MỤC LỤC

MỞ ĐẦU

Thị trường chứng khoán vẫn luôn chiếm vị thế nhất định trong thị trường đầu tư, không chỉ đem lại nguồn thu nhập thụ động mà còn giúp cho nền kinh tế phát triển một cách mạnh mẽ. Do đó đã thu hút được đông đảo các nhà đầu tư từ các cá nhân nhỏ lẻ cho đến các công ty lớn, bởi đây là thị trường tiềm năng đem đến nguồn tiền tích cực. Cùng với sự phát triển của công nghệ, việc truy xuất các dữ liệu giao dịch, các thông tin trực quan của thị trường đã giúp các nhà đầu tư có cơ sở và quyết định đúng đắn để bắt đầu công cuộc đầu tư của mình, chính vì lẽ đó mà sự phát triển của các sàn chứng khoán trở nên mạnh mẽ và năng động hơn bao giờ hết

Tuy nhiên, thị trường chứng khoán có tính chất ngẫu nhiên và phi tuyến tính, điều đó có nghĩa là việc dự đoán xu hướng của chứng khoán chỉ bằng những dữ liệu lịch sử giá là một bài toán đầy thách thức, đồng thời thị trường chứng khoán cũng bị ảnh hưởng bởi nhiều yếu tố như thị trường cung cầu, tình hình thế giới, tình hình chính trị xã hội, biến động kinh tế, báo cáo tài chính. Dẫn đến việc giải quyết bài toán này gần như là không thể ở những năm về trước.

Nhưng nhờ có sự bùng nổ của lĩnh vực trí tuệ nhân tạo, cụ thể là sự phát triển nhảy vọt của những kỹ thuật học sâu, các bài toán đầy thách thức ở các lĩnh vực khác nhau đã lần lượt được giải quyết. Các ứng dụng về xử lý ảnh trong lĩnh vực y khoa như dự đoán bệnh qua ảnh CT, ứng dụng xử lý ngôn ngữ tự nhiên, nhận diện giọng nói đã được phát triển và cải thiện vượt bậc. Với những đột phá về công nghệ và kỹ thuật này đã góp phần tạo nên một nền tảng vững chắc để xây dựng mô hình dự đoán và các tri thức để cải tiến.

I. TỔNG QUAN

1.1. Tổng quan về thị trường chứng khoán

1.1.1. Các thông tin tổng quát

Chứng khoán là loại tài sản, bao gồm: Cổ phiếu, trái phiếu, chứng chỉ quỹ; Chứng quyền, chứng quyền có bảo đảm, quyền mua cổ phần, chứng chỉ lưu ký; Chứng khoán phát sinh; Các loại chứng khoán khác do Chính phủ quy định.

***) Tầm quan trọng của thị trường chứng khoán:**

- **Đối với nhà đầu tư:** Thị trường chứng khoán là một kênh đầu tư tiềm năng và phong phú với đa dạng danh mục đầu tư. Các loại chứng khoán này khác nhau về tính chất, giá cả, mức độ rủi ro và tính thanh khoản, do vậy các nhà đầu tư có thể lựa chọn mã cổ phiếu phù hợp với sở thích và khả năng của mình. Việc tham gia vào thị trường chứng khoán cũng dễ dàng, thủ tục đơn giản giúp cho các cá nhân nhỏ lẻ hay các tổ chức có nguồn vốn lớn có thể tiếp cận một cách nhanh chóng.
- **Đối với doanh nghiệp:** Thị trường chứng khoán giúp các doanh nghiệp đa dạng các hình thức huy động vốn đầu tư bằng việc phát hành cổ phiếu hay trái phiếu, điều này giúp cho cho doanh nghiệp có được một số vốn đầu tư dài hạn mà còn tránh được các khoản vay ngân hàng với lãi suất cao. Hơn thế nữa, việc doanh nghiệp có chứng khoán niêm yết trên sàn giao dịch giúp tạo được niềm tin và sự uy tín đối với công chúng, nhờ đó mà doanh nghiệp có thể huy động được nguồn vốn một cách linh hoạt, hiệu quả và rẻ hơn. Ngoài ra việc mở của trị trường chứng khoán còn giúp cho doanh nghiệp thu hút được thêm các nguồn vốn từ thị trường quốc tế, mở rộng cơ hội của doanh nghiệp.
- **Đối với nền kinh tế:** Thị trường chứng khoán tạo ra các công cụ có tính thanh khoản cao, có thể tích tụ, tập trung và phân phối vốn, chuyển thời hạn của vốn phù hợp với yêu cầu phát triển kinh tế, giúp tạo vốn cho nền kinh tế quốc dân. Nhờ có thị trường chứng khoán, Chính phủ có thể huy động các nguồn lực tài chính mà không bị áp lực về lạm phát, đặc biệt khi nguồn vốn đầu tư từ khu vực nhà nước còn hạn chế.

***) Khái niệm về cổ phiếu**

Cổ phiếu là loại chứng khoán xác nhận quyền và lợi ích hợp pháp của người sở hữu đối với một phần vốn cổ phần của tổ chức phát hành. Hay nói cách khác cổ phiếu là giấy chứng nhận số tiền mà cổ đông đầu tư vào doanh nghiệp.

Các công ty cổ phần phát hành cổ phiếu để huy động vốn vào việc kinh doanh của họ. Cổ phiếu được mua và bán chủ yếu trên các sàn giao dịch chứng khoán. Tại thị trường chứng khoán Việt Nam, một cổ phiếu cơ sở là đại diện cho 10.000

đồng vốn điều lệ của doanh nghiệp. Các nhà đầu tư mua bán cổ phiếu vì các mục đích sau:

- Khai thác lợi nhuận từ việc chênh lệch giá mua và bán, hưởng cổ tức – phần lợi nhuận sau thuế để chia cho cổ đông của công ty.
- Để tham gia biểu quyết, nắm quyền quyết định, điều hành của công ty/doanh nghiệp.

1.2. Bài toán dự đoán xu hướng giá chứng khoán

Một trong những bài toán được phần đông những nhà khoa học, nhà kinh tế quan tâm nhất chính là bài toán dự đoán xu hướng giá chứng khoán. Ngay từ những ngày đầu khi thị trường chứng khoán đầu tiên được thành lập vào những năm 1600, các nhà kinh tế đã thấy được sự tiềm năng mà chứng khoán đem lại.

Lần lượt các nhà tiên phong về phương pháp dự đoán chuỗi thời gian đã xuất hiện, có thể kể đến mô hình AutoRegressive được nhà thống kê Udny Yule và các đồng nghiệp phát minh vào những năm 1920. Mô hình này là nền tảng cho một số mô hình thống kê và kinh tế lượng sau này như ARMA, ARIMA.

Điều đó cho thấy từ trước những năm phát triển vượt bậc của lĩnh vực trí tuệ nhân tạo, sự quan tâm của các nhà khoa học đối với bài toán này là không hề nhỏ. Trong những năm gần đây, các mô hình máy học đã được ứng dụng vào bài toán này để hỗ trợ các nhà đầu tư tạo ra lợi nhuận, tuy nhiên với những mô hình máy học truyền thống thì độ chính xác vẫn còn những hạn chế nhất định. Tuy nhiên với sự phát triển của những mô hình học sâu, việc nhận dạng được những mẫu phi tuyến tính trong chuỗi thời gian của chứng khoán đã trở nên dễ tiếp cận hơn bao giờ hết. Một hướng tiếp cận khá phổ biến và hiệu quả trong những năm gần đây cho bài toán dự đoán chuỗi thời gian là sử dụng mô hình LSTM đây là mô hình học sâu thu hút được nhiều sự quan tâm của các nhà nghiên cứu trong và ngoài nước. LSTM được sử dụng rất nhiều cho các bài toán có dữ liệu thời gian hay tuần tự như dịch máy, nhận diện giọng nói, dự báo thời tiết với độ chính xác cao.

II. CƠ SỞ NGHIÊN CỨU

2.1. Các kiến trúc Recurrent Neural Networks

2.1.1 Giới thiệu

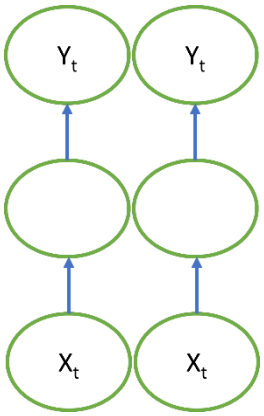
Mạng Recurrent Neural Networks (RNN) được phát triển dựa trên công trình của David Rumelhart vào năm 1986. RNN ra đời với ý tưởng chính là sử dụng

một bộ nhớ để lưu lại chuỗi các thông tin từ những bước xử lý trước đó để đưa ra dự đoán cho bước hiện tại.

