TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN HỌC MÁY**

**Phân lớp văn bản bằng Deep Learning**

*Người hướng dẫn*: **TS BÙI THANH HÙNG**

*Người thực hiện*: **LÂM KIM PHÚ – 51600065**

**TRẦN THỊ MỸ NHI – 518H0543**

Nhóm **: 22**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN HỌC MÁY**

**Phân lớp văn bản bằng Deep Learning**

Người hướng dẫn: **TS BÙI THANH HÙNG**

Người thực hiện: **LÂM KIM PHÚ – 51600065**

**TRẦN THỊ MỸ NHI – 518H0543**

Nhóm **: 22**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Nhóm chúng em xin chân thành cảm ơn những lời chỉ tận tình và sâu sắc nhất của thầy Bùi Thanh Hùng. Qua bài báo cáo tiểu luận này, tụi em đúc kết từ những gì thầy dạy cho tụi em qua các buổi học, và qua những góp ý của thầy.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của chúng tôi và được sự hướng dẫn của TS Bùi Thanh Hùng;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 24 tháng 04 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Lâm Kin Phú*

*Trần Thị Mỹ Nhi*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc70284973)

[PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc70284974)

[MỤC LỤC 1](#_Toc70284975)

[DANH MỤC CÁC HÌNH VẼ 4](#_Toc70284976)

[DANH MỤC CÁC BẢNG 5](#_Toc70284977)

[PHẦN I: PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN 6](#_Toc70284978)

[1..1 Giới thiệu về bài toán 6](#_Toc70284979)

[1.2 Phân tích yêu cầu của bài toán 6](#_Toc70284980)

[PHẦN II: PHƯƠNG PHÁP GIẢI QUYẾT 7](#_Toc70284981)

[2.1 Các phương pháp giải quyết bài toán 7](#_Toc70284982)

[2.2 Phương pháp đề xuất giải quyết bài toán 12](#_Toc70284983)

[2.3 Phương pháp giải quyết bài toán 12](#_Toc70284984)

[2.3.1 Mô hình tổng quát 12](#_Toc70284985)

[2.3.2 Đặc trưng của mô hình đề xuất 13](#_Toc70284986)

[2.3.2.1. Cào dữ liệu 13](#_Toc70284987)

[2.3.2.2 Trích xuất đặc trưng 13](#_Toc70284988)

[2.3.2.4 Phương pháp huấn luyện 13](#_Toc70284989)

[PHẦN III: THỰC NGHIỆM 16](#_Toc70284990)

[3.1 Dữ liệu 16](#_Toc70284991)

[3.2 Xử lý dữ liệu 17](#_Toc70284992)

[3.3 Công nghệ sử dụng 17](#_Toc70284993)

[3.4 Cách đánh giá 18](#_Toc70284994)

[3.5 Kết quả đạt được 18](#_Toc70284995)

[PHẦN IV 19](#_Toc70284996)

[KẾT LUẬN 19](#_Toc70284997)

[4.1 Kết quả đạt được 19](#_Toc70284998)

[4.2 Hướng phát triển 19](#_Toc70284999)

[PHỤ LỤC 21](#_Toc70285000)

[TỰ ĐÁNH GIÁ 22](#_Toc70285001)

6

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

**CÁC CHỮ VIẾT TẮT**

LSTM - Long Short Term Memory networks

RNN - Recurrent Neural Network

CNN - Convolution Neural Network

DANH MỤC CÁC BẢNG

[Bảng 3.1 Ví dụ cho chèn bảng 1](#_Toc387689363)

[Bảng 3.2 Kết quả 1](#_Toc387689363)

PHẦN I: PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN

1..1 Giới thiệu về bài toán

**Text classification sử dụng LSTM Deep Learning.**

Xây dựng mô hình phân loại văn bản tin tức tiếng Việt cho trang báo điện tử VNExpress. Mỗi khi một bài báo được đăng, chương trình cần phải tự động xác định được bài báo đó nằm trong danh mục nào. Các danh mục gồm có: Kinh tế, Thể thao, Giáo dục, Sức khỏe, Du lịch, Pháp luật,…

1.2 Phân tích yêu cầu của bài toán

Yêu cầu bài toán bao gồm:

* Crawl dữ liệu tự động từ trang báo VNExpress
* Tiền xử lý dữ liệu
* Đưa ra dữ liệu được gắn nhãn tự động và độ chính xác( 80% trở lên)

PHẦN II: PHƯƠNG PHÁP GIẢI QUYẾT

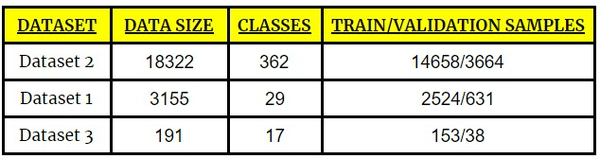
2.1 Các phương pháp giải quyết bài toán

Bài viết của tác giả Akshat Maheshwari( được dịch bởi Nguyễn Hoàng Nam) chia sẻ về kinh nghiệm của mình trong bài toán phân loại văn bản.

