



fit@hcmus

DỰ ĐOÁN LIÊN KẾT ĐỒ THỊ TRI THỨC DỰA TRÊN MẠNG NƠ RON TÍCH CHẬP

Sinh viên thực hiện: Lê Nhựt Nam (18120061) Phan Anh Hào (18120356)

Giảng viên hướng dẫn: Th.S Lê Ngọc Thành Th.S Tạ Việt Phương

Bộ môn Khoa học Máy tính, Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, Đại học Quốc Gia TP.HCM, Việt Nam

Ngày 12 tháng 8 năm 2022

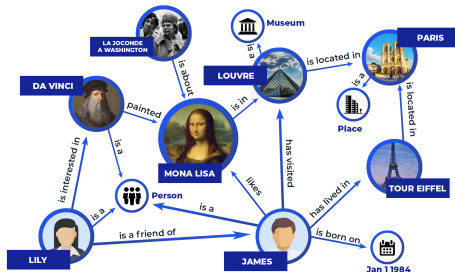
Nội dung trình bày

- ① Giới thiệu
- ② Kiến thức nền tảng
- ③ Công trình liên quan
- ④ Phương pháp đề xuất
- ⑤ Thực nghiệm và đánh giá
- ⑥ Kết luận và định hướng nghiên cứu tương lai
- ⑦ Tài liệu tham khảo

Giới thiệu

Tri thức trong dạng đồ thị:

- Các nút thể hiện những thực thể thế giới thật; Được gán nhãn với kiểu của chúng.
- Các cạnh nối thể hiện mối quan hệ giữa các thực thể.



Hình 1: Minh họa đồ thị tri thức¹

¹Introduction to Question Answering over Knowledge Graphs - Yashu Seth

Động lực nghiên cứu khoa học

Hiện nay, có nhiều kho dữ liệu đồ thị tri thức lớn (FreeBase, Wikidata, Dbpedia, YAGO, NELL) với những đặc điểm chung:

- Lớn: Chứa hàng triệu thực thể, hàng nghìn quan hệ và hàng tỷ bộ ba.
- **Chưa hoàn thiện:** Những liên kết giữa những thực thể với nhau bị mất.



Cơ sở dữ liệu FreeBase: Gần 80 triệu thực thể; Gần 30 nghìn quan hệ; Hơn 3 tỷ bộ ba dữ kiện. Tuy nhiên có gần 93.8% thông tin con người bị thiếu mất nơi sinh và hơn 78.5% không có thông tin về quốc tịch.

Nhúng đồ thị tri thức

Các thành phần cơ bản của mô hình nhúng đồ thị tri thức:

- Các véc-tơ nhúng: Không gian véc-tơ thực; không gian véc-tơ phức; Đa tạp; Nhóm.
- Hàm tính điểm: Dựa trên khoảng cách (distance); dựa trên độ tương đồng (similarity)
- Chiến lược phát sinh mẫu âm: Phá hỏng bộ ba dương theo phân phối đồng nhất; Bernoulli; Phát sinh mẫu âm đối kháng; Phát sinh mẫu âm tự đối kháng.
- Hàm mất mát: Margin-based loss; Logistic-based loss; Binary cross-entropy loss.

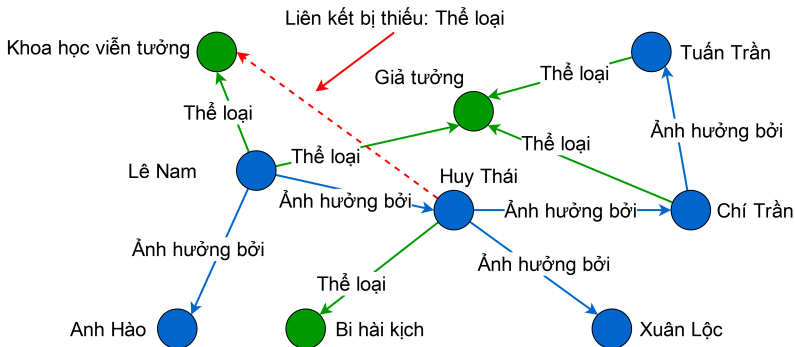
Phát biểu bài toán

Bài toán dự đoán liên kết trên đồ thị tri thức (tĩnh):

