TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Ngô Phù Hữu Đại Sơn

Sử dụng GNN biểu diễn các Embeddings cho mô hình Transformer

ĐỒ ÁN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

Tp. Hồ Chí Minh, tháng 07/2022

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Ngô Phù Hữu Đại Sơn- 18120078

Sử dụng GNN biểu diễn các Embeddings cho mô hình Transformer

ĐỒ ÁN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

NGƯỜI HƯỚNG DẪN
Nguyễn Ngọc Thảo

Tp. Hồ Chí Minh, tháng 07/2022

Lời cảm ơn

Trải qua thời gian dài học tập trong trường, đã đến lúc những kiến thức của em được vận dụng vào thực tiễn công việc. Em lựa chọn làm khóa luận tốt nghiệp để tổng hợp lại kiến thức của mình. Đề tài của em là: "Sử dụng GNN biểu diễn các Embeddings cho mô hình Transformer". Trong suốt quá trình làm khóa luận, em đã nhận được sự hướng dẫn, giúp đỡ quý báu của các thầy cô, các anh chị và các bạn. Em xin được bày tỏ lời cảm ơn chân thành tới:

Th.S Nguyễn Ngọc Thảođã hướng dẫn và truyền đạt những kinh nghiệm quý báu cho em trong suốt thời gian làm khóa luận tốt nghiệp của mình.

Em cũng cảm ơn gia đình và bạn bè đã giúp đỡ em hoàn thành tốt khóa luận.

Khóa luận của em còn những hạn chế về năng lực và những thiếu sót trong quá trình nghiên cứu. Em xin lắng nghe và tiếp thu những ý kiến của giáo viên phản biện để hoàn thiện, bổ sung kiến thức.

Em xin chân thành cảm ơn!

Mục lục

Danh sách hình

Danh sách bảng

Chương 1

Giới thiệu

1.1 Đặt vấn đề

Ngôn ngữ là một loại phương tiện giúp con người có thể giao tiếp và truyền đạt suy nghĩ, ý kiến của mình cho những người xung quanh. Theo trang Ethnologue.com, tính đến năm 2022, trên thế giới có 7151 ngôn ngữ. Vì vậy, một người không thể nào học và hiểu hết mọi ngôn ngữ trên thế giới. Từ đó, có thế độ cần thiết của việc dịch từ một ngôn ngữ sang một ngôn ngữ khác. Ngành khoa học máy tính, cụ thể hơn là xử lý ngôn ngữ tự nhiên, chúng ta cũng tâm đến bài toán trên.

1.2 Bài toán dịch máy (Machine Translation)

Bài toán dịch máy là một lĩnh vực trong ngành khoa học máy tính. Đầu vào được nhập vào máy tính là một đoạn văn bản từ một ngôn ngữ (ngôn ngữ nguồn). Và qua quá trình xử lý, đưa ra được đoạn văn bản tương ứng ở một ngôn ngữ khác (ngôn ngữ đích). Khác với các mô hình xử lý ngôn ngữ khác, khi chúng chỉ cần một bộ ngữ liệu của một ngôn ngữ. Các mô hình dịch máy cần ích nhất hai bộ ngữ liệu của ngôn ngữ nguồn và ngôn ngữ đích. Bộ ngữ liệu bao gồm tập các đoạn văn bản. Với mỗi

đoạn văn bản từ ngôn ngữ nguồn sẽ được ánh xạ đến một đoạn văn bản có ý nghĩa tương ứng ở ngôn ngữ đích. Các bộ ngữ liệu này có thể tổng hợp từ các nguồn khác nhau: từ các bản dịch (subtitle) của các bộ phim, bản dịch sách và cả các bộ dữ liệu được thiết kế riêng cho bàn toàn dịch máy (các bản dịch từ các chuyên gia).