Các loại mô hình RNN

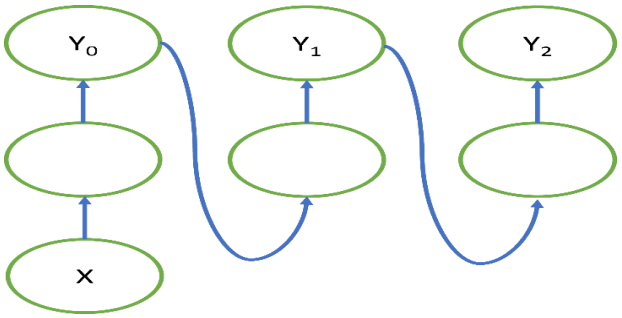
Dựa vào số lượng xử lý của chuỗi đầu vào và chuỗi đầu ra, mạng RNN được chia thành 4 loại chính với nhiều kiến trúc khác nhau.

One-to-One: Đây là mạng RNN truyền thống sử dụng kiến trúc One-to-One, cặp (xt , yt) duy nhất.



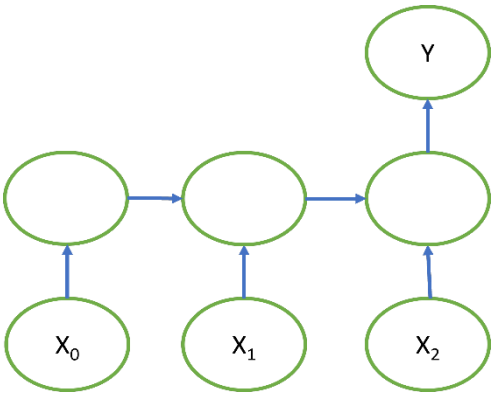
Hình 2.1 – Mô phỏng One-to-One RNN

One-to-Many: Đối với One-to-Many, một đầu vào duy nhất tại xt có thể tạo ra nhiều đầu ra: (yt0, yt1, yt2). Music Generation là một ví dụ.



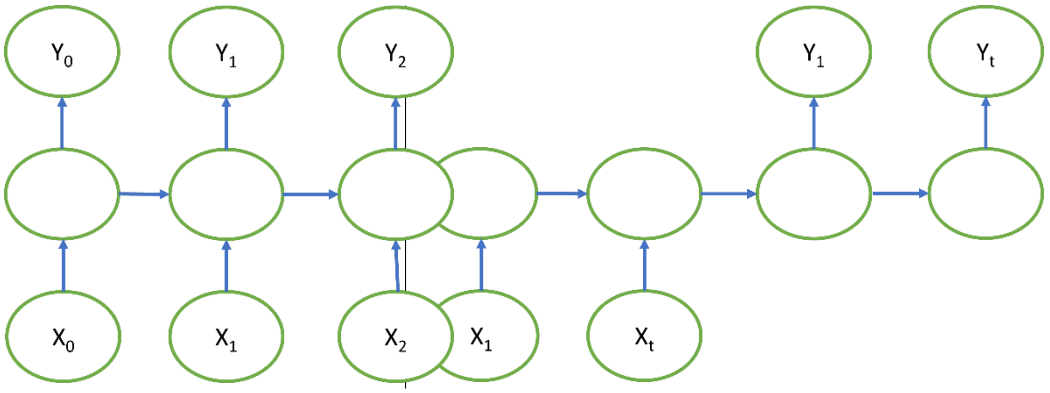
Hình 2.2 – Mô phỏng One-to-Many RNN

Many-to-One: Trong trường hợp này, nhiều đầu vào từ các bước thời gian khác nhau tạo ra một đầu ra duy nhất. Ví dụ: (xt , xt+1, xt+2) có thể tạo ra một đầu ra duy nhất yt . Các mạng như vậy được sử dụng trong phân tích cảm xúc (sentiment anslysis) hoặc phát hiện cảm xúc (emotion detection),.. trong đó nhãn phụ thuộc vào một chuỗi các từ.



Hình 2.3 – Mô phỏng Many-to-One RNN

Many-to-Many: Có nhiều loại mô hình thuộc nhóm. Hình 2.4 là một ví dụ, trong đó ba đầu vào tạo ra ba đầu ra, hai đầu vài tạo ra hai đầu ra, lưu ý là số lượng đầu vào và đầu ra có thể khác nhau. Một ví dụ thực tế nhưng dịch máy, ví dụ: hệ thống dịch các từ tiếng Anh sang tiếng Pháp hoặc ngược lại.



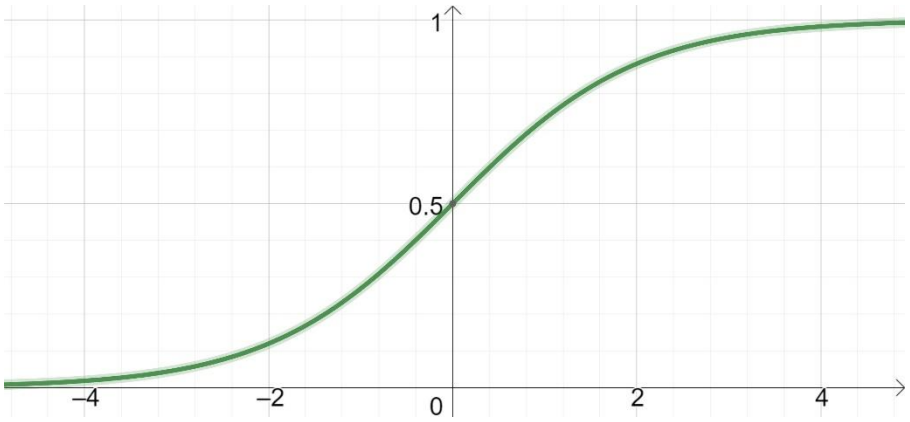
Hình 2.4 – Mô phỏng Many-to-Many RNN

Các hàm kích hoạt phổ biến trong các mô hình RNNs

Các hàm kích hoạt (Activation function) quyết định khi nào một nơ-ron được kích hoạt, khi nào thông tin được truyền qua nơ-ron khác dựa trên một phép biến đổi phi tuyến được tính toán bằng giá trị đầu vào, đầu ra sau khi biến đổi được dùng làm đầu vào cho nơ-ron tiếp theo.

Hàm Sigmoid: Nếu đầu vào lớn, hàm số sẽ cho đầu ra gần với 1; với đầu vào nhỏ (rất âm), hàm số sẽ cho đầu ra gần với 0. Trước đây, hàm kích hoạt này được sử dụng nhiều vì có đạo hàm rất đẹp. Tuy nhiên gần đây, hàm sigmoid chỉ được sử dụng ở output layer khi yêu cầu của đầu ra là các giá trị nhị phân.

$$Sigmoid(x) = \frac{1}{1+e^{(-x)}} \quad (2.1)$$

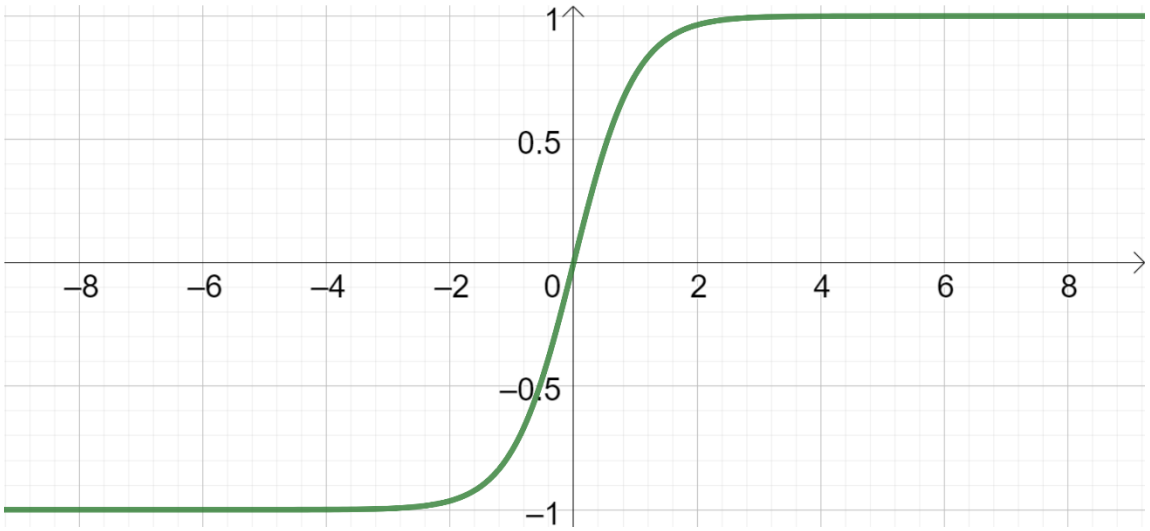


Hình 2.5 – Hàm kích hoạt Sigmoid

Hàm Tanh: Một hàm tương tự thường được sử dụng và mang lại hiệu quả tốt hơn là hàm tanh, hàm số này có tính chất đầu ra chạy từ -1 đến 1, khiến cho nó có tính chất zero-centered, thay vì chỉ dương như hàm sigmoid.