- Đối với bài toán dự đoán thực thể đầu hay xếp hạng thực thể đầu, cho trước bộ ba có dạng (?, quan hệ, thực thể đuôi), dự đoán các thực thể đầu tiềm năng tương ứng.
- Tương tự đối với bài toán dự đoán thực thể đuôi, cho trước bộ ba có dạng (thực thể đầu, quan hệ, ?), dự đoán các thực thể đuôi tiềm năng tương ứng.
- Đối với bài toán dự đoán quan hệ, cho trước (thực thể đầu, ?, thực thể đuôi), dự đoán các quan hệ tiềm năng tương ứng.

Trong khóa luận này tập trung vào bài toán dự đoán thực thể.

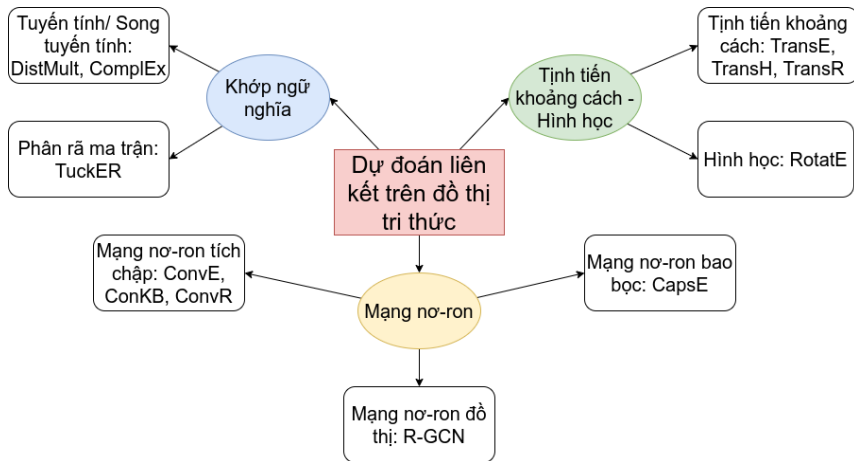
Phát biểu bài toán



Hình 2: Ví dụ dự đoán liên kết trên đồ thị tri thức đơn giản.²

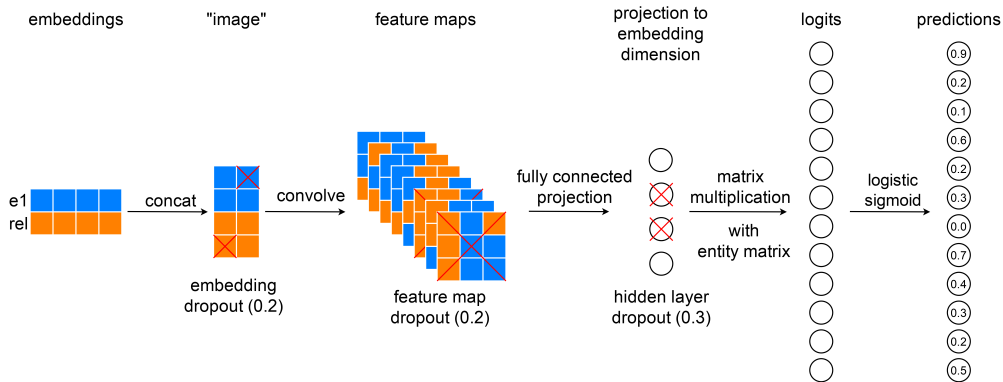
²Course CS224W: Machine Learning with Graphs, Stanford / Fall 2021, slide 10 - Knowledge Graph Embeddings và slide 11 - Reasoning over Knowledge Graphs

