

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

Hoàng Minh Thanh - Phan Minh Tâm

DỰ ĐOÁN LIÊN KẾT TRONG
ĐỒ THỊ PHỨC
KNOWLEDGE GRAPH EMBEDDING
FOR LINKING PREDICTION

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH HOÀN CHỈNH ĐẠI HỌC

Tp. Hồ Chí Minh, tháng 09/2020

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

Hoàng Minh Thanh - 18424062

Phan Minh Tâm - 18424059

**DỰ ĐOÁN LIÊN KẾT TRONG
ĐỒ THỊ PHỨC
KNOWLEDGE GRAPH EMBEDDING
FOR LINKING PREDICTION**

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH HOÀN CHỈNH ĐẠI HỌC

GIÁO VIÊN HƯỚNG DẪN

Th.S Lê Ngọc Thành

Tp. Hồ Chí Minh, tháng 09/2020

Lời cảm ơn

Tôi xin chân thành cảm ơn t
sdfsdf

Mục lục

Lời cảm ơn	i
Mục lục	ii
1 Giới thiệu	1
2 Tổng quan	2
2.1 Cơ sở lý thuyết	2
2.2 Các nghiên cứu liên quan	2
3 Phương pháp đề xuất	3
3.1 AnyBURL	3
3.2 CGAT	3
3.2.1 Mô hình GAT	4
4 Kết quả thực nghiệm và phân tích	6
4.1 Tập dữ liệu	6
4.2 Phương pháp đánh giá	6
4.3 Phương pháp huấn luyện	7
4.4 Kết quả và phân tích	7
5 Kết luận	8
Danh mục công trình của tác giả	9

Tài liệu tham khảo	10
A Kết quả training	12
B Tập dữ liệu	13

Danh sách hình

Danh sách bảng

4.1	Mô tả tập dữ liệu	6
4.2	Kết quả thực nghiệm trên tập WN18RR và FB15K-237. Kết quả tốt nhất được bôi đen và kết quả tốt thứ hai được <u>gạch chân</u>	7

Tóm tắt

Dùng sfdefaultdf s df s defcounter

Chương 1

Giới thiệu

Ngôn [1] ngữ để viết và trình bày báo cáo khóa luận tốt nghiệp, đồ án tốt nghiệp, thực tập tốt nghiệp (sau đây gọi chung là báo cáo) là tiếng Việt hoặc tiếng Anh. Trường hợp chọn ngôn ngữ tiếng Anh để viết và trình bày báo cáo, sinh viên cần có đơn đề nghị, được cán bộ hướng dẫn (CBHD) đồng ý và nộp cho bộ phận Giáo vụ của Khoa vào thời điểm đăng ký đề tài để xin ý kiến. Báo cáo viết và trình bày bằng tiếng Anh phải có bản tóm tắt viết bằng tiếng Việt.

Chương 2

Tổng quan

sdfsdf

2.1 Cơ sở lý thuyết

sdfsdfsdfsdf

2.2 Các nghiên cứu liên quan

Sau thời kỳ ngủ đông của AI, các kỹ thuật học sâu có một bước tiến đáng kể với các mô hình CNN [0], RNN [0], Attention [8], Transformer [0] đã đạt được những kết quả rất lớn với độ chính xác cao cả trong nghiên cứu cũng như khi áp dụng vào công nghiệp. Các mô hình học sâu, cụ thể là CNN đã rất thành công để giải quyết các vấn đề về phân loại hình ảnh [0], phân đoạn ngữ nghĩa [0] và dịch máy, chúng tạo ra một bản đồ đặc trưng biểu diễn cấu trúc thông tin dưới dạng lưới. CNN cũng đã được áp dụng vào trong cấu trúc đồ thị với mô hình GCN [0].

