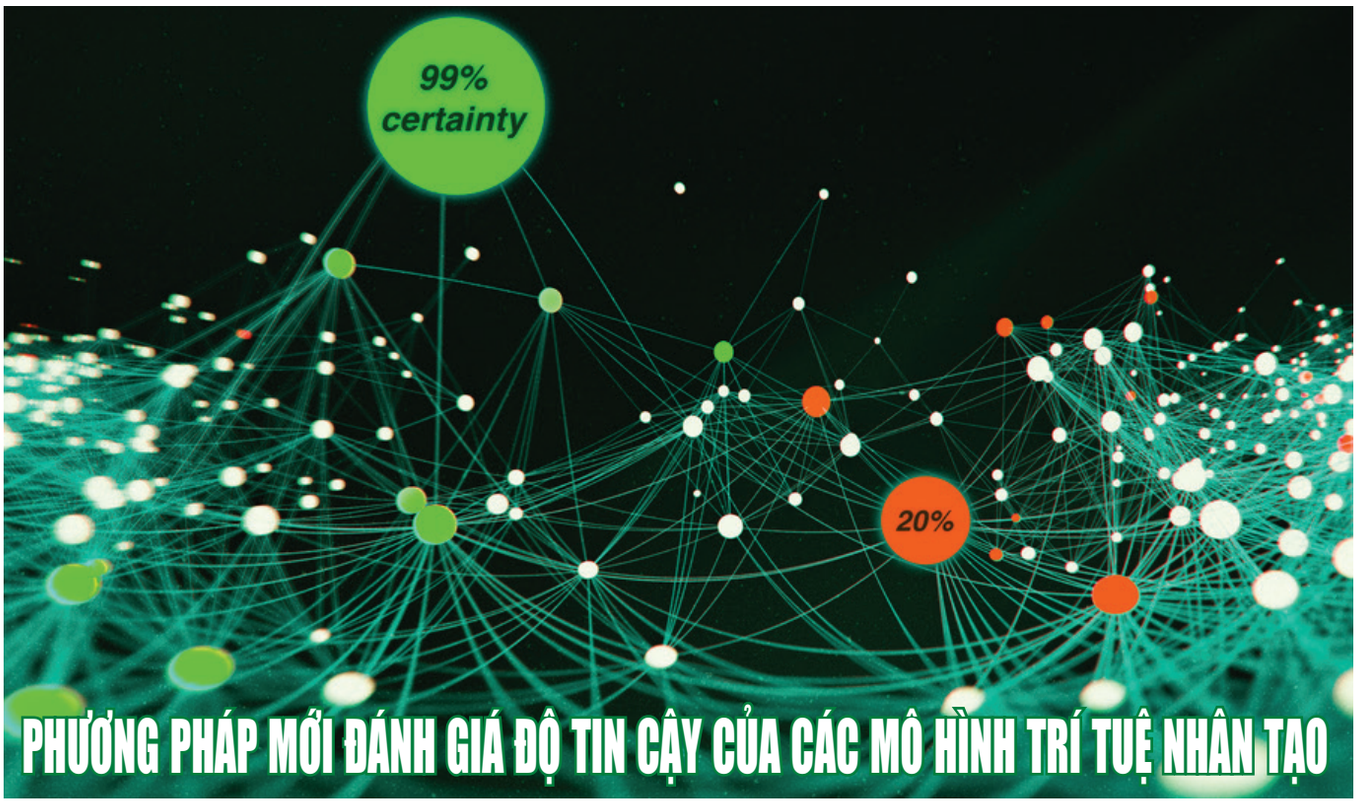

PHƯƠNG PHÁP MỚI ĐÁNH GIÁ ĐỘ TIN CẬY CỦA CÁC MÔ HÌNH TRÍ TUỆ NHÂN TẠO

ThS Đặng Thế Hân

Khoa Công nghệ Thông tin, Trường Đại học Nguyễn Tất Thành

Trí tuệ nhân tạo (AI) ngày càng phổ biến trong nhiều lĩnh vực, ảnh hưởng đến cuộc sống hàng triệu người, nhưng các mô hình AI, đặc biệt là mô hình AI nền tảng có thể cung cấp thông tin sai lệch gây hậu quả nghiêm trọng. Để giải quyết vấn đề này, các nhà nghiên cứu thuộc Viện Công nghệ Massachusetts (MIT), Mỹ đã phát triển hai phương pháp mới nhằm đánh giá độ tin cậy của các mô hình AI. Phương pháp đầu tiên đánh giá độ tin cậy của mô hình AI nền tảng bằng cách sử dụng tập hợp các mô hình khác nhau để kiểm tra tính nhất quán của kết quả, giúp quyết định xem mô hình có phù hợp để áp dụng trong bối cảnh cụ thể hay không. Phương pháp thứ hai là IF-COMP, dựa trên nguyên tắc độ dài mô tả tối thiểu (MDL) để định lượng độ không chắc chắn trong dự đoán của mô hình máy học, đem lại hiệu quả và độ chính xác cao hơn so với các phương pháp hiện có, giúp người dùng xác định mức độ tin cậy của dự đoán. Hai phương pháp này không chỉ cải thiện độ tin cậy và an toàn của các hệ thống AI mà còn có thể áp dụng cho nhiều loại mô hình khác nhau, mở ra tiềm năng phát triển các hệ thống AI đáng tin cậy, an toàn và hiệu quả hơn trong tương lai.

Đánh giá độ tin cậy của mô hình trí tuệ nhân tạo nền tảng, tạo sinh

Làn sóng AI hiện nay là các mô hình được đào tạo trên một tập hợp dữ liệu rộng lớn không có nhãn có thể được sử dụng cho các nhiệm vụ khác nhau, với sự tinh chỉnh tối thiểu. Đây được gọi là mô hình AI nền tảng (Foundation model). Được đào tạo trên các tập dữ liệu lớn, các mô hình AI nền tảng là các mạng nơ-ron học sâu quy mô lớn đã thay đổi cách các nhà khoa học dữ liệu tiếp cận máy học. Thay vì phát triển AI từ đầu, các nhà khoa học dữ liệu sử dụng mô

