Giải thích mạng nơ-ron đồ thị bằng đại diện có cùng phân phối (Generating In-Distribution Proxy Graphs for Explaining Graph Neural Networks)

23C11007 - Vũ Công Minh

Ngày 9 tháng 2 năm 2025

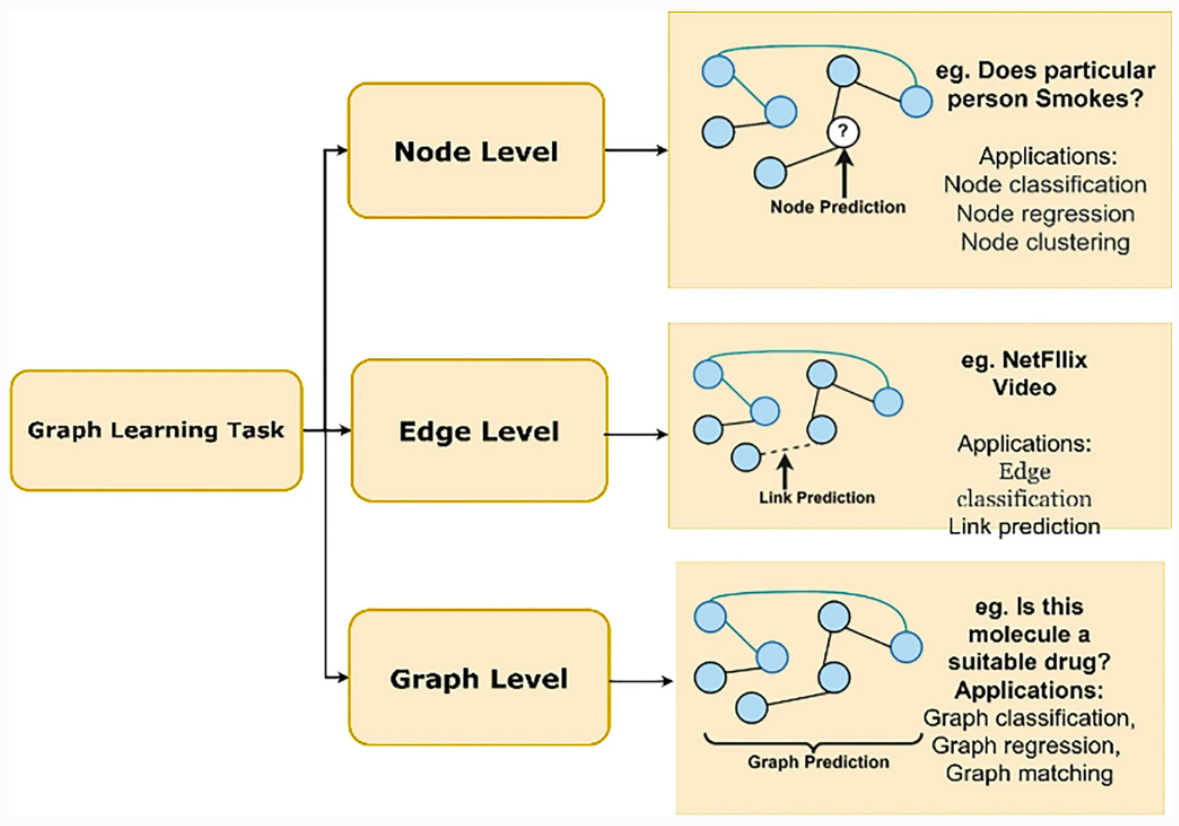
Tóm tắt nội dung

Với sự phát triển của Trí tuệ nhân tạo như hiện nay, mạng nơ-ron đồ thị (GNN) được xem là một trong những hướng nghiên cứu đầy tiềm năng bởi khả năng tích hợp thông tin, mở rộng và áp dụng trong nhiều lĩnh vực có cấu trúc phức tạp như y sinh, mạng xã hội, ... Chính vì nhu cầu triển khai GNN vào các ứng dụng quan trọng, các mô hình phức tạp như GNN cần đòi hỏi thêm về khả năng giải thích nhằm hỗ trợ người dùng hiểu và ra những quyết định chính xác hơn. Một phương pháp phổ biến để giải thích mô hình GNN là xác định các đồ thị con có thể giải thích bằng cách so sách nhãn của nó với đồ thị gốc. Tuy nhiên, công việc này bị ảnh hưởng lớn bởi sự chuyển dịch phân phối trong quá trình huấn luyện dẫn tới việc dự đoán nhãn không đạt được kết quả cao. Chính vì vậy, Zhuomin và các cộng sự đã đề xuất một phương pháp mới bằng cách tạo các đồ thị đại diện (proxy graph) từ những đồ thị con có thể giải thích, vừa đảm bảo khả năng giải thích cũng như phạm vi phân phối không quá lệch nhau trong quá trình huấn luyện. Trong phạm vi môn học, nhóm sẽ tiến hành tìm hiểu, trình bày lại những đóng góp của nhóm tác giả theo cách hiểu của mình, đồng thời diễn giải chi tiết hơn về các mô hình và phương pháp được liệt kê trong bài nghiên cứu. Ngoài ra, nhóm sẽ thử nghiệm thêm với một tập dữ liệu khác về giao thông cũng như bổ sung kết quả thực nghiệm mà tác giả không đề cập trong bài.

