TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VIỆN TRÍ TUỆ NHÂN TẠO

BÁO CÁO MÔN HỌC
HỌC SÂU

Đề tài: Hệ thống dịch máy

Nhóm thực hiện:

1. Ngô Văn Kiệt - 22022643
2. Mai Thanh Tùng - 21021645
3. Nguyễn Bình Minh - 22022579
4. Phạm Thành Nam - 22022618

Giảng viên hướng dẫn:

TS. Triệu Hải Long

Hà Nội, 12.2024

Mở đầu

Dịch máy hiện nay đang trở thành một công cụ không thể thiếu trong việc xóa bỏ rào cản ngôn ngữ và thúc đẩy sự giao tiếp toàn cầu, đặc biệt trong bối cảnh kỷ nguyên số hóa đang phát triển mạnh mẽ. Nó không chỉ hỗ trợ kết nối giữa các cá nhân mà còn mở rộng cơ hội hợp tác trong các lĩnh vực như kinh doanh, giáo dục, và nghiên cứu khoa học.

Trong dự án này, chúng em tập trung nghiên cứu và phát triển một mô hình Dịch Máy Thần Kinh (Neural Machine Translation - NMT) tối ưu hóa dành riêng cho các trang web. Mục tiêu của dự án là xây dựng một hệ thống dịch tự động có độ chính xác cao, đảm bảo khả năng dịch linh hoạt giữa các ngôn ngữ nguồn và ngôn ngữ đích. Hệ thống này được thiết kế dựa trên các kỹ thuật tiên tiến trong lĩnh vực Deep Learning nhằm mang lại hiệu quả vượt trội so với các phương pháp truyền thống.

Dự án sử dụng kiến trúc Transformer, một công nghệ nổi bật trong xử lý ngôn ngữ tự nhiên (NLP) nhờ khả năng nắm bắt ngữ cảnh dài hạn và mối quan hệ phức tạp giữa các từ trong câu. Tập dữ liệu song ngữ lớn đã được thu thập từ nhiều nguồn đáng tin cậy và trải qua các bước tiền xử lý kỹ lưỡng để đảm bảo chất lượng dữ liệu đầu vào. Các thành phần cốt lõi của Transformer, bao gồm cơ chế Attention và kiến trúc Encoder-Decoder, đã được tinh chỉnh để tối ưu hóa quá trình huấn luyện và dự đoán.

Hiệu quả của mô hình được đánh giá dựa trên các chỉ số chuẩn ngành như BLEU (Bilingual Evaluation Understudy) và ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Kết quả cho thấy mô hình Transformer đạt hiệu suất tốt, với khả năng dịch chính xác, tự nhiên và phù hợp với ngữ cảnh. Dự án không chỉ chứng minh tiềm năng ứng dụng của công nghệ NMT mà còn mở ra cơ hội cải thiện trải nghiệm người dùng trong việc truy cập thông tin đa ngôn ngữ trên các trang web.

Báo cáo gồm .. chương:

Phần 1: Giới thiệu chung

Phần 2: Tổng quan lý thuyết

Phần 3: Phương pháp

Phần 4: Bộ dữ liệu

Phần 5: Thực nghiệm và đánh giá

Phần 6: Các thách thức và hướng phát triển

Phần 7: Kết luân

Nhiệm vụ các thành viên

Thành viên	Nhiệm vụ	Đóng góp
Ngô Văn Kiệt	Tiền xử lý, code model seq2seq	25%
Mai Thanh Tùng	Tiền xử lý, code model transformer	25%
Nguyễn Bình Minh	Tổng hợp kết quả, lập kế hoạch, code giao diện, tối ưu hệ thống	25%
Phạm Thành Nam	Thu thập dữ liệu, trích xuất nội dung URL, làm slide	25%

Bảng 1: Phân công nhiệm vụ trong nhóm

Mục lục

1 Giới thiệu chung

1.1 Bối cảnh chung

Trong xu hướng toàn cầu hóa và chuyển đổi số, giao tiếp ngôn ngữ giữa các quốc gia và vùng lãnh thổ trở thành nhu cầu thiết yếu. Yêu cầu về thông dịch trở nên mạnh mẽ hơn khi con người đang hướng tới một thế giới được hòa nhập chung sống mà không có bất kì rào cản về ngôn ngữ nào.

Và dịch máy từ lâu đã trở thành vấn đề được bàn tán khá nhiều. Từ người dùng phổ thông cho đến các chuyên gia trong lĩnh vực dịch thuật. Dịch máy bắt đầu từ sự phát triển của mạng lưới Internet và toàn cầu hóa. Không khó để có thể tiếp cận được các công cụ dịch máy trực tuyến. Bởi nó vừa dễ dàng tiếp cận vừa dễ dàng sử dụng mà không có bất cứ quy tắc nào về tuổi tác hay đối tượng người dùng.