Công trình liên quan



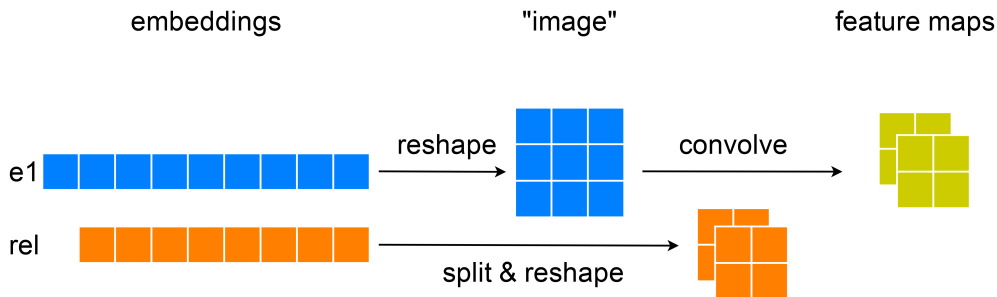
Hình 3: Các công trình liên quan

Mô hình dựa trên mạng nơ-ron



Hình 4: Mô hình dựa trên mạng nơ-ron tích chập: ConvE[1]

Mô hình dựa trên mạng nơ-ron



Hình 5: Mô hình dựa trên mạng nơ-ron tích chập: ConvR[2]

Cơ chế hiệu chỉnh

Với các feature maps \mathbf{C} thu được từ phép tích chập, tổng hợp trung bình $\mathbf{Z} = [z_1, z_2, \dots, z_n]$

$$z_i = \frac{1}{h' \times w'} \sum_{a=1}^{h'} \sum_{b=1}^{w'} c_{a,b}^{(i)} \quad (1)$$

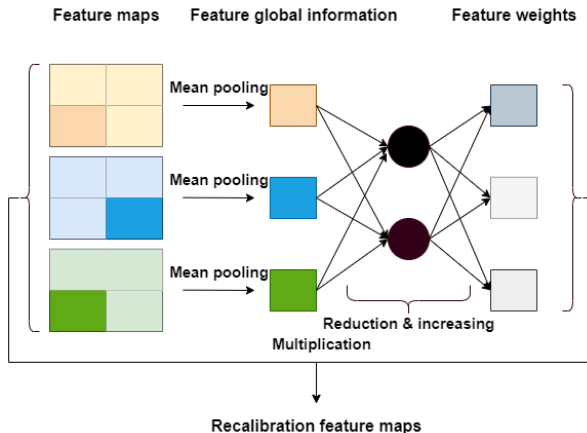
Truyền qua phép giảm và tăng (hình dạng giống với Autoencoder):

$$\mathbf{U} = \sigma(\mathbf{W}_b g(\mathbf{W}_a \mathbf{Z})) \quad (2)$$

Feature maps hiệu chỉnh được tính toán:

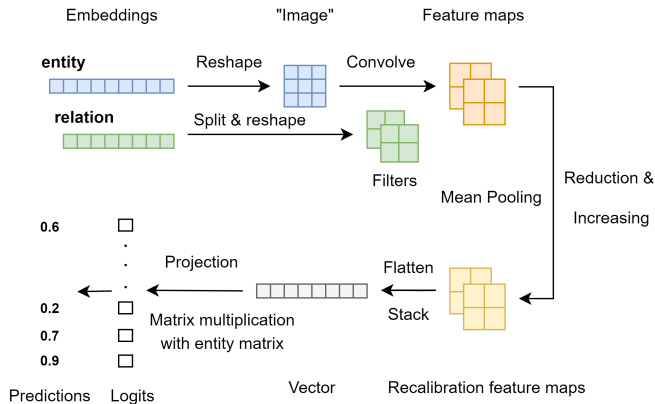
$$\mathbf{V} = \mathbf{CU} = c_i \times u_i \quad (3)$$

Cơ chế hiệu chỉnh



Hình 6: Minh họa cho cơ chế hiệu chỉnh.