Có 3 phương pháp được đặt ra:

* Convolution Neural Network (CNN)
* Recurrent Neural Network (RNN)
* Hierarchical Attention Network (HAN)

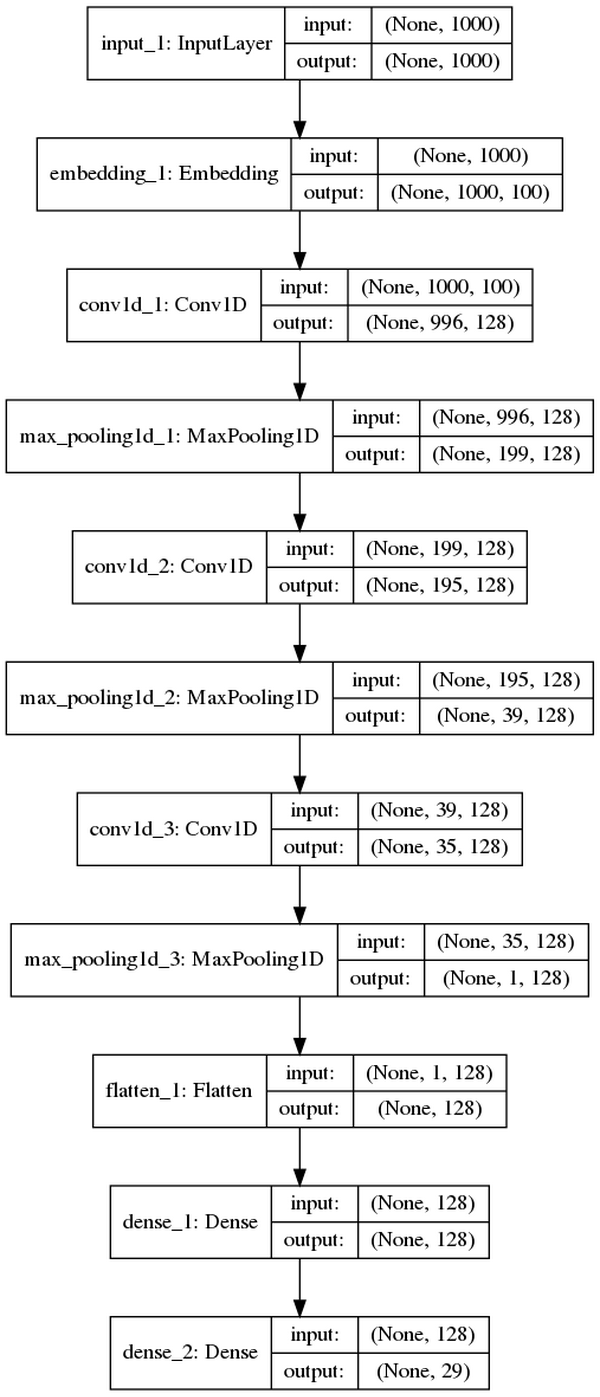
Dữ liệu được sử dụng như bảng dưới đây



**Phân loại văn bản sử dụng Convolution Neural Network (CNN):**

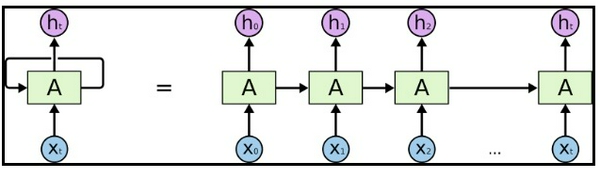
CNN là một class các mạng nơ-ron nhân tạo, feed-forward (nơi các kết nối giữa các nút không tạo thành một chu kỳ) và sử dụng một biến thể của multilayer perceptrons được thiết kế để yêu cầu tiền xử lý tối thiểu. Chúng được lấy cảm hứng từ vỏ não thị giác động vật.