Các trang điện tử, là cửa ngõ chính trong kỷ nguyên số, thường yêu cầu khả năng hỗ trợ đa ngôn ngữ để phục vụ người dùng trên toàn thế giới. Tuy nhiên, nhiều hệ thống dịch hiện tại gặp khó khăn trong việc dịch các ngôn ngữ phức tạp hoặc ngôn ngữ ít được hỗ trợ.

Vì vậy, dự án này của chúng em được khởi tạo để giải quyết những hạn chế này bằng việc ứng dụng Deep Learning, cụ thể là Neural Machine Translation (NMT). Phương pháp này đã chứng minh được khả năng cải thiện hiệu suất và chất lượng dịch thông qua việc học tập dữ liệu lớn và cơ chế Attention.

1.2 Mục tiêu của dự án

Dự án của chúng em đặt mục tiêu xây dựng một mô hình dịch máy không chỉ chính xác mà còn linh hoạt, phù hợp với các đặc thù ngôn ngữ và bối cảnh sử dụng. Hệ thống được thiết kế để đáp ứng nhu cầu thực tiễn của người dùng thông qua các tính năng sau:

- Thân thiện và dễ sử dụng: Hệ thống sẽ được tối ưu hóa để người dùng có thể dễ dàng truy cập và thao tác, ngay cả với những người không có nền tảng kỹ thuật.
- Hỗ trợ dịch trang web trực tiếp qua URL: Người dùng chỉ cần cung cấp đường link URL của trang web cần dịch, hệ thống sẽ tự động xử lý và hiển thị phiên bản dịch.
- Đảm bảo độ chính xác và giữ nguyên ngữ cảnh: Hệ thống tập trung vào việc cung cấp bản dịch có độ chính xác tương đối cao, đồng thời duy trì ngữ nghĩa và ngữ cảnh của nội dung gốc, giúp bản dịch trở nên tự nhiên và dễ hiểu.

Với các tiêu chí trên, dự án không chỉ hướng tới việc tạo ra một công cụ dịch thuật mạnh mẽ mà còn góp phần nâng cao trải nghiệm người dùng khi tiếp cận thông tin đa ngôn ngữ.

1.3 Phạm vi ứng dụng

Hệ thống dịch máy do chúng em phát triển được thiết kế để hỗ trợ các ngôn ngữ chính như Tiếng Anh, Tiếng Việt và một số ngôn ngữ khác, đáp ứng nhu cầu dịch thuật đa dạng trong nhiều lĩnh vực. Dự án không chỉ tập trung vào tính chính xác mà còn đảm bảo khả năng ứng dụng linh hoạt, mang lại giá trị thiết thực cho người dùng.
Các ứng dụng tiềm năng của hệ thống bao gồm:

- Thương mại điện tử: Hỗ trợ người mua hàng và nhà cung cấp giao tiếp dễ dàng, phá bỏ rào cản ngôn ngữ, từ đó mở rộng cơ hội kinh doanh giữa các quốc gia.
- Nền tảng giáo dục: Cung cấp tài liệu học tập bằng nhiều ngôn ngữ, hỗ trợ học sinh, sinh viên và giáo viên tiếp cận nguồn tài nguyên phong phú trên toàn cầu.
- Truyền thông và tin tức: Dịch nhanh chóng các bài báo, tin tức hoặc nội dung trực tuyến, giúp người đọc tiếp cận thông tin từ nhiều nguồn khác nhau một cách dễ dàng và chính xác.

Dự án nhấn mạnh tầm quan trọng của công nghệ Học sâu (Deep Learning) trong việc phát triển các giải pháp dịch máy mạnh mẽ và linh hoạt. Bằng cách áp dụng các mô hình hiện đại như Transformer, chúng em hướng đến việc mang lại trải nghiệm dịch thuật hiệu quả, phù hợp với nhu cầu thực tế và không ngừng cải thiện chất lượng dịch thuật.

2 Tổng quan lý thuyết

2.1 Dịch máy (Machine Translation - MT)

Dịch máy là lĩnh vực nghiên cứu trong trí tuệ nhân tạo (AI) và xử lý ngôn ngữ tự nhiên (NLP), nhằm tự động hóa việc dịch văn bản từ ngôn ngữ nguồn sang ngôn ngữ đích. Tầm quan trọng của MT đã tăng đáng kể trong bối cảnh toàn cầu hóa, nơi giao tiếp đa ngôn ngữ đóng vai trò quan trọng trong kinh doanh, giáo dục, và giao lưu văn hóa.

