

ĐẠI HỌC QUỐC GIA TP. HCM  
**TRƯỜNG ĐẠI HỌC BÁCH KHOA**  
**ĐOÀN**

**BÁO CÁO CHUYÊN ĐỀ**

**Tên đề tài:**

**Estimating spatio-temporal Traffic Speed Based on LSTM for an urban network**

Nghiên cứu sinh:

Huỳnh Nam

Giảng viên hướng dẫn:

PGs.TS. Phạm Trần Vũ

TS Lê Thanh Vân

Thành phố Hồ Chí Minh – Tháng 01/2018

**Abstract:** In crowded cities, the space mean speed plays an important role in describing traffic state, knowing space mean speed in advance can help to alert the travelers and guide them to avoid bad traffic condition in some areas. With the popularity of GPS enabled devices, GPS data can be used to identify vehicle speed, as well as to predict the space mean speed in specific locations. In the paper, we propose a method for predicting space mean speed: Long Short Term Memory Recurrent Network with multiple features which is considered current speed would be result by sequential multiple decision factors previously. Besides, by statistical technique and comparative experiments is also proving that space mean speed is affected by decision factor in location, period of time, daily work time. To validate the proposed method, Global Position System (GPS) data was collected from historical floating car collecting data system in the urban of Ho Chi Minh City, Vietnam. The result obtained is promising to implement an Intelligent Transport System (ITS) application to determine space mean speed.

**Tóm tắt:** Với các thành phố lớn, thông tin về vận tốc luồng giao thông đóng một vai trò quan trọng trong việc mô tả trạng thái giao thông, với các thông tin phân tích cụ thể về vận tốc có thể giúp cảnh báo và hướng dẫn người tham gia giao thông tránh các khu vực có tình trạng giao thông ùn tắc. Thông qua việc ứng dụng thiết bị định vị toàn cầu GPS, việc thu thập dữ liệu GPS có thể xác định vận tốc của các phương tiện tham gia giao thông, với các dữ liệu thu thập đó được sử dụng như dữ liệu lịch sử để dự báo vận tốc luồng giao thông tại các khu vực cụ thể. Báo cáo này đề xuất phương pháp dự báo vận tốc luồng giao thông dựa trên mạng LSTM bằng cách xem vận tốc hiện tại là kết quả của chuỗi các vận tốc trước đó theo thời gian. Bên cạnh đó, bằng các kỹ thuật thống kê và thực nghiệm so sánh cũng đã chứng minh rằng vận tốc tại một khu vực cụ thể chịu ảnh hưởng của các yếu tố về khung giờ và ngày làm việc trong tuần, các yếu tố này sẽ được xem như là các tham số tham gia vào việc cải thiện kết quả dự báo. Để đánh giá mô hình, nghiên cứu xây dựng thực nghiệm dựa trên dữ liệu GPS được thu thập trong hệ thống giao thông tại thành phố Hồ Chí Minh. Kết quả đạt được đã khẳng định sự hiệu quả của mô hình trong việc xây dựng ứng dụng giao thông thông minh liên quan đến dự báo vận tốc luồng giao thông.

**Index Terms:** GPS, LSTM, RNN, ITS, floating car data, space mean speed

## MỤC LỤC

I.	GIỚI THIỆU ĐỀ TÀI .....	4
II.	NỘI DUNG TRIỂN KHAI .....	6
2.1	Các nghiên cứu liên quan .....	6
2.2	Giới thiệu về dữ liệu GPS .....	12
2.3	Thu thập và tiền xử lý dữ liệu .....	13
2.3	Tìm hiểu đặc trưng segment .....	15
2.4	Ước lượng vận tốc theo segment .....	17
2.4.1	Mô hình ước lượng LTSM.....	17
2.4.2	Ước lượng vận tốc segment với mô hình LSTM .....	21
2.4.2.1	Xây dựng mô hình .....	21
2.4.2.2	Mô hình LSTM .....	24
2.4.2.3	Mean Absolute Percentage Error .....	24
2.4.2.4	Mean Square Error .....	24
2.4.2.5	Đánh giá mô hình.....	25
III.	KẾT LUẬN & KIẾN NGHỊ .....	28
	TÀI LIỆU THAM KHẢO .....	29

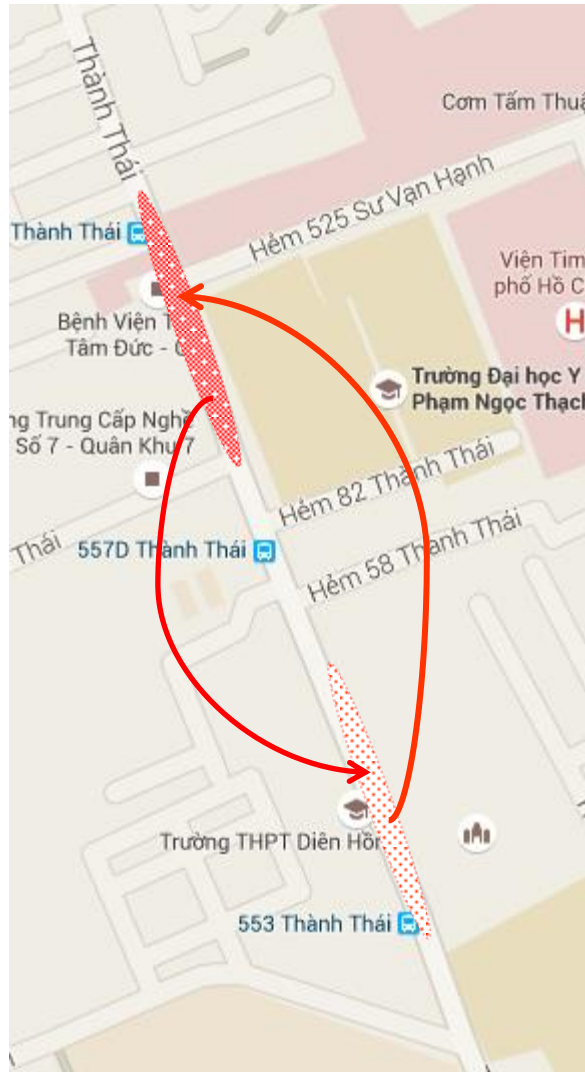
## I. GIỚI THIỆU ĐỀ TÀI

Các vấn đề về giao thông như giải quyết tình trạng ùn tắc, các vấn đề về điều tiết phương tiện... là những vấn đề mà các thành phố, đô thị lớn trên thế giới rất quan tâm, trong đó có Việt Nam. Bên cạnh các giải pháp đầu tư thiết kế cơ sở hạ tầng kỹ thuật giao thông như xây dựng đường, xây dựng metro, xây dựng các hệ thống giao thông công cộng... thì việc ứng dụng công nghệ thông tin vào xây dựng hệ thống giao thông thông minh (ITS) là một việc làm cấp thiết dựa trên các dữ liệu giao thông thu thập được nhằm mô hình hóa hiện trạng giao thông, từ đó giúp các nhà quản lý đưa ra những quyết định có cơ sở khoa học trong việc lập kế hoạch, điều phối... một cách hợp lý.

Hệ thống giao thông thông minh (Intelligent Transport System - ITS) là một hệ thống ứng dụng công nghệ cao về điện tử, tin học và viễn thông để điều hành và quản lý hệ thống giao thông vận tải. Đây là giải pháp ứng dụng kết hợp nhiều lĩnh vực công nghệ đang phát triển mạnh mẽ trong thời gian gần đây như hệ thống định vị toàn cầu (GPS), hệ thống thông tin địa lý (GIS), cơ sở dữ liệu và các giải pháp truyền thông, internet và di động [12].

Trong những năm gần đây, bài toán giao thông thông minh được quan tâm rất lớn do giúp giải quyết các vấn đề đang gặp phải như ách tắc luồng di chuyển, hiện trạng giao thông trên các tuyến đường, phân luồng. Trong đó, tìm hiểu đặc điểm giao thông và dự đoán được trạng thái đường dựa trên các đặc trưng đại diện như vận tốc, mật độ theo thời gian là một bài toán khó và cần được nghiên cứu nhằm góp phần hỗ trợ định tuyến, phân luồng giao thông [1]. Hiện nay, tại các đô thị lớn, hệ thống ITS đã được triển khai và áp dụng thành công đặc biệt là ở các nước đang phát triển như Mỹ, Nhật Bản... và các nước ở châu Âu. Đa số các nghiên cứu về giao thông trên thế giới tập trung vào nghiên cứu bản chất dòng giao thông, một định hướng khá hợp lý trong ngữ cảnh các nước tiên tiến, với phương tiện chính là ô tô.

Giao thông là quá trình mang tính ảnh hưởng tác động lẫn nhau theo thời gian và không gian. Điều này có thể thấy rõ trong việc những đoạn đường trong trạng thái ùn tắc sẽ ảnh hưởng đến trạng thái giao thông đến những đoạn đường lân cận.



Hình 1: Sự ảnh hưởng của trạng thái giao thông

Nếu như xem luồng giao thông như dòng chảy chất lỏng thì trong các điều kiện bình thường luồng giao thông trở nên ổn định với 3 thông số (flow, speed, density). Trong trường hợp tổng quát dòng giao thông sẽ trở nên bị tác động khi xuất hiện một lực cưỡng bức hạn chế (định lượng) hoặc làm thay đổi (hành vi – định tính) luồng di chuyển (có thể là sự kiện bất thường). Trong các thành phố lớn, mật độ đường sá trở nên dày đặc, đặc biệt đó là các giao lộ, đây là một trong những tác nhân chính gây nên sự xuất hiện các lực cưỡng bức xảy ra trên các luồng giao thông. Do đó, vấn đề xây dựng mô hình theo dõi và cảnh báo đặc điểm các luồng giao thông giữa các tuyến đường là một trong những vấn đề cần thiết.