hình nền tảng làm điểm khởi đầu để phát triển các mô hình máy học giúp các ứng dụng mới nhanh hơn và tiết kiệm chi phí hơn. Những mô hình AI nền tảng được đào tạo trên một phổ dữ liệu tổng quát, không có gắn nhãn và có khả năng thực hiện nhiều tác vụ chung khác nhau như hiểu ngôn ngữ, tạo văn bản hay hình ảnh và trò chuyện bằng ngôn ngữ tự nhiên. Hiện nay, những mô hình AI tạo sinh này đóng vai trò là xương sống cho các công cụ AI mạnh mẽ như ChatGPT, DALL-E hay BERT. Tuy nhiên, chúng có thể cung cấp thông tin không chính xác hoặc gây hiểu lầm cho người dùng. Trong những tình huống quan trọng về an toàn, chẳng hạn như trong giao thông hay công nghệ xe tự lái, những chỉ dẫn sai lầm này có thể gây ra hậu quả nghiêm trọng.

Để giúp ngăn ngừa những vấn đề như vậy, các nhà nghiên cứu từ MIT và Phòng thí nghiệm AI Watson của MIT-IBM đã phát triển một phương pháp để đánh giá độ tin cậy của các mô hình AI nền tảng trước khi chúng được triển khai cho một nhiệm vụ cụ thể. Kỹ thuật này có thể giúp quyết định xem có nên áp dụng một mô hình AI trong một bối cảnh nhất định hay không, mà không cần phải thử nghiệm nó trên một tập dữ liệu thực tế. Điều này đặc biệt hữu ích khi các tập dữ liệu có thể không truy cập được do lo ngại về quyền riêng tư, chẳng hạn như trong các hoạt động chăm sóc sức khỏe. Ngoài ra, kỹ thuật này có thể được sử dụng để xếp hạng các mô hình dựa trên điểm tin cậy, cho phép người dùng chọn mô hình AI tốt nhất cho nhiệm vụ của họ. Tất cả các mô hình AI đều có thể sai, nhưng sẽ rất hữu ích nếu biết khi nào các mô hình sẽ sai. Vấn đề định lượng sự không chắc chắn hoặc độ tin cậy trở nên khó khăn hơn đối với các mô hình AI nền tảng vì khả năng thể hiện các kết quả có tính trừu tượng của chúng là khó so sánh.

