TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT

VIỆN ĐÀO TẠO CÔNG NGHỆ THÔNG TIN-CHUYỂN ĐỔI SỐ

BÁO CÁO TỔNG KẾT

HỌC SÂU TRONG PHÂN TÍCH DỮ LIỆU

GVHD: Ths. Hồ Ngọc Trung Kiên

SVTH:

Nguyễn Hữu Nghĩa	**2124802050013**
Phạm Tuấn Vũ	**2124802050008**
Trần Trung Nguyên	**2124802050007**

Bình Dương, ngày 26/1/2025

MỤC LỤC

DANH MỤC HÌNH

CHƯƠNG 1 TỔNG QUAN VỀ HỌC SÂU

1.1 Giới thiệu về học sâu

1.1.1 Lịch sử ra đời

Trong những năm gần đây, mạng nơ-ron nhân tạo sâu (bao gồm cả mạng nơ-ron hồi quy) đã giành chiến thắng trong nhiều cuộc thi về nhận dạng mẫu và học máy. Khảo sát lịch sử này tóm tắt một cách cô đọng các công trình có liên quan, phần lớn là từ thiên niên kỷ trước. Người học nông và người học sâu được phân biệt bởi độ sâu của các đường dẫn phân bổ tín chỉ của họ, là các chuỗi liên kết nhân quả có thể học được giữa các hành động và hiệu ứng. Tôi xem xét học có giám sát sâu (cũng tóm tắt lại lịch sử của truyền ngược), học không giám sát, học tăng cường & tính toán tiến hóa và tìm kiếm gián tiếp các chương trình ngắn mã hóa các mạng sâu và lớn.

Học sâu là một trong những xu hướng mới nhất trong nghiên cứu Học máy và Trí tuệ nhân tạo. Đây cũng là một trong những xu hướng nghiên cứu khoa học phổ biến nhất hiện nay. Các phương pháp học sâu đã mang lại những tiến bộ mang tính cách mạng trong thị giác máy tính và học máy. Thỉnh thoảng, các kỹ thuật học sâu mới và mới lại ra đời, vượt trội hơn học máy hiện đại và thậm chí cả các kỹ thuật học sâu hiện có. Trong những năm gần đây, thế giới đã chứng kiến nhiều đột phá lớn trong lĩnh vực này. Vì học sâu đang phát triển với tốc độ rất lớn nên việc theo dõi những tiến bộ thường xuyên, đặc biệt là đối với các nhà nghiên cứu mới, khá khó khăn.

Thuật ngữ "Deep Learning" (DL) lần đầu tiên được giới thiệu cho Machine Learning (ML) vào năm 1986 và sau đó được sử dụng cho Mạng nơ-ron nhân tạo (ANN) vào năm 2000 (Schmidhuber, 2015). Các phương pháp học sâu bao gồm nhiều lớp để tìm hiểu các tính năng của dữ liệu với nhiều cấp độ trừu tượng (LeCun et al., 2015). Phương pháp tiếp cận DL cho phép máy tính học các khái niệm phù hợp bằng cách xây dựng chúng từ những khái niệm đơn giản hơn (Goodfellow et al., 2016). Đối với Mạng nơ-ron nhân tạo (ANN), Deep Learning (DL) hay còn gọi là học phân cấp (Deng và Yu, 2014) là về việc gán tín chỉ trong nhiều giai đoạn tính toán một cách chính xác, để chuyển đổi kích hoạt tổng hợp của mạng (Schmidhuber, 2014). Để tìm hiểu các hàm phức tạp, kiến trúc sâu được sử dụng với nhiều cấp độ trừu tượng, tức là các phép toán phi tuyến tính; ví dụ: ANN có nhiều lớp ẩn (Bengio, 2009). Tóm lại, Deep Learning là một lĩnh vực phụ của Machine Learning, sử dụng nhiều cấp độ tiếp tục và trừu tượng hóa thông tin phi tuyến tính, để học và biểu diễn

các tính năng có giám sát hoặc không giám sát, phân loại và nhận dạng mẫu (Deng và Yu, 2014)

Học sâu là một lĩnh vực nghiên cứu phát triển mạnh với nhiều ứng dụng thành công trong các lĩnh vực khác nhau. Bài viết được viết nhằm cung cấp một đánh giá hiện đại về học sâu. Ở một mức độ nào đó, chúng tôi sẽ trình bày tổng quan lịch sử cần thiết để hiểu các khái niệm đặt nền móng cho Deep Learning ngày nay. Chúng tôi sẽ đề cập đến các phương pháp khác nhau giúp có thể đào tạo thành công các mô hình học sâu ở quy mô rất cao trong nhiều phương pháp thực hành hiện đại khác nhau.