Dưới đây là cấu trúc Model CNN:



**Phân loại văn bản sử dụng Recurrent Neural Network (RNN) - Mạng nơ-ron hồi quy**

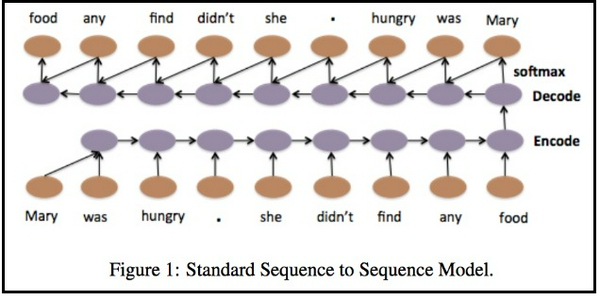
Đây là một mạng nhân tạo có các két nối giữa các node tạo thành một đồ thị theo dọc theo một chuỗi. Mô hình RNN dưới đây có thể trông khá phức tạp. RNN là một chuối các khối mạng nỏ-ron được liên kết với nhau như một chuỗi. Mối một khối sẽ chuyển một tin nhắn đến khối tiếp theo.



Vấn đề được giải quyết bằng cách sử dụng mạng nơ-ron hồi quy và attention dựa trên LSTM.

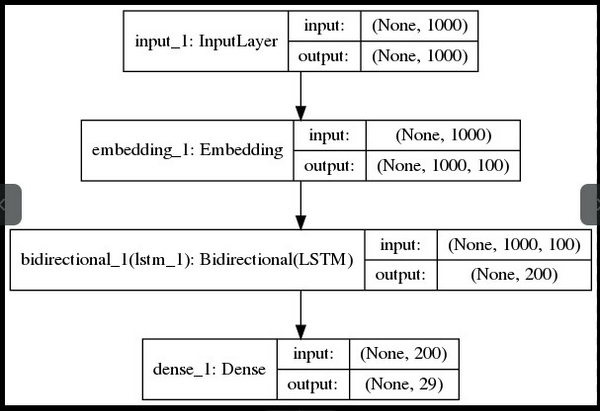
Bằng cách sử dụng bộ mã hóa LSTM, tôi dự định mã hóa tất cả thông tin của văn bản trước khi chạy mạng feed-forword để phân loại văn bản.

Dưới đây là hình ảnh từ [A Hierarchical Neural Autoencoder for Paragraphs and Documents](https://arxiv.org/pdf/1506.01057v2.pdf).



Sử dụng LSTM hai chiều và nối cả đầu ra cuối cùng của các đầu ra LSTM

Kiến trúc mạng RNN:

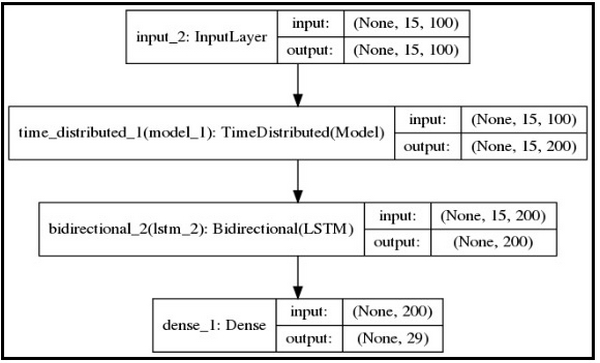


**Phân loại văn bản sử dụng HAN (Mạng chú ý đến phân cấp)**

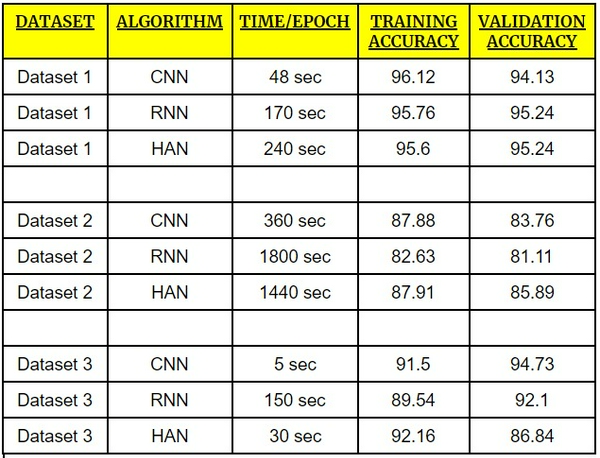
Tài liệu tham khảo từ bài báo [*Hierarchical Attention Networks for Document Classification*](https://www.cs.cmu.edu/~diyiy/docs/naacl16.pdf). Nó là một hướng dẫn tuyệt vời cho phân loại văn bản sử dụng HAN.

Xây dựng kiến trúc mạng LSTM, tôi phải xây dựng dữ liệu đầu vào là 3D thay vì 2D như ở 2 kiến trúc mạng bên trên.