Để có thể hiểu hơn sâu hơn bài toán và tìm ra hướng giải quyết, ta cần phải hiểu được cách tự nhiên mà con người dịch một đoạn văn bản ngôn ngữ nguồn sang ngôn ngữ đích như thế nào. Quá trình này có thể chia làm hai bước:

- Trích xuất ngữ nghĩa (context) của ngôn ngữ nguồn thành thông tin.
- Chuyển hóa thông tin thu thập được thành ngôn ngữ đích.

1.3 Các cách tiếp cận trước

Với tính chất trên của bài toán, ta có thể thấy bài toán này là một bài toán sequence-to-sequence(S2S) và có thể giải quyết bằng kiến trúc encoder-decoder. Các mô hình sử dụng các mạng nơ ron hồi quy (Recursive Neural Network) sử dụng cơ chế long-short-term memory và cả cơ chế attention.

Các mô hình hồi quy (Recurrent model) cho các các kết quả tốt trong bài toán dịch máy. Tuy nhiên, các mô hình này lại sử dụng cơ chế hồi quy. Ở mỗi bước tính toán, mô hình sử dụng thông tin ẩn (hiden state) được tổng hợp từ đầu văn bản đến hiện tại h_t để để làm đầu vào tính toán. Quá trình này lặp lại cho mỗi bước. Từ đó, ta có thể thấy mô hình toán toán bước tiếp theo phải phụ thuộc vào bước trước đó, dẫn đến không thể song song quá trình tính toán này được. Điều này khiến cho việc tối ưu thời gian huấn luyện lẫn hiệu quả tính toán của mô hình trước nên khó khăn.

1.4 Transformer

Trong khi đó, Transformer - một phương pháp dựa hoàn toàn trên cơ chế attention, bỏ qua các cấu trúc của mạng nơ ron hồi quy và mạng nơ ron tích chập phức tạp, giúp đơn giản hóa mô hình những vẫn thể hiện được độ hiệu quả của mô hình. Theo **Paper attention is all your need**, mô hình cho kết quả 41.8 BLEU khi huấn luyện trên tập WMT 2014, cao hơn tất cả các mô hình dịch máy trước đó.

Nhờ không sử dụng kiến trúc RNN, Transformer tránh được nhược điểm chí mạng của các mô hình loại này. Dựa hoàn toàn vào cơ chế attention, cụ thể hơn là giới thiệu kiến trúc multihead-attention giúp việc song song tính toán mô hình. Từ đó mà tăng độ hiệu quả huấn luyện cũng như hiệu quả tính toán.

1.5 Tokenization & Word Embeddings

Trong bài toán dịch máy, các đoạn văn bản thô cần phải được tiền xử lý, chuyển đổi thành các dạng dữ liệu mà máy tính của thể hiểu được. Quá trình đó được thực hiện như sau:

- Tokenization là quá trình tách đoạn văn bản ra thành các từ thành phần (token). Các token này đã được quy định sẵn trong một tập các từ đã biết trước (vocabulary). Với các từ lạ, không thuộc trong vocabulary sẽ được đánh dấu là UNK (unknown).

- Các token sau khi được tách ra vẫn chưa thể đưa vào mô hình do chúng vẫn ở dạng dữ liệu mà các mô hình chưa thể hiểu. Do đó, với mỗi token, ta cần chuyển chúng thành các vector N chiều (N chọn trước) mang tính chất và đại diện cho từ đó. Với mỗi chiều của vector được biểu diễn bằng một số thực. Các vector này được gọi là các Word Embeddings. Các Word Embeddings sẽ là đầu vào của mô hình. Một số phương pháp để tính toán các Word Embeddings trước

đây gồm: ...

1.6 Graph Convolution Network (GCN)

Khai thác quan hệ giữa các điểm dữ liệu với nhau là một đề tài được nhiều sự quan tâm trong học máy. Trong các mô hình học sâu trước đây, ta chỉ có thể trích xuâts được các mối quan hệ thông thường từ các bộ dữ liệu Euclidian.