2.2 Các phương pháp dịch máy truyền thống

2.2.1 Dịch máy dựa trên quy tắc (Rule-Based Machine Translation - RBMT)

Dịch máy dựa trên quy tắc (RBMT) là hình thức dịch máy xuất hiện đầu tiên, sử dụng cơ sở dữ liệu đã có từ trước để thực hiện dịch. Phương pháp RBMT sử dụng tập hợp quy tắc ngữ pháp, từ vựng, và câu trúc của ngôn ngữ để thực hiện dịch. Quy trình bao gồm 3 bước chính:

- **Phân tích ngôn ngữ mục tiêu**: Xác định các thành phần của văn bản gốc như cấu trúc câu và ngữ pháp.
- **Chuyển đổi**: Dựa vào quy tắc ngôn ngữ đã được định nghĩa.
- **Tổng hợp**: Tạo văn bản trong ngôn ngữ đích.

Đặc điểm: Phương pháp RBMT cho kết quả đều đặn nhưng phụ thuộc lớn vào tập hợp quy tắc, yêu cầu nhiều công sức để xây dựng và duy trì.

2.2.2 Dịch máy dựa trên thống kê (Statistical Machine Translation - SMT)

Dịch máy dựa trên thống kê (SMT) dựa vào mô hình xác suất để tìm kiếm và đề xuất dịch. Các bước thực hiện bao gồm:

- Thu thập dữ liệu song ngữ.
- Huấn luyện mô hình xác suất dựa trên các cặp câu.
- Tối ưu hóa dựa trên xác suất của các bản dịch.

Đặc điểm: SMT yêu cầu khối lượng dữ liệu huấn luyện lớn, nhưng mang lại tính linh hoạt trong nhiều ngữ cặp.

2.3 Dịch máy dựa trên học sâu (Neural Machine Translation - NMT)

2.3.1 Giới thiệu NMT

Dịch máy dựa trên học sâu (NMT) là phương pháp tiên tiến sử dụng mô hình học sâu (Deep Learning) dựa trên mạng não nhân tạo. Khác với SMT, NMT không dựa vào từng thành phần rời rạc mạng não nhân tạo. Khác với SMT, NMT không dựa vào từng thành phần rời rạc mà xem văn bản như một thể thống nhất.

2.3.2 Kiến trúc tổng quan

- **Encoder-Decoder:** Mô hình bao gồm hai phần chính:
  - **Encoder:** Mã hóa chuỗi ngôn ngữ nguồn thành vector đại diện.
  - **Decoder:** Giải mã vector đại diện thành chuỗi ngôn ngữ đích.
- **Attention Mechanism:** Tối ưu hóa bằng cách tăng tầm nhìn vào từng phần của chuỗi nguồn khi dịch.
- **Transformer Architecture:** Cơ sở cho nhiều mô hình hiện đại, như GPT hoặc BERT.

2.3.3 Ưu điểm của NMT

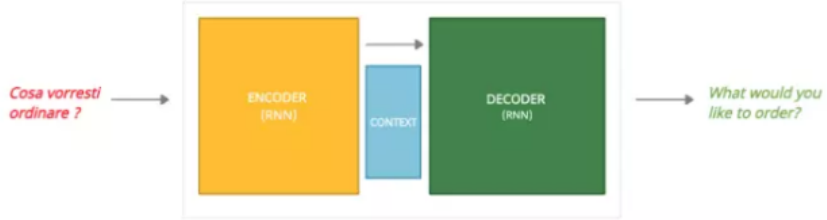
- **Hiệu suất cao trong việc dịch ngữ cặp phức tạp:** Mô hình có khả năng xử lý các ngữ cặp phức tạp và mối quan hệ giữa các từ trong câu, cải thiện độ chính xác và tự nhiên khi dịch các văn bản dài hoặc có ngữ nghĩa đa chiều.
- **Tự động hóa việc học đại diện tự nhiên cho từ vựng và ngữ pháp:** Mô hình tự động học các đại diện từ vựng và ngữ pháp, giúp nâng cao chất lượng dịch mà không cần can thiệp thủ công.
- **Dễ dàng tích hợp và cải thiện nhờ khả năng học từ dữ liệu:** Hệ thống có thể dễ dàng tích hợp vào các ứng dụng hiện tại và cải thiện theo thời gian nhờ khả năng học từ dữ liệu mới.

3 Phương pháp

3.1 Seq2Seq

3.1.1 Seq2Seq truyền thống

Chúng ta có thể sử dụng mô hình Seq2Seq cho nhiều bài toán như Machine Translation (dịch máy), Named Enity Recognition (Nhận diện thực thể có tên) hay Sentiment Classification (Phân loại cảm xúc), ... Mô hình seq2seq sử dụng kiến trúc encoder và decoder có độ dài đầu ra và đầu vào khác nhau. Như hình vẽ dưới đây:



Hình 1: Kiến trúc encoder và decoder

Bộ Mã hóa (Encoder) xử lý tất cả các đầu vào bằng cách chuyển đổi chúng thành 1 vector duy nhất, được gọi là context vector. Context vector này chứa tất cả thông tin bộ mã hóa có thể phát hiện ra từ đầu vào và giúp bộ giải mã đưa ra được quyết định chính xác, và context vector này sau đó cũng hoạt động như trạng thái ẩn đầu tiên của bộ giải mã.. Cuối cùng vector được gửi đến bộ giải mã (Decoder) để tạo ra chuỗi đầu ra.