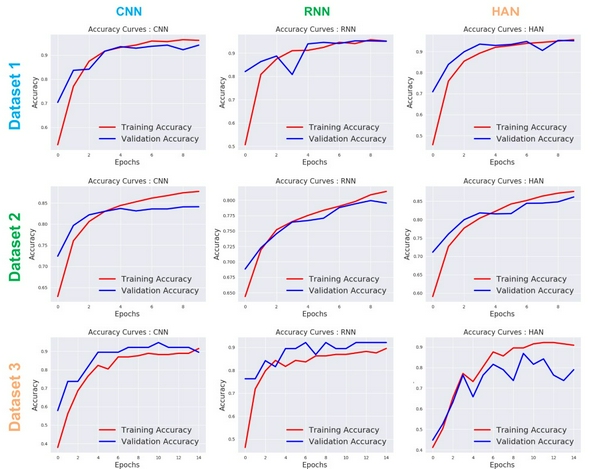
Dưới đây là kiến trúc HAN Model:



Kết quả nhận được:

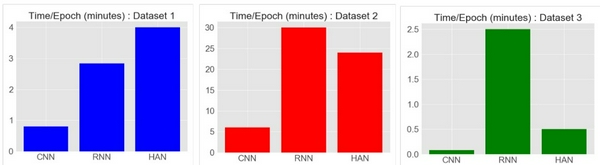


Plots hình ảnh của Accuracy và Loss



**Kết luận:**

* Dựa trên các hình ảnh plot ở trên, CNN đã đạt được độ chính xác khá tốt với tính nhất quán cao, còn RNN và HAN đã được độ chính xác cao nhưng không nhất quán trong toàn bộ tập dữ liệu.
* CNN vượt trội hơn về thời gian training, tuy nhiên HAN có thể hoạt động tốt hơn CNN và RNN nếu ta có một tập dữ liệu lớn.
* Ở bộ data1 và data2, HAN đã thực hiện rất tốt khi có một bộ training lớn, tuy nhiên ở data3 thì HAN đã không làm được tốt như vậy (bộ dữ liệu training rất ít)
* Khi bộ dữ liệu training ít (data3) thì CNN đã đạt đc độ chính xác xác thực tốt nhất.



Link bài viết: [***Báo cáo phân loại văn bản sử dụng CNN, RNN & HAN (Phần 1: CNN) - Noron.vn***](https://www.noron.vn/post/bao-cao-phan-loai-van-ban-su-dung-cnn-rnn--han-phan-1-cnn-1g62w4nxcq9), [***Báo cáo phân loại văn bản sử dụng CNN, RNN & HAN (Phần 2: RNN & HAN) - Noron.vn***](https://www.noron.vn/post/bao-cao-phan-loai-van-ban-su-dung-cnn-rnn--han-phan-2-rnn--han-1g62w4nxcr5?ref=97249675912513823)

2.2 Phương pháp đề xuất giải quyết bài toán

Phương pháp đề xuất: LSTM (RNN)

Đối với RNN, đó là một tập hợp các trọng số được áp dụng tạm thời (theo thời gian). Một đầu vào đi vào, và được nhân với trọng số. Các mạng lưu một trạng thái nội bộ và đưa ra một số loại đầu ra. Sau đó, phần dữ liệu tiếp theo xuất hiện và được nhân với trọng số. Tuy nhiên, trạng thái bên trong được tạo ra từ phần dữ liệu cuối cùng cũng xuất hiện và được nhân với một trọng số khác nhau. Chúng được thêm vào và đầu ra đến từ một kích hoạt được áp dụng cho tổng, nhân với trọng số khác. Trạng thái nội bộ được cập nhật và quá trình lặp lại RNN thực sự tốt cho xử lý ngôn ngữ tự nhiên. Có thể tưởng tượng rằng từ tiếp theo trong câu sẽ bị ảnh hưởng nhiều bởi những từ đi trước nó, vì vậy sẽ rất hợp lý khi đưa trạng thái nội bộ đó về phía trước và có một trọng số nhỏ có thể áp dụng cho bất kỳ đầu vào nào.