1 Giới thiệu

1.1 Mạng nơ-ron đồ thị GNN

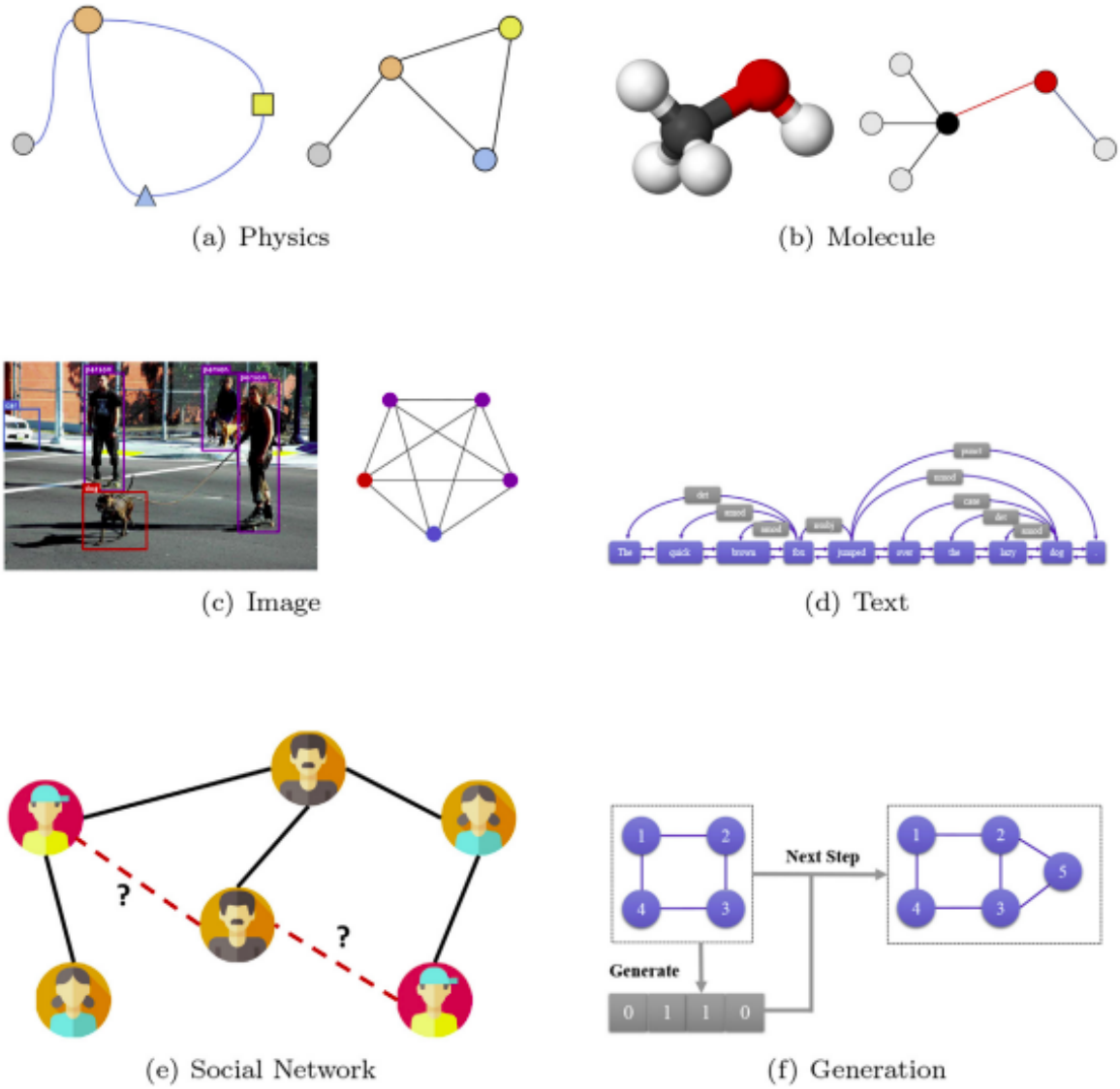
Hiện nay, mạng nơ-ron đồ thị (Graph Neural Network - GNN) ngày càng nhận được sự quan tâm bởi khả năng xử lý của nó trên các dữ liệu lớn có nhiều liên kết phức tạp như mạng xã hội, liên kết phân tử, bản đồ đường đi, ... [Zhou et al.(2018)]. Ý tưởng cốt lõi của GNNs là tận dụng cấu trúc đồ thị của dữ liệu để học các biểu diễn (representations) hiệu quả cho các đối tượng (nút) và liên kết của chúng (cạnh). Tahy vì xử lý dữ liệu một cách độc lập, GNNs xem xét mối quan hệ giữa các phần tử dữ liệu thông qua các kết nối trong đồ thị. Nguyên lý hoạt động của GNNs dựa trên việc truyền thông tin giữa các nút lân cận trong đồ thị. Mỗi nút sẽ tổng hợp thông tin từ các nút hàng xóm của nó và sử dụng thông tin này để cập nhật biểu diễn của chính nó. Quá trình này được lặp lại nhiều lần, cho phép thông tin lan truyền khắp đồ thị và các nút có thể học được các biểu diễn phức tạp, phản ánh cấu trúc và mối quan hệ của đồ thị. Những năm gần đây chứng kiến sự bùng nổ của GNNs với nhiều kiến trúc mới được đề xuất, chẳng hạn như Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), và nhiều biến thể khác. Sự phát triển này không chỉ giúp cải thiện hiệu suất của GNNs trên các bài toán khác nhau mà còn mở rộng phạm vi ứng dụng của chúng.



