THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): *https://www.youtube.com/watch?v=DcmCpZr3Mks*
- Link slides (dạng .pdf đặt trên Github): *https://github.com/vtmphuong/CS2205.APR2023/blob/main/Ph%C6%B0%C6%A1ng%20V%C5%A9%20Th%E1%BB%8B%20Minh%20-%20xCS2205.DeCuong.FinalReport.Slide.pdf*

• Họ và Tên: Vũ Thị Minh Phương • MSSV: 220104012	• Lớp: CS2205.APR2023 • Tự đánh giá (điểm tổng kết môn): 8/10 • Số buổi vắng: 1 • Link Github: https://github.com/vtmphuong/CS2205.APR2023

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ỨNG DỤNG DỊCH MÁY ANH - ĐỨC DỰA VÀO MÔ HÌNH TRANSFORMER

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ENGLISH - GERMAN TRANSLATION APPLICATION BASED ON THE TRANSFORMER MODEL

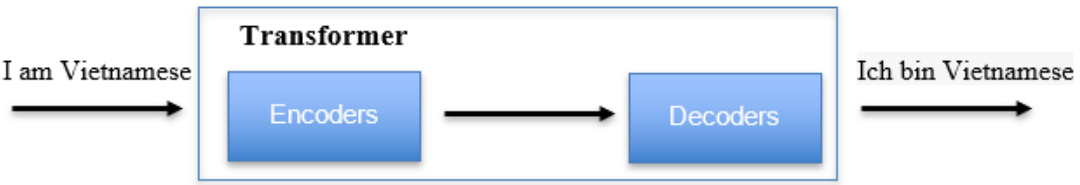
TÓM TẮT *(Tối đa 400 từ)*

Trong thời đại toàn cầu hóa hiện nay, giao tiếp đa ngôn ngữ đã trở thành một yêu cầu quan trọng trong việc xây dựng và duy trì các mối quan hệ quốc tế. Dịch máy là một lĩnh vực nghiên cứu ngày càng phát triển để giải quyết khó khăn trong việc hiểu và dịch qua lại giữa các ngôn ngữ khác nhau. Trong nghiên cứu này, chúng tôi đề xuất xây dựng một hệ thống dịch máy từ tiếng Anh sang tiếng Đức dựa trên mô hình Transformer. Mô hình này sử dụng cấu trúc chú ý (self-attention) để tạo ra sự tương tác giữa các từ trong một câu. Điều này cho phép mô hình "tự chú ý" đến các phần quan trọng của câu trong quá trình dịch, giúp cải thiện khả năng hiểu ngữ cảnh và xử lý các mối quan hệ phức tạp trong câu. Hệ thống dịch máy này nhằm nâng cao chất lượng, hiệu suất cao hơn so với các phương pháp truyền thống, đồng thời cung cấp khả năng xử lý hiệu quả cho các câu có độ dài biến đổi.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Hiện nay, có nhiều ngôn ngữ khác nhau với số lượng ngôn ngữ lớn như vậy đã gây ra rất nhiều khó khăn trong việc trao đổi thông tin. Để có thể trao đổi thông tin phải cần đến một hệ thống dịch các văn bản, tài liệu từ tiếng này sang tiếng khác. Vì vậy, con người đã nghĩ đến việc thiết kế một hệ thống tự động trong việc dịch. RNN, CNN, LSTM là những phương pháp được sử dụng rộng rãi trong mô hình ngôn ngữ và dịch máy. Tuy nhiên các phương pháp này còn gặp nhiều khó khăn trong việc xử lý thông tin từ xa và gặp vấn đề về hiệu suất tính toán song song. Năm 2017, Google công bố bài báo “Attention Is All You Need” thông tin về Transformer như tạo ra bước ngoặt