Đối chiếu với tình hình nghiên cứu về ITS trên thế giới, Việt Nam chúng ta có nhiều điểm khác biệt. Ngoài việc thua kém về khoa học và công nghệ (KH&CN) cũng như cơ sở hạ

tăng, tình trạng giao thông của chúng ta cũng khác họ rất nhiều. Dòng giao thông ở Việt Nam là hỗn hợp với nhiều loại phương tiện cùng tham gia giao thông, luồng di chuyển của các phương tiện và hành vi tham gia giao thông cũng rất phức tạp và khó nắm bắt. Tuy nhiên, hiện nay các nguồn dữ liệu giao thông hiện nay của Việt Nam cũng khá dồi dào: GPS, video từ camera, dùng sóng, từ tính, sóng âm,... vấn đề đặt ra là làm thế nào để có thể tổng hợp và tích hợp được dữ liệu từ các nguồn này vào việc nghiên cứu xây dựng hệ thống ITS phù hợp với hiện trạng đặc thù giao thông tại Việt Nam. Từ đó, hệ thống sẽ phục vụ được những vấn đề cấp thiết như khắc phục thiệt hại kinh tế, lãng phí nhiên liệu, tiết kiệm thời gian... của những người tham gia giao thông và nhà quản lý. Ở đây, có thể đưa ra những bài toán nhỏ trong hệ thống ITS: mô phỏng hiện trạng giao thông, đánh giá ùn tắc giao thông, dự báo thời gian di chuyển của các phương tiện công cộng, điều phối luồng giao thông, tìm đường đi ngắn nhất, tìm đường đi dựa theo thói quen người dùng...

Trong nghiên cứu này, dữ liệu được sử dụng được thu thập chủ yếu từ các thiết bị GPS gắn trên các phương tiện tham gia giao thông tại thành phố Hồ Chí Minh. Các dữ liệu này được thu thập, xử lý và đồng bộ hóa theo thời gian thực tại phòng nghiên cứu HPCC trường đại học Bách Khoa, thành phố Hồ Chí Minh. Với ngữ cảnh giao thông tại thành phố Hồ Chí Minh, có nhiều vấn đề tác động đến sự chính xác của dữ liệu thu thập như chất lượng cơ sở hạ tầng, yếu tố thời tiết, yếu tố con người... Ngoài ra, hiện nay có rất ít những nghiên cứu về giao thông tại thành phố Hồ Chí Minh. Do đó, việc khai thác tri thức từ những thông tin về hành vi giao thông trên những dữ liệu thời gian thực này là một thách thức không hề đơn giản.

## **II. NỘI DUNG TRIỂN KHAI**

### **2.1 Các nghiên cứu liên quan**

Có một vài cách tiếp cận trong các nghiên cứu về giao thông thông minh, mà nổi bật là sử dụng mô hình toán học và học máy.

Với hướng tiếp cận sử dụng các phương pháp toán các đặc tính của các luồng giao thông chủ yếu dựa trên Traffic Theory, có rất nhiều thành phần đóng vai trò tham số đo lường trong lý thuyết giao thông như: số lượng xe trên một đơn vị thời gian, vận tốc... [1]. Trong lý thuyết ba pha, ba pha trong giao thông bao gồm trạng thái thông thoáng (free flow: F) và hai pha ùn tắc, trong hai pha ùn tắc được chia ra cụ thể là: trạng thái di chuyển chậm (synchronized flow: S) và trạng thái ùn tắc diện rộng (wide moving jam: J). 3 pha này biểu thị cho các đặc tính về

định lượng của hiện tượng ùn tắc giao thông [2]. Trong việc xây dựng mô hình bài toán giao thông thì các nghiên cứu dạng này sẽ xoay quanh các đại lượng luồng giao thông ( $q$ ) đặc trưng cho số lượng phương tiện trong một đơn vị thời gian, vận tốc ( $v$ ) và mật độ ( $k$ ) đặc trưng cho số lượng phương tiện trên một đơn vị diện tích. Tuy nhiên, các tác giả thường sử dụng một giả định về mối liên hệ tuyến tính giữa vận tốc và mật độ. Khi đó, thì theo phương trình Greenshields' ta có:

$$v = v_f - \left[ \frac{v_f}{k_j} \right] \cdot k$$

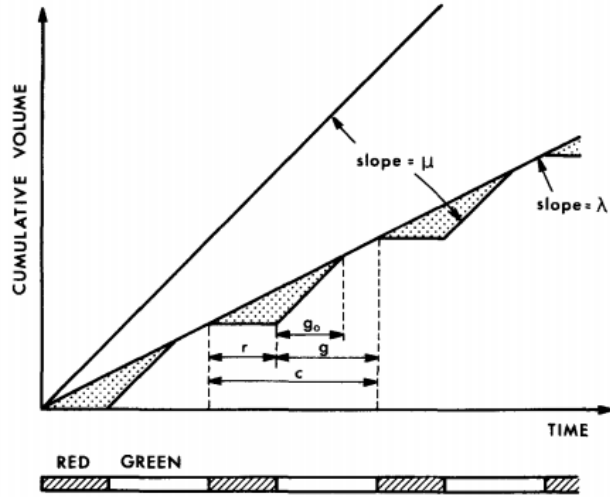
Trong đó,  $v$  là vận tốc đo được,  $v_f$ ,  $k_f$  là các hệ số điều chỉnh trong trạng thái phương tiện di chuyển thông thoáng (free flow),  $k$  được xem là mật độ đo được. Một giải pháp trong việc xấp xỉ các giá trị  $v_f$ ,  $k_f$  là sử dụng phương pháp hồi quy tuyến tính  $y = ax + b$ ; với  $y$  biểu thị cho  $k$  và  $x$  biểu thị cho  $v$ . Khi đó, ta có thể xấp xỉ hệ số  $a$  và  $b$  thông qua các giá trị đo được như sau:

$$\begin{aligned} b &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Từ đó, ta sẽ có phương trình thể hiện mối liên hệ giữa vận tốc và luồng giao thông như sau:

$$q = k_j \cdot v - \left[ \frac{k_j}{v_f} \right] v^2$$

Khi đó, khả năng giải phóng lưu lượng giao thông tại các giao lộ có đèn tín hiệu theo thời gian được xem như việc giải phóng hàng đợi. Vì khi hiện tượng ùn tắc xảy ra thì flow bằng 0 nên việc giải phóng hàng đợi liên quan đến giải phóng hàng đợi tại giao lộ.



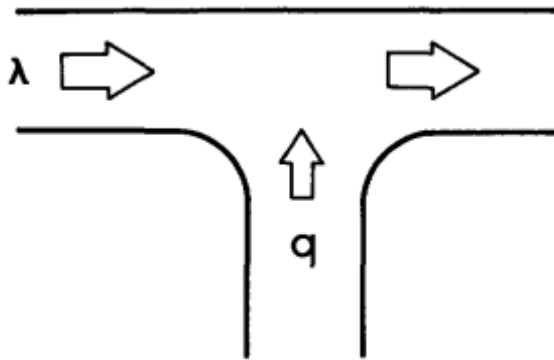
Hình 2: Mô hình thời gian chờ tại giao lộ có đèn tín hiệu

Nếu xem tỉ lệ luồng giao thông đến giao lộ là phân bố Poisson thì ta có mô hình thời gian chờ trung bình của phương tiện tại giao lộ như sau:

$$wI = wC + wQ = \frac{r^2}{2c(1-\rho)} + \frac{R^2}{2\lambda(1-R)}$$

Trong đó,  $wI$  là thời gian chờ trung bình của một phương tiện tại giao lộ,  $wC$  là thời gian chờ trung bình của một phương tiện phụ thuộc vào yếu tố đèn giao thông,  $wQ$  là thời gian chờ trung bình của một phương tiện phụ thuộc vào yếu tố (inter-arrival time) hàng đợi.

Tuy nhiên, với mô hình các giao lộ không có đèn tín hiệu, bài toán càng trở nên phức tạp trong việc tính toán đại lượng time-gap.



Hình 3: Mô hình thời gian chờ tại giao lộ không có đèn tín hiệu

Khi hiện tượng ùn tắc xảy ra việc giải phóng hàng đợi liên quan đến giải phóng time-gap. Nếu xem phân bố các time-gap là một phân bố theo quá trình Poisson thì ta có phương trình sau:



$$E[W] = E[N]E[T | T < t_{cr}]$$

$$= \frac{\int_0^{t_{cr}} f_T(t) dt}{\int_{t_{cr}}^{\infty} f_T(t) dt} \frac{\int_0^{t_{cr}} t f_T(t) dt}{\int_0^{t_{cr}} f_T(t) dt}$$

Trong đó,  $E[W]$  là giá trị kì vọng cho thời gian giải phóng giao thông tại một giao lộ.

Một số tác giả đã tổng quát hóa giả định tuyến tính giữa vận tốc và mật độ bằng cách thực hiện mở rộng mô hình Greenshields' [14]. Giả sử mối quan hệ giữa vận tốc và mật độ có thể xấp xỉ thuộc họ đa thức dựa trên mô hình Greenshields' mở rộng là Pipes' generalized model ta biểu diễn  $v$  và  $k$  theo công thức sau [15]:

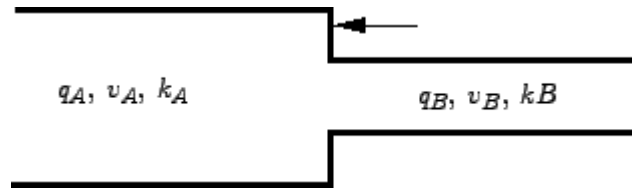
$$v = v_f \left[ 1 - \left( \frac{k}{k_j} \right)^n \right]$$

Trong đó,  $n$  là tham số xấp xỉ cho mô hình luồng giao thông.

Ngoài ra, do đặc điểm giao thông chịu nhiều ảnh hưởng từ các phương tiện, điều này gây tác động trực tiếp lên hành vi tham gia giao thông. Khi có một sự biến đổi hành vi di chuyển khác thường của một cá thể phương tiện sẽ dẫn đến sự ảnh hưởng đến các cá thể khác và lan rộng toàn bộ luồng giao thông [16]. Do đó, việc tính toán các thông số trên cả đoạn đường sẽ dẫn đến các kết quả không chính xác. Một vài nghiên cứu đề xuất áp dụng mô hình multi-regime cho việc xử lý các đoạn nhỏ giữa hai giao lộ dựa trên việc gom cụm các khoảng phụ thuộc theo không gian và thời gian sao cho có sự tương đồng nhất về vận tốc và khoảng cách giữa các phương tiện.