Trong thực tế, không phải mọi dữ liệu đều biểu diễn ở dạng Euclidian. Do đó, để khai thác được thông tin từ các bộ dữ liệu đồ thị (non-Euclidian), ta cần sử dụng các mô hình Graph Neural Network (GNN). Trong vài năm qua, nhiều biến thể của đã được phát triển. Trong đó, Graph Convolution Network (GCN) là một biến thể quan trọng được xem như là biến thể cơ bản nhất của GNN.

GCN ứng dụng phép tích chập được giới thiệu trong convolution layer của CNN:

- Đối với phép tích chập trên CNN, một đoạn dữ liệu của lớp hiện tại sẽ được nhân vô hướng (tích chập) với một bộ lọc (*filter* hoặc *kernel*). Bộ lọc sẽ trượt trên bộ dữ liệu và các đầu ra của phép tích chập sẽ được truyền vào lớp tiếp theo của mạng.

- Với phép tích chập trên GCN. Ta xét từng nút(node) của đồ thị. Với mỗi nút, ta nhóm nút này với các nút kề của chính nó lại và thực hiền phép tích chập tương tự với một màn lọc. Do số lượng nút kề của mỗi nút là khác nhau, nên kích thước của filter sẽ không cố định. Phép tích chập sẽ được thực hiện trên tất cả các nút của đồ thị.

1.7 Word GCN

Ngôn ngữ là một dạng dữ liệu non-Euclidian. Do đó, trích xuất thông tin của chúng bằng các mô hình GCN được kì vọng là có kết quả tốt hơn

các phương pháp trước đó. Các loại thông tin có thể khai thác bao gồm:

- Thông tin về cú pháp (syntactic) của ngôn ngữ. Các từ trong một câu sẽ sẽ có các quy định về các mối quan hệ của chúng trong câu. Biểu diễn các mối quan hệ này là các cạnh còn các từ trong câu là các nút của đồ thị. Từ đó mà ta có thể khai thác được các đặc trưng cú pháp của các từ. Mô hình SynGCN sẽ giúp ta huấn luyện được các embeddings mang các thông tin trên

- Thông tin về ngữ nghĩa (semantic) của ngôn ngữ. Ý nghĩa của các từ sẽ có các quan hệ với nhau như: đồng nghĩa, trái nghĩa, kế thừa,... Nhờ các mối quan hệ đó, ta có thể biểu diễn được đồ thị ngữ nghĩa của các từ trong bộ từ vựng. Từ đó mà ta có thể khai thác được các đặc trưng ngữ nghĩa của các từ. Mô hình SemGCN sẽ giúp ta huấn luyện được các embeddings mang thông tin ngữ nghĩa.

1.8 Mô hình đề xuất

Ở mô hình đề xuất, tôi muốn tích hợp các embeddings được huấn luyện thông qua WordGCN để đưa vào mô hình các mô hình dịch máy. Mô hình dịch máy để tích hợp vào là transformer do tính hiệu quả cao của mô hình transformer và khả năng tính toán song song của nó.

Chương 2

Các công trình liên quan

2.0.1 Blockchain

Trong những năm gần đây, với sự mở rộng và phát triển nhanh chóng của các loại tiền ảo (cryptocurrency) như Bitcoin[], Etherum[], Zcash[], ...etc, khiến cho mức độ quan tâm đến công nghệ nền tảng của chúng, blockchain, tăng lên đáng kể. Trên thực tế, blockchain đã cho thấy công nghệ này không chỉ đóng vai trò quan trọng trong các lĩnh vực liên quan đến tài chính, mà các ứng dụng của nó đã và đang dần phổ biến rộng rãi trên nhiều lĩnh vực phi tài chính khác. Shen[] đã tổng hợp các ứng dụng của blockchain trong việc phát triển đô thị thông minh và sắp xếp chúng thành 9 danh mục, bao gồm: quản lý nhà nước(governance and citizen engagement), giáo dục, chăm sóc sức khỏe, kinh tế, giao thông vận tải, năng lượng, quản lý nước và chất thải, công trình công cộng và bảo vệ môi trường. Jaroodi[] cũng đã chỉ ra các lợi ích cũng như thách thức của việc sử dụng blockchain trong các lĩnh vực: tài chính, chăm sóc sức khỏe, các hoạt động thương mại, sản xuất, năng lượng, nông nghiệp và thực phẩm, tự động hóa, xây dựng, truyền thông và các ứng dụng giải trí.