Hình 1: Các nhiệm vụ học trong GNN

Mạng nơ-ron đồ thị đặc biệt thích hợp với các dữ liệu có không gian phi Euclid. Khả n ăng tổng quát hóa tốt cũng là một điểm mạnh của GNNs. Mô hình được huấn luyện trên một tập đồ thị có thể được áp dụng cho các đồ thị mới với cấu trúc tương tự. Điều này cho thấy GNNs có khả năng thích ứng linh hoạt với dữ liệu mới, giúp tiết kiệm thời gian và công sức huấn luyện lại mô hình từ đầu. Bên cạnh đó, GNNs cũng tồn tại một số hạn chế về tài nguyên tính toán và thời gian vì trong thực tế các mạng lưới thông tin thật sự rất thưa và phức tạp. Mặc dù còn một số thách thức, GNNs đã chứng minh được tính hiệu quả của mình trong một loạt các ứng dụng thực tế. [Zhou et al.(2021)] Trong lĩnh vực mạng xã hội, GNNs được sử dụng để phân tích cấu trúc mạng, dự đoán hành vi người dùng, gợi ý bạn bè và phát hiện tin giả. Trong hóa học và sinh học, GNNs giúp dự đoán tính chất phân tử, tìm kiếm thuốc mới, phân tích tương tác protein và hiểu các quá trình sinh học phức tạp. Trong xử

lý ngôn ngữ tự nhiên, GNNs được áp dụng để phân tích cú pháp, trích xuất thông tin, dịch máy và hiểu ngữ nghĩa của câu. Trong lĩnh vực giao thông, GNNs giúp dự đoán lưu lượng giao thông, tối ưu hóa lộ trình và quản lý mạng lưới giao thông thông minh. Ngoài ra, GNNs còn được ứng dụng trong thương mại điện tử để gợi ý sản phẩm cho người dùng, phân tích hành vi mua hàng và phát hiện gian lận. Sự đa dạng trong ứng dụng thực tế cho thấy tiềm năng to lớn của GNNs trong việc giải quyết các bài toán phức tạp trong nhiều lĩnh vực khác nhau.

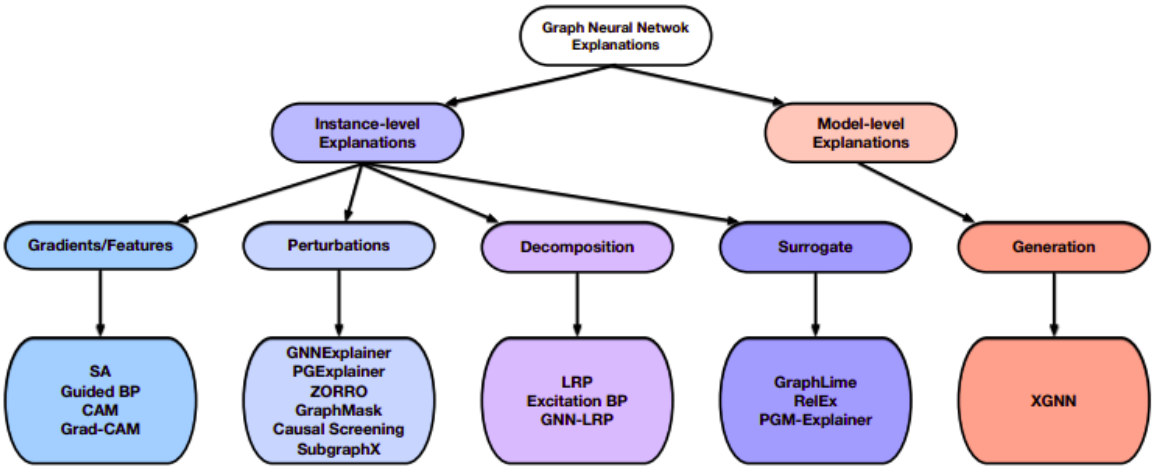


Hình 2: Những ứng dụng của GNN

1.2 Khả năng giải thích của GNN

Việc giải thích các mô hình học sâu nói chung và GNNs nói riêng có vai trò vô cùng quan trọng, đặc biệt trong bối cảnh chúng ngày càng được ứng dụng rộng rãi trong nhiều lĩnh vực như hiện nay. Những mô hình ấy được coi là "hộp đen"(blackbox) bởi tính phức tạp trong cách chúng tính toán và đưa ra quyết định. Việc diễn giải giúp làm sáng tỏ quá trình này, cho phép con người hiểu được "lý do"đằng sau các dự đoán. Điều này đặc biệt quan trọng trong các ứng dụng nhạy cảm như y tế, tài chính, pháp lý, nơi mà độ tin cậy và tính minh bạch là yếu tố then chốt. Không những thế, việc phân tích các lời giải thích, chúng ta có thể phát hiện ra các lỗi trong mô hình hoặc trong dữ liệu huấn luyện,

từ đó giúp cải thiện hiệu suất mô hình hơn. Theo [Yuan et al.(2020)], khả năng diễn giải của mô hình nơ-ron đồ thị được chia thành các thành phần sau:



Hình 3: Các khía cạnh của việc cách diễn giải của GNN

- Mức độ giải thích:
  - Instance-level Explanations: Giải thích lý do đưa ra dự đoán cho một trường hợp cụ thể (ví dụ: tại sao một nút được phân loại như vậy).
  - Model-level Explanations: Hiểu tổng quan về cách thức hoạt động của toàn bộ mô hình (ví dụ: đặc trưng nào quan trọng nhất cho việc dự đoán).
- Phương pháp tiếp cận:
  - Gradients/Features: Dựa trên gradient của mô hình để xác định các thành phần đặc trưng.
  - Perturbations: Tạo ra các phiên bản nhiễu loạn của đồ thị đầu vào.
  - Decomposition: Phân tích mức độ đóng góp của từng phần tử vào dự đoán.
  - Surrogate: Sử dụng mô hình đơn giản hơn để xấp xỉ và giải thích.
  - Generation: Tạo ra một đồ thị mới để giải thích.