Ngoài ra, nếu xem  $x$  là luồng giao thông cung cấp cho giao lộ  $f$ , vậy thì  $y = f(x)$  sẽ là luồng giao thông đầu ra sau khi thông qua giao lộ. Như vậy ta thấy rằng khi  $y$  có giá trị bằng  $x$  tức là phương trình trên trở thành  $x = f(x)$  thì giao thông bình thường. Tuy nhiên khi  $y < x$  thì rõ ràng là hiện tượng ùn tắc hoặc tác nhân gây ùn tắc bắt đầu xuất hiện. Điều này đúng khi ta xem xét 2 ngữ cảnh, ngữ cảnh đầu tiên là khi hiện tượng ùn tắc xảy ra hiển nhiên  $y < x$ , ngữ cảnh thứ 2 xảy ra khi các phương tiện giao thông từ  $x$  rẽ qua nhánh khác, điều này dẫn đến tác nhân gây ra ùn tắc do xuất hiện lực ngăn cản luồng từ  $x$ . Trường hợp  $y > x$  cũng sẽ giống trường hợp  $y < x$  vì khi luồng từ  $y$  nhiều hơn  $x$  có nghĩa là xuất hiện nhiều nguồn cung cấp lưu

lượng từ các nhánh khác  $x$  vào  $y$  dẫn đến tạo ra lực cản làm suy yếu việc giải phóng luồng từ  $x$  như mô hình shockwave



Hình 3: Shockwave: Đặc tính dòng

Với những khả năng mạnh mẽ của lý thuyết toán trong việc xử lý các vấn đề về những ma trận có số chiều lớn và mô tả mối quan hệ không tuyến tính giữa các biến số, hướng tiếp cận tính toán thông minh (CI) [7] cũng được ưa chuộng hơn các mô hình thống kê. Ở đây có thể kể đến là mô hình ARIMA [9]. [5] Nghiên cứu này đề xuất mô hình ARMAX trong dự báo lưu lượng giao thông dựa trên dữ liệu hiện tại và quá khứ được tổng hợp thành hồ sơ về lưu lượng giao thông được biểu diễn trung bình cho mỗi ngày, dữ liệu thực nghiệm được lấy từ 6 trạm theo dõi lưu lượng phương tiện (VDS) được đặt tại tuyến đường. Từ các hồ sơ dữ liệu đó, nghiên cứu sẽ xây dựng các tham số hồi quy nhằm thể hiện mô tả tốt nhất cho dữ liệu đầu vào và kết quả dự đoán đạt được. Sau đó, kết quả đạt dự đoán sẽ được cải thiện bằng cách giải phương trình Bezout thông qua các bước trước đó trong mô hình dự báo.

Đối với giao thông tại Việt Nam khi còn nhiều hạn chế về mặt kỹ thuật cơ sở hạ tầng và chịu ảnh hưởng bởi nhiều yếu tố từ các thông tin không thu thập được (motobike) là rất nhiều khi so sánh với thông tin thu được (floating car data) nên việc đo lường để định lượng các thông số tham gia vào việc xây dựng mô hình giao thông như luồng, mật độ và vận tốc là rất khó khăn.

Tuy nhiên, với các mô hình toán trên khi áp dụng vào thực tế gặp phải hạn chế về việc thu thập dữ liệu liên quan để cung cấp các thông tin sau: dữ liệu về chu kỳ đèn tín hiệu, vận tốc trung bình trên tuyến đường kết nối hai giao lộ, tỉ lệ luồng đến và tỉ lệ luồng khởi hành tại các giao lộ.

Bên cạnh đó các công thức theo mô hình toán học thì chỉ phù hợp trên điều kiện lý tưởng khi dữ liệu tuân thủ đúng phân phối Poisson, và luồng dữ liệu phải phù hợp với mô hình tuyến tính. Chính vì vậy với các bài toán này, phương pháp khả thi để giải quyết là dùng máy học, khai phá dữ liệu và được đa số các nhóm nghiên cứu áp dụng, đồng thời chuyển đại lượng

đặc trưng cho tuyến đường về vận tốc vì tốc độ di chuyển trên đường phản ánh khá rõ tình trạng giao thông trên tuyến đường đó.

Trong các phương pháp được sử dụng để dự đoán về vận tốc, mạng nơ ron nhân tạo (ANN) là một phương pháp có độ chính xác cao dựa trên cơ chế hoạt động của não người [17,18,28]. ANN được ứng dụng để huấn luyện từ tập dữ liệu học từ đó sinh ra mô hình dự đoán cho các dữ liệu mới chưa biết nhãn. Liên quan đến các bài toán giao thông do tính thay đổi động của dữ liệu, mô hình RNN [27] (Resursive neural net work) có khả năng liên kết chuỗi vì thể hiện mối liên hệ chặt chẽ giữa input và output theo thời gian [19, 22-24]. Điều này khi sử dụng mạng nơ ron cổ điển sẽ không thấy được điều này do không không xét về tính thứ tự. Với mạng RNN, vấn đề này sẽ được giải quyết do RNN có sự “ghi nhớ” về vị trí.

Tuy nhiên khi áp dụng thực tế, RNN vẫn thể hiện hạn chế khi không có khả năng “nhớ” dài, tức là không liên kết được tốt giữa những trạng thái trước đó cho dự báo vị trí kế tiếp [20]. Đây là vấn đề về “vanishing” hay “exploding gradient” được đề cập trong nghiên cứu [13]. Tuy nhiên, hạn chế này sẽ được giải quyết trong mô hình Long-short term memory network [21,29].

Trong những năm gần đây, mạng học sâu (deep learning) đóng một vai trò quan trọng trong việc giải quyết các vấn đề liên quan đến hệ thống giao thông thông minh (ITS), trong đó có bài toán dự báo luồng giao thông và vận tốc giao thông. Mạng LSTM là một nhánh của mạng RNN thì được đánh giá cao trong lĩnh vực này trong những năm gần đây. Những nguyên nhân mà LSTM được quan tâm vì có những thách thức sau trong vấn đề dự báo vận tốc luồng giao thông: (1) sự phức tạp trong sự phụ thuộc giữa các tuyến đường gần nhau, (2) sự phụ thuộc của vận tốc theo thời gian là không tuyến tính và (3) trạng thái giao thông tại thời điểm hiện tại phụ thuộc vào các trạng thái trước đó theo một thời gian dài [14]. Mô hình LSTM có khả năng bắt được mối quan hệ không tuyến tính của sự biến đổi về luồng giao thông và xác định “optimal time lag” một cách tự động. Một nghiên cứu gần đây sử dụng mô hình LSTM với một lớp đầu vào, lớp ẩn LSTM và một lớp đầu ra để dự báo vận tốc lưu thông với dữ liệu được thu thập bởi các thiết bị “microwave detectors” tại Bắc Kinh. Sau đó, nghiên cứu sử dụng kết quả MAPE và MSE để kiểm chứng mô hình và kết quả đạt được đã thể hiện rằng mô hình LSTM có độ chính xác cao hơn và ổn định hơn các mô hình ANN truyền thống [8]. Bên cạnh đó mô hình LSTM cũng được sử dụng trong việc kết hợp DCNNs để sản sinh ra mô hình SRCNs. Mô hình tổng hợp có khả năng bắt được sự phụ thuộc về không gian thông qua mô hình DCNNs và thời gian thông qua mô hình SRCNs từ tập hợp dữ liệu các hình ảnh

tỉnh, tập dữ liệu này mô tả trạng thái giao thông giữa 278 tuyến đường kết nối tại Bắc Kinh [4]. Một trong những thách thức trong bài toán dự báo luồng giao thông là khai thác và kết hợp được mối quan hệ về thời gian, không gian như những đặc tính trong mô hình. Có thể giải quyết vấn đề đó bằng cách sử dụng mô hình Bi-directional để duyệt ngẫu nhiên trên đồ thị nhằm khai thác yếu tố về không gian và sử dụng kiến trúc “Encoder-Decoder” để trích xuất mối quan hệ về thời gian. Mô hình này được xem như là quá trình khuếch tán trên một đồ thị có hướng [14]. Khi các thông tin về không gian và thời gian được đưa vào mạng học sâu như Bidirectional Long Short Term Memory (BDLSTMs) thì cả mối tương quan giữa vận tốc tại các khu vực khác nhau và sự phụ thuộc về vận tốc về thời gian có thể được nhận diện. Ở đó, mỗi mạng “Deep stacked bidirectional and unidirectional LSTM” được đề xuất như một kiến trúc học sâu [7]. Bên cạnh đó, một nghiên cứu khác cũng chỉ ra rằng mạng học sâu (deep learning architecture) có khả năng giải quyết vấn đề phi tuyến trong việc mô tả sự chuyển đổi trạng thái giao thông dựa trên tác động của không gian và thời gian như: trạng thái thông thoáng, trạng thái di chuyển chậm, trạng thái ùn tắc và trạng thái ùn tắc [15].

## 2.2 Giới thiệu về dữ liệu GPS

Hệ thống định vị toàn cầu là hệ thống định vị của các vệ tinh nhân tạo trong không gian xác định thông tin vị trí và thời gian ở mọi điều kiện thời tiết, bất cứ nơi đâu trên trái đất hoặc gần trái đất. Để xác định vị trí của một đối tượng bất kỳ, đối tượng đó cần đảm bảo ít nhất 4 vệ tinh có thể nhìn thấy đối tượng đó. Ban đầu hệ thống định vị được bộ quốc phòng Mỹ phóng lên với 24 vệ tinh chủ yếu phục vụ cho quân sự, sau đó chính phủ Mỹ cho phép sử dụng hoàn toàn miễn phí cho các lĩnh vực dân sự. Về sau, các nước khác cũng xây dựng hệ thống định vị riêng của mình như Nga, liên minh Châu Âu, Trung Quốc, Ấn Độ.

Hệ thống định vị toàn cầu được ứng dụng trên nhiều lĩnh vực trên thế giới gồm quân sự, xây dựng, thương mại. Đặc biệt, trong lĩnh vực giao thông, hệ thống GPS được sử dụng rất rộng rãi trong giao thông hàng hải, hàng không, đường bộ. Để biết được thông tin định vị các ứng dụng cần có thiết bị GPS để thu tín hiệu GPS. Các thiết bị GPS sẽ nhận thông tin từ tín hiệu từ các vệ tinh (tín hiệu vô tuyến). Độ sai lệch về thời gian gửi nhận tín hiệu cho biết khoảng cách từ thiết bị GPS đến các vệ tinh, từ nhiều vệ tinh khác nhau thì thiết bị GPS bằng phép tính lượng giác để xác định vị trí của thiết bị.