2.3 Phương pháp giải quyết bài toán

2.3.1 Mô hình tổng quát

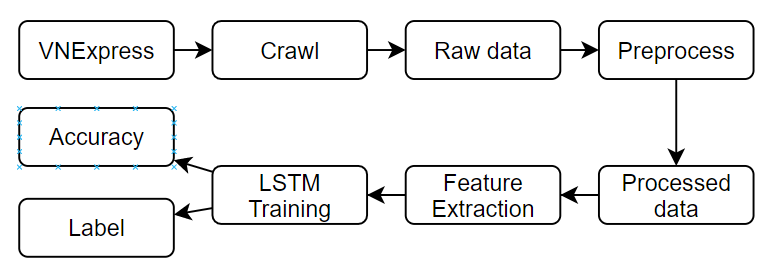
Mô hình tổng quát giải quyết bài toán Text Classification được trình bày theo sơ đồ dưới đây. Trong mô hình này có gồm có 4 phần chính như sau:

Phần 1: Thu thập dữ liệu tự động (crawl)

Phần 2: Tiền xử lý dữ liệu

Phần 3: Trích xuất thuộc tính bằng CountVectorizer

Phần 4: Huấn luyện bằng mô hình học sâu LSTM



2.3.2 Đặc trưng của mô hình đề xuất

2.3.2.1. Cào dữ liệu

* Dữ liệu được lấy từ trang https://vnexpress.net/ . Đầu tiên truy cập vào trang <https://vnexpress.net/> rồi lấy url của các topic (Ví dụ: <https://vnexpress.net/thoi-su>) từ navigation bar của trang. Sau đó vào trang từng của topic để lấy dữ liệu về bằng thẻ <p class="description">. Dữ liệu bao gồm “Topic” chứa topic của bài báo, “Title” chứa tiêu đề bài báo, “Description” chứa phần mô tả sơ lược bài báo.
* Ở phần “Description” có một số bài có gắn location stamp vì thế có một bước để loại bỏ location stamp. Nếu không bỏ thì phần “Description” chữ ở location stamp và phần text nó sẽ bị dính vào nhau.

2.3.2.2 Trích xuất đặc trưng

Trích xuất đặc trưng, vector hóa văn bản thành ma trận, trong đó mỗi chữ sẽ có số tương ứng.