Chương 3

Phương pháp đề xuất

sdfsdf

3.1 AnyBURL

Anytime Bottom-Up Rule Learning for Knowledge Graph Completion
[3] sdfsdf

Dùng lệnh để trích dẫn một hoặc nhiều tài liệu Lưu ý khi trích dẫn tài liệu tham khảo, cần viết câu sao cho bỏ phần trong cặp ngoặc vuông đi thì câu vẫn đầy đủ ý nghĩa. Ví dụ, thay Một ví dụ khác, thay vì viết “... như trong công trình nghiên viết “... nh’

Để chèn mã nguồn, cần dùng package listings

3.2 CGAT

Ở phần này chúng tôi sẽ giới thiệu tóm lược về mạng đồ thị chú ý (GATs [9]) và cải tiến của chúng tôi trên lớp GAT mà chúng tôi gọi là *mạng cộng tác đồ thị chú ý* CGAT, và sau đó chúng tôi áp dụng vào để xây dựng đồ thị theo mô hình KGAT [6] để tối ưu quá trình dự đoán các mối quan hệ.

3.2.1 Mô hình GAT

Mạng đồ thị tích chập (GCNs [0]) giúp tổng hợp thông tin bằng cách tính trung bình thông tin từ các thực thể lân cận, tuy nhiên cách này sẽ làm cho các thực thể có trọng số ngang bằng nhau không biểu diễn đúng thông tin trong thế giới thực. Để giải quyết vấn đề đó, GATs [9] ra đời để đối xử với các node lân cận bằng sự quan trọng của chúng.

Đầu vào của mô hình là vector biểu diễn đặc trưng của từng thực thể (entity) $E = \vec{e}_1 + \vec{e}_2 + \dots + \vec{e}_N$. Và mục tiêu của chúng ta là biến đổi thành một đặc trưng đầu ra mới $E'' = \vec{e}_1'' + \vec{e}_2'' + \dots + \vec{e}_N''$; với \vec{e}_i và $\vec{e}_i'' \in \mathcal{R}^k$ tương ứng là vector nhúng đầu vào và vector đầu ra của thực thể e_i , N là số lượng của các thực thể (nodes), k là đặc trưng đầu vào.

Mô hình sẽ đi qua hai quá trình biến đổi vector đặc trưng \vec{e}_i và có thể tóm lược như sau :

$$\vec{e}_i \longrightarrow \vec{e}_i' \longrightarrow \vec{e}_i'' \quad (3.1)$$

Ở quá trình biến đổi đầu tiên, mô hình sẽ tổng hợp thông tin từ các thực thể lân cận và ghép chồng lên nhau để tạo ra vector \vec{e}_i' sau đó mô hình sẽ dùng vector \vec{e}_i' để coi là vector nhúng của thực thể cho lớp mới và tiếp tục quá trình tổng hợp từ các thông tin lân cận và tạo ra vector \vec{e}_i'' cuối cùng.

Đầu tiên để tham số hóa quá trình biến đổi tuyến tính, ta cần một trọng số $W \in \mathbb{R}^{N_e \times k}$ ánh xạ vector đầu vào thành một vector mới với miền không gian lớn hơn và một hàm chú ý a chúng ta tùy chọn :

$$e_{ij} = a(W\vec{e}_i, W\vec{e}_j) \quad (3.2)$$

trong đó e_{ij} là giá trị chú ý của một cạnh (e_i, e_j) trong đồ thị \mathcal{G} hay e_{ij} thể hiện sự quan trọng của đặc trưng cạnh (e_i, e_j) so với thực thể e_i . Sau đó, chúng ta áp dụng hàm *softmax* qua tất cả các giá trị nhúng của hàng

xóm để tạo ra α_{ij} . Quá trình tổng hợp các sự chú ý được biểu diễn sự chú ý được thể hiện ở biểu thức sau :

$$\vec{a}_{ij} = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W} \vec{e}_j \right) \quad (3.3)$$