Phương pháp đề xuất



Hình 7: Kiến trúc của mô hình đề xuất ACRM.

Phương pháp đề xuất

Hàm tính điểm được định nghĩa như sau:

$$f_r(h, t) = f(\mathbf{W}\mathbf{y} + b)^\top \mathbf{t} \quad (4)$$

Hàm mất mát Binary Cross-Entropy

$$\mathcal{L}(h, r) = -\frac{1}{|\mathcal{E}|} \sum_{t \in \mathcal{E}} q_t^{h,r} \log(p_t^{h,r}) + (1 - q_t^{h,r}) \log(1 - p_t^{h,r}) \quad (5)$$

Bộ tối ưu tham số học: Adam

Thuật toán huấn luyện

Algorithm 1: MÃ GIẢI CHO MÔ HÌNH ACRM

Input: Số lượng thực thể n_e ; số lượng quan hệ n_r ; số chiều nhúng thực thể d_e ; số chiều nhúng quan hệ d_r ; số lần duyệt *iterations*

Output: Điểm số cho các bộ ba

```

1 Khởi tạo tập véc-tơ nhúng thực thể  $E$ ; Khởi tạo tập véc-tơ nhúng quan hệ  $R$ ;
2 for iter to iterations do
3   for  $(h, r, t) \in \mathcal{G}$  do
4     Véc-tơ nhúng thực thể  $\mathbf{h} \in E$ ; Véc-tơ nhúng quan hệ  $\mathbf{r} \in R$ ;
5     Thay đổi dạng của  $\mathbf{h}$  thành một ma trận 2D  $\mathbf{S}$ ;
6     Véc-tơ  $\mathbf{r}$  được cắt thành các ma trận  $c$  có cùng kích thước để tạo ra các bộ lọc  $\mathbf{R}^{(i)}$ ;
7     Thực hiện tích chập  $\mathbf{S}$  với  $\mathbf{R}$ , thu được các bản đồ đặc trưng  $\mathbf{C}^{(i)}$ ;
8     Tính toán tổng hợp trung bình  $\mathbf{Z} = [z_1, z_2, \dots, z_n]$ ;
9     Thực hiện phép giảm và tăng để tính toán trọng số đặc trưng:  $\mathbf{U}$ ;
10    Tính toán bản đồ đặc trưng hiệu chỉnh:  $\mathbf{V}$ ;
11    Tính toán độ tương đồng với tất cả các thực thể mục tiêu  $\mathbf{t}$ ;
12    Áp dụng hàm sigmoid để đưa ra điểm số;
13  Cập nhật các tham số huấn luyện một cách tuần tự bằng cách cực tiểu hàm mất mát binary cross-entropy
  
```

Các bộ dữ liệu đánh giá

Bảng 1: Thông tin các bộ dữ liệu được sử dụng cho thực nghiệm

Dataset	#Ent	#Rel	#Train	#Valid	#Test	#Total
FB15k	14,951	1,345	483,142	50,000	59,071	592,213
FB15k-237	14,541	237	272,115	17,535	20,466	310,116
WN18	40,943	18	141,442	5,000	5,000	151,442
WN18RR	40,943	11	86,835	3,034	3,134	93,003

Các độ đo đánh giá

Có 3 độ đo phổ biến được dùng để đánh giá mô hình:

- Mean rank (MR)

$$MR = \frac{1}{|Q|} \sum_{q \in Q} q \quad (6)$$

- Mean Reciprocal Rank (MRR)

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{q} \quad (7)$$

- $Hits@K$

$$Hits@K = \frac{|q \in Q : q < K|}{|Q|} \quad (8)$$

trong đó: q thể hiện phần tử dự đoán; Q thể hiện tất cả các phần tử dự đoán được cho bởi thuật toán.