Chúng ta đang sống trong thời đại mà học sâu đang có những bước đột phá lớn và thành công trong nhiều lĩnh vực, dù là trong lĩnh vực thị giác máy tính và nhiều ứng dụng của nó cho mục đích y tế, tự động hóa công nghiệp, hỗ trợ lái xe và xử lý ngôn ngữ tự nhiên.

Tuy nhiên, deep learning không phải là hiện tượng mới; nó chỉ có những cái tên khác nhau trong suốt nhiều năm. Chúng ta biết nó đầu tiên là điều khiển học từ năm 1940 đến năm 1960, với ý tưởng về một perceptron tái tạo bộ não người, thay vì đuôi gai lấy đầu vào và sợi trục, chúng ta có một nút lấy đầu vào và tính toán một hàm tuyến tính trong một cơ thể tế bào tương tự như khớp thần kinh Hình 1. Sau đó, đề cử của nó chuyển sang chủ nghĩa kết nối vào những năm 1980 và 1990 với việc đào tạo lan truyền ngược bị giới hạn cho một vài lớp mạng nơ-ron. Mạng nơ-ron vẫn rất khó đào tạo do chi phí tính toán của chúng, cho đến khi đổi tên cuối cùng thành deep learning vào năm 2006 với công trình của Hinton và cộng sự đánh dấu một kỷ nguyên mới của mạng nơ-ron. Bài báo nói trên đề xuất một chiến lược mới để đào tạo hiệu quả các mạng lưới niềm tin sâu sắc bằng cách sử dụng đào tạo trước và tinh chỉnh.

Việc triển khai đầu tiên của thuật toán perceptron là với Mark I Perceptron vào năm 1957, có khả năng đào tạo một tế bào thần kinh, sau đó, vào năm 1960, Widrow và Hoff đã phát triển Adaline cho một mạng một lớp và Madaline cho một mạng nhiều lớp. Những mô hình đó đã được đào tạo mà không có sự lan truyền ngược mà không được giới thiệu cho đến năm 1986 với công trình của Rumelhart và cộng sự. Dẫn đến thành tựu của LeCun et al. vào năm 1998 đã sử dụng thành công học dựa trên gradient để đào tạo mạng nơ-ron tích chập để nhận dạng chính xác các chữ cái viết tay của bảng chữ cái bằng cách sử dụng bộ dữ liệu MNIST.

Cho đến năm 2006, các mạng nơ-ron sâu được coi là không khả thi do chi phí tính toán cao, công trình của Hinton et al. đã đề xuất một chiến lược để đào tạo chúng

một cách hiệu quả. Bốn năm sau là bước đột phá đầu tiên trong lĩnh vực nhận dạng giọng nói, khi sử dụng mạng nơ-ron đầu tiên đã làm giảm tỷ lệ lỗi khoảng 30%, nhưng phổ biến nhất là nhận dạng hình ảnh với phân loại ImageNet với mạng nơ-ron tích chập sâu vào năm 2012.

Sự gia tăng kích thước bộ dữ liệu đóng một vai trò quan trọng trong sự tiến bộ của học sâu, một trong những bộ dữ liệu lớn đầu tiên là MNIST (Viện Tiêu chuẩn và Công nghệ Quốc gia Sửa đổi) một bộ dữ liệu chứa hàng chục nghìn chữ số viết tay được quét và nhãn của chúng, được sử dụng rất phổ biến để kiểm tra các thuật toán học máy khác nhau. Khi dữ liệu kỹ thuật số trở nên sẵn có hơn, có các bộ dữ liệu lớn hơn và phức tạp xuất hiện như ImageNet với hàng triệu ví dụ được gắn nhãn xác định lại mạng nơ-ron hiện đại. Nhưng khi kích thước của các bộ dữ liệu tăng lên và các mạng trở nên sâu hơn, tính toán càng cao khiến việc triển khai các mạng nơ-ron trên CPU không đủ. Do đó, các nhà nghiên cứu, nhận thấy sự tương đồng trong thuật toán đồ họa tính toán và thuật toán mạng nơ-ron, bắt đầu triển khai mạng nơ-ron trên GPU ngay sau khi chúng trở nên linh hoạt về khả năng lập trình. Steinkraus et al. vào năm 2005 lần đầu tiên triển khai thành công mạng nơ-ron được kết nối hoàn toàn 2 lớp và đạt được tốc độ gấp 3 lần so với triển khai dựa trên CPU, chứng minh rằng NN rất phù hợp cho tính toán song song mà GPU cung cấp. Sau đó, việc sử dụng GPU đã vượt quá mục đích ban đầu của chúng trong việc tăng tốc đồ họa 3D và GPU đa năng GP-GPU với ngôn ngữ lập trình CUDA của NVIDIA đã có sẵn cho phép sử dụng rộng rãi trong học sâu, vào năm 2009 Raina et al. đã giảm thời gian đào tạo cho các mạng tin tưởng sâu từ vài tuần xuống còn khoảng một ngày, nó nhanh hơn khoảng 70 lần.