Một nhược điểm dễ nhận thấy là khi đầu vào có trị tuyệt đối lớn (rất âm hoặc rất dương), đạo hàm của cả sigmoid và tanh sẽ rất gần với 0. Điều này đồng nghĩa với việc các hệ số tương ứng với unit đang xét sẽ gần như không được cập nhật khi sử dụng công thức cập nhật gradient desent. Thêm nữa, khi khởi tạo các hệ số cho multilayer neural network với hàm kích hoạt sigmoid, chúng ta phải tránh trường hợp đầu vào một hidden layer nào đó quá lớn, vì khi đó đầu ra của hidden layer đó sẽ rất gần với 0 hoặc 1, dẫn đến đạo hàm bằng 0 và gradient desent hoạt động không hiệu quả.

$$Tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (2.2)$$

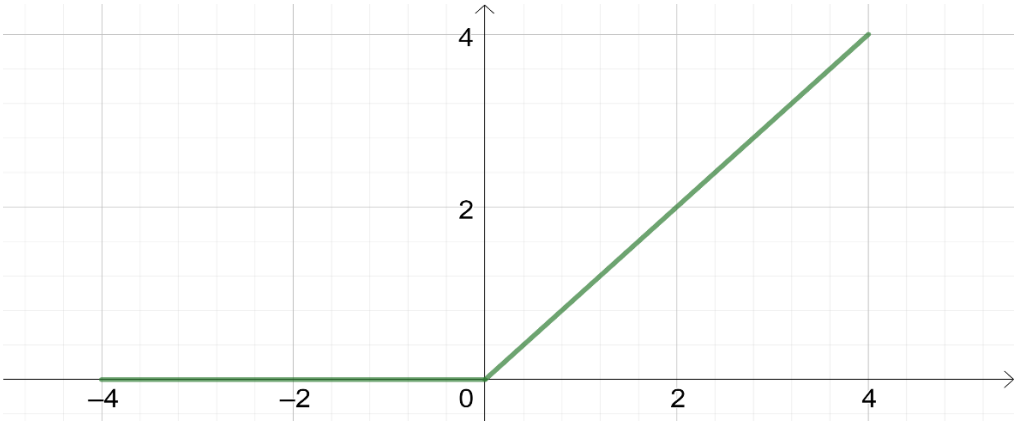


Hình 2.6 – Hàm kích hoạt Tanh

Hàm ReLU: ReLU (Rectified Linear Unit) được sử dụng rộng rãi gần đây vì tính đơn giản của nó. Hàm ReLU có công thức toán học rất đơn giản, rất lợi về mặt

tính toán, đạo hàm của nó bằng 0 tại các điểm âm, bằng 1 tại các điểm dương. ReLU được chứng minh giúp cho việc huấn luyện các multilayer neural network và deep network (rất nhiều hidden layer) nhanh hơn rất nhiều so với hàm tanh.

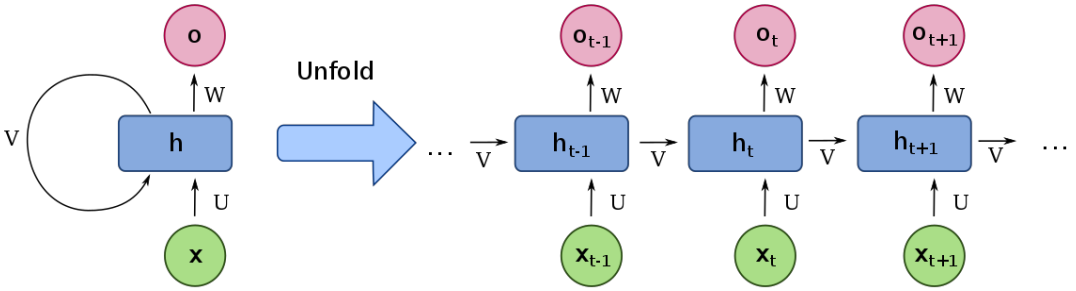
$$\text{ReLu}(x) = \max(0, x) \qquad (2.3)$$



Hình 2.7 – Hàm kích hoạt ReLU

2.1.2. Mạng Recurrent Neural Networks

Khác với mạng nơ-ron truyền thống có tất cả các đầu vào và cả đầu ra là độc lập với nhau tức là chúng không liên kết thành chuỗi với nhau, mạng nơ-ron tuần hoàn là một loại mạng nơ-ron trong đó đầu ra của bước trước được cung cấp làm đầu vào cho bước hiện tại. Một vòng lặp cho phép thông tin có thể truyền đi từ bước này sang bước tiếp theo.



Hình 2.8 – Mô hình RNN truyền thống

Hình 2.8 mô tả dữ liệu được đưa vào một cách tuần tự, ở mỗi bước thời gian t ta sẽ đưa dữ liệu đầu vào xt và nhận đầu ra là ot với g là hàm kích hoạt thường là Tanh hoặc ReLU), có thể thấy ht mang thông tin từ trạng thái trước đó ht−1 và đầu vào của trạng thái hiện tại xt nên ht giống như một bộ nhớ các đặc điểm của các đầu ra từ x1 đến xt .

$$h_t = g_h(W_{hx}x_t + U_{hh}h_{t-1} + b_h) \quad (2.4)$$

$$o_t = g_o(W_{oh}h_t + b_o) \quad (2.5)$$

Các vấn đề về Gradient trong quá trình huấn luyện của mạng RNN

Về lý thuyết, RNN có thể sử dụng thông tin theo chuỗi dài tùy ý, nhưng trên thực tế, chúng bị hạn chế khi chỉ nhìn lại một vài bước trong các trường hợp dữ liệu quá dài mang một lượng thông tin lớn thì mạng RNN không thể nhớ được các thông tin xa trước đó. Vanishing Gradient (Gradient biến mất) và Exploding Gradient (Gradient bùng nổ) là những vấn đề thường gặp phải khi sử dụng các kỹ thuật tối ưu hóa trọng số dựa trên gradient để huấn luyện mạng nơ-ron xuất phát từ việc lựa chọn hàm kích hoạt không hợp lý hoặc số lượng các lớp ẩn của mạng quá lớn.

Ta có thể coi gradient là độ dốc của một hàm. Gradient càng cao, độ dốc càng lớn và mô hình càng có thể học nhanh hơn. Nhưng nếu độ dốc bằng 0, mô hình ngừng học. Một gradient chỉ đơn giản là đo lường sự thay đổi của tất cả các trọng số liên quan đến sự thay đổi của sai số.

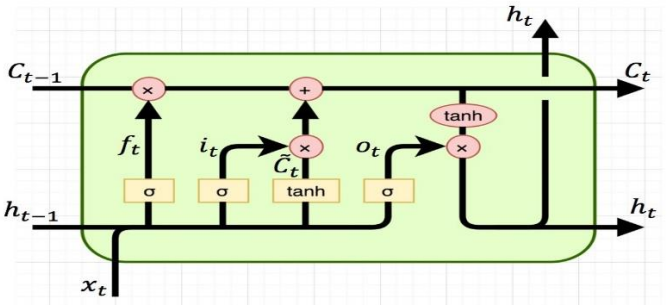
Vấn đề Exploding Gradient xảy ra khi thuật toán đánh trọng số quá cao, may mắn thay vấn đề này có thể được giải quyết bằng kỹ thuật Gradient Clipping.

Vấn đề Vanishing Gradient xảy ra khi các giá trị của gradient quá nhỏ và kết quả là mô hình ngừng học hoặc mất quá nhiều thời gian. Đây là một vấn đề lớn trong những năm 1990 và khó giải quyết hơn nhiều so với vấn đề Exploding Gradient. Ta có thể xử lý vấn đề này bằng cách sử dụng hàm kích hoạt ReLu có đạo hàm bằng 0 tại các điểm âm hoặc 1 tại các điểm dương, nên ta phần nào có thể kiểm soát vấn đề mất mát đạo hàm.