Mô hình GAT sẽ tiến hành *lớp chú ý đa đỉnh*(multi-head attention) để ổn định quá trình học bằng cách ghép A đầu chú ý với nhau :

$$\vec{x}_i = \parallel_{a=1}^A \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^a \mathbf{W}^a \vec{x}_j \right) \quad (3.4)$$

trong đó phép \parallel biểu diễn quá trình ghép chồng lên nhau và σ là bất kỳ hàm biến đổi phi tuyến tính nào, α_{ij}^a là hệ số chú ý được chuẩn hóa của cạnh (e_i, e_j) được tính từ lớp thứ a^{th} cơ chế chú ý. Cuối cùng \vec{x}_i được coi là vector thực thể nhúng mới và cho vào lớp chú ý đa đỉnh với đầu ra thay vì ghép chồng thì được tính trung bình như công thức sau :

$$\vec{x}_i'' = \sigma \left(\frac{1}{A} \sum_{a=1}^A \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^a \mathbf{W}^a \vec{x}_j \right) \quad (3.5)$$

[6] ===== GAT [9]

TransE [4]

Attention [8]

CAttention [5]

ConvKB [7]

Chương 4

Kết quả thực nghiệm và phân tích

sdfsdf

4.1 Tập dữ liệu

Tập dữ liệu bao gồm :

Dataset	Entities	Relations	Edges			
			Training	Validation	Test	Total
WN18RR	40,943	11	86,835	3034	3134	93,003
FB15k-237	14,541	237	272,115	17,535	20,466	310,116

Bảng 4.1: Mô tả tập dữ liệu

4.2 Phương pháp đánh giá

sdfsdf

	WN18RR					FB15K-237				
	MR	MRR	Hits@N			MR	MRR	Hits@N		
			@1	@3	@10			@1	@3	@10
TransE	7000	0.444	41.2	47	50.4	7000	0.444	41.2	47	50.4
ConvKB	7000	0.444	41.2	47	50.4	7000	0.444	41.2	47	50.4
Mô hình của chúng tôi										
AnyBURL	7000	<u>0.444</u>	41.2	47	<u>50.4</u>	7000	0.444	41.2	47	50.4
CGAT	7000	0.444	41.2	47	50.4	7000	0.444	41.2	47	50.4

Bảng 4.2: Kết quả thực nghiệm trên tập WN18RR và FB15K-237. Kết quả tốt nhất được **bôi đen** và kết quả tốt thứ hai được gạch chân

4.3 Phương pháp huấn luyện

sdfsdf

4.4 Kết quả và phân tích

sdfsdf

Chương 5

Kết luận

EXPERIMENTS RESULTS AND ANALYSIS

Dùng lện

Danh mục công trình của tác giả

1. Tạp chí ABC
2. Tạp chí XYZ

Tài liệu tham khảo

Đồ thị tri thức

- [1] Rossi, Andrea et al. “Knowledge Graph Embedding for Link Prediction: A Comparative Analysis”. In: *arXiv preprint arXiv:2002.00819* (2020).

Phương pháp AnyBURL

- [3] Meilicke, Christian et al. “Anytime Bottom-Up Rule Learning for Knowledge Graph Completion.” In: *IJCAI*. 2019, pp. 3137–3143.

Phương pháp CGAT

- [4] Bordes, Antoine et al. “Translating embeddings for modeling multi-relational data”. In: *Advances in neural information processing systems*. 2013, pp. 2787–2795.
- [5] Cordonnier, Jean-Baptiste, Loukas, Andreas, and Jaggi, Martin. “Multi-Head Attention: Collaborate Instead of Concatenate”. In: *arXiv preprint arXiv:2006.16362* (2020).
- [6] Nathani, Deepak et al. “Learning attention-based embeddings for relation prediction in knowledge graphs”. In: *arXiv preprint arXiv:1906.01195* (2019).

- [7] Nguyen, Dai Quoc et al. “A novel embedding model for knowledge base completion based on convolutional neural network”. In: *arXiv preprint arXiv:1712.02121* (2017).
- [8] Vaswani, Ashish et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.
- [9] Veličković, Petar et al. “Graph attention networks”. In: *arXiv preprint arXiv:1710.10903* (2017).

Phụ lục A

Kết quả training

Đây là phụ lục.

Phụ lục B

Tập dữ liệu

Đây là phụ lục 2.