Theo Geoffrey Hinton (được coi là cha đẻ của Deep Learning), deep learning (DL) cho phép các mô hình tính toán bao gồm nhiều lớp xử lý để học các biểu diễn của dữ liệu với nhiều cấp độ trừu tượng. Do thực tế này, DL đã cho thấy sự áp dụng rộng rãi trong các lĩnh vực khác nhau với các tính năng phức tạp, trong đó biểu diễn ab-stracted có thể giảm bớt đáng kể quá trình xử lý và phân tích xuôi dòng, ví dụ: chúng tôi đã áp dụng lớp DNN và bộ mã hóa tự động (AE) khác nhau để cải thiện độ chính xác phát hiện phần mềm độc hại.

Phương pháp DL là một phương pháp học đại diện nhiều cấp độ, thu được bằng cách kết hợp các mô-đun phi tuyến tính nhưng đơn giản thay đổi biểu diễn ở một cấp độ (đầu vào thô ở cấp độ bắt đầu) thành biểu diễn ở cấp độ cao hơn, trừu tượng hơn một chút. Khía cạnh chính của DL là các tính năng trong các lớp này không được thiết kế bởi các kỹ sư mà được học từ dữ liệu bằng cách sử dụng một

chuyên gia học tập có mục đích chung. Công thức DL này, còn được gọi là mạng nơ-ron sâu (DNN) rất hữu ích đối với dữ liệu mà các tính năng không liên quan từ góc độ miền. Đối với văn bản và dữ liệu không gian nhịp độ hoặc dữ liệu chuỗi khác, kiến trúc DL như Mạng nơ-ron lặp lại, hoặc các biến thể của nó như Bộ nhớ dài hạn-ngắn hạn và Đơn vị tái phát có cổng phổ biến hơn. Điều này là do, trong DNN thông thường, sự lan truyền ngược bị hỏng, do vòng lặp lặp lại, được giải quyết bằng một kỹ thuật gọi là Truyền ngược qua thời gian. Tương tự, đối với hình ảnh và video, kiến trúc DL như Mạng nơ-ron tích chập (CNN) hoạt động tốt hơn vì chúng có thể hiểu mối tương quan không gian của cường độ pixel trong hình ảnh tốt hơn và hiệu quả hơn.

1.2 Các định nghĩa

1.2.1 Học máy

Học máy là một lĩnh vực điện toán gần đây đã chứng kiến sự bùng nổ trong các ứng dụng và nhận dạng hộ gia đình. Một khái niệm lỏng lẻo của học máy là sử dụng máy tính cho các tác vụ mà chúng có thể học hỏi và trở nên giỏi hơn mà không cần lập trình rõ ràng. Một số ví dụ về điều này đặc biệt phù hợp trong ngành y tế, nơi thị giác máy tính đã trở nên nhanh hơn và đáng tin cậy hơn con người trong công việc. Tuy nhiên, những ý tưởng đằng sau học máy không phải là mới; Alan Turing, đôi khi được gọi là cha đẻ của học máy, đã nói về việc máy tính có thể cải thiện hiệu suất của chúng trong các nhiệm vụ được giao mà không cần lập trình bổ sung kể từ năm 1950, và John McCarthy, một người sáng lập thường xuyên khác của thế giới, đã đặt ra thuật ngữ trí tuệ nhân tạo vào năm 1955. Vì vậy, trong khi sự gia tăng phổ biến của công cụ tính toán mạnh mẽ này đã rõ rệt hơn trong những thập kỷ gần đây, các khái niệm này đã tồn tại hơn nửa thế kỷ. Ngoài ra còn có nhiều thuật ngữ khác nhau liên quan đến eld này, chẳng hạn như khoa học dữ liệu, học thống kê, học sâu và các thuật ngữ khác. Danh sách các thuật ngữ này, cũng như một số biệt ngữ liên quan đến học máy, có thể được tìm thấy trong phần Lingo và Jargon ở cuối các ghi chú này. Có thể hữu ích nếu bạn tự làm quen với các thuật ngữ này trước khi sử dụng các ghi chú này.