May mắn thay, những vấn đề của mạng RNN đã được giải quyết thông qua các biến thể khác của RNN, phổ biến có thể kể đến là mạng LSTM của Sepp Hochreiter và Juergen Schmidhuber và mạng GRU.

2.2. Giới thiệu mạng Long Short-Term Memory

Mạng Long Short-Term Memory (LSTM) là một phiên bản mở rộng của RNN được đề xuất bởi Sepp Hochreiter và Ju¨rgen vào năm 1997. LSTM được thiết kế để giải quyết vấn đề Gradient Vanishing trong mạng RNN do các bài toán phụ thuộc xa (Long-term dependencies) gây ra. Mạng LSTM có thể bao gồm nhiều tế bào LSTM (LSTM memory cell) liên kết với nhau và kiến trúc cụ thể của mỗi tế bào được biểu diễn như trong Hình 2.9. Kiến trúc LSTM bao gồm một trạng thái tế bào (cell state) C_t và ba cổng (gate) - một cổng quên (forget gate) f_t một cổng vào (input gate) i_t và một cổng ra (output gate) o_t.

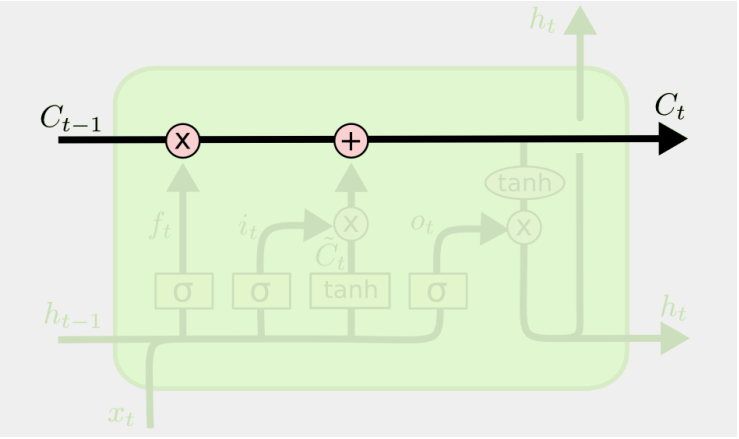


Hình 2.9 – Sơ đồ biểu diễn kiến trúc bên trong của một tế bào LSTM

Tại mỗi bước thời gian t, các cổng đều lần lượt nhận giá trị đầu vào x_t và giá trị h_{t-1} có được từ giá trị đầu ra của trạng thái trước đó.

Trạng thái tế bào (Cell state)

Trạng thái tế bào chính là đường chạy ngang như trong Hình 2.10. Các thông tin có thể dễ dàng truyền đi thông suốt qua nhiều tế bào LSTM mà không sợ bị thay đổi do trạng thái tế bào đi xuyên suốt qua các nơ-ron và chỉ tương tác tuyến tính đôi chút.

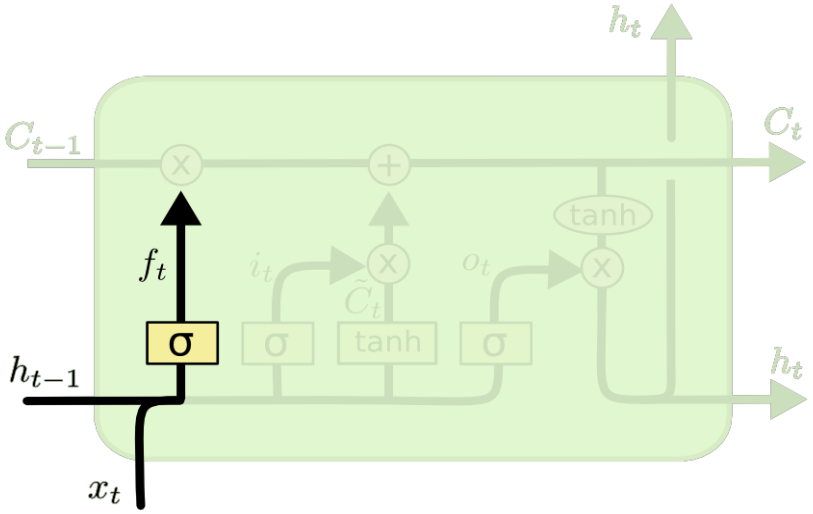


Hình 2.10 – Sơ đồ biểu diễn trạng thái tế bào LSTM

Cổng quên (Forget gate)

Cổng quên kiểm tra trạng thái ẩn trước đó h_{t-1} và đầu vào hiện tại x_t, sau đó xác định bộ phận nào sẽ được loại bỏ khỏi trạng thái tế bào. Quyết định này sẽ phụ thuộc bởi hàm sigmoid f_t luôn giữ kết quả là một con số có giá trị từ trong khoảng [0,1] trong đó giá trị 1 là dấu hiệu để "giữ lại" toàn bộ thông tin và giá trị 0 là dấu hiệu để "loại bỏ" toàn bộ thông tin. Công thức được biểu diễn như sau:

$$ft = \sigma(Wf * h_{t-1} + Wf * xt + bf) \quad (2.6)$$



Hình 2.11 – Sơ đồ biểu diễn cổng quên trong LSTM

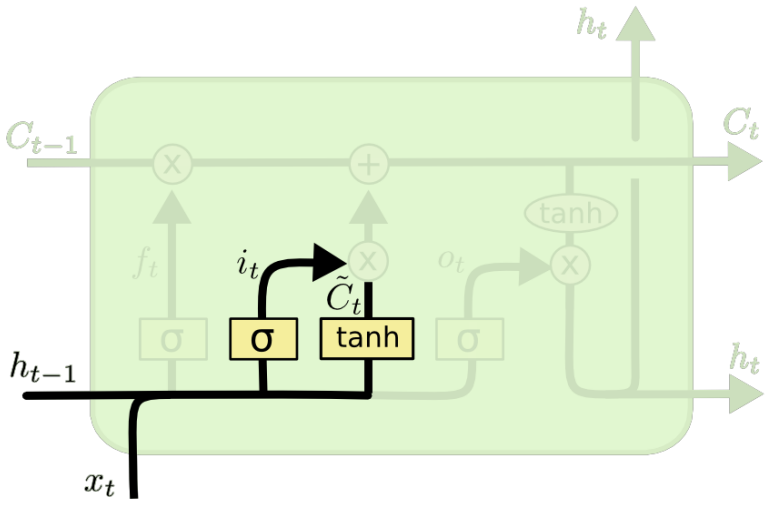
Cổng đầu vào (Input gate)

Cổng đầu vào quyết định thông tin mới nào sẽ được lưu trữ lại trạng thái tế bào. Hàm Sigmoid với công thức như bên dưới sẽ quyết định giá trị nào sẽ được cập nhật:

$$i_t = \sigma(W_i * h_{t-1} + W_i * x_t + b_i) \quad (2.7)$$

Sau đó, hàm Tanh tạo một vector $\tilde{C}_t$ kết hợp với i_t để cập nhật cho trạng thái tế bào. Công thức $\tilde{C}_t$ được biểu diễn như sau:

$$\tilde{C}_t = \tanh(W_c * h_{t-1} + W_c * x_t + b_c) \quad (2.8)$$



Hình 2.12 – Sơ đồ biểu diễn cổng đầu vào trong LSTM

Trạng thái tế bào mới C_t được cập nhật theo những thông tin được mang theo từ trạng thái tế bào trước đó C_{t-1} kết hợp với thông tin được tính toán tại trạng thái hiện tại và sẽ được truyền vào những tế bào LSTM tiếp theo. Công thức tính C_t được biểu diễn như sau:

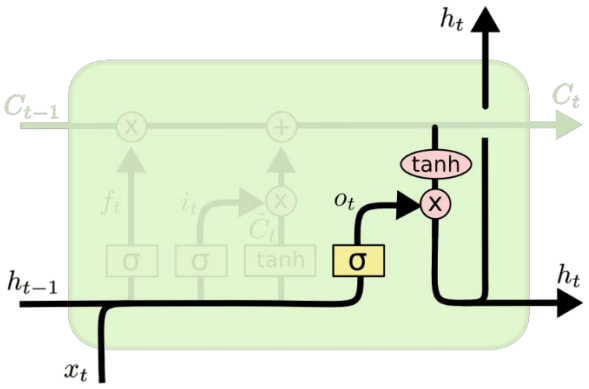
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (2.9)$$