Blockchain trong việc bảo vệ dữ liệu và thông tin cá nhân

Bảo mật thông tin cá nhân khi sử dụng internet đã và đang là một vấn đề nan giải khi các mạng xã hội liên tục thu thập thông tin người dùng, bao gồm các thông tin cá nhân, hoạt động và thói quen. Người sử dụng các mạng xã hội gặp khó khăn trọng việc quản lý các loại thông tin mà nhà cung cấp dịch vụ được quyền thu thập và mục đích sử dụng các thông tin này và thường không thể rút lại các quyền truy cập đã cho phép. Zyskind[] đề xuất một ý tưởng lưu trữ các chính sách truy cập thông tin trên một Blockchain và để cho các node của Blockchain truy cập thông tin từ một bảng băm phân tán (Distributed hash table). Khi người dùng muốn cấp hoặc hủy bỏ quyền truy cập vào thông tin cá nhân, Blockchain đóng vai trò như một nhà phân phối chỉ cho phép bên thứ 3 truy cập các thông tin đã được cấp quyền. Wang[] đề xuất một mô hình kết hợp hệ thống lưu trữ phân tán (decentrailize storage system), Etherum Blockchain và kĩ thuật mã hóa dựa trên thuộc tính (attribute-based ecryption). Mô hình này cho phép người dùng tìm kiếm trên các thông tin đã mã hóa dựa vào các hợp đồng thông minh (smart contract) của Etherum. Đồng thời cũng cho phép người sở hữu thông tin phân phối các khóa bí mật cho người dùng và chia sẻ thông tin thông qua các chính sách truy cập đặc tả.

Blockchain trong các hệ thống IOT

Các hệ thống lưu trữ phân tán - (Decentralize storage systems)

Các hệ thống lưu trữ phân tán hướng đến việc chia nhỏ dữ liệu và phân phối chúng lên các node trong một mạng lưới có sẵn thay vì dồn nén tất cả dữ liệu trong một server duy nhất. Dựa vào bảng A, ta nhận thấy các hệ thống lưu trữ phân tán có một số ưu điểm so với cách lưu trữ tập trung truyền thống.

- Khi một hệ thống bị tấn công, việc phân tán dữ liệu sẽ giảm thiểu

được rủi ro về thông tin bị lộ một cách toàn vẹn.

- Ngoài ra, tính tồn tại của một hệ thống lưu trữ phân tán cũng cao hơn, khi một hoặc nhiều node bị ngắt kết nối, các node còn lại có thể được lập trình để hoạt động độc lập. Điều này tốt hơn nhiều so với việc một hệ thống lưu trữ tập trung sẽ ngưng hoạt động khi server lưu trữ bị ngắt kết nối.

- Cuối cùng, một hệ thống lưu trữ phi tập trung sẽ thân thiện hơn với người dùng thông qua việc cho phép người sử dụng dịch vụ trả phí theo dung lượng đã sử dụng thay vì phải trả trước để mua dung lượng như cách làm hiện tại của các hệ thống lưu trữ tập trung.

2.0.2 Mã hóa bất đối xứng

2.0.3 Chữ kí điện tử

Chương 3

Phương pháp

3.0.1 Tạo và chuyển dữ liệu cho chủ dữ liệu

3.0.2 Chia sẻ dữ liệu ngang hàng

Chương 4

Ứng dụng và cài đặt

4.0.1 Web Application (cho nhà cung cấp nội dung)

4.0.2 Mobile Application (cho người dùng)

Tài liệu tham khảo