Học máy là một ngành tập trung vào hai câu hỏi liên quan đến nhau: Làm thế nào để xây dựng các hệ thống máy tính tự động cải thiện thông qua kinh nghiệm? và Các định luật tính toán-thông tin-lý thuyết thống kê cơ bản chi phối tất cả các hệ thống học tập, bao gồm máy tính, con người và tổ chức là gì? Nghiên cứu về học máy rất quan trọng để giải quyết các câu hỏi khoa học và kỹ thuật cơ bản này và đối với phần mềm máy tính thực tế cao mà nó đã sản xuất và triển khai trên nhiều ứng

dụng. Học máy đã tiến bộ dramati trong hai thập kỷ qua, từ sự tò mò trong phòng thí nghiệm đến một công nghệ thực tế được sử dụng rộng rãi cho thương mại. Trong trí tuệ nhân tạo (AI), học máy đã nổi lên như một phương pháp được lựa chọn để phát triển phần mềm thực tế cho thị giác máy tính, nhận dạng giọng nói, xử lý thiết bị ngôn ngữ tự nhiên, điều khiển robot và các ứng dụng khác. Nhiều nhà phát triển hệ thống AI hiện nay nhận ra rằng, đối với nhiều ứng dụng, việc đào tạo một hệ thống bằng cách cho nó thấy hành vi đầu vào-đầu ra mong muốn có thể dễ dàng hơn nhiều so với việc lập trình nó theo cách thủ công bằng cách dự đoán phản hồi mong muốn cho tất cả các đầu vào có thể. Tác động của việc học tập cũng đã được cảm nhận rộng rãi trong khoa học máy tính và trên một loạt các công ty liên quan đến các vấn đề sử dụng nhiều dữ liệu, chẳng hạn như dịch vụ tiêu dùng, chẩn đoán lỗi trong các hệ thống phức tạp và kiểm soát chuỗi hậu cần. Đã có một loạt các tác động tương tự trên các khoa học thực nghiệm, từ sinh học đến vũ trụ học đến khoa học xã hội, vì các phương pháp học máy đã được phát triển để phân tích dữ liệu thực nghiệm thông lượng cao theo những cách mới.

1.2.2 Học sâu

1.2.3 Trí tuệ nhân tạo

Trí tuệ nhân tạo (AI) đã phát triển từ khái niệm AI mạnh, bắt chước trí thông minh của con người, thành sự kết hợp của AI yếu có thể giải quyết một số vấn đề nhất định. Các nghiên cứu về AI yếu khám phá các cách xây dựng các thuật toán có thể học hỏi từ dữ liệu và đưa ra dự đoán. Học máy là một nhánh của khoa học máy tính xây dựng các thuật toán được hướng dẫn bởi dữ liệu. Trong số đó, mạng nơ-ron (NN), bao gồm các nút và trọng số, là một trong những loại thuật toán AI đầu tiên được phát triển. Sức mạnh tính toán của các mạng này phụ thuộc vào chất lượng và số lượng dữ liệu đào tạo, cho phép các mạng này cập nhật trọng số của các kết nối. Các cấu trúc mạng đơn giản chỉ có một vài lớp được gọi là mạng nơ-ron học "nông", trong khi các cấu trúc mạng sử dụng nhiều lớp lớn và được gọi là "sâu" học mạng nơ-ron. Các cấu trúc học sâu được gọi là mạng nơ-ron tích chập (CNN), có thể trích xuất nhiều đặc điểm từ các lớp bộ lọc trừu tượng, chủ yếu được sử dụng để xử lý các hình ảnh lớn và phức tạp. Học sâu đang được tăng tốc bởi sự phát triển của các thuật toán lan truyền ngược tự học giúp tinh chỉnh dần dần kết quả từ dữ liệu, cũng như bằng cách tăng sức mạnh tính toán. Do những tiến bộ công nghệ nhanh chóng này, AI, được đại diện bởi học sâu, có thể được sử dụng cho các vấn đề thực tế và được áp dụng trên tất cả các lĩnh vực của xã hội.