Cổng đầu ra (Output gate)

Và cuối cùng, cổng đầu ra sẽ quyết định xuất cái gì. Hàm Sigmoid O_t sẽ lọc trạng thái tế bào để tiếp tục quyết định phần nào sẽ được truyền đến đầu ra. Để tạo đầu ra h_t , nhân o_t với trạng thái tế bào đi qua hàm Tanh để giữ tất cả các giá trị trong khoảng [-1,1]. Chi tiết có thể được biểu diễn như sau:

$$O_t = \sigma(W_O * h_{t-1} + W_O * x_t + b_O) \quad (2.10)$$

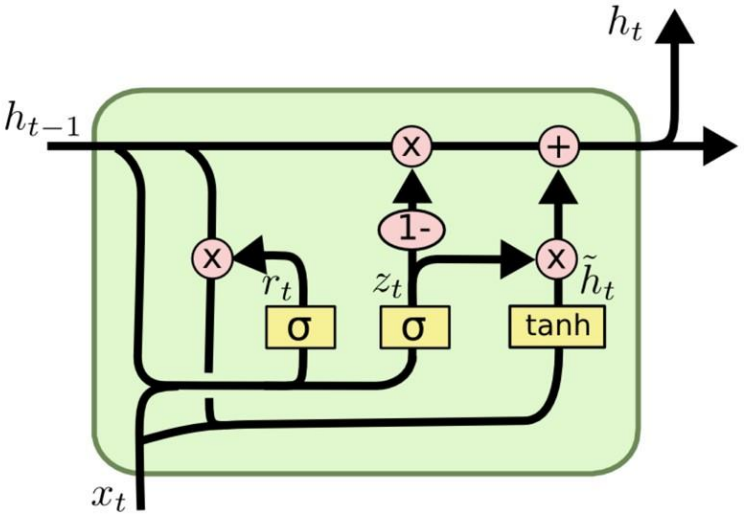
$$h_t = o_t * tanh(C_t) \quad (2.11)$$



Hình 2.13 – Sơ đồ biểu diễn cổng đầu ra trong LSTM

2.3. Giới thiệu mạng Gated Recurrent Unit

Mạng Gated Recurrent Unit (GRU) được đề xuất bởi Cho và các cộng sự vào năm 2014. Đây là một biến thể đơn giản hơn của LSTM nhưng vẫn giữ được các đặc điểm tương tự như LSTM vì nó cũng hoạt động để giải quyết vấn đề bộ nhớ ngắn hạn của các mô hình RNN. Thay vì sử dụng thông tin điều chỉnh "trạng thái tế bào (state cell)", nó sử dụng các trạng thái ẩn và thay vì ba cổng, nó có hai - một cổng đặt lại (reset gate) và một cổng cập nhật (update gate). Tương tự như các cổng trong LSTM, các cổng đặt lại và cập nhật kiểm soát số lượng và thông tin nào cần giữ lại, gửi đi.



Hình 2.14 – Sơ đồ biểu diễn mạng GRU

Ở cổng đặt lại (reset gate), hàm r_t quyết định bao nhiêu thông tin quá khứ cần quên, công thức của nó như sau:

$$r_t = \sigma(W_r * [h_{t-1}, x_t] + b_r) \quad (2.12)$$

Sau đó, cổng cập nhật (update gate) sẽ quyết định loại bỏ thông tin nào thông qua hàm z_t và thông tin mới nào cần thêm qua hàm h_t , cuối cùng hàm $h_{t'}$ đưa ra dữ liệu đầu ra.

$$z_t = \sigma(W_z * [h_{t-1}, x_t] + b_z) \quad (2.13)$$

$$_t = tanh(W_h * [h_{t-1}, x_t] + b_h)$$

$$h_t = (1 - z_t) * h_t + z_t * ht'$$

Thời gian huấn luyện mô hình GRU nhanh hơn LSTM và hiệu quả hơn do đã được cắt giảm bớt 1 cổng tính toán so với LSTM. Tuy nhiên, LSTM hoạt động tốt hơn GRU khi bài toán yêu cầu bộ nhớ dài hạn (long-term memory).

III. ĐỀ XUẤT VÀ THỬ NGHIỆM MÔ HÌNH HỌC SÂU

3.1. Đề xuất mô hình

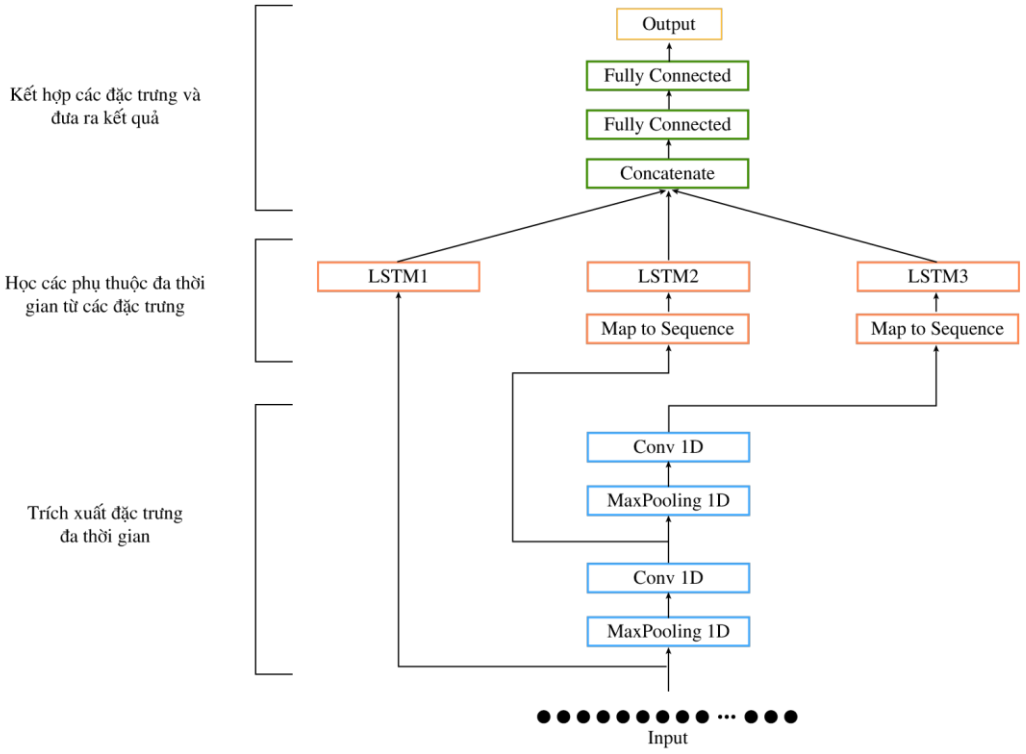
***) Kiến trúc mô hình CNN - LSTM kết hợp**

-Tổng quan

Mô hình CNN - LSTM kết hợp bao gồm 3 lớp thành phần, mỗi lớp đóng một vai trò cốt yếu cho mô hình tổng thể. Bằng việc sử dụng 3 kiến trúc mô hình khác nhau, mô hình đề xuất có khả năng trích xuất được những đặc trưng thời gian khác

nhau sau đó kết hợp những đặc trưng đó để đưa ra kết quả cuối cùng. Các tầng thành phần bao gồm:

- **Lớp Convolutional 1D**: Đóng vai trò trích xuất đặc trưng đa thời gian (Multiple time-scale learning) khác nhau qua số lượng lớp CNN khác nhau, kết hợp với các giá gốc hàng ngày để phản ánh những thay đổi của giá cổ phiếu trong thời gian ngắn hạn, trung hạn, dài hạn.
- **Lớp LSTM**: Đóng vai trò trong việc sử dụng các đặc trưng đa thời gian từ tầng CNN trước đó để trích xuất được sự phụ thuộc giữa các đặc trưng đa thời gian (Dependencies in Multiple Time-scale Features).
- **Lớp Fully connected**: Kết hợp các đặc trưng (Feature fusion) đã trích xuất



Hình 3.1: Mô hình CNN - LSTM kết hợp

được từ lớp LSTM trước đó rồi thông qua tầng fully connected để đưa ra dự đoán giá cổ phiếu cho ngày tiếp theo.