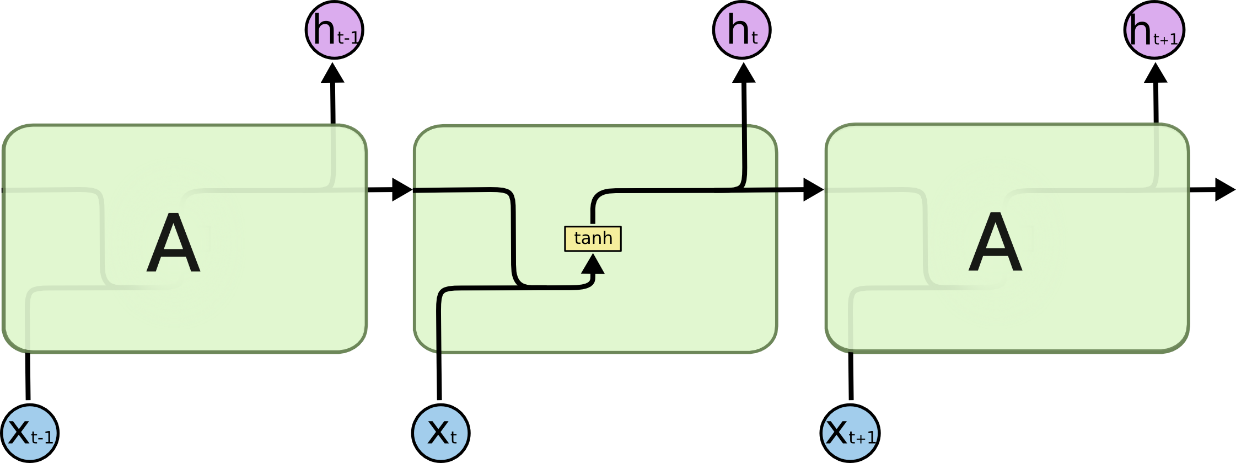
2.3.2.4 Phương pháp huấn luyện

**LSTM – Long Short Term Memory networks**

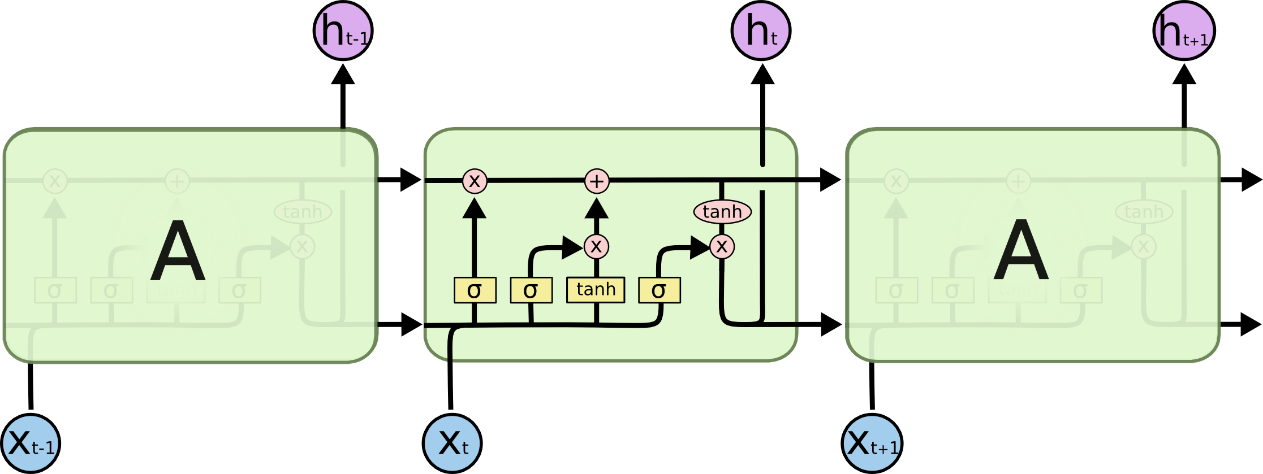
Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-dun này có cấu trúc rất đơn giản, thường là một tầng *tanh*.



LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.



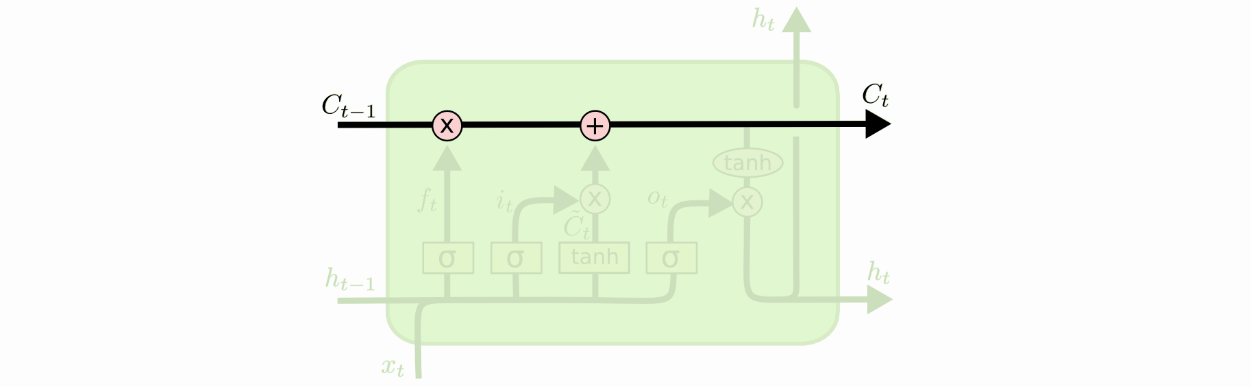
Ở sơ đồ trên, mỗi một đường mang một véc-tơ từ đầu ra của một nút tới đầu vào của một nút khác. Các hình trong màu hồng biểu diễn các phép toán như phép cộng véc-tơ chẳng hạn, còn các ô màu vàng được sử dụng để học trong các từng mạng nơ-ron. Các đường hợp nhau kí hiệu việc kết hợp, còn các đường rẽ nhánh ám chỉ nội dung của nó được sao chép và chuyển tới các nơi khác nhau.



**Ý tưởng cốt lõi của LSTM**

Chìa khóa của LSTM là trạng thái tế bào (cell state) - chính đường chạy thông ngang phía trên của sơ đồ hình vẽ.

Trạng thái tế bào là một dạng giống như băng truyền. Nó chạy xuyên suốt tất cả các mắt xích (các nút mạng) và chỉ tương tác tuyến tính đôi chút. Vì vậy mà các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay đổi.



LSTM có khả năng bỏ đi hoặc thêm vào các thông tin cần thiết cho trạng thái tế báo, chúng được điều chỉnh cẩn thận bởi các nhóm được gọi là cổng (gate).

Các cổng là nơi sàng lọc thông tin đi qua nó, chúng được kết hợp bởi một tầng mạng sigmoid và một phép nhân.

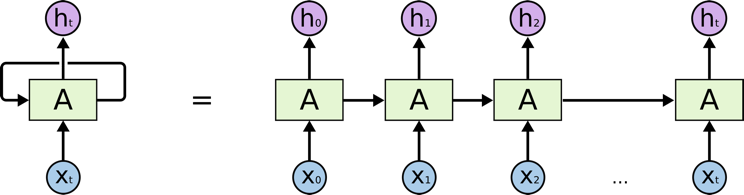


Tầng sigmoid sẽ cho đầu ra là một số trong khoản [0, 1], mô tả có bao nhiêu thông tin có thể được thông qua. Khi đầu ra là 0 thì có nghĩa là không cho thông tin nào qua cả, còn khi là 1 thì có nghĩa là cho tất cả các thông tin đi qua nó.