Các mô hình máy học truyền thống được đào tạo để thực hiện một nhiệm vụ cụ thể. Các mô hình này thường đưa ra dự đoán cụ thể dựa trên đầu vào biết trước. Tuy nhiên, đối với những mô hình AI nền tảng thì khác, chúng thường được đào tạo trước bằng dữ liệu nguồn chung, trong bối cảnh mà người tạo ra không biết tất cả các tác vụ hạ nguồn mà mô hình sẽ được áp dụng. Người dùng sẽ điều chỉnh mô hình cho các tác vụ cụ thể của họ sau khi mô hình đã được đào tạo. Không giống như các mô hình máy học truyền thống, các mô hình AI nền tảng không đưa ra đầu ra có tính cố định, thay vào đó, chúng tạo ra kết quả dựa trên một điểm dữ liệu đầu vào và có tính tạo sinh.

Để đánh giá độ tin cậy của một mô hình AI nền tảng, các nhà nghiên cứu đã sử dụng phương pháp tiếp cận tổng hợp bằng cách đào tạo một số mô hình có chung nhiều thuộc tính nhưng hơi khác nhau để đo lường sự đồng thuận. Nếu tất cả các mô hình AI nền tảng đó đưa ra các biểu diễn kết quả nhất quán cho bất kỳ dữ liệu nào trong tập dữ liệu thử nghiệm, thì có thể kết luận rằng mô hình AI này đáng tin cậy.

Để so sánh và kiểm tra độ tin cậy của các kết quả có tính trừu tượng, nhóm nghiên cứu sử dụng phương pháp gọi là tính nhất quán của vùng lân cận. Họ chuẩn bị một tập hợp các điểm tham chiếu đáng tin cậy để thử nghiệm trên tập hợp các mô hình AI. Sau đó, đối với mỗi mô hình, họ điều tra các điểm tham chiếu nằm gần của mô hình đó để kiểm tra. Bằng cách xem xét tính nhất quán của các điểm lân cận, họ có thể ước tính độ tin cậy của các mô hình AI.

Đánh giá độ tin cậy của mô hình máy học

Các nhà nghiên cứu của MIT cũng đã giới thiệu một phương pháp mới có thể cải thiện và đo lường mức độ không chắc chắn trong việc dự đoán của các mô hình máy học một cách chính xác và hiệu quả hơn các phương pháp đã có. Phương pháp mới này có khả năng mở rộng quy mô nên có thể áp dụng cho các mô hình học sâu quy mô lớn đang ngày càng được triển khai rộng trong chăm sóc sức khỏe và các tình huống quan trọng khác về an toàn. Phương pháp này có thể cung cấp cho người dùng cuối (phần nhiều trong số họ không có chuyên môn về máy học) thông tin tốt hơn, giúp họ xác định xem có nên tin tưởng vào dự đoán của mô hình hay không hoặc liệu mô hình có nên được triển khai cho một nhiệm vụ cụ thể hay không.

Các phương pháp định lượng sự không chắc chắn thường yêu cầu các phép tính thống kê phức tạp không phù hợp với các mô hình máy học có hàng triệu tham số. Các phương pháp này cũng yêu cầu người dùng đưa ra các giả định về mô hình và dữ liệu được sử dụng để đào tạo mô hình. Các nhà nghiên cứu của MIT đã áp dụng một cách tiếp cận khác. Họ sử dụng nguyên tắc được gọi là độ dài mô tả tối thiểu (MDL - minimum description length principle), nguyên tắc này không yêu

cầu có các giả định (điều có thể cản trở độ chính xác mà các phương pháp khác thường dùng). MDL được sử dụng để định lượng tốt hơn và hiệu chỉnh sự không chắc chắn cho các điểm kiểm tra mà mô hình được yêu cầu dán nhãn. Kỹ thuật này được gọi là IF-COMP, giúp MDL đủ nhanh để sử dụng cho các loại mô hình học sâu quy mô lớn được triển khai trong nhiều bối cảnh thực tế.

MDL liên quan đến việc xem xét tất cả các nhãn mà một mô hình có thể đưa ra cho một điểm kiểm tra. Nếu có nhiều nhãn thay thế phù hợp với điểm này, độ tin cậy của nó vào nhãn mà nó đã chọn sẽ giảm theo. Ví dụ, một mô hình máy học dự đoán hình ảnh y tế cho thấy có hiện tượng tràn dịch màng phổi. Nếu các nhà nghiên cứu cho mô hình biết là hình ảnh này có hiện tượng phù nề thì mô hình máy học sẽ cập nhật thêm tình huống (thêm nhãn), từ đó thì mô hình này sẽ ít tin tưởng hơn vào dự đoán ban đầu của mình.