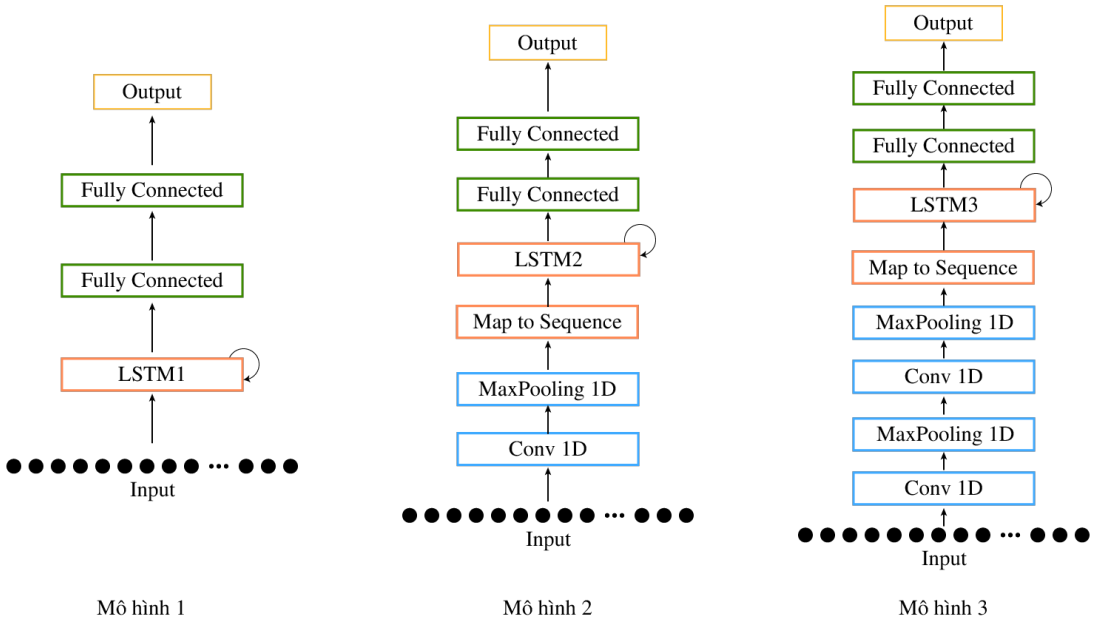
- Các lớp thành phần đề xuất

Như đã giới thiệu ở trên, mô hình CNN - LSTM kết hợp có 3 lớp thành phần chính, mỗi thành phần là một kiến trúc mạng khác nhau để trích xuất đặc trưng sau đó kết hợp với nhau tạo thành một mô hình để đưa ra dự đoán.

+ Lớp Convolutional 1D

Việc trích xuất thông tin qua ba thang thời gian:

Ngắn hạn (Mô hình 1): Bằng cách giữ nguyên giá trị giá cả hàng ngày để làm đầu vào cho lớp thành phần sau (LSTM), Mô hình 1 được coi là đặc trưng thời gian ngắn hạn, nó phản ánh những thay đổi cục bộ, ảnh hưởng quan trọng đến việc dự đoán.



Hình 3.2: Mô hình thành phần

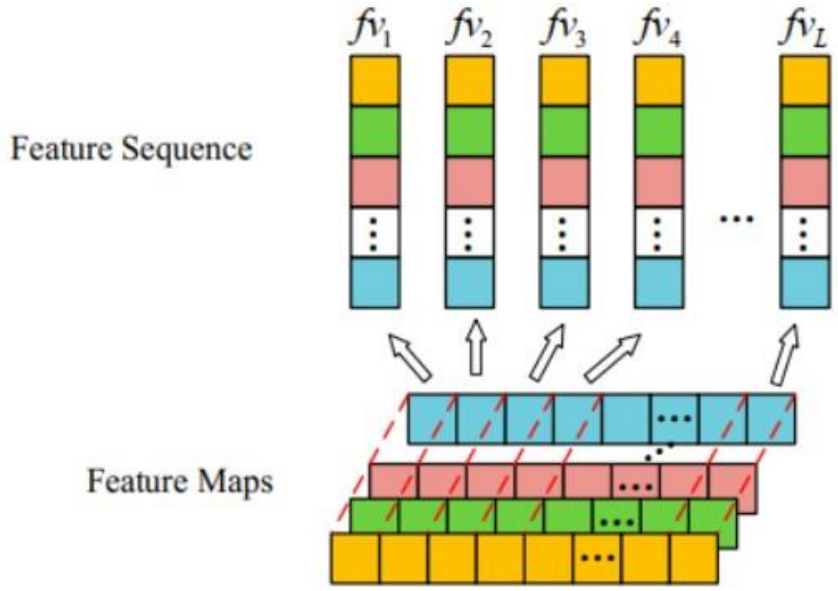
- Trung hạn (Mô hình 2): Từ dữ liệu giá cả hàng ngày, Mô hình 2 trích xuất đặc trưng qua một lớp CNN để là đầu ra cho lớp tiếp theo. Mô hình trích xuất các đặc trưng theo bước thời gian dài hơn so với Mô hình 1.
- Dài hạn (Mô hình 3): Tương tự như ở Mô hình 2, Mô hình 3 dùng thêm một lớp CNN để trích xuất đặc trưng, từ đó có thể trích xuất đặc trưng trên thang đo thời gian phức tạp hơn.

Bằng cách kết hợp các lớp CNN khác nhau này, mô hình 1, 2, 3 có thể phản ánh được sự thay đổi của giá cổ phiếu theo thời gian ngắn hạn, trung hạn, dài hạn.

+ Lớp LSTM

Mỗi mô hình ở trên đều sử dụng thêm một lớp LSTM để học sự phụ thuộc theo thời gian của các đặc trưng. Mô hình chuyển đổi các biểu đồ đặc trưng (feature map) được trích xuất từ các lớp CNN thành các chuỗi đặc trưng (feature sequence) phù hợp cho các lớp LSTM. Các biểu đồ đặc trưng đại diện cho các đặc trưng được học bởi CNN **hình 3.3**. Các màu khác nhau biểu thị cho biểu đồ đặc trưng được học bởi

các bộ lọc khác nhau. Các điểm trong biểu đồ đặc trưng được sắp xếp theo thứ tự thời gian từ trái qua phải.



Hình 3.3: Chuyển đổi biểu đồ đặc trưng thành chuỗi đặc trưng

Chuỗi đặc trưng được sử dụng làm dữ liệu đầu vào cho lớp LSTM. Vectơ đặc trưng trong chuỗi đặc trưng được biểu thị bởi fv_t, t tương ứng với thứ tự của nó trong chuỗi. Vectơ đặc trưng t là sự kết hợp của cột thứ t của tất cả các biểu đồ đặc trưng.

+ Lớp Fully connected

Lớp Concatenate được sử dụng để kết hợp các đầu ra từ ba mô hình LSTM. Các đầu ra của mô hình LSTM được kết hợp để tạo ra một vectơ đặc trưng chung. Sau đó, vector đặc trưng chung này đi qua một lớp fully connected để dự đoán giá cổ phiếu.

- Phương pháp đánh giá

Khi xây dựng một mô hình máy học, chúng ta cần một phép đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh khả năng của các mô hình. Có rất nhiều cách đánh giá một mô hình hồi quy. Tuỳ vào những bài toán khác nhau mà chúng ta sử dụng các phương pháp khác nhau.

+ MAPE

Sai số phần trăm tỉ lệ tuyệt đối trung bình (Mean Absolute Percentage Error - MAPE) là một trong các phương pháp đánh giá thường được sử dụng cho bài toán hồi quy. MAPE đo lường hiệu suất của mô hình bằng cách tính toán phần trăm lỗi tuyệt đối, với công thức được định nghĩa như trong (3.1)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \qquad (3.1)$$

Ưu điểm:

- Được biểu diễn dưới dạng phần trăm, không phụ thuộc vào kích thước mẫu có thể dùng để so sánh các dự đoán trên các mẫu có kích thước khác nhau.
- Dễ hiểu, dễ hình dung.

Nhược điểm:

- Nhận giá trị không xác định khi mẫu (y_i) có giá trị bằng 0, hoặc nhận giá trị rất lớn khi giá trị mẫu rất gần 0.

+ MAE

Sai số tuyệt đối trung bình (Mean Absolute Error - MAE) đo sai số trung bình trong một tập hợp các dự đoán mà không phải xem xét hướng của chúng. Nói cách khác MAE là giá trị tuyệt đối trung bình về độ lệch giữa dự đoán và dữ liệu thực tế, trong đó tất cả các độ lệch riêng lẻ có trọng số bằng nhau. Công thức của MAE được định nghĩa bởi (3.2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \qquad (3.2)$$

Ưu điểm:

- Vì lấy giá trị tuyệt đối, tất cả các sai số được tính theo cùng một thang đo tuyến tính, do đó không đặt nhiều trọng số cho các giá trị ngoại lai.
- Dễ dàng so sánh giữa các mẫu khác nhau với nhau.