Nghiên cứu học thuật về trí tuệ nhân tạo (AI), được định nghĩa rộng rãi là khả năng của một cỗ máy để thực hiện các nhiệm vụ thường liên quan đến hành vi thông minh của con người, bắt đầu từ những năm 1940 và 1950, khi các nhà khoa học máy tính đầu tiên, chẳng hạn như Turing và McCarthy, lần đầu tiên đề xuất các định nghĩa chính thức về trí thông minh máy móc.1,2 Lĩnh vực xuất hiện trong những thập kỷ tiếp theo bao gồm 2 trường phái tư tưởng chính, kết nối và tượng trưng. Cách tiếp cận trường phái kết nối đối với AI cho rằng, giống như tế bào thần kinh sinh học, đơn vị cơ bản của hệ thống AI phải là các nút được kết nối với nhau với sức mạnh khác nhau, trong khi cách tiếp cận trường phái biểu tượng cho rằng các hệ thống thông minh có thể được thiết kế để xử lý trực tiếp các khái niệm tượng trưng hoặc trừu tượng — các nhà bình luận sau này sẽ lưu ý rằng 2 mô hình không nhất thiết phải loại trừ lẫn nhau.3 Hai mô hình này làm nền tảng cho phần lớn công việc dẫn đến AI hiện đại

Mặc dù đã được nghiên cứu trong hơn một thế kỷ, sức mạnh tính toán và những tiến bộ cần thiết cho ứng dụng rộng rãi của AI chỉ mới xuất hiện gần đây.4 Học máy (ML), một tập hợp con của AI liên quan đến các thuật toán có thể tạo ra các mô hình dự đoán bắt nguồn từ việc tiếp xúc với dữ liệu đào tạo thay vì thiết kế tiên nghiệm đầy đủ, có liên quan đặc biệt đến y học và là trọng tâm của đánh giá này. Hầu hết các bác sĩ lâm sàng đều có một số quen thuộc với ML ngoài cuộc sống nghề nghiệp của họ. Trợ lý kỹ thuật số, chẳng hạn như Alexa của Amazon; Phần mềm GPS; hệ thống quản lý nội dung; và các nền tảng truyền thông xã hội đều sử dụng ML để đưa ra dự đoán chính xác cho việc sử dụng của chúng và giúp những người dùng đó điều hướng hiệu quả lượng dữ liệu khổng lồ có sẵn trong thời đại hiện đại. Cần có nhiều nỗ lực hơn nữa để giáo dục các bác sĩ về ML và cách sử dụng công nghệ trong chăm sóc sức khỏe.

Các khoa của ML làm cho nó trở thành một giải pháp hấp dẫn khi xem xét những thách thức do khối lượng và độ phức tạp của kiến thức y học hiện đại đặt ra. Tính đến năm 2020, khối lượng kiến thức y tế ước tính tăng gấp đôi sau mỗi 2 đến 3 tháng và do một cá nhân không thể giữ lại tất cả kiến thức đó, y học hiện đại ngày càng yêu cầu các bác sĩ lâm sàng tiêu hóa hiệu quả và sử dụng thực tế khối lượng dữ liệu có sẵn cho họ.5 Trong thập kỷ qua, ML đã nổi lên như một phần ngày càng hữu ích về mặt lâm sàng trong giải pháp cho thách thức này. Đánh giá này nêu bật những điều cơ bản của ML và tóm tắt các ứng dụng gần đây của nó trong chẩn đoán ung thư, tiên lượng và lựa chọn điều trị (Hình 1), cũng như những bài học chính cho các bác sĩ lâm sàng. Phần lớn nghiên cứu ML cho đến nay liên quan đến dữ liệu hồi cứu;

trừ khi có quy định khác, các bài báo được thảo luận ở đây có bản chất hồi cứu và tiền lâm sàng.

1.2.4 Cấu trúc dữ liệu

1.2.5 Loại dữ liệu

1.3 Các ứng dụng của học sâu

1.4 So sánh học sâu và học máy