Thông thường, chỉ cần nhận tín hiệu từ ba vệ tinh thì thiết bị GPS có thể xác định được vị trí trên trái đất bao gồm kinh độ, vĩ độ. Nếu nhận được ít nhất 4 vệ tinh thì thiết bị có thể xác

định thêm độ cao so với mặt đất. Khi đã xác định được vị trí, thiết bị còn có thể tính toán các thông tin khác như vận tốc di chuyển, hướng di chuyển và các thông tin khác. Độ chính xác của các thiết bị GPS chịu ảnh hưởng bởi địa hình, trạng thái của khí quyển, độ chính xác trung bình của các thiết bị GPS là trong vòng bán kính 15m.



Hình 4: Dữ liệu GPS

### 2.3 Thu thập và tiền xử lý dữ liệu

Dữ liệu về giao thông chủ yếu được thu thập từ các thiết bị định vị Global Positioning System (GPS) nên độ chính xác phụ thuộc rất nhiều vào thiết bị, đôi khi lại bị nhiễu bởi thời tiết hoặc bị che bởi các thiết bị cao tầng. Có rất nhiều trường hợp không thu thập được dữ liệu trong một khoảng thời gian dài hoặc dữ liệu vận tốc thu về có nhiều, một vài điểm ghi nhận vận tốc rất cao hơn 90km/h tại khu vực nội ô. Với những trường hợp thiết bị có số lượng dữ liệu ghi nhận thấp hoặc vận tốc sai lệch (nhiều được phát hiện bởi phương pháp thống kê), chúng tôi sẽ loại bỏ để giúp tập dữ liệu huấn luyện có chất lượng tốt.

Các thuộc tính chính của dữ liệu gồm có:

- DeviceId: mỗi xe được gắn một thiết bị GPS, mỗi thiết bị đó có mã nhất định. Giá trị này hiện tại được thể hiện bằng bảng số xe mà thiết bị được gắn để dễ dàng quản lý hơn
- Latitude: vị trí theo tọa độ vĩ độ hiện thời của xe buýt.
- Longitude: vị trí theo tọa độ kinh độ hiện thời của xe buýt.
- Speed: vận tốc tức thời của xe.
- Satellite: thông tin vệ tinh.
- Lock: trạng thái hoạt động của xe, nếu giá trị là 1 thì xe đang hoạt động; ngược lại giá trị là 0.
- Tracktime: thời điểm ghi nhận thông tin vị trí hiện thời của xe.

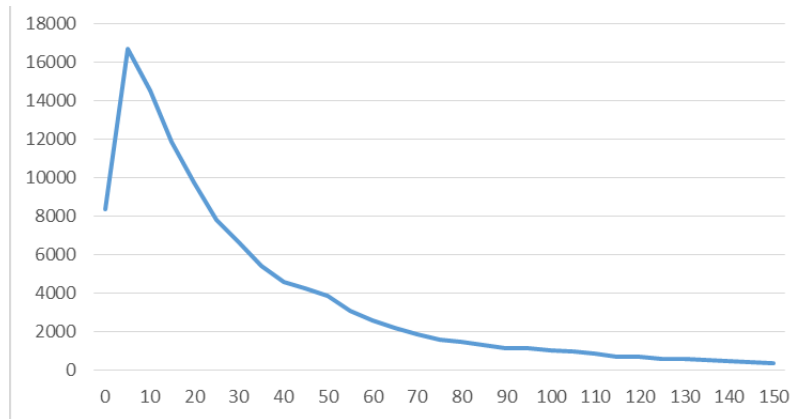
Trong mục này, các định nghĩa về các thuật ngữ sử dụng trong nghiên cứu sẽ được trình bày như GPS data point, frame, day period, segment, space mean speed data point:

- GPS data point: Được xem như là bộ ( $Lat, Lon, T$ ) và mỗi bộ này sẽ được ghi nhận tự động trong hệ thống từng 10 giây.
- Khung giờ (frame): Là khoảng thời gian 15 phút, do đó một ngày sẽ có 96 khung giờ. Khung giờ được đánh số từ 0 đến 95.
- Chu kì ngày: Chu kì ngày được đánh chỉ số từ 1 (thứ 2) đến 8 ( chủ nhật) . Chu kì ngày được chia làm hai nhóm: Ngày làm việc (1-2-3-4-5-6) và ngày cuối tuần (7-8).
- Segment: Là một đường thẳng không giao nhau được đánh dấu bởi segment\_id. Segment được mô tả bởi cặp điểm đầu ( $lat_s, lon_s$ ) và điểm cuối ( $lat_e, lon_e$ ). Dữ liệu này được thu thập từ Open Street Map (OSM). Một tuyến đường là tập hợp của nhiều segment và một segment có thể thuộc nhiều tuyến đường.
- Space mean speed data point: Rất khó để xác định vận tốc trung bình trên cả tuyến đường trong điều kiện môi trường không đồng nhất. Hệ thống sẽ tính vận tốc trung bình trên từng segment theo mỗi khung giờ. “Space mean speed data point” được định nghĩa như một bộ (segment\_id, day period, frame, speed).

Trong phạm vi của đề tài, toàn bộ dữ liệu GPS thu thập được sẽ được gom theo segment (được tính dựa trên latitude và longitude) theo khung thời gian của ngày. Dữ liệu phục vụ việc xây dựng mô hình và kiểm thử được thu thập trên 123,857 trong phạm vi 1600km<sup>2</sup> tại TP Hồ Chí Minh, ghi nhận liên tục từ 1/1/2017 đến 9/5/2017, gồm có 146,683,630 bản ghi. Tuy nhiên sau khi xử lý nhiễu do số lượng tín hiệu gửi về không đều trong một số segment, dữ liệu mẫu có thể sử dụng được gồm có 1,366,761 bản ghi. Mỗi bản ghi có các thuộc tính sau:

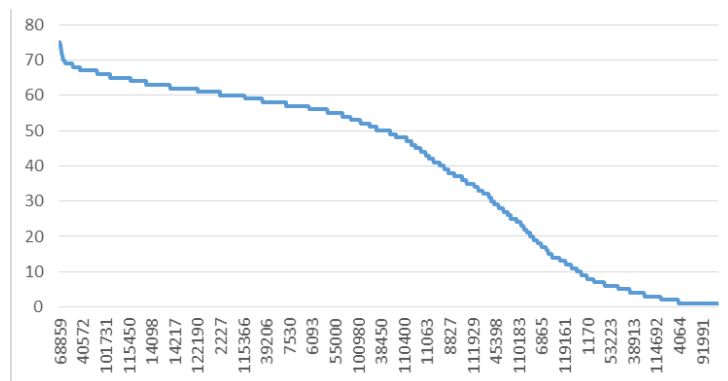
- segment\_id: Định danh của segment
- date: Ngày ghi nhận dữ liệu
- frame: khung giờ quan sát dữ liệu
- speed: vận tốc di chuyển được đo dựa trên thông số của thiết bị GPS, đơn vị km/h.

Hình 1 mô tả số tần suất của của các segment có độ dài từ 5 đến 150m. Trục X biểu thị độ dài của segment và trục Y thể hiện số lượng segment. Có tổng cộng 88% segment có độ dài từ 5 mét đến 150 mét.



Hình 5: Thống kê độ dài các segment

Hình 6 bên dưới mô tả tần suất số tín hiệu trung bình trên từng segment trong từng ngày. Có nhiều segment có số tín hiệu ghi nhận rất thấp, không đủ thông tin để rút trích hoặc xây dựng pattern cho segment, vì vậy những segment đó sẽ không được sử dụng.



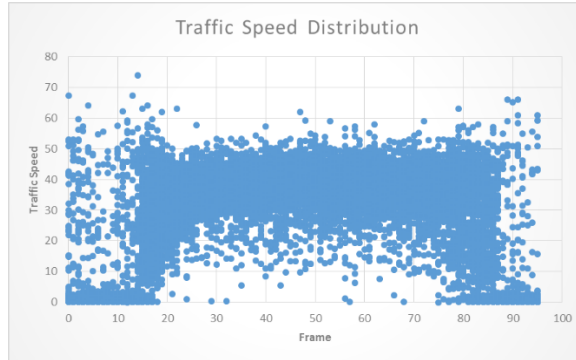
Hình 6: Tần suất tín hiệu trung bình trên segment

## 2.3 Tìm hiểu đặc trưng segment

Bản đồ nền Việt Nam được dựng dựa trên nền bản đồ mã nguồn mở Open Street Map, người dùng có thể bổ sung đường, cầu mới hoặc điều chỉnh các làn đường. Trong phạm vi đề tài, để thuận lợi cho việc quan sát và tính toán, nhóm nghiên cứu chia bản đồ làm 2,449,668 segment, trong đó vận tốc tại hai điểm bất kỳ trong segment được xem là tương tự nhau. Như vậy mỗi segment sẽ có các khung vận tốc gắn liền với khung thời gian. Từ trực quan quan sát cho thấy tồn tại mối liên hệ giữa vận tốc segment và khung giờ theo ngày, tuy nhiên chúng ta cần phải phân tích dữ liệu thu thập để có luận cứ rõ ràng hơn về mối quan hệ.

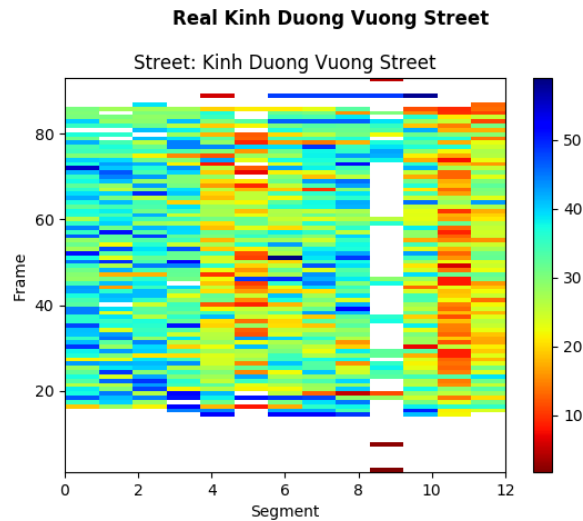
Tiến hành thống kê 32,679,313 mẫu tin từ 37 ngày. Sau đó, gom nhóm các mẫu tin theo [segment\_id, frame, work\_day]. Từ biểu đồ ta thấy sự phân bố về vận tốc phụ thuộc vào từng

khung thời gian cụ thể trong ngày. Trục hoành mô tả khung thời gian từ [0...95] và trục tung thể hiện vận tốc.