Cấu hình thực nghiệm

Môi trường phần mềm: Ubuntu 18.04.5 LTS, Python 3, Pytorch 1.8.1 LTS
 Phần cứng: Intel i7-10700K (16) @ 5.100GHz, GPU NVIDIA GeForce RTX 3070
 8GB VRAM

Bảng 2: Cấu hình tối ưu cho mô hình ACRM trên bốn bộ dữ liệu, trong đó d_e là kích thước nhúng thực thể, c là số bộ lọc, $h \times w$ là kích thước của bộ lọc và p_1, p_2, p_3 lần lượt là bỏ qua đầu vào, bỏ qua bản đồ đặc trưng và bỏ qua ẩn

Bộ dữ liệu	d_e	c	$h \times w$	p_1	p_2	p_3
FB15k	200	100	3×3	0.2	0.3	0.3
WN18	200	100	3×3	0.3	0.2	0.3
FB15k-237	100	100	5×5	0.3	0.2	0.3
WN18RR	200	200	3×3	0.2	0.2	0.5

Kết quả thực nghiệm

Bảng 3: Kết quả dự đoán liên kết trên FB15k và WN18

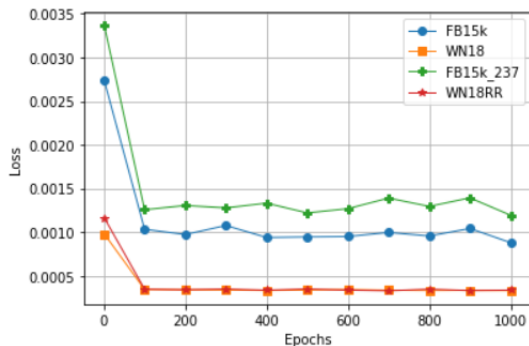
Mô hình	FB15k				WN18			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE	0.380	0.231	0.472	0.641	0.454	0.089	0.823	0.934
TransD	-	-	-	0.773	-	-	-	0.922
STransE	0.543	-	-	0.797	0.657	-	-	0.934
DistMult	0.654	0.546	0.733	0.824	0.822	0.728	0.914	0.936
ComplEx	0.692	0.599	0.759	0.840	0.941	0.936	0.936	0.947
HoIE	0.524	0.402	0.613	0.739	0.938	0.930	0.945	0.949
CrossE	0.728	0.634	0.802	-	0.830	0.741	0.931	-
ConvR	0.770	0.706	0.814	0.880	0.950	0.945	0.953	0.958
R-GCN	0.696	0.601	0.760	0.842	0.814	0.686	0.928	0.955
ConvE	0.657	0.558	0.723	0.831	0.943	0.935	0.946	0.956
ACRM	0.767	0.700	0.816	0.883	0.950	0.946	0.954	0.958

Kết quả thực nghiệm

Bảng 4: Kết quả dự đoán liên kết trên FB15k-237 và WN18RR

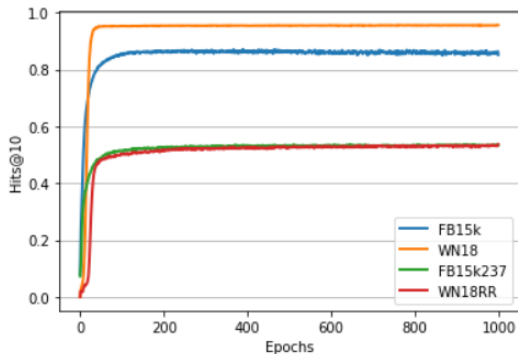
Mô hình	FB15k-237				WN18RR			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE	0.257	0.174	0.284	0.420	0.182	0.027	0.295	0.444
TransD	-	-	-	-	-	-	-	-
STransE	-	-	-	-	-	-	-	-
DistMult	0.241	0.155	0.263	0.419	0.430	0.390	0.440	0.490
Complex	0.247	0.158	0.275	0.428	0.440	0.410	0.460	0.510
HoIE	-	-	-	-	-	-	-	-
CrossE	0.299	0.211	0.331	0.474	-	-	-	-
ConvR	0.348	0.258	0.382	0.528	0.474	0.442	0.489	0.534
R-GCN	0.249	0.151	0.263	0.417	-	-	-	-
ConvE	0.325	0.237	0.356	0.501	0.430	0.400	0.440	0.520
ACRM	0.35	0.26	0.382	0.535	0.475	0.444	0.490	0.539

Kết quả thực nghiệm



Hình 8: Giá trị mất mát qua từng lần lặp huấn luyện.