Nhược điểm:

- Nếu mô hình bị ảnh nhiều bởi các giá trị ngoại lai, MAE sẽ không hiệu quả. Các lỗi lớn đến từ các giá trị ngoại lai sẽ được tính trọng số giống hệt các lỗi nhỏ hơn. Điều này có thể dẫn đến việc mô hình có thể thường sẽ dự đoán tốt, nhưng cũng thường xuyên đưa ra vài dự đoán rất kém.

+ RMSE

Sai số trung bình bình phương gốc (Root Mean Squared Error - RMSE) tính căn bậc hai của trung bình sai số bình phương giữa giá trị dự đoán và thực tế của mẫu.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.3)$$

Ưu điểm

RMSE đảm bảo cho mô hình hạn chế dự đoán các giá trị ngoại lai với sai số lớn, vì RMSE đặt trọng số lớn hơn cho các lỗi lớn do phần bình phương trong công thức tính toán.

Nhược điểm:

- Nếu mô hình đưa ra một dự đoán rất tệ, phần bình phương của hàm sẽ phóng đại sai số. Tuy nhiên, trong nhiều trường hợp thực tế, không quantâm nhiều đến những ngoại lệ này và hướng tới một mô hình toàn diện, hoạt động đủ tốt đối với số đông.

Điểm giống nhau của MAE và RMSE là cả hai đều là những giá trị không âm, và với những giá trị đánh giá thấp hơn sẽ tốt hơn. Sự khác biệt quan trọng của RMSE so với MAE là do các lỗi được bình phương trước khi lấy trung bình, nên RMSE cho trọng số tương đối cao với các lỗi lớn. Điều này có nghĩa là RMSE hữu ích hơn khi gặp các lỗi lớn.

+ MSE

Sai số trung bình bình phương (Mean Squared Error - MSE) tính trung bình bình phương sai số giữa giá trị dự đoán và thực tế của mẫu. Ngoài việc được sử dụng làm một phương pháp đánh giá, MSE còn thường được sử dụng để làm hàm mất mát cho các bài toán hồi quy.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3.4)$$

+ AMSE

RMSE, MAPE, MAE,... là các phương pháp đánh giá thường được dùng cho các bài toán hồi quy. Tuy nhiên những phương pháp này không xem xét đến "hướng" của kết quả dự đoán. Để giải quyết vấn đề đó, luận văn đề xuất một metric đánh giá mới để đo hiệu suất mô hình trong việc dự đoán xu hướng giá chứng khoán, cụ thể là phương pháp Sai số trung bình bình phương có điều chỉnh (Adjusted Mean Squared Error - AMSE) với công thức được xác định như sau:

$$AMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 * c(y_i, y_{i+1}, \hat{y}_i, \hat{y_{i+1}}) \quad (3.5)$$

Với:

$$c(y_i, y_{i+1}, \hat{y}_i, \hat{y_{i+1}}) = \begin{cases} 1, & \text{nếu } sign(y_{i+1} - y_i) = sign(\hat{y}_{i+1} - \hat{y}_i), \\ 2, & \text{nếu } sign(y_{i+1} - y_i) \neq sign(\hat{y}_{i+1} - \hat{y}_i) \end{cases} \quad (3.6)$$

Trong đó:

$$sign(x) = \begin{cases} 1, & \text{nếu } x \geq 0, \\ 0, & \text{nếu } x < 0 \end{cases} \quad (3.7)$$

AMSE có thể giải quyết vấn đề trên bằng cách nhân một hàm c vào MSE. Hàm c cho phép AMSE phạt gấp hai lần cho dự đoán sai "hướng" so với dự đoán cùng độ lệch nhưng đúng "hướng".

+ Accuracy

Thông thường bài toán dự đoán xu hướng chứng khoán có thể được xem như bài toán phân lớp (classification). Để có thể đánh giá chất lượng của dự đoán, ta sử dụng Accuracy như một phương pháp đánh giá. Accuracy thể hiện tỷ lệ dự mẫu dự đoán đúng trên tổng số lượng mẫu. Accuracy càng cao thể hiện hiệu suất dự đoán của mô hình càng tốt. Accuracy được tính bằng công thức sau:

$$Accuracy = \frac{N_{correct}}{N_{all}} \quad (3.8)$$

Với $N_{correct}$ đại diện cho số mẫu có cùng xu hướng với số mẫu thực tế, N_{all} đại diện cho toàn bộ mẫu:

3.2. Thử nghiệm mô hình học sâu

Chuẩn bị dữ liệu

Bộ dữ liệu chứng khoán Việt Nam (CKVN) gồm dữ liệu lịch sử của các mã chứng khoán thuộc VN50 index (50 cổ phiếu hàng đầu tại Việt Nam) từ lúc được phát hành đến thời điểm hiện tại. Gồm các trường dữ liệu: Giá mở cửa (Open), giá đóng cửa (Close), giá cao (High), giá thấp (Low), khối lượng (Volume), được lấy từ thư viện vnquant của tác giả Phạm Đình Khánh.

Dữ liệu gồm nhiều đặc trưng, các đặc trưng lại có độ lớn nhỏ khác nhau. Điều này tác động sự hiệu quả của các thuật toán ví dụ như quá trình hội tụ, thời gian thực hiện hay ảnh hưởng đến sự khái quát hóa của mô hình và độ chính xác của thuật

toán. Với các giá trị đầu vào lớn có thể dẫn đến sự bùng nổ tham số. Vì vậy người ta thường điều chỉnh để các đặc trưng đầu vào có cùng một khoảng tỉ lệ. Hai phương pháp thường dùng để điều chỉnh tỉ lệ dữ liệu thường được sử dụng là bình thường hóa dữ liệu (data normalization) và chuẩn hóa dữ liệu (data standardization).

Dữ liệu sau khi chuẩn hóa và xử lý được chia thành ba tập: tập huấn luyện (train dataset), tập kiểm định (validation dataset) và tập kiểm thử (test dataset). Tập huấn luyện được sử dụng để huấn luyện và tinh chỉnh tham số mô hình. Tập kiểm định nhằm lựa chọn các mô hình tốt cũng như ngăn chặn mô hình bị học tập quá mức (overfitting). Tập kiểm thử dùng để so sánh khả năng dự báo và kiểm tra khả năng khái quát hóa của mô hình. Dữ liệu được chia làm ba phần, trong đó 76.66% dữ liệu dùng cho tập huấn luyện, 17.54% dữ liệu dùng cho tập kiểm định và 5.8% dữ liệu dùng cho tập kiểm thử.

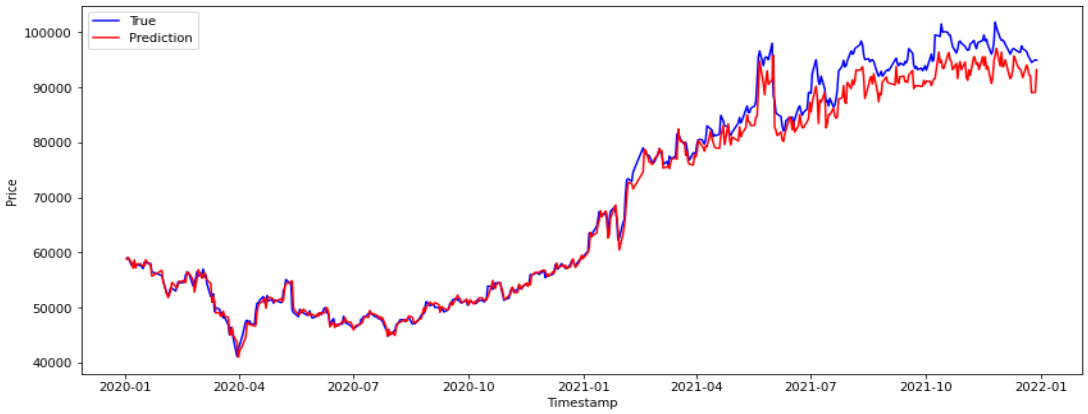
***Quá trình huấn luận mô hình thành phần và kết quả thực nghiệm**

Dựa trên kết quả thực nghiệm, tác giả nhận định rằng, nếu chỉ sử dụng dữ liệu của một mã cổ phiếu (Đơn) để dự đoán giá trị của ngày tiếp theo thì kết quả dự đoán của mô hình sẽ không thật sự tốt. Bởi các lý do sau:

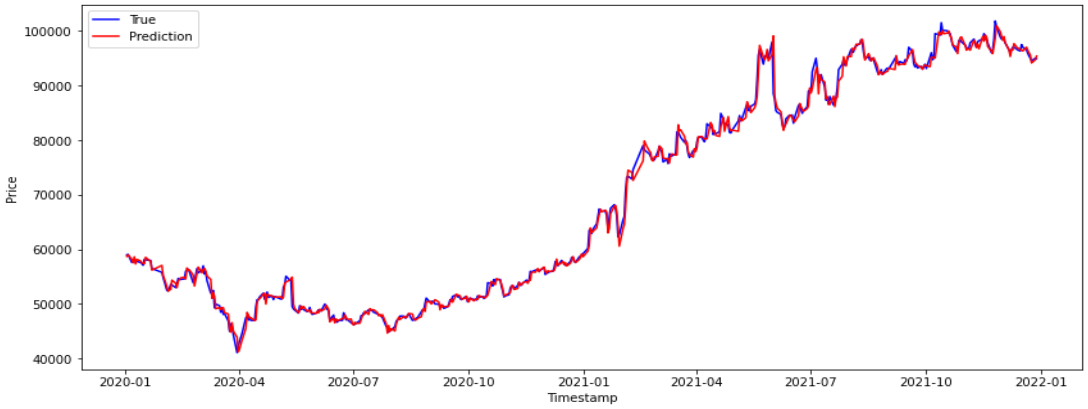
- Mô hình sẽ không dự đoán tốt ở hiện tại nếu trong dữ liệu quá khứ không có những mẫu (pattern) gần tương tự.
- Dữ liệu của một mã cổ phiếu không đủ nhiều để phản ánh được sự biến động của thị trường.

Để giải quyết được những vấn đề đó, tác giả đề xuất cách gộp những dữ liệu của các cổ phiếu trong VN50 index và dữ liệu từ thị trường chứng khoán nước ngoài

S&P100 để mô hình có thể học được những mẫu đa dạng góp phần giảm thiểu sai lệch với giá trị thực tế.



(a) Kết quả giá High trên tập dữ liệu đơn của FPT



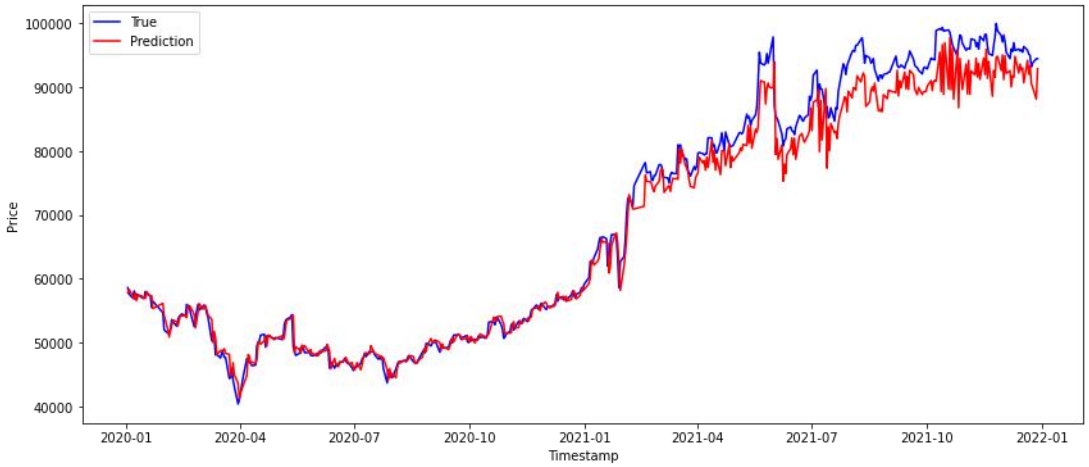
(b) Kết quả giá High trên tập dữ liệu gộp của FPT

Hình 3.4: So sánh kết quả giá High của tập dữ liệu đơn và gộp trên mã FPT

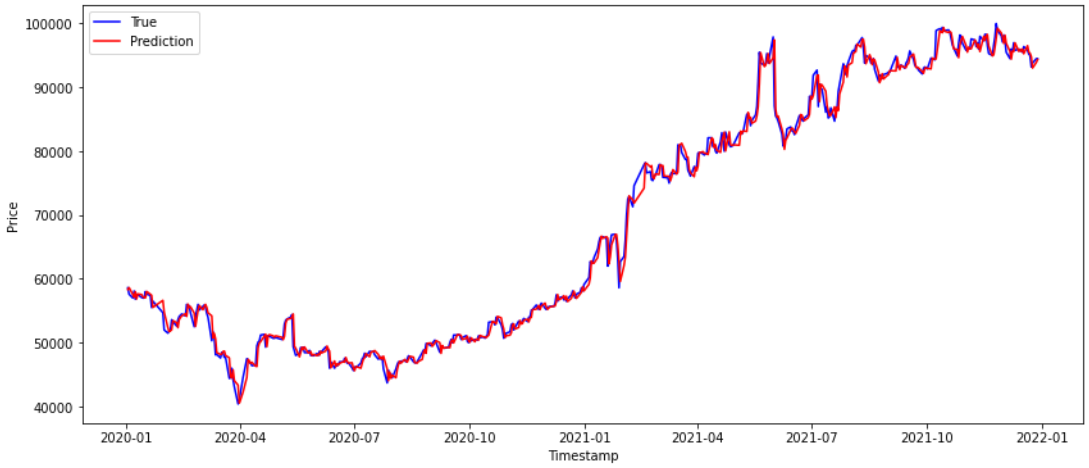
(a) Kết quả giá Low trên tập dữ liệu đơn của FPT

(b) Kết quả giá Low trên tập dữ liệu gộp của FPT

Hình 3.5: So sánh kết quả giá Low của tập dữ liệu đơn và gộp trên mã FPT



(a) Kết quả giá Close trên tập dữ liệu đơn của FPT



(b) Kết quả giá Close trên tập dữ liệu gộp của FPT

Hình 3.6: So sánh kết quả giá Close của tập dữ liệu đơn và gộp trên mã FPT

IV. KẾT LUẬN

Bài báo cáo đã nghiên cứu về mô hình học sâu và áp dụng cho bài toán dự đoán xu hướng giá chứng khoán. Đề xuất được phương pháp đánh giá mô hình, phương pháp xây dựng tập dữ liệu và phương pháp kết hợp phân tích kỹ thuật với mô hình học sâu để cho ra kết quả tốt cho bài toán.

Tuy nhiên thị trường chứng khoán luôn biến động, khó khăn để dự đoán, do nhiều lý do ảnh hưởng của nhiều các yếu tố bên ngoài, chẳng hạn bị ảnh hưởng bởi tình hình chính trị (đối với một số quốc gia cụ thể), tình hình kinh tế, nền kinh tế toàn cầu, lãi suất ngân hàng, các yếu tố tâm lý của nhà đầu tư.

Một số yếu tố cơ bản như chỉ số tài chính, chỉ số kỹ thuật, chỉ số kinh tế vĩ mô được chứng minh là yếu tố quan trọng ảnh hưởng đến việc tăng hoặc giảm giá cổ phiếu. Tuy nhiên, chưa có câu trả lời chính xác cho câu hỏi về các yếu tố ảnh hưởng nhiều nhất lên biến động giá cổ phiếu, do đó có rất nhiều nghiên cứu khác nhau, lựa chọn các yếu tố khác nhau làm biến đầu vào cho các mô hình dự đoán giá cổ phiếu khác nhau. Việc sử dụng các dữ liệu đầu vào cho cùng một mô hình cũng đưa ra kết quả khác nhau. Do đó, lựa chọn dữ liệu đầu vào cũng là một thách thức rất lớn để xây dựng một mô hình dự đoán giá chứng khoán hiệu quả.

TÀI LIỆU THAM KHẢO

[1] Lstm là gì? https://dominhhai.github.io/vi/2017/10/what-is-lstm/.

[2] Luật số 54/2019/qh14 của quốc hội: Luật chứng khoán. https:// vanban.chinhphu.vn/

[3] Neural network | deep learning cơ bản. https://nttuan8.com/bai-3-neural-network/.

[4] Trung tâm lưu ký chứng khoán việt nam tin tức. https://www.vsd..vn

[5] Đệm và sải bước — Đắm mình vào học sâu 0.14.4 documentation. https://d2l.aivivn.com/chapter_convolutional-neural-networks/