Hình 7: Phân bố vận tốc theo khung giờ

Sau khi sử dụng hai tiêu chí đơn giản chọn lọc bộ dữ liệu để kiểm tra mô hình, chúng tôi xây dựng bộ dữ liệu trên segment có mã số là 68859, segment này có độ dài là 125 mét. Hình bên dưới thể hiện biểu đồ phân bố vận tốc theo khung giờ của các segment lân cận segment 68859. Trục tung biểu diễn cho khung giờ trong ngày từ 0 đến 95, trục hoành biểu thị cho danh sách các segment tạo thành trên đường Kinh Dương Vương và dải màu thì cho biết giá trị vận tốc thay đổi từ 0 đến 65 km/h. Như vậy, dựa trên dải màu thay đổi theo khung giờ, đồ thị cho ta thấy có mối liên hệ về vận tốc giữa không gian các segment gần nhau theo thời gian.

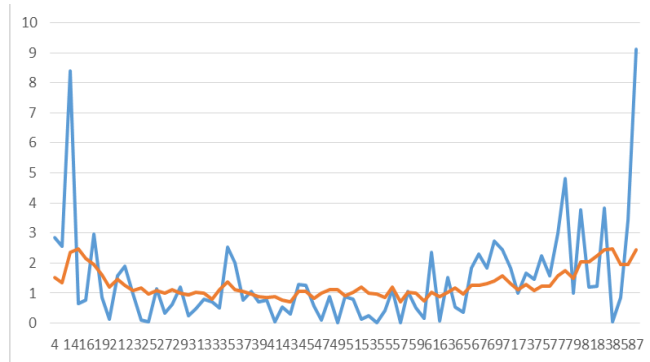


Hình 8: Phân bố vận tốc giữa các segment lân cận

Dựa trên công cụ thống kê trên tập dữ liệu mẫu của segment 68859, nghiên cứu phân tích để chọn lọc ra chuỗi dữ liệu thời gian của vận tốc có ý nghĩa trong thực tế và có tính ổn định theo khung giờ. Đầu tiên, đặt  $\Delta = |\text{Mean} - \text{Median}|$ , tham số này thể hiện tính ổn định của tập dữ liệu trong tập mẫu. Tiếp theo, giá trị standard error (SE) =  $\frac{S}{\sqrt{n}}$ , ở đó S là độ lệch chuẩn và n



là kích thước của tập mẫu dữ liệu. Giá trị này diễn tả sự phân tán trung bình giữa các cá thể xung quanh giá trị trung bình trong tập dữ liệu.



Hình 9: Mối quan hệ về vận tốc thời gian khung giờ

Hình 9 phía trên thể hiện mối quan hệ giữa hai giá trị SE và  $\Delta$  theo khung giờ trong ngày. Đường màu xanh biểu diễn cho đồ thị của  $\Delta$  và đường màu cam giải thích cho đồ thị SE. Trong khi đó, trục tung là giá trị của  $\Delta$  và SE, còn trục hoành là các giá trị của khung giờ từ 0 đến 95. Giá trị  $\Delta$  dao động khoảng từ 0.5 đến 2.43 còn giá trị của SE thì thuộc đoạn [0.7, 1.47]. Căn cứ vào các giá trị  $\Delta$  và SE cho biết sự biến thiên của vận tốc thì ổn định từ khung giờ 20 đến 80, tương đương từ 5h đến 20h hằng ngày. Từ đó dẫn đến kết luận tiêu chí về khung giờ và chu kì ngày đóng vai trò quan trọng trong việc xây dựng chuỗi dữ liệu về thời gian trong mô hình dự báo bởi mạng LSTM.

Mối liên hệ giữa vận tốc của segment, khung giờ và ngày trong tuần sẽ được sử dụng cho bài toán dự đoán vận tốc theo segment nhằm tăng độ chính xác của kết quả dự đoán. Điều này được thể hiện rõ trong phần thực nghiệm của phần nội dung kế tiếp của đề tài.

## 2.4 Ước lượng vận tốc theo segment

### 2.4.1 Mô hình ước lượng LTSM

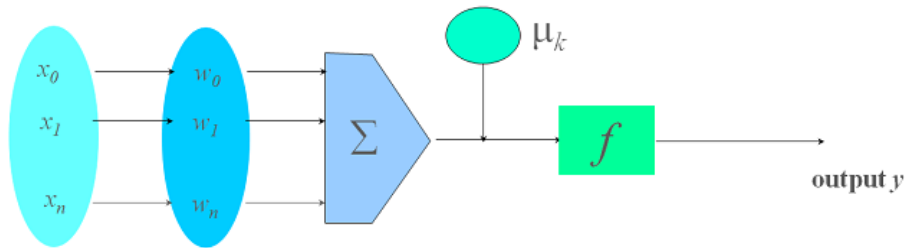
#### 2.4.1.1 Mô hình mạng nơ-ron

Mạng nơ-ron nhân tạo là một mô hình xử lý thông tin được lấy cảm hứng từ mạng nơ-ron sinh học. Yếu tố chính của mô hình này dựa trên cách thức giao tiếp và xử lý thông tin của các tế bào thần kinh. Mạng nơ-ron được thiết kế nhằm giải quyết một bài toán cụ thể như nhận dạng mẫu hoặc phân loại dữ liệu, thông qua một quá trình huấn luyện liên quan đến điều chỉnh các kết nối tồn tại giữa các tế bào thần kinh (các nút mạng) [17].

Mạng nơ-ron được phát triển vào năm 1943 bởi Warren Sturgis McCulloch và Walter Pits. Nhưng các công nghệ sẵn có tại thời điểm đó đã không cho phép họ đi xa hơn. Hiện nay,

với sự phát triển của siêu máy tính trong công nghệ, các lĩnh vực mạng lưới thần kinh được phát triển mạnh mẽ và có nhiều ứng dụng thực tế.

Một mạng nơ-ron có cấu trúc như sau:



Hình 10: Cấu trúc mạng nơ-ron

Trong đó:

- $x_i$ : giá trị đầu vào
- $w_i$ : trọng số
- $\Sigma$ : hàm tổng
- $\varphi$ : độ lệch
- $y$ : giá trị đầu ra
- $f$ : hàm kích hoạt

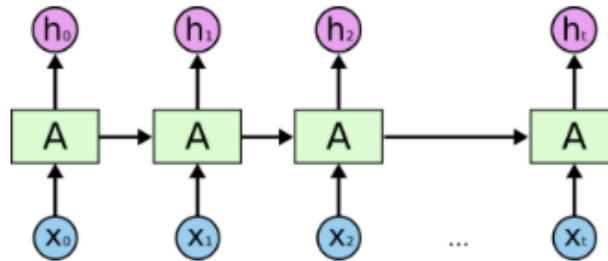
Quá trình huấn luyện được thực hiện với một tập dữ liệu đầu vào. Các giá trị trọng số của mạng nơ-ron được chọn ngẫu nhiên. Mỗi một dòng dữ liệu được học với mạng nơ-ron và sinh giá trị đầu ra được xem là giá trị dự đoán. Giá trị này sẽ được so sánh với giá trị thực tế để tính sai số. Trong mạng nơ-ron lan truyền ngược, các giá trị trọng số được tinh chỉnh từ các sai số cho đến khi nào đạt giá trị ngưỡng của sai số chấp nhận được.

Mặc dù mạng nơ-ron cho kết quả khá tốt tuy nhiên khi ứng dụng với bài toán giao thông, cụ thể là với việc dự đoán vận tốc của segment khi biết trước được một vài thông tin lịch sử trước đó thì đã bộc lộ sự hạn chế. Mạng nơ-ron thông thường không chú ý đến tính thứ tự và không có sự ghi nhớ. Từ khuyết điểm này, chúng tôi đề xuất dùng phương pháp học sâu Long-Short Term Memory sẽ được trình cụ thể hơn ở phần sau.

#### 2.4.1.2 Mô hình học sâu LSTM

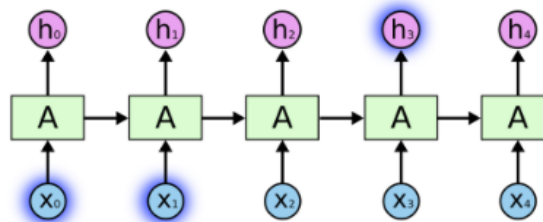
Những khuyết điểm của mô hình mạng nơ-ron truyền thống khuyến khích các nhà khoa học tiếp tục nghiên cứu và đề xuất liên tục các cải tiến. Một thành công mới được phát triển là mô hình học sâu RNN (Recurrent Neural Network) được đề xuất bởi Hopfield và các cộng sự vào những năm 1980. Một mô hình RNN được minh họa bao gồm mạng nơ-ron A với các dữ liệu đầu vào  $x_t$ , các kết quả đầu ra  $h_t$ . Thông tin được truyền liên tục qua các lần lặp kế tiếp.

Mô hình RNN có thể được hiểu bao gồm nhiều bản copy được liên kết với nhau, trong đó kết quả của bản copy này sẽ được truyền đến bản copy sau liền kề.



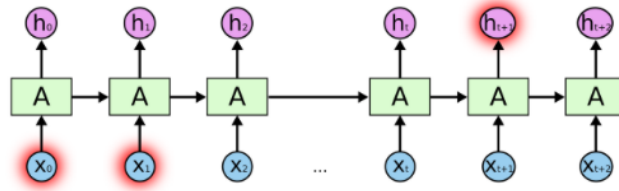
Hình 11: Mô hình mạng RNN

Một trong những đặc điểm nổi bật của mô hình RNN là khả năng liên kết các thông tin trước đó để dự đoán cho khả năng kết quả hiện tại. Ví dụ trong lĩnh vực xử lý ngôn ngữ tự nhiên, với yêu cầu dự đoán cho từ cuối cùng của một câu khi đã biết trước các từ in đậm “Lan là một cô bé đáng yêu có hai **bím tóc**”. Trong ngữ cảnh này, chúng ta có thể biết rất rõ ràng khả năng từ “tóc” là từ cuối cùng của câu mà không cần thêm ngữ cảnh gì tiếp theo. Trong những trường hợp này, mối liên hệ giữa những từ trước đó, đặc biệt từ “bím” với từ “tóc” khá rõ, và mô hình RNN làm rất tốt việc này.