3.1.2 Vấn đề trong Seq2seq truyền thống

Model truyền thống cũng khá là tốt tuy nhiên, RNN gặp một số vấn đề như:

- Vanishing Gradient: (Đạo hàm triệt tiêu) hiện tượng gradient bị biến mất khi câu quá dài, tức là RNN không thể học được những phụ thuộc xa.

- Exploding Gradient :(bùng nổ gradient) Đây là hiện tượng gradient quá lớn do tích tụ gradient ở những lớp cuối đặc biệt hay xảy ra đối với câu dài.

- Không tương thích với dữ liệu có cấu trúc (structured bias) : Ví dụ câu " Tôi rất thích học môn Toán" ở đây từ "học " và "môn Toán" có mối quan hệ với nhau so với các từ khác. Tuy nhiên cơ chế của RNN là học tuần tự từ trái sang phải (inductive bias) thiếu đi mất những cơ chế để mô hình học được những từ thực sự có liên quan đến nhau.

Ở vấn đề Vanishing gradient và Exploding Gradient gần như đã được giải quyết bởi LSTM/GRU. Trong khi đó vấn đề không tương thích với dữ liệu có cấu trúc chưa được giải quyết và **Attention đã giải quyết được vấn đề đó**.

3.2 Cơ chế Attention

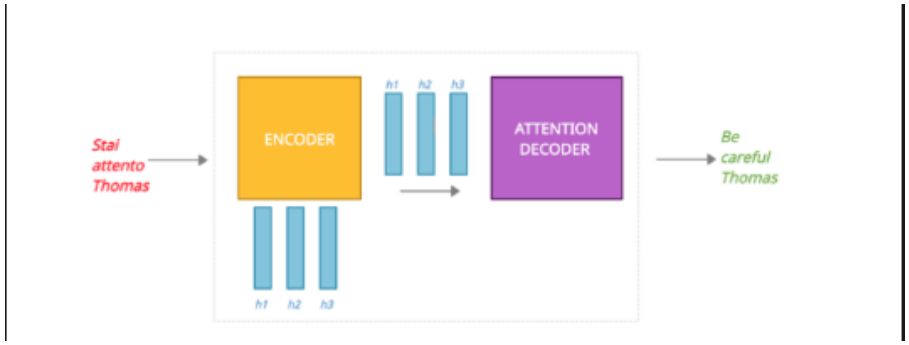
Như vừa đề cập ở trên Attention có thể giúp giải quyết vấn đề không tương thích với dữ liệu có cấu trúc, bây giờ chúng ta cùng xét 1 ví dụ.

Input: Tôi rất thích học môn Toán

Ở đây ta có thể thấy "học " và "môn Toán" có mối quan hệ với nhau so với các từ khác .Vì vậy, thay vì nhìn vào tất cả những từ trong đầu vào thì chúng ta có thể tăng tầm quan trọng của một vài từ cụ thể của đầu vào có ý nghĩa đối với đầu ra giúp mô hình dự đoán kết quả chính xác hơn. Đó là ý tưởng cơ bản của cơ chế Attention.

Model Seq2seq ở đây cũng bao gồm:Encoder, Attention Decoder.

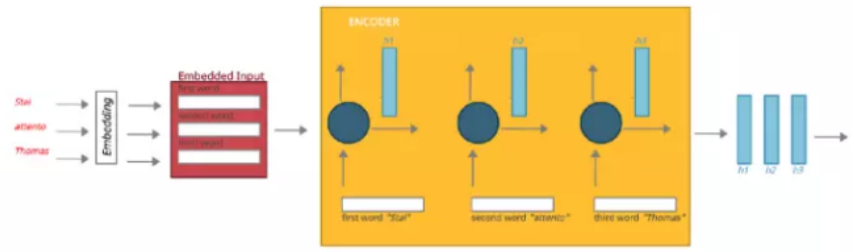
Khác với model Seq2seq truyền thống, ở đây sử dụng cả hidden state và context vector, ở trên chúng ta chỉ dùng mỗi context vector. Như hình ở dưới đây:



Ở model Seq2seq truyền thống chúng ta chỉ thấy mỗi hidden state 3 được đưa vào Decoder để dự đoán đầu ra, còn ở đây chúng ta có thể thấy tất vả hidden state được đưa vào Decoder.