mới trong lĩnh vực xử lý ngôn ngữ tự nhiên. Kiến trúc transformer gồm hai thành phần chính: encoder và decoder cho phép thực hiện các phép tính song song nên giảm đáng kể thời gian huấn luyện, tận dụng được sức mạnh tính toán của multi-GPU. Transformer ra đời kế thừa ý tưởng từ self attention từ LSTM, loại bỏ hoàn toàn tính tuần tự phụ thuộc hoàn toàn vào cơ chế Attention để tính toán ra được mối tương quan giữa input và output. Trong nghiên cứu này, chúng tôi xây dựng một hệ thống dịch bằng cách sử dụng mô hình Transformer. Đầu vào(input) sẽ là một dữ liệu tiếng Anh và đầu ra(output) là dữ liệu tiếng Đức tương ứng.



Ảnh minh họa

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Nghiên cứu để hiểu rõ về cấu trúc và hoạt động của mô hình Transformer, cơ chế Self-attention
- Xây dựng bộ dữ liệu và huấn luyện dữ liệu
- Áp dụng mô hình Transformer để xây dựng được hệ thống dịch giúp tiết kiệm thời gian và cho ra kết quả chính xác cao

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Để thực hiện được mục tiêu đã đề ra, chúng tôi đã lên kế hoạch nghiên cứu như sau:

Nội dung:

- Nghiên cứu về dịch máy và mô hình Transformer, cơ chế Self-attention
- Xây dựng tập dữ liệu huấn luyện từ tiếng Anh sang tiếng Đức

- Huấn luyện và tinh chỉnh mô hình Transformer
- Đánh giá và so sánh hiệu suất của hệ thống dịch máy với các mô hình khác
- Xây dựng chương trình minh họa

Phương pháp:

- Tìm hiểu và hệ thống các kiến thức về:
  - Tổng quan về dịch máy, một số các cách tiếp cận dịch máy
  - Kiến trúc mô hình Transformer, bộ mã hóa, bộ giải mã của mô hình
  - Cách thức hoạt động và tầm quan trọng của cơ chế Self-attention
- Xử lý tập dữ liệu
  - Chuẩn bị bộ dữ liệu phù hợp có sẵn tiếng Anh và tiếng Đức hoặc tự xây dựng bộ dữ liệu gồm: Tập huấn luyện, Tập đánh giá, Tập kiểm tra
  - Huấn luyện bộ dữ liệu trên
- Xây dựng chương trình thử nghiệm
  - Cài đặt thư viện torchtext, transformer,...
  - Viết chương trình sử dụng mô hình transformer bằng code Python
  - Chạy thử nghiệm chương trình transformer và mô hình RNN
- Đánh giá kết quả chạy thử nghiệm
  - Dựa trên số điểm của Bilingual Evaluation Understudy Score (BLEU score)
  - Đánh giá và so sánh hiệu suất của Transformer và RNN

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Trau dồi được thêm kiến thức về dịch máy và Transformer
- Cài đặt và thử nghiệm thành công mô hình Transformer áp dụng cho cặp ngôn ngữ Anh – Đức
- Hiệu suất dịch máy sử dụng mô hình Transformer cao hơn mô hình dịch máy sử dụng mô hình trước (RNN, LSTM,..)
- Có thể ứng dụng được hệ thống dịch cho các cặp ngôn ngữ khác

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N. Gomez, A., Kaiser L., & Polosukhin, I. (2017, June 12). *[1706.03762] Attention Is All You Need*. arXiv. Retrieved May 12, 2023, from https://arxiv.org/abs/1706.03762

[2] Bui, M. Q. (2021, March 29). *Tản mạn về Self Attention*. Viblo. Retrieved May 12, 2023, from https://viblo.asia/p/tan-man-ve-self-attention-07LKXoq85V4

[3] Nguyen, A. V. (2020, May 1). *Transformers - "Người máy biến hình" biến đổi thế giới NLP*. Viblo. Retrieved May 12, 2023, from https://viblo.asia/p/transformers-nguoi-may-bien-hinh-bien-doi-the-gioi-nlp-924lJPOX

PM