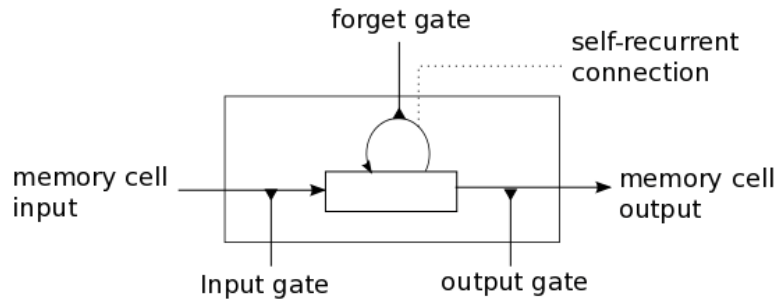
Hình 12: Mối quan hệ giữa output và input giữa các bản copy

Tuy nhiên trong một số trường hợp chúng ta cần phải có thêm ngữ cảnh để có thể có kết quả dự đoán tốt. Ví dụ như trong câu “Tôi lớn lên ở Việt Nam...Tôi nói trôi chảy tiếng Việt”. Những từ “Tôi nói trôi chảy tiếng” thể hiện từ sau đó sẽ có khả năng là tên của một ngôn ngữ nào đó, nhưng nếu chúng ta muốn chính xác là tiếng Việt, chúng ta phải xem xét thêm ý “Tôi lớn lên ở Việt Nam”. Như vậy trong trường hợp này khoảng cách thể hiện giữa thông tin trước đó và thông tin cần dự đoán tương đối dài. Với những thông tin yêu cầu có khoảng ghi nhớ dài như vậy thì mô hình RNN sẽ không thể liên kết và không giải quyết được.



Hình 13: Hạn chế của mạng RNN

Để giải quyết bài toán này, năm 1997 Hochreiter và các cộng sự đã đề xuất mô hình LSTM (Long short term memory) với một kiểu kiến trúc mới gọi là phần tử ghi nhớ (memory cell) gồm có 4 thành phần chính: cổng input, một nơ ron tự liên kết chính nó, một cổng “quên” và một cổng output. Nhiệm vụ chính của nơ-ron tự liên kết là với giá trị trọng số là 1.0, nó phải đảm bảo trạng thái của phần tử ghi nhớ là một hằng số cho dù có bất kì sự tác động từ bên ngoài nào. Các cổng có nhiệm vụ điều chỉnh các giao tiếp giữa các thành phần của phần tử ghi nhớ với môi trường. Cổng input có thể cho phép một tín hiệu đến thay đổi trạng thái của phần tử ghi nhớ hoặc block nó. Ngược lại, cổng output cho phép trạng thái có thể có tác động đến một nơ-ron khác hoặc ngăn chặn nó. Cổng “quên” điều chỉnh sự tự liên kết, cho phép phần tử ghi nhớ quên hay nhớ các trạng thái trước đó nếu cần thiết.



Hình 14: Thiết kế của phần tử ghi nhớ (memory cell)

Các phương trình dưới đây mô tả sự cập nhật các lớp của các phần tử ghi nhớ tại mỗi bước lặp thời gian  $t$ , trong đó

- $x_t$  là input vào của phần tử ghi nhớ tại  $t$ .
- $W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o$  và  $V_o$  là các ma trận trọng số,  $b_f, b_c$  và  $b_o$  là các vecto bias.

Các giá trị  $i_t$  của cổng input, và  $\tilde{C}_t$  giá trị ứng viên cho trạng thái của phần tử ghi nhớ tại thời điểm  $t$ :

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

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

Tính giá trị  $f_t$  của cổng “quên” tại thời điểm  $t$ :

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

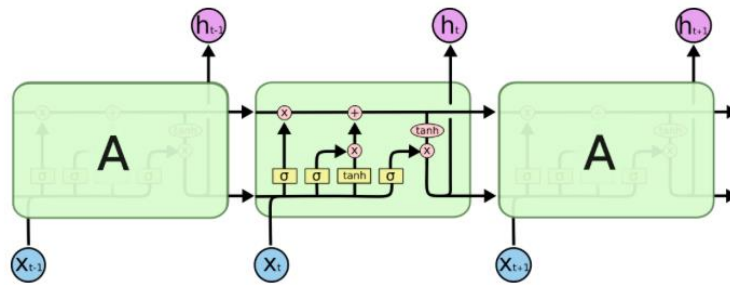
Trạng thái mới của phần tử ghi nhớ tại thời điểm  $t$  được tính như sau:

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

Các giá trị tại cổng output và các giá trị output được tính dựa trên hai công thức như sau:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_t + b_o)$$

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

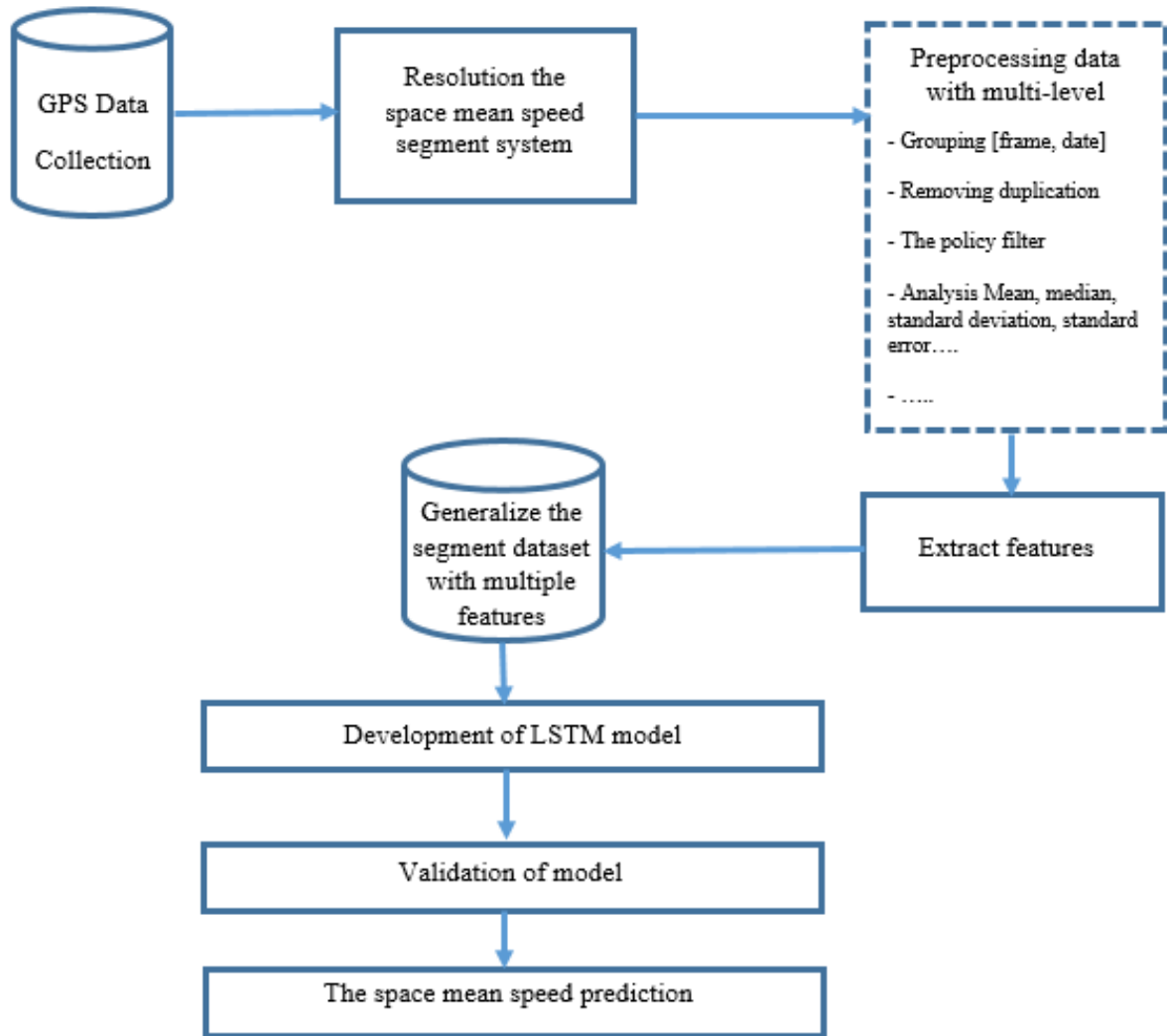


Hình 15: Minh họa mạng LSTM

## 2.4.2 Ước lượng vận tốc segment với mô hình LSTM

### 2.4.2.1 Xây dựng mô hình

Từ những ưu điểm của LSTM được phân tích, chúng tôi sử dụng phương pháp học sâu LSTM để dự đoán vận tốc của segment nhằm phân tích trạng thái của segment dựa trên trạng thái của những segment trước đó. Hình 13 mô tả các bước công việc được chuẩn bị và thực hiện cho bài toán dự đoán vận tốc của segment.



Hình 16: Minh họa luồng công việc thực thi ước lượng vận tốc với LSTM

Đầu tiên, dữ liệu GPS được thu thập và tiền xử lý dữ liệu để chọn lọc bộ dữ liệu tốt vì với các phương pháp máy học và học sâu, chất lượng dữ liệu có sự ảnh hưởng rất lớn đến mức độ hiệu quả của mô hình xây dựng được. Đặc biệt với dữ liệu thu thập từ hệ thống GPS, độ tin cậy của dữ liệu rất thấp, theo như tập dữ liệu chúng tôi thu thập được thì sau khi xử lý loại bỏ nhiễu, sai lệch thì chỉ còn sử dụng được khoảng 17%.

Sau khi đã chọn lọc được dữ liệu, các đặc trưng ảnh hưởng lên vận tốc sẽ được xem xét và trích xuất để xây dựng các vector dữ liệu đầu vào cho mô hình LSTM. Bước kế tiếp là xây dựng mô hình và kiểm thử, và cuối cùng là sử dụng mô hình để ước lượng và đo độ chính xác trên tập dữ liệu kiểm tra.