3.2.1 Encoder

Trước khi được đưa vào encoder, dữ liệu của chúng ta mỗi từ sẽ được chuyển thành 1 vector thông qua quá trình embedding. Từ đầu tiên, “Stai” trong câu dưới đây, sau khi được embedding sẽ được chuyển đến encoder. Ở đây RNN tạo ra hidden state 1, tương tự cho từ thứ hai và thứ ba, các hidden state này được tạo ra từ đầu vào và các đầu vào trước đó của nó. Khi tất cả các từ trong câu của chúng ta đã được xử lý, các trạng thái ẩn (h1, h2, h2) sẽ được chuyển đến Attention Decoder.



3.2.2 Attention Decoder

Ở bước đầu tiên, bộ giải mã attention thực hiện embedding hidden state từ encoder, RNN xử lý các đầu vào và tạo ra một vector trạng thái ẩn của bộ Decoder mới (h4).

Mỗi trạng thái ẩn của bộ mã hóa được ấn định một điểm số theo công thức concat của Luong Attention paper như sau:

$$\text{score}(h_t, \bar{h}_s) = \begin{cases} h_t^\top \bar{h}_s & \text{dot} \\ h_t^\top W_a \bar{h}_s & \text{general} \\ v_a^\top \tanh(W_a[h_t; \bar{h}_s]) & \text{concat} \end{cases}$$

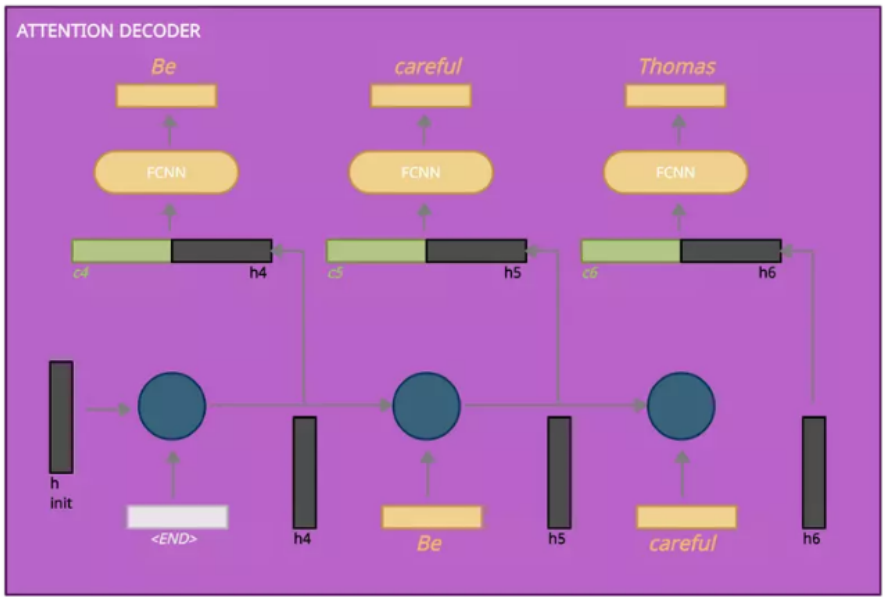
mỗi score sẽ được đưa qua công thức softmax (attention weighted):

$$a_t(s) = \text{align}(h_t, \bar{h}_s) = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_{s'}))}$$

Trạng thái ẩn của bộ mã hóa và điểm softmax liên quan được nhân với nhau, Các trạng thái ẩn(hidden states) thu được được thêm vào để có được vectơ ngữ cảnh (context vector) (c4)