Kết quả thực nghiệm



Hình 9: Giá trị Hits@10 qua từng lần lặp huấn luyện.

Kết quả thực nghiệm

Bảng 5: Ảnh hưởng của kích thước bộ lọc đối với hiệu suất của mô hình trên FB15k-237 và WN18RR

Kích thước bộ lọc	FB15k-237		WN18RR	
	MRR	Hits@10	MRR	Hits@10
2×2	0.347	0.529	0.468	0.533
3×3	0.349	0.530	0.475	0.539
4×4	0.348	0.534	0.473	0.535
5×5	0.350	0.535	0.474	0.534

Kết quả thực nghiệm

Bảng 6: Dự đoán thực thể đầu và đuôi trên các bộ dữ liệu

Bộ dữ liệu	MRR		Hits@1		Hits@3		Hits@10	
	Head	Tail	Head	Tail	Head	Tail	Head	Tail
FB15k	0.791	0.742	0.726	0.674	0.842	0.79	0.907	0.859
WN18	0.95	0.949	0.946	0.945	0.953	0.954	0.958	0.958
FB15k- 237	0.451	0.249	0.357	0.163	0.493	0.271	0.64	0.429
WN18- RR	0.499	0.451	0.468	0.42	0.515	0.465	0.556	0.522

Kết luận

- Kết quả thực nghiệm của mô hình đề xuất đạt kết quả cải thiện trên bốn bộ dữ liệu FB15k, FB15k-237, WN18, WN18RR.
- Tích chập thích nghi sử dụng các nhúng quan hệ làm bộ lọc đã cải thiện được sự tương tác giữa thực thể và quan hệ.
- Với cơ chế hiệu chỉnh, bản đồ đặc trưng giữa các kênh có thể sử dụng thông tin ngữ cảnh bên ngoài trường tiếp nhận cục bộ giúp cho mô hình nơ-ron tích chập ConvR cải thiện được hiệu suất trên các độ đo chuẩn.

Hướng nghiên cứu tương lai

- Biểu diễn đối tượng trong mô hình ACRM chỉ có thể tương tác với một véc-tơ ẩn. Điều này làm cho mô hình không thể trích xuất thông tin từ các véc-tơ ẩn khác. Trong tương lai, nhóm dự định tạo ra một cơ chế hiệu chỉnh phức tạp hơn để cải thiện nhược điểm còn tồn tại trong mô hình ACRM.
- Đồng thời cải tiến cấu trúc tích chập kết hợp với những tiếp cận khác để tạo ra những hướng đi mới đóng góp nhiều hơn cho bài toán dự đoán liên kết trên đồ thị tri thức.

Công bố khoa học

Trong quá trình nghiên cứu và thực hiện khóa luận này, các kết quả nghiên cứu đã được chấp nhận công bố như sau:

- ① **ACRM: Integrating Adaptive Convolution with Recalibration Mechanism for Link Prediction** tại hội nghị 15th International Conference on Knowledge Science engineering and Management (KSEM 2022), August 6-8, 2022, Singapore, Rank B (CORE2021)

Tài liệu tham khảo



Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel.
Convolutional 2d knowledge graph embeddings.

In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.



Xiaotian Jiang, Quan Wang, and Bin Wang.

Adaptive convolution for multi-relational learning.

In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 978–987, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Cảm ơn quý thầy, cô và các bạn đã lắng nghe!