Một LSTM gồm có 3 cổng như vậy để duy trì và điều hành trạng thái của tế bào.

**Áp dụng LSTM vào bài toán**

Layer long-short term memory với số unit là 64, số lượng unit sẽ định hình kích thước ma trận ở layer Dense kế tiếp. Về bản chất layer long-short term memory là một cấu trúc mạng có khả năng tuần hoàn. Từ một đầu vào xt thông qua những biến đổi tuyến tính và phi tuyến, mạng sẽ tạo ra đầu ra là một nhân tố ẩn h(t) và nhân tố này tiếp tục được sử dụng là nguyên liệu đầu vào cho một cấu trúc mạng tiếp theo. Dạng trải dài tổng quát của LSTM có dạng như bên dưới:



Ở đây A tượng trưng cho một cấu trúc mạng tuần hoàn có khả năng thu nhận thông tin từ trạng thái trước đó.

PHẦN III: THỰC NGHIỆM

3.1 Dữ liệu

* Dữ liệu được lấy từ trang https://vnexpress.net/ . Đầu tiên truy cập vào trang <https://vnexpress.net/> rồi lấy url của các topic (Ví dụ: <https://vnexpress.net/thoi-su>) từ navigation bar của trang. Sau đó vào trang từng của topic để lấy dữ liệu về bằng thẻ <p class="description">. Dữ liệu bao gồm “Topic” chứa topic của bài báo, “Title” chứa tiêu đề bài báo, “Description” chứa phần mô tả sơ lược bài báo.
* Ở phần “Description” có một số bài có gắn location stamp vì thế có một bước để loại bỏ location stamp. Nếu không bỏ thì phần “Description” chữ ở location stamp và phần text nó sẽ bị dính vào nhau.
* Dữ liệu được lấy về có dung lượng khoảng 100KB và được đặt tên là “raw\_data.csv”, bao gồm 3 cột “Topic”, “Title”, “Description”, và khoảng 430 – 460 dòng dữ liệu tùy theo thời điểm crawl.

3.2 Xử lý dữ liệu

Tiền xử lý dữ liệu là một bước rất quan trọng trong việc giải quyết bất kỳ vấn đề nào trong lĩnh vực Học Máy. Hầu hết các bộ dữ liệu được sử dụng trong các vấn đề liên quan đến Học Máy cần được xử lý, làm sạch và biến đổi trước khi một thuật toán Học Máy có thể được huấn luyện trên những bộ dữ liệu này.

Các bước xử lý dữ liệu:

* Bước 1: lowercase tất cả dữ liệu.
* Bước 2: loại bỏ dấu câu và các kí tự đặc biệt trong dữ liệu.
* Bước 3: tokenize dữ liệu theo thư viện underthesea.
* Bước 4: loại bỏ stopword.

Dữ liệu qua xử lý được lưu vào file “preprocessing.csv” trong thư mục “data”.

3.3 Công nghệ sử dụng

Ngôn ngữ lập trình: Python

Các thư viện sử dụng giải quyết bài toán:

* beautifulsoup4 4.9.3
* Keras 2.4.3
* matplotlib 3.4.1
* nltk 3.5
* numpy 1.20.1
* pandas 1.2.4
* regex 2021.4.4
* requests 2.22.0
* tensorflow 2.4.1
* underthesea 1.3.1

Các công cụ sử dụng giải quyết bài toán: Pycharm, Google Colaboratory

3.4 Cách đánh giá

Cách đánh giá bằng độ đo gì, độ đo đó được tính bằng công thức nào?

3.5 Kết quả đạt được

Trình bày các tham số cụ thể hóa các thực nghiệm: ví dụ như số epoch, số tầng, số lớp, tham số x, …

* Số epoch: 5
* Batch size: 64
* Độ chính xác: khoảng 70%
* Bài báo được gắn nhãn gần như chính xác
* Code cho model Sequental LSTM

|  |
| --- |
| model = Sequential()  model.add(Embedding(MAX\_NB\_WORDS, EMBEDDING\_DIM, input\_length=X.shape[1]))  model.add(SpatialDropout1D(0.2))  model.add(LSTM(100, dropout=0.2, recurrent\_dropout=0.2))  model.add(Dense(6, activation='softmax'))  model.compile(loss='categorical\_crossentropy', optimizer='adam', metrics=['accuracy'])  print(model.summary()) |

PHẦN IV

KẾT LUẬN

4.1 Kết quả đạt được

Kết quả đạt được:

* Độ chính xác khoảng 70% (thấp hơn mục tiêu đề ra).
* Bài báo được gắn nhãn gần như chính xác.
* Hiểu được cách hoạt động của thuật toán.