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j$$

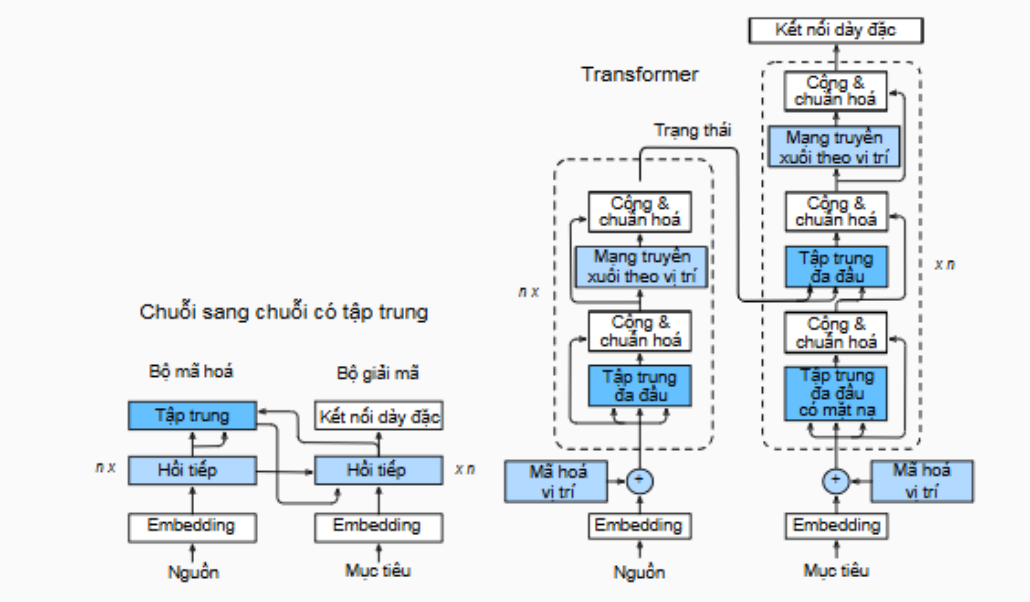
Context Vector (c4) và attention weighted được đưa vào 1 lớp RNN => Vector output, Vector này được chuyển qua một mạng fully connected neural network, được nhân với ma trận trọng số Wc và sử dụng tanh activation. Đầu ra của lớp được kết nối đầy đủ này sẽ là từ đầu ra đầu tiên của chúng ta trong chuỗi đầu ra (đầu vào: "Stai" -> đầu ra: "Be"). Ở timestep tiếp theo bắt đầu với đầu ra của bước đầu tiên ("Be") và với trạng thái ẩn của bộ giải mã (h5) được tạo ra. Tất cả được thực hiện tương tự như trên.



3.3 Transformer

- Mô hình Transformer được giới thiệu trong bài báo *"Attention is All You Need"* năm 2017, là bước tiến quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Transformer nhanh chóng trở thành nền tảng của nhiều hệ thống dịch máy hiện đại, như Google Translate, nhờ vào khả năng xử lý dữ liệu nhanh chóng và hiệu quả cao.

- Encoder only Transformer Models : Là cấu trúc Transformer chỉ sử dụng duy nhất các lớp Encoder, thường được thấy trong các tác vụ phân loại. Cấu trúc của mô hình này có thể thấy như hình sau:



Transformer Encoder-Decoder : Kiến trúc Transformer bao gồm cả Encoder và Decoder, mỗi phần được tạo thành từ nhiều lớp Self-attention và các mạng Feed Forward. Trong đó cơ chế Self-attention là trái tim của Transformer, cho phép mô hình đánh giá tầm quan trọng của các từ khác nhau trong một câu dựa trên mối liên kết của chúng với nhau. Ngoài ra, Transformer còn sử dụng Positional Encoding giúp mô hình theo dõi được vị trí của từ trong câu, quan trọng trong việc thứ tự các từ ảnh hưởng đến nghĩa của câu.

3.3.1 Encoder

Encoder nhận câu nguồn (đầu vào) và chuyển đổi thành một biểu diễn ngữ nghĩa trung gian. Mỗi lớp Encoder gồm hai thành phần chính:

- Multi-Head Self-Attention Mechanism: Tính toán mối quan hệ giữa từ nguồn.

- Feedforward Neural Network (FFN): Xử lý và trích xuất các đặc trưng từ biểu diễn ngữ nghĩa.

3.3.2 Decoder

Decoder nhận biểu diễn trung gian từ Encoder và sinh ra câu đích. Mỗi lớp Decoder gồm ba thành phần chính:

- Masked Multi-Head Self-Attention: Tập trung vào những từ đã sinh trong bước trước đó.

- Encoder-Decoder Attention: Học mối quan hệ giữa biểu diễn trung gian và ngữ cảnh nguồn.

- Feedforward Neural Network: Giống Encoder.

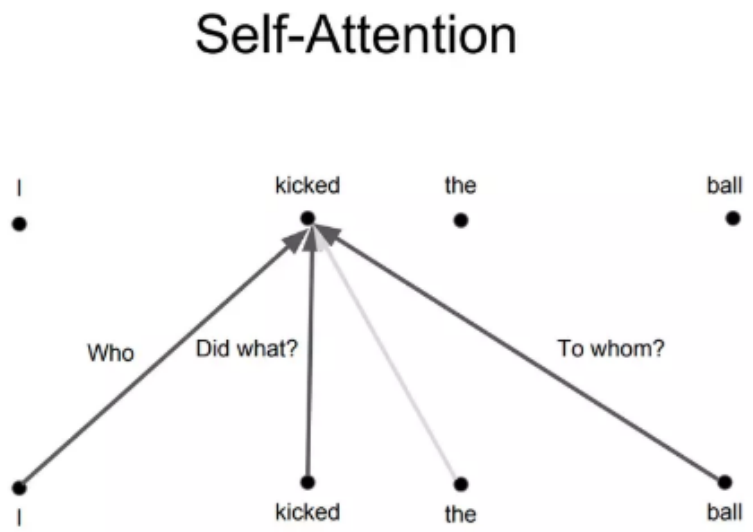
3.3.3 Cơ chế Self-Attention

Cơ chế Self-Attention là trái tim của Transformer, đồng vai trò chính trong việc xác định tầm quan trọng của mỗi từ trong câu nguồn. Self-Attention dùng ba ma trận Query (Q), Key (K) và Value (V) để tính trọng số giữa từ trong câu.

Các bước: - Tính độ tương đồng giữa Query và Key.

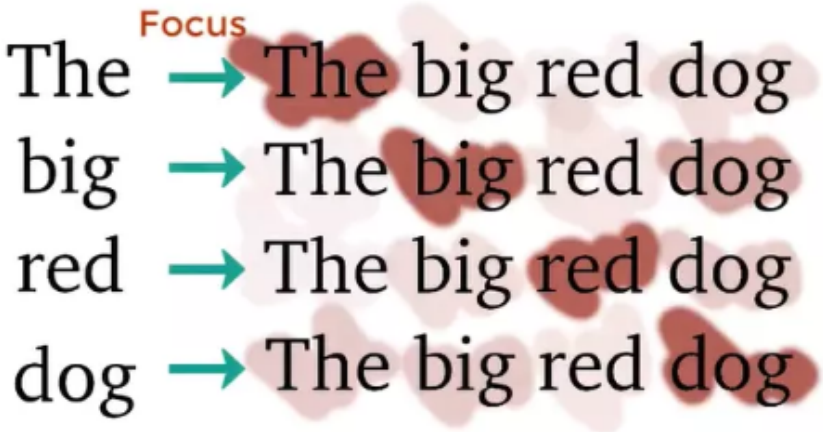
- Chuẩn hóa trọng số bằng Softmax.

- Nhân trọng số với Value để tạo biểu diễn cuối.



3.3.4 Multi-Head Attention

Multi-Head Attention giúp mô hình học nhiều mối quan hệ khác nhau trong dữ liệu bằng cách sử dụng nhiều đầu Attention song song.



3.3.5 Position Encoding

Transformer không xử lý tuần tự như RNN, do đó nó cần Position Encoding để biểu diễn thông tin vị trí từ trong câu. Thông thường, Position Encoding được tính bằng các hàm sinus và cosinus với tần số khác nhau.

4 Bộ dữ liệu

4.1 Mô tả bộ dữ liệu

Bộ dữ liệu Multi30k là một tập dữ liệu phổ biến được sử dụng trong các bài toán dịch máy và các nhiệm vụ liên quan đến xử lý ngôn ngữ tự nhiên. Bộ dữ liệu này bao gồm:

- Các câu chú thích hình ảnh bằng nhiều ngôn ngữ (ví dụ: Anh, Đức, Việt).

- Tổng cộng khoảng 30,000 mẫu, mỗi mẫu là một cặp câu (hoặc chuỗi câu) trong các ngôn ngữ khác nhau, phù hợp để xây dựng và huấn luyện các mô hình dịch máy hoặc các nhiệm vụ đa ngôn ngữ.

- Nguồn gốc dữ liệu: Bộ dữ liệu được lấy từ nguồn 'bentrevett/multi30k' thông qua nền tảng Hugging Face. Đây là một phiên bản chuẩn hóa, dễ dàng tích hợp vào các dự án nghiên cứu.

4.2 Sử dụng

- Phân chia dữ liệu: Bộ dữ liệu được chia thành ba phần chính:

- Train set: Được sử dụng để huấn luyện mô hình.

- Validation set: Dùng để tinh chỉnh các siêu tham số.

- Test set: Đánh giá hiệu suất mô hình.

- Tiền xử lý dữ liệu: Các câu được token hóa (tokenization), sau đó chuyển đổi thành các chỉ mục (indexing) sử dụng bộ từ vựng. Quá trình này được hỗ trợ bởi thư viện như 'torchtext' hoặc 'transformers'.

- Tích hợp vào pipeline: Dữ liệu sau khi xử lý được đưa vào mô hình deep learning để thực hiện các bước huấn luyện và dự đoán.

- Các thách thức: Bộ dữ liệu này có kích thước nhỏ so với các tập dữ liệu lớn hiện nay, do đó mô hình có thể gặp khó khăn trong việc tổng quát hóa. Bộ dữ liệu chưa được dịch chuẩn do đó lại càng ảnh hưởng tới hiệu suất mô hình.

5 Thực nghiệm và đánh giá

Mô hình	BLEU Score	ROUGE 1 Score
Seq2Seq	0.35	0.41
Transformer	0.52	0.88
MarianMT	0.37	0.74
MarianMT(finetune)	0.69	0.89

Bảng 2: Kết quả đánh giá

5.1 Đánh giá Seq2seq

- BLEU Score: 0.35 (thấp so với các mô hình khác).

- ROUGE Score: 0.41 (cũng khá thấp).

- Đánh giá: Đây là mô hình cơ bản nhất, và các chỉ số cho thấy nó không hoạt động tốt bằng các mô hình khác.

5.2 Đánh giá Transformer

- BLEU Score: 0.52 (cao hơn đáng kể so với Seq2seq).

- ROUGE Score: 0.88 (rất tốt).

- Đánh giá: Transformer thể hiện hiệu quả vượt trội, đặc biệt với ROUGE Score. Điều này cho thấy mô hình nắm bắt được ngữ cảnh và nội dung văn bản tốt hơn.

5.3 Đánh giá MarianMT

- BLEU Score: 0.37 (chỉ nhỉnh hơn Seq2seq một chút).

- ROUGE Score: 0.74 (tốt hơn Seq2seq, nhưng thấp hơn Transformer).

- Đánh giá: MarianMT chưa fine-tune không đạt hiệu suất cao như Transformer. Tuy nhiên, ROUGE Score cao hơn một chút so với BLEU cho thấy mô hình vẫn có khả năng tạo ra văn bản giống về nội dung.

5.4 MarianMT (finetune)

- BLEU Score: 0.69 (cao nhất trong các mô hình).

- ROUGE Score: 0.89 (cao nhất, ngang Transformer).

- Đánh giá: Sau khi fine-tune, MarianMT vượt trội so với các mô hình khác. BLEU và ROUGE Score đều rất cao, cho thấy mô hình vừa nắm bắt tốt cú pháp vừa thể hiện nội dung chính xác.

5.5 Kết luận

- MarianMT sau khi fine-tune là mô hình tốt nhất, với BLEU Score 0.69 và ROUGE Score 0.89.

- Transformer cũng là một lựa chọn rất mạnh, đặc biệt nếu bạn không có điều kiện fine-tune thêm.

- Seq2seq không phù hợp nếu bạn cần độ chính xác và chất lượng cao.

- MarianMT trước khi fine-tune có hiệu suất trung bình, nhưng có tiềm năng nếu được tinh chỉnh.

6 Các thách thức và hướng phát triển

6.1 Thách thức tồn đọng

Do việc sử dụng bộ dữ liệu multi30k là quá nhỏ nên mô hình dễ bị overfitting và không học hỏi được nhiều nếu để nó dịch một trang văn bản.

Hạn chế về thời gian và kiến thức khiến chúng em chưa đưa ra những kết quả và giải pháp tối ưu nhất cho mô hình bài toán dịch máy.

6.2 Hướng phát triển

- Trong tương lai, chúng em sẽ hiệu chỉnh các mô hình với độ phức tạp lớn hơn, sử dụng các kĩ thuật khác để giảm ảnh hưởng tiêu cực của overfitting, nghiên cứu và ứng dụng thêm các mô hình và bộ dữ liệu khác trên Hugging Face. Thử nghiệm các mô hình phức tạp và tinh vi hơn các mô hình hiện tại.

- Kết hợp dịch văn bản với nhận dạng hình ảnh để tạo ra các hệ thống dịch ngữ cảnh, phù hợp với các ứng dụng như mô tả hình ảnh đa ngôn ngữ.

- Xây dựng các ứng dụng học ngôn ngữ hoặc hỗ trợ dịch thuật tự động cho sinh viên và chuyên gia.

- Triển khai mô hình thành dịch vụ API hoặc ứng dụng web, giúp người dùng truy cập dễ dàng trên nhiều nền tảng.

7 Kết luận

Trong dự án này, chúng em đã thực hiện đa dạng các mô hình deep learning và điều chỉnh để phù hợp cho bài toán dịch văn bản. Thông qua trải nghiệm và phân tích chúng em cũng đã miêu tả đầy đủ lý do chọn mô hình và các tham số phục vụ cho yêu cầu của Web URL Translation.

Trong tương lai, dự định của chúng em sẽ nhằm nghiên cứu, thực hiện nhiều hơn các mô hình Deep Learning khác, thử nghiệm trên nhiều bộ dữ liệu và các tác vụ phức tạp hơn.

Tài liệu

1. https://viblo.asia/p/attention-trong-seq2seq-model-m68Z0J4z5kG
2. https://tiensu.github.io/blog/58$_attention$/
3. https://huggingface.co/datasets/bentrevett/multi30k
4. https://d2l.aivivn.com/chapter$_attention - mechanisms/transformer_vn.html$
5. https://viblo.asia/p/transformers-nguoi-may-bien-hinh-bien-doi-the-gioi-nlp-924lJPOXKPM