Trong máy học, gắn nhãn dữ liệu là quá trình gán nhãn cho dữ liệu thô để giúp cung cấp bối cảnh cho máy học và học sâu. Việc gắn nhãn dữ liệu là quá trình xác định dữ liệu thô như hình ảnh, tệp văn bản, video… và thêm một hoặc nhiều nhãn có ý nghĩa và thông tin để cung cấp ngữ cảnh và phân loại đầu vào cho mô hình máy học để nó có thể học từ đó. Các nhãn này đóng vai trò là hướng dẫn thiết yếu cho các mô hình máy học, cho phép chúng diễn giải dữ liệu và dự đoán một cách hiệu quả.

Với MDL, nếu một mô hình tự tin khi gắn nhãn một điểm dữ liệu, thì mô hình sẽ sử dụng một mã rất ngắn để mô tả điểm đó. Nếu nó không chắc chắn về quyết định của mình vì điểm đó có thể có nhiều nhãn khác, thì mô hình sẽ sử dụng một mã dài hơn để suy đoán những khả năng khác có thể xảy ra. Lượng mã được sử dụng để gắn nhãn một điểm dữ liệu được gọi là độ phức tạp của dữ liệu ngẫu nhiên và việc kiểm tra từng điểm dữ liệu bằng MDL sẽ đòi hỏi một lượng tính toán khổng lồ.

Với IF-COMP, các nhà nghiên cứu đã phát triển một kỹ thuật xấp xỉ có thể ước tính chính xác độ phức tạp của dữ liệu ngẫu nhiên bằng một hàm đặc biệt, được gọi là hàm ảnh hưởng. Họ cũng sử dụng một kỹ thuật thống kê gọi là mức nhiệt độ, giúp cải thiện việc hiệu chuẩn đầu ra của mô hình. Sự kết hợp giữa các hàm ảnh hưởng và mức nhiệt độ này cho phép tạo ra sự xấp xỉ gần đúng có độ tin cậy cao của độ phức tạp về dữ liệu ngẫu nhiên. IF-COMP có thể tạo ra sự định lượng độ không chắc chắn được hiệu chuẩn tốt một cách hiệu quả, giúp phản ánh độ tin cậy thực sự của mô hình. Kỹ thuật này cũng có thể xác định xem mô hình có dán nhãn sai một số điểm dữ liệu nhất định hay không hoặc tiết lộ điểm dữ liệu nào là điểm ngoại lệ.

Các công cụ kiểm toán đang trở nên cần thiết hơn trong các vấn đề máy học khi chúng ta sử dụng một lượng lớn dữ liệu chưa được kiểm tra để tạo ra các mô hình sẽ được áp dụng cho các vấn đề liên quan đến con người. IF-COMP không phụ thuộc vào mô hình, do đó, nó có thể cung cấp các định lượng không chắc chắn chính xác cho nhiều loại mô hình máy học. Điều này có thể cho phép triển khai nó trong nhiều bối cảnh thực tế hơn, cuối cùng giúp nhiều mô hình AI thực hiện quyết định tốt hơn.

Trong tương lai, các nhà nghiên cứu quan tâm đến việc áp dụng cách tiếp cận của họ vào các mô hình ngôn ngữ lớn và nghiên cứu các trường hợp sử dụng tiềm năng khác cho MDL ✍

TÀI LIỆU THAM KHẢO

1. A. Zewe (2024), “When to trust an AI model”, *MIT News*, https://news.mit.edu/2024/when-to-trust-ai-model-0711, truy cập ngày 15/07/2024.

2. A. Zewe (2024), “How to assess a general-purpose AI model’s reliability before it’s deployed”, *MIT News*, https://news.mit.edu/2024/how-assess-general-purpose-ai-models-reliability-its-deployed, truy cập ngày 17/07/2024.

3. R. Meritt (2023), “What are foundation models?”, *Nvidia*, https://blogs.nvidia.com/blog/what-are-foundation-models/, truy cập ngày 25/05/2024.

4. D. Ackerman (2020), “A neural network learns when it should not be trusted”, *MIT News*, https://news.mit.edu/2020/neural-network-uncertainty-1120, truy cập ngày 05/05/2024.