Bảng 1 sau đây mô tả các tập dữ liệu thực nghiệm có định dạng như sau:

1.Sequential type with previous steps = k
Dataset: $speed_0, speed_1, \dots, speed_{k-1} \rightarrow speed_k$ $speed_1, speed_2, \dots, speed_k \rightarrow speed_{k+1}$ ..... $speed_n, speed_{n+1}, \dots, speed_{n+k-1} \rightarrow speed_{n+k}$
2. Frame type with previous steps = k
Given $fs_k = (frame_k, speed_k)$ Dataset: $fs_0, fs_1, \dots, fs_{k-1} \rightarrow speed_k$ $fs_1, fs_2, \dots, fs_k \rightarrow speed_{k+1}$ ..... $fs_n, fs_{n+1}, \dots, fs_{n+k-1} \rightarrow speed_{n+k}$
1. Workday type with previous steps = k
Given $dfs_k = (day\_in\_week_k, frame_k, speed_k)$ Dataset: $dfs_0, dfs_1, \dots, dfs_{k-1} \rightarrow speed_k$ $dfs_1, dfs_2, \dots, dfs_k \rightarrow speed_{k+1}$ ..... $dfs_n, dfs_{n+1}, \dots, dfs_{n+k-1} \rightarrow speed_{n+k}$

Tập dữ liệu 1 chỉ quan tâm đến các giá trị vận tốc lịch sử, nghĩa là dùng 1 hoặc n các giá trị vận tốc trước để dự đoán cho một giá trị tại vị trí n+1.

Tập dữ liệu 2 bổ sung thêm yếu tố ảnh hưởng khung giờ.

Tập dữ liệu 3 dùng khung giờ, ngày trong tuần, và vận tốc của các khung thời gian trước của segment trước để dự đoán.

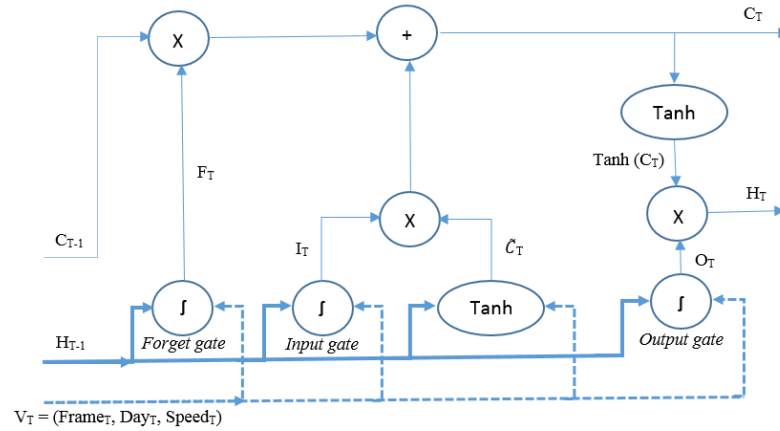
Cụ thể chúng tôi xây dựng ba tập dữ liệu thực nghiệm như sau:

- CASE2: [speed0, speed1, speed2, speed3, predictedSpeed]. Trong đó, [speed0, speed1, speed2, speed3]: là chuỗi vận tốc (tương ứng 4 input) dùng để dự báo vận tốc liền sau **predictedSpeed**
- CASE3: [(speed0, frame0), (speed1, frame1), (speed2, frame2), (speed3, frame3), predictedSpeed]. Trong đó, input đầu là 4, mỗi input được điều chỉnh như một vector gồm 2 thành phần (speed, frame)
- CASE4: [(daily0, frame0, speed0), (daily1, frame1, speed1), (daily2, frame2, speed2), (daily3, frame3, speed3), predictedSpeed]: Trong đó, input đầu là 4, mỗi input được điều chỉnh như một vector gồm 3 thành phần (daily, frame, speed).

### 2.4.2.2 Mô hình LSTM

Mô hình LSTM được xây dựng dựa trên các thông kê dữ liệu để đưa ra các tham số đầu vào cũng như cấu hình hệ thống. Thực nghiệm chạy trên thư viện Keras sử dụng nền Theano, với giá trị look\_back = 4 giá trị này thể hiện độ dài của chuỗi vận tốc liên quan có tác động đến vận tốc hiện tại dự báo, batch\_size = 1 giá trị này cho biết số mẫu trên mỗi lần cập nhật tổng hợp thông số cho mạng LSTM, epochs = 100 là giá trị thể hiện số lần huấn luyện mô hình.

Mạng LSTM được xây dựng có các số lượng tham số đầu vào tại thời điểm T được xem như một vector  $V_T = \{\text{frame}_T, \text{day}_T, \text{speed}_T\}$  chứa các tham số về thời gian tham gia để điều chỉnh khối nhớ của mạng LSTM trong quá trình huấn luyện, nhằm cải thiện độ chính xác của mô hình, mạng có một tầng ẩn có 13 khối nhớ và cuối cùng là có một tham số đầu ra.



Hình 17: Mô hình mạng LSTM

### 2.4.2.3 Mean Absolute Percentage Error

Nếu đặt giá trị Error được định nghĩa là trị tuyệt đối của hiệu số của cặp giá trị { Observed, Predicted } và giá trị Error được tính theo % là  $|O-P| / O$ . Khi giá trị Error càng tiến đến 0% thì độ chính xác của mô hình sẽ càng tăng. Giá trị sai số phần trăm trung bình (MAPE) được định nghĩa bởi công thức:

$$MAPE = \frac{100}{n} \sum_{t=1}^n \frac{|O_t - P_t|}{O_t}$$

Trong đó,  $O_t$  là giá trị quan sát được tại thời điểm t.

Công cụ thống kê này nhằm kiểm chứng hiệu quả của mô hình dựa trên giá trị dự báo và giá trị quan sát

### 2.4.2.4 Mean Square Error

Giá trị mean square error (MSE) thể hiện chất lượng của mô hình dự báo. Giá trị này luôn là số dương và càng tiến về 0 thì mô hình càng tốt.



Giả sử  $\hat{Y}$  là vector biểu diễn “n” giá trị dự báo và Y thể hiện cho vector của các giá trị thực thì MSE được cho bởi công thức:

$$MSE = \frac{1}{n} \sum_{i=0}^n (\hat{Y}_i - Y_i)^2$$

#### 2.4.2.5 Đánh giá mô hình

Trong phần đánh giá mô hình này, nghiên cứu so sánh sự hiệu quả của các mô hình dựa trên các tập dữ liệu chuỗi vận tốc, chuỗi dữ liệu đính kèm khung giờ và chu kỳ ngày trong tuần. Dữ liệu đưa vào mạng LSTM huấn luyện được chia theo tỉ lệ 80% cho huấn luyện và 20% cho kiểm chứng mô hình.

Hình 18 biểu diễn kết quả vận tốc dự báo giữa các mô hình. Trục hoành thể hiện thời gian được chia theo khung giờ và trục dọc thể hiện giá trị vận tốc. Đường màu xanh thể hiện cho vận tốc thực, các đường màu cam, màu xám và màu vàng biểu diễn cho giá trị vận tốc dự báo được tương ứng với từng mô hình đã đề cập bên trên.



Hình 18: Dự báo vận tốc dựa trên các đặc tính thời gian

Kết quả thu được bởi mô hình dự báo LSTM có đính kèm thông tin về ngày làm việc trong tuần từ khung giờ 25 đến 74, tức là tương đương từ 6h15 đến 18h30 thì gần đúng hơn với giá trị thực khi đem so sánh với các mô hình khác. Nguyên nhân là mô hình có khả năng nhận diện tốt chu kỳ làm việc vào các ngày trong tuần tại khu vực thành phố Hồ Chí Minh từ thứ Hai đến thứ Sáu vào khung giờ 6h đến 19h. Kết quả thu được cũng đã cho thấy sự hiệu quả của quá trình tiền xử lý dữ liệu và chọn lọc dữ liệu có ý nghĩa để đưa vào quá trình huấn luyện.

Bảng 2 bên dưới biểu diễn sự so sánh kết quả dự báo giữa các mô hình như ARIMA, ANN và LSTM, trong đó các đặc tính về thời gian được đính kèm. Các thước đo MAPE và MSE được dùng để đánh giá sự hiệu quả của các mô hình.

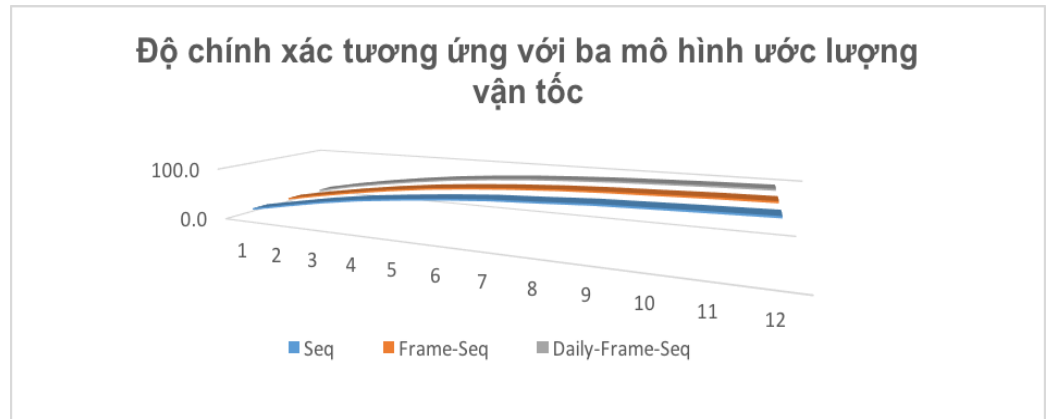
LSTM						
	Training			Validation		
	Seq	Frame	Day	Seq	Frame	Day
RMSE	6.5	6.5	6.5	5.8	5.8	5.8
MAPE	13.7	13.7	13.6	15.4	15.4	15.2
ANN						
	Training			Validation		
	Seg	Frame	Day	Seg	Frame	Day
RMSE	5.8	5.8	5.8	6.5	6.5	6.5
MAPE	13.8	13.8	13.9	15.9	15.9	16.1
ARIMA (5,1,0)						
	Training			Validation		
	Seg	Frame	Day	Seg	Frame	Day
RMSE		-	-	10.1	-	-
MAPE	1304.92	-	-		-	-

Bên cạnh đó, độ chính xác trung bình khi loại trừ các giá trị không thể dự báo là khoảng 85% với độ sai số là khoảng 7 km (được so sánh trên vận tốc lớn nhất là 65 km/h). Kết quả dự báo cho thấy với mô hình bổ sung các tham số về thời gian đã mang lại sự hiệu quả hơn trong việc dự báo vận tốc luồng giao thông, bởi vì các đặc trưng về thời gian đã góp phần vào việc cải thiện khả năng nhớ các bước trước đó. Nghiên cứu cũng so sánh kết quả giữa mô hình LSTM và mạng ANN truyền thống, kết quả mang đến mạng LSTM hiệu quả hơn và cho độ chính xác cao hơn mạng ANN. Ngoài ra, mạng LSTM có tính ổn định hơn mạng ANN khi so sánh giữa các mô hình, tuy nhiên các tham số về thời gian thì không đóng góp vào quá trình cải thiện chất lượng dự báo của mạng ANN truyền thống.

