim) mà còn sắp xếp lại cấu trúc câu một cách tự nhiên theo văn phong tiếng Việt.
**Ví dụ 2: Phân tích lỗi ngữ cảnh**
**Source:** The study was completed in May 2015 with 476 children aged 12 years at 9 secondary schools, applying the methodology of investigation and classification of WHO-2013 with the Dean's fluorosis index.
**Reference:** Nghiên cứu được hoàn thành vào tháng 5 năm 2015 với 476 trẻ em trong độ tuổi 12 tại 9 trường trung học, áp dụng phương pháp điều tra và phân loại của WHO-2013 theo chỉ số nhiễm fluor của Dean.
**NMT (ViT5):** Nghiên cứu được tiến hành vào tháng 5 năm 2015 với 476 trẻ em từ 12 tuổi tại 9 trường mầm non, áp dụng phương pháp điều tra, phân loại theo WHO-2013 với chỉ số fluor Dean.
**Phân tích lỗi:** Mô hình đã dịch sai từ secondary thành trường mầm non.
**Đề xuất khắc phục:** Lỗi này cho thấy sự thiếu hụt ngữ cảnh trong dữ liệu. Việc finetuning trên một bộ dữ liệu lớn hơn, đa dạng hơn về các độ tuổi và loại trường học có thể giúp mô hình hiểu và dịch chính xác hơn.

**Ví dụ 3: Phân tích lỗi sai lệch cuối câu (Divergence)**
**Source:** From January 2018 to April 2021 at Cardiovascular and Thoracic Center - Viet Duc University Hospital, 38 patients underwent video-assisted thoracoscopic surgery for primary spontaneous pneumothorax.
**Reference:** Từ tháng 1 năm 2018 đến tháng 4 năm 2021 tại Trung tâm Tim mạch và Lồng ngực - Bệnh viện Hữu nghị Việt Đức, 38 bệnh

nhân được phẫu thuật nội soi lồng ngực có video hỗ trợ điều trị tràn khí màng phổi tự phát nguyên phát.
**NMT (ViT5):** Từ tháng 1 năm 2018 đến tháng 4 năm 2021 tại Trung tâm Tim mạch và Lồng ngực - Bệnh viện Hữu nghị Việt Đức, 38 bệnh nhân đã trải qua phẫu thuật lồng ngực và *tử vong và tử vong và tử vong*.
Phân tích:
Mô hình đã dịch tốt và chính xác toàn bộ phần đầu của câu, bao gồm cả thời gian, địa điểm, và thủ thuật y tế.
Tuy nhiên, ở cuối câu, thay vì dịch for primary spontaneous pneumothorax, mô hình đột ngột sai lệch và sinh ra một chuỗi lặp lại vô nghĩa và sai sự thật (*và tử vong*).
=> Đây là một ví dụ điển hình của lỗi "divergence". Sau khi đã sinh ra một chuỗi dài và phức tạp, trạng thái ẩn của decoder có thể đã mất đi thông tin từ câu gốc. Nó có thể đã rơi vào một trạng thái lỗi, nơi token death, một từ cũng thường xuất hiện trong văn bản y khoa, có xác suất cao và bị lặp lại. Điều này cho thấy sự khó khăn của mô hình trong việc duy trì sự chú ý (attention) trên toàn bộ câu nguồn khi câu đích đã trở nên quá dài.

### 4.3. Thảo luận về Hiệu năng

Quá trình huấn luyện mô hình "from scratch" tốn rất nhiều thời gian (nhiều ngày) và tài nguyên tính toán do phải học từ đầu trên một bộ dữ liệu lớn với seq_len dài. Ngược lại, quá trình finetuning ViT5 nhanh hơn đáng kể, đạt được kết quả tốt chỉ sau vài giờ huấn luyện, chứng tỏ đây là phương pháp tiếp cận hiệu quả hơn về mặt tài nguyên.

### 4.4. Hạn chế

- **Chiến lược Decoding Đơn giản:** Trong phần đánh giá mô hình "from scratch", bọn em chỉ sử dụng Greedy Search do hạn chế về thời gian triển khai. Các phương pháp cao cấp hơn như Beam Search có thể cho kết quả tốt hơn.
- **Hạn chế Tài nguyên:** Cấu hình mô hình đã phải được giảm nhẹ để phù hợp với môi trường GPU miễn phí. Mô hình lớn hơn có thể cho kết quả tốt hơn.

# Chương 5: Kết luận

### 5.1. Tóm tắt các kết quả chính

Nghiên cứu này đã triển khai và đánh giá thành công hai phương pháp tiếp cận cho bài toán dịch máy Anh-Việt chuyên ngành y khoa. Các kết quả chính bao gồm:

Mô hình Transformer được huấn luyện từ đầu đã đạt được điểm BLEU là 15, cho thấy khả năng học được các quy luật dịch thuật cơ bản nhưng còn nhiều hạn chế.

Mô hình ViT5-base sau khi được finetuning đã đạt được hiệu năng vượt trội với điểm BLEU lên tới 43.94, chứng tỏ sức mạnh của phương pháp học chuyển giao.

Phân tích định tính cho thấy mô hình finetuned có khả năng xử lý tốt các câu dài, các thuật ngữ y khoa phức tạp và tạo ra các bản dịch tự nhiên, mạch lạc. Tuy nhiên, mô hình vẫn còn gặp phải một số lỗi về ngữ cảnh và sai lệch trong quá trình giải mã.

## 5.2. Kết luận

Qua quá trình thực nghiệm, chúng tôi rút ra kết luận cốt lõi: Học chuyển giao thông qua việc finetuning các mô hình ngôn ngữ lớn là một phương pháp cực kỳ mạnh mẽ và hiệu quả để giải quyết các bài toán NLP chuyên ngành.

Việc tận dụng kiến thức nền sâu rộng về ngôn ngữ đã được huấn luyện trước không chỉ giúp mô hình đạt được chất lượng dịch vượt trội mà còn tiết kiệm đáng kể thời gian và tài nguyên tính toán so với việc huấn luyện một mô hình từ đầu. Kết quả của bài tập này là một minh chứng rõ ràng cho xu hướng phát triển hiện nay trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên.