**Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI**

**Giới thiệu**

* Các phương pháp hoạt động của AI trong lĩnh vực y tế tốt nhất hiện nay đều thiếu khả năng diễn giải. Tính không rõ ràng trong mô hình này được gọi là black-box.
* Trong nhiều trường hợp, điều bắt buộc là các quyết định của AI phải truy nguyên được nguồn gốc, các phương thức liên quan giúp các chuyên gia trong lĩnh vực y tế có thể hiểu, xác nhận hoặc bác bỏ chúng. Vì vậy, việc sử dụng đồ thị hoặc biểu diễn đồ thị có thể có lợi trong việc diễn giải.
* Bài báo nêu ra 3 thách thức trong việc thiết lập cấu trúc học máy đa phương thức và có thể giải thích được trong lĩnh vực y tế. Mỗi giai đoạn trong quy trình có thể thực hiện riêng biệt, nhưng việc tổng hợp lại các cách tiếp cận tối ưu hóa được lợi ích.

**Thử thách 1: Xây dựng Multi-modal Embedding Space**

* Lý do: Học trên dữ liệu hợp nhất từ các nguồn khác nhau hoặc các phương thức khác nhau có thể đem lại kết quả tốt hơn việc học trên 1 cấu trúc dữ liệu.
* Thử thách ở đây là thu hẹp khoảng cách về mặt ngữ nghĩa trong kết hợp dữ liệu nhưng các phương pháp hiện nay là tổng hợp lại các đặc trưng không chú trọng khoảng cách hay ngữ nghĩa của chúng.
* Để nắm được những đặc trưng như tính quan hệ, thứ bậc, loại suy, các hướng giải quyết có khả năng là Joint embedding và Graph Representation Learning.

**Thách thức 2: Distributed Graph Representation Learning**

* Lý do: hạn chế bộ nhớ chứa dữ liệu hoặc hiệu suất xử lý, các vấn đề pháp lý trong việc trao đổi dữ liệu, nhất là trong trong lĩnh vực y tế, nơi dữ liệu nhạy cảm, bảo mật và bộ dữ liệu rất lớn.
* Thử thách chính ở phần này là giữ không gian biểu diễn trên các ranh giới vị trí sao cho giống với sự căn chỉnh đặc trưng trong cài đặt MM.
* Distributed Learning không đảm bảo độ trễ giữa các kết nối → Thách thức trong duy trì hiệu suất và độ chính xác, không đảm bảo nắm bắt được toàn bộ thông tin.
* Quy trình được đề xuất: Cải thiện kỹ thuật *neighbour aggregation*, lựa chọn và thử nghiệm(mean, max, avg), hay sử dụng LSTM cho việc tự động hoá lựa chọn, tổng hợp thông tin - được sử dụng thành công trong GraphSage.
* Tối ưu hoá các kĩ thuật tổng hợp góp phần tối đa hoá hiệu suất và độ chính xác, do đó tránh được các quy tắc logic cứng nhắc áp đặt trong phần tổng hợp và lựa chọn đặc trưng, trong khi có thể thu được các hướng tiếp cận khác.

**Thử thách 3: Explainable AI / Interpretable ML**

* Ba phương pháp có thể giải thích nổi bật của Convolutional NN bao gồm: Contrastive Gradient-based saliency map (CG), Class Activation Mapping (CAM), và các biến thể của chúng như là: Gradient weighted (Grad-CAM) và constrastive EB (c-EB) được mở rộng cho các GCNN.
* Mô hình GCN có thể được xem là mô hình hộp đen, thiếu tính minh bạch. Vì vậy số lượng nghiên cứu giúp diễn giải mô hình ngày càng tăng. Trong đó, nghiên cứu được đề xuất của bài báo này là GNN-LRP, được hình thành từ 2 phương pháp là back propagation technique với max-pooling sử dụng ReLu và phân rã Taylor. Ý tưởng tổng thể là xác định đầu vào liên kết với một activation nào đó.
* Phương pháp GCN-LRP giải thích dự đoán của mô hình bằng cách đưa ra các cạnh liên kết của các node đầu ra từ các node đầu vào.
  + Để đi sâu vào GCN-LRP thì chúng em implement mô hình theo paper dưới để củng cố sự khả thi của paper trên

**Layerwise Relevance Visualization in Convolutional Text Graph Classifiers**

**Giới thiệu**

* Trình bày phương pháp giúp diễn giải các đặc trưng mà góp phần trong việc đưa ra quyết định của mô hình trong các lớp hidden hay lớp visible trong GCN.
* Layerwise Relevance Visualization (LRV) trực quan hóa sự giải thích trong lớp hidden.

**Phương pháp**

* **Graph Convolutional Network**
* The Graph G (A, H), với ma trận A là ma trận kề chuẩn hóa và H là Embedding matrix.
* **Layer Wise Relevance**
* LRP giải thích sự kích hoạt của noron bằng cách tính toán xem mỗi noron của nó đã đóng góp bao nhiêu vào quá trình kích hoạt. Các đóng góp được truyền ngược theo từng lớp.
* Layer-wise Relevance Propagation (LRP) hoạt động bằng cách truyền ngược dự đoán trong mạng noron.
* LRP không chỉ bảo đảm sự liên quan giữa các noron với nhau mà còn bảo đảm sự liên quan giữa các lớp với nhau.

**Layerwise Relevance Visualization**

* Nếu ma trận kề là đầu vào, sẽ có nhận định rằng nó chỉ đang nhận đươc relevance mass→ Vì vậy, tác giả đề xuất xây dựng 1 lớp FC biểu diễn trọng số của giá trị trong ma trận kề A. Lớp GCN gồm 2 lớp con FC, trong đó lớp đầu tiên thực hiện trên các đặc trưng và lớp thứ hai thực hiện trên các giá trị cạnh kề.
* Sử dụng hàm activation ReLU để tránh giá trị âm.
* Sử dụng giá trị ngay khi truyền ngược giữa các layer (R) và giá trị tương quan giữa các cạnh (e) để diễn giải mô hình.

**Thí nghiệm**

* GCN-LRP sentence classifier
* Bộ dữ liệu: PubMed 200k RCT, chứa 5 nhãn Background, Objective, Method, Result và Conclusion.
* Tiền xử lý:
* Sử dụng ScispaCy để tokenization tạo đầu vào cho mô hình.
* Node Embedding sử dụng pretrained model FastText.
* Thông tin loại cạnh bị xóa.
* Kiến trúc:
* FC -> GCN -> GCN -> FC (classification)
* Sử dụng Adam optimizer, no bias.
* Clamp để tránh giá trị âm
* LRP:
* Bắt đầu từ lớp FC output với ReLu, trước khi chuẩn hóa Softmax.
* Lưu trữ các giá trị R và e đã đề cập ở phần Layerwise Relevance Visualization
* Phân tích định tính:
* Trực quan hóa sự đóng góp của từng lớp theo các cạnh liên quan
* Xác nhận định lượng:
* Xóa các phần tăng trưởng của cạnh toàn cục liên quan nhất.
* Làm điều tương tự với các cạnh ít liên quan nhất
  + Để quan sát sự đóng góp của các cạnh ảnh hưởng đến hiệu suất của mô hình.

**Kết quả của thực nghiệm**

* Metric weighted F1 score = 0.822
* Hiệu suất của mô hình giảm nhanh hơn khi loại bỏ các cạnh đóng góp nhiều trong đưa ra quyết định.