Bảng 3 trình bày kết quả thực nghiệm với 3 tập dữ liệu trên, độ chính xác được sử dụng để đánh giá mô hình.

<b>Khoảng sai số so với giá trị dự báo (km)</b>	<b>%Case2</b>	<b>%Case3</b>	<b>%Case4</b>
< 1	17.2	17.8	17.2
< 2	31.7	32.3	32.6
< 3	45.8	45.4	45.6
< 4	57.4	57.0	57.0
< 5	66.2	66.8	66.8
< 6	74.0	74.3	74.3
< 7	80.2	80.2	80.2
< 8	84.0	85.0	84.6

Hình bên dưới biểu thị cho độ chính xác của mô hình dự báo với sai số vận tốc từ 0 đến 8 km. Trục hoành biểu diễn cho khoảng sai số và trục tung diễn giải cho độ chính xác lần lượt của các mô hình Case2, Case3 và Case4 tương ứng với: đường màu xanh, đường màu cam và đường màu xám.



Hình 19: Độ chính xác theo khoảng sai số

Kết quả thực nghiệm cho thấy việc ước lượng gần đúng (sai số 1km/h) chỉ đạt 17.2% độ chính xác với tập dữ liệu Case2 và Case4, và đạt 17.8 với Case3 khi sử dụng bổ sung yếu tố khung giờ. Với sai số 8km/h, mô hình cho độ chính xác lần lượt với ba tập dữ liệu là 84%, 85% và 84.6%. Do segment chúng tôi sử dụng thực nghiệm không có thể hiện sự khác biệt rõ ràng về tình trạng ùn tắc nên khi bổ sung yếu tố khung giờ và ngày trong tuần vào mô hình huấn luyện, kết quả được cải thiện nhỏ chứ không nhiều, dao động từ 0-1%. Để chứng tỏ sự ảnh hưởng của các yếu tố này, đề tài cần chọn được một segmentID thể hiện khá rõ việc ùn tắc trong ngày vào các giờ cao điểm. Mặc dù còn hạn chế về việc chọn lựa dữ liệu cho phần thực nghiệm, tuy nhiên mô hình cũng thể hiện và khẳng định được việc ước lượng vận tốc của segment có phụ thuộc vào khung giờ và ngày trong tuần. Điều này được thể hiện qua kết quả đường màu xám và màu đỏ trong hình 14 có độ chính xác cao hơn đường màu xanh là đường chỉ dựa vào vận tốc để ước lượng.

Hiện nay nhằm phục vụ công việc chọn lọc dữ liệu tốt hơn, nhóm nghiên cứu đã tiến hành bổ sung thông tin khu vực vào từng segment, do số lượng dữ liệu segment lớn nên việc này đòi hỏi tốn khá nhiều thời gian và nhân lực dẫn đến ảnh hưởng phần nào kết quả thực nghiệm của đề tài.

### III. KẾT LUẬN & KIẾN NGHỊ

Đề tài đã phát triển tập hợp dữ liệu với nhiều đặc tính thời gian và lưu trữ theo không gian, thông qua các công cụ thống kê để chọn lọc dữ liệu khi đưa vào mô hình LSTM nhằm cải thiện khả năng dự báo. Kỹ thuật chính của nghiên cứu là xem vận tốc hiện tại chịu ảnh hưởng của một chuỗi các vận tốc trước đó. Nghiên cứu cũng so sánh các kết quả và đưa ra các tham số đóng góp vào trong quá trình huấn luyện mạng LSTM một cách trực tiếp. Kết quả đạt được hứa hẹn có thể áp dụng trong các ứng dụng APTS nhằm để dự báo vận tốc trong môi trường không đồng nhất về điều kiện giao thông như tại các thành phố đông dân.

Ngoài ra, đề tài tập trung vào việc phát hiện mối liên hệ giữa các segment dựa trên vận tốc nhằm dự đoán luồng di chuyển giao thông trong phạm vi thành phố Hồ Chí Minh. Do thời gian thực hiện đề tài có hạn và phụ thuộc lớn vào chất lượng dữ liệu thu thập, và một phần nguyên nhân khác do nguồn dữ liệu cung cấp không ổn định nên phần thực nghiệm chưa thể hiện được hết tất cả ưu điểm của đề tài. Nghiên cứu vẫn đang tiếp tục hoàn thiện bộ dữ liệu có chất lượng tốt và bổ sung thêm thông tin ngữ cảnh khu vực đường, phố cho dữ liệu để có thể kiểm chứng một cách rõ ràng hơn.

## TÀI LIỆU THAM KHẢO

1. Md Aftabuzzaman, Measuring Traffic Congestion – A Critical Review, Australia, 2007.
2. Wu, Y. J., F. Chen, C. T. Lu, B. Smith, and Y. Chen, Traffic Flow Prediction for Urban Network using Spatio-Temporal Random Effects Model, Transportation Research Board (TRB) 91st Annual Meeting, Washington DC, USA, 12-4179, 2012
3. Tingting Zhao, Yi Zhang , Bingyan Huang, RMT-Based Urban Traffic Cross-Correlation Analysis and ITS Application on Traffic Incident Impact Analyses, 2012
4. Short-Term Traffic Flow Forecasting: An Experimental Comparison of Time-Series Analysis and Supervised Learning, Tsinghua Science and Technology, Volume 17, Issue 1, 2012.
5. Lin Xu, Yang Yue Qingquan Li, Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data, **Procedia - Social and Behavioral Sciences**, Volume 96 , 2084--2095, 2013.
6. Xiao Zhang, Enrique Onieva, Asier Perallos, Eneko Osaba, Victor C.S. Lee Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction, Transportation Research Part C: Emerging Technologies , Volume 43, Part 1, 127—142, 2014.
7. Yi Liu, Xuesong Feng, Quan Wang, Hemeizi Zhang, Xinye Wang, Prediction of Urban Road Congestion Using a Bayesian Network Approach, *Procedia - Social and Behavioral Sciences*, Volume 138, 671-678, 2014.
8. Xiangjie Kong, Zhenzhen Xu, Guojiang Shen, Jinzhong Wang, Qiuyuan Yang, Benshi Zhang , Urban traffic congestion estimation and prediction based on floating car trajectory data, *Future Generation Computer Systems* , Volume 61 Pages 97-107, 2016.
9. Emilian Necula, Analyzing traffic pattern on street segments based on GPS data using R, *Transportation Research Procedia*, Volume 10, 276-285, 2015
10. Cheng-Jie Jin, Wei Wang, Rui Jiang, H.M. Zhang, Hao Wang, Mao-Bin Hu, Understanding the structure of hyper-congestion traffic from empirical and experimental evidences, *Transportation Research Part C: Emerging Technologies*, Volume 60, 324--338, 2015.
11. Guido Gentile , Using the General Link Transmission Model in a Dynamic Traffic Assignment to simulate congestion on urban networks, 2013, *Transportation Research Procedia*, Volume 5, 66--81, 2015.
12. Berk Anbaroglu, Benjamin Heydecker, Tao Cheng, Spatio-Temporal clustering for non-recurrent traffic congestion detection on urban road networks, *Transportation Research Part C: Emerging Technologies*, Volume 48, 47—65, 2014.
13. Suhas Prakash Kakliji, Mining GPS Data for Traffic Congestion Detection and Prediction, *International Journal of Science and Research*, Volume 4 Issue 9, September 2015.
14. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y. Dynamical Model of Traffic Congestion and Numerical Simulation, *Physical Review E*, 1994.
15. Adolf D. May, *Fundamentals of Traffic Flow*. Prentice - Hall, Inc. Englewood Cliff New Jersey 07632, second edition, 1990.

16. Yosef Sheffi, Urban Transportation Networks: Equilibrium Analysis With Mathematical Programming Methods, published by Prentice Hall, 1985.
17. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition, Elsevier Inc, 2012.
18. Alfréd Csikós, Zsolt János Viharos and Krisztián Balázs Kis, "Traffic speed prediction method for urban networks — an ANN approach," 2015 International Conference on Models and Technologies for Intelligent Transportation Systems, August 27, 2015.
19. Christopher Olah, "Understanding LSTM Networks," August 27, 2015
20. Razvan Pascanu, Tomas Mikolov and Yoshua Bengio, "On the difficulty of training recurrent neural networks," e-print arXiv:1211.5063, November 2012.
21. Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, pp. 1735-1780, November 15, 1997.
22. Yuhua Jia, Jianping Wu and Ming Xu, "Traffic Flow Prediction with Rainfall Impact Using a Deep Learning Method," Journal of Advanced Transportation, vol. 2017 (2017), 2017.
23. KrantiKumar, M.Parida and V.K.Katiyar, "Short Term Traffic Flow Prediction for a Non Urban Highway Using Artificial Neural Network," Procedia - Social and Behavioral Sciences, vol. 104, pp. 755-764, December 2, 2013.
24. Haiyang Yu, Zhihai Wu, Shuqin Wang, Yunpeng Wang and Xiaolei Ma, "Spatiotemporal Recurrent Convolutional Networks for Traffic Prediction in Transportation Networks," Sensors, vol. 17, May 4, 2017.
25. Mohammadhadi Fouladgar, Mostafa Parchami, Ramez Elmasri and Amir Ghaderi, "Scalable Deep Traffic Flow Neural Networks for Urban Traffic Congestion Prediction," arXiv:1703.01006, Mar 3, 2017
26. Xiaolei Ma, Zhimin Tao, Yinhai Wang Haiyang Yu and Yunpeng Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," Transportation Research Part C, vol. 54, pp. 187-197, May ,2015.
27. Rafal Jozefowicz, Wojciech Zaremba and Ilya Sutskever, "An empirical exploration of recurrent network architectures," Journal of Machine Learning Research, 2015.
28. Qing Ye, W. Y. Szeto and S. C. Wong, "Short-term traffic speed forecasting based on data recorded at irregular intervals," IEEE transactions on intelligent transportation systems, vol. 13, December, 2012.
29. Gers, F. A., Schmidhuber, J., & Cummins, F. , Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471, 2000.