Hạn chế:

* Dữ liệu chuẩn bị tương đối ít nên độ chính xác chưa được cao.
* Bài báo còn bị gắn nhãn sai.

4.2 Hướng phát triển

Hướng phát triển trong tương lai: độ chính xác của thuật toán đạt được trên 85%

Trong tương lai sẽ tìm ra nguyên nhân cụ thể dẫn đến độ chính xác không được cao và sẽ cải thiện.

Các phương pháp cải thiện độ chính xác:

* **Điều chỉnh thông số Hyper-Parameter**: Hyper-parameters là các biến được đặt trước khi train và xác định cấu trúc mạng và cách mạng được train. (ví dụ như learning rate, bath size, num of epochs).
* **Cải thiện tiền xử lý dữ liệu**: Việc cải thiện tiền xử lý phụ thuộc vào tập dữ liệu của bạn như xóa một số ký tự đặc biệt, ...
* **Sử dụng**[**Dropout Layer**](https://keras.io/layers/core/#dropout)**:**Đây là kỹ thuật thường được dùng để tránh overfitting do đó có thể cải thiện được mô hình sử dụng.

**TÀI LIỆU THAM KHẢO**

1. [Báo cáo phân loại văn bản sử dụng CNN, RNN & HAN (Phần 1: CNN) - Noron.vn](https://www.noron.vn/post/bao-cao-phan-loai-van-ban-su-dung-cnn-rnn--han-phan-1-cnn-1g62w4nxcq9)
2. [Báo cáo phân loại văn bản sử dụng CNN, RNN & HAN (Phần 2: RNN & HAN) - Noron.vn](https://www.noron.vn/post/bao-cao-phan-loai-van-ban-su-dung-cnn-rnn--han-phan-2-rnn--han-1g62w4nxcr5?ref=97249675912513823)
3. [Sự khác biệt cơ bản giữa CNN và RNN là gì? (qastack.vn)](https://qastack.vn/ai/4683/what-is-the-fundamental-difference-between-cnn-and-rnn?fbclid=IwAR3nAzw0F8gVtzuyUFtG1u7yyni1tdeV4Zy1jKO7SaTpxs0Km68to5SFcb4)
4. [Multi-Class Text Classification with LSTM | by Susan Li | Towards Data Science](https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17)
5. [[RNN] LSTM là gì? (dominhhai.github.io)](https://dominhhai.github.io/vi/2017/10/what-is-lstm/?fbclid=IwAR1BBGFFgEUuJ61fbwfN1bSxpvL-q6VWCLL4QXEZO2Ue5OUNo2CfiEf9NuA#3-m%E1%BA%A1ng-lstm)
6. [Text Classification With Python and Keras | Kaggle](https://www.kaggle.com/sanikamal/text-classification-with-python-and-keras)
7. [Text classification with an RNN  |  TensorFlow Core](https://www.tensorflow.org/tutorials/text/text_classification_rnn)

PHỤ LỤC

TỰ ĐÁNH GIÁ

(Với nhóm có 2 thành viên)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Phần | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| I (1) | Phân tích yêu cầu của bài toán | 1 | 1 |  |
| *II (3.25)* | Phương pháp giải quyết bài toán | | | |
|  | **2.1 Các phương pháp giải quyết bài toán** | 1 | 1 |  |
|  | **2.2 Phương pháp đề xuất giải quyết bài toán** | 0.25 | 0.25 |  |
|  | **2.3 Phương pháp giải quyết bài toán** | 2 | 1.5 |  |
| III (4.75) | Thực nghiệm | | | |
|  | **3.1 Dữ liệu** | 0.5 | 0.5 |  |
|  | **3.2 Xử lý dữ liệu** | 1 | 1 |  |
|  | **3.3 Công nghệ sử dụng** | 0.25 | 0.25 |  |
|  | **3.4 Cách đánh giá** | 0.5 |  |  |
|  | **3.5 Kết quả đạt được** | 2.5 | 0.5 |  |
| IV (0.25) | **Kết luận** | 0.25 |  |  |
|  | Điểm nhóm | 0.5 | 0.5 |  |
|  | Báo cáo theo mẫu | 0.25 | 0.25 |  |
| Tổng điểm đạt được | | 6.5 | |  |