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

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

Sử dụng GNN biểu diễn các Embeddings cho mô hình *transformer*

1 THÔNG TIN CHUNG

Người hướng dẫn:

- TS. Nguyễn Ngọc Thảo (Khoa Công nghệ thông tin)
- ThS. Tạ Việt Phương (Trường Đại học Công nghệ Thông tin)

Sinh viên thực hiện:

1. Ngô Phù Hữu Đại Sơn (MSSV: 18120078)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *01/2022* đến *07/2022*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Các mô hình giải quyết bài toán Machine Translation sử dụng kiến trúc Attention đang thể hiện các kết quả rất tốt trong các thí nghiệm và thực tế. Một trong những phần quan trọng của những mô hình này là các mô hình encoding các từ thành các word embeddings để có thể chuyển từ ngữ thành các đầu vào có thể tính toán.

Đồ thị xuất hiện một cách tự nhiên trong nhiều lĩnh vực ứng dụng, từ phân tích xã hội, sinh học, hóa học đến thị giác máy tính. Đồ thị cho phép nắm bắt các mối quan hệ cấu trúc giữa các dữ liệu và do đó cho phép thu thập nhiều thông tin chi tiết hơn.

Một trong các phương pháp giúp encode các từ ngữ thành các vector là sử dụng Graph Convolution Neural NetWork (GCN). Lợi ích của việc sử dụng *GCN* là:

- Thể hiện được các Syntactic Context giữa các từ ngữ trong câu.
- Thể hiện được các Semantic Context giữa các từ ngữ với nhau.

Nhờ vào đó, các embeddings được học ra có khả năng biểu diễn ngữ nghĩa và cấu trúc trong câu tốt hơn, phù hợp cho các bài toán về Machine Translation.

2.2 Mục tiêu đề tài

Mục tiêu của đề tài này bao gồm: (1) - Nghiên cứu, khỏa sát các thuật toán biểu diễn từ ngữ thành các embeddings và (2) - Kết hợp cài đặt các thuật toán giúp nâng cao hiệu xuất của mô hình *transformer* để giải quyết bài toán Machine Translation.

Việc nghiên cứu khảo sát các thuật toán nhằm đưa ra được các đánh giá về ưu điểm và nhược điểm của chúng. Từ đó, cho thấy được độ hiệu quả và tiềm năng của *GCN* trong bài toán machine translation.

Việc kết hợp cài đặt nhằm chứng minh tính hiệu quả của mô hình đề xuất.

2.3 Phạm vi của đề tài

Nghiên cứu ở ([1]) đã chỉ các các hướng nghiên cứu cho mô hình transfomer. Trong đó: (1) - Cải thiện hiệu suất cho mô hình *transformer* , (2) - Khái quát hóa mô hình giúp huấn luyện với tập dữ liệu nhỏ hơn. (3) - Tăng độ thích nghi của mô hình vào các tác vụ thực tế hơn.

Đề tài của khóa luận này có phạm vi nghiên cứu giúp cải thiện hiệu suất của mô hình *transformer* dựa trên embeddings đã được huấn luyện trước([2])

Đề tài thực hiện bài toán dịch cụ thể từ tiếng Anh sang tiếng Đức.

2.4 Cách tiếp cận dự kiến

Phương pháp chính. Theo các đề xuất ở ([2]):

- Mô hình *SynGCN* để huấn luyện các word embeddings theo Syntactic Context.
- Sau đó sử dụng mô hình *SemGCN* để huấn luyện các word embeddings từ *SynGCN* theo Semantic Context.

Phương pháp đề xuất trong khóa luận nhắm tích hợp các embeddings sau khi được vào huấn luyện ở mô hình *SynGCN+SemGCN* sẽ được sử dụng để huấn luyện mô hình *transformer* được đề xuất ở ([3]) với kì vọng sẽ tăng được hiệu xuất của mô hình.

Dữ liệu thực nghiệm. Sử dụng tập dữ liệu Multi30k để huấn luyện mô hình có thể dịch từ tiếng Anh sang tiếng Đức. Sử dụng 80% dữ liệu để huấn luyện và 20% dữ liệu để kiểm thử.

Phương pháp đối sánh. Sử dụng mô hình transfomer làm mô hình baseline. So sánh hiệu suất giữa mô hình đề xuất và mô hình baseline.

2.5 Kết quả dự kiến của đề tài

Sau khi tiến hành, nghiên cứu này kỳ vọng sẽ đạt được các kết quả sau:

- Nắm được ý tưởng của việc sử dụng *GCN* để huấn luyện các word embeddings.
- Cài đặt được mô hình baseline.
- Tích hợp *SynGCN* và *SemGCN* vào mô hình *transformer* . So sánh hiệu suất của mô hình đề xuất với mô hình baseline.

2.6 Kế hoạch thực hiện

Kế hoạch thực hiện khóa luận bao gồm các giai đoạn được trình bày như sau:

Giai đoạn	Thời gian	Công việc
1	01/01/2022 - 31/01/2022	Tìm hiểu kiến thức nền tảng về *transformer* Cài đặt mô hình *transformer*
2	01/02/2022 - 28/02/2022	Tìm hiểu kiến thức nền tảng về word embeddings Tìm hiểu kiến thức nền tảng về *GCN* Tìm hiểu kiến thức nền tảng của *SynGCN* và *SemGCN*
3	01/03/2022 - 31/03/2022	Tích hợp mô hình *SynGCN* và *SemGCN* vào mô hình *transformer*
5	01/04/2022 - 30/04/2022	Chạy các thực nghiệm trên mô hình Phân tích và đánh giá kết quả
6	01/05/2022 - 31/05/2022	Viết luận văn Làm slide thuyết trình Tập thuyết trình

Bảng 1: Bảng kế hoạch thực hiện

Tài liệu

[1] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021.

[2] S. Vashishth, M. Bhandari, P. Yadav, P. Rai, C. Bhattacharyya, and P. Talukdar, "Incorporating syntactic and semantic information in word embeddings using graph convolutional networks," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, (Florence, Italy), pp. 3308–3318, Association for Computational Linguistics, July 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

TP. Hồ Chí Minh, 04/04/2022

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)