1.3 Đóng góp của bài báo

Bài báo này tập trung vào phân tích và giải quyết một thách thức quan trọng: vấn đề "ngoài phân bố"(OOD - Out-of-Distribution) trong các mô hình GNN có khả năng giải thích. Vấn đề này có ý nghĩa then chốt để nâng cao độ tin cậy và khả năng diễn giải của GNN trong các ứng dụng thực tế. Những đóng góp chính của chúng tôi bao gồm:

- Phân tích và giải quyết vấn đề OOD: Chúng tôi đã tiến hành phân tích một cách có hệ thống và đề xuất giải pháp cho vấn đề dữ liệu nằm ngoài phân bố, một thách thức lớn đối với việc tạo ra lời giải thích đáng tin cậy cho GNN. Việc giải quyết vấn đề này giúp tăng cường tính ứng dụng thực tế của GNN.
- Phương pháp tham số mới: Chúng tôi giới thiệu một phương pháp tham số cải tiến, kết hợp các bộ tự mã hóa đồ thị để tạo ra các "đồ thị đại diện phân bố trong". Các đồ thị này không chỉ nằm trong phân bố dữ liệu gốc mà còn bảo toàn thông tin giải thích thiết yếu. Điều này giúp tạo ra các lời giải thích chính xác và dễ hiểu hơn cho các ứng dụng GNN.
- Thực nghiệm tàn diện: Thông qua các thực nghiệm toàn diện trên nhiều tập dữ liệu thực tế, chúng tôi chứng minh tính hiệu quả của phương pháp được đề xuất. Kết quả cho thấy phương pháp của chúng tôi có tính ứng dụng cao và vượt trội trong việc tạo ra các lời giải thích chất lượng.

2 Cơ sở lý thuyết

2.1 Nút thắt thông tin trong đồ thị

2.2 Đồ thị đại diện

2.3 Bài toán tối ưu hoá 2 cấp

3 Giải thích mô hình với Đồ thị đại diện

3.1 Nút thắt thông tin trong đồ thị

3.2 Đồ thị đại diện

3.3 Bài toán tối ưu hoá 2 cấp

4 Thực nghiệm

4.1 Chuẩn bị dữ liệu và các mô hình

4.2 Đánh giá định lượng (RQ1)

4.3 Đánh giá độ chuyển dịch phân phối (RQ2)

4.4 Đánh giá tầm ảnh hưởng của các thành phần (RQ3)

4.5 Thực nghiệm bổ sung

5 Kết luận

Nhìn chung, đây là một nghiên cứu khá thú vị về vấn đề ít được chú trọng, đặc biệt là trong việc huấn luyện các mô hình mạng nơ-ron đồ thị có khả năng giải thích: hiện tượng OOD(Out-Of-Distribution). Hiện tượng này xảy ra khi mô hình được huấn luêyjn trên một tập dữ liệu nhất định, nhưng sau đó được áp dụng trên một dữ liệu mới có phân bố quá khác biệt. Việc này dẫn đến kết quả không chính xác và lời giải thích không đáng tin cậy. Và để giải quyết vấn đề này, nhóm tác giả đã cải tiến GIB (Graph Information Bottleneck - một nguyên lý thông tin trong diễn giải đồ thị) bằng cách tạo ra các Đồ thị đại diện nằm trong cùng phân phối với đồ thị huấn luyện (In-Distributed Proxy Graph). Cách tiếp cận này vừa giúp mô hình học được cách biểu diễn đồ thị chính xác hơn, giúp việc phân loại cạnh và khả năng giải thích của nó tốt hơn so với các mô hình khác được đề cập trên cả về tập dữ liệu thực tế lẫn tập dữ liệu nhân tạo. Điều này giúp mở rộng hơn các hướng phát triển mới nhằm giải quyết vấn đề OOD không chỉ ở cấp độ các nút mà còn các cấp độ cao hơn như đồ thị con,... hoặc các cấu trúc dữ liệu khác như hình ảnh, âm thanh, chuỗi thời gian,...

Tài liệu

[Yuan et al.(2020)] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2020. Explainability in Graph Neural Networks: A Taxonomic Survey. *CoRR* abs/2012.15445 (2020). arXiv:2012.15445 https://arxiv.org/abs/2012.15445

[Zhou et al.(2021)] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2021. Graph Neural Networks: A Review of Methods and Applications. arXiv:1812.08434 [cs.LG] https://arxiv.org/abs/1812.08434

[Zhou et al.(2018)] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Graph Neural Networks: A Review of Methods and Applications. *CoRR* abs/1812.08434 (2018). arXiv:1812.08434 http://arxiv